Machine Learning Developments in Dependency Modelling and Feature Extraction

Dorota Toczydlowska

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
of
University College London

Department of Statistical Science
University College of London

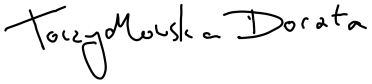
December 16, 2018

To my Grandma, Anna Stupkiewicz.

Dla mojej Babci, Anny Stupkiewicz.

Statement of Originality

I, Dorota Toczydlowska, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.



Acknowledgements

I would like to express my sincere appreciation to all the people who accompanied me, even for a moment, in this amazing yet challenging journey.

First and foremost, I would like to thank my supervisor, prof. Gareth W. Peters. I am deeply grateful for his generosity in guidance, support and never-ending patience. I consider myself very fortunate to have met such an outstanding person as Gareth and been under his supervision as a young researcher. His constant encouragement, motivation and belief, as well as selflessness in sharing his exceptional scientific knowledge and contagious passion for research, contributed greatly to both my professional and personal growth.

I gratefully acknowledge the funding received for my research studies from the UK PhD Centre in Financial Computing. I thank prof. Philip Treleaven, the director of the UK PhD Centre for Financial Computing for giving me an opportunity to start the PhD programme at University College London and for introducing me to many industrial partnerships and projects during my stay in London. I greatly appreciate his patience and encouragement during my PhD studies.

I am truly grateful to all the research environments I have been exposed to during my PhD. Every interaction with the passionate and knowledgeable individuals I have met has refined my knowledge, broadened my research interests and boosted creativity. I thank Eduardo Epperlein and my colleagues from Risk Methodology at Nomura International for providing me with an opportunity to dig into the problems of financial markets in a friendly working atmosphere and the additional funding for my PhD research. I would like to thank prof. Tomoko Matsui at the Institute of Statistical Mathematics (ISM) for her ongoing collaboration and opportunities to visit ISM. I benefited greatly from the hospitability of prof. Tomoko Matsui and the remarkable scientific environment of my visits. I want to express my sincere gratitude to prof. Pavel Shevchenko at Macquire University for his guidance and research discussions, especially while writing my very first publication.

I would like to thank my family for all their immeasurable love and support over these years. This accomplishment would not have been possible without them. I am especially thankful to my mother for all her calming words and for teaching me patience.

Last, but not least, I am thankful for the great joys and happiness brought to me by my wonderful friends. I thank them for all the patience, support and belief during these years as well as many excuses to laugh to tears.

Abstract

Three complementary feature extraction approaches are developed in this thesis which addresses the challenge of dimensionality reduction in the presence of multivariate heavy-tailed and asymmetric distributions.

First, we demonstrate how to improve the robustness of the standard Probabilistic Principal Component Analysis by adapting the concept of robust mean and covariance estimation within the standard framework. We then introduce feature extraction methods that extend the standard Principal Component Analysis by exploring distribution-based robustification. This is achieved via Probabilistic Principal Component Analysis (PPCA), in which new, statistically robust variants are derived, also treating missing data.

We propose a novel generalisation to the t-Student Probabilistic Principal Component methodology which (1) accounts for asymmetric distribution of the observation data, (2) is a framework for grouped and generalised multiple-degree-of-freedom structures, which provides a more flexible framework to model groups of marginal tail dependence in the observation data, and (3) separates the tail effect of the error terms and factors. The new feature extraction methods are derived in an incomplete data setting to efficiently handle the presence of missing values in the observation vector. We discuss statistical properties of their robustness.

In the next part of this thesis, we demonstrate the applicability of feature extraction methods to the statistical analysis of multidimensional dynamics. We introduce the class of Hybrid Factor models that combines classical state-space model formulations with incorporation of exogenous factors. We show how to utilize the information obtained from features extracted using introduced robust PPCA in a modelling framework in a meaningful and parsimonious manner.

In the first application study, we show the applicability of robust feature extraction methods in the real data environment of financial markets and combine the obtained results with a stochastic multi-factor panel regression-based state-space model in order to model the dynamic of yield curves, whilst incorporating regression factors. We embed the rank-reduced feature extractions into a stochastic representation of state-space models for yield curve dynamics and compare the results to classical multi-factor dynamic Nelson-Siegel state-space models. This leads to important new representations of yield curve models that can have practical importance for addressing questions of financial stress testing and monetary policy interventions which can

efficiently incorporate financial big data. We illustrate our results on various financial and macroeconomic data sets from the Euro Zone and international markets.

In the second study, we develop a multi-factor extension of the family of Lee-Carter stochastic mortality models. We build upon the time, period and cohort stochastic model structure to include exogenous observable demographic features that can be used as additional factors to improve model fit and forecasting accuracy. We develop a framework in which (a) we employ projection-based techniques of dimensionality reduction that are amenable to different structures of demographic data; (b) we analyse demographic data sets from the patterns of missingness and the impact of such missingness on the feature extraction; (c) we introduce a class of multi-factor stochastic mortality models incorporating time, period, cohort and demographic features, which we develop within a Bayesian state-space estimation framework. Finally (d) we develop an efficient combined Markov chain and filtering framework for sampling the posterior and forecasting. We undertake a detailed case study on the Human Mortality Database demographic data from European countries and we use the extracted features to better explain the term structure of mortality in the UK over time for male and female populations. This is compared to a pure Lee-Carter stochastic mortality model, demonstrating that our feature extraction framework and consequent multi-factor mortality model improves both in-sample fit and, importantly, out-of-sample mortality forecasts by a non-trivial gain in performance.

Impact Statement

The research presented in thesis provides benefits both inside and outside academia and contributes in methodological and application-related developments which address important questions of the financial and insurance markets.

In the era of big data, probabilistic models continue to grow in prominence in these sectors, as they are seen to be useful techniques in machine learning, when interpretation of the results obtained plays a crucial role in the modelling process. The information contained in big datasets is of key importance for efficient investment decision-making, risk management procedures, such as effective risk mitigation and risk transfer, reliable insurance provision, stress testing, or understanding the influence of unconventional monetary policies.

Therefore, the methods which study such data for real-life applications should acknowledge the methodological developments. We extend existing feature extraction techniques by linear projections in a probabilistic setting which are particularly relevant when one wishes to apply feature extraction in contexts which require one to ensure that both interpretation and quantification of the probabilistic confidence of the results are accurate for practical applications.

We illustrate the applicability of the investigated methods by two case studies. The first study, published in (Toczydlowska and Peters 2018), provide a coherent approach which uses global macroeconomic and financial market data sets in the analysis of one of the core reference interest rates, the London Interbank Offered Rate (Libor). We show how to measure Euro-Zone market conditions and their decomposition into influences from the domestic and international markets.

In our second study, published as (Toczydlowska et al. 2017), we investigated various ideas for extracting meaningful information from the demographic data of European countries and applying them as explanatory variables to mortality estimation and forecasting of British female log mortality data. The new model provides both the improvement of the forecast and, what is more, the valuable information about the mortality forces which affect the British population.

List of Publication

Stochastic Period and Cohort Effect State-Space Mortality Models Incorporating Demographic Factors via Probabilistic Robust Principal Components, D. Toczydlowska, G. W. Peters, M. Ch. Fung and P. V. Shevchenko, *Risks*, 5(3), 42, 2017;

Financial Big Data Solutions for State Space Panel Regression in Interests Dynamics, D. Toczydlowska, G.W. Peters, *Econometrics* , 6(3), 34, 2018;

Parsimonious Feature Extraction Methods: Extending Robust Probabilistic Projections with Generalized Skew t-Student Distribution, D. Toczydlowska, G. W. Peters and P. V. Shevchenko, in preparation, 2019;

Finance of Green Bonds: Quantitative and Qualitative Aspects, G.W. Peters, D. Toczydlowska, V. Zhang, G. Bagnarosa and A. Macrina, *Springer*, in preparation, early 2019;

Contents

1 Introduction **33**

1.1 Motivations . . . 33

1.2 Related Works . . . 36

1.2.1 Feature Extraction and Extensions to the Classical Principal Component Analysis . . . 36

1.2.2 State Space Models and Filtering Techniques . . . 42

1.3 Contributions . . . 56

1.3.1 Methodology Related Contributions . . . 56

1.3.2 Application Related Contributions . . . 58

1.4 Thesis Structure . . . 59

I Parsimonious Feature Extraction via Probabilistic Principal Component Analysis **61**

2 Introduction to Probabilistic Principal Component Analysis **65**

2.1 Principal Component Analysis Revisited: Non-Stochastic Principal Component Analysis . . . 65

2.2 Extending Principal Component Analysis to the Probabilistic Setting . . . 67

2.3 Extending Stochastic Principal Component Analysis to Factor Analysis . . . 67

2.3.1 Link to Principal Component Analysis . . . 69

2.3.2 PCA as a Limiting Case of Factor Analysis . . . 69

2.4 Probabilistic Principal Component Analysis with Gaussian Distribution . . . 69

2.5 Introduction to the Expectation-Maximisation Algorithm . . . 70

2.5.1 The Convergence Types for the Expectation-Maximisation Algorithm . . . 73

2.6 The Application of Expectation-Maximisation Algorithm to Probabilistic Principal Component Analysis . . . 75

2.6.1 Expectation step and its maximum . . . 76

2.6.2 The maximum likelihood estimation - the convergence of EM algorithm . . . 77

3 Robust Probabilistic Principal Component Analysis **81**

3.1 Robust Gaussian Probabilistic Principal Component Analysis: Utilising the Concept of Robust Estimation . . . 82

3.1.1 Robust Estimators for the Mean and Covariance Matrix . . . 82

3.1.1.1 Concept of robustness. . . . 83

3.1.1.2 M-estimators . . . 84

3.1.1.3 S-estimators . . . 89

3.1.1.4 Additional Comments on Robust Estimators 91
3.2 Extension 1: t-Student Probabilistic Principal Component Analysis 93
3.2.1 The t-Student Principal Component Analysis Model 93
3.2.2 Complete Data EM Algorithm for t-Student Probabilistic Principal Component Analysis 94
3.2.2.1 EM Algorithm for Probabilistic Principal Component with Independent t-Student Distribution 95
3.2.2.2 EM Algorithm for Probabilistic Principal Component with Identical and Conditionally Independent t-Student Distribution 97
3.2.3 Remark On Links Between **W** and Population Covariance in t-Student PPCA 98
3.2.4 The Statistical Robustness of t-Student PPCA 100
3.3 Case Study: Comparison of Probabilistic Principal Component Analysis Frameworks on Synthetic Data 102
3.3.1 Testing PPCA Algorithms For Non-Perturbed Sample 103
3.3.2 Testing Robustness of PPCA Algorithms For the Perturbed Sample 109
3.4 Extension 2: Generalized Skew t-Student Probabilistic Principal Component Analysis 115
3.4.1 Probabilistic Principal Component Analysis with Generalized Skew t-Student Distribution 116
3.4.1.1 The model of Independent Generalized Skew t-Student Probabilistic Principal Component 116
3.4.1.2 The model of Identical and Conditionally Independent Generalized Skew t Student Probabilistic Principal Component Analysis 117
3.4.1.3 Special Cases: Probabilistic Principal Component Analysis with Skew t-Student Distribution and Grouped t-Student Distribution . 118
3.4.2 Complete Data EM Algorithm for Generalized Skew t-Student Probabilistic Principal Component 120
3.4.2.1 Expectation-Maximisation Algorithm for Probabilistic Principal Component Analysis with Independent Generalized Skew t-Student Distribution 121
3.4.2.2 Expectation-Maximisation Algorithm for Probabilistic Principal Component Analysis with Identical and Conditionally Independent Generalized Skew t-Student Distribution 129
3.5 The Optimisation Steps of EM Algorithm for Generalized Skew t-Student Probabilistic Principal Component Analysis 136
3.5.1 Approximated Solutions to the Integrals in E-step of EM Algorithm for Generalized Skew t-Student Probabilistic Principal Component 136
3.5.2 Optimisation Framework for Iterative Root Search in M-step of EM Algorithm for Generalized Skew t-Student Probabilistic Principal Component Analysis 140
3.6 Generalized Expectation-Maximisation Algorithm 142

3.7 Case Study: Investigating the Imapct of Model Assumptions on Principal Components 145
3.7.1 The Misspecification of the Sample Covariance Matrix 145
3.7.2 Performance of the PPCA frameworks 152
3.7.2.1 The Estimation of the Static Parameters of the Model in Equation (2.7) Under Various Model Assumptions. 153
3.7.2.2 Results of the Model Selection for the Degrees of Freedom: t-Student IND PPCA and Special Case 3 GStS IND PPCA 158

4 Robust Probabilistic Principal Component Analysis in Presence of Missing Data 161
4.1 Novel Extension of Probabilistic Principal Component Analysis for Missing Data Setting 161
4.2 Extending EM Algorithm for Gaussian Probabilistic Principal Component Analysis to Incomplete Data Setting 162
4.3 Incomplete Data EM Algorithm for t-Student Probabilistic Principal Component Analysis 163
4.3.1 EM Algorithm for Probabilistic Principal Component with Independent t-Student Distribution in the Presence of Missing Data 164
4.3.2 EM Algorithm for Probabilistic Principal Component with Identical and Conditionally Independent t-Student Distribution in the Presence of Missing Data 165
4.4 Incomplete Data EM Algorithm for Generalized Skew t-Student Probabilistic principal Component Analysis 167
4.4.1 EM Algorithm for Probabilistic Principal Component Analysis with Independent Generalized Skew t-Student Distribution in the Presence of Missing Data 167
4.4.1.1 EM Algorithm for Special Cases of Probabilistic Principal Component with Independent Generalized Skew t-Student Distribution in the Presence of Missing Data 169
4.4.2 EM Algorithm for Probabilistic Principal Component Analysis with Identical and Conditionally Independent Generalized Skew t-Student Distribution in the Presence of Missing Data 175
4.4.2.1 EM Algorithm for Special Cases of Probabilistic Principal Component with Identical and Conditionally Independent Generalized Skew t-Student Distribution in the Presence of Missing Data 177

II State Space Models 183

5 Hybrid Factor Model 187
5.1 Model Assumptions 187

6 Optimal Filtering Problem **190**
6.1 Optimal Filtering Problem Statement 190

7 Kalman Filter and Gaussian Linear State Space Models **193**
7.1 Recursions of Kalman Filter for Gaussian Linear State-Space Models 193
7.1.1 Irregular dimensionality of the observation vector 196
7.2 Statistical Properties & Key Interpretations 197
7.2.1 Minimum Mean Square Error Estimation by Kalman Filter for Gaussian Linear Models 197
7.2.2 Kalman Filter as Minimum Mean Square Linear Filter for Non-Gaussian Data 198
7.2.3 The Interpretation of the Kalman Gain 198
7.2.4 Main Characteristic of Dynamic Systems 199
7.3 Forecasting via Kalman Filter 200
7.4 Smoothing via Kalman Filter 201
7.5 Estimation of Static Parameters via Marginalized Likelihood for Linear Gaussian State-Space Models 203

8 Bayesian Inference for Linear Gaussian State-Space Models **206**
8.1 Introduction to Bayesian Inference 206
8.2 Bayesian Model Development and Inference for Linear Gaussian State-Space Models 208
8.2.1 Combining Gibbs Sampler and Kalman Filter: Forward-Backward Filtering for Latent State Dynamics 208
8.2.2 Forecasting via FFBS 210

9 In-sample Model Assessment of the Hybrid Factor Model **212**

10 Monte Carlo Methods for State-Space Models **214**
10.1 Classical Monte Carlo Problem 214
10.1.1 Importance Sampling 215
10.1.2 Rejection Sampling 216
10.2 Markov Chain Monte Carlo 217
10.2.1 Metropolis-Hastings Algorithm 218
10.2.2 Gibbs Sampler Algorithm 220

III Applications **223**

11 Financial Big Data Solutions for State Space Panel Regression in Interest Rates Dynamics **225**
11.1 Introduction 225
11.1.1 Multifactor Models For Yield Curve Dynamics 226
11.1.2 Feature Extraction Methods For Financial Data 227
11.1.3 Contributions & Structure 228

11.2 International Macroeconomic and Financial Big Data Sets . . . 229
11.3 Multifactor Model with Macroeconomic Factors for EUR Libor Yield Curve 230
11.3.1 The dynamic Nelson-Siegel model . . . 231
11.3.2 Extending the Dynamic Nelson-Siegel to Macroeconomic Factor Model . . 232
11.3.3 Estimation based on the Kalman Filter . . . 234
11.3.4 Model Selection Methodology For Nelson-Siegel Class of Models . . . 235
11.4 Feature Extraction using Cross-Country Macroeconomic Data Sets . . . 235
11.4.1 Features Extraction for Country-Specific Sovereign Yield Curves . . . 237
11.4.2 Features Extraction for Country-Specific Inflation-Linked Yield Curves . . . 241
11.4.3 Features Extraction for Macroeconomic Proxies of Euro-Zone Activity . . . 244
11.4.4 Simmilarity of Extracted Features Across Various Financial Data Sets 245
11.5 Hybrid Multi-Factor Dynamic Nelson-Siegel Yield Curve Model of Euro Libor . . 248
11.5.1 The Specification of the Optimal Models for Eur Libor Yield Dynamic Analysis 248
11.5.2 The Comparison of the Baseline Model Selection Methodologies . . . 250
11.5.3 In-Sample Fit of the Best Models for Euro Libor Yield Curve . . . 251
11.5.3.1 The Calibration of the Euro Libor Yield Curve using 1Y ICE Euro Libor or 1Y swap rates . . . 253
11.5.3.2 Filtering of Latent Variables . . . 256
11.5.4 Forecasting Performance of the Extended Nelson-Siegel Models with Macroeconomic Factors . . . 259

12 Stochastic Period and Cohort Effect State-Space Mortality Models Incorporating Demographic Factors 265
12.1 Motivations and Related Works . . . 265
12.2 Period and Cohort Effect Stochastic Mortality Models: State-Space Formulation . . 267
12.3 Demographic Factor Model Extension to the Period-Cohort Stochastic Mortality State-Space Models . . . 270
12.4 Data . . . 273
12.4.1 Preprocessing of Data . . . 274
12.4.2 Missing Data . . . 274
12.5 Feature Extraction from European Demographic Data via Probabilistic Principal Component Analysis . . . 275
12.5.1 The Assessment of the Methodology . . . 275
12.5.2 Births . . . 276
12.5.3 Life Expectancy at Birth . . . 279
12.5.4 Deaths . . . 283
12.5.5 Death Rates . . . 288
12.6 Bayesian Modelling and Sampling of Demographic Factor Model Extension to the Period-Cohort Stochastic Mortality State-Space Models . . . 292
12.6.1 Bayesian Model Development and Inference for Stochastic Mortality Models 293

12.6.2 Forward-Backward Filtering for Latent State Dynamics . . . 293
12.6.3 Posteriors for Static Parameters in the Cohort Model . . . 294
12.6.4 Posteriors for Static Parameters in the Extended Models . . . 295
12.6.5 Application of the Constraints . . . 296
12.7 Stochastic Mortality Models for UK utilizing Factor Extraction from European Demographic Data . . . 296
12.7.1 Description of the Models . . . 297
12.7.2 Setup . . . 298
12.7.3 Estimation of Static Parameters . . . 299
12.7.4 Filtering of Latent Variables . . . 301
12.7.5 In-Sample and Out-of-Sample Performance . . . 307
12.7.5.1 Comparison of the Models . . . 308
12.8 Additional Remarks on Modelling and Forecasting Results . . . 308
12.8.1 The Affect of the Stratification and Identification Constraints on Estimation of Stochastic Lee-Carter Type Models . . . 309
12.8.1.1 The Estimates of the Static Parameters and Filtered Latent Variables 311
12.8.2 The Estimates of the Intercept for the Factor Model DFM-PC-D-r . . . 314

IV Conclusion **317**

13 Conclusion **319**
13.1 Probabilistic Principal Component Analysis and Its Estimation Framework Conclusions . . . 319
13.2 Application-related Conclusions . . . 321
13.2.1 Modelling Yield Curves Using International Exogenous Marco & Financial Factors . . . 321
13.2.2 Incorporating Demographic Information into Mortality Model . . . 322
13.3 Future Extensions . . . 323

References **325**

Appendix **338**

A Proofs of Steps of EM Algorithms **341**
A.1 Supporting Lemmas . . . 341
A.2 Gaussian Probabilistic Principal Component Analysis . . . 341
A.2.1 Proof of Theorem 4 . . . 341
A.2.2 Proof of Theorem 5 . . . 342
A.2.3 Proof of Theorem 19 . . . 343
A.2.4 Proof of Theorem 20 . . . 344

A.3 t-Student Probabilistic Principal Component Analysis . . . 345
A.3.1 Proof of Theorem 9 . . . 345
A.3.2 Proof of Theorem 10 . . . 346
A.3.3 Proof of Theorem 11 . . . 348
A.3.4 Proof of Theorem 12 . . . 350
A.3.5 Proof of Theorem 21 . . . 353
A.3.6 Proof of Theorem 22 . . . 354
A.3.7 Proof of Theorem 23 . . . 356
A.3.8 Proof of Theorem 24 . . . 356
A.3.9 Supporting Lemmas . . . 359
A.4 Generalized Skew t-Student Probabilistic Principal Component Analysis . . . 366
A.4.1 IND: Complete Data Setting . . . 366
A.4.1.1 Proof of Theorem 13 . . . 366
A.4.1.2 Proof of Corollary 2 . . . 368
A.4.1.3 Proof of Theorem 14 . . . 369
A.4.1.4 Proof of Corollary 4 . . . 370
A.4.1.5 Proof of Corollary 6 . . . 374
A.4.2 IND: Incomplete Data Setting . . . 377
A.4.2.1 Proof of Theorem 25 . . . 377
A.4.2.2 Proof of Corollary 18 . . . 379
A.4.2.3 Proof of Theorem 26 . . . 381
A.4.2.4 Proof of Corollary 15 . . . 382
A.4.2.5 Proof of Corollary 19 . . . 387
A.4.3 ID: Complete Data Setting . . . 391
A.4.3.1 Proof of Theorem 15 . . . 391
A.4.3.2 Proof of Theorem 16 . . . 392
A.4.3.3 Proof of Corollary 11 . . . 393
A.4.3.4 Proof of Corollary 13 . . . 395
A.4.4 ID: Incomplete Data Setting . . . 398
A.4.4.1 Proof of Theorem 27 . . . 398
A.4.4.2 Proof of Theorem 28 . . . 399
A.4.4.3 Proof of Corollary 22 . . . 400
A.4.4.4 Proof of Corollary 26 . . . 403
A.4.5 Supporting Lemmas for Generelized Skew t-Student IND PPCA . . . 405
A.4.6 Supporting Lemmas for Generalized Skew t-Student ID PPCA . . . 411
A.4.7 Proof of Proposition 1 . . . 415

B Supplementary Material to Chapter 11 **420**
B.1 Data Description . . . 420
B.1.1 Euro Libor Yield Curve . . . 420

B.1.2 Country-Specific Sovereign Yield Curves 421
B.1.3 Measures of Inflation Risk 424
B.1.3.1 Country-Specific Consumer Price Index 425
B.1.3.2 Country-Specific Inflation-Linked Yield Curves 426
B.1.4 Country-Specific Productivity Proxies 429
B.1.5 Euro-Zone Liquidity & Credit Proxies and Foreign Exchange 430
B.2 Supplementary Material to Section 11.4 432
B.3 Model Fit Statistics 432

C Supplementary Material to Chapter 12 **437**
C.1 Observed Patterns of Missingness 437
C.1.1 Births 437
C.1.2 Deaths 439
C.2 Description of Stochastic Mortality Models Utilizing Factor Extraction from European Demographic Data 441
C.2.1 DFM-PC-B Model 442
C.2.2 The Models of the Class DFM-PC-D and DFM-PC-Mx 442

List of Figures

3.1 The median of the loglikelihood functions of EM algorithm for: Gaussian PPCA , robust Gaussian PPCA, t-Student IND PPCA and t-Student IID PPCA for different dimensionality of the observation (d) and sample size (N) and proportion of missing values in sample set (*missing*). 104

3.2 The logarithm of the squared Frobenius norm of the difference between the three first eigen vectors of the true matrix $\mathbf{C} = \mathbf{W}^T\mathbf{W} + \sigma^2\mathbf{I}_d$ and its estimators $\hat{\mathbf{C}} = \hat{\mathbf{W}}^T\hat{\mathbf{W}} + \hat{\sigma}^2\mathbf{I}_d$ given by by EM algorithm for: Gaussian PPCA, Robust Gaussian PPCA, t-Student IND PPCA and t-Student IID PPCA for different dimensionality of a single observation (d) and the sample size (N) and the proportion of the missing values in the sample set (*missing*). 104

3.3 The logarithm of the squared Frobenius norm of the difference between the true value of the vector of means $\boldsymbol{\mu}$ and its estimator given by by EM algorithm for: Gaussian PPCA , Robust Gaussian PPCA, t-Student IND PPCA and t-Student IID PPCA for different dimensionality of a single observation (d) and the sample size (N) and the proportion of the missing values in the sample set (*missing*). 105

3.4 The median of the loglikelihood function of the EM algorithm for t-Student IND PPCA across the grid of degrees of freedom over different dimensionality of the observations (d) and the sample size (N) and the proportion of the missing values in the sample set. 107

3.5 The median of the loglikelihood function of the EM algorithm for t-Student IID PPCA across the grid of degrees of freedom over different dimensionality of the observations (d) and the sample size (N) and the proportion of the missing values in the sample set. 107

3.6 The median logarithm of the likelihood function of EM algorithm for robust Gaussian PPCA across the grid of degrees of freedom over different dimensionality of the observation (d) and sample size (N) and proportion of missing values in sample set (*missing*). 108

3.7 The logarithm of the squared Forbenious norm of the difference between the three first eigen vectors of the true matrix $\mathbf{C} = \mathbf{W}^T\mathbf{W} + \sigma^2\mathbf{I}_d$ and its estimate $\hat{\mathbf{C}} = \hat{\mathbf{W}}^T\hat{\mathbf{W}} + \hat{\sigma}^2\mathbf{I}_d$ given by the following EM algorithms: Gaussian PPCA, Robust Gaussian PPCA, t-Student IND PPCA and t-Student IID PPCA for different dimensionality of a single observation (d) and the sample size (N) and the proportion of the sample which is perturbed with the Laplace noise and the proportion of the missing values in the sample set. 111

3.8 The logarithm of the squared Forbenious norm of the difference between the true vector of means μ and its estimator given by the following EM algorithms: Gaussian PPCA, Robust Gaussian PPCA, t-Student IND PPCA and t-Student IID PPCA for different dimensionality of a single observation (d) and the sample size (N) and the proportion of the sample which is perturbed with the Laplace noise and the proportion of the missing values in the sample set. 112
3.9 The optimal breakdown points for **robust Gaussian PPCA** across different proportions of perturbed samples with the Laplace noise over different dimensionality of the observation (d) and sample size (N) and the proportion of the missing values in the sample set. 113
3.10 The optimal degrees of freedom for **t-Student IND PPCA** across different proportions of perturbed samples with the Laplace noise over different dimensionality of the observation (d) and sample size (N) and the proportion of the missing values in the sample set. 113
3.11 The optimal degrees of freedom for **t-Student IID PPCA** across different proportions of perturbed samples with the Laplace noise over different dimensionality of the observation (d) and sample size (N) and the proportion of the missing values in the sample set. 114
3.12 The lower triangle of the sample covariance matrix of the data which is simulated from the models **M0**-**M3** (row-wise) with assumed different degrees of freedom. . 148
3.13 The lower triangle of the sample covariance matrix of the data which is simulated from the model **M4**. 149
3.14 The discrepancies between the first five eigenvectors of the sample covariance matrix of the models **M1**-**M3** and the corresponding eigenvectors given by the model **M0**. 150
3.15 The the decimal logarithm ($\log_{10}$) of the subsequent eigenvalues (x-axis) of the sample covariance matrix provided by the models **M0**-**M3** with assumed different degrees of freedom. 151
3.16 The the decimal logarithm ($\log_1 0$) of the subsequent eigenvalues (x-axis) of the sample covariance matrix provided by the models **M0**-**M3** with assumed different degrees of freedom. 152
3.17 The average discrepancies with 95% empirical quantiles (the red lines) between the true μ and its estimate obtained by three PPCA frameworks 154
3.18 The average discrepancies with 95% empirical quantiles (the red lines) between the true μ and its estimate obtained by three PPCA frameworks 155
3.19 The average discrepancies with 95% empirical quantiles (the red lines) between the first singular vector of the true matrix $\mathbf{W}$ and the first singular vector of its estimate obtained by three PPCA frameworks 155
3.20 The average discrepancies with 95% empirical quantiles (the red lines) between the first singular vector of the true matrix $\mathbf{W}$ and the first singular vector of its estimate obtained by three PPCA frameworks 156

3.21 The average discrepancies with 95% empirical quantiles (the red lines) between the second singular vector of the true matrix **W** and the second singular vector of its estimate obtained by three PPCA frameworks 156
3.22 The average discrepancies with 95% empirical quantiles (the red lines) between the second singular vector of the true matrix **W** and the second singular vector of its estimate obtained by three PPCA frameworks 157
3.23 The average discrepancies with 95% empirical quantiles (the red lines) between the true σ^2 and its estimate obtained by three PPCA frameworks 157
3.24 The average discrepancies with 95% empirical quantiles (the red lines) between the true σ^2 and its estimate obtained by three PPCA frameworks 158
3.25 The histogram of chosen values for ν in t-Student IND PPCA given 20 simulations for different variants of the model**M3**. 159
3.26 The histogram of chosen values for ν in t-Student IND PPCA given 20 simulations for different variants of the model **M4** 159
3.27 The heatmap of chosen values for ν_ϵ and ν_x in Special Case 3 of the GStS IND PPCA given 20 simulations for different variants of the model**M3**. 160
3.28 The heatmap of chosen values for ν_ϵ and ν_x in Special Case 3 of the GStS IND PPCA given 20 simulations for different variants of the model **M4**. 160

11.1 The logarithms of the three highest eigenvalues of the covariance matrix and the optimal values of the degrees of freedom for t-Student PPCA frameworks over years for the data sets **D2** which consists of country-specific sovereign yields. 238
11.2 The first eigenvector of a covariance matrix for the **D2** set of country-specific sovereign yield curves across different years partitioned into country-specific subvectors and their term structure. 239
11.3 The first eigenvector of a covariance matrix for the **D2** data set of country-specific sovereign yield curves without South African or Brazilian yields in 2007 partitioned into country-specific subvectors and their term structure. 241
11.4 The first eigenvector of a covariance matrix for the **D3** set of inflation-linked mixed yield curves across different years partitioned into country-specific subvectors and their term structure. 242
11.5 The first eigenvector of a covariance matrix for the **D3b** set of inflation-linked bond-only yield curves across different years partitioned into country-specific subvectors and their term structure. 242
11.6 The first eigenvector of a covariance matrix for the **D3s** set of inflation-linked swap-only yield curves across different years partitioned into country-specific subvectors and their term structure. 243
11.7 The evolution of three highest eigenvalues of covariance matrix for **D3**, **D3b** and **D3s** data sets of country-specific inflation-linked yield curves over years (x axis). . 243

11.8 The evolution of degrees of freedom for **D3**, **D3b** and **D3c** data sets of country-specific inflation-linked yield curves over years (x axis). 243
11.9 The first eigenvector of a covariance matrix for of **D4** data set of Euro-zone activity proxies across different years partitioned into country-specific subvectors and their term structure. 244
11.10 The evolution of three highest eigenvalues of covariance matrix (a) and the evolution of degrees of freedom of t-Student PPCA framework (b) over years (x axis) for the data sets which consists of country-specific soveregin yield. 245
11.11 The RV coefficients between various factors over different years. 247
11.12 The logarithms of mean square errors (MSE) per year and tenor of best Nelson-Siegel and Extended Nelson-Siegel models. 254
11.13 The market observed yields of the Euro Libor rates at various maturities over time against the conditional mean of in-sample 1 step ahead prediction $\mathbf{y}_{t|t-1}$ with corresponding predictive 97.5% confidence intervals. 256
11.14 The conditional mean of filtered latent states $\boldsymbol{\beta}_{t|t}$ of the coefficient loadings with corresponding confidence intervals of the Nelson-Siegel models over time which are optimal in term of the AIC criterion. 257
11.15 The conditional mean of filtered latent states $\tilde{\boldsymbol{\beta}}_{t|t}$ with corresponding confidence intervals of the Extended Nelson-Siegel models over times which are optimal in term of the AIC criterion. 259
11.16 The logarithms of Mean Square Errors of 1 day prediction per year (column-wise) and tenor of optimal Nelson-Siegel (NS) and Extended Nelson-Siegel (ENS) models. 262
11.17 The logarithms of Mean Square Errors of 1 week prediction per year and tenor of optimal Nelson-Siegel(NS) and Extended Nelson-Siegel (ENS) models. 263
11.18 The logarithms of Mean Square Errors of 1 month prediction per year and tenor of optimal Nelson-Siegel (NS) and Extended Nelson-Siegel (ENS) models. 263
11.19 The Euro Libor rates at 1M, 3M, 1Y and 25Y maturities with the one step ahead out-of-sample prediction given by the conditional mean $\mathbf{y}_{T+s+1|T+s}$ with the corresponding predictive 97.5% confidence intervals over considered period. . . . 264

12.1 The Mahalanobias distances obtained using Probabilistic Principal Component Analysis (PPCA) for Females and Males Births over time. 277
12.2 The eigenvalues obtained using PPCA for Females and Males population of Number of Births for different percentages of maximal missing entries in rows. 278
12.3 The eigenvectors over the joint distribution of countries obtained using PPCA for Females in Births. 279
12.4 The scores over time obtained using PPCA for Females in Births. 279
12.5 The Mahalanobias distances obtained using PPCA for Females and Males of Life Expectancy at Birth over time. 280

12.6 The eigenvalues obtained using PPCA for Females and Males of Life Expectancy at Birth data for different percentages of maximal missing entries in rows. 281
12.7 The eigenvectors over the joint distribution of countries obtained using PPCA for female population of Life Expectancy at Birth. Every row of subfigure corresponds to different eigenvector. 281
12.8 The eigenvectors over the joint distribution of countries obtained using PPCA for male population of Life Expectancy at Birth. 282
12.9 The scores over time obtained using PPCA for female population of Life Expectancy at Birth. 282
12.10 The scores over the time obtained using PPCA for Males of Life Expectancy at Birth. 283
12.11 The Mahalanobias distances obtained using PPCA for Females and Males of Number of Death over time. 284
12.12 The eigenvalues of Deaths counts obtained using PPCA for Females and Males over different cases of missing entries (x axis). 285
12.13 The scores over time obtained using PPCA for female population of Number of Deaths. 285
12.14 The eigenvectors of Death counts (y axis) over age groups and countries obtained using PPCA for Females. 286
12.15 The eigenvectors of Death counts over age groups and countries obtained using PPCA for Males. 287
12.16 The scores over time obtained using PPCA for male population of Number of Deaths. 288
12.17 The Mahalanobias distances obtained using PPCA for female and male population of Death Rates over time. 289
12.18 The eigenvalues obtained using PPCA for female and male population of Death Rates for different percentages of maximal missing entries in rows. 289
12.19 The eigenvectors over the joint distribution of countries obtained using PPCA for female population of Death Rates. 290
12.20 The eigenvectors over the joint distribution of countries obtained using PPCA for male population of Death Rates. 291
12.21 The scores over time obtained using PPCA for female population of Death Rates. . 292
12.22 The scores over time obtained using PPCA for male population of Death Rates. . . 292
12.23 Bayesian posterior mean estimators with 95% posterior credible intervals for the estimation of the age-specific diagonal elements of the transition matrix $\boldsymbol{\Omega}$ (x axis) under DFM-PC-B. 301
12.24 Bayesian posterior estimators with 95% posterior credible intervals for the estimation of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ under different models (colours of lines) for British female mortality data (1922–2002). 302
12.25 Bayesian posterior mean estimators with 95% posterior credible intervals for the estimation of the diagonal elements of the transition matrix $\boldsymbol{\Omega}$ under DFM-PC-D-r, DFM-PC-D-s, DFM-PC-Mx-r and DFM-PC-Mx-s models. 302

12.26 The Bayesian posterior mean estimates with 95% posterior credible intervals for κ_t (upper panel) and cohort effect state process γ_t^0 (lower panel) under different models (colours of lines) for British female log death rates during 1922–2002. . . . 303
12.27 The Bayesian posterior mean estimates for $\boldsymbol{\varrho}_t$ across age groups (y axis) over time (x axis) under DFM-PC-B model for British female log death rates during 1922–2002. 303
12.28 The Bayesian posterior mean estimates for the cohort effect latent processes vector $\boldsymbol{\gamma}_t$ across age groups (y axis) over time (x axis) under different models for British female log death rates during 1922–2002. 304
12.29 The Bayesian posterior mean estimates for $\boldsymbol{\varrho}_t$ across countries over time under the models from the classes DFM-PC-D and DFM-PC-Mx for British female log death rates during 1922–2002. 306
12.30 The Bayesian posterior mean estimates with 95% posterior credible intervals for $\boldsymbol{\varrho}_t$ across countries (different panels) overtime (x axis) under the models from the classes DFM-PC-D and DFM-PC-Mx (colours of lines) for British female log death rates during 1922–2002. 307
12.31 Bayesian posterior estimators with 95% posterior credible intervals for the estimation of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ under different models (colours of lines) for British female mortality data (1922–2002). 313
12.32 The Bayesian posterior mean estimates with 95% posterior credible intervals for κ_t (**upper panel**) and cohort effect state process γ_t^0 (**lower panel**) under different models (colours of lines) for British female log death rates during 1922–2002. . . . 314
12.33 The Bayesian posterior mean estimates with 95% posterior credible intervals average over time for $\boldsymbol{\alpha} + \tilde{\mathbf{F}}_t \boldsymbol{\varrho}_t$ for DFM-PC-D-r. 315
12.34 The Bayesian posterior mean estimates of $\boldsymbol{\alpha} + \tilde{\mathbf{F}}_t \boldsymbol{\varrho}_t^T$ over time for DFM-PC-D-r. Colours of lines are related to the age groups (the elements of the vector $\boldsymbol{\alpha}$). 315
12.35 10-year out-of-sample forecasted log death (y axis) rates of different age groups (different subplots) under different models (colours of lines) with corresponding prediction intervals. Calibration period: 1922–2002 316

A1 The Euro Libor rates and the indicators of missing values over time. 421
A2 The sovereign yield curves of 12 countries listed in Table 11.1. 424
A3 The indicators of missing values for sovereign yields of 12 countries listed in Table 11.1. 425
A4 The proportion (%) of missing values present in the time series of sovereign bond rates across different maturities for a sample period 01/2006 - 12/2016 (2870 days excluding weekends and holidays). 426
A5 The percentage of the aggregated missing entries per year and the percentage of missing entries per year and country of yields across maturities from sovereign yields. 427

A6 The comparison of inflation-lined yield curves (%) obtained from inflation-linked government bonds or inflation-linked swap rates of 8 countries and Euro- zone listed in Table 11.1. . . . 428
A7 The proportion of missing values present in the time series of inflation-linked sovereign bond or swap rates across different maturities and countries for a sample period 01/2006 - 12/2016 (2'870 days excluding weekends and holidays). . . . 429
A8 The indicators of missing values of swap-based and sovereign yield curves comprised of inflation-linked rates at different maturities for countries listed in Table 11.1. . . . 429
A9 Daily Consumer Price Indexes, quarterly Gross Domestic Products, Unemployment rate in % and Labour Productivity of 12 countries listed in Table 11.1. . . . 430
A10 The raw and smoothed open interests of Euro-Bund futures contractsm 5Y Marikt Itraxx Index, the spread between 3 month Euribor and Eurorepo rates and the exchange rates against EUR for AUD, GBP and USA over time. . . . 431
A11 The yield curves of 12 countries from the data set **D2** for selected years across different maturities over time. . . . 432
A12 The Euro Libor yield curve with the one step ahead out-of-sample prediction given by the conditional mean $\mathbf{y}_{T+s+1|T+s}$ with corresponding 97.5% confidence intervals over examined period. . . . 433

A1 The percentage of missing entries per observation vector over time for the Births counts and Deaths counts for female and male population. . . . 438
A2 he indicator of a missing value per country over time for the Births counts for female population. . . . 439
A3 Percentages of missing values per observation for Number of Deaths for Females per age groups over time. . . . 440
A4 Percentages of missing values per observation for Number of Deaths for Males per age groups over time. . . . 440
A5 Percentages of missing values for Number of Deaths for Females per country and age group. . . . 441
A6 Percentages of missing values or Number of Deaths for Males per country and age group. . . . 441

List of Tables

11.1 The list of collected financial and macroeconomic time series investigated in the study. 230

11.2 The model specification and in-sample statistics for the best standard Nelson-Siegel model and the Extended Nelson-Siegel models chosen by **diagAll** per calendar year for the two sets of Euro Libor yield. 252

11.3 The forecasting performance of the the best Extended Nelson-Siegel and Nelson-Siegiel models for the Euro Libor yield curve with **1 year ICE Euro Libor rate** and **1 year swap rate** per year. 261

12.1 The availability of the demographic data per country (Human Mortality Database). 273

12.2 Bayesian posterior mean estimators with 95% posterior credible intervals for the estimation of the static parameters $\left(\lambda,\ \theta,\ \eta, \sigma_\epsilon^2,\ \sigma_\gamma^2,\ \sigma_\kappa^2, \sigma_\varrho^2\right)$ of $\log m_{x,t}$. 300

12.3 The MAP estimates of the static parameters $\left(\lambda,\ \theta,\ \eta, \sigma_\epsilon^2,\ \sigma_\gamma^2,\ \sigma_\kappa^2, \sigma_\varrho^2\right)$ of $\log m_{x,t}$. . . 301

12.4 Mean square error of the fit of the models to the data (MSE), deviance information criterion (DIC) and mean square errors of predictions using forecasting distributions given by MCMC samples (MSEP_{MCMC}) and Kalman Filter (MSEP_{Kalman}). 308

12.5 Mean square error of the fit of the models to the data (MSE) and mean square errors of predictions using forecasting distributions given by MCMC samples (MSEP_{MCMC}) and Kalman Filter (MSEP_{Kalman}). 311

A1 The availability and the Bloomberg identifiers (tickers) of Euro Libor yield curves sourced from Bloomberg. 420

A2 The availability and Bloomberg identifiers of the sovereign yield curves and inflation-linked yield curves sourced from Bloomberg. 421

A2 The availability and the reference to the Bloomberg terminal (tickers) of the sovereign yield curves and inflation-linked yield curves sourced from Bloomberg (cont.). . . 422

A2 The availability and the reference to the Bloomberg terminal (tickers) of the sovereign yield curves and inflation-linked yield curves sourced from Bloomberg (cont.). . . 423

A3 The availability and Bloomberg identifiers of the national macroeconomic measures of economic activity. 428

A4 The list of yearly features from the data sets **D2-D3** which are chosen as exogenous variables in the calibration of the Extended Nelson-Siegel model introduced in Section 11.3. 434

A5 The comparison of the optimal models selected using two employed strategies - by minimum AIC (minAIC) and the lowest complexity (diagAll) for the Euro Libor yield with **1Y swap rate**. 435

A6 The comparison of the optimal models selected using two employed strategies - by minimum AIC (minAIC) and the lowest complexity (diagAll) for the Euro Libor yield with **1Y ICE rate**. 436

1. Introduction

This chapter presents an overview of the thesis. Section 1.1 outlines the key motivations of the research conducted into extensions to Probabilistic Principal Component Analysis and their applications to the modelling of important financial problems. This is followed by Section 1.2, which reviews related literature on feature extraction concepts investigated in the thesis. It also provides an overview of the relevant background to the evolution of state-space models and their estimation techniques, which are used in the second part of this thesis to illustrate the applicability of the developed methods to addressing several practical problems from financial mathematics and insurance. We detail the research contributions of the studied problems in Section 1.3. The final part of this chapter reviews the structure of the thesis.

1.1. Motivations

The primary concern of the research presented in this thesis is to develop a range of different and flexible feature extraction methods that can each parsimoniously handle the characteristics of real data, such as various patterns of outliers, missing data or asymmetric distribution of data. The second objective is to examine the quality of the information extracted by parsimonious feature extraction methods from large datasets and its benefits in the modelling of fundamental market dynamics.

The increased collection of and access to complex financial and macroeconomic data is transforming the way in which the financial services operate. In the era of big data, probabilistic models continue to grow in prominence in this sector, as they are seen to be useful techniques in machine learning, when interpretation of the results obtained plays a crucial role in the modelling process. Therefore, the methods which study such data for real-life applications should acknowledge these methodological developments. Our work is novel in its focus on efficient and statistically robust feature extraction methods. This work deals with challenging, but not uncommon situations, where the data is noisy and does not provide complete information on every observation, i.e., the components of the observation set are not consistently available or randomly missing, censored or truncated in some manner.

The principal motivation of this thesis is to develop and investigate novel feature-extraction and dependence-modelling techniques in a probabilistic setting. The methods explored are particularly relevant when one wishes to apply feature extraction in contexts which require one to ensure that both interpretation and quantification of the probabilistic confidence of the results are accurate for practical applications. The investigated frameworks facilitate incorporation of prior assumptions about the data distribution into the model, tailoring the model to our knowledge, to ensure robust analysis of large incomplete datasets. The proposed approaches enjoy many encouraging properties which enable one to determine, interpret and respond to different interactions present in large datasets and which appear as a consequence of the data's complexity.

Furthermore, we wish to represent various sources of uncertainty present in the dataset, such as multiple patterns of noise related to the process of generating data, or too small a sample size. These characteristics can be explicitly taken into account when estimating the static parameters of a model. We focus on the linear latent variable models and investigate the extensions to the probabilistic paradigm of Principal Component Analysis formalised by (Tipping and Bishop 1999b) and (Roweis 1998).

In many applications, the stochastic noise or observation errors in the data set are assumed to be, in some sense, "well-behaved"; for instance, additive, light-tailed, symmetric and zero-mean. When non-robust feature extraction methods are naively utilised in the presence of violations of these implicit statistical assumptions, the information contained in the extracted features cannot be relied upon, resulting in misleading regression models and predictions. Therefore, it is critical to ensure that the feature extraction is appropriately performed. In this thesis, we relax the inherent assumption of "well-behaved" observation noise by developing a class of robust estimators that can withstand violations of such assumptions, which routinely arise in real data sets, especially in the crucial areas of demographic and mortality projections and financial interest-rate modelling.

These robust feature-extraction methods are primarily based on the assumption that observations are independent over time and that the marginal distributions of their components have the same profile of heavy tails. This reasoning might be criticised for having a limited ability to capture various tail-dependence patterns in multivariate data analysis. Therefore, we want to investigate an approach that will be able to accommodate a broader range of dependence and marginal distribution assumptions in the data-generating mechanism. This will include grouped and generalised multiple degrees of freedom structures within the family of Generalised, Skewed and Grouped Student's-t copula models, which provide a flexible framework for modelling groups of marginal tail dependence in the observation data.

The second motivation of this thesis is to incorporate the complexity present in large data sets into the stochastic modelling. We focus on three real-world applications: (1) the use of international macroeconomic and financial big datasets to study the effects of the global economy on the dynamics of the EUR Libor market, (2) assessing the influence of the global economy on the yield curve of green bonds, a novel class of financial instruments issued to support the funding of projects directly related to climate and environmental concerns; and (3) investigating the impact of global mortality trends given by various sets of international demographic data and their potential influence on the mortality experience of one country - the United Kingdom - in our case study. In particular, in this latter case, we focus on the ability to both predict and better explain mortality projections and life expectancy patterns in one population relative to many other populations.

The increasing volume of market data poses both an opportunity and a challenge for financial and insurance institutions to deepen their understanding of market-wide and country-specific sources of risk in different financial and insurance markets and their implications for modelling and forecasting market dynamics. Such information is of key importance for efficient investment decision-making, risk management procedures, such as effective risk mitigation and risk transfer, reliable insurance provision, stress testing, or understanding the influence of unconventional

monetary policies. Therefore, we utilise growing data sets, with increasing dimensionality and numbers of observations, to understand and reinterpret critical financial or demographic applications, such as interest rate dynamics or longevity rates.

Even though large datasets are now available, the naive incorporation of the multitude of available features into a stochastic model would result in far too many parameters to perform estimation. To illustrate this point, let us consider the case in which the size of the available observation set cannot justify the number of parameters needed in the model; such a situation would result in an overfitted model which utilises residual variation and consequently fails to provide sufficient accuracy of the out-of-sample prediction. Therefore, we employ the introduced class of probabilistic, statistically robust feature extraction approaches to reduce dimensionality and capture core information present in the datasets under examination. This gives the desired result of allowing the information to be more parsimoniously included in the models. The incorporation of features improves the predictive power of the models, and permits interpretation of the investigated market dynamics.

The first application discussed focuses on providing a coherent approach which utilises global macroeconomic and financial market data sets to model one of the core reference interest rates, the London Interbank Offered Rate (Libor). Until recently, the Libor rate was widely considered a risk-free interest rate, and is still used as a reference for pricing and hedging for the majority of financial instruments. Therefore, the rate acts as an important economic indicator. In order to use the information available in a range of financial data sets in a meaningful and parsimonious way to help explain the dynamics of Libor, we apply feature extraction, specifically tailored to handle irregular time series (missing data) and outliers.

However, the incorporation of large financial and macroeconomic datasets into the Libor rate stochastic model would lead to an increased possibility of over-parameterisation of the model. In addition, the noise present in these datasets, as well as various dependencies between them, may impair the modelling quality. We demonstrate how to use the derived market features to avoid these difficulties in the modelling process.

The second application study concerns modelling the “term-structure" of age-specific mortality rates by gender and country. It represents a vital topic in the current global environment, where projections of population demographics are increasingly important for pension planning, economic planning and insurance provisions – particularly in relation to ageing populations in developed countries. Most of the works on the topic, and the market practices, do not introduce large datasets into mortality rate modelling due to the excessive number of parameters to estimate compared to the limited sizes of the measurement sets. In order to address this issue, we develop a multi-factor extension of the stochastic mortality models. We build upon the time, period and cohort stochastic model structure to extend it to exogenous, observable, demographic features that can be used as additional factors and demonstrate our feature-extraction framework. Consequently, the multi-factor mortality model investigated produces a non-trivial gain in performance, both of the in-sample fit and, importantly, out-of-sample mortality forecasts.

1.2. Related Works

The objectives of this thesis split the conducted research into two distinct areas; feature-extraction methodologies, and state-space modelling inference to illustrate the application of these methodologies to important questions of relevance to financial and insurance markets. Therefore, we discuss the related works in two separate sections. Firstly, we review the essential literature on the evolution of Principal Component Analysis. Subsequently, we highlight the fundamental concepts of state-space models and the development of their learning algorithms.

1.2.1. Feature Extraction and Extensions to the Classical Principal Component Analysis

Principal Component Analysis (PCA), as defined in (Hotelling 1933) or (Jolliffe 2002), and related matrix factorisation methodologies are widely used in data-rich environments for dimensionality reduction, data compression, feature-extraction techniques or data de-noising. The methodologies identify a lower-dimensional linear subspace to represent the data, which captures second-order dominant information contained in high-dimensional data sets. PCA is limited to finding a lower-dimensional data representation only in linear subspaces. The general matrix representation of the problem assumes that the observation data can be decomposed into the sum of a lower-rank matrix and an additive noise matrix. Hence, PCA can be viewed as a matrix factorisation problem which aims to learn the lower-dimensional representation of the data, preserving its Euclidean structure. Since the distance in the Euclidean space is defined by the $\mathbf{L}_2$-norm, the error function between the data and its assumed representation is quadratic, and the classical optimisation problem of PCA is defined as follows

$$(\mathbf{U}, \mathbf{V}) = \underset{\mathbf{V},\mathbf{U}}{\operatorname{argmin}} ||\mathbf{Y} - \mathbf{U}\mathbf{V}^T||^2_{\mathbf{L}_2} \tag{1.1}$$

for $N \times d$ real-valued matrix $\mathbf{Y}$, $N \times k$ real-valued matrix $\mathbf{U}$ and $d \times k$ real-valued orthogonal projection matrix $\mathbf{V}$. Given the above formulation of the problem, the matrix $\mathbf{V}$ corresponds to the eigenvectors of the covariance matrix of the dataset $\mathbf{V}$, as explained in (Jolliffe 2002).

When looking at the PCA from a probabilistic perspective, the quadratic error function is obtained under the assumption of the additive Gaussian noise. The basis vectors, which span the lower-dimensional subspace, are the eigenvectors of the data covariance matrix. The probabilistic interpretation of PCA is broadly discussed and justified in the main body of the thesis.

PCA emerges as an optimal solution to many problems. Given the orthonormal constraints imposed on the linear transformation matrix, the methodology minimises the mean square error of the representation of the data, and maximises its linearly transformed variance. The lower-rank representation of the data provided by PCA is uncorrelated, which will imply independence only when data follows the Gaussian distribution. Also, under the Gaussian model of the data, the entropy of the representation obtained by PCA is minimised. Therefore, applying PCA to normally-distributed data will enjoy many optimal properties. Also, the characteristics of the quadratic error function of PCA, such as its smoothness and convexity, enable the PCA problem

to be solved by efficient methodologies, such as Singular Value Decomposition. Please refer to (Young and Calvert 1974) or (Devijver and Kittler 1987) for more details on the properties of the classical PCA.

In the presence of incomplete data matrix $\mathbf{Y}$, the problem in Equation (1.1) can be efficiently solved by the famous algorithm introduced by (Wiberg 1976). The Wiberg algorithm approaches the difficulty of factorising an incomplete data matrix under Euclidean norm assumption, by eliminating the estimation of either matrix $\mathbf{U}$ or matrix $\mathbf{V}$. The method assumes a mapping between the matrices and solves the optimisation problem in Equation (1.1) by a two-step procedure. The formulations and the numerical performance of the framework are analysed in (Okatani and Deguchi 2007).

However, in the presence of either a non-Gaussian distribution of the data, or in the presence of impurities or outliers which corrupt the data, the standard PCA methodology provides biased information about the lower-rank representation. In such a situation, the selection of an alternative dimensionality reduction methodology starts with identifying the characteristics of the data.

One of the most relevant characteristics of the data is the type of noise which is attached to a generative process of the data. We can categorise noises according to many aspects. The noise can be univariate or multivariate, additive or multiplicative; it can be serially independent or can exhibit a dependency structure between its occurrences. Furthermore, following the classification by (Zhou et al. 2010), one may classify the types of noises according to two other characteristics: its abundance (whether the noise has sparse or dense values) or its magnitude (whether the values of the noise are small or large).

In the latter, we focus on additive white noise; that is, the process added to the sample set, which is independent in time and has a zero-mean vector and some covariance matrix. Considering the other characteristics of the noise pattern, the standard PCA belongs to the class of methodologies which are efficient in the presence of a data corruption process which is dense and relatively small in proportion to the leading eigenvalues. It is a consequence of the quadratic loss function of the methodology. Therefore, it is only applicable to cases of so-called 'Gaussian' noise, as other types of perturbations significantly impact the derivation of principal vectors.

Multiple enhancements to the classical PCA methodology have been introduced to address the problem of different types of noise in the data. Robust variants of PCA modify the standard error function in Equation (1.1) to incorporate different noise assumptions and, consequently, they redefine the distance measure between each observation and its lower-rank approximation.

Since the standard problem of the PCA corresponds to finding the directions which maximise the covariance of the projected data as shown in (Jolliffe 2002), one group of improvements focuses on robustifying the calculation of the estimators of mean and covariance matrices. The standard approach to calculating the sample mean and covariance matrix, which is used in a standard estimation framework for PCA, such as SVD, can be significantly influenced by distant data points. Therefore the works of (Croux and Haesbroeck 2000),(Croux and Ruiz-Gazen 2005), (Hubert et al. 2005) or (Croux et al. 2007) among others, investigate the use of robust estimators as defined by

(Huber 1964),(Rousseeuw and Yohai 1984),(Rousseeuw 1985), (Tyler 1987b), in order to decrease the vulnerability of the covariance estimation.

The use of a robust estimator is, in fact, equivalent to introducing observation-specific weights to the loss function of the standard PCA. The majority of these approaches are based on down-weighting, or even removing, observations with outlying distances; this categorises them as a non-probabilistic set of methodologies, as they do not directly incorporate any assumption about the noise distribution.

The early study of (Xu and Yuille 1995) generalises the loss function in Equation (1.1) and introduces additional binary variables which discard observations labelled as outliers. The work of (Xu and Yuille 1995) is criticised for removing part of the sample set and losing information about the remaining elements of observations, which are not distant. The study of (De la Torre and Black 2002) investigates an M-estimation weighting algorithm for PCA, following the reasoning of the Generalized Weighted SVD framework with partial solutions given by (Greenacre 1984). The authors assume additional weights on elements of the squared error function of PCA, which leads to the formulation of the weighted least squares problem of PCA.

The concept of observation-specific weights is also addressed in (Park et al. 2004). The authors introduce Weighted PCA, where the weights, which are inversely proportional to the distance between the data points, specify the distance measure on the sample. Through the iterations, the observation set is centred according to the measure defined by the weights. The study of (Zhao et al. 2007) uses the fact that non-Gaussian data can be locally viewed as Gaussian: alternatively, any manifold can locally resemble an element of Euclidean space. The study investigates an algorithm based on a local weighting procedure given by the so-called local Laplacian scatter matrix. The above methodologies preserve the simplicity of PCA, as the underlying optimisation problems are formulated by a quadratic function with additional weighting procedures.

The other class of approaches to the robust PCA investigates different types of measures which assess the distance between a set of observations and its projection. Consequently, new frameworks provide procedures which are efficient in the presence of various assumptions on the noise.

When the observation data reveals a noise pattern which is sparse and of arbitrary amplitude, it impacts a small subset of the matrix elements. To illustrate this point, let us consider sudden illiquidity events on the currency markets, which cause abrupt jumps in the volatility of currency quotes. Such events are unexpected, short-lived, and sparse; that is, they appear only in a very small subset of the market participants and are characterized by relatively high differences. We can think about many factors, such as political, economic or social developments, which implicate these changes.

If the analysed dataset describes the prices of different currencies over time, sparse events of increased volatility appear as significant spikes on the data surface. What is more, these spikes impact only a limited number of instruments. Consequently, we observe significant relative differences between the majority of the prices described by the dataset and a few quotes. This characteristic defines these events as sparse. Sparse outliers easily corrupt classical PCA; the

corresponding distances are quadratically magnified, which impacts the calculation of eigenvectors and eigenvalues..

To handle a sparse noise component, a series of studies suggest replacing the quadratic loss function in the PCA problem with the $\mathbf{L}_1$ norm. The main disadvantage of this approach is the fact that the $\mathbf{L}_1$ term induces the sparsity to the solution of the optimisation problem, as shown in (Tibshirani 1996). The authors highlight a lack of smoothness in the methodology, with $\mathbf{L}_1$-type loss function, since the absolute value is not continuously differentiable. Therefore, the majority of works on PCA with $\mathbf{L}_1$ loss function recommend algorithms or simplifications to achieve the absolute convexity of the optimisation problem.

The work of (Croux and Filzmoser 1998) extends the application of the $\mathbf{L}_1$ regression studied in (Ukkelberg and Borgen 1993) to the PCA setting. The $\mathbf{L}_1$ regression is sensitive to points which have high leverage and belong to the regression plane. Therefore, the authors suggest using weights which are separable; that is, for the element of matrix $\mathbf{Y}$ in the ith row and jth column, the recommended weight w_{ij} is a product of w_j and w_i, the row- and column-wise weights. The assumption imposed on weight, however, has a poor performance, especially when there are missing elements in the dataset. When the value of a cell is missing, either a row-wise or column-wise weight corresponding to the missing entry is set to zero. An additional disadvantage of this approach is the fact that the method associates the information aggregated in the columns or rows of the datasets as outliers. If an element of matrix $\mathbf{Y}$ is identified as an outlier, then its entire row or column is labelled as outliers as well.

Therefore, the authors of (Ke and Kanade 2005) suggest a more element-wise framework for the $\mathbf{L}_1$ PCA. The algorithm overcomes the difficulty of the $\mathbf{L}_1$ norm by considering various sub-problems which allow for an optimisation framework by linear or quadratic programming. The algorithm stems from the fact that, if either matrix $\mathbf{U}$ or $\mathbf{V}$ is known, the equivalent solution to the problem defined in Equation (1.1) with $\mathbf{L}_1$ norm becomes convex. In addition, the authors introduce a sparse PCA scheme with a weighted $\mathbf{L}_1$ norm to further increase the robustness of the methodology. As a last contribution, the study advises how to account for missing data in the described optimisation framework.

A different optimisation problem is explored in(Ding et al. 2006), which induces the sparsity of the solution, using a rotationally invariant $\mathbf{L}_1$ norm which is expressed by the sum of the square roots of errors in the observations. As an alternative solution to the factorisation of the incomplete data matrix with the sparse noise assumption, the authors (Eriksson and Hengel 2012) generalise the Wiberg algorithm of (Wiberg 1976) and extend the approach to the $\mathbf{L}_1$ norm.

The Laplace loss function is very convenient to work with datasets which exhibit large, sparse noises, but it is not efficient when the perturbation present in the data is dense. The work of (Jolliffe et al. 2003) investigates the idea of combining the classical PCA with a flexible framework which induces sparsity to the solution. The authors introduce $\mathbf{L}_1$ penalties imposed on the eigenvectors stored in the columns of the matrix $\mathbf{V}$. The Lagrangian multiplier which corresponds to the constraint term is treated as a tuning parameter which decides the importance of the sparsity in

the investigated solution. (C. Croux et al. 2013) extends this idea by replacing the sample mean and covariance with more robust equivalents.

A great overview of projection pursuit methods which address the problem of representing a data matrix by sparse and dense components is presented in (Bouwmans et al. 2017). The authors discuss in detail the evolution of the concept and provide many useful references, with emphasis on the different assumptions of the works. Starting with the studies of (Wright et al. 2009), (Candes et al. 2009) or (Zhou et al. 2010), the alternative PCA methodology aims to combine both sparse and dense noise patterns. The studies investigate a model for PCA which decomposes the data into the lower-rank matrix, a small and dense noise matrix and a large and sparse noise matrix. Most of the methods are very sensitive to the initialisation step. For instance, the frameworks investigated by (Zhou et al. 2010) requires a good estimate of the magnitude of the dense noise, which is usually difficult to obtain. The work of (Xu et al. 2012) proposes a formulation of the problem which focuses on the exact recovery of the eigen-representation of the data matrix, rather than the recovery of the data matrix as is broadly approached in other studies.

Up to this moment, all discussed methodologies and approaches to a robust PCA use a non-probabilistic paradigm or implicitly refer to the assumptions on a distribution of the observation's noise. Since PCA is a deterministic method, it is difficult when one wants to incorporate the probabilistic inference and specify various a priori assumptions about the distribution of the observation process. The probabilistic setting is applicable when one wants to assess and test the statistical properties of the model and is particularly relevant in handling incomplete observations.

The other class of methodologies, which extends PCA to its probabilistic interpretation, were introduced by (Tipping and Bishop 1999b) and (Roweis 1998) as Probabilistic Principal Component Analysis (PPCA). In addition to modelling the noise vector, which is assumed to follow a Gaussian distribution, PPCA imposes a prior assumptions on the distribution of latent variables. These random vectors specify a linear projection of the original data points onto the lower-rank space. PPCA can be easily tailored to handle incomplete information in the sample data and allows the utilisation of the probabilistic assumptions about both the type of missingness as well as the distribution of missing values. The Expectation-Maximisation algorithm, formalised by (Dempster et al. 1977) and discussed in detail, with its extensions, in (McLachlan and Krishnan 1997), is especially suited for inference of probabilistic models with unobserved or hidden variables and is broadly utilised as an estimation technique for PPCA. The reader may refer to the works of (Tipping and Bishop 1999b), (Roweis 1998) and (Ilin and Aiko 2010) for the derivations of PPCA in complete and incomplete data settings.

Similarly to PCA, its probabilistic version, the PPCA, is not robust in the classical setting of PPCA statistical model assumptions. The detailed discussion on the statistical properties of the robustness of Gaussian PPCA is given in (Xie and Xing 2015). The authors show that the estimators of the model's parameters obtained by Gaussian PPCA are very sensitive to the small amount of contamination that may be present in the domain of an underlying distribution. The effect of the contamination on the results obtained by Gaussian PPCA is unbounded, which is not a desirable

property of any estimator, since one would like to ensure that the error of estimation is restricted. The authors of (Xie and Xing 2015) prove, using an influence function, that the Gaussian PPCA is sensitive to the presence of outliers.

One of the natural extensions of Gaussian PPCA is to replace the Gaussian distribution with a heavy-tailed density. The studies of (de Ridder and Franc 2003), (Khan and Dellaert 2003), (Archambeau et al. 2006), (Fang and Jeong 2008) and (Chen et al. 2009) address this problem and explore the use of the t-Student assumption on the noise distribution and its impact on the robustness of the methodology with respect to dense noise. If one considers the work of (Ke and Kanade 2005) from a probabilistic perspective, the study proposes to use the Laplace distribution to improve the resistance of PPCA to large, sparse noises. (Gao 2008) formulates the PPCA problem with the Laplace error term and Gaussian latent variables. Other works introducing sparsity to the solution of PPCA are (Guan and Dy 2009) and (Bach and Archambeau 2009), both of which incorporate sparse prior distribution on the model's parameters via a variational Expectation-Maximisation. (Austin and Ghosh 2015) add sparse domain constraints on the distribution of latent variables.

Similarly to the sparse version of PCA, the PPCA model with assumed Laplace distribution is not resistant to dense perturbations. The authors of (Xie and Xing 2015) improve the robustness of PPCA to both sparse and dense outliers of significant magnitude by assuming that the error term and latent random vector follow a Cauchy distribution. The investigated distribution has similarly heavy tails to the Laplace distribution and hence is feasible to model the outliers with exceedingly distant values. Contrary to the Laplace distribution, the Cauchy distribution does not impose sparsity in the solution. Another significant contribution of (Xie and Xing 2015) is the focus on the statistical interpretation of the robustness achieved by Gaussian, Laplace and Cauchy PPCA.

Many studies suggest combining PPCA with the mixture models, as defined in (Redner and Walker 1984), (Peel and McLachlan 2000), to model arbitrary probability density functions of the observation process or the distribution of the noise which corrupts the data. The use of the mixture of Gaussians is a universal approximation technique for any continuous distribution. The core idea behind the procedure is partitioning a set of observations into sub-populations which are modelled by different parametric distributions. The author of (Archambeau 2005) discusses robust models in the context of a finite mixture and compares the interpretation of the Mixture Models PPCA to probabilistic versions of graphical models or the weighted vector quantisation-based kernel density estimator. The models' parameters are shown to be feasibly learnt through the Expectation-Maximization algorithm with explicit formulas. Also, the authors show the extension to the Bayesian formulations. The study of (Archambeau 2005) improves the robustness of the Mixture Models PPCA by utilising the concept of t-Student mixture models as in (Peel and McLachlan 2000). The authors of (Chen et al. 2006) propose the efficient Bayesian estimation framework, which assigns the prior distribution on the model's parameters. The study investigates the utilisation of the concept of infinite Gaussian mixture models to the PPCA setting. The motivation of mixtures of PPCA as an approach to model nonlinear relationships by alignment of a collection of local models is studied in (Tipping and Bishop 1999a), among many other studies.

1.2.2. State Space Models and Filtering Techniques

Often, it is crucial to use the sufficient characterisation of large datasets in statistical modelling. The term 'sufficient characterisation' may mean a model or a functional representation of the data set. State-space models, as defined in (Harvey 1990) or (Hamilton 1994),(West and Harrison 1997), (Bar-Shalom et al. 2001) among others, are both powerful and flexible tools which can represent a wide range of linear or non-linear time series, including stationary or non-stationary time series, as well as regression model structures with changing coefficients and unobserved components in univariate and multivariate frameworks. In the state-space model approach, the relationships between a measurement (the response variable), a latent or explanatory variable and a noise component are expressed in the form of a dynamic system. This framework has been more broadly introduced to statistics by (Kalman 1960) and originates from earlier pioneering works on stochastic filtering theory, such as (Wiener and Hopf 1931) and (Kolmogorov 1941) and from physics and system theory, as in (Zadeh and Desoer 1963).

The econometric approach to state-space models, with comprehensive analysis of various cases, is broadly discussed in (Harvey 1990), (Hamilton 1994) or (Bar-Shalom et al. 2001) where the authors demonstrate the treatment of the time series from the standpoint of linear Gaussian models. The books of (Bar-Shalom et al. 2001) and (Durbin and Koopman 2012) further extend the discussion to address the assumptions required to treat non-Gaussian or non-linear dynamics. Standard references to the Bayesian approach in state-space modelling include (Kitagawa and Gersh 1996), (Kim and Nelson 1999) or (Fruhwirth-Schnatter 2006).

The inference of latent variables in linear Gaussian dynamic systems is successfully carried out by the Kalman Filter, first derived by T.N. Thiele in 1880 (see discussion in (Lauritzen 1981)) but more comprehensibly introduced by (Kalman 1960) and (Kalman and Bucy 1961). This recursive procedure estimates the current state of an unobserved process, given the previous and current observations. Therefore, this is explicitly structured to allow for a recursive formulation that is often termed 'on-line' estimation of the hidden dynamics. The Kalman Filter plays a fundamental role in both filtering and forecasting of Gaussian linear models. Under the assumption of the Gaussian distribution on the additive terms of the observation and state equations in the linear model, the filter provides an optimal estimation and forecast of the observations in future time steps, as well as the estimation of the latent state vector in both in-sample and predictive settings. The obtained estimates of hidden states maximise the likelihood of the model, as shown in (Duncan and Horn 1972). When the assumption on the Gaussian distribution of the model is relaxed, the filter provides optimal estimates in the context of the mean square error. Additionally, the construction of the filter gives a straightforward framework for the estimation of static parameters in the model. The framework of the Kalman Filter yields a closed-form expression of the marginal likelihood function conditioned on the parameters, given the assumptions mentioned above on the distribution of the error terms. This either results in the complete system of linear equations for the state estimation or forms a recursive procedure to calculate the estimates of the parameters by the gradient descent methods. If the Gaussian distribution of the error terms is not assumed, the Kalman Filter provides the recursive least-squares estimation of the parameters of a classical linear regression model.

The discussions in (Chapter 6, Anderson and Moore (1979)) or (Sayed and Kailath 1994) focus on improvements to the numerical stability and computational efficiency of the classical algorithm of the filter, known as the covariance filter. Variants of the algorithm include the information filter of (Fraser 1967) or the square root filter of (Thornton 1976) and (Thornton and Bierman 1980) as the first studies on the topic, which are also broadly discussed in (Bar-Shalom et al. 2001) .

Following its first introduction, many extensions appeared to the classical application of the Kalman Filter. Early works include (Harvey and Philips 1979), where the authors demonstrate the treatment of ARMA errors in state-space modelling or the manuscript of (Jones 1981), which investigates modelling unequally-spaced observations in time by continuous state-space model formulations. (Naveau et al. 2005) extends the framework to the Skew-Gaussian distribution and provides closed-form formulations of the filtering recursions. The Kalman Filter under a Bayesian paradigm was initially derived by (Meinhold and Singpurwalla 1983).

More complex data structures are efficiently modelled by non-Gaussian dynamic models; for instance, heavy-tailed distributions or non-linear dependencies of either noise vectors or hidden processes. This generalised class of state-space models is broadly overviewed in (Kitagawa and Gersh 1996), (Kim and Nelson 1999) or (Durbin and Koopman 2012).

Many pioneering works have studied these optimal nonlinear filtering problems, with the majority of them from the stochastic differential equations perspective, such as (Zadeh 1953), (Bucy and Joseph 1968), (Bucy 1969), (Wonham 1965), (Zakai 1969) or most importantly, (Kushner 1967) and (Stratonovich 1968). In this thesis, we focus on discrete time problems, and in the remainder of this part, we revisit the optimal solutions to intractable discrete-time filter problems and their approximations. For the treatment of filtering problems in the diffusion setting, we refer the reader to the discussion in (Bain and Crisan 2009) or (Crisan and Rozovskii 2011) and references therein.

Recursive equations of the Kalman Filter in the discrete-time setting are explicitly formulated within the class of linear Gaussian models. However, for the majority of non-Gaussian or non-linear models, the corresponding conditional posterior densities and their moments may not be analytically tractable or may not have sufficient statistics to formulate the filtering recursions explicitly. The same applies to the Bayesian formulations of these problems when one incorporates a priori knowledge about the distribution of the static parameters of the models.

As highlighted in (Sorenson 1985), the approximations to solutions of the filtering recursions can be categorised as local or global. Local approximations exploit the information available at a particular point of the approximated function's domain and project this information to the neighbourhood of this point. Consequently, they are valid only in close proximity to the point at which they are generated. Global approximations find the most accurate representation of the function over its whole domain.

Numerical Approximation Methods.

The most popular estimation and filtering methodology, among local approximation approaches, is the Extended Kalman Filter as in (Jazwinski 1970). The procedure exploits the first-order linearisation of the dynamic model and approximates its filtering distribution by

Gaussian approximations. The first-order linearisation may not provide a sufficient representation of the model to accurately estimate highly non-linear state-space models when the higher-order terms of Taylor expansion of the dynamic model are not negligible. It can lead to large errors in the true mean and covariance of the state distribution which is transformed by linear combinations of Gaussian distributions. Such a situation may lead to suboptimal solutions of the filtering recursions, as well as divergence of the filter. The authors of (Julier and Uhlmann 1997) address this issue by suggesting the filtering algorithm called the Unscented Kalman Filter. The procedure uses a deterministic adaptive sampling technique to specify the linearisation points which capture up to the third order statistics of the true posterior distribution of the non-linear problem. Instead of a crude linearisation of the system, the authors propose to use a set of weighted points to parameterise the true means and covariances of the filtering distributions. The method ensures that the obtained approximated posterior means and covariance matrices capture the true values. Other works which described higher-order local approximations include the studies of (Athans et al. 1968), (Tanizaki and Mariano 1996), (Wan and van der Merwe 2000), (van der Merwe 2004), among many others.

Global approximation methodologies comprise the group of frameworks which, instead of approximating a system itself, directly approximate either the densities required for the filtering recursion steps or the corresponding integrals. We can distinguish two classes; deterministic methods and Monte Carlo approximation techniques.

Deterministic approaches, which approximate the solution to either an integral problem or a density function, are based on quantisation of the state space at points distributed through the region containing non-negligible probability. The difference between these two approaches lies in the motivation behind the choice of the points and approximation functions. In the density function approximation, the points and approximate values of a density function are chosen to match an interpolated function, in order to make the integrations tractable. The core idea behind integral approximation is to match the regions of interest of the density function with interpolating functions from a particular numerical integration scheme.

The earliest works on deterministic approximations of densities include the point-mass method, which originates from the works of (Bucy 1969) or (Bucy and Senne 1971). The procedure discretizes the state space and the resulting distribution representation, using a rectangular grid and approximate continuous density function by weighted Dirac deltas to minimise a distance measure to the true density, and discrete probability masses. (Kramer and Sorenson 1988) suggest an improvement to the point-mass approach of (Bucy and Senne 1971) by introducing the piecewise-constant density approximation. Another idea, known as the Gaussian sum filter, uses a mixture of Gaussian distributions which correspond to different masses of points in the probability space, as proposed by (Sorenson and Alspach 1971) or (Alspach 1974). The reader may notice that Gaussian mixtures were also mentioned in the previous review of robust PPCA treatment when linear combinations of Gaussian distributions were adopted to approximate the heavy-tailed distribution of an error term and a latent random vector.

With the increase in computational power at the beginning of the 1980s, more sophisticated and accurate deterministic techniques were investigated which utilised the developments of advanced numerical integration. These new methodologies focus on approximating the filtering recursions expressed by expectations of conditional distributions and, therefore, are represented as integrals. We can exemplify the first works on the topic which use different schemes to obtain approximations of nonlinear filtering recursions. The study of (Kitagawa 1983) recommends the spline and step function interpolation of non-Gaussian recursions, while the Laplace approximation is supported by (Tierney et al. 1989), with various extensions in (Mackay 1998). (Schnatter 1992) first proposed a Gauss-Hermite polynomial quadrature for a special case of a dynamic linear model with a trend. (Frühwirth-Schnatter 1993) generalised the method to the broader range of models with linear transition equations. The work of (Ito and Xiong 2000) further extends the concept of the Gauss-Hermite quadrature applied to non-linear filtering problems. The authors introduce the Gauss–Hermite filter and the Central Difference Filter, which are recommended for non-Gaussian dynamic models with Gaussian filtering distributions. A comprehensive review of the early works on estimation by non-Gaussian state-space models can be found in (Kitagawa and Gersh 1996) or (Ito and Xiong 2000).

Gaussian mixtures are also used to approximate the expectations required by Kalman Filter recursions. (Harrison and Stevens 1976) propose a multi-state model which, at each time step of filtering, approximates a posterior density function by a discrete Gaussian mixture model. The procedure results in exponential growth of the Gaussian components with time, since the Kalman filter is applied to each combination of new and previous models over time. The problem of state augmentation was addressed by (Kitagawa 1994), who proposed a two-filter formula which re-approximates true filtering distributions at each time step by a fixed number of new Gaussian components. It is worth noting that the described difficulty of polynomial growth of state to track is not solely a problem of non-linear filtering, as it is also present in other statistical areas, such as estimation or classification.

Numerical Approximations by Monte Carlo Methods.

The class of Monte Carlo (MC) methods comprises a set of procedures which allows samples to be taken from various complex distributions which are otherwise not tractable to perform an integration. Some of these methods, but not all of them, satisfy important features from a statistical perspective that the resulting estimator of the integrals satisfies the Weak or Strong Law of Large Numbers or Central Limit Theorem. However, the primary objective of this class is an integration of intractable functions in some sense. The properties of these methods, such as convergence, unbiasedness or consistency of the estimation, additionally provide statistical properties of the estimation and the notion of precision for the approximation, given a number of generated samples. The modern formulations of MC methods originate from the early works of (Metropolis and Ulam 1949) , (Metropolis et al. 1953), (Hammersley and Morton 1954) or (Rosenbluth and Rosenbluth 1955).

There are different Monte Carlo techniques which evolved in parallel in the literature, since they address different statistical problems. Some of them are built for on-line recursive (sequential) estimation while others focus on static batch estimation. Some methods target low-dimensional problems, while others target high- dimensional contexts. In this section, we will not definitively describe all different classes of Monte Carlo methods; we will discuss potential classes of methods which are relevant to the filtering problems explored in this thesis, such as Simple Monte Carlo Methods and Markov Chain Monte Carlo methods, applied to estimation of state-space models, as well as the class of Particle Filters. These methods are broadly discussed in (Rober and Casella 2004) or (Evans and Swartz 2000) and have been successfully applied to estimate non-Gaussian and non-linear state-space models.

The most widely-known Monte Carlo technique, known as the Simple Monte Carlo, translates the theory of the Central Limit Theorem into a sampling procedure to approximate solutions to difficult integration problems. The standard Monte Carlo method samples from a posterior distribution of the state vector to approximate required filtering recursions by empirical measures. As explained in (Shiryaev 1996), Central Limit Theorem (CLT) ensures that an estimator of the integral, which is its approximate solution, converges in distribution to its true value with some approximation error, which decreases proportionally to the number of samples. The general framework for applying MC methods to approximate filtering recursions of state-space models is outlined in (Tanizaki 1993). However, if the true posterior distribution belongs to a complex high-dimensional distribution or is unknown, the simple MC method might be infeasible or inefficient.

If it is difficult to sample from the true distribution, one can address this problem by the rejection sampling algorithm discussed in (von Neumann 1951) and reviewed in (Hammersley and Handscomb 1964) or (Handschin and Mayne 1969). The procedure requires the existence of an auxiliary density function, which is an upper bound of a target density function on its whole domain. The algorithm is justified by the Fundamental Theorem of Simulation. The samples are drawn from this distribution and rejected with a rate based on the ratio between the two functions and a constant. The constant term is specified before the procedure starts and determines the rejection rate of the samples drawn from the proposal distribution, as well as the speed of the algorithm. However, finding a strictly dominating density in high dimensions in rejection sampling might be very difficult.

Standard rejection-sampling might turn out to be computationally expensive, even for one-dimensional problems, since the procedure requires calculating the ratio between the proposal and the true density at each iteration. In the early work of (Devroye 1986), the author addresses this problem by adapting the squeezing principle of (Marsaglia 1977) to the rejection resampling concept in the univariate setting. By finding two functions which are upper and lower bounds of a true distribution, a way is found to design a more efficient sampling procedure, with the rejection being decided by these two specified functions. (Wild and Gilks 1993) further extends this idea in the setting of log-concave probability density functions. The procedure, known as adaptive rejection sampling, uses the fact that log-concave functions have their logarithms bounded by

the tangent lines. The key idea behind the algorithm is the application of an adaptive squeezing function which is updated at each iteration by the information given in rejected points. A review of rejection sampling techniques applied to Bayesian inference can be found in (Tanizaki 1993) and (Tanizaki and Mariano 1998).

The concept of using auxiliary distribution is also explored within importance sampling techniques formulated in (Marshall 1956). The methods have been discussed in the context of inference for non-linear and non-Gaussian models in (Handschin and Mayne 1969) (Kloek and van Dijk 1978), (Geweke 1988), (Shao 1989), (Geweke 1989), as early examples. The procedure draws random realisations from a proposal distribution which overweights the region of importance and, as a result, reduces computational intensiveness of the procedure. The choice of the proposal distribution determines the properties of the sampling procedure and the obtained estimator; for instance, its variance. For a given sample, a ratio between the underlying probability density function and the auxiliary function acts as a weight which assigns its importance. Therefore, the assumption of the auxiliary function is a crucial element of the efficiency of the methodology - a naive choice of the proposal function can result in variance of the estimator which is higher than the variance of the estimator obtained from simple MC samples.

Also, importance sampling starts to be expensive in high-dimensional problems, since it samples from the whole domain, without learning to target the region of importance. It is not trivial to find an importance sampling distribution which ensures that the ratio of the weights is not degenerated; in other words, is not high for one sample and close to zero for many others. A small distortion in mass with respect to the true distribution in the ratio causes discrepancy in the weights and results in higher variance of the estimates when one integrates those samples.

In order to improve the variance of the estimator obtained via importance-sampling techniques, (Rubin 1987) (Rubin 1988) introduces the importance sampling procedure with a rejection step, known as Sampling-Importance Resampling or simply resampling. Its application to Bayesian inference has been studied since the early works of (Smith and Gelfand 1992), (Kitagawa 1996), (Kitagawa and Gersh 1996) and (Kitagawa 1998). The idea of resampling originates from the statistical bootstrap formulated in (Efron 1979), where one explores the properties of the estimator from the empirical distribution perspective. The first step of a resampling algorithm constructs an empirical distribution based on samples and their weights obtained by importance sampling. Next, the samples are redrawn according to this empirical distribution.

The additional resampling step adds more variability to the estimation but removes samples with assigned high-probability mass and outside the region of importance, as shown in (Chopin 2004). If the variability of the weights which determine the importance of a sample is small, resampling is unnecessary. A popular procedure which assesses the variance of the importance weights is the Effective Sample Size criterion proposed in (Kong 1992).

There are many variants of the classical resampling algorithm which was first applied to numerical integration in (Rubin 1987); we may distinguish multinomial resampling, discussed in (Rubin 1987), residual resampling as in (Higuchi 1997), minimum variance resampling described in (Kitagawa 1996) or local MC resampling as in (Liu and Chen 1998).

Markov Chain Monte Carlo.

The other class of MC techniques, Markov Chain Monte Carlo (MCMC), was initially used for addressing batch estimation problems in high dimensions, which remains its most common use. There are many types of MCMC algorithms which share the same objective: to obtain a sequence of samples with some dependency structure (depending on the MCMC algorithm), typically from a high-dimensional distribution.

The concept of MCMC originates from the studies of (Metropolis and Ulam 1949) and (Metropolis et al. 1953) and was further introduced to probabilistic inference by (Neal 1993). MCMC explores the structures between the elements of a random vector, described by the underlying joint distribution, and approximates its expectation by using a Markov chain mechanism. Therefore, instead of drawing independent samples from an underlying distribution, as in MC methods, MCMC algorithms draw sequential elements of a Markov chain which converges to its stationary distribution. The construction of the chain ensures that the invariant distribution corresponds to the underlying distribution of the integration problem. From the Markov chain theory, the irreducible, aperiodic chain with invariant distribution converges to its invariant distribution. This convergence and its speed depend on the number of samples, the dimensionality of the problem and the assumed type of the dependency structure between different states of the chain, defined by its transition kernel. An early review of the most popular MCMC algorithms can be found in (Roberts and Smith 1994), (Tierney 1994) or (Robert 1998) who carefully analyse and develop properties of the resulting chains, such as convergence and central limit theorems. We refer the reader to many excellent tutorials and studies which review different MCMC algorithms, such as (Gilks et al. 1995), (Fishman 1996), (Chen et al. 2000) (Rober and Casella 2004),(Liu 2004), (Gamerman and Lopes 2006), (Särkkä 2013) or (Brooks et al. 2011).

A range of various MCMC algorithms investigates different variants of transition kernels. We can distinguish two especially popular algorithms. The Metropolis-Hastings algorithm, studied by (Metropolis et al. 1953) and (Hastings 1970), associates the transitioning from one state of chain to another with a rejection step. It requires specification of an auxiliary function which, combined with an actual probability density function, defines the kernel of a Markov chain. The algorithm specifies a rejection step which ensures that the detailed balance conditions of the chain are satisfied.

(Roberts and Smith 1994) and (Tierney 1994) show that, under mild regularity conditions on the target distribution, which provide irreducibility and aperiodicity of the chain, the chain converges to its invariant distribution. (Tierney 1994) shows that, by the design of the kernel, the Markov chain constructed by the Metropolis-Hastings algorithm satisfies balance conditions and is aperiodic. If, in addition, the support of a true density function is included in the support of the auxiliary function which defines the kernel, we obtain the convergence of the chain. As discussed by (Chib and Greenberg 1995), the properties of the chain, such as the rate of convergence, are dependent on the construction of the kernel and its proposal distribution. The studies of (Mengersen and Tweedie 1996) and (Roberts and Tweedie 1996) focus on analysing the convergence and its rate under various conditions imposed on the kernel. The authors of (Mengersen and

Tweedie 1996) show that, in a univariate setting, a sufficient condition to obtain a geometrical convergence rate of the algorithm is an invariant distribution with exponentially decreasing tails. The results are further extended to the multidimensional setting in (Roberts and Tweedie 1996), where the authors define additional sufficient conditions on the smoothness of the invariant distribution required for its geometrical convergence.

The next popular MCMC algorithm, the Gibbs sampler, introduced by (Geman and German 1984), defines the transition kernel as a product of full conditional probability densities of the random vector's elements. We recommend the work of (Casella and George 1992), which provides a comprehensive and intuitive introduction to the technique. The Gibbs sampler is a special case of the Metropolis-Hastings algorithm, with the acceptance rate equal to one; see, for example, (Robert 1998). By construction, it is a multidimensional framework. The algorithm reduces sampling from a highly dimensional joint distribution to a sequence of less complex, one-dimensional problems. If the support of a true distribution can be decomposed into the Cartesian product of the supports of full conditionals, the generated chain is irreducible and recurrent, which implies convergence for aperiodic chains. The conditions which need to be satisfied to ensure the recovery of joint distribution from its full conditionals are summarized by the Hammersley–Clifford Theorem. (Roberts and Smith 1994) discuss sufficient conditions necessary for a generated chain to be aperiodic and irreducible. The works of (Schervish and Carlin 1992) and (Roberts and Polson 1994) study the convergence properties of chains generated by the Gibbs Sampler and the conditions which imply a geometric rate of their convergence.

The works of (Liu et al. 1994) and (Liu et al. 1995) study the effects of a covariance structure of the target distribution on the convergence of the Gibbs sampler. The authors analyse various methods of grouping and collapsing techniques which improve the speed of convergence. The 'grouped' or 'blocked' Gibbs sampler algorithm is based on generating a component of a Markov chain by sampling from full conditionals of two or more variables together. The collapsed Gibbs sampler marginalizes full conditionals in some of the steps. The authors show that both algorithms improve the rate of convergence of the chain to its invariant distribution.

Another set of methods which improves the efficiency and accuracy of sampling techniques, in general, is Rao-Blackwellization. These marginalisation techniques for MC methods originate from the works of (Rao 1945), (Blackwell 1947) and (Kolmogorov 1950) and ensure the variance improvement of the estimators obtained by simulation-based approximations. The framework is based on the formula of variance decomposition and suggests the use of a conditional expectation of an estimator as an estimator with lower variance. The Rao-Blackwellization estimator is not applicable to all estimation problems, since its advantage lies primarily in obtaining analytical solutions to the portion of integrals required by a filtering problem. The authors of (Liu et al. 1994) and (Liu et al. 1995) highlight the improved convergence of the Gibbs sampler chain by adopting Rao-Blackwellization in the Gibbs sampling. A similar argument is given in (Gelfand and Smith 1990) but in the setting of independent realisations of a chain. (Casella and Robert 1996) discusses efficient algorithms which adopt Rao-Blackwellization in rejection sampling and Metropolis-Hastings algorithms.

Conceptually, the Gibbs sampler is a straightforward procedure, given that formulations of conditional densities are simple enough to generate random draws. One can address the application of the Gibbs sampler in the presence of complex full conditionals by adopting a rejection step procedure within the sampling framework. (Carlin et al. 1992) recommends using a carefully selected accept/reject method to draw from non-standard distributions. Motivated by the efficiency of the Gibbs sampler, (Gilks and Wild 1992) introduces an adaptive rejection sampling step applicable to sampling from univariate, log-concave probability density functions, known as adaptive rejection sampling. One of the shortcomings of this procedure is its applicability only to univariate full-conditional densities which are log-concave. (Gilks et al. 1995) extends the framework to combine an adaptive rejection step and the Metropolis-Hasting algorithm step within the Gibbs sampler to sample from non-log-concave full conditionals. The first application of the Gibbs sampler with an adaptive rejection sampling step includes inference of generalised linear models by (Dellaportas and Smith 1993) and the calculation of likelihood function for Bayesian hierarchical models from an exponential family with conjugate prior distributions by (George et al. 1993).

Another idea is to use the formulation of the Gibbs as a special case of the Metropolis-Hastings algorithm: combining a Metropolis-Hastings step within the Gibbs sampler algorithm is a straightforward solution to sampling from complex full conditionals. The early works of (Tierney 1994) and (Muller 1991) explore various hybrids of these two MCMC frameworks. Examples of filters with hybrid MCMC steps for the inference of non-linear state-space models include the works of (Geweke and Tanizaki 1999) and (Geweke and Tanizaki 2001), which explore the application of the Metropolis-Hastings algorithm within Gibbs sampling. The method handles cases exceptionally well when the acceptance probability of a candidate for a chain is zero, and a rejection sampling step cannot be carried out.

Bayesian Parameters Learning.

So far, we have discussed simulation-based filtering techniques assuming that the static parameters of dynamic models are known. In frequentist inference, optimal estimates of the parameters are primarily obtained by the maximum likelihood estimation. If the likelihood function of a model cannot be explicitly formulated or the optimisation is infeasible, one can use the Expectation-Maximisation algorithm of (Dempster et al. 1977), among many other approximation methodologies. The computational intensiveness of these procedures increases with the complexity of an applied filtering methodology and the dimensionality of the measurement and state-space. One of the methods which mitigate this problem is described in the studies of (Kitagawa 1998). The suggested method imposes simple artificial time evolutions on the distribution of the parameters. The added random noise component addresses the issue of likelihood sensitivity to sampling error.

However, if the inference accounts for the uncertainty of the model's static parameters, the estimation becomes even more difficult, especially when the filtering problem is high-dimensional. The estimation of parameters with assigned a priori distribution concentrates on obtaining their maximum a posteriori values or conditional expectations, or, alternatively, conditional quantiles.

Standard MC methods are rarely used in the context of joint Bayesian state inference and the learning of parameters, due to the complex formulations of the underlying posterior distribution and high dimensionality of the state-space, which includes additional random components, the vector of parameters.

The MCMC methods are more convenient to address the Bayesian parameters learning and filtering of state variables. The study of (Carlin et al. 1992) illustrates the broad applicability of the MCMC frameworks, especially of the Gibbs sampler, to the inference of state-space models. (Carlin 1992) extends this framework to the Bayesian setting when an inference of latent vectors and parameter learning is performed simultaneously. The framework utilises the Kalman Filter recursion within an MCMC procedure. The majority of the simulation-based filters exemplified previously are not designed to perform simultaneous estimation of state vectors and parameters specified by Bayesian state-space model formulation. The authors of (Carlin 1992) outline a general formulation of inference for state-space models in a Bayesian setting, which utilises the Gibbs sampler in the Kalman Filter for the class of models which are conditionally Gaussian. The objective of the sampler is to approximate the posterior mean of the filter; that is, to generate samples from a posteriori distribution. The algorithm of (Carlin 1992) assumes the samples to be drawn one at the time. The single-state sampling procedure results in a very slow convergence of the sampler for more complex, non-Gaussian state-space models.

To achieve faster convergence of the Gibbs filter, (Fruhwirth-Schnatter 1994) and (Carter and Kohn 1994a) suggest using a multi-state sampler which draws all states at the iteration to evaluate the smoothing means in a Bayesian framework. This framework reduces the correlation between subsequent iterations of the random draws and speeds up the convergence of the chain. Alternatively, (de Jong and Shephard 1995) argues that applying the multi-state Gibbs sampler to generate samples from a distribution of noise processes provides a less complex framework. For non-standard formulations of full conditionals for the parameters, (Carlin 1992) proposes the rejection sampling procedure descried above within the Gibbs sampler as an alternative treatment for non-linear models, though it is not computationally efficient. This problem can also be addressed by blending MCMC procedures.

The filtering procedure with parameter learning proposed by (Fruhwirth-Schnatter 1994) and (Carter and Kohn 1994a) belongs to the batch estimation group of problems. It means that the estimation framework is not updated after the arrival of a new observation. Instead, the algorithm needs to be rerun on the new set of data; it is not a time-efficient solution. As remarked above, the previously mentioned MCMC filters, which might be reformulated for Bayesian inference, might also not be efficient. With the increase of dimensionality of the state or measurement space of dynamic models, an inference by MCMC often has inferior efficiency and a slow convergence, due to increasing correlation between samples from posterior distributions of the state vectors.

What is more, these MCMC solutions applied to a Bayesian inference belong to the class of off-line or batch algorithms. The frameworks are not suitable for online learning when the data arrives sequentially, and the integrals need to be solved progressively in time. In such online applications, MCMC is deemed computationally expensive. The dimensionality of a random

vector to be propagated in a sequential filtering problem increases over time. It makes MCMC even more computationally intensive, as one needs to ensure a sufficient number of samples to reach the invariant distribution at each time-step.

Sequential Monte Carlo.

The studies of (Gordon et al. 1993), (Kitagawa 1996) and (Hurzeler and Kunsch 1995) independently formalise the class of filters which exploits the concept of on-line Bayesian filtering for general state-space models. The methods approximate the solution to the filtering recursions of state vectors by a set of candidates with discrete probability masses. The procedure involves three steps: (1) generating particles, the candidates for realisations of the state vector in time t, given information up to t using a predictive density of the filter, which is a function of the filtered states from the previous step and a realisation of the error term; (2) computing the likelihood of the information which arrives in time t for each particle; (3) resampling the particles according to probabilities proportional to the corresponding likelihood values. (Gordon et al. 1993) introduce this framework as the bootstrap filter. The procedure follows closely the idea of a sampling-resampling framework discussed by (Smith and Gelfand 1992) but applied to filtering recursions.

This class of MCMC methods, known as Sequential Monte Carlo, includes sampling techniques for efficient filtering steps, which update a posterior distribution of the dynamic model in the light of a new observation. They follow a similar concept of on-line inference to that of the Kalman Filter, but address the problem of dynamic modelling of arbitrary state space models. These methods exploit time-ordering of sequences of random vectors, determined by an underlying model structure and filtering recursions. Extensive literature on past and recent developments of SMC can be found in (Doucet et al. 2001), (Arulampalam et al. 2004), (Doucet and Johansen 2009),(Pollock 2010), (Carvalho et al. 2010), (Lopes and Tsay 2011), among many other available resources.

The core method behind SMC is Sequential Importance Sampling (SIS), proposed by (Kong et al. 1994), which is a variant of importance sampling. By exploring the time order of random vectors, the procedure enables sequential construction of an importance function and sampling of a current state vector without modifying the candidates of its past realisations. At each time point, the procedure results in a set of pairs; a sample and its importance weight. The weights specify the contribution of each set of augmented samples to the posterior distribution and reflect their accuracy in predicting the next observation.

Degeneracy Problem of Sequential Monte Carlo.

The early work of (Kong et al. 1994) highlights the main problem with SIS filters and shows that the variance of sequential importance weights increases over recursive steps of the filter. Therefore, particles appear with associated weights which are close to zero. Since the algorithm tends to update a sequence of samples which have a negligible contribution to the posterior

distribution, it impacts its computational efficiency and might result in a lower accuracy of the obtained approximation. The problem is known as the degeneracy phenomenon.

One way to reduce the possibility of degeneracy is to add a resampling step to the procedure. The combination of SIS and resampling lead to the rise of a new class of Monte Carlo algorithms, called Sequential Importance Resampling (SIR). The resampling step aims to eliminate samples which do not belong to the region of interest, defined by the importance function, and increases the probability of samples which do belong. The methodology is similar to the standard resampling procedure but explores the dynamic properties of models under consideration. (Kong et al. 1994) introduces a procedure of sequential imputing for estimation of missing data using the resampling procedure. (Liu and Chen 1995) applies an analogous procedure to solve the problem of a blind convolution, formulated as a linearly degraded moving average in the Bayesian setting.

The other way to slow down the degeneracy of SIS filters is to use a proposal function which is optimal, in the sense of minimising the variance of importance weights. The concept was first discussed in (Zaritskii et al. 1976). The optimal function formulates a filtering recursion in a way that the weights can be specified using the available observation, but before the samples are propagated. Hence, the representation of the optimal importance function proposes a framework which allows changing the order of sampling and resampling. It yields a new direction of particle filters, where the resampling step is performed before the samples are generated from the posterior. The use of an optimal importance function ensures that the samples of current states are simulated from particles with high associated predictive likelihood, in contrast to the proposals of the state vector generated by SIR, which are sampled blindly; that is, ignoring the current observation.

The conditional densities required by the factorisation of the optimal importance function are often infeasible to sample from or are not tractable. As a possible solution, one can resort to the same linearisation techniques applied to the standard Kalman Filter, such as Taylor expansions of the dynamic model or of the optimal importance function, as illustrated in (Doucet et al. 2000) or an MCMC step as suggested in (van der Merwe et al. 2000). Another idea, the Auxiliary Particle Filter of (Pitt and Shephard 1999) and (Pitt and Shephard 2001), introduces an auxiliary variable which enforces the tractability of analytically intractable conditional distributions required by the factorisation of an optimal importance function.

Even though the resampling step improves the efficiency of sequential importance sampling, it suffers from one significant drawback. In general, as commented in (Doucet and Johansen 2009), any sampling procedure fails to accurately represent a distribution on a high-dimensional space, given finite sample size. The resampling procedure in the particle filters discussed encourages the situation where the space of a dynamic model is represented by only a few series of particles with significant importance weights, incremented over a large number of time steps. The problem is often referred to as sample impoverishment. The diversity of the paths is exceptionally low when the process noise is small. Various filters have been introduced to improve the diversity of particles. The comparison studies of (Arulampalam et al. 2004) and (Doucet and Johansen 2009) discuss the most popular solutions, such as the jittering and prior editing methods discussed in (Gordon et al. 1993), the prior boosting method investigated in (Gordon et al. 1995), the resample-move algorithm

with MCMC step of (Gilks and Berzuini 2001), the regularised particle filter suggested in (Musso et al. 2001) or block sampling strategies of (Doucet et al. 2006). (Doucet et al. 2000) and (Doucet and Johansen 2009) show how to improve the accuracy of the particle filters by Rao-Blackwellization. The authors detail the steps of the marginalisation procedure for popular classes of state-space models, such as conditional linear Gaussian models or finite-space Hidden Markov models.

Particle Filters with embedded MCMC.

When the state-space of the dynamic models has high dimensionality, it can be beneficial to integrate particle filters with MCMC methods to ensure the accuracy of the estimation. Also, the MCMC step can improve the degeneracy issue of the filter and ensure the diversity of generated particles. The MCMC is added to either the sampling or resampling step of the sequential procedures. (Burzuini et al. 1997) combines the SIS step with the Metropolis-Hastings algorithm instead of the resampling step. The proposal function in a given time step acts as the stationary distribution of the underlying Markov chain and defines its transition kernel. The Gibbs sampler, as an update step, is employed in (MacEachern et al. 1999) and is argued by (Fearnhead 1998) as a move which should lead to increased efficiency by the Rao-Blackwellisation argument. (Pitt and Shephard 1999) adapts MCMC to sequential filtering, with auxiliary variables included in both true density and a proposal function.

The resample move algorithm of (Gilks and Berzuini 2001) uses MCMC to improve the diversity among generated particles. The additional step has a final decision; whether a state vector moves to a new position. The invariant distribution of the kernel corresponds to the posterior distribution of the filtering problem. What is more, the constructed Markov chain is not assumed to be ergodic.

MCMC Filters with SMC.

In the context of particle filters, it is worth mentioning emerging solutions introduced by (Andrieu and Doucet 2010), which address the problem of low efficiency of MCMC filters applied to high-dimensional problems. The procedures belong to the class of off-line methods which use SMC techniques to decrease the computational complexity of the MCMC methods. The core idea of the frameworks is to apply SMC methods to target the full posterior density of the filtering problem by a sequence of approximations over time. The obtained product serves as a proposal function to construct an underlying Markov chain. The authors illustrate this reasoning on the Metropolis-Hastings sampler, which is simplified to a two-step iterative procedure. At each iteration, the filtered values of the state vector are sampled from the full importance distribution obtained from an SMC method. The candidates are accepted with the probability corresponding to the discrepancy between the approximated marginalised likelihood functions in current and previous iteration steps, also provided by the sequential sampling step.

The concept of (Andrieu and Doucet 2010) to adapt SMC methods to improve the computational intensiveness of the MCMC filter can also be applied to parameter learning in Bayesian formulations of general state-space models. The authors discuss the Metropolis-Hastings

sampler with the update step formulated as a product of a proposal function for the marginal probability density of the unknown parameters and a posterior density of the states. SMC methods are used to approximate the full posterior density of the state vector and the likelihood function, which are both conditioned on the vector of parameters. The only degree of freedom is the choice of the proposal distribution for the parameters. If the conditional distributions of the parameters are feasible to sample from, one can overcome this difficulty by replacing the Metropolis-Hastings algorithm with the Gibbs sampler. The Gibbs sampler is a well-known framework which exploits the structure of dynamic models in the context of parameter learning. The steps of this alternative filter are also provided in (Andrieu and Doucet 2010).

(Peters 2005) and (Del Moral et al. 2006) generalised SMC methods into Sequential Monte Carlo Samplers and proposed a procedure which samples sequentially from a sequence of distributions which lie on a common space, similarly to the problems defined by expensive MCMC. In contrast, the measurable space of the standard SMC increases with every time step.

Bayesian Parameter Learning by Sequential Monte Carlo Methods

The pioneering work of (Kitagawa 1996) shows how to construct an approximation to a marginalised likelihood function of a general state-space model by the likelihood of the particles to obtain a frequented estimation of static parameters. The procedure may not be especially efficient, since the maximisation of the likelihood function will require multiple runs of the Particle Filter, equal to the number of steps of numerical maximisation algorithms. The other solution is to adopt the discretization of the parameters' space and perform the filtering on the pre-specified grid of static parameters. Both of these solutions might become more expensive with the increase of dimensionality of the parameter space. One can address this problem by imposing a dynamic on the vector of static parameters as in (Kitagawa 1998) by augmenting the state-space. The study of (Liu and West 2001) explores the artificial time evolution of the parameters proposed by (Kitagawa 1998) to mitigate the degeneracy issue of particle filters with unknown parameters. However, this application is shown to produce parameter estimates which are largely determined by the most recent samples of the parameter vector and need a large sample. The sequential filter proposed by (Liu and West 2001) combines kernel-smoothing methods with the Auxiliary Particle Filter of (Pitt and Shephard 1999). The use of kernel methods in the context of particle attrition originates from the work of (Gordon et al. 1993), where the authors populate the particles of the state-space vectors with the addition of small random corruption.

Another idea is to treat the parameter estimation in the Bayesian setting. Sequential parameter learning follows the intuition of SMC. Standard particle filters struggle to estimate parameters accurately, since the extended state process is not ergodic, which implies that the Markovian conditions are not satisfied. The algorithms explore the parameter space only at the initialisation step, when the initial values of the parameters are sampled from a prior distribution. Consequently, the approximation of the posterior distributions of the parameters may collapse to a single value which corresponds to one of the initial candidates. Therefore, the presence of sample impoverishment is even more frequent for these inference problems.

Particle filters with MCMC are especially applicable to Bayesian inference with parameter learning in dynamic state-space models. The general idea behind the algorithm of (Gilks and Berzuini 2001) is to slow down the progressive impoverishment of the particle representation, which is even more significant when unknown parameters are included in a dynamic model. A special case of the algorithm is described in (Chopin 2002) in the context of parameter learning in static Bayesian models. The study of (Polson et al. 2008) expresses the filtering distribution by a mixture of smoothing distributions with some predetermined lag. The authors derive a lag-smoothing algorithm based on the algorithm of (Carter and Kohn 1994a) for sequential parameter learning and state vector filtering.

If possible, the Rao-Blackwellization concept is applied to integrate out the unknown static parameters of a dynamic model and employ their sufficient statistics instead. The idea has appeared independently in various simulation-based works. The study of (Strovik 2002) applies Rao-Blackwellization to the SIR filter and reformulates posterior distributions investigating static parameters using their sufficient statistics. The author compares the new framework with its standard version and a full MCMC filter based on the Gibbs sampler. The results are illustrated on various classes of dynamic models, such as linear partial Gaussian state-space models with t-Student innovations, or generalised linear models with a state vector which belongs to the exponential family of distributions. (Djuric and Miguez 2002) illustrated the marginalisation concept of Bayesian inference and parameter learning through SIS applied to non-linear Gaussian models with additive noise. It is further extended by (Saha et al. 2010) to integrate out the unknown parameters in the SIR filter. (Fearnhead 2002) demonstrates how to adapt sufficient statistics within particle filters with embedded MCMC in the context of parameter estimation. The author extends the algorithms of (Gilks and Berzuini 2001) and (Pitt and Shephard 1999) to distributions with unknown static parameters and shows how to marginalise out random parameters and provide only summarised information to the embedded MCMC steps.

1.3. Contributions

The thesis contributes to the existing literature in two respects; by extending existing methodological developments on feature extraction by linear projections and by real-life application studies which address important questions of the financial and insurance markets. Therefore, the contributions of the conducted research can be divided into two groups: methodological contributions and application-related contributions.

1.3.1. Methodology Related Contributions

In the first part of this thesis, we introduce two classes of solutions which improve the robustness of PPCA to perturbations in an observation set. The common objective of all investigated frameworks is to find directions within the dataset which explain the highest portion of its variability in the presence of missing observations. The first approach adopted in our study explores a deterministic enhancement of the existing PPCA methods. Under the assumptions of the Gaussian distribution, we study the application of robust moment estimation techniques

defined in (Maronna 1976) and (Huber and Ronchetti 2009). The class of estimators considered decreases the sensitivity of the mean and covariance estimation in the iterative steps of the EM algorithm for the Gaussian PPCA model of (Tipping and Bishop 1999b). Therefore, the crucial components of the estimation framework become less vulnerable to outlying data points and better reflect the information about the true distribution of measurements and latent variables. In addition, we study the effect of robust normalisation of the data as the initialisation step of the EM algorithm.

The remaining extensions to the classical PPCA method belong to the class of probabilistic robustification techniques, since the developed procedures directly change the assumptions about the process of generating observation data. The thesis investigates various extensions to the t-Student PPCA model and its efficient estimation frameworks in the presence of missing data. The described procedures present a general framework one may adopt to incorporate various probabilistic assumptions about the PPCA model.

We begin with a novel, identical and conditionally independent t-Student PPCA method in which the marginal density over the data follows a multivariate t-Student distribution. The original framework of (Lange et al. 1989), (Archambeau et al. 2006), (Fang and Jeong 2008) or (Chen et al. 2009) forces error terms and latent random vectors to be independently distributed over time. We relax this assumption and argue that it limits the random vectors to have time-specific tail dependence, and therefore, time-specific probability density functions. We show that, by requiring random vectors to be identically distributed, their realisations become only conditionally independent over time. The thesis studies the impact of the assumptions on the statistical properties of the extracted features.

Following that, we introduce a new class of PPCA models which addresses a wider range of limitations of existing feature extraction methods. Unlike previous robust approaches to PPCA and PCA, we propose a probabilistic framework, which has significant flexibility in the model's assumptions of considered data, such as various sources of perturbations, different marginal profiles of tail dependence or asymmetric distributions. To achieve this flexibility, the Grouped t-copula of (Daul et al. 2003) and (Luo and Shevchenko 2012) and Skew t-Student of (Azzalini and Capitanio 2003) or (Gupta 2003) are combined in the probabilistic feature extraction model. We investigate two variants of the introduced class of models, called the Generalized Skew t-Student PPCA (GStS PPCA): in these, the realisations of unobserved components are either independent or only conditionally independent over time. Furthermore, the defined framework can be simplified to the following special cases: (1) the generalized t-Student PPCA, which allows for different assumptions about the profiles of heavy tails of the error term and the latent vector; (2) the Skew t-Student PPCA, which extends the generalized t-Student PPCA by accounting for asymmetric distributions of unobserved components; (3) Grouped t-Student PPCA, which extends the generalized t-Student PPCA by allowing the marginal elements of unobserved vectors to have individual or grouped profiles of heavy-tails dependency structures. The combination of these three special cases forms the class of the Generalized Skew t-Student PPCA.

Our next contribution is the derived efficient Expectation-Maximisation (EM) algorithm of (Dempster et al. 1977) for parameter estimation, which, within the class of elliptical families, ensures convergence to a local maximum. The standard EM algorithm iterates over two steps: the expectation step (E-step), which defines a loss function between the old and new candidates of the model's parameters, and the maximisation step (M-step), which finds new parameters to optimise the loss function. In contrast to the standard Gaussian PPCA and t-Student PPCA, the steps of the EM algorithm of GStS PPCA do not have explicit formulas. The integration problems, solutions of which formulate the E-step and M-step, are not obtained in closed forms. We propose a two-dimensional quadrature-based approximation which avoids the computational intensiveness of the studied estimation algorithm. Furthermore, if the maximisation of the loss function is not trivial or is infeasible, we introduce a gradient-based optimisation step within the M-step to find the new updates of the static parameters. The framework handles the presence of missing data, and we comment how the procedure can be adjusted to various assumptions about the patterns of missing data.

1.3.2. Application Related Contributions

In the second part of the thesis, we illustrate how to use introduced feature extraction frameworks for econometric studies on big financial datasets. Our first contribution is extending the state-space structure of the Nelson-Siegel Yield Curve model to incorporate additional factors efficiently. We discuss the methodology of incorporating meaningful market information into an econometric model of a yield curve. We show how to extend the model which calibrates and forecasts the Euro Libor yield curve in order to utilise available market information. Such a model can serve as a useful tool for a parsimonious stress-testing exercise. The introduction of the methodology is followed by a discussion of an optimal estimation technique based on the Kalman Filter, studied by (Kalman and Bucy 1961). We perform comprehensive selection and structuring of state-space models. The in-sample and out-of-sample analyses confirm that our methodology outperforms the standard method.

The utilisation of market information in yield-curve modelling is well documented in the current literature. In the study of (Diebold et al. 2006), the authors extended the dynamic Nelson-Siegel model of (Nelson and Siegel 1987), by including observable macroeconomic variables in order to investigate the interaction between the macroeconomy and a yield curve. Nevertheless, given current availability of the data, we find this approach disadvantageous for two reasons, as mentioned earlier: the high possibility of over parametrisation of the model and hidden links between various data sets, which may spoil the modelling quality. Therefore, instead of incorporating raw data, we discuss how to use the derived market features. Our attention is focused on investigating features which measure Euro-Zone market conditions and their decomposition into their influences from the domestic and international markets.

The Euro Libor yield curve represents the Euro-zone market. We use the following indicators of the economic condition of the Euro Zone: the open interest of Euro-Bund futures, which are considered the benchmark for long-term Euro-denominated government debts; the spread between

3-month Eur Repo and 3 month Eurobor as a proxy for short-term liquidity; and 5-year Markit iTraxx, a credit proxy of the Euro-zone as the most actively traded contract at 5Y maturity, which comprises 125 equally-weighted credit default swaps on investment-grade European corporate bonds. We collect the data from a total of 12 leading domestic and global economies, such as sovereign yield curves, inflation yield curves, foreign exchange rates and various country-specific productivity proxies. The study presents an extensive analysis of the real data for what we believe is the most complete analysis of yield-curve modelling to date in the literature, which incorporates country-specific financial and macroeconomic factors and comprehensively examines variants of the models' structure.

This thesis also explores the impact of global demography trends on the mortality experience in one country; in our case study, the United Kingdom. To achieve this in a manner suitable for incorporation in multi-age stochastic mortality models, we need to perform a parsimonious feature-extraction method in order to reduce the large dimensional sets of data to a form suitable for inclusion in such a mortality model. Therefore, we introduce a methodology which is not exclusive to one type of demographic data and is capable of handling the analysis jointly over many different exogenous variables.

Our study is split into several tasks: the first is to explain a canonical and principled approach to combining such demographic time-series data into the stochastic mortality models, for which we develop and present number of structural approaches.

The second aspect is that large demographic data sets are now available, but the naive incorporation of such features into a stochastic mortality model would result in far too many parameters to perform estimation; the models would be overfitting and would not provide good generalisation properties for out-of-sample forecast performance. Therefore, we utilise the introduced robust feature-extraction approaches to reduce dimensionality and capture core information present in the demographic data that can be more parsimoniously included in stochastic mortality models.

Another challenge is the issue of uncertainty of parameters in mortality modelling, which we address by adapting the Bayesian Inference framework based on efficient Markov Chain Monte Carlo as in (Fung et al. 2018). The estimation of the model is achieved via a Rao-Blackwellised Gibbs sampler. We sample the static parameters via conjugate Gibbs sampling steps, which are followed by the Forward-Backward filtering sampler for state variables to inference from the resulting posteriors.

1.4. Thesis Structure

The remaining thesis is partitioned into three main parts, which are structured as follows.

Part I: Parsimonious Feature Extraction via Probabilistic Principal Component Analysis.

Part I of the thesis develops advanced probabilistic models of PPCA. Chapter 2 provides a detailed introduction to the concept of Principal Component Analysis (PCA), its links to the factor

analysis and extensions to the classical Probabilistic Principal Component Analysis (PPCA), for which assumptions and the estimation framework are given in the last section of the chapter.

We demonstrate our extensions to the classical PPCA in Chapter 3 under the assumption that the observation vector of the response variable is complete; that is, there are no missing entries. We start with the simple robustification of the classical PPCA, denoted as Gaussian PPCA, by employing robust estimators of mean and covariance. The background of the estimation using robust statistical theory as well as the estimation steps of the Robust Gaussian PPCA are provided in Section 3.1.

Probabilistic extensions to the Gaussian PPCA are investigated in Section 3.2 and Section 3.4 where we introduce the classes of PPCA methodologies which assume t-Student distributions or Generalized Skew t-Student distributions of the underlying data. We distinguish between two cases; when the observation vectors are independent over time or conditionally independent but identically distributed over time. The corresponding estimation frameworks are provided separately for each case. The derived feature extraction methodologies are further extended to incomplete data settings in Chapter 4.

Part II: State Space Models.

The second part focuses on the modelling framework that is used to include the information, extracted by the dimensionality reduction methods, derived in Part I. We revisit several methodologies which apply to the inference of models that are relevant to future extensions to model formulation and estimation methods developed in this thesis. We define the class of models which we study in this thesis in Chapter 5.

Chapter 6 discusses the concept of a general state-space representation and briefly reviews key ideas behind its optimal estimation approach. It is followed by a discussion of the framework and the properties of the Kalman Filter in Chapter 7. The modelling framework is further extended to Bayesian formulations in Chapter 8 with the background on the relevant estimation methods based on Monte Carlo techniques given in Chapter 10.

Part III: Applications.

The real data case studies conducted are summarized in Part III. The first real data study is presented in Chapter 11, where we analyse the implications of interactions in international financial and macroeconomic large data sets on the dynamic of Libor yield curve over time. Chapter 12 demonstrates the utilization of international demographic data sets in modelling the United Kingdom mortality rate. Each application related includes a corresponding literature review and objectives of the problems investigated, as well as descriptions of the modelling techniques employed.

The remaining chapter of this thesis, Chapter 13, concludes and reflects on the future direction of the research on the topics investigated.

Proofs of the derived methodologies, as well as the detailed data description used in the application studies, are provided in the Appendix.

Part I

Parsimonious Feature Extraction via Probabilistic Principal Component Analysis

Summary

In the following part we introduce progressively the feature extraction methods that should be considered and have been developed to deal with real data issues such as missing data and outliers which may affect the feature extraction process.

Chapter 2 reviews the concept of feature extraction techniques by means of deterministic Principal Component Analysis (PCA) and their extension to the probabilistic setting. We revisit the fundamental ideas behind PCA in both deterministic and probabilistic paradigms in order to discuss Probabilistic Principal Component Analysis (PPCA) defined by (Tipping and Bishop 1999b) which serves as a background knowledge to the novel robust extensions introduced in this thesis.

The other chapters of Part I focus on deriving feature extraction methodologies which adapt PPCA and extend the framework in order to incorporate various assumptions about the data which corrupt the standards methodology such as missingness, skewness, outliers and noise and their non-trivial dependency structures.

Firstly, Section 3.1 demonstrates how to utilize the concept of robust mean and covariance estimation within the Gaussian PPCA framework. Next, we move our attention to the extensions which directly implicate the distribution assumptions about the PPCA model.

We introduce a class of models which imposes different enhancements to PPCA using the multivariate t-Student distribution and its generalizations. All outlined methodologies assume various structures of heavy tails and hence are robust to the distant observations in a sample sets outliers which falsify the results obtained by standard PPCA. We investigate two separate frameworks, (1) one which assumes the observations to be independent over time, and consequently results in time-specific probability density functions of the latent variables, and (2) the second one which enforces the identical distribution of the latent variables over time and consequently, relaxes the assumption on time independence of subsequent observations.

The justifications and detailed assumptions of the determined class of robust PPCA models as well as the corresponding estimation procedures by means of Expectation-Maximisation algorithms in complete data setting are described in Chapter 3. Chapter 4 provides the estimation procedures for the class of robust PPCA methods when some components of the observation set are missing. The proofs to the derived properties are located in Appendix A.

2. Introduction to Probabilistic Principal Component Analysis

In order to infer from high dimensional data, we resort to feature extraction methods based on minimizing some pre-specified projection pursuit index. As mentioned in Chapter 1, our attention is placed to linear methods of dimensionality reduction, more precisely, those expressible as linear projections as defined in (Friedman and Tukey 1974) which includes Principal Component Analysis (PCA) and its extensions and robust alternatives. In the application studies in Part III we focus on the practical aspects of PPCA and utilize the basis vectors which span the projected lower rank space as the most meaningful factors in terms of the volatility of the considered data set.

2.1. Principal Component Analysis Revisited: Non-Stochastic Principal Component Analysis

Let us denote the $N \times d$ matrix $\mathbf{Y}$ to represent the original data set, where a row of the matrix is a single d - dimensional observation in a given moment of time. The goal of Principal Component Analysis is to identify the most meaningful unit length basis to re-express a data set $\mathbf{Y}$. The purpose of a new basis is to better filter out the noise and reveal hidden structure (orthonormal basis). Therefore, PCA looks for the given projection of the observation data

$$\mathbf{Y}_{N\times d}\mathbf{W}_{d\times d} = \mathbf{X}_{N\times d} \tag{2.1}$$

where $\mathbf{W}$ is a $d \times d$ matrix and denotes a linear projection. The columns of $\mathbf{W}$ are the new basis vectors, that is $\mathbf{W}^T\mathbf{W} = \mathbb{I}_d$, and express rows of $\mathbf{X}$.

The goal of re-expressing $\mathbf{Y}$ in such a manner can be interpreted in the sense that PCA aims to lower the redundancy in the data set by removing the linear dependencies. In mathematical terms, the goal can be written for $i,\ j$ columns of $\mathbf{X}$

$$[\mathbf{X}]_{\cdot,i}^T [\mathbf{X}]_{\cdot,i} = [\mathbf{W}]_{\cdot,i}^T \mathbf{C}_Y [\mathbf{W}]_{\cdot,i} \text{ and } [\mathbf{X}]_{\cdot,i}^T [\mathbf{X}]_{\cdot,j} = [\mathbf{W}]_{\cdot,i}^T \mathbf{C}_Y [\mathbf{W}]_{\cdot,j} = 0$$

where $\mathbf{C}_Y = \mathbf{Y}^T\mathbf{Y}$. We seek a linear combination given by Equation (2.1) that maximizes the overall variance of $\mathbf{X}$ given by $\mathbf{C}_X = \mathbf{X}^T\mathbf{X}$. The solution to the problem is found by a maximiser of the following Lagrangian expression given by

$$Q(\mathbf{W}) = \mathbf{W}^T\mathbf{C}_Y\mathbf{W} - \mathbf{\Lambda}\left(\mathbf{W}^T\mathbf{W} - \mathbb{I}_d\right),$$

for $\mathbf{\Lambda}_{d\times d}$ being a diagonal $d \times d$ matrix with Lagrangian coefficients. One can then solve optimization by differentiating and finding turning points to solve for the roots of the quadratic form of this objective function as following

$$\frac{\partial Q}{\partial \mathbf{W}} = 2\mathbf{C}_Y\mathbf{W} - 2\mathbf{\Lambda}\mathbf{W} = 0 \Rightarrow \mathbf{C}_Y\mathbf{W} = \mathbf{\Lambda}\mathbf{W}$$

We see that $\mathbf{W}$ is a matrix in which columns are eigenvectors of $\mathbf{C}_Y$, whereas $\mathbf{\Lambda}$ is a matrix of corresponding eigenvalues with the number of the non-zero elements equal to the rank of $\mathbf{C}_Y$. The columns of $\mathbf{X}$ are indeed orthogonal since

$$[\mathbf{X}]_{\cdot,i}^T [\mathbf{X}]_{\cdot,j} = [\mathbf{W}]_{\cdot,i}^T \mathbf{C}_Y [\mathbf{W}]_{\cdot,j} = [\mathbf{W}]_{\cdot,i}^T \lambda_j [\mathbf{W}]_{\cdot,j} = \lambda_j [\mathbf{W}]_{\cdot,i}^T [\mathbf{W}]_{\cdot,j} = \mathbf{0},$$

and correspond to unequal eigenvalues. It is easily proven that $\mathbf{X}$, defined by $\mathbf{W}$ - the eigenvectors of $\mathbf{C}_Y$, maximizes the total trace of $\mathbf{C}_X$, its determinant and maximizes the Euclidean distance between the columns of $\mathbf{X}$, see (Basilevsky 1994). Also, the representation minimizes the mean square error between the observation and its projection as it is equivalent problem to maximizing the variance of $\mathbf{X}$.

We wish to find estimates of $\mathbf{W}$ and $\mathbf{X}$ which minimizes the sum of squares, $\boldsymbol{\epsilon} = \mathbf{Y} - \mathbf{X}\mathbf{W}^T$, both of $\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}$ and $\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T$. Assuming that the residuals have homogeneous covariance matrix, that is $\boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = \sigma^2\mathbb{I}_d$ we have

$$Q(\mathbf{W},\mathbf{X}) = \boldsymbol{\epsilon}^T\boldsymbol{\epsilon} = \sigma^2\mathbb{I}_d = \left(\mathbf{Y} - \mathbf{X}\mathbf{W}^T\right)^T \left(\mathbf{Y} - \mathbf{X}\mathbf{W}^T\right) = \mathbf{Y}^T\mathbf{Y} + \mathbf{W}\mathbf{X}^T\mathbf{X}\mathbf{W}^T - \mathbf{W}\mathbf{X}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}\mathbf{W}^T. \tag{2.2}$$

Since both $\mathbf{W}$ and $\mathbf{X}$ are treated as parameters to be estimated, we minimize Equation (2.2) by computing the partial derivatives of function Q with respect to $\mathbf{W}$ and $\mathbf{X}$ and setting them to zero

$$\frac{\partial Q}{\partial \mathbf{W}} = -2\mathbf{Y}^T\mathbf{X} + 2\mathbf{W}\mathbf{X}^T\mathbf{X} = 0$$

and since by construction $\mathbf{Y}\mathbf{W} = \mathbf{X}$ we obtain that

$$\mathbf{Y}\mathbf{Y}^T\mathbf{X} = \mathbf{Y}\mathbf{W}\mathbf{X}^T\mathbf{X} \;\Rightarrow\; \mathbf{Y}\mathbf{Y}^T\mathbf{X} = \mathbf{X}\mathbf{X}^T\mathbf{X}.$$

As we are looking for uncorrelated explanatory variables, for $\mathbf{\Lambda} = \mathbf{X}^T\mathbf{X}$ we get

$$\mathbf{Y}\mathbf{Y}^T\mathbf{X} = \mathbf{X}\mathbf{\Lambda}$$

which shows that $\mathbf{X}$ and $\mathbf{\Lambda}$ are eigenvectors and eigenvalues of the $N \times N$ matrix $\mathbf{Y}\mathbf{Y}^T$. What is more, differentiating Q with respect to $\mathbf{X}$ gives

$$\frac{\partial Q}{\partial \mathbf{X}} = -2\mathbf{Y}\mathbf{W} + 2\mathbf{X}\mathbf{W}^T\mathbf{W} = 0.$$

By applying $\mathbf{YW} = \mathbf{X}$, we obtain that

$$\mathbf{Y}^T\mathbf{YW} = \mathbf{W}\mathbf{\Lambda},$$

what shows that $\mathbf{W}$ and $\mathbf{\Lambda}$ are eigenvectors and eigenvalues of the $d \times d$ covariance matrix of $\mathbf{Y}$, $\mathbf{C}_Y = \mathbf{Y}^T\mathbf{Y}$.

2.2. Extending Principal Component Analysis to the Probabilistic Setting

Principal Component Analysis can also be considered from the population distribution point of view, i.e., instead of the deterministic sample matrix of observations $\mathbf{Y}$, we instead consider a $d-$dimensional random vector $\mathbf{Y}_t$ which is linearly transformed into uncorrelated d-dimensional random variable $\mathbf{X}_t$. **At this stage the Principal Component Analysis does not require any assumption about the distribution of a random vector $\mathbf{Y}_t$.** The only assumption we make refers to the projection matrix $\mathbf{W}$ and where we require its orthonormality. If the random vector $\mathbf{Y}_t$ has a known mean equal to zero and covariance matrix $\mathbf{C}_Y$, the model transforms to

$$\mathbf{Y}_t\mathbf{W} = \mathbf{X}_t \tag{2.3}$$

and implies $\mathbf{X}_t$ is a d-dimensional multivariate random variable with a diagonal covariance matrix $\mathbf{\Lambda}$. If in addition we assume $\mathbf{Y}_t$ to be normally distributed, the lack of correlation imposes independence.

If we consider N realisations of the random variable $\mathbf{Y}_t$, $\mathbf{y}_{1:N} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$, which are placed in rows of the $N \times d$ matrix $\mathbf{Y}$, we have an algebraic problem as introduced in Section 2.1. The conceptual difference is that in the case of stochastic PCA we work with an estimator of a covariance matrix, e.g., a sample estimator $\mathbf{S}_Y = \frac{1}{N}\mathbf{Y}^T\mathbf{Y}$.

2.3. Extending Stochastic Principal Component Analysis to Factor Analysis

In this section we no longer assume the underlying process to be perfectly observed as would be the assumption typically made in the stochastic version of PCA above. The implication of this can be interpreted as follows: we no longer assume that the underlying time series of the considered data set is perfectly observed with no observation error. Instead there is an observation error present and the covariance matrix used in the PCA (deterministic or stochastic-population estimator based analysis) no longer explains all variation in the response or the time series. This is practically important to consider in feature extraction in practice. In this section we briefly introduce this relaxation and show its relationship to stochastic PCA above.

To discuss the PCA by means of Factor Analysis we need to introduce an additional notation and variable to our model, that is, an d-dimensional error term, $\boldsymbol{\epsilon}_t$, and rewrite Equation (2.3) as

$$\mathbf{Y}_t = \mathbf{X}_t\mathbf{W}^T + \boldsymbol{\epsilon}_t,$$

where $\mathbf{Y}_t$, $\mathbf{X}_t$ and $\boldsymbol{\epsilon}_t$ are d-dimensional random vectors and $\mathbf{X}_t$ and $\boldsymbol{\epsilon}_t$ are mutually independent. Given N realisations of the random vectors, which are placed in the rows of the matrices $\mathbf{Y}$, $\mathbf{X}$, $\boldsymbol{\epsilon}$ respectively, the above problem has the following matrix form

$$\mathbf{Y}_{N\times d} = \mathbf{X}_{N\times d}\mathbf{W}^T_{d\times d} + \boldsymbol{\epsilon}_{N\times d}.$$

Factor analysis assumes the diagonal covariance structure of $\boldsymbol{\epsilon}_t$. It differs from the PCA model discussion from the previous sections as the components given by $\mathbf{X}_t$ and $\mathbf{W}$ accounts for correlation between elements of $\mathbf{Y}_t$ and only part of the variation (in standard PCA $\mathbf{X}_t$ and $\mathbf{W}$ account for the entire variance) since

$$\mathbb{E}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right] = \mathbb{E}\left[\left(\mathbf{X}_t\mathbf{W}^T + \boldsymbol{\epsilon}_t\right)^T\left(\mathbf{X}_t\mathbf{W}^T + \boldsymbol{\epsilon}_t\right)\right] = \mathbf{W}\boldsymbol{\Lambda}\mathbf{W}^T + \boldsymbol{\Psi}.$$

If we assume multivariate distribution of $\mathbf{X}_t \sim \mathcal{N}(\mathbf{0}, \mathbb{I}_d)$ and $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi})$ we obtain conditional independence of $\mathbf{Y}_t$ given latent variable $\mathbf{X}_t$, i.e.,

$$\mathbf{Y}_t|\mathbf{X}_t, \mathbf{W}, \boldsymbol{\Psi} \sim \mathcal{N}\left(\mathbf{X}_t\mathbf{W}^T, \boldsymbol{\Psi}\right). \tag{2.4}$$

as $\boldsymbol{\Psi}$ is diagonal. Recall that the variable $\mathbf{X}_t$ reproduces all correlations between components of $\mathbf{Y}_t$. Imposing normality assumptions on $\mathbf{Y}_t$ and $\mathbf{X}_t$ enables performing the maximum likelihood estimation of $\mathbf{X}_t$, $\mathbf{W}$ and $\boldsymbol{\Psi}$ with optimality properties.

The marginal distribution of $\mathbf{Y}_t$ is then calculated by the integration of the joint distribution of $\mathbf{Y}_t$ and $\mathbf{X}_t$ (which is given via chain rule)

$$\begin{aligned}
&\pi_{\mathbf{Y}_t,\mathbf{X}_t|\mathbf{W},\boldsymbol{\Psi}}(\mathbf{y}_t, \mathbf{x}_t) = \pi_{\mathbf{Y}_t|\mathbf{X}_t,\mathbf{W},\boldsymbol{\Psi}}(\mathbf{y}_t)\ \pi_{\mathbf{X}_t|\mathbf{W},\boldsymbol{\Psi}}(\mathbf{x}_t) \text{ and hence} \\
&\pi_{\mathbf{Y}_t,\mathbf{X}_t|\mathbf{W},\boldsymbol{\Psi}}(\mathbf{y}_t, \mathbf{x}_t) = (2\pi|\boldsymbol{\Psi}|)^{-d}\,|\boldsymbol{\Psi}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\mathbf{y}_t - \mathbf{x}_t\mathbf{W}^T)\boldsymbol{\Psi}^{-1}(\mathbf{y}_t - \mathbf{x}_t\mathbf{W}^T)^T - \frac{1}{2}\mathbf{x}_t\mathbf{x}_t^T\right\}
\end{aligned} \tag{2.5}$$

with respect to the random variable $\mathbf{X}_t$, using the assumptions on the random vector $\mathbf{X}_t$ and the conditional distribution of $\mathbf{Y}_t$ given in Equation (2.4), that is

$$\pi_{\mathbf{Y}_t|\mathbf{W},\boldsymbol{\Psi}}(\mathbf{y}_t) = \int_{\mathbb{R}^d}\pi_{\mathbf{Y}_t,\mathbf{X}_t|\mathbf{W},\boldsymbol{\Psi}}(\mathbf{y}_t, \mathbf{x}_t)d\mathbf{x}_t = (2\pi)^{-\frac{d}{2}}\,|\mathbf{C}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\mathbf{y}_t\mathbf{C}^{-1}\mathbf{y}_t^T\right\} \tag{2.6}$$

for $\mathbf{C} = \mathbf{W}\mathbf{W}^T + \boldsymbol{\Psi}$ where $|\mathbf{C}|$ denotes the determinant of the matrix. Hence, $\mathbf{Y}_t|\mathbf{W}, \boldsymbol{\Psi} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{W}\mathbf{W}^T + \boldsymbol{\Psi}\right)$. Notice that since $\boldsymbol{\Psi}$ is diagonal, the correlation structure between components $\mathbf{Y}_t$ is specified by the matrix $\mathbf{W}$.

2.3.1. Link to Principal Component Analysis

If we assume that the error term $\boldsymbol{\epsilon}_t$ is homogeneous, that is $\boldsymbol{\Psi} = \sigma^2 \mathbb{I}_d$ for $\sigma^2 > 0$, then the problem of finding $\mathbf{W}$ by means of PCA given $\mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2 \mathbb{I}_d$ is identifiable (see further discussions in (Tipping and Bishop 1999b).

Having the eigendecomposition of the covariance matrix, $\mathbf{C} = \mathbf{U}_{d\times d}\mathbf{L}_{d\times d}\mathbf{U}^T$, for diagonal matrix $\mathbf{L}$ and orthonormal matrix $\mathbf{U}$, we have

$$\mathbf{0} = (\mathbf{C} - \mathbf{L})\mathbf{U} = \left(\mathbf{W}^T\mathbf{W} + \sigma^2\mathbb{I}_d - \mathbf{L}\right)\mathbf{U} = \left(\mathbf{W}\mathbf{W}^T - \left(\mathbf{L} - \sigma^2\mathbb{I}_d\right)\right)\mathbf{U}.$$

Thus, the matrix $\boldsymbol{\Lambda} = \left(\mathbf{L} - \sigma^2\mathbb{I}_d\right)$ and $\mathbf{U}$ are matrices of eigenvalues and corresponding eigenvectors of $\mathbf{W}\mathbf{W}^T$. Since $\lambda_i = l_i - \sigma^2 \geq 0$, the scalar σ^2 can be chosen as the smallest diagonal element of $\boldsymbol{\Lambda}$. Then the factors loadings are given by $\mathbf{P} = \mathbf{U}\boldsymbol{\Lambda}^{\frac{1}{2}}$.

2.3.2. PCA as a Limiting Case of Factor Analysis

The assumption of the isotropic error term is crucial in order to establish the link between Factor Analysis and PCA. Standard derivation of PCA does not account for any error term. However, we can perceive PCA as a limiting case for $\sigma^2 \to 0$. Then, as noted in (Roweis 1998), PCA is a limiting case of the linear Gaussian model as the covariance matrix becomes infinitesimally small and equal in all directions. This has an effect that the likelihood of a point $\mathbf{Y}_t$ is dominated by the squared residuals between the observation and its projection $\mathbf{X}_t\mathbf{W}^T$. As the σ^2 tends to zero, the posteriori over states $\mathbf{X}_t$ collapses to a single point and its covariance becomes zero since

$$\mathbf{X}_t|\mathbf{Y}_t, \mathbf{W}, \sigma^2 \sim \mathcal{N}\left(\mathbf{y}_t\mathbf{W}\mathbf{M}^{-1}, \sigma^2\mathbf{M}^{-1}\right) \xrightarrow{\sigma^2\to 0} \delta\left(\mathbf{X}_t - \mathbf{Y}_t\mathbf{W}\left(\mathbf{W}^T\mathbf{W}\right)^{-1}\right)$$

for $\mathbf{M} = \mathbf{W}^T\mathbf{W} + \sigma^2\mathbb{I}_k$. The form of conditional probability $\mathbf{X}_t|\mathbf{Y}_t, \mathbf{W}, \sigma^2$ is justified in the next section.

2.4. Probabilistic Principal Component Analysis with Gaussian Distribution

The combination of PCA and the Factor Analysis leads one to then naturally consider the framework referred to as "Probabilistic PCA" which can be estimated via Expectation Maximisation algorithm as introduced by (Tipping and Bishop 1999b) and described below. This is an exceptionally efficient and numerically stable approach to apply in practice which provides straightforward framework to deal with missing observations and various assumptions about them, however, this property of the framework is broadly discussed in Chapter 4.

In this section further, we seek k - dimensional uncorrelated latent vector $\mathbf{X}_t$ which provides the most meaningful model of $\mathbf{Y}_t$,

$$\mathbf{Y}_t = \boldsymbol{\mu} + \mathbf{X}_t\mathbf{W}_{d\times k}^T + \boldsymbol{\epsilon}_t, \tag{2.7}$$

for a vector of constants $\boldsymbol{\mu} \in \mathbb{R}^d$ and d-dimensional error term random vector $\boldsymbol{\epsilon}_t$. **We aim for $\mathbf{W}^T\mathbf{W} = \mathbb{I}_k$ (i.e. orthonormality of the projection matrix), however it is not assumed in the estimation process**. We assume the multivariate normal priori distributions of the k dimensional latent variable $\mathbf{X}_t \sim \mathcal{N}(0, \mathbb{I}_k)$ and the error term $\boldsymbol{\epsilon}_t \sim \mathcal{N}\left(0, \sigma^2\mathbb{I}_d\right)$.

The Gaussian PPCA assumes that $\mathbf{X}_t$ and $\boldsymbol{\epsilon}_t$ are k- and d-dimensional zero mean Gaussian random vectors with covariance matrices equal to $\mathbb{I}_k$ and $\sigma^2\mathbb{I}_d$ respectively, which are mutually independent and independent over time. Given the model in Equation (2.7), $\mathbf{Y}_t$ is also $d-$ dimensional random vector which follows Gaussian distribution and is independent over time, with mean $\boldsymbol{\mu}$ and the covariance matrix $\mathbf{C} = \mathbf{W}^T\mathbf{W} + \sigma^2\mathbb{I}_d$.

Our goal is to estimate the projection matrix $\mathbf{W}$, the vector $\boldsymbol{\mu}$ and the scalar σ^2 given the marginal distribution of $\mathbf{Y}_t$

$$\mathbf{Y}_t|\Psi \sim \mathcal{N}\left(\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^T + \sigma^2\mathbb{I}_d\right),$$

for the vector of static parameters $\Psi = \left[\mathbf{W}, \boldsymbol{\mu}, \sigma^2\right]$ of the model in Equation (2.7). In order to calculate the covariance matrix we have to estimate the vector of parameters Ψ and filter realisations of the latent vector $\mathbf{X}_t$ employing the Expectation-Maximisation (EM) algorithm. The steps and derivation of the algorithm have been described in (Rubin and Thayer 1982) or (Tipping and Bishop 1999b) where no missigness in assumed. The authors use the results introduced by (Dempster et al. 1977) which we detail in the following section.

2.5. Introduction to the Expectation-Maximisation Algorithm

The Expectation-Maximisation (EM) Algorithm of (Dempster et al. 1977) is a broadly known iterative procedure which finds Maximum Likelihood Estimation (MLE) estimates of parameters in probabilistic models when the direct optimisation of a likelihood function is not feasible. Although the EM algorithm was originally a method for computing MLE estimates, it can be easily modified to find Maximum a Posterior (MAP) estimates - an equivalent estimation approach to MLE in Bayesian inference. In the following section, we explain the basic concepts behind the method in a frequentists setting. The reader is advised to see the discussion on the merits and various extensions to the EM algorithm which is provided in (Liu and Rubin 1995), (McLachlan and Krishnan 1997), (Rober and Casella 2004) or (Särkkä 2013), among many other tutorial books.

We begin with introducing the necessary notation. Let $\mathbf{Y}_t \in \mathbb{R}^d$ denotes a d-dimensional random vector of some probabilistic model with static parameters $\Psi \in \Omega$. The model yields a marginal distribution of $\mathbf{Y}_t$ which is defined by a probability density function $\pi_{\mathbf{Y}_t|\Psi}(\mathbf{y}_t)$. Given N realisations of the random vector $\mathbf{Y}_t$ in time $t = 1, \ldots, N$, denoted by $\mathbf{y}_{1:N} = \left\{\mathbf{y}_1, \ldots, \mathbf{y}_N\right\}$, the loglikelihood function of the model is specified as follows

$$l(\Psi; \mathbf{y}_{1:N}) := \log \pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N}) \tag{2.8}$$

The MLE estimates of $\Psi \in \Omega$ are computed by maximising this likelihood function.

In order to iteratively find a stationary point of the function in Equation (2.8), the EM algorithms exploits the artificial formulation of probabilistic models which regards $\mathbf{Y}_t$ as incomplete

information about the studied model. The framework assumes the existence of a random vector which represents the complete model dynamics. The random vector $\mathbf{Y}_t$ acts as an observable elements of this vector. Consequently, a studied model can be described by two components, the observed random vector $\mathbf{Y}_t$ and a hidden k-dimensional random vector $\mathbf{X}_t \in \mathbb{R}^k$. The joint distribution of the random vectors is defined by the probability density function $\pi_{\mathbf{Y}_t,\mathbf{X}_t|\Psi}(\mathbf{y}_t, \mathbf{x}_t)$.

Each iteration of the EM algorithm, which find maximizers of $l(\Psi; \mathbf{y}_{1:N})$ with respect to Ψ, consists of two steps: an expectation step (E-step) and a maximisation step (M-step). The E-step infers missing values or latent variables, $\mathbf{X}_{1:N}$, by finding their distribution given the known observed values $\mathbf{Y}_{1:N}$, and current estimates of parameters. At i-th iteration, the E-step specifies an estimate of complete information formulated as a function of parameters, that is

$$Q(\tilde{\Psi}, \Psi^*) := \mathbb{E}_{\mathbf{X}_{1:N}|\mathbf{Y}_{1:N},\tilde{\Psi}}\Big[\log \pi_{\mathbf{Y}_{1:N},\mathbf{X}_{1:N}|\Psi^*}(\mathbf{y}_{1:N}, \mathbf{x}_{1:N})\Big] \tag{2.9}$$

where Ψ^* is a candidate for MLE estimator of Ψ and $\tilde{\Psi} \in \Omega$. Next, the M-step re-estimates the parameters by optimising a complete likelihood of a model which distribution is specified in the E-step. Let us denote $\Psi^{(i)} = \tilde{\Psi}$. The optimisation problem is defined as following

$$\Psi^{(i+1)} = \underset{\Psi^* \in \Omega}{\operatorname{argmax}}\, Q(\Psi^{(i)}, \Psi^*)$$

where $\Psi^{(i+1)}$ is the updated parameter $\tilde{\Psi}$ for the next iteration. The key idea behind the steps of the algorithm is to use the following representation of the loglikelihood function in Equation (2.8), which exploits Bayes' rule applied to $\pi_{\mathbf{Y}_t,\mathbf{X}_t|\Psi}(\mathbf{y}_t, \mathbf{x}_t)$, that is

$$\begin{aligned} l(\Psi, \mathbf{y}_{1:N}) := \log \pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N}) &= \log\Big(\frac{\pi_{\mathbf{Y}_{1:N},\mathbf{X}_{1:N}|\Psi}(\mathbf{y}_{1:N}, \mathbf{x}_{1:N})}{\pi_{\mathbf{X}_{1:N}|\mathbf{Y}_{1:N},\Psi}(\mathbf{x}_{1:N})}\Big) \\ &= \log \pi_{\mathbf{Y}_{1:N},\mathbf{X}_{1:N}|\Psi}(\mathbf{y}_{1:N}, \mathbf{x}_{1:N}) - \log \pi_{\mathbf{X}_{1:N}|\mathbf{Y}_{1:N},\Psi}(\mathbf{x}_{1:N}). \end{aligned} \tag{2.10}$$

We observe that the left side of Equation (2.10) is invariant under the expectation with respect to the conditional distribution $\mathbf{X}_{1:N}|\mathbf{Y}_{1:N}, \Psi^*$ since

$$\begin{aligned} \mathbb{E}_{\mathbf{X}_{1:N}|\mathbf{Y}_{1:N},\tilde{\Psi}}\Big[l(\Psi = \Psi^*, \mathbf{y}_{1:N})\Big] &= \mathbb{E}_{\mathbf{X}_{1:N}|\mathbf{Y}_{1:N},\tilde{\Psi}}\Big[\big(\log \pi_{\mathbf{Y}_{1:N},\mathbf{X}_{1:N}|\Psi}(\mathbf{y}_{1:N}, \mathbf{x}_{1:N}) - \log \pi_{\mathbf{X}_{1:N}|\mathbf{Y}_{1:N},\Psi}(\mathbf{x}_{1:N})\big)\Big] \\ \log \pi_{\mathbf{Y}_{1:N}|\Psi=\Psi^*}(\mathbf{y}_{1:N}) &= \underbrace{\mathbb{E}_{\mathbf{X}_{1:N}|\mathbf{Y}_{1:N},\tilde{\Psi}}\Big[\log \pi_{\mathbf{Y}_{1:N},\mathbf{X}_{1:N}|\Psi^*}(\mathbf{y}_{1:N}, \mathbf{x}_{1:N})\Big]}_{Q(\Psi,\Psi^*)} - \underbrace{\mathbb{E}_{\mathbf{X}_{1:N}|\mathbf{Y}_{1:N},\tilde{\Psi}}\Big[\log \pi_{\mathbf{X}_{1:N}|\mathbf{Y}_{1:N},\Psi^*}(\mathbf{x}_{1:N})\Big]}_{H(\Psi,\Psi^*)}, \end{aligned} \tag{2.11}$$

for some vector of $\tilde{\Psi} \in \Omega$. In their study, (Dempster et al. 1977) shows that the maximizers of $\log \pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N})$ can be specified by iteratively optimising the first component of the representation in Equation (2.11), the expectation $Q(\tilde{\Psi}, \Psi^*)$ defined in Equation (2.9), using the steps of the EM algorithm. We denote $\left\{\Psi^{(i)}\right\}_{i\in\mathbb{N}_0}$ as a sequence of updates of the parameter Ψ which is specified by iterations of the EM algorithm and further referred as the EM sequence. Let $\left\{l^{(i)}\right\}_{i\in\mathbb{N}_0}$ denotes

a sequence of real numbers which is obtained by applying the loglikelihood function in Equation (2.8) to the candidates of parameters in $\left\{\Psi^{(i)}\right\}_{i\in\mathbb{N}_0}$, that is

$$l^{(i)} := l(\Psi^{(i)}; \mathbf{y}_{1:N}) \text{ for } i \in \mathbb{N}_0.$$

In their work, (Dempster et al. 1977) shows that the sequences $\left\{l^{(i)}\right\}_{i\in\mathbb{N}_0}$ are non-decreasing and consequently, each iteration of the EM algorithm results in the update of the parameter Ψ which increases the loglikelihood in Equation (2.8) or leaves it unchanged. We present this outcome in Theorem 1.

Theorem 1. *(Dempster et al. 1977) Let $\left\{\Psi^{(i)}\right\}_{i\in\mathbb{N}_0}$ be a sequence of candidates specified by subsequent iterations of the EM algorithm. The sequence $\left\{l^{(i)}\right\}_{i\in\mathbb{N}_0}$ is defined as a values of the function $l(\Psi, \mathbf{y}_{1:N})$ in Equation (2.8) applied to the elements of $\left\{\Psi^{(i)}\right\}_{i\in\mathbb{N}_0}$ such that*

$$l^{(i)} := l(\Psi^{(i)}; \mathbf{y}_{1:N}) \text{ for } i \in \mathbb{N}_0.$$

Then, the sequence $\left\{l^{(i)}\right\}_{i\in\mathbb{N}_0}$ is non-decreasing.

Proof. The proof of the theorem can be found in (Theorem 1, Dempster et al. (1977)). We will detail its most important steps. At the i-th update of the EM algorithm, the loglikelihood in Equation (2.8) would be expressed as

$$\begin{aligned} l(\Psi) := l(\Psi, \mathbf{y}_{1:N}) &= \mathbb{E}_{\mathbf{X}_{1:N}|\mathbf{Y}_{1:N},\Psi^{(i)}}\left[\log \pi_{\mathbf{Y}_{1:N},\mathbf{X}_{1:N}|\Psi}(\mathbf{y}_{1:N}, \mathbf{x}_{1:N})\right] - \mathbb{E}_{\mathbf{X}_{1:N}|\mathbf{Y}_{1:N},\Psi^{(i)}}\left[\log \pi_{\mathbf{X}_{1:N}|\mathbf{Y}_{1:N},\Psi}(\mathbf{x}_{1:N})\right] \\ &:= Q(\Psi^{(i)}, \Psi) - H(\Psi^{(i)}, \Psi). \end{aligned}$$

We will show that $l(\Psi^{(i)}) \leq l(\Psi^{(i+1)})$. Since the new candidate $\Psi^{(i+1)}$ is specified as a maximizer of $Q(\Psi^{(i)}, \Psi^*)$ with respect to Ψ^*, it is straightforward that for any $\Psi^* \in \Omega$, the following inequality holds

$$Q(\Psi^{(i)}, \Psi^{(i+1)}) \geq Q(\Psi^{(i)}, \Psi^*). \tag{2.12}$$

On the other hand, for any $\Psi^* \in \Omega$, the difference $H(\Psi^{(i)}, \Psi^*) - H(\Psi^{(i)}, \Psi^{(i)}) \leq 0$ since

$$\begin{aligned} H(\Psi^{(i)}, \Psi^*) - H(\Psi^{(i)}, \Psi^{(i)}) &= \mathbb{E}_{\mathbf{X}_{1:N}|\mathbf{Y}_{1:N},\Psi^{(i)}}\left[\log \pi_{\mathbf{X}_{1:N}|\mathbf{Y}_{1:N},\Psi^*}(\mathbf{x}_{1:N})\right] - \mathbb{E}_{\mathbf{X}_{1:N}|\mathbf{Y}_{1:N},\Psi^{(i)}}\left[\log \pi_{\mathbf{X}_{1:N}|\mathbf{Y}_{1:N},\Psi^{(i)}}(\mathbf{x}_{1:N})\right] \\ &= \mathbb{E}_{\mathbf{X}_{1:N}|\mathbf{Y}_{1:N},\Psi^{(i)}}\left[\log\left(\frac{\pi_{\mathbf{X}_{1:N}|\mathbf{Y}_{1:N},\Psi^*}(\mathbf{x}_{1:N})}{\pi_{\mathbf{X}_{1:N}|\mathbf{Y}_{1:N},\Psi^{(i)}}(\mathbf{x}_{1:N})}\right)\right] \\ &\leq \log \mathbb{E}_{\mathbf{X}_{1:N}|\mathbf{Y}_{1:N},\Psi^{(i)}}\left[\frac{\pi_{\mathbf{X}_{1:N}|\mathbf{Y}_{1:N},\Psi^*}(\mathbf{x}_{1:N})}{\pi_{\mathbf{X}_{1:N}|\mathbf{Y}_{1:N},\Psi^{(i)}}(\mathbf{x}_{1:N})}\right] = \log 1 = 0, \end{aligned}$$

where we applied Jensen's inequality to the logarithm as a convex function. The illustrated properties of the functions Q and H, prove that the loglikelihood function in Equation (2.8) is not decreased after i-th iteration of the EM algorithm. □

Therefore, the EM algorithm monotonically increases the likelihood function during each iteration. The equality $l(\Psi^{(i)}, \mathbf{y}_{1:N}) = l(\Psi^{(i+1)}, \mathbf{y}_{1:N})$ holds if and only if $Q(\Psi^{(i)}, \Psi^{(i)}) = Q(\Psi^{(i)}, \Psi^{(i+1)})$. We observe that loglikelihood will increase, when the inequality in Equation (2.12) is strict.

2.5.1. The Convergence Types for the Expectation-Maximisation Algorithm

Iterative application of an algorithm that increases evaluation of an objective function is a minimum requirement one would need for an optimisation technique such the considered EM algorithm given in Theorem 1. However, this results does not ensure that the algorithm converges or, what is more, converges to the stationary points of the likelihood function. For instance, this also does not a saddle point or an extremum of the function which could be a local or global maximum in the contact of our optimisation problem. The other convergence results of the EM algorithm are dependent on various characteristics of the Q function, or equivalently the likelihood function, such as smoothness or properties of the parameter space. Majority of them are easily satisfied for the models from the natural exponential family as shown in (Dempster et al. 1977) and (Wu 1983),however, the more general classes of models might be less straightforward to verify and satisfied required conditions.

The studies of (Dempster et al. 1977), (Wu 1983) and (Boyles 1983) investigate additional assumptions which need to be satisfied in order to ensure that the sequence $\left\{l^{(i)}\right\}_{i\in\mathbb{N}_0}$ coverages to a stationary point or, more specifically, a local or global maximum of the likelihood function in Equation (2.10).

The convergence of the EM algorithm is studied by (Dempster et al. 1977), (Wu 1983) for models which satisfy the following conditions

(1): $\Omega \subseteq \mathbb{R}^{|\Psi|}$ where $|\Psi|$ denotes number of scalars to estimate;
(2): $\Omega_{\Psi_0} = \left\{\Psi \in \Omega : l(\Psi; \mathbf{y}_{1:N}) \geq l(\Psi_0; \mathbf{y}_{1:N})\right\}$ is compact for any $\Psi_0 \in \Omega$ such that $l(\Psi_0; \mathbf{y}_{1:N}) > -\infty$;
(3): the log-likelihood function $l(\Psi; \mathbf{y}_{1:N})$ is continuous in $\Psi \in \Omega$ and differentiable in the interior of Ω.

The first and second condition ensures that a monotonic sequence $\left\{l^{(i)}\right\}_{i\in\mathbb{N}_0}$, which starts from $l\left(\Psi^{(0)}, \mathbf{y}_{1:N}\right)$ and has arguments in $\Omega_{\Psi^{(0)}}$, is convergent. The last point provides smoothness of the objective function of the algorithm and ensures that the loglikelihood is integrable. It implies that the steps of the EM algorithm are feasible to be obtained as integrals of an integrable function (it does not mean they are tractable). As a consequence, $\left\{l^{(i)}\right\}_{i\in\mathbb{N}_0}$ is bounded above for any starting point which belongs to the compact set Ω, increasing and integrable. Following (Wu 1983), by combining the conditions (1)-(3) with the results of Theorem 1, we obtain the weakest type of convergence for the EM algorithms what we summarize in Lemma 1

Lemma 1. *(Wu 1983) Let the loglikelihood function in Equation (2.8) and the parameter space Ω satisfy conditions (1)-(3). It implies that the EM sequence $\left\{l^{(i)}\right\}_{i\in\mathbb{N}_0}$ is bounded for any initial value $\Psi^{(0)} \in \Omega$. Then the EM sequence $\left\{l^{(i)}\right\}_{i\in\mathbb{N}_0}$ converges to some limit, denoted by $\tilde{l}$.*

The convergence implied by Lemma 1 does not guarantee that the limit $\tilde{l}$ is a stationary point of the loglikelihood function defined in Equation (2.8). The authors of (Wu 1983) and (Boyles 1983) show that if the function Q is smooth enough, then the limit $\tilde{l}$ is either a stationary point of the loglikelihood function or its local or global maximum, which we state in Theorem 2.

Theorem 2. *(Wu 1983) If $Q(\Psi, \Psi^*)$ is continuous both in Ψ and Ψ^* and the condition (1)-(3) are satisfied, then all limit points of the sequence $\left\{\Psi^{(i)}\right\}_{i\in\mathbb{N}_0}$ are the stationary points of the loglikelihood function in Equation (2.8) and the sequence $\left\{l^{(i)}\right\}_{i\in\mathbb{N}_0}$ converges to $\tilde{l}$ such that $\tilde{l} := l(\tilde{\Psi}; \mathbf{y}_{1:N})$ for a stationary point $\tilde{\Psi}$.*

The work of (Wu 1983) highlights that the family of joint distributions of $\mathbf{Y}_{1:N}, \mathbf{X}_{1:N}$ which satisfies the smoothness conditions of Theorem 2 is the exponential family. This condition is also satisfied for the models developed in Section 3.2 where the function Q is in a form of a rational polynomial of the parameters for both Ψ and Ψ^*. Given the model assumptions, the points of the discontinuity of this polynomial cannot be reached and we obtain continuity of Q on the whole surface of a considered parameter space.

We remark that the convergence in Theorem 2 is guaranteed only to a stationary point of the loglikelihood, which might be a global or a local maximum, or a saddle point. If a loglikelihood function has multiple stationary points then the convergence of the algorithm depends on the initial values of the parameters. The techniques such as simulating the algorithms many times for various starting points can give some assurance of convergence to the global maximum. In (Wu 1983), the author propose an additional condition which indicates if an EM sequence reached a local maximum of the loglikelihoods function in Equation (2.8). If the function Q which satisfies Theorem 2 and

$$\sup_{\Psi^*\in\Omega} Q(\tilde{\Psi}, \Psi^*) > Q(\tilde{\Psi}, \tilde{\Psi})$$

for $\tilde{\Psi}$ which belongs to the set of stationary points of the likelihood function which are not local maximas, then the sequence in Theorem 2 converges to a local maximum. However, the additional condition is typically of a little practical use since it difficult to verify for more complex models.

In addition to the convergence of $\left\{l(i)\right\}_{i\in\mathbb{N}_0}$ to a local maximum $\tilde{l}$, the authors of (Wu 1983) study the conditions which ensure to obtain the convergence of $\left\{\Psi^{(i)}\right\}_{i\in\mathbb{N}_0}$ to $\tilde{\Psi}$ where $\tilde{\Psi}$ is a stationary point of the loglikelihood in Equation (2.8). The author imposes stronger conditions on smoothness of the likelihood function and the function Q, such as continuous partial derivative of Q. In addition, the work investigates various cases of the shape of the likelihood function which impact on the size and density of the the set with its stationary points. The most general convergence results for the convergence of sequences generated by the EM algorithm are summarized in the following theorem

Theorem 3. *(Wu 1983) Let us denote $\mathcal{S}_a$ as a subset of all stationary points of the loglikelihood function in Equation (2.8) which belong to the interior of Ω and result in the loglikelihood function equal to $a \in \mathbb{R}$, that is*

$$l(\Psi; \mathbf{y}_{1:N}) = a \text{ for any } \Psi \in \Omega_a.$$

Suppose that $\frac{\partial Q(\Psi,\Psi^*)}{\partial \Psi^*}$ is continuous in Ψ and Ψ^*. Then the EM sequence $\left\{\Psi^{(i)}\right\}_{i\in\mathbb{N}_0}$ converges to a stationary point of the loglikelihood in Equation (2.8), denoted by $\tilde{\Psi} \in \mathcal{S}_{\tilde{l}}$, such that $\tilde{l} = l(\tilde{\Psi}; \mathbf{y}_{1:N})$ is a limit of the corresponding sequence $\left\{l^{(i)}\right\}_{i\in\mathbb{N}_0}$, if one of the following conditions holds:

(1): *$\mathcal{S}_{\tilde{I}}$ consists of a single point,*

(2): *the elements of the sequence $\left\{\Psi^{(i)}\right\}_{i\in\mathbb{N}_0}$ satisfy that*

$$||\Psi^{(i+1)} - \Psi^{(i)}|| \to 0 \text{ for } i \to \infty,$$

where $||\cdot||$ denotes the norm on Ω, and $\mathcal{S}_{\tilde{I}}$ is discrete.

2.6. The Application of Expectation-Maximisation Algorithm to Probabilistic Principal Component Analysis

The EM algorithm uses the logarithm of the the complete data likelihood, i.e. the joint distribution of $\mathbf{Y}_{1:N}, \mathbf{X}_{1:N}|\Psi$ given by

$$\begin{aligned}\pi_{\mathbf{Y}_{1:N},\mathbf{X}_{1:N}|\Psi}\left(\mathbf{y}_{1:N},\mathbf{x}_{1:N}\right) &= \prod_{t=1}^{N}\pi_{\mathbf{Y}_t,\mathbf{X}_t|\Psi}\left(\mathbf{y}_t,\mathbf{x}_t\right) = \prod_{t=1}^{N}\pi_{\mathbf{Y}_t|\mathbf{X}_t,\Psi}\left(\mathbf{y}_t\right)\,\pi_{\mathbf{X}_t|\Psi}\left(\mathbf{x}_t\right)\\ &= (2\pi)^{-N\frac{d+k}{2}}\left(\sigma^2\right)^{-N\frac{d}{2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{t=1}^{N}\left(\mathbf{y}_t-\boldsymbol{\mu}-\mathbf{x}_t\mathbf{W}^T\right)\left(\mathbf{y}_t-\boldsymbol{\mu}-\mathbf{x}_t\mathbf{W}^T\right)^T-\frac{1}{2}\sum_{t=1}^{N}\mathbf{x}_t\mathbf{x}_t^T\right\},\end{aligned} \tag{2.13}$$

where we used the assumption about time independence of the subsequent in time realisations and applied the formulation of the joint distribution of the $\mathbf{Y}_t, \mathbf{X}_t|\Psi$ from Equation (2.5). The EM algorithm maximises the expression (2.13) which is integrated with respect to the unobserved values of the random vector $\mathbf{X}_t$, denoted by the function Q. The procedure iterates over two steps, where the maximizers to the function Q of Gaussian PPCA are found in closed form. The steps are summarized as follows

1. Expectation step: Expectation of the loglikelihood function of the join distribution of $\mathbf{Y}_{1:N}, \mathbf{X}_{1:N}|\Psi$ for a fixed vector of static parameters Ψ^* with respect to the conditional distribution $\mathbf{X}_{1:N}|\mathbf{Y}_{1:N}, \Psi$

$$Q\left(\Psi,\Psi^*\right) = \mathbb{E}_{\mathbf{X}_{1:N}|\mathbf{Y}_{1:N},\Psi}\left[\log\pi_{\mathbf{Y}_{1:N},\mathbf{X}_{1:N}|\Psi^*}\left(\mathbf{y}_{1:N},\mathbf{x}_{1:N}\right)\right]. \tag{2.14}$$

2. Maximisation step: Finding $\mathbf{W}^*, \boldsymbol{\mu}^*$ and σ^{*2} that maximize $Q\left(\Psi|\Psi^*\right)$

$$\left(\mathbf{W}^*,\boldsymbol{\mu}^*,\sigma^{*2}\right) = \underset{\mathbf{W}^*\in\mathbb{R}^{d\times k},\boldsymbol{\mu}^*\in\mathbb{R}^d,\sigma^{*2}>0}{\operatorname{argmax}} Q\left(\Psi,.\Psi^*\right)$$

The Expectation step (E-step) provides with the expectation (2.14) of complete data likelihood (2.13) based on $\mathbf{Y}_{1:N}$ and assumes $\mathbf{W}$, $\boldsymbol{\mu}$ and σ^2 to be known. It uses the observed data, current estimates of parameters and the distribution of missing values conditioned on these elements. The Maximisation step (M-step) maximizes the expectation (2.14) with respect to $\mathbf{W}^*$, $\boldsymbol{\mu}^*$ and σ^{*2} as if complete data information is available. Recall that the vectors Ψ and Ψ^* are, respectively, an old and updated version of the static parameters.

2.6.1. Expectation step and its maximum

We derive the steps of the algorithm using the assumptions of the normality of $\mathbf{Y}_t$ and $\mathbf{X}_t$. As mentioned in (Dempster et al. 1977) and (Wu 1983), the convexity of regular exponential families, for which the normal distribution is a sub-family, ensures the uniqueness of the maximizers computed in the M-step. Also, the normal distribution provides us with the closed forms of the moments used in subsequent steps of the algorithm. Hence, it simplifies the computations. Theorem 4 and Theorem 5 below provide the formulations of the E-step and M-step of the EM algorithm for Gaussian PPCA respectively, assuming that the data is fully observable, i.e there is no missing elements in the the observation vector $\mathbf{Y}_t$ over time.

Theorem 4. *The E-step of the EM algorithm for Gaussian Probabilistic Principal Component Analysis given the model's assumption from Section 2.4 and N realisations of the observation vector $\mathbf{Y}_t$ denoted by $\mathbf{y}_{1:N} = \{\mathbf{y}_1, \dots, \mathbf{y}_N\}$ is obtained in the closed form as follows*

$$
\begin{aligned}
Q(\Psi, \Psi^*) = \mathbb{E}_{\mathbf{X}_{1:N}|\mathbf{Y}_{1:N},\Psi}\left[\log \pi_{\mathbf{Y}_{1:N},\mathbf{X}_{1:N}|\Psi^*}(\mathbf{y}_{1:N}, \mathbf{x}_{1:N})\right] &= -\frac{N(d+k)}{2}\log 2\pi - \frac{Nd}{2}\log \sigma^{*2} \\
- \frac{1}{2}\sum_{t=1}^{N}\Bigg\{ \frac{1}{\sigma^{*2}} Tr\left\{\mathbf{y}_t^T\mathbf{y}_t\right\} - \frac{2}{\sigma^{*2}}\mathbf{y}_t\boldsymbol{\mu}^{*T} + \frac{1}{\sigma^{*2}}\boldsymbol{\mu}^*\boldsymbol{\mu}^{*T} - \frac{2}{\sigma^{*2}} Tr\left\{\mathbf{W}^*\mathbb{E}_{\mathbf{X}_t|\mathbf{Y}_t,\Psi}\left[\mathbf{X}_t^T\right]\mathbf{y}_t\right\} \\
+ \frac{2}{\sigma^{*2}}\mathbb{E}_{\mathbf{X}_t|\mathbf{Y}_t,\Psi}\left[\mathbf{X}_t\right]\mathbf{W}^{*T}\boldsymbol{\mu}^{*T} + Tr\left\{\left(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{W}^* + \mathbb{I}_k\right)\mathbb{E}_{\mathbf{X}_t|\mathbf{Y}_t,\Psi}\left[\mathbf{X}_t^T\mathbf{X}_t\right]\right\}\Bigg\}
\end{aligned}
$$

for the corresponding moments of the conditional distribution $\mathbf{X}_t|\mathbf{Y},\Psi$ being the following

$$
\begin{aligned}
\mathbb{E}_{\mathbf{X}_t|\mathbf{Y}_t,\Psi}\left[\mathbf{X}_t\right]_{1\times k} &= (\mathbf{y}_t - \boldsymbol{\mu})\mathbf{W}\mathbf{M}^{-1}, \\
\mathbb{E}_{\mathbf{X}_t|\mathbf{Y}_t,\Psi}\left[\mathbf{X}_t^T\mathbf{X}_t\right]_{k\times k} &= \sigma^2\mathbf{M}^{-1} + \mathbf{M}^{-1}\mathbf{W}^T(\mathbf{y}_t - \boldsymbol{\mu})^T(\mathbf{y}_t - \boldsymbol{\mu})\mathbf{W}\mathbf{M}^{-1}
\end{aligned}
$$

where $\mathbf{M} = \mathbf{W}^T\mathbf{W} + \sigma^2\mathbb{I}_k$.

Proof. The proof of the Theorem 4 is given in Appendix A.2 in Subection A.2.1. □

Theorem 5. *The maximizers of the function $Q(\Psi, \Psi^*)$ defined in Theorem 4 are the solution to the following system of equations $\nabla_{\Psi^*}Q = \mathbf{0}$ and are given by*

$$
\begin{aligned}
\boldsymbol{\mu}^* &= \bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N};\Psi)\left(\mathbb{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\right) + \boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T} \\
\mathbf{W}^* &= \bar{\mathbf{C}}_{\mu,\mu^*}(\mathbf{y}_{1:N};\Psi,\Psi^*)\mathbf{W}\mathbf{M}^{-1}\left(\sigma^2\mathbf{M}^{-1} + \mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\mu}(\mathbf{y}_{1:N};\Psi)\mathbf{W}\mathbf{M}^{-1}\right)^{-1} \\
\sigma^{*2} &= \frac{1}{d}Tr\Big\{\bar{\mathbf{C}}_{\mu^*}(\mathbf{y}_{1:N};\Psi,\Psi^*) - 2\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\mu,\mu^*}(\mathbf{y}_{1:N};\Psi,\Psi^*) \\
&\quad + \mathbf{W}^*\left(\sigma^2\mathbf{M}^{-1} + \mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\mu}(\mathbf{y}_{1:N};\Psi)\mathbf{W}\mathbf{M}^{-1}\right)\mathbf{W}^{*T}\Big\}
\end{aligned}
$$

where

$$
\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N};\Psi)_{1\times d} = \frac{1}{N}\sum_{t=1}^{N}\mathbf{y}_t,
$$

$$\bar{\mathbf{S}}(\mathbf{y}_{1:N};\Psi)_{d\times d} = \frac{1}{N}\sum_{t=1}^{N}\mathbf{y}_t^T\mathbf{y}_t,$$

$$\bar{\mathbf{C}}_{\mu}(\mathbf{y}_{1:N};\Psi)_{d\times d} = \bar{\mathbf{S}}(\mathbf{y}_{1:N};\Psi) - 2\boldsymbol{\mu}^T\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N};\Psi) + \boldsymbol{\mu}^T\boldsymbol{\mu},$$

$$\bar{\mathbf{C}}_{\mu^*}(\mathbf{y}_{1:N};\Psi,\Psi^*)_{d\times d} = \bar{\mathbf{S}}(\mathbf{y}_{1:N};\Psi) - 2\boldsymbol{\mu}^{*T}\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N};\Psi) + \boldsymbol{\mu}^{*T}\boldsymbol{\mu}^*,$$

$$\bar{\mathbf{C}}_{\mu,\mu^*}(\mathbf{y}_{1:N};\Psi,\Psi^*)_{d\times d} = \bar{\mathbf{S}}(\mathbf{y}_{1:N};\Psi) - \left(\boldsymbol{\mu}+\boldsymbol{\mu}^*\right)^T\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N};\Psi) + \boldsymbol{\mu}^{*T}\boldsymbol{\mu}.$$

Proof. The proof of the Theorem 5 is given in Appendix A.2 in Subection A.2.2. □

The parameters $\mathbf{W}, \boldsymbol{\mu}, \sigma^2$ are estimated iteratively by two steps specified in Theorem 4 and Theorem 5 as illustrated in Algorithm 1. Recall that since the data does not contain any missing values, the calculation of the sample mean and variance can be performed only once during the initialisation step of the algorithm.

Algorithm 1 EM Algorithm for Gaussian PPCA with no missing observations

1: **for** $j = 1, \ldots, d$ **do**
2: Compute $\Theta_N\left([\mathbf{Y}]_{\cdot,j}\right) = (\hat{\mu}_j, \hat{\sigma}_j^2)$
3: Standardize data $[\tilde{\mathbf{Y}}]_{\cdot,j} = \frac{[\mathbf{Y}]_{\cdot,j} - \hat{\mu}_j}{\sigma_j}$
4: **end for**
5: Initialise: $i = 0$, $\mathbf{W}^{(0)} = \mathbf{W}_0$, $\sigma^{2(0)} = \sigma_0^2$,
6: Calculate: $\bar{\boldsymbol{\mu}} = \frac{1}{N}\sum_{t=1}^{N}\tilde{\mathbf{Y}}_t$, $\bar{\mathbf{S}} = \frac{1}{N}\sum_{t=1}^{N}\tilde{\mathbf{Y}}_t^T\tilde{\mathbf{Y}}_t$
7: **repeat**

Calculate: $\bar{\mathbf{C}}_{\mu^{(i)}}$, $\bar{\mathbf{C}}_{\mu^{(i+1)}}$, $\bar{\mathbf{C}}_{\mu^{(i)},\mu^{(i+1)}}$

$$\boldsymbol{\mu}^{(i+1)} = \bar{\boldsymbol{\mu}}\left(\mathbb{I}_d - \mathbf{W}^{(i)}\mathbf{M}^{(i)\,-1}\mathbf{W}^{(i)\,T}\right) + \boldsymbol{\mu}^{(i)}\mathbf{W}^{(i)}\mathbf{M}^{(i)\,T}\mathbf{W}^{(i+1)\,T}$$

$$\mathbf{W}^{(i+1)} = \bar{\mathbf{C}}_{\mu^{(i)},\mu^{(i+1)}}\mathbf{W}^{(i)}\mathbf{M}^{(i)\,-1}\left(\sigma^{2\,(i)}\mathbf{M}^{(i)\,-1} + \mathbf{M}^{(i)\,-1}\mathbf{W}^{(i)}\bar{\mathbf{C}}_{\mu^{(i)}}\mathbf{W}^{(i)}\mathbf{M}^{(i)\,-1}\right)^{-1}$$

$$\sigma^{2\,(i+1)} = \frac{1}{d}\operatorname{Tr}\left\{\bar{\mathbf{C}}_{\mu^{(i+1)}} - 2\mathbf{W}^{(i+1)}\mathbf{M}^{(i)\,-1}\mathbf{W}^{(i)\,T}\bar{\mathbf{C}}_{\mu^{(i)},\mu^{(i+1)}}\right\} + \frac{1}{d}\operatorname{Tr}\left\{\mathbf{W}^{(i+1)\,T}\mathbf{W}^{(i+1)}\left(\sigma^{2\,(i)}\mathbf{M}^{(i)\,-1} + \mathbf{M}^{(i)\,-1}\mathbf{W}^{(i)\,T}\bar{\mathbf{C}}_{\mu^{(i)}}\mathbf{W}^{(i)}\mathbf{M}^{(i)\,-}\right.\right.$$

8: $i = i + 1$
9: **until** a convergence criterion is satisfied

2.6.2. The maximum likelihood estimation - the convergence of EM algorithm

In the work of (Dempster et al. 1977), the author proves that the EM algorithm for the normal distribution always converges to a local maximum. Recall that the solution provided by a EM algorithm converges to the solution obtained by maximizing the likelihood of marginal distribution $\mathbf{Y}_t|\Psi$ when $\mathbf{Y}_t$ is assumed to be centred, given by

$$\log \pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N}) = -\frac{N}{2}\left(d\log(2\pi) + \ln|\mathbf{C}| + \operatorname{Tr}\left\{\mathbf{C}^{-1}\bar{\mathbf{S}}(\mathbf{y}_{1:N};\Psi)\right\}\right) \tag{2.15}$$

where $\bar{\mathbf{S}}(\mathbf{y}_{1:N};\Psi)$ is defined as in Theorem 4. The maximum log-likelihood estimator (MLE) of $\mathbf{W}$ given by the likelihood (2.15) is a solution to the following fix point equation

$$\frac{\partial \log \pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N})}{\partial \mathbf{W}} = N\left(\mathbf{C}^{-1}\bar{\mathbf{S}}(\mathbf{y}_{1:N};\Psi)\mathbf{C}^{-1}\mathbf{W} - \mathbf{C}^{-1}\mathbf{W}\right) = \mathbf{0} \Longrightarrow \mathbf{W} = \bar{\mathbf{S}}(\mathbf{y}_{1:N};\Psi)\mathbf{C}^{-1}\mathbf{W} \tag{2.16}$$

For the convenience of the notation, let us denote $\mathbf{S} := \bar{\mathbf{S}}(\mathbf{y}_{1:N}; \Psi)$. Following (Tipping and Bishop 1999b), we may distinguish three possible cases to the above solution. The solution to $\mathbf{W} = \mathbf{0}$ is treated as as a minimum of the log-likelihood.

When $\mathbf{S} = \mathbf{C}$, it implies that $d - k$ smallest eigenvalues of $\mathbf{S}$ are equal to σ^2 and the problem is indefinable since $\mathbf{W}\mathbf{W}^T = \mathbf{S} - \sigma^2 \mathbb{I}_d$. Given the eigendecomposition of $\mathbf{S}$, that is

$$\mathbf{S} = \mathbf{U}_{d\times d}\mathbf{\Lambda}_{d\times d}\mathbf{U}^T_{d\times d} \tag{2.17}$$

for orthonormal matrix $\mathbf{U}$ such that $\mathbf{U}^T\mathbf{U} = \mathbb{I}_d$ and diagonal matrix $\mathbf{\Lambda}$ with non-negative entries, the matrix $\mathbf{W}$ is equal to $\mathbf{W} = \mathbf{U}\left(\mathbf{\Lambda} - \sigma^2\mathbb{I}_d\right)^{\frac{1}{2}}\mathbf{R}^T$ where $\mathbf{R}_{k\times d}$ is an arbitrary rotation matrix. The case is proven in Subsubsection 2.3.2.

The last case examines $\mathbf{W} \neq \mathbf{0}$ and $\mathbf{S} \neq \mathbf{C}$. In order to compute the solution to (2.16) we use the singular value decomposition of $\mathbf{W}$, that is

$$\mathbf{W} = \mathbf{V}_{d\times k}\mathbf{L}_{k\times k}\mathbf{R}^T_{k\times k}$$

where $\mathbf{V}$ and $\mathbf{R}$ are real valued orthogonal matrix by columns, and $\mathbf{L}$ is non-negative diagonal matrix. Using $\mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbb{I}_d$ we apply the above facts to the problem defined by the fix point equation

$$\begin{aligned}
&\mathbf{W}_{d\times k} = \mathbf{S}\mathbf{C}^{-1}\mathbf{W},\\
&\mathbf{V}\mathbf{L}\mathbf{R}^T = \mathbf{S}\left(\mathbf{V}\mathbf{L}\mathbf{R}^T\mathbf{R}\mathbf{L}\mathbf{V}^T + \sigma^2\mathbb{I}_d\right)^{-1}\mathbf{V}\mathbf{L}\mathbf{R}^T,\\
&\mathbf{V}\mathbf{L}\mathbf{R}^T = \mathbf{S}\left(\mathbf{V}\mathbf{L}^2\mathbf{V}^T + \sigma^2\mathbb{I}_d\right)^{-1}\mathbf{V}\mathbf{L}\mathbf{R}^T,\\
&\mathbf{V}\mathbf{L}^2\mathbf{V}^T + \sigma^2\mathbb{I}_d = \mathbf{S}\\
&\mathbf{U}\mathbf{L}^2\mathbf{L} + \sigma^2\mathbb{I}_d\mathbf{V}\mathbf{L} = \mathbf{S}\mathbf{V}\mathbf{L},\\
&\mathbf{V}\left(\mathbf{L}^2 + \sigma^2\mathbb{I}_k\right)\mathbf{L} = \mathbf{S}\mathbf{V}\mathbf{L}
\end{aligned}$$

Notice that

$$\mathbf{S}\mathbf{v}_j = (\sigma^2 + l_j^2)\mathbf{v}_j,$$

where $\mathbf{v}_j = [\mathbf{V}]_{,j}$ and $l_j = [\mathbf{L}]_{j,j}$. Hence, the vectors $\mathbf{v}_j$ are eigenvectors of the estimated covariance matrix $\mathbf{S}$. Using the eigendecomposition of $\mathbf{S}$ given by (2.17), we see that $\mathbf{v}_j$ corresponds to the eigenvectors of $\mathbf{S}$, $\mathbf{u}_j$ with eigenvalues $\lambda_j = l_j^2 + \sigma^2$. Since $\mathbf{L}$ has a different dimension than $\mathbf{\Lambda}$, we express it as

$$\mathbf{L} = (\mathbf{K} - \sigma^2\mathbb{I}_k)^{\frac{1}{2}},$$

where $[\mathbf{K}]_{j,j} = \lambda_j$ is the jth eigenvalue of $\mathbf{S}$ and corresponds to the jth eigenvector of $\mathbf{S}$, given by $\mathbf{u}_j$. In the case in which $l_j = 0$, the eigenvector $\mathbf{v}_j$ is arbitrary and $[\mathbf{K}]_{j,j} = \sigma^2$. The scalar σ^2 is estimated as the average of the eigenvalues in which the index $i > k$ taken from eigen-decomposition of $\mathbf{S}$. $i > k$ eigenvalues of $\mathbf{S}$.

The remaining question is if the EM algorithm converges to the global maximum. The authors of (Tipping and Bishop 1999b) shows that if the stationary points of the likelihood function with respect to $\mathbf{W}$ which are spanned by minor eigenvectors (the eigenvectors with corresponding negligible eigenvalues), are stable, then the convergence is not guaranteed.

However, we observe that any eigenvector which does not correspond to the k highest eigenvalues of $\mathbf{S}$ belongs to the set of saddle points of the logliklihood and does not provide stable solutions. (Tipping and Bishop 1999b) shows that if singular values of the estimates of $\mathbf{W}$ correspond to the smallest $d - k$ eigenvalues of $\mathbf{S}$, the Hessian of the loglikehood function is neither positive nor definite definite. When the singular-values of the estimated matrix $\mathbf{W}$ are smaller that the singular values of the eigenvectors of the matrix $\mathbf{S}$, the sign of the likelihood gradient in the neighbourhood of the remaining eigenvectors of $\mathbf{S}$, is not negative. Therefore, the authors concludes that the maximum of the likelihood function is achieved only when the singular values and singular vectors of $\mathbf{W}$ equals k highest eigenvalues of $\mathbf{S}$ and corresponding eigenvectors.

3. Robust Probabilistic Principal Component Analysis

In the following chapter, we discuss several novel extensions to the Probabilistic Principal Component Analysis frameworks which seek to extract most relevant representations of the data set in efficient and statistically robust manner. We still assume all components of the observation process to be observable, i.e. the sample, which contains N realisations of the process $\mathbf{Y}_t$, denoted by $\mathbf{y}_{1:N} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$, has no unobserved or missing elements in each vector $\mathbf{y}_t$ for $t = 1, \ldots, N$. This assumption is relaxed in Chapter 4.

We demonstrate in Section 3.1 how to utilize the widely used concept of robust estimation of sample moments within Gaussian Probabilistic Principal Component Analysis. The later part of the chapter discuss how to ensure the robustness of the PPCA methodology by various model based assumptions.

We start in Section 3.2 by introducing the distribution-based robustification by replacing the Gaussian distribution assumption with the t-Student in order to account for heavy tail assumptions. This extension results in the methodology being statistically robust to the distant observations in a sample sets outliers. We investigate two t-Student frameworks, one in which the observations are assumed to be independent over time, which leads to different probability density functions over time, and the second in which the observation vectors are assumed to be identically distributed and conditionally independent over time. The first concept has been partially introduced by (Lange et al. 1989) whereas the latter is our novel contribution to the existing literature on the Probabilistic Principal Component Analysis. The assumed independence over time of the random vectors limits the latent vectors to have time-specific tail dependence and therefore, the time-specific probability functions. On the other hand, when we restrict the random vectors to be identically distributed over time, the realisations of the variables are only conditionally independent.

Next, we define in Section 3.4 the Generalized Skew t-Student PPCA model. The robust alternatives of deterministic Principal Component Analysis (PCA) or Probabilistic Principal Component Analysis (PPCA) are largely based on an assumption that the marginal distributions of the individual components of the observation vector have the same profile of heavy tails. This reasoning might be criticized for having a limited ability to capture various tail dependence patterns in a multivariate data analysis. Therefore we introduce a probabilistic model which allows to capture different profiles of the tail dependence in different sub-spaces of the data random vectors. In developing such a framework, we also allow for asymmetric distributions to be considered. We distinguish two cases of the distribution, when the random vectors $\mathbf{X}_t$ and $\boldsymbol{\epsilon}_t$ are either independently and non-identically distributed or identically and conditionally independently distributed over time.

The extended frameworks of the derived methodologies so far discussed are then further extended to provide a novel framework for feature extraction that allows for the consistent statistical incorporation of missingness and this is broadly discussed in Chapter 4.

3.1. Robust Gaussian Probabilistic Principal Component Analysis: Utilising the Concept of Robust Estimation

In the presence of distant data points, so called outliers, the sample estimators of moments, which are a crucial element of standard Principal Component Analysis and the EM algorithm for Gaussian Probabilistic Principal Component Analysis as shown in Theorem 5, may loss the information about the true distribution of the observation vector. Again, we seek a computationally efficient and parsimonious solution to this estimation, we will therefore again remain in the class of solutions characterized by the EM algorithm. However, we will need to develop a modification of this approach to accommodate robustness to outliers to reduce the degradation in the feature extraction that would otherwise arise if the standard EM approach from Section 2.4 was applied. In the following section we introduce the concepts of the statistical robustness in estimation that and its most popular in practice techniques. We start by observing that one can approach the robustification of such feature extraction methods in two ways:

- by robust standardisation of data in the initialisation of EM algorithm,
- and by replacing the sample mean and covariance matrix in Theorem 5, $\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N;\Psi})$ and $\bar{\mathbf{S}}(\mathbf{y}_{1:N;\Psi})$, by their robust equivalents.

3.1.1. Robust Estimators for the Mean and Covariance Matrix

We first introduce what exactly is meant by statistically robust feature extraction. This requires a short set of formal definitions for key statistical concepts used to define and study classes of robust estimators. We refer the reader to the detail discussion in (Huber and Ronchetti 2009) on the broader aspect of the robust estimation methodology.

In this section we assume that the random vector $\mathbf{Y}_t$ is independently and identically distributed over time with empirical distribution denoted by F_N and obtained from sample realizations denoted by $\mathbf{y}_{1:N}$. The true population distribution and density functions of $\mathbf{Y}_t$ are denoted by capital letter F and f.

Recall, that the subscript index, that denotes sequence of observation order or time, the notation for the random vectors and observations, could have been dropped in this subsection, since the discussed random vectors are assumed to be independently and identically distributed. However, for the clarity of the notation in this thesis and the later discussion, we keep the lower index. Hence, the sample set $\mathbf{y}_{1:N}$ can be viewed as N realisations of a one random vector over different time points.

We define an estimator $\hat{\Theta}$ of a parameter $\Theta \in \Omega$ as a functional on the domain of distribution functions. We can also analogously think of the definition of an estimator as a function of the N realisations of the d-dimensional random vector $\mathbf{Y}_t$, sample $\mathbf{y}_{1:N} = \left\{\mathbf{y}_1, \ldots, \mathbf{y}_N\right\}$, which construct an population distribution function F_N . Definition 1 formalizes the notion of an estimator.

Definition 1. *The estimator of the parameter Θ is a mapping $\hat{\Theta}$ from the data space to the parameter space. The data space is represented by a distribution function. Therefore $\hat{\Theta}$ is a statistical functional on a set of distribution functions.*

3.1.1.1. Concept of robustness.

We define a robustness according to two measures: a measure of local robustness and secondly a measure of global robustness. Both, the local and global measures can be explained using a ϵ-contamination set of a distribution functions, defined in Definition 2.

Definition 2. *Denote by $\mathcal{F}_\epsilon$ an ϵ-contamination neighbourhood of a probability distribution F, defined as*

$$\mathcal{F}_\epsilon = \{G : G = (1-\epsilon)F + \epsilon H, \text{ for a distribution } H\}$$

given fraction of the contamination $0 \leq \epsilon \leq 1$.

One can then measure the local robustness of an estimator $\hat{\Theta}$ by influence function according to the specification in Definition 3.

Definition 3 (Influence function)**.** *The influence function of an estimator $\hat{\Theta}$ on the domain of the distribution function F defined according to a local point y_0 is given by*

$$IF(\mathbf{y}_0, \hat{\Theta}, F) = \lim_{\epsilon \to 0} \frac{\hat{\Theta}\left((1-\epsilon)F + \epsilon\delta_{\mathbf{y}_0}\right) - \hat{\Theta}(F)}{\epsilon}$$

where $\delta_{\mathbf{y}_0}$ is a probability measure which puts mass 1 at the point $\mathbf{y}_0$.

Let us remark that a influence function is a derivative of the functional $\hat{\Theta}$ at F in the direction of an distribution function with all probability mass in the point $\mathbf{y}_0$ The influence function is a crucial tool used in robust statistics to study the asymptotic variance and efficiency of an estimator as

$$\sqrt{N}\left(\hat{\Theta}(F_N) - \hat{\Theta}(F)\right) \approx \sqrt{N}\int_{\mathbb{R}^d} IF\left(\mathbf{Y}_t, \hat{\Theta}, F\right) dF_N(\mathbf{Y}_t)$$

where we used Tylor expansion of the empirical distribution function F_N around true population distribution function F.

The influence function provides us with the knowledge relating to how the ϵ contamination on a point $\mathbf{y}_0$ changes the information about the true distribution of the random variable $\mathbf{Y}_t$ as capture by the summary contained in the estimator functional is given by $\hat{\Theta}$. Thus, it is perceived as a local measure of robustness. To measure global robustness, one can examine a breakdown point ϵ^* of estimator $\hat{\Theta}$ at the true population distribution F, given in Definition 4.

Definition 4 (Breakdown point)**.** *The finite- sample breakdown point ϵ^* of an estimator Θ_N at the true population distribution F is defined as*

$$\epsilon^* := \sup\{\epsilon \leq 1 : \sup_{G \in \mathcal{F}_\epsilon} |\hat{\Theta}(G)| < \infty\}$$

Intuitively, it is understood as the maximal contamination which does not cause the estimator to loose valid information about the true distribution F. We may define a finite sample definitions of the breakdown point as

Definition 5. *The breakdown point ϵ_N^* of estimator $\hat{\Theta}$ at the empirical population distribution F_N is defined as*

$$\epsilon_N^*(\hat{\Theta}, \mathbf{y}_{1:N}) := \max\{n_1 \in \mathbb{Z}, \max_{i_1,\ldots,i_{n_1}} \sup_{\mathbf{z}_1,\ldots,\mathbf{z}_{n_1}} |\hat{\Theta}(\tilde{F}_N)| < \infty\}$$

where some subset of sample points $\mathbf{y}_{i_1}, \ldots, \mathbf{y}_{n_1}$ were replaced with arbitrary points $\mathbf{z}_1, \ldots, \mathbf{z}_{n_1}$ from a contamination distribution H and the new set of samples constructed the mepirical distribution function $\tilde{F}_N$).

After introducing the above formal definitions of robustness related to estimation, we next briefly overview the most popular estimators of a covariance matrix. We discuss their robust characteristics according to introduced measures. Let us denote $\boldsymbol{\mu}$, $\mathbf{C}$ as a mean and covariance matrix of $\mathbf{Y}_t$. In the reminded of this section we assume that $\boldsymbol{\mu}$ is a real-valued d-dimensional vector and $\mathbf{C}$ is a $d \times d$ symmetric matrix which is positive-definite.

3.1.1.2. M-estimators

The study of (Maronna 1976) and (Huber and Ronchetti 2009) introduces one of the first classes of robust estimators, called M-estimators, which are a generalized version of Maximum Likelihood Estimation (MLE) estimator where

$$\hat{\Theta}^{MLE}(F_N) = \operatorname*{argmax}_{\Theta\in\Omega} \prod_{t=1}^{N} f(\mathbf{y}_t) \Leftrightarrow \hat{\Theta}_N^{MLE} = \operatorname*{argmin}_{\Theta\in\Omega} \sum_{t=1}^{N} -\log f(\mathbf{y}_t). \tag{3.1}$$

The method of maximum likelihood which is the most popular technique in deriving estimators in a frequentist setting. The MLE estimator reflects a point estimate of Θ for which the realisations $\mathbf{y}_{1:N}$ are the most likely. We refer the reader to the broad literature on the estimation techniques such as (Casella and Berger 2001) or (Lehmann and Casella 1998).

The M-estimators are constructed by replacing the density function in Equation (3.1), f, with a proposal function $\rho : \mathbb{R}_+ \to \mathbb{R}$. The objective of ρ is improve the robustness of the estimation by downweighting contributions of extremely distant data points in calculating an estimator. The M-estimators are defined in Definition 6.

Definition 6 (M-estimators)**.** *The M-estimator $\hat{\Theta}$ of a parameter Θ is defined as a solution to the following problem*

$$\hat{\Theta}(F_N) = \operatorname*{argmin}_{\Theta\in\Omega} \sum_{t=1}^{N} \rho(D_t), \tag{3.2}$$

for a function $\rho : \mathbb{R}_+ \to \mathbb{R}$ where $D_t^2 = (\mathbf{y}_t - \boldsymbol{\mu})\,\mathbf{C}^{-1}\,(\mathbf{y}_t - \boldsymbol{\mu})^T$ is a Mahalanobias distance of the sample point $\mathbf{y}_t$.

Remark 1. *If ρ is a continuous function, denoting its derivative $\rho' = \psi$, the estimator $\hat{\Theta}_N$ is specified by finding a solution to the following system of equations with respect to $\boldsymbol{\mu}$ and $\mathbf{C}$*

$$\begin{aligned}\sum_{t=1}^{N} \frac{\psi(D_t)}{D_t} \mathbf{C}^{-\frac{1}{2}} (\mathbf{y}_t - \boldsymbol{\mu}) &= \mathbf{0}, \\ \sum_{t=1}^{N} \frac{\psi(D_t)}{t_n} \mathbf{C}^{-\frac{1}{2}} (\mathbf{y}_t - \boldsymbol{\mu})^T (\mathbf{y}_t - \boldsymbol{\mu}) \mathbf{C}^{-\frac{1}{2}} &= \mathbf{0}\end{aligned}.$$

The above system of equations is a robust analogue of the typical normal type equations one would solve in MLE estimations.

Remark 2. *If we additionally assume that $\mathbf{Y}_t$ is a random vector from elliptical family with density of the form*

$$f(\mathbf{y}_t; \boldsymbol{\mu}, \mathbf{C}) = c^{-1} |\mathbf{C}|^{-\frac{1}{2}} g\left((\mathbf{y}_t - \boldsymbol{\mu}) \mathbf{C}^{-1} (\mathbf{y}_t - \boldsymbol{\mu})^T \right). \tag{3.3}$$

with $g : \mathbb{R}_+ \to \mathbb{R}_+$ being a density generator of random variable $\mathbf{Y}$, the solution to the problem (3.2) is equivalent to

$$\begin{aligned}\sum_{t=1}^{N} \frac{\psi(D_t)}{D_t} \mathbf{C}^{-\frac{1}{2}} (\mathbf{y}_t - \boldsymbol{\mu}) &= \mathbf{0}, \\ \sum_{t=1}^{N} \frac{\psi(D_t)}{D_t} \mathbf{C}^{-\frac{1}{2}} (\mathbf{y}_t - \boldsymbol{\mu})^T (\mathbf{y}_t - \boldsymbol{\mu}) \mathbf{C}^{-\frac{1}{2}} &= \mathbb{I}_d.\end{aligned} \tag{3.4}$$

More generalized notation is used in (Maronna 1976) and (Huber and Ronchetti 2009) who introduce functions $u_1, u_2 : \mathbb{R}_+ \to \mathbb{R}$ and rewrite Equation (3.4) as is shown in Definition 7.

Definition 7. *The M-estimator $\hat{\Theta}$ of a parameter Θ of a location and scatter of random variable $\mathbf{Y}_t$ from elliptical family, is defined as a solution to*

$$\begin{aligned}\sum_{t=1}^{N} u_1(D_t) (\mathbf{y}_t - \boldsymbol{\mu}) &= \mathbf{0}, \\ \sum_{t=1}^{N} u_2\left(D_t^2\right) \mathbf{C}^{-\frac{1}{2}} (\mathbf{y}_t - \boldsymbol{\mu})^T (\mathbf{y}_t - \boldsymbol{\mu}) \mathbf{C}^{-\frac{1}{2}} &= \mathbb{I}_d,\end{aligned} \tag{3.5}$$

for functions $u_1, u_2 : \mathbb{R}_+ \to \mathbb{R}$.

The authors provide conditions for u_1 and u_2 to ensure existence and uniqueness of (3.5) and its quadratic convergence to a normal asymptotic distribution. (Maronna 1976) proves the breakdown point of (3.5) to be very sensitive to dimensionality of the data and show in fact that its breakdown points is equal to $\epsilon^* = \frac{1}{d+1}$.

Recall that if we are motivated by the standardisation of the sample data, in such cases the goal is then reduced to computing the variance of element of the random vector, rather than the joint covariance matrix. In such a case, due to the low dimensional requirements of each individual estimator, which can be treated independently, the M-esitmator family is then particularly useful.

What is more, the robustness of M-estimators does not depend on the sample size what is an appealing characteristic of an estimator when only limited number of observation are available.

Remark 3. *As an example of function u_1 and u_2, (Huber 1964) gives the following*

$$u_1(s) = \frac{\psi_H(s,k)}{s},\ u_2(s) = \frac{\psi_H(s,k^2)}{s},\ \rho_H(s) = \begin{cases} \frac{1}{2}s^2, & |s| < k, \\ \frac{1}{2}|s| - \frac{1}{2}k^2, & |s| \geq k, \end{cases},$$

for $\psi_H(s,k) = \rho'_H(s,k)$ and the tuning constant k.

The M-estimator with the function ρ_H, known as the Huber M-estimator, is proven in (Huber 1964) to result in minimal maximal asymptotic variance of all affine invariant estimators under the assumption that $\mathbf{Y}_t$ is Gaussian.

The other practical example of M-estimator has been introduced by (Tyler 1987a). The author investigates the estimation only of covariance matrix and proposes a function $u_2(s) = \frac{d}{s}$ where d is a dimensionality of the observation process. The obtained estimator, know as the Tyler M-estimator, is investigated in details by (Frahm and Jaekel 2010).

Definition 8. *Tyler's estimator of a covariance matrix $\mathbf{C}$ is obtained as a sollution to the following equation problem*

$$\frac{d}{N}\sum_{t=1}^{N}\frac{(\mathbf{y}_t - \boldsymbol{\mu})^T(\mathbf{y}_t - \boldsymbol{\mu})}{(\mathbf{y}_t - \boldsymbol{\mu})\mathbf{C}^{-1}(\mathbf{y}_t - \boldsymbol{\mu})^T} = \mathbf{C}. \tag{3.6}$$

Please refer to (Tyler 1987a) or discussion and references in (Frahm and Jaekel 2010) for the remarks on the statistical properties of the Tyler M-estimator. Recall that the function u_2 is not bounded and hence does not satisfy the conditions described in (Maronna 1976) or (Huber and Ronchetti 2009) which ensure the uniqueness of an M-estimator. (Tyler 1987a) shows that the existence uniqueness of the estimator specified by Definition 8 is achieved up to the scaling parameter for for any constitutions distribution of the data. Also, given this assumption on the distribution function, the author shows that the estimator is asymptotically normally and strongly consistent.

The asymptotic variance of Tyler's M-estimator, which was analysed i.e. by the author of (Tyler 1987b), is proven to have the lowest maximum bias among all normally distributed estimators for symmetric elliptical family random variables. Another advantage of a Tyler's estimator is the fact that its maximal breakdown point of a point contamination is equal to $\epsilon^* = \frac{1}{d}$ as proven in (Tyler 1987b). The latter study shows that for any other contamination, the maximal breakdown point is between

$$\frac{1}{d+1} \leq \epsilon^* \leq \frac{1}{d},$$

where d is the dimensionality of $\mathbf{Y}_t$. Furthermore, it can be shown that the Tyler's estimator correspond to the MLE estimator of angular Gaussian distribution. To prove this property we use the following Lemma.

Lemma 2. *If the random vector $\mathbf{Y}_t$ follows a generalized elliptical distribution it can be expressed according to the following equivalence in distribution representation*

$$\mathbf{Y}_t =_d \boldsymbol{\mu} + R_t \mathbf{C}^{\frac{1}{2}} \mathbf{U}_t,$$

for $\mathbf{U}_t$ comprised of d-dimensional independent and identically distributed uniform random variables on the unit sphere and R_t being a univariate random variable.

Proof. Please refer to (Frahm 2004). □

Given the representation of the generalized elliptical distribution in Lemma 2, the univariate random variable R_t is an component of the model which generates the tails the distribution of random vector $\mathbf{Y}$.

Remark 4. *If R_t is absolutely continuous and independent from $\mathbf{U}_t$, the generator of the probability density function of centred random vector $\mathbf{Y}_t$, denoted by $\tilde{\mathbf{Y}}_t$, is given by*

$$f(\tilde{\mathbf{y}}_t) := \frac{\boldsymbol{\Gamma}\left(\frac{r}{2}\right)}{2\pi^{\frac{r}{2}}} \tilde{\mathbf{y}}_t^{-\frac{r-1}{2}} f_R(\sqrt{\tilde{\mathbf{y}}_t}),$$

for $r = rank(\mathbf{C})$ and f_R being probability density function of the random variable R_t. Then $\mathbf{Y}_t$ is said to be symmetrically distributed.

Remark 5. *If the variable R_t is negative and has is mutually dependent with uniformly distributed $\mathbf{U}_t$, the distribution of $\mathbf{Y}_t$ is called the generalised elliptical distribution. The generalised elliptical family allows to model asymmetric and tail dependent distributions of $\mathbf{Y}_t$.*

Theorem 6 (Distribution - free Tyler estimator)**.** *Let $||\cdot||_F$ denotes the Frobenius norm. The estimator introduced by Definition 8 is distribution-free, i.e it does not depend on values of the random variable R_t.*

Proof. Following Lemma 2 we observe that

$$\mathbf{Z}_t := \frac{\mathbf{Y}_t - \boldsymbol{\mu}}{||\mathbf{Y}_t - \boldsymbol{\mu}||_F} = \frac{R_t \mathbf{U}_t \mathbf{C}^{\frac{1}{2}}}{||R_t \mathbf{U}_t \mathbf{C}^{\frac{1}{2}}||_F} \overset{d}{=} sign(R_t) \frac{\mathbf{U}_t \mathbf{C}^{\frac{1}{2}}}{||\mathbf{U}_t \mathbf{C}^{\frac{1}{2}}||_F}.$$

Therefore, the distribution of $\mathbf{Z}_t$ is robust against extreme realisations of R_t as it does not depend on its values. Rewriting (3.6) using N realisations of $\mathbf{Z}_t$, denoted by $\mathbf{z}_1, \dots \mathbf{z}_N$, we obtain

$$\mathbf{C} = \frac{d}{N} \sum_{t=1}^{N} \frac{\mathbf{z}_t^T \mathbf{z}_t}{\mathbf{z}_t \mathbf{C}^{-1} \mathbf{z}_t^T} = \frac{d}{N} \sum_{t=1}^{N} \frac{\mathbf{C}^{-\frac{1}{2}} \mathbf{u}_t^T \mathbf{u}_t \mathbf{C}^{-\frac{1}{2}}}{\mathbf{u}_t \mathbf{u}_t^T}. \tag{3.7}$$

□

Tyler's estimator is invariant under any change of distribution of R_t and, consequently, it is distribution-free.

Theorem 7. *The estimator introduced by Definition 8 is the MLE estimator of angular Gaussian Distribution.*

Proof. The probability density function of $\mathbf{V}_t := \frac{\mathbf{Y}_t - \boldsymbol{\mu}}{||\mathbf{Y}_t - \boldsymbol{\mu}||_F}$ is given by

$$f_v(\mathbf{v}_t) = \frac{\Gamma\left(\frac{d}{2}\right)}{2\pi^{\frac{d}{2}}} \frac{|\mathbf{C}|^{-\frac{1}{2}}}{||\mathbf{v}_t \mathbf{C}^{-\frac{1}{2}}||_F^d}.$$

Having N realisations of the random vector $\mathbf{V}_t$, given by $\mathbf{v}_1, \ldots, \mathbf{v}_N$, the log-likelihood function is the following

$$\begin{aligned} \log \pi_{\mathbf{V}_{1:N}|\mathbf{C}}(\mathbf{v}_{1:N}) &= \sum_{t=1}^{N} \log f_v(\mathbf{v}_t) = \text{const} + \frac{N}{2} + \log|\mathbf{C}^{-1}| - \frac{d}{2}\sum_{t=1}^{N} \log\left(\mathbf{v}_t \mathbf{C}^{-1} \mathbf{v}_t^T\right) \\ &= \text{const} + \frac{N}{2} + \log|\mathbf{C}^{-1}| - \frac{d}{2}\sum_{t=1}^{N} \log\left(\mathbf{z}_t \mathbf{C}^{-1} \mathbf{z}_t^T\right) \end{aligned} \tag{3.8}$$

Since the set of all symmetric positive-definite $d \times d$ matrices, denoted by $\mathcal{P}^d$, is an open set, the maximizer of $\log \pi_{\mathbf{V}_{1:N}|\mathbf{C}}(\mathbf{v}_{1:N})$ with respect to $\mathbf{C}$ is the stationary point. By setting

$$\frac{\partial \log \pi_{\mathbf{V}_{1:N}|\mathbf{C}}(\mathbf{v}_{1:N})}{\partial \mathbf{C}^{-1}} = \mathbf{0} \iff N\mathbf{C} - d\sum_{t=1}^{N} \frac{(\mathbf{y}_t - \boldsymbol{\mu})^T (\mathbf{y}_t - \boldsymbol{\mu})}{(\mathbf{y}_t - \boldsymbol{\mu}) \mathbf{C}^{-1} (\mathbf{y}_t - \boldsymbol{\mu})^T} = \mathbf{0}$$

which is precisely equal to (3.6). □

The solution to the estimation problem of Tyler's estimator is in a form $x = f(x)$ what allows one to apply a numerical iterative search solution based on a fix-point iteration scheme in which the $i+1$-th step would be performed as follows

$$\mathbf{C}^{(i+1)} = \frac{d}{N}\sum_{t=1}^{N} \frac{(\mathbf{y}_t - \boldsymbol{\mu})^T (\mathbf{y}_t - \boldsymbol{\mu})}{(\mathbf{y}_t - \boldsymbol{\mu}) \mathbf{C}^{(i)(-1)} (\mathbf{y}_t - \boldsymbol{\mu})^T}. \tag{3.9}$$

Lemma 3. *The fixed-point iterations according to steps performed by algorithm (3.9) converges to $\alpha\mathbf{C}$ for a scalar $\alpha > 0$.*

Proof. Since $f(\mathbf{C}) = \frac{d}{N}\sum_{t=1}^{N} \frac{(\mathbf{y}_t - \boldsymbol{\mu})^T (\mathbf{y}_t - \boldsymbol{\mu})}{(\mathbf{y}_t - \boldsymbol{\mu}) \mathbf{C}^{-1} (\mathbf{y}_t - \boldsymbol{\mu})^T}$ being continuous on $\mathcal{P}^d$, the algorithm converges to $\mathbf{C}$ for $\alpha > 0$. □

Following (Tyler 1987a), let us define the function $M : \mathcal{P}^d \to \mathcal{P}^d$ such that

$$M(\boldsymbol{\Gamma}) = \frac{d}{N}\sum_{t=1}^{N} \frac{\boldsymbol{\Gamma}^{\frac{1}{2}} (\mathbf{y}_t - \boldsymbol{\mu})^T (\mathbf{y}_t - \boldsymbol{\mu}) \boldsymbol{\Gamma}^{\frac{1}{2}}}{(\mathbf{y}_t - \boldsymbol{\mu}) \boldsymbol{\Gamma} (\mathbf{y}_t - \boldsymbol{\mu})^T}$$

The fixed point iteration step (3.9) can be rewritten as

$$\mathbf{C}^{(i+1)} = \mathbf{C}^{(i)\frac{1}{2}} M\left(\mathbf{C}^{(i)(-1)}\right) \mathbf{C}^{(i)\frac{1}{2}}$$

Denoting $\boldsymbol{\Gamma}^{(i)} = \mathbf{C}^{(i)(-1)}$ and $\mathbf{M}^{(k)} = M\left(\boldsymbol{\Gamma}^{(i)}\right)$ we get

$$\boldsymbol{\Gamma}^{(i+1)} = \boldsymbol{\Gamma}^{(i)\frac{1}{2}} \mathbf{M}^{(i)(-1)} \boldsymbol{\Gamma}^{(i)\frac{1}{2}}$$

Since the Tyler's estimator of a covariance matrix is unique up to the scaling parameter as noted in Lemma 3, the author of (Tyler 1987b) suggests to restrict the search space to the positive definite symmetric matrices with trace equal to 1. The author argues that this constrain would provide

the unique to the solution to Equation (3.9). The problem in Equation (3.9) is then reformulated as follows

$$\mathbf{\Gamma}^{(i+1)} = \frac{\mathbf{\Gamma}^{(i)\frac{1}{2}}\mathbf{M}^{(i)(-1)}\mathbf{\Gamma}^{(i)\frac{1}{2}}}{\text{Tr}\left\{\mathbf{\Gamma}^{(i)\frac{1}{2}}\mathbf{M}^{(i)(-1)}\mathbf{\Gamma}^{(i)\frac{1}{2}}\right\}} = \frac{\mathbf{\Gamma}^{(i)\frac{1}{2}}\mathbf{M}^{(i)(-1)}\mathbf{\Gamma}^{(i)\frac{1}{2}}}{\text{Tr}\left\{\mathbf{\Gamma}^{(i)}\mathbf{M}^{(i)(-1)}\right\}}.$$

what ensures that $\text{Tr}\left\{\mathbf{\Gamma}^{(i+1)}\right\} = 1$. In his study, Tyler proves that the convergence of the sequence $\mathbf{\Gamma}^{(i+1)}$ to a non singular matrix $\mathbf{\Gamma}$ implies $M(\mathbf{\Gamma}) = \mathbb{I}_d$

Theorem 8 ((Tyler 1987a)). *If the following conditions hold:*

1. *the sample $\mathbf{y}_1, \ldots, \mathbf{y}_N$ does not contains values equal to $\boldsymbol{\mu}$;*
2. *the empirical distribution measure F_N of the sample satisfies $F_N(\mathbf{A}) < \frac{1}{d}$, for $\mathbf{A}$ being a any proper subspace of $\mathbb{R}^d$;*
3. *for some the mth smallest eigenvalue of $\mathbf{\Gamma}^{(i)}$, it is holds that $\lambda_{m,d} > \frac{d}{rank(\mathbf{A})} F_N(\mathbf{A})$ where $\mathbf{A}$ is any proper subspace of $\mathbb{R}^d$;*

then $\mathbf{\Gamma}^{(i)} \to \mathbf{\Gamma}$ and $M(\mathbf{\Gamma}) = \mathbb{I}_d$.

Remark 6. *As just illustrated with the restriction of (Tyler 1987b), it is common to deal with the lack of uniqueness of the solution obtained from iterations according to Equation (3.9) by imposing additional constraints. The other examples of the constraints which can be considered are $|\mathbf{C}| = 1$ as in (Frahm and Jaekel 2010) or $Tr\left\{\mathbf{C}\right\} = 1$ as in (Tyler 1987b) or (Sun et al. 2016).*

3.1.1.3. S-estimators

Since the robustness of M-estimators is dependent on the dimensionality of $\mathbf{Y}_t$, we discuss an extension to the problem (3.2) and other class of robust estimators which is shown to be less sensitive to the dimensionality of the observation process, known as S-estimators. The concept of S-estimators was firstly introduced by (Rousseeuw and Yohai 1984) and extended to multivariate setting by (Davies 1987)

Definition 9 (S - estimator). *The S-estimator $\hat{\Theta}$ of a parameter $\Theta = [\boldsymbol{\mu}, \mathbf{C}]$, that belongs to the parameter space $\Omega := \mathbb{R}^d \times \mathcal{P}^d$, is defined as a solution to*

$$\begin{aligned} \hat{\Theta}(F_N) &= \min_{\Theta^* \in \Omega} |\mathbf{C}| \ \textit{subject to} \\ &\int_{\mathbb{R}^d} \rho\left(D_t(\mathbf{y}_t)\right) dF(\mathbf{y}_t) \geq b_0, \end{aligned} \tag{3.10}$$

where D_t is a Mahalanobias distance of the observation point $\mathbf{y}_t$ under an assumption of the distribution of the random vector $\mathbf{Y}_t$, as defined in Definition 6, and the constant b_0 is the expectation of $\rho(D_t(\mathbf{Y}_t))$, that is

$$b_0 = \mathbb{E}_{\mathbf{Y}_t|\Theta}\left[\rho\left(D_t(\mathbf{Y}_t)\right)\right] = \mathbb{E}_{\mathbf{Y}_t|\Theta}\left[\rho\left(||\mathbf{Y}_t - \mu||_F\right)\right] = \frac{\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)} \int_0^\infty \rho(r) f(r) r^{d-1} dr.$$

Function ρ is defined as in Definition 6.

Remark 7. *If ρ is continuous then the problem (3.10) is an equality.*

In the study of (Davies 1987), the author investigates the properties of S-estimators under the assumption of elliptically distributed $\mathbf{Y}_t$. The work of (Davies 1987) gives the general assumptions on function $\rho : \mathbb{R}_+ \to [0, 1]$, e.g. being continuous on its domain and zero at the point c such that $0 < c < \infty$. The author proves that under these assumptions and if $Nb_0 \geq d + 1$, the problem (3.10) has at least one solution for non singular estimate $\hat{\Theta} = \left[\hat{\mu}, \hat{\mathbf{C}}\right]$.

The work of (Davies 1987) proves consistency of the estimator $\hat{\Theta}$ and its uniqueness under these assumptions. When additionally ρ has a continuous third derivative, the solution to (3.10), is asymptotically normal with the quadratic convergence convergence.

Let us denote $\epsilon = 1 - b_0$. Following (Davies 1987), if $N(1 - 2\epsilon) \geq N + 1$, and the sample is in general position, that is no more that d points of the sample lies on a $(d-1)$-dimensional hyperplane, then the finite - sample breakdown point is equal to

$$\epsilon_N^* = \frac{[N\epsilon] + 1}{N} \Rightarrow \lim_{n \to \infty} \epsilon_n^* = \epsilon = \epsilon^*.$$

where $[\cdot]$ denotes the integer component.

The alternative definition of the class of S-estimators is introduced in (Lopuhaa 1989). The authors investigate the relation between S-estimators and M-estimators. In this study, they assume a normalized version of the function ρ from Definition 9 in (Davies 1987)– the two definitions of ρ are equivalent under the following transformation

$$\rho \to 1 - \frac{\rho}{\sup \rho}.$$

Under the assumption of an elliptical distribution for $\mathbf{Y}_t$, the author shows that any solution to the problem (3.10) satisfies (3.2). Furthermore, one can rewrite the problem (3.10) in the similar manner to (3.4) using again a Lagrangian formulation:

$$\log \mathcal{L}_N(\Theta, \lambda) := \log |\mathbf{C}| - \lambda \left[\frac{1}{N} \sum_{t=1}^{N} \rho(D_t) - b_0 \right],$$

to obtain the equivalent set of estimation equations

$$\left(\hat{\Theta}(F_N), \lambda\right) = \underset{\Theta \in \Omega, \lambda \in \mathbb{R}}{\operatorname{argmin}} \log \mathcal{L}_N(\Theta,\ \lambda) \Rightarrow \begin{cases} \frac{1}{N} \sum_{t=1}^{N} u(D_t)(\mathbf{y}_t - \boldsymbol{\mu}) = 0, \\ \frac{d}{N} \sum_{t=1}^{N} \frac{u(D_t)}{v(D_t)} (\mathbf{y}_t - \boldsymbol{\mu})^T (\mathbf{y}_t - \boldsymbol{\mu}) = \mathbf{C}, \end{cases} \tag{3.11}$$

for $u(s) = \frac{\psi(s)}{s}$ and $v(s) = \psi(s)s - \rho(s) + b_0$. The term $-\rho(s) + b_0$ replaces as a substitute for the Lagrangian multiplier λ. Hence, every solution to the problem in Equation (3.10) is a solution to Equation (3.11) what determines the M-estimator problem defined by Equation (3.5).

Consequently, the author of (Lopuhaa 1989) argues that under this refined formulation, one can show that the S-estimators achieve the same asymptotic variance as corresponding M-estimators. However, as dimensionality of data increases, this class of estimators higher higher breakdown point than M-estimators.

Remark 8. *An example of functions which satisfy the conditions of uniqueness and existence of the S-estimators in Equation*(3.10) *are Tuckey's biweight functions given by*

$$\rho_B(s) = \begin{cases} \frac{s^2}{2} - \frac{s^4}{2k^2} + \frac{s^6}{6k^4}, & |s| \leq k, \\ \frac{k^2}{6}, & |s| \geq k \end{cases} \Rightarrow \psi_B(s) = s\left(1 - \left(\frac{s}{k}\right)^2\right)^2 \mathbf{1}_{[-k,k]}(s). \tag{3.12}$$

The breakdown point of the S-estimator for $\rho = 1 - \frac{\rho_B}{\sup \rho_B}$ *is equal to*

$$\epsilon^* = \frac{[N(1-b_0)]+1}{N} = \frac{\left[\frac{N}{\sup \rho_B}\mathbb{E}_{\mathbf{Y}_t|\mathbf{C},\boldsymbol{\mu}}\left[\rho_B\left(\mathbf{Y}_t - \boldsymbol{\mu}\right)\right]\right]+1}{N}.$$

3.1.1.4. Additional Comments on Robust Estimators

A special type of function used for the solution of (3.10) is the function $\rho(s) = \mathbf{1}_{\{|s|\leq c\}}$ with tuning parameter $c > 0$. As the function is not continuous, we must stay with inequality in the defined problem, which provides us with the solution $\hat{\Theta}(F_N)$

$$\begin{aligned} &\hat{\Theta}(F_N) = \min_{\Theta\in\Omega} |\Sigma| \text{ subject to} \\ &\frac{1}{N}\sum_{t=1}^{N} \mathbf{1}_{\{D_t\leq c\}} \geq b_0 \end{aligned} \tag{3.13}$$

or equivalently

$$\begin{aligned} &\hat{\theta}(F_N) = \min_{\Theta\in\Omega} |\mathbf{C}| \text{ subject to} \\ &\int \mathbf{1}_{\{D_t\leq c\}} dF_N \geq b_0 \end{aligned} \tag{3.14}$$

where c is chosen to ensure consistency of $\hat{\Theta}$ under assumed distribution of Mahalanobias distance D_t. It might fixed as $1-\alpha$ quantile of the distribution of Mahalanobias distance D_t for the random variable $\mathbf{Y}_t$. The parameter b_0 dictates the minimum size for data subset for computing determinant of covariance. Hence the upper limit of b_0 is whole sample size, n.

The estimator defined above is known as Minimum Volume Ellipsoid Estimator and was firstly introduced by (Rousseeuw 1985). The author recommends to use $h = \left[\frac{N}{2}\right] + 1$ and further proves that under the assumption that sample is in the general position, the finite sample breakdown point of the MVE estimator $\hat{\Theta}(F_N)$ to be equal to

$$\epsilon^* = \frac{\left[\frac{N}{2}\right] - d + 1}{N} \tag{3.15}$$

and thus asymptotically

$$\lim_{N\to\infty} \epsilon^* = \frac{1}{2}. \tag{3.16}$$

On the other hand, if one assumes there are at most α proportion of outliers in the data set, b_0 should be set to $b_0 = [N(1-\alpha)]+1$ what asymptotically results in the breakdown point equal to α.

The conditions for the existence of MVE estimator and its uniqueness can be found in (Davies 1992). Very briefly, to ensure the existence, h observation which built the estimator must not lie in a $(d-1)$ - dimensional hyperplane, that si, they are in a general position. However, it is shown that the MVE estimator is not unique for a general probability measure. citepDavis2003 suggests to use ordering as identifier of uniqueness and proves that under the aforementioned condition about the lower dimension hyperplane, there is a unique solution to the problem (3.14) for an empirical distribution. In the same study, it is shown that MVE estimator of mean and covariance matrix converges to its asymptotic distribution, which is non- Gaussian, at rate $N^{-\frac{1}{3}}$.

On the other hand, MVE estimator is known for its low maximum possible asymptotic bias as shown in (Croux and Haesbroeck 2001). The bias of the estimator is a measure of its behaviour when the fraction of contamination is below the breakdown point. Desirable it should be as small as possible. This property of MVE estimator makes it readily applicable as initial step for algorithms which computes more efficient estimators such as M- estimators or S-estimators.

Another S-estimator type robust statistics in a sense of the definition (3.10) is Minimum Covariance Determinant estimator. This estimator of scatter has been also introduced by (Rousseeuw 1985). It corresponds to finding b_0 points out of N for which a sample covariance matrix has the lowest determinant. The MCD estimate of location is then a sample average over these points and covariance matrix estimate is a sample covariance. The solution of MCD problem is defined similarly to (3.13) except for inequality, that is, the estimates $\hat{\Theta}(F_N)$ is obtained as the solution to the following

$$\begin{aligned} &\hat{\Theta}(F_N) = \min_{\theta \in \Omega} |\mathbf{C}| \text{ subject to} \\ &\frac{1}{N} \sum_{t=1}^{N} \mathbf{1}_{\{D_t \leq c\}} = b_0 \end{aligned} \tag{3.17}$$

or equivalently

$$\begin{aligned} &\hat{\Theta}(F_N) = \min_{\Theta \in \Omega} |\mathbf{C}| \text{ subject to} \\ &\int \mathbf{1}_{\{D_t \leq c\}} dF_N = b_0 \end{aligned} \tag{3.18}$$

For $\alpha = 1 - \frac{b_0}{n}$, when $\alpha \to 0$, MVE estimator of location is a center of ellipsoid defined by $\int \mathbf{1}_{\{d \leq c\}} dF_N$ covering chosen data points. The MCD estimator converges to the sample mean as shown in (Rousseeuw 1985).

(Croux and Haesbroeck 1999) focuses on the influence function of the MCD estimator of covariance matrix. It is shown that under the assumption of an elliptical distribution, the MCD problem (3.17) has an unique solution. Also, the authors show that the influence function of MCD estimator is bounded and provide the expression for asymptotic its variance.

(Rousseeuw 1985) proves the same breakdown point limits of MCD and MVE estimators. The authors highlights that MCD estimator achieves asymptotic normality with convergence rate $n^{-\frac{1}{2}}$. It is main advantage of MCD estimator over MVE, as the estimator obtains the same breakdown

point with better efficiency. Its consistency and asymptotic distribution are in detailed studied and proved by (Butler et al. 1993) and (Croux and Haesbroeck 1999).

3.2. Extension 1: t-Student Probabilistic Principal Component Analysis

In the following part, we outline a novel distribution-related methodology which employs the t-Student distribution assumptions in order to enhance the robustness of the Probabilistic Principal Component Analysis framework from Section 2.4. We discuss the existing solutions as well as provide our novel model extending the standard methods in an efficient and statistically robust manner.

The existing literature introduces the t-Student distribution in the PPCA model as an example of robust framework of Gaussian PPCA. The reader may refer to (Lange et al. 1989), (Archambeau et al. 2006), (Fang and Jeong 2008) or (Chen et al. 2009). The methodology assumes that the latent vectors and the error terms, $\mathbf{X}_t$ and $\boldsymbol{\epsilon}_t$, follow a multivariate t-Student distributions with a common degree of freedom denoted by ν. The realisations of these random vectors are further assumed to be independent in time, what forces each of their realisations to be generated from random vectors with different probability functions. Consequently, under these assumptions the observation process, $\mathbf{Y}_t$, which follows the model in Equation (2.7), is independent over time.

The assumptions provide the PPCA framework with the flexibility to treat each of the observations individually, however, it adds additional degrees of freedoms to the estimation problem which may cause unnecessary difficulties when the sample size is not sufficiently large. It also may overparameterize the problem in general, which may not bode well in the performance of the feature extraction. Further, the statistical characteristic about the observation set as well as the latent processes over time are limited since every realisation of the variables over time is distributed differently, an assumption often not true at least locally in time.

As a possible remedy to the listed limitations of Independent t-Student PPCA (t-Student IND PPCA), we propose an extension to the methodology by assuming $\mathbf{X}_t$ and $\boldsymbol{\epsilon}_t$ to be *identically distributed*, that is, their realisations have the same probability distribution over time and therefore are treated as realisations of one random variable. The reader will see that the assumption of the identical distribution restricts $\mathbf{X}_t$ and $\boldsymbol{\epsilon}_t$ to be only conditionally independent over time. The following section is based on our study discussed in (Toczydlowska and Peters 2018).

3.2.1. The t-Student Principal Component Analysis Model

Let Ψ be denoted as a vector which contains all static parameters of the model in Equation (2.7) and additional parameter ν which determines the tail behaviour and tail dependence of the multivariate t-Student distribution, that is $\Psi = [\mathbf{W}, \boldsymbol{\mu}, \sigma^2, \nu]$. Let $\mathbf{y}_{1:N}$ represents the set of of N realisations of $\mathbf{Y}_t$, which is available. Consider the definition of the t-Student random variable density (see (Gupta and Nagar 1999)), which is rewritten as a mixture of Gaussian and Gamma random variables. One may then define the sequence of scalar Gamma random

variables $U_t \sim \Gamma(\frac{\nu}{2}, \frac{\nu}{2})$ for $t = 1, \dots N$, whose elements are independent over time. The stochastic representations of the t-Student random vectors $\mathbf{X}_t$ and $\boldsymbol{\epsilon}_t$ may then be expressed by

$$\mathbf{X}_t = \frac{1}{\sqrt{U_t}}\mathbf{Z}_{x,t}, \ \boldsymbol{\epsilon}_t = \sqrt{\frac{\sigma^2}{U_t}}\mathbf{Z}_{\epsilon,t}, \tag{3.19}$$

where $\mathbf{Z}_{\epsilon,t}$ and $\mathbf{Z}_{x,t}$ are mutually independent d and k dimensional standard normal random vectors, respectively, which are independent over time. The joint probability density functions of the model in Equation (2.7) under the t-Student IND PPCA case is then given by

$$\pi_{\mathbf{Y}_{1:N},\mathbf{X}_{1:N},U_{1:N}|\Psi}(\mathbf{y}_{1:N}, \mathbf{x}_{1:N}, u_{1:N}) = \prod_{t=1}^{N} \pi_{\mathbf{Y}_t|\mathbf{X}_t,U_t,\Psi}(\mathbf{y}_t)\pi_{\mathbf{X}_t|U_t,\Psi}(\mathbf{x}_t)\pi_{U_t|\Psi}(u_t), \tag{3.20}$$

where the conditional distributions of $\mathbf{Y}_t|\mathbf{X}_t, U_t, \Psi$ and $\mathbf{X}_t|U_t, \Psi$ are Gaussian such that

$$\mathbf{Y}_t|\mathbf{X}_t, U_t, \Psi \sim \mathcal{N}\left(\boldsymbol{\mu} + \mathbf{X}_t\mathbf{W}^T, \frac{\sigma^2}{U_t}\mathbb{I}_d\right) \text{ and } \mathbf{X}_t|U_t, \Psi \sim \mathcal{N}\left(\mathbf{0}, \frac{1}{U_t}\mathbb{I}_k\right).$$

We extend the concept of the t-Student IND PPCA by proposing a more parsimonious model requiring a reduction of integrals from N to 1 when marginalizing to get the distribution of $\mathbf{Y}_{1:N}$. The latent variable U is no longer time dependent and the methodology assumes one realisation of $U \sim \Gamma\left(\frac{\nu}{2}, \frac{\nu}{2}\right)$ for the sample set. Consequently, the variables $\mathbf{X}_t$ and $\boldsymbol{\epsilon}_t$ are identically distributed. The stochastic representations of $\mathbf{X}_t$ and $\boldsymbol{\epsilon}_t$ under the new PPCA model, denoted as the t-Student Identical and Conditionally Independent PPCA model (t-Student IID PPCA), is given by

$$\mathbf{X}_t = \frac{1}{\sqrt{U}}\mathbf{Z}_{x,t}, \ \boldsymbol{\epsilon}_t = \sqrt{\frac{\sigma^2}{U}}\mathbf{Z}_{\epsilon,t}. \tag{3.21}$$

The vectors $\mathbf{X}_t$ and $\boldsymbol{\epsilon}_t$ are independent over time only when conditioned on U. Under these assumptions, the observation vector $\mathbf{Y}_t$ is still conditionally multivariate Gaussian, such that

$$\mathbf{Y}_t|\mathbf{X}_t, U, \Psi \sim \mathcal{N}\left(\boldsymbol{\mu} + \mathbf{X}_t\mathbf{W}^T, \frac{\sigma^2}{U}\mathbb{I}_d\right),$$

The observation vector $\mathbf{Y}_t$ is no longer marginally independent over time which results in the following joint probability density function of the model (2.7) under the t-Student IID PPCA case

$$\begin{aligned}\pi_{\mathbf{Y}_{1:N},\mathbf{X}_{1:N},U|\Psi}(\mathbf{y}_{1:N}, \mathbf{x}_{1:N}, u) &= \pi_{\mathbf{Y}_{1:N}|\mathbf{X}_{1:N},U,\Psi}(\mathbf{y}_{1:N}) \cdot \pi_{\mathbf{X}_{1:N}|U,\Psi}(\mathbf{x}_{1:N}) \cdot \pi_{U|\Psi}(u) \\ &= \pi_{U|\Psi}(u) \prod_{t=1}^{N} \pi_{\mathbf{Y}_t|\mathbf{X}_t,U\Psi}(\mathbf{y}_t) \cdot \pi_{\mathbf{X}_t|U,\Psi}(\mathbf{x}_t).\end{aligned} \tag{3.22}$$

3.2.2. Complete Data EM Algorithm for t-Student Probabilistic Principal Component Analysis

In the following section, we derive novel closed-form solutions to evaluate each step of the EM algorithm for t-Student PPCA cases discussed in Subsection 3.2.1. The proofs of the presented Theorems from the following subsection are detailed in Section A.3 in Appendix A.

The EM algorithm for the variants of t-Student PPCA is used to estimate the parameters $\mathbf{W}, \boldsymbol{\mu}, \sigma^2$ and filtered values of unobserved process $\mathbf{X}_t$. Recall, that the models introduced in Section 3.2.1 depend on one additional parameter, ν. The additional parameter corresponds to the degrees of freedom which determine a t-Student probability function and its tail behaviour. We will show in Section 3.3 that each of the different model variants, that have been presented previously in Section 3.2.1 for the t-Student PPCA, will produce a flat likelihood based profile for the degrees of freedom, making identification and estimation of this parameter challenging. Therefore we employ an independent grid-based search to choose ν by fixing the parameter and proceeding with EM algorithms for $\mathbf{W}, \boldsymbol{\mu}, \sigma^2$ which are derived in this section. The value of the scale parameter which obtains the highest log-likelihood is then chosen as an estimate of ν.

Therefore, the steps of the algorithm for the t-Student IND PPCA , under the fixed value of ν, are the following

Step 1: Expectation Step (E-step)

We calculate the expectation of the logarithm of the joint probability density function of all random vectors required by the model in Equation (2.7) under the distribution of all unobserved variables $\mathbf{X}_{1:N}, U_{1:N} | \mathbf{Y}_{1:N}$. The expectation is a function of two vectors with static parameters $\Psi = [\mathbf{W}, \boldsymbol{\mu}, \sigma^2]$ and its future update $\Psi^* = [\mathbf{W}^*, \boldsymbol{\mu}^*, \sigma^{*2}]$, that is

$$Q_{TS,ind}(\Psi, \Psi^*) = \mathbb{E}_{\mathbf{X}_{1:N}, \mathbf{U}_{1:N} | \mathbf{Y}_{1:N}, \Psi} \Big[\log \pi_{\mathbf{Y}_{1:N}, \mathbf{X}_{1:N}, U_{1:N} | \Psi^*} (\mathbf{y}_{1:N}, \mathbf{x}_{1:N}, u_{1:N}) \Big].$$

Step 2: Maximisation Step (M-step)

We find an update of the vector Ψ at each iteration of the EM algorithm via the maximisation of the defined function $Q_{TS,ind}$ with respect to Ψ^*, that is

$$\hat{\Psi}^* = \underset{\Psi^*}{\operatorname{argmax}}\, Q_{TS,ind}(\Psi, \Psi^*).$$

The steps of the EM Algorithm for t-Student IID PPCA are analogous but require replacing the vector $U_{1:N}$ with U as the model assumes one realisation of the Gamma variable per a sample set. Its objective function is denoted by $Q_{TS,id}$.

3.2.2.1. EM Algorithm for Probabilistic Principal Component with Independent t-Student Distribution

The $Q_{TS,ind}$ function of the E-step for the t-Student IND PPCA is provided in Theorem 9 whereas the derivation of its maximizers is given in Theorem 10.

Theorem 9. *Let the observation process $\mathbf{Y}_t$ follows the t-Student IND PPCA model defined in Subsection 3.2.1 in Equation (3.19) . Given N realisations of the observation vector $\mathbf{Y}_t$, denoted by $\mathbf{y}_{1:N} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$, the E-step of the corresponding EM algorithm is formulated as following*

$$
\begin{aligned}
Q_{TS,ind}(\Psi, \Psi^*) &= \mathbb{E}_{\mathbf{X}_{1:N}, U_{1:N}|\mathbf{Y}_{1:N}, \Psi}\left[\log \pi_{\mathbf{Y}_{1:N}, \mathbf{X}_{1:N}, U_{1:N}|\Psi^*}(\mathbf{y}_{1:N}, \mathbf{X}_{1:N}, U_{1:N})\right] \\
&= -\frac{1}{2}\sum_{t=1}^{N}\left\{\left(d + k - \frac{\nu^*}{2}\right)\log 2\pi + d\log\sigma^{*2} - (d + k + \nu^* - 2)\mathbb{E}_{U_t|\mathbf{Y}_t, \Psi}\left[\log U_t\right]\right. \\
&+ \frac{1}{\sigma^{*2}}\mathbb{E}_{U_t|\mathbf{Y}_t, \Psi}\left[U_t\right](\mathbf{y}_t - \boldsymbol{\mu}^*)(\mathbf{y}_t - \boldsymbol{\mu}^*)^T - \frac{2}{\sigma^{*2}}\mathbb{E}_{\mathbf{X}_t, U_t|\mathbf{Y}_t, \Psi}\left[U_t\mathbf{X}_t\right]\mathbf{W}^{*T}(\mathbf{y}_t - \boldsymbol{\mu}^*)^T \\
&\left. + Tr\left\{\left(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{W}^* + \mathbb{I}_k\right)\mathbb{E}_{\mathbf{X}_t, U_t|\mathbf{Y}_t, \Psi}\left[U_t\mathbf{X}_t^T\mathbf{X}_t\right]\right\} + \nu^*\mathbb{E}_{U_t|\mathbf{Y}_t, \Psi}\left[U_t\right] - 2\log\left(\frac{\left(\frac{\nu^*}{2}\right)^{\frac{\nu^*}{2}}}{\Gamma\left(\frac{\nu^*}{2}\right)}\right)\right\},
\end{aligned}
\tag{3.23}
$$

where the corresponding moments of the conditional distribution $\mathbf{X}_t, U_t|\mathbf{Y}_t, \Psi$ are given by

$$
\begin{aligned}
&\mathbb{E}_{U_t|\mathbf{Y}_t, \Psi}\left[U_t\right] = \frac{\nu + d}{\nu + (\mathbf{y}_t - \boldsymbol{\mu})\mathbf{C}^{-1}(\mathbf{y}_t - \boldsymbol{\mu})^T}, \\
&\mathbb{E}_{U_t|\mathbf{Y}_t, \Psi}\left[\log U_t\right] = \psi\left(\frac{\nu + d}{2}\right) - \log\left(\frac{\nu + (\mathbf{y}_t - \boldsymbol{\mu})\mathbf{C}^{-1}(\mathbf{y}_t - \boldsymbol{\mu})^T}{2}\right), \\
&\mathbb{E}_{\mathbf{X}_t, U_t|\mathbf{Y}_t, \Psi}\left[U_t\mathbf{X}_t\right]_{1\times k} = \mathbb{E}_{U_t|\mathbf{Y}_t, \Psi}\left[U_t\right](\mathbf{y}_t - \boldsymbol{\mu})\mathbf{W}\mathbf{M}^{-1} \\
&\mathbb{E}_{\mathbf{X}_t, U_t|\mathbf{Y}_t, \Psi}\left[U_t\mathbf{X}_t^T\mathbf{X}_t\right]_{k\times k} = \sigma^2\mathbf{M}^{-1} + \mathbb{E}_{U_t|\mathbf{Y}_t, \Psi}\left[U_t\right]\mathbf{M}^{-1}\mathbf{W}^T(\mathbf{y}_t - \boldsymbol{\mu})^T(\mathbf{y}_t - \boldsymbol{\mu})\mathbf{W}\mathbf{M}^{-1},
\end{aligned}
$$

for $\mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbb{I}_d$, $\mathbf{M}_{k\times k} = \mathbf{W}^T\mathbf{W} + \sigma^2\mathbb{I}_k$ and $\psi(\cdot)$ being digamma function.

Proof. The proof of the Theorem 9 is given in Subection A.3.1 in Appendix A.3. □

Theorem 10. *The maximizers of the function $Q_{TS,ind}(\Psi, \Psi^*)$ specified in Theorem 9 are the solutions to the system of equations given by $\nabla_{\Psi^*}Q_{TS,ind} = \mathbf{0}$. The maximizers with respect to $\boldsymbol{\mu}^*$, $\mathbf{W}^*$ and σ^{*2} are obtained in closed form given by*

$$
\begin{aligned}
\boldsymbol{\mu}^* &= \bar{u}(\mathbf{y}_{1:N}; \Psi)^{-1}\left(\bar{\boldsymbol{\mu}}_{ts}(\mathbf{y}_{1:N}; \Psi)\left(\mathbb{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\right) + \bar{u}(\mathbf{y}_{1:N}; \Psi)\boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\right), \\
\mathbf{W}^* &= \bar{\mathbf{C}}_{\mu,\mu^*}(\mathbf{y}_{1:N}; \Psi, \Psi^*)\mathbf{W}\mathbf{M}^{-1}\left(\sigma^2\mathbf{M}^{-1} + \mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\mu}(\mathbf{y}_{1:N}; \Psi)\mathbf{W}\mathbf{M}^{-1}\right)^{-1}, \\
\sigma^{*2} &= \frac{1}{d}Tr\left\{\bar{\mathbf{C}}_{\mu^*}(\mathbf{y}_{1:N}; \Psi^*) - 2\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\mu,\mu^*}(\mathbf{y}_{1:N}; \Psi, \Psi^*)\right. \\
&\quad \left. + \mathbf{W}^*\left(\sigma^2\mathbf{M}^{-1} + \mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\mu}(\mathbf{y}_{1:N}; \Psi)\mathbf{W}\mathbf{M}^{-1}\right)\mathbf{W}^{*T}\right\},
\end{aligned}
$$

where $D_t = (\mathbf{y}_t - \boldsymbol{\mu})\mathbf{C}^{-1}(\mathbf{y}_t - \boldsymbol{\mu})^T$, $\mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbb{I}_d$, $\mathbf{M}_{k\times k} = \mathbf{W}^T\mathbf{W} + \sigma^2\mathbb{I}_k$ and

$$
\begin{aligned}
\bar{u}(\mathbf{y}_{1:N}; \Psi) &= \frac{1}{N}\sum_{t=1}^{N}\frac{\nu + d}{\nu + D_t}, \\
\bar{\boldsymbol{\mu}}_{ts}(\mathbf{y}_{1:N}; \Psi) &= \frac{1}{N}\sum_{t=1}^{N}\mathbf{y}_t\frac{\nu + d}{\nu + D_t},
\end{aligned}
$$

$$\bar{\mathbf{S}}_{ts}(\mathbf{y}_{1:N};\Psi) = \frac{1}{N}\sum_{t=1}^{N}\frac{\nu+d}{\nu+D_t}\mathbf{y}_t^T\mathbf{y}_t,$$

$$\bar{\mathbf{C}}_{\mu,\mu^*}(\mathbf{y}_{1:N};\Psi,\Psi^*) = \bar{\mathbf{S}}_{ts}(\mathbf{y}_{1:N};\Psi) - \boldsymbol{\mu}^{*T}\bar{\boldsymbol{\mu}}_{ts}(\mathbf{y}_{1:N};\Psi) - \bar{\boldsymbol{\mu}}_{ts}(\mathbf{y}_{1:N};\Psi)^T\boldsymbol{\mu} + \bar{u}(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^{*T}\boldsymbol{\mu},$$

$$\bar{\mathbf{C}}_{\mu}(\mathbf{y}_{1:N};\Psi) = \bar{\mathbf{S}}_{ts}(\mathbf{y}_{1:N};\Psi) - 2\boldsymbol{\mu}^T\bar{\boldsymbol{\mu}}_{ts}(\mathbf{y}_{1:N};\Psi) + \bar{u}(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^T\boldsymbol{\mu},$$

$$\bar{\mathbf{C}}_{\mu^*}(\mathbf{y}_{1:N};\Psi,\Psi^*) = \bar{\mathbf{S}}_{ts}(\mathbf{y}_{1:N};\Psi) - 2\boldsymbol{\mu}^{*T}\bar{\boldsymbol{\mu}}_{ts}(\mathbf{y}_{1:N};\Psi) + \bar{u}(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^{*T}\boldsymbol{\mu}^*.$$

Proof. The proof of Theorem 10 is given in Subection A.3.2 in Appendix A.3. □

3.2.2.2. EM Algorithm for Probabilistic Principal Component with Identical and Conditionally Independent t-Student Distribution

Since the model from Subsection 3.2.1 for t-Student IID PPCA assumes one realisation of the random variable U to determine the distribution of $\mathbf{Y}_t$ over time, the E-step of the EM algorithm cannot be formulated as a sum over various conditional expectations given realisations and latent variables in different time points as in Theorem 9. Hence, we give the integral representation of the E-step for t-Student IID PPCA in Theorem 11 which has maximizers that can be obtained in closed forms with respect to Ψ^* as presented in Theorem 12.

Theorem 11. *Let the observation process $\mathbf{Y}_t$ follows the t-Student IID PPCA model defined in Subsection 3.2.1 in Equation (3.21) . Given N realisations of the observation vector $\mathbf{Y}_t$, denoted by $\mathbf{y}_{1:N} = \{\mathbf{y}_1,\ldots,\mathbf{y}_N\}$, the E-step of the corresponding EM algorithm is formulated as following*

$$\begin{aligned}Q_{TS,id}(\Psi,\Psi^*) &= \mathbb{E}_{\mathbf{X}_{1:N},U|\mathbf{Y}_{1:N},\Psi}\left[\log\pi_{\mathbf{Y}_{1:N},\mathbf{X}_{1:N},U|\Psi^*}(\mathbf{y}_{1:N},\mathbf{X}_{1:N},U)\right]\\ &= \frac{C_H(\mathbf{y}_{1:N};\Psi)}{\pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N})}\int_{\mathbb{R}_+}\tilde{\pi}_{U|\Psi}(u)\Big(\log\pi_{U|\Psi^*}(u) + \sum_{t=1}^{N}w(\mathbf{y}_t,u;\Psi,\Psi^*)\Big)\,du\end{aligned}$$

for $w:\mathbb{R}^d\times\mathbb{R}_+\to\mathbb{R}$ and the probability density function $\tilde{\pi}_{U|\Psi}(u)$, defined in Lemma A3 and Lemma A4 in Appendix A.3, respectively. The function $\tilde{\pi}_{U|\Psi}(u)$ is a density function of $U\sim\Gamma(\alpha,\beta(\mathbf{y}_{1:N};\Psi))$ where $\alpha = \frac{\nu}{2}+\frac{dN}{2}$ and

$$\beta(\mathbf{y}_{1:N};\Psi) := \frac{\nu}{2} + \frac{1}{2\sigma^2}\sum_{t=1}^{N}(\mathbf{y}_t-\boldsymbol{\mu})\left(\mathbb{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^T\right)(\mathbf{y}_t-\boldsymbol{\mu})^T,$$

$$C_H(\mathbf{y}_{1:N};\Psi) := (2\pi)^{-\frac{dN}{2}}(\sigma^2)^{-N\frac{d-k}{2}}\left|\mathbf{M}\right|^{-\frac{N}{2}}\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}\frac{\Gamma(\alpha)}{\Gamma\left(\frac{\nu}{2}\right)}\beta(\mathbf{y}_{1:N};\Psi)^{-\alpha}.$$

Proof. The proof of the Theorem 11 is given in Appendix A Subection A.3.3. □

Since the M-step specifies the gradient of the function $Q_{TS,id}$ in Theorem 11 with respect to the parameters Ψ^*, some of the terms in the integration problem will disappear. Therefore, when proceeding to the M-step it is sufficient to drop some integral terms which happen to be non-closed form, but are not required explicitly in the EM algorithm iterations and hence can be ignored for the purpose of this algorithms derivation.

Theorem 12. *The maximizers of $Q_{TS,id}\left(\Psi,\Psi^*\right)$ specified in Theorem 11 are the solution to the problem given by system of equations $\nabla_{\Psi^*}Q_{TS,id}=\mathbf{0}$. The maximizers with respect to $\boldsymbol{\mu}^*$, $\mathbf{W}^*$ and σ^{*2}, which can be obtained in closed form, are given by*

$$\begin{aligned}
\boldsymbol{\mu}^* &= \bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N};\Psi)\left(\mathbb{I}_d-\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\right)+\boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T},\\
\mathbf{W}^* &= \bar{\mathbf{C}}_{\mu,\mu^*}(\mathbf{y}_{1:N};\Psi,\Psi^*)\mathbf{W}\mathbf{M}^{-1}\left(\mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\mu}(\mathbf{y}_{1:N};\Psi)\mathbf{W}\mathbf{M}^{-1}+\frac{\beta(\mathbf{y}_{1:N};\Psi)}{\alpha}\sigma^2\mathbf{M}^{-1}\right)^{-1},\\
\sigma^{*2} &= \frac{1}{d}\,Tr\left\{\frac{\alpha}{\beta(\mathbf{y}_{1:N};\Psi)}\left(\bar{\mathbf{C}}_{\mu^*}(\mathbf{y}_{1:N};\Psi,\Psi^*)-2\bar{\mathbf{C}}_{\mu,\mu^*}(\mathbf{y}_{1:N};\Psi,\Psi)\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\right)\right\},\\
&\quad+\frac{1}{d}\,Tr\left\{\left(\frac{\alpha}{\beta(\mathbf{y}_{1:N};\Psi)}\mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\mu}(\mathbf{y}_{1:N};\Psi)\mathbf{W}\mathbf{M}^{-1}+\sigma^2\mathbf{M}^{-1}\right)\mathbf{W}^{*T}\mathbf{W}^*\right\},
\end{aligned}$$

where $\mathbf{M}=\mathbf{W}^T\mathbf{W}+\sigma^2\mathbb{I}_k$, $\alpha=\frac{\nu}{2}+\frac{dN}{2}$ *and*

$$\begin{aligned}
\beta(\mathbf{y}_{1:N};\Psi) &= \frac{\nu}{2}+\frac{1}{2\sigma^2}\sum_{t=1}^{N}\left(\mathbf{y}_t-\boldsymbol{\mu}\right)\left(\mathbb{I}_d-\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^T\right)\left(\mathbf{y}_t-\boldsymbol{\mu}\right)^T,\\
\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N};\Psi)_{1\times d} &= \frac{1}{N}\sum_{t=1}^{N}\mathbf{y}_t,\\
\bar{\mathbf{S}}(\mathbf{y}_{1:N};\Psi)_{d\times d} &= \frac{1}{N}\sum_{t=1}^{N}\mathbf{y}_t^T\mathbf{y}_t,\\
\bar{\mathbf{C}}_{\mu}(\mathbf{y}_{1:N};\Psi)_{d\times d} &= \bar{\mathbf{S}}(\mathbf{y}_{1:N};\Psi)-2\boldsymbol{\mu}^T\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N};\Psi)+\boldsymbol{\mu}^T\boldsymbol{\mu},\\
\bar{\mathbf{C}}_{\mu^*}(\mathbf{y}_{1:N};\Psi,\Psi^*)_{d\times d} &= \bar{\mathbf{S}}(\mathbf{y}_{1:N};\Psi)-2\boldsymbol{\mu}^{*T}\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N};\Psi)+\boldsymbol{\mu}^{*T}\boldsymbol{\mu}^*,\\
\bar{\mathbf{C}}_{\mu,\mu^*}(\mathbf{y}_{1:N};\Psi,\Psi^*)_{d\times d} &= \bar{\mathbf{S}}(\mathbf{y}_{1:N};\Psi)-\left(\boldsymbol{\mu}+\boldsymbol{\mu}^*\right)^T\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N};\Psi)+\boldsymbol{\mu}^{*T}\boldsymbol{\mu}.
\end{aligned}$$

Proof. The proof of Theorem 12 is given in Subection A.3.4 in Appendix A.3. □

3.2.3. Remark On Links Between W and Population Covariance in t-Student PPCA

In Subsection 2.3.1, we showed, following the reasoning of (Tipping and Bishop 1999b), that the k first eigenvectors of the population covariance matrix $\mathbf{C}$ correspond to the k singular vectors of the projection matrix $\mathbf{W}$ of the Gaussian PPCA model since $\mathbf{C}=\mathbf{W}\mathbf{W}^T+\sigma^2\mathbb{I}_d$. This result allows one to avoid estimation of covariance matrix and use instead the singular vectors of the matrix $\mathbf{W}$ as most meaningful directions of the data variance. The same relation can be proven for t-Student IND PPCA as demonstrated next. To see this consider,

$$\mathbf{Y}_t=\boldsymbol{\mu}+\frac{1}{\sqrt{U_t}}\mathbf{Z}_{x,t}\mathbf{W}^T+\sqrt{\frac{\sigma^2}{U_t}}\mathbf{Z}_{\epsilon,t}=\boldsymbol{\mu}+\frac{1}{\sqrt{U_t}}\left(\mathbf{Z}_{x,t}\mathbf{W}^T+\sqrt{\sigma^2}\mathbf{Z}_{\epsilon,t}\right)=\boldsymbol{\mu}+\frac{1}{\sqrt{U_t}}\mathbf{Z}_{y,t}$$

for $\mathbf{Z}_{y,t}\sim\mathcal{N}\left(0,\mathbf{W}\mathbf{W}^T+\sigma^2\mathbb{I}_d\right)$ being a sum of two independent multivariate Gaussian random variables. Notice that we utilized the stochastic representation of $\mathbf{X}_t$ and $\boldsymbol{\epsilon}_t$ specified in Equation (3.19) as well as the general model of PPCA from Equation (2.7). Hence, the population covariance

of the marginal distribution of $\mathbf{Y}_t$ is $\mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbb{I}_d$ and the same argument applies as valid for Gaussian PPCA.

The specification of the population covariance of t-Student IID PPCA model defined in Subsection 3.2.1 requires more advanced derivations since the observations are conditionally independent over time. Given the likelihood of the model in Equation (3.22) and applying Lemma A2 and Lemma A4, we obtain the following formulations of the likelihood function

$$\begin{aligned}
\pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N}) &= \int_{\mathbb{R}_+}\int_{\mathbb{R}^{N\times k}} \pi_{\mathbf{Y}_{1:N},\mathbf{X}_{1:N},U|\Psi}(\mathbf{y}_{1:N},\mathbf{x}_{1:N},u)\, d\curvearrowleft_{1:N}\, du \\
&\overset{\text{Lemma A2}}{=} \int_{\mathbb{R}_+} \pi_{U|\Psi}(u) \prod_{t=1}^{N} h(\mathbf{y}_t, u; \Psi)\, du \\
&\overset{\text{Lemma A4}}{=} C_H(\mathbf{y}_{1:N};\Psi)\int_{\mathbb{R}_+}\tilde{\pi}_{U|\Psi}(u)\, du = (2\pi)^{-\frac{dN}{2}}(\sigma^2)^{-N\frac{d-k}{2}}\left|\mathbf{M}\right|^{-\frac{N}{2}}\frac{\left(\frac{v}{2}\right)^{\frac{v}{2}}\Gamma(\alpha)}{\Gamma\left(\frac{v}{2}\right)}\beta(\mathbf{y}_{1:N};\Psi)^{-\alpha},
\end{aligned} \tag{3.24}$$

for $\alpha = \frac{v}{2} + \frac{dN}{2}$ and $\beta : \mathbb{R}^{N\times d} \to \mathbb{R}$ defined as

$$\beta(\mathbf{y}_{1:N};\Psi) := \frac{v}{2} + \frac{1}{2\sigma^2}\sum_{t=1}^{N}(\mathbf{y}_t - \boldsymbol{\mu})\left(\mathbb{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^T\right)(\mathbf{y}_t - \boldsymbol{\mu})^T.$$

Therefore, the function C_H is a density function of the random matrix $\mathbf{Y}_{1:N}$. In the next step we show that it can be seen as a density function of dN-dimensional t-Student random variable.

Let us define the random variable $\tilde{\mathbf{Y}} = \text{vec}\ \mathbf{Y}_{1:N}$ where the vec operator of an arbitrary $N \times d$ matrix $\mathbf{A}$ is defined such that $\text{vec}\,\mathbf{A}$ is a $1 \times dN$ vector obtained by stacking the columns of $\mathbf{A}$ underneath each other in order from left to right. Let $\otimes$ be a Kronecker product. The function β can be reformulated such that

$$\beta(\mathbf{y}_{1:N};\Psi) := \frac{v}{2} + \frac{1}{2\sigma^2}(\tilde{\mathbf{y}} - \tilde{\boldsymbol{\mu}})\tilde{\mathbf{V}}(\tilde{\mathbf{y}} - \tilde{\boldsymbol{\mu}})^T$$

where $\tilde{\mathbf{y}}$ is a single realisation of $\tilde{\mathbf{Y}}$, $\tilde{\boldsymbol{\mu}}_{1\times dN} := \mathbf{1}_N \otimes \boldsymbol{\mu}$ and $\tilde{\mathbf{V}}_{dN\times dN} := \mathbb{I}_N \otimes \mathbf{V}$ is a square, block-diagonal matrix where $\mathbf{V} = \mathbb{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^T$. Then

$$\begin{aligned}
\pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N}) = \pi_{\tilde{\mathbf{Y}}|\Psi}(\tilde{\mathbf{y}}) &= (2\pi)^{-\frac{dN}{2}}(\sigma^2)^{-N\frac{d-k}{2}}\left|\mathbf{M}\right|^{-\frac{N}{2}}\frac{\left(\frac{v}{2}\right)^{\frac{v}{2}}\Gamma(\frac{v+dN}{2})}{\Gamma\left(\frac{v}{2}\right)}\left(\frac{v}{2} + \frac{1}{2\sigma^2}(\tilde{\mathbf{y}} - \tilde{\boldsymbol{\mu}})\tilde{\mathbf{V}}(\tilde{\mathbf{y}} - \tilde{\boldsymbol{\mu}})^T\right)^{-\frac{v+dN}{2}} \\
&= (2\pi)^{-\frac{dN}{2}}(\sigma^2)^{-N\frac{d-k}{2}}\left|\mathbf{M}\right|^{-\frac{N}{2}}\frac{\left(\frac{v}{2}\right)^{\frac{v}{2}}\Gamma(\frac{v+dN}{2})}{\Gamma\left(\frac{v}{2}\right)}\left(\frac{v}{2}\right)^{-\frac{v+dN}{2}}\left(1 + \frac{1}{v\sigma^2}(\tilde{\mathbf{y}} - \tilde{\boldsymbol{\mu}})\tilde{\mathbf{V}}(\tilde{\mathbf{y}} - \tilde{\boldsymbol{\mu}})^T\right)^{-\frac{v+dN}{2}} \\
&= (\sigma^2)^{-N\frac{d-k}{2}}\left|\mathbf{M}\right|^{-\frac{N}{2}}\left|\sigma^2\tilde{\mathbf{V}}^{-1}\right|^{\frac{1}{2}}\frac{\Gamma(\frac{v+dN}{2})}{\Gamma\left(\frac{v}{2}\right)\pi^{\frac{dN}{2}}v^{\frac{dN}{2}}\left|\sigma^2\tilde{\mathbf{V}}^{-1}\right|^{\frac{1}{2}}}\left(1 + \frac{1}{v\sigma^2}(\tilde{\mathbf{y}} - \tilde{\boldsymbol{\mu}})\tilde{\mathbf{V}}(\tilde{\mathbf{y}} - \tilde{\boldsymbol{\mu}})^T\right)^{-\frac{v+dN}{2}} \\
&= \tilde{C}\,\frac{\Gamma(\frac{v+dN}{2})}{\Gamma\left(\frac{v}{2}\right)\pi^{\frac{dN}{2}}v^{\frac{dN}{2}}\left|\sigma^2\tilde{\mathbf{V}}^{-1}\right|^{\frac{1}{2}}}\left(1 + \frac{1}{v\sigma^2}(\tilde{\mathbf{y}} - \tilde{\boldsymbol{\mu}})\tilde{\mathbf{V}}(\tilde{\mathbf{y}} - \tilde{\boldsymbol{\mu}})^T\right)^{-\frac{v+dN}{2}},
\end{aligned}$$

for a normalising constant $\tilde{C} := (\sigma^2)^{\frac{kN}{2}} \left|\mathbf{M}\right|^{-\frac{N}{2}} \left|\mathbf{V}\right|^{-\frac{N}{2}}$. Hence, $\pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N})$ is proportional to the probability density function of the dN-dimensional t-Student random variable with location $\tilde{\boldsymbol{\mu}}$ and the shape matrix $\tilde{\boldsymbol{\Sigma}} = \sigma^2 \tilde{\mathbf{V}}^{-1}$ for $\tilde{\mathbf{V}}^{-1} = \mathbb{I}_N \otimes \mathbf{V}^{-1}$. The last step is to show the link between the singular decomposition of the matrix $\mathbf{W}$ and the eigen decomposition of the matrix $\tilde{\mathbf{V}}^{-1}$. Given the SVD decomposition of the matrix $\mathbf{W} = \mathbf{U}_{d\times d}\mathbf{D}_{d\times k}\mathbf{U}^*_{k\times k}$, the matrix $\mathbf{V}$ can be represented such that

$$
\begin{aligned}
\mathbf{V} &= \mathbb{I}_d - \mathbf{U}\mathbf{D}\mathbf{U}^* \left(\mathbf{U}^{*T}\mathbf{D}^T\mathbf{U}^T\mathbf{U}\mathbf{D}\mathbf{U}^* + \sigma^2\mathbb{I}_k\right)^{-1}\mathbf{U}^{*T}\mathbf{D}^T\mathbf{U}^T \\
&= \mathbb{I}_d - \mathbf{U}\mathbf{D}\mathbf{U}^* \left(\mathbf{U}^{*T}\mathbf{D}^T\mathbf{D}\mathbf{U}^* + \sigma^2\mathbf{U}^{*T}\mathbf{U}^*\right)^{-1}\mathbf{U}^{*T}\mathbf{D}^T\mathbf{U}^T \\
&= \mathbb{I}_d - \mathbf{U}\mathbf{D}\mathbf{U}^* \left(\mathbf{U}^{*T}\left(\mathbf{D}^T\mathbf{D} + \sigma^2\right)\mathbf{U}^*\right)^{-1}\mathbf{U}^{*T}\mathbf{D}^T\mathbf{U}^T \\
&= \mathbb{I}_d - \mathbf{U}\mathbf{D}\left(\mathbf{D}^T\mathbf{D} + \sigma^2\right)^{-1}\mathbf{D}^T\mathbf{U}^T \\
&= \mathbf{U}\mathbf{U}^T - \mathbf{U}\tilde{\mathbf{D}}\mathbf{U}^T = \mathbf{U}\left(\mathbb{I}_d - \tilde{\mathbf{D}}\right)\mathbf{U}^T = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T,
\end{aligned}
$$

for the diagonal matrices $\tilde{\mathbf{D}}_{d\times d} := \mathbf{D}\left(\mathbf{D}^T\mathbf{D} + \sigma^2\right)^{-1}\mathbf{D}^T$ and $\boldsymbol{\Lambda}_{d\times d} := \mathbb{I}_d - \tilde{\mathbf{D}}$. The element in i row and i column of the matrix $\boldsymbol{\Lambda}$ equals $\lambda_{ii} = 1 - \frac{d_{ii}^2}{d_{ii}^2+\sigma^2} = \frac{\sigma^2}{d_{ii}^2+\sigma^2}$ if $i \leq k$, $\lambda_{ii} = 1$ if $i > k$, otherwise 0, where d_{ii} is an element of matrix $\mathbf{D}$ in i row and i column. Consequently, we obtain the eigen decomposition of the shape matrix $\tilde{\boldsymbol{\Sigma}} = \sigma^2\tilde{\mathbf{V}}^{-1}$ which is also explained by the singular decomposition of the matrix $\mathbf{W}$, since

$$
\tilde{\mathbf{V}}^{-1} = \mathbb{I}_N \otimes \mathbf{V}^{-1} = \mathbb{I}_N \otimes \left(\mathbf{U}\boldsymbol{\Lambda}^{-1}\mathbf{U}^T\right) = \left(\mathbb{I}_N \otimes \mathbf{U}\right)\left(\mathbb{I}_N \otimes \boldsymbol{\Lambda}^{-1}\right)\left(\mathbb{I}_N \otimes \mathbf{U}^T\right).
$$

We can remark that each realisation of the random vector $\mathbf{Y}_t$ which follows t-Student IID PPCA model from Subsection 3.2.1, is characterised by the same location vector $\boldsymbol{\mu}$ and shape matrix $\sigma^2\mathbf{V}^{-1}$ which eigenvectors correspond to left singular vectors of the matrix $\mathbf{W}$ and its eigenvalues are a function of singular values of the matrix $\mathbf{W}$ and the model's parameter σ^2. The eigenvalues of the shape matrix are equal to $d_{ii}^2 + \sigma^2$ if $i \leq k$, otherwise σ^2.

3.2.4. The Statistical Robustness of t-Student PPCA

The general form of Maximum Likelihood Estimation (MLE) problem for any statistical model is the following

$$
\Psi_N = \underset{\Psi}{\operatorname{argmax}} \log\left(\pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N})\right)
$$

where $\pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N})$ denote the likelihood of $\mathbf{Y}_t$, which correspond to the observation process and follows some assumed model, givne its N realisation $\mathbf{y}_{1:N}$.

Under the Gaussian PPCA model from Section 3.1, we can specify the marginal distribution of the observation vector $\mathbf{Y}_t$ to be Gaussian with the mean vector $\boldsymbol{\mu}$ and a covariance matrix $\mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbb{I}_d$. The normal equations which correspond to the MLE, are specified by the optimisation problem

$$
\hat{\boldsymbol{\mu}},\ \hat{\mathbf{C}} = \underset{\boldsymbol{\mu},\mathbf{C}}{\operatorname{argmax}} \sum_{t=1}^{N} \log\left(\pi_{\mathbf{Y}_t|\Psi}(\mathbf{y}_t)\right) = \underset{\boldsymbol{\mu},\mathbf{C}}{\operatorname{argmin}} \left\{ N\log\left|\mathbf{C}\right| + \sum_{t=1}^{N} \left(\mathbf{y}_t - \boldsymbol{\mu}\right)\mathbf{C}^{-1}\left(\mathbf{y}_t - \boldsymbol{\mu}\right) \right\}.
$$

After differentiating the objective function to obtain the maximum one obtains the following system of equations

$$\begin{cases} -N\boldsymbol{\mu} + \sum_{t=1}^{N} \mathbf{y}_t = \mathbf{0} \\ N\mathbf{C}^{-1} - \sum_{t=1}^{N} \mathbf{C}^{-1}(\mathbf{y}_t - \boldsymbol{\mu})^T(\mathbf{y}_t - \boldsymbol{\mu})\mathbf{C}^{-1} = \mathbf{0} \end{cases} \iff \begin{cases} \hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{t=1}^{N} \mathbf{y}_t, \\ \hat{\mathbf{C}} = \frac{1}{N}\sum_{t=1}^{N} (\mathbf{y}_t - \boldsymbol{\mu})^T(\mathbf{y}_t - \boldsymbol{\mu}). \end{cases}$$

The MLE estimators of the Gaussian PPCA model can be obtained in closed form. However, the information captured by provided eigen vectors can lose the intended interpretation (statistical summary) if the data is corrupted by even a single observation. Please refer to early studies on the estimation of location vector and covariance matrix, discussed in Section 3.1 or recent works referring to robustness of the Gaussian PPCA itself, such as (Xie and Xing 2015).

The generalized version of MLE by M-estimators is defined by the optimisation problem from Subsubsection 3.1.1.2 in Equation (3.1) which results in the system of normal equations

$$\begin{cases} \sum_{t=1}^{N} \frac{\psi(D_t)}{D_t} \mathbf{C}^{-\frac{1}{2}} (\mathbf{y}_t - \boldsymbol{\mu}) = \mathbf{0}, \\ \sum_{t=1}^{N} \frac{\psi(D_t)}{D_t} \mathbf{C}^{-\frac{1}{2}} (\mathbf{y}_t - \boldsymbol{\mu})^T (\mathbf{y}_t - \boldsymbol{\mu}) \mathbf{C}^{-\frac{1}{2}} = \mathbf{0}. \end{cases}$$

The robustness of the obtained estimators and their properties vary as one chooses different functional forms for Ψ as discussed in Subsection 3.1.1.

The analogous system of normal equation for estimation of the mean and covariance matrix of random vectors which follows the t-Student IND PPCA model, is given by the optimisation problem

$$\hat{\boldsymbol{\mu}}, \hat{\mathbf{C}} = \underset{\boldsymbol{\mu},\mathbf{C}}{\operatorname{argmin}} \left\{ N \log |\mathbf{C}| + (v+d) \sum_{t=1}^{N} \log \left(1 + \frac{1}{v}(\mathbf{y}_t - \boldsymbol{\mu})\mathbf{C}^{-1}(\mathbf{y}_t - \boldsymbol{\mu})\right) \right\},$$

which after differentiation will produce the system of equations given by

$$\begin{cases} \sum_{t=1}^{N} \frac{\mathbf{C}^{-1}(\mathbf{y}_t - \boldsymbol{\mu})}{v + (\mathbf{y}_t - \boldsymbol{\mu})\mathbf{C}^{-1}(\mathbf{y}_t - \boldsymbol{\mu})^T} = \mathbf{0}, \\ N\mathbf{C}^{-1} - (v+d)\sum_{t=1}^{N} \frac{\mathbf{C}^{-1}(\mathbf{y}_t - \boldsymbol{\mu})^T(\mathbf{y}_t - \boldsymbol{\mu})\mathbf{C}^{-1}}{v + (\mathbf{y}_t - \boldsymbol{\mu})\mathbf{C}^{-1}(\mathbf{y}_t - \boldsymbol{\mu})^T} = \mathbf{0}, \end{cases} \iff \begin{cases} \sum_{t=1}^{N} \frac{\mathbf{y}_t - \boldsymbol{\mu}}{v + (\mathbf{y}_t - \boldsymbol{\mu})\mathbf{C}^{-1}(\mathbf{y}_t - \boldsymbol{\mu})^T} = \mathbf{0}, \\ \mathbf{C} = \frac{v+d}{N}\sum_{t=1}^{N} \frac{(\mathbf{y}_t - \boldsymbol{\mu})^T(\mathbf{y}_t - \boldsymbol{\mu})}{v + (\mathbf{y}_t - \boldsymbol{\mu})\mathbf{C}^{-1}(\mathbf{y}_t - \boldsymbol{\mu})^T}. \end{cases} \tag{3.25}$$

We observe that if $v \to 0$, the bottom equation from Equation (3.25) corresponds to the formulation of Tyler estimator of the covariance from Definition 8 which is unique up to the scaling parameter. Hence the Tyler Estimator of a covariance matrix is a limiting case of the t-Student estimator.

Lastly, given the likelihood function of the t-Student IID PPCA model derived in Equation (3.24), the objective function of the MLE problem of the mean and covariance matrix is formulated as follows

$$\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}} = \underset{\boldsymbol{\mu},\boldsymbol{\Sigma}}{\operatorname{argmax}} \log \left(\pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N})\right)$$

$$= \underset{\boldsymbol{\mu},\boldsymbol{\Sigma}}{\operatorname{argmin}} \left\{ N \log |\boldsymbol{\Sigma}| + (v + dN) \log \left(1 + \frac{1}{v} \sum_{t=1}^{N} (\mathbf{y}_t - \boldsymbol{\mu}) \boldsymbol{\Sigma}^{-1} (\mathbf{y}_t - \boldsymbol{\mu})^T \right) \right\},$$

where $\boldsymbol{\Sigma} = \sigma^2 \mathbf{V}^{-1}$. The resulting system of normal equation equations is given by

$$\begin{cases} \frac{\sum_{t=1}^{N} (\mathbf{y}_t - \boldsymbol{\mu})}{1 + \frac{1}{v} \sum_{t=1}^{N} (\mathbf{y}_t - \boldsymbol{\mu}) \boldsymbol{\Sigma}^{-1} (\mathbf{y}_t - \boldsymbol{\mu})^T} = \mathbf{0}, \\ N\boldsymbol{\Sigma}^{-1} - (v + dN) \frac{\sum_{t=1}^{N} \boldsymbol{\Sigma}^{-1} (\mathbf{y}_t - \boldsymbol{\mu})^T (\mathbf{y}_t - \boldsymbol{\mu}) \boldsymbol{\Sigma}^{-1}}{1 + \frac{1}{v} \sum_{t=1}^{N} (\mathbf{y}_t - \boldsymbol{\mu}) \boldsymbol{\Sigma}^{-1} (\mathbf{y}_t - \boldsymbol{\mu})^T} = \mathbf{0}, \end{cases} \iff \begin{cases} \boldsymbol{\mu} = \frac{1}{N} \sum_{t=1}^{N} \mathbf{y}_t, \\ \boldsymbol{\Sigma} = \frac{\frac{v+dN}{N} \sum_{t=1}^{N} (\mathbf{y}_t - \boldsymbol{\mu})^T (\mathbf{y}_t - \boldsymbol{\mu})}{1 + \frac{1}{v} \sum_{t=1}^{N} (\mathbf{y}_t - \boldsymbol{\mu}) \boldsymbol{\Sigma}^{-1} (\mathbf{y}_t - \boldsymbol{\mu})^T}. \end{cases} \tag{3.26}$$

We observe that the estimators obtained under the t-Student IID PCA model correspond to a weighted version of the estimators provided by the Gaussian PPCA with weights which summarize the outlyingness of the whole sample set. Let us remark that the t-Student IND PPCA model adjusts each elements of the sample set.

3.3. Case Study: Comparison of Probabilistic Principal Component Analysis Frameworks on Synthetic Data

In the next section we demonstrate the performance of the feature extraction frameworks introduced so far, the standard Gaussian PPCA in Section 2.4, the robust Gaussian PPCA in Section 3.1 and the two t-Student PPCA frameworks detailed in Section 3.2, as well as the standard Gaussian Probabilistic Principal Component Analysis methodology introduced by (Tipping and Bishop 1999b) and its robust alternative which adapts the robust estimation of mean vectors and covariance matrices described in Section 3.1. The synthetic data case studies address the following questions:

Q1: How sensitive are the investigated methodologies to the initialisation step of the algorithms? Does the dimensionality of the data, the number of observations, a ration of missing values and proportion of outliers have an impact on the stability of the algorithms?

Q2: How successful are the newly derived methodologies in estimating the components of model (1) under different proportions of missing values and other perturbation to the data sample? Does the type of the perturbation (row-wise or element-wise) have an impact on the performance of the methodologies?

Hence, the first subsection discuss the stability of investigated the EM algorithms and their ability to estimate the parameters of the model in Equation (2.7) when the assumptions of the PPCA model agree with the distribution of the underlying data. Hence, we test Gaussian PPCA and robust Gaussian PPCA algorithms given the samples from a multivariate Gaussian distribution, and we examine the t-Student frameworks given the data which follows multivariate t-Student distributions discussed in Section 3. We study the performance of the algorithms for the different dimensionality of a single observation and the sample size as well as various proportions of missing values in the data set without corrupting the data sample with any additional noise.

Next, we focus on the efficiency of the algorithms in estimating the parameters of the model in Equation (2.7) given the sample data which is corrupted. We generate the data from a multivariate Gaussian distribution which then is perturbed row-wise (the whole observation is perturbed) or element-wise (only single elements of an observation vector are perturbed) by multivariate Laplace distribution with zero mean and the covariance matrix of Gaussian data. We study the performance of the algorithm under different dimensionality of the data, the number of observations, proportion of missing values and proportions of the corrupted sample.

We test the algorithms given 100 simulations and consider various input parameters: sample size $N = 200,\ 500$; the dimensionality of a single observation $d = 5,\ 10,\ 20$; the proportion of missing values $missing = 0\%,\ 20\%,\ 50\%$ and proportion of the perturbed sample $perturbed = 20\%,\ 50\%$. We examine the robust Gaussian PPCA framework for breakdown points $bp = 0.1, 0.2, 0.3, 0.4, 0.45$. We choose the grid of degrees of freedom for t-Student PPCA frameworks to be flexible and reflect both heavy tail assumptions and less robust frameworks, that is $dof = \left\{\{1, \ldots, 10\}, 15, 20, 30, 50, 75, 100\right\}$. In addition, we examine the effect of the robust initialisation on the performance of the algorithms with corresponding input variable $robustInit = \text{TRUE},\ \text{FALSE}$.

3.3.1. Testing PPCA Algorithms For Non-Perturbed Sample

We examine the recovery of the following static parameters: the mean vector $\boldsymbol{\mu}_{1\times d}$ and the $k = 3$ first eigenvectors of the covariance matrix $\mathbf{C}_{d\times d} = \mathbf{W}^T\mathbf{W} + \sigma^2\mathbf{I}_d$ for $d \times k$ real matrix $\mathbf{W}$. Additionally, we remove some portion of data equal to the parameter *missing* in order to test how the proposed estimation framework handles presence of incomplete observations.

We simulate 100 estimations of the true parameters for each of the investigated PPCA frameworks and various assumptions on the dimensionality of the data, sample size, the proportion of missingness. We set the sample set to be fixed per dimensionality in order to examine the sensitivity of the algorithms to the initialisation step. Therefore, each of the plots in this sections shows the median of results with corresponding 97.5% empirical quantiles. For t-Student PPCA cases, we keep the parameter of degrees of freedom fixed to its true value $v = 4$.

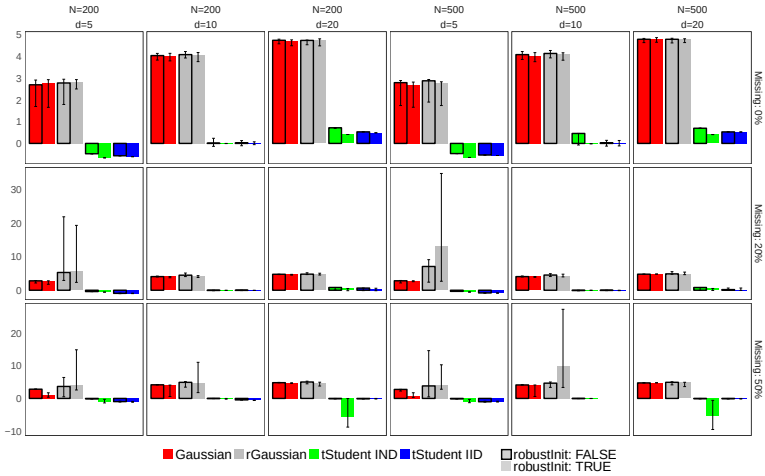


Figure 3.1. The median of the loglikelihood functions of EM algorithm for: Gaussian PPCA (red), robust Gaussian PPCA (grey), t-Student IND PPCA (green) and t-Student IID PPCA (blue) for (column-wise) different dimensionality of the observation (d) and sample size (N) and (row-wise) proportion of missing values in sample set (*missing*). The bars with solid black borders correspond to the results of the algorithms which have not been initialized with the robust step. The vertical solid black lines highlight the 2.5% and 97.5% empirical quantiles of the obtained results.

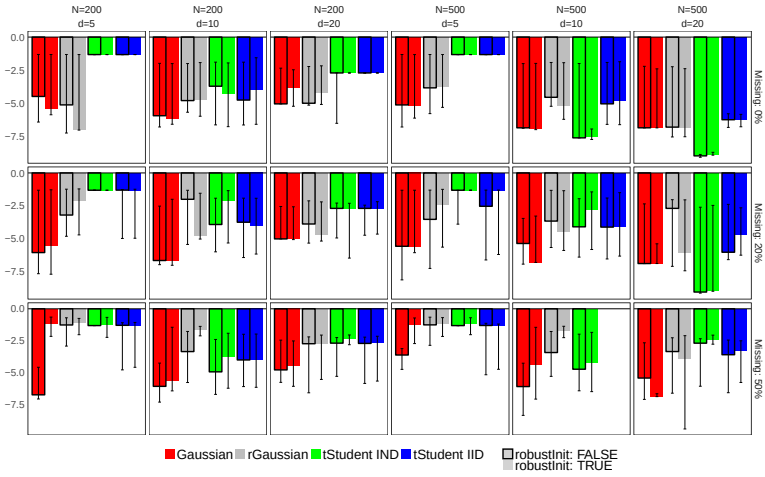


Figure 3.2. The logarithm of the squared Frobenius norm of the difference between the three first eigen vectors of the true matrix $\mathbf{C} = \mathbf{W}^T\mathbf{W} + \sigma^2\mathbf{I}_d$ and its estimators $\hat{\mathbf{C}} = \hat{\mathbf{W}}^T\hat{\mathbf{W}} + \hat{\sigma}^2\mathbf{I}_d$ given by by EM algorithm for: Gaussian PPCA (red), Robust Gaussian PPCA (grey), t-Student IND PPCA (green) and t-Student IID PPCA (blue) for (column-wise) different dimensionality of a single observation (d) and the sample size (N) and (row-wise) the proportion of the missing values in the sample set (*missing*). The bars with solid black borders correspond to the resuts fo the algorithms which have robust initialisation. The vertical solid black lines highlights the 2.5% and 97.5% quantiles of the results.

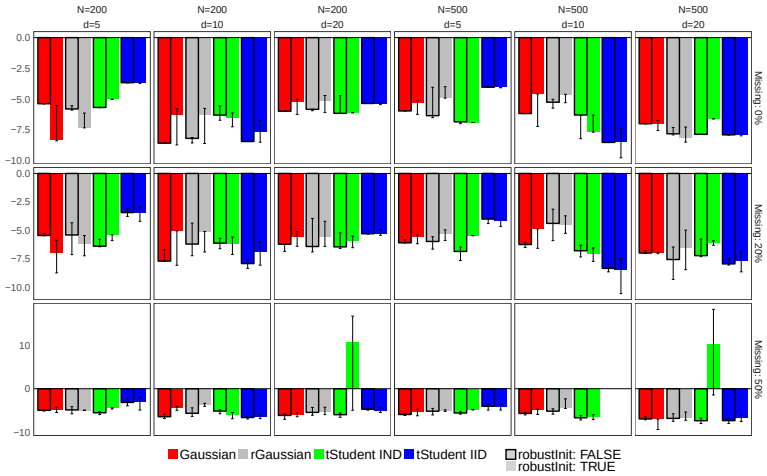


Figure 3.3. The logarithm of the squared Frobenius norm of the difference between the true value of the vector of means $\boldsymbol{\mu}$ and its estimator given by by EM algorithm for: Gaussian PPCA (red), Robust Gaussian PPCA (grey), t-Student IND PPCA (green) and t-Student IID PPCA (blue) for (column-wise) different dimensionality of a single observation (d) and the sample size (N) and (row-wise) the proportion of the missing values in the sample set (*missing*). The bars with solid black borders correspond to the results for the algorithms which have robust initialisation. The vertical solid black lines highlight the 2.5% and 97.5% quantiles of the results.

Figure 3.1 illustrates the final loglikelihood functions of the EM algorithms for different variants of PPCA framework. The bars show the median loglikelihood of the corresponding EM algorithms over 100 simulations per case study whereas the vertical solid lines denote 2.5% and 97.5% empirical quantiles of the results. The non-zero interquartile range informs us about the sensitivity of the algorithms to the initialisation step. Recall that we test the algorithms on the datasets which correspond to the PPCA assumptions, hence the magnitude of resulting loglikelihood across different PPCA framework does not inform us which algorithm fits the data most accurately. We can notice that the algorithms are the most affected by changes to the initialisation when the dimensionality of the data is the lowest regardless of the sample size, especially when the proportion of missing values is the highest. The wide interquartile intervals of robust Gaussian PPCA (grey bars) for $d = 5$ and non zero missingness shows that this methodology is the most sensitive to the initialisation. It is the intuitive outcome since the multivariate frameworks of the robust covariance estimation are very sensitive to their own initialisation what increases the overall sensitivity of the robust Gaussian PPCA framework. On the other hand, the methodology provides solutions with much higher likelihood that the standard Gaussian PPCA for lower dimensionality cases with a presence of missing values

The t-Student PPCA algorithms are less affected by the initialisation step for most of the cases but for the higher dimensionality of the sample data and highest portion of missing values we notice greater sensitivity of t-Student IND PPCA when we do not proceed with robust initialisation. For the other cases related to the two t-Student PPCA frameworks, the robust initialisation neither does decrease nor increase the sensitivity of the algorithm. In addition, robustly initialized t-Student PPCA tends to find solutions in local maximums of likelihoods which have higher values than their equivalents provided by non-robustly initialised algorithms, especially for t-Student IND PPCA.

The convergence of the algorithms to different local maxima of the corresponding loglikelihood function of the EM algorithms results in different estimations of the model's (1) static parameters and therefore the eigenvectors of the covariance matrix differ. The Figure 3.2 illustrates the discrepancies between the matrix with three most meaningful eigenvectors of the true covariance matrix and its estimation across different cases and methodologies. We use the squared Frobenius norm of the difference between matrices as a measure of the distance.

As expected, the sensitivity of the estimation algorithms to the initialisation step influences the variability of the estimators. The plots in Figure 3.2 shows the median of the norms with the corresponding 97.2% empirical quantiles. The standard EM algorithm for Gaussian PPCA outperforms its robust equivalent, especially for the incomplete data cases. Even though the robust Gaussian algorithms converged to the higher likelihood of the data, its quality of the recovery of the matrix with eigenvectors is worse than the standard Gaussian PPCA. Recall that the Gaussian PPCA is the least affected methodology by the change of the input parameters: dimensionality, sample sizes and the proportion of missing values.

One may notice the improvement of estimation of the eigenvectors obtained by t-Student frameworks with the increase of the sample size, especially for t-Student IND PPCA. Both methodologies provide the most accurate estimate of the eigenvectors for higher dimensions of the data. The median of the results between the frameworks initialised robustly or not exhibit small despondencies, but the robust initialisation increases their interquartile range, and consequently the certainty.

Figure 3.3 illustrates the estimation of the mean vector $\boldsymbol{\mu}$. Similarly to the matrix with eigenvectors, we use the squared Frobenius norm of the difference between the true value of the vector of means $\boldsymbol{\mu}$ and its estimate to asses the distance between the vectors. The plots show the median of the logarithms of the norms with corresponding 97.5.% empirical confidence intervals. The accuracy of the estimation of $\boldsymbol{\mu}$ is less volatile in different cases than for the corresponding matrices with eigenvectors. We observe a small decrease of the confidence of the results provided by the algorithms with the robust initialisation, with the exception of the t-Student IND PPCA for the highest proportion of missing values and dimensionality. The algorithms tend to provide poor accuracy of the estimation of $\boldsymbol{\mu}$ what explains the low loglikelihood values of the corresponding EM algorithms shown in Figure 3.1. We can try to improve the results by restraining the convergence criterion.

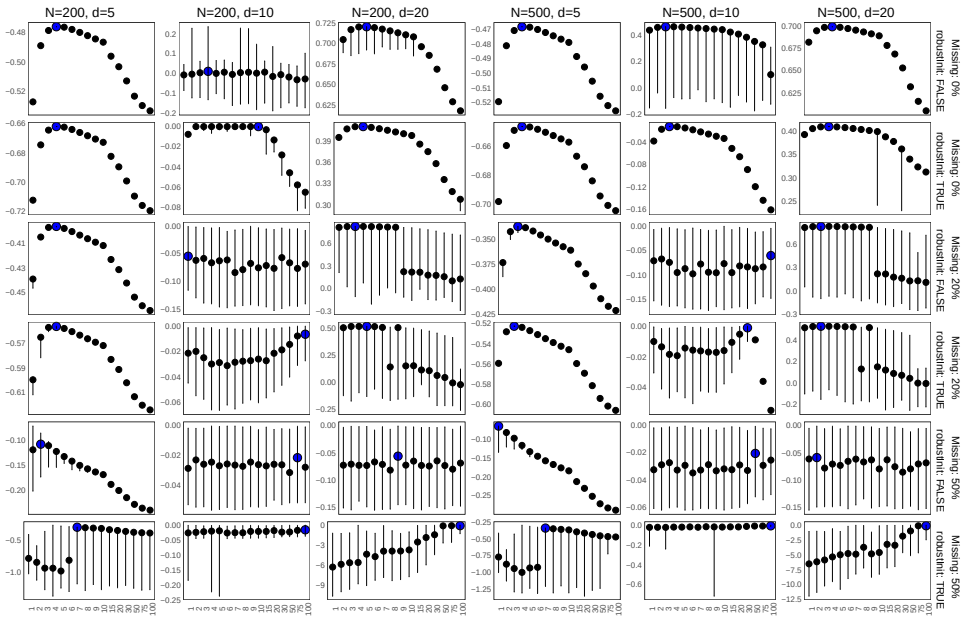


Figure 3.4. The median of the loglikelihood function of the EM algorithm for t-Student IND PPCA across the grid of degrees of freedom (x-axis) over different dimensionality of the observations (d) and the sample size (N) (column-wise) and the proportion of the missing values in the sample set (row-wise). The vertical solid black lines highlight the 2.5% and 97.5% empirical quantiles of the results. The blue colour of the points corresponds to the degree of freedom parameter which is optimal in terms of MLE model selection criterion. The true value of the parameter is equal to 4.

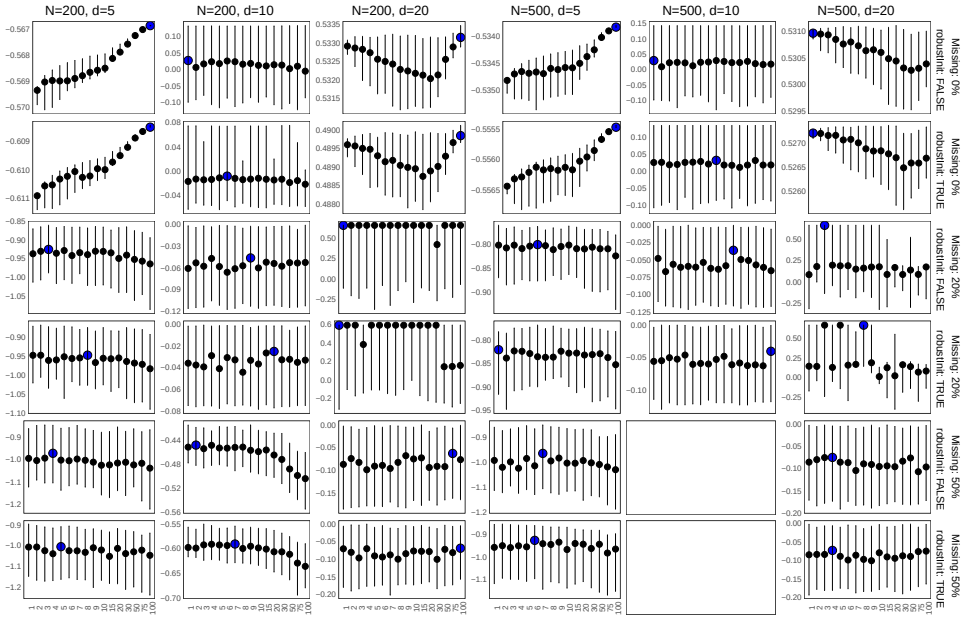


Figure 3.5. The median of the loglikelihood function of the EM algorithm for t-Student IID PPCA across the grid of degrees of freedom (x-axis) over different dimensionality of the observations (d) and the sample size (N) (column-wise) and the proportion of the missing values in the sample set (row-wise). The vertical solid black lines highlight the 2.5% and 97.5% empirical quantiles of the results. The blue colour of the points corresponds to the degree of freedom parameter which is optimal in terms of MLE model selection criterion. The true value of the parameter is equal to 4.

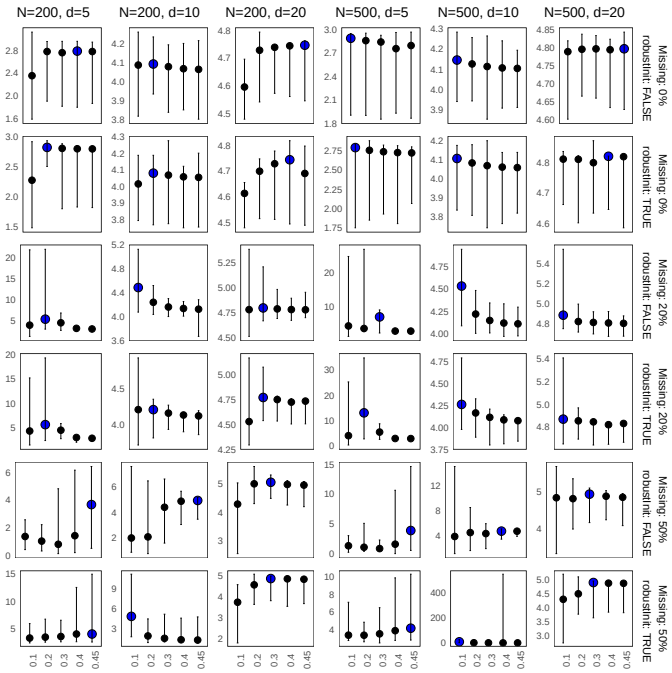


Figure 3.6. The median logarithm of the likelihood function of EM algorithm for robust Gaussian PPCA across the grid of degrees of freedom (x-axis) over different dimensionality of the observation (d) and sample size (N) and (row-wise) proportion of missing values in sample set (*missing*). The blue colour of bars corresponds to the degree of freedom parameter per case which results in the maximal median value of the corresponding loglikelihood. The bars with red borders correspond to the results fo the algorithms which have robust initialisation. The vertical solid black lines highlights the 2.5% and 97.5% empirical quantiles of the results.

We want to verify what is the impact of different assumption on heavy tails of a sample data on the loglikelihood of the t-Student PPCA and robust Gaussian PPCA EM algorithms. Figure 3.4 and Figure 3.5 present the median logarithm of the likelihood function per dimensionality, sample size and proportion of missingness across different values of degrees of freedom for t-Student IND and t-Student IID PPCA algorithms whereas Figure 3.6 illustrates the results for robust Gaussian PPCA across various breakdown points. The maximum median of logarithms is highlighted by blue colour. The row-wise order of the panels corresponds to the different proportions of missing values and the type of the initialisation step

One may notice that the algorithms remind sensitive to the initialisation step regardless of the parameters of degrees of freedom or a breakdown point, especially with the increase of the dimensionality and the ratio of missing value. In addition, the t-Student PPCA algorithms rarely provide the highest values of the loglikelihood functions for their true parameter of degrees of freedom. The corresponding likelihood functions are very flat across considered values of degrees of freedom and therefore the estimation of the parameters is sensitive to numerical error. Consequently, we recommend against the estimation of these parameters via Maximum Loglikelihood Estimation in EM algorithm as it is proceeded with the rest of the linear static parameters of the model (1). Therefore, we use the parameter of degrees of freedom as a model selection criterion and specify its value on a grid.

With regards to the determining the value of the breakdown points in robust Gaussian PPCA, the choice is related to model selection methodology as it is not a component of the model (1)

or assumptions on Gaussian PPCA. Figure 3.6 shows that the optimal breakdown points are not significantly affected by the initialisation methodology. However, we observe that the selection of the values is affected by the dimensionality of the data and the proportion of missing values.

The breakdown point denotes the allowed maximal proportion of a sample corrupted with outlying data which can be handled before by a covariance estimator before giving an incorrect result. The plots in 3.6 further confirm that the robust Gaussian PPCA methodology is more sensitive to initialisation step with an increase of missing values, especially for lower breakdown points. The profile loglikelihood of the corresponding EM algorithms is not flat and provides us with clear information which value of the parameter to select, however, the interquartile range of the results tends to be wide especially for the higher proportion of missing values and small dimensionality.

3.3.2. Testing Robustness of PPCA Algorithms For the Perturbed Sample

In order to investigate the efficiency of the PPCA frameworks in a presence of perturbation to the sample data, we generate the observation data which follows multivariate Gaussian distribution and then corrupt some portion of the sample corresponding to the parameter *perturbed* with the realisation of zero mean multivariate Laplace random vector with covariance matrix $\mathbf{C}$ and the dimensionality which is consistent with the observation vector.

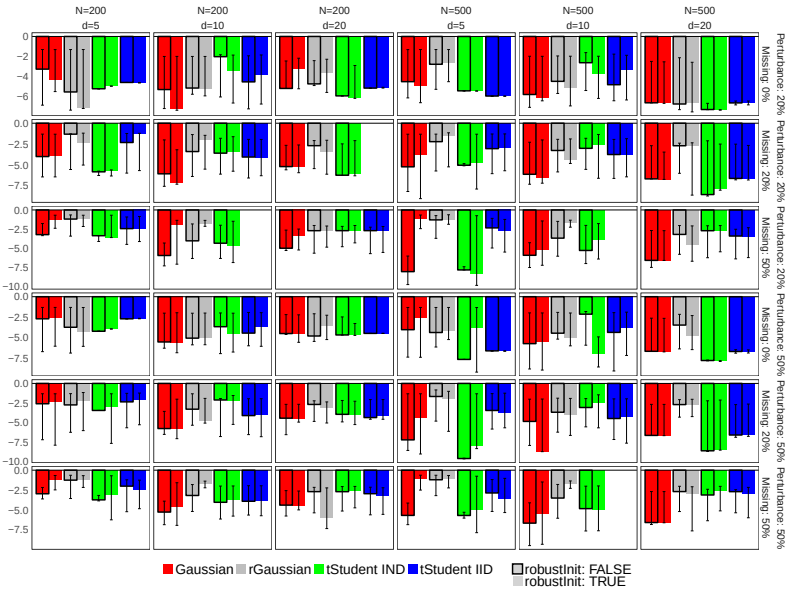
We perturb the observation set in two ways: (1) row-wise when we draw the whole observations and perturb it with the corresponding realisation of the Laplace random vector; or (2) we draw single elements from the sample set which we perturb with corresponding elements from a realisation of the Laplace random vector. The studies aim to identify which methodology is the most robust in the presence of two different perturbation patterns. We examine the methodology under different input parameters: length of the sample, the dimensionality of an observation, the proportion of missing values and the proportion of the corrupted sample.

The studies show that in the presence of the element-wise perturbation, the best methodology which estimates the mean vector $\boldsymbol{\mu}$ is t-Student IND regardless of the input parameters. The panels in Figure 3.7 show the accuracy of the estimations of the three most meaningful eigenvectors of the covariance matrix $\mathbf{C}$ for two patterns of the perturbation: row-wise in the top panel and element-wise in the bottom panel. They illustrate the logarithm of the median of discrepancies between the estimation and the true value of the parameter across different cases and methodologies. We use the squared Frobenius norm of the difference between matrices as a measure of the distance. We remark that the Gaussian PPCA is comparably accurate in the estimation of the matrix with eigenvectors that the robust frameworks when the dimensionality of the data is higher. The robust Gaussian PPCA provides the poorest recovery. The t-Student IND PPCA obtains the best results for the lower dimensionality of the data when the sample is perturbed element-wise or for higher dimensionality when the sample is perturbed row-wise. Also, the high proportion of missing values decrease the performance of the algorithms. With regards to the t-Student IID PPCA, the methodology is most efficient in high dimensional and obtains the best performance among examined frameworks when the data is corrupted element-wise.

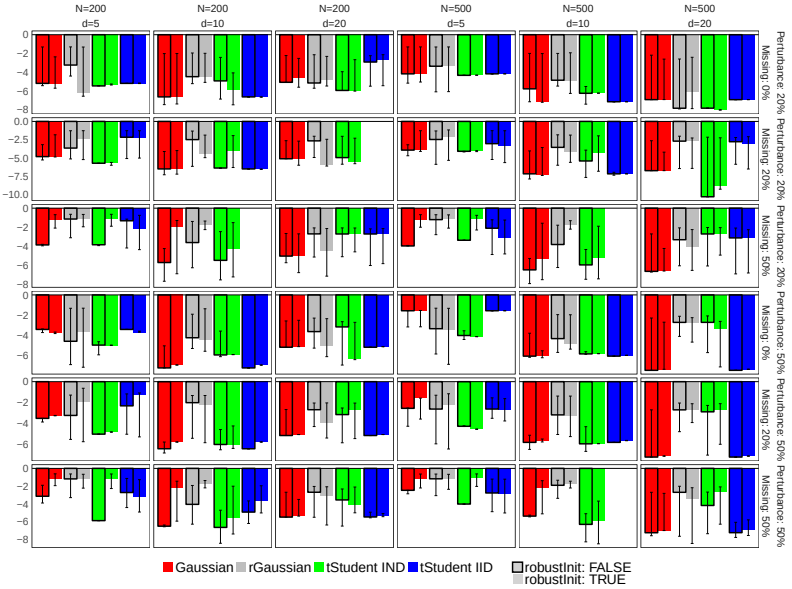
The estimation of the vector of means, $\boldsymbol{\mu}$ is more consistent across the methodologies and perturbations patterns. The panels in Figure display the accuracy of the estimations for row-wise and element-wise perturbation of the sample set, in the top and bottom panel, respectively. We notice, that similarly to the non-perturbed case, the t-Student IND PPCA tends to provide the poor estimation of the parameter when the dimensionality of the data and the proportion of missing values are high. In addition, we observe that in a presence of the row-wise perturbation and high dimensionality of data, the t-Student IID PPCA is the most accurate methodology.

The selected breakdown points for robust Gaussian PPCA are illustrated in Figure 3.9 for row-wise and element-wise pattern of perturbation in left and right panel, respectively. The proportion of perturbed sample is indicated in the x-axis of the plots. The blue color of bars corresponds to the selected parameters when the algorithm is initialized in a robust way. The optimal values are mostly consistent within the perturbation patterns. We notice that the discrepancies between the values of parameters provided by the robust or non-robust initialisation increase when the proportion of missing values is higher and dimensionality of data is low, especially in the element-wise case. What is more, we observe that the increase of the selected breakdown points follows the increase of the missingness. It is not a surprising outcome since the missing values are projected using the conditional means. The robust covariance estimation algorithms detects more variability and outgoingness in the smoothed data set with filled missing entries and hence requires higher breakdown point which specifies the percentage of sample which is drawn to calculate the sample estimators of a covariance matrix.

With regards to the other robust PPCA methodologies, Figure 3.10 and Figure 3.11 shows corresponding optimal degrees of freedom selected for t-Student IND PPCA and t-Student IID PPCA, respectively. The selected values of the parameter are less consistent across different patterns of perturbation as well as dimensionality of the data, the sample size and proportion of missing values than the selected breakdown points. It is a consequence of the flat profile of loglikelihood of the corresponding EM algorithms and therefore their numerical sensitivity. There is small consistency of the results with the proportion of the perturbation, especially for higher dimensionality of the data. However, we remark that t-Student IID PPCA selects higher values of degrees of freedom for element-wise perturbation when the proportion of missing values is low whereas lower degrees of freedom when the missingness is the highest. The selected of degrees of freedom are in more agreement for the row-wise perturbation pattern.

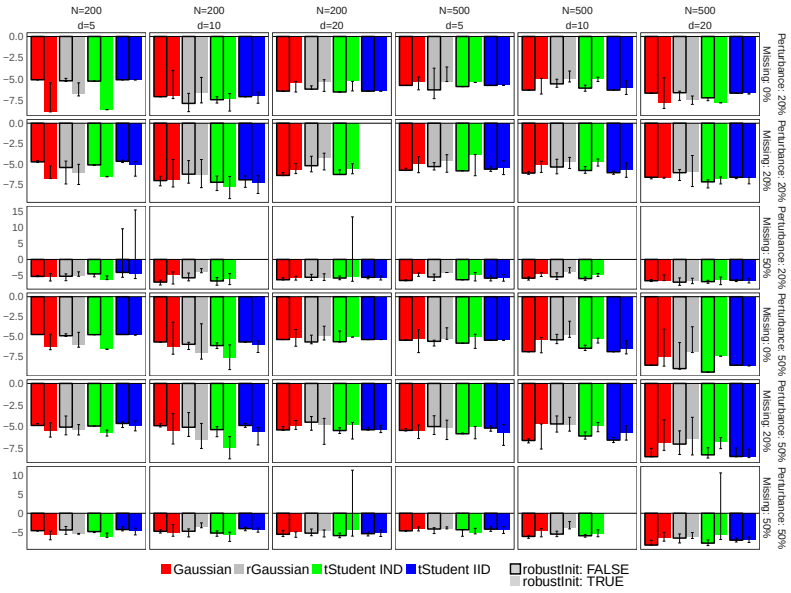


(a) row-wise

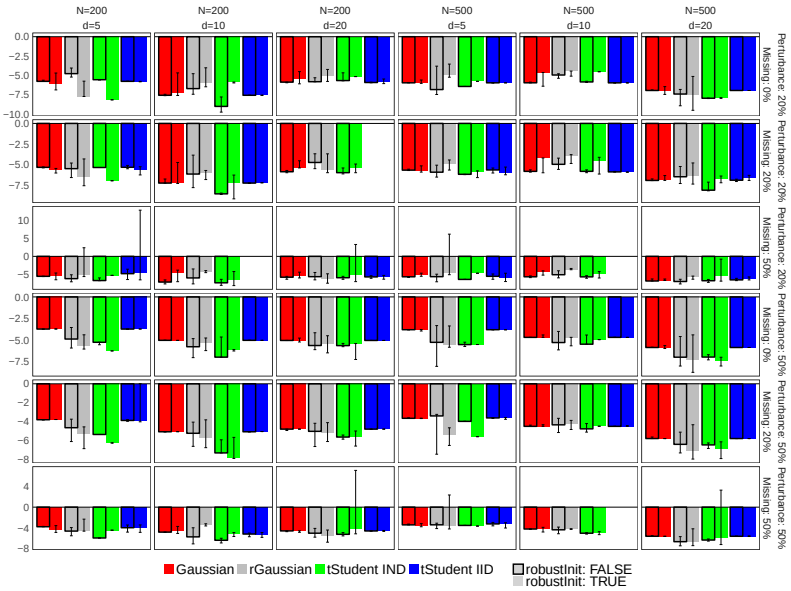


(b) element-wise

Figure 3.7. The logarithm of the squared Forbenious norm of the difference between the three first eigen vectors of the true matrix $\mathbf{C} = \mathbf{W}^T\mathbf{W} + \sigma^2\mathbf{I}_d$ and its estimate $\hat{\mathbf{C}} = \hat{\mathbf{W}}^T\hat{\mathbf{W}} + \hat{\sigma}^2\mathbf{I}_d$ given by the following EM algorithms: Gaussian PPCA (red), Robust Gaussian PPCA (grey), t-Student IND PPCA (green) and t-Student IID PPCA (blue) for (column-wise) different dimensionality of a single observation (d) and the sample size (N) and (row-wise) the proportion of the sample which is perturbed with the Laplace noise and the proportion of the missing values in the sample set. The bars with solid black borders correspond to the resuts fo the algorithms which have robust initialisation. The vertical solid black lines highlights the 2.5% and 97.5% quantiles of the results. The top and bottom panels correspond to **the row-wise and element-wise perturbation patterns**, respectively.

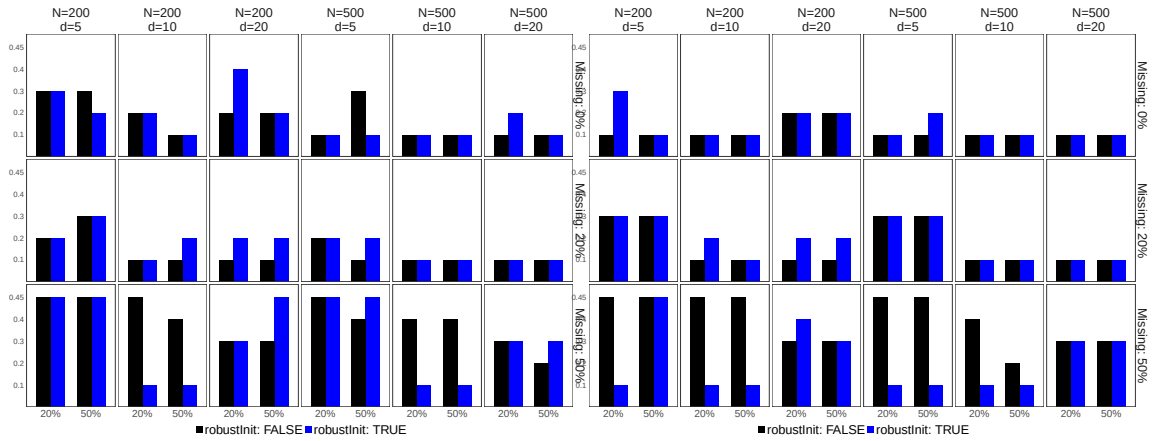


(a) row-wise



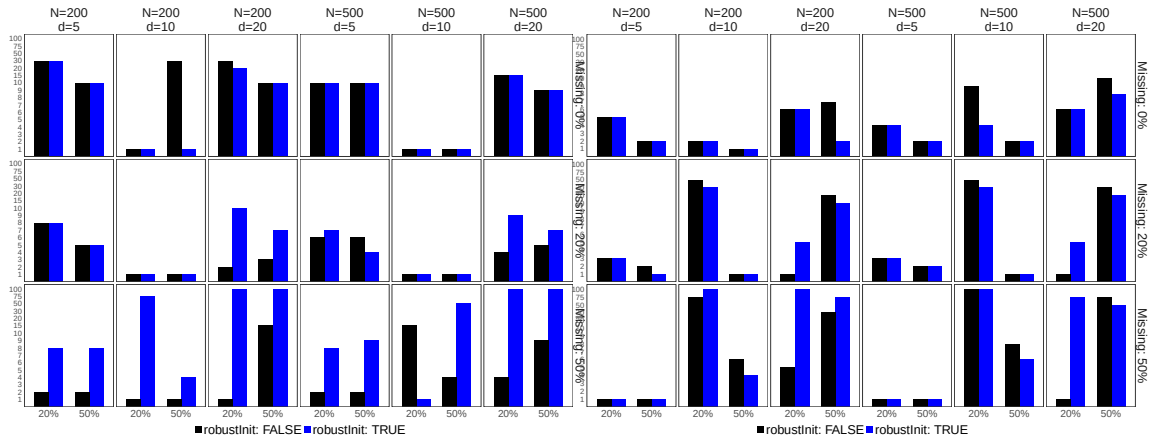
(b) element-wise

Figure 3.8. The logarithm of the squared Forbenious norm of the difference between the true vector of means μ and its estimator given by the following EM algorithms: Gaussian PPCA (red), Robust Gaussian PPCA (grey), t-Student IND PPCA (green) and t-Student IID PPCA (blue) for (column-wise) different dimensionality of a single observation (d) and the sample size (N) and (row-wise) the proportion of the sample which is perturbed with the Laplace noise and the proportion of the missing values in the sample set. The bars with solid black borders correspond to the results of the algorithms which have robust initialisation. The vertical solid black lines highlight the 2.5% and 97.5% quantiles of the results. The top and bottom panels correspond to **the row-wise** and **element-wise perturbation patterns**, respectively.



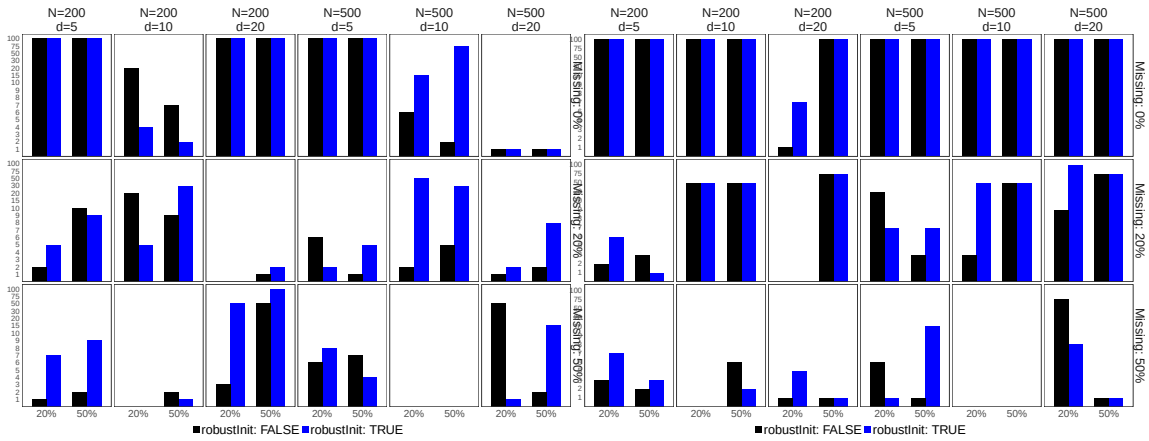
(a) row-wise perturbation **(b)** element-wise perturbation

Figure 3.9. The optimal breakdown points for **robust Gaussian PPCA** (y-axis) across different proportions of perturbed samples with the Laplace noise (x-axis) over different dimensionality of the observation (d) and sample size (N) and (row-wise) the proportion of the missing values in the sample set. The blue color of bars corresponds to the algorithms with the robust initialisation. The left and right panels correspond to the results for row-wise and element-wise perturbations, respectively.



(a) row-wise perturbation **(b)** element-wise perturbation

Figure 3.10. The optimal degrees of freedom for **t-Student IND PPCA** (y-axis) across different proportions of perturbed samples with the Laplace noise (x-axis) over different dimensionality of the observation (d) and sample size (N) and (row-wise) the proportion of the missing values in the sample set. The blue color of bars corresponds to the algorithms with the robust initialisation. The left and right panels correspond to the results for row-wise and element-wise perturbations, respectively



(a) row-wise perturbation **(b)** element-wise perturbation

Figure 3.11. The optimal degrees of freedom for **t-Student IID PPCA** (y-axis) across different proportions of perturbed samples with the Laplace noise (x-axis) over different dimensionality of the observation (d) and sample size (N) and (row-wise) the proportion of the missing values in the sample set. The blue color of bars corresponds to the algorithms with the robust initialisation. The left and right panels correspond to the results for row-wise and element-wise perturbations, respectively.

3.4. Extension 2: Generalized Skew t-Student Probabilistic Principal Component Analysis

In this section we define the Generalized Skew t-Student PPCA model. Following the concept of Grouped t-copula discussed in (Daul et al. 2003) and (Luo and Shevchenko 2012), we introduce a Probabilistic Principal Component model which allows one to develop more flexible tail dependence structures than the standard t-Student models. This extension is important as the features obtained from the novel baseline t-Student PPCA models presented previously may be misspecified for the given data generating process. In such cases, even if robust estimation is performed, it will not ensure that the appropriate feature summaries are obtained by the PPCA feature extraction summaries. Therefore, to allow a greater degree of flexibility, especially when asymmetry is present in tail dependence between pairs or sub-sets of the multivariate random observation vectors, one can develop novel extensions based on grouped and generalised t-Student PPCA models. This flexible framework will be performed in the following section.

Let us recall the baseline PPCA model in Equation (2.7). Given the stochastic representation of the t-Student random variables, as a scale mixture of a Gaussian random variable and Gamma variable, as formulated in (Gupta and Nagar 1999) or (Kotz and Nadarajah 2004), we assume that $\mathbf{X}_t$ and $\boldsymbol{\epsilon}_t$ are represented by independent Gamma random variables. Consequently, the vectors themselves are mutually independent and have individual dependency structures. The assumption provides the model with the additional flexibility to determine which component of the model impacts on tails behaviour of the observation vector. As an additional extension to the existing literature on robust dimensionality reduction, we introduce the coefficient of skewness which specifies the strength of asymmetry in the distribution of the unobserved random vectors using the definition of the Skew t distribution of (Azzalini and Capitanio 2003) or (Gupta 2003).

Lastly, we want to highlight that under appropriate assumptions on the deterministic parameters of the introduced models, the generalized PPCA framework reduces to various special cases such as the PPCA model under a Grouped t-Student distribution (when skewness is equal to zero), or the PPCA model under a Skew t-Student distribution (when degrees of freedom are equal per marginal). The investigated family of models will be also shown to include an alternative variant of the t-Student PPCA outlined in Section 3.2 where $\mathbf{X}_t$ and $\boldsymbol{\epsilon}_t$ are mutual independent. These special cases of Generalized Skew t-Student PPCA and their estimation frameworks are our additional novelty and further contribution to the existing literature.

Therefore, unlike previous robust approaches to PPCA and PCA we use a probabilistic framework, which has significant advantages:

- **(1):** it adds more flexibility to a model for a considered data, that is, it detects the characteristics such as skewness our various tails structures;
- **(2):** the framework handles presence of missing data and can be adjusted to various assumptions about the patterns of missigness;
- **(3):** we derive an efficient Expectation-Maximisation (EM) algorithm for parameter estimation which within the class of elliptical families ensures convergence to local maximum.

We again consider two cases of the distributions: the first in which the random vectors $\mathbf{X}_t$ and $\boldsymbol{\epsilon}_t$ are independently and non-identically distributed, and the second in which they are identically and conditionally independently distributed over time – analogous to the set up already introduced for the t-Student PPCA model in Section 3.2. The two groups of models are discussed in Subsubsection 3.4.1.1 and Subsubsection 3.4.1.2, respectively.

3.4.1. Probabilistic Principal Component Analysis with Generalized Skew t-Student Distribution

We define two classes of models for Generalized Skew t-Student PPCA, one which assumes that the latent variables follows the Independent Generalized Skew t-Student Distribution (Generalized Skew t-Student IND PPCA), and the second one denoted by Identical and Conditionally Independent Generalized Skew t-Student Distribution (Generalized Skew t-Student IID PPCA).

3.4.1.1. The model of Independent Generalized Skew t-Student Probabilistic Principal Component

Let us denote two mutually independent and identically distributed over time uniform random variables $S_{\epsilon,t}, S_{x,t} \sim \mathcal{U}(0,1)$. For the convince of the notation, we determinate d- and k-dimensional random vectors $\mathbf{U}_t$ and $\mathbf{V}_t$, respectively,

$$\begin{aligned} \mathbf{U}_t &= \left(\frac{\chi^{-1}_{\nu^1_\epsilon}(S_{\epsilon,t})}{\nu^1_\epsilon}, \ldots, \frac{\chi^{-1}_{\nu^d_\epsilon}(S_{\epsilon,t})}{\nu^d_\epsilon} \right)_{1\times d} \\ \mathbf{V}_t &= \left(\frac{\chi^{-1}_{\nu^1_x}(S_{x,t})}{\nu^1_x}, \ldots, \frac{\chi^{-1}_{\nu^k_x}(S_{x,t})}{\nu^k_x} \right)_{1\times k} \end{aligned} \tag{3.27}$$

for vectors of non-negative real numbers $\boldsymbol{\nu}_\epsilon = \{\nu^1_\epsilon, \ldots, \nu^d_\epsilon\}$ and $\boldsymbol{\nu}_x = \{\nu^1_x, \ldots, \nu^k_x\}$ and χ^{-1}_ν denoting the quantile function of the chi-square distribution with ν degrees of freedom. Notice, that each of the vectors $\mathbf{U}_t$ and $\mathbf{V}_t$ follows a multivariate Gamma distribution and are mutually independent and independent in time. However, the components of the vectors are perfectly dependent. That is, hey are constructed to be co-monotonic as transformations of the same uniform variable in time t.

Let us denote d-dimensional and k-dimensional real valued model parameter vectors, $\boldsymbol{\delta}_\epsilon$ and $\boldsymbol{\delta}_x$. Let $\mathbf{Z}_{x,t}$ and $\mathbf{Z}_{\epsilon,t}$ be mutually independent standard normal multivariate variables, k- and d-dimensional respectively which are independent of $\mathbf{U}_t$ and $\mathbf{V}_t$ (or $S_{x,t}$ and $S_{\epsilon,t}$ likewise). Under the **Independent Generalized Skew t-Student Distribution**, the stochastic representation of the d-dimensional error term $\boldsymbol{\epsilon}_t$ and the k-dimensional latent variable $\mathbf{X}_t$ is given by

$$\begin{aligned} \mathbf{X}_t &= \mathbf{V}_t^{-1} \circ \boldsymbol{\delta}_x + \sqrt{\mathbf{V}_t^{-1}} \circ \mathbf{Z}_{x,t}, \\ \boldsymbol{\epsilon}_t &= \mathbf{U}_t^{-1} \circ \boldsymbol{\delta}_\epsilon + \sqrt{\sigma^2 \mathbf{U}_t^{-1}} \circ \mathbf{Z}_{\epsilon,t}. \end{aligned} \tag{3.28}$$

The operator $\circ$ denotes the Hadamard product, that is, for two d-dimensional vectors $\mathbf{a}$ and $\mathbf{b}$, the product of the two vectors results in the d-dimensional vector $\mathbf{a} \circ \mathbf{b} = \left(a^1 b^1, \ldots, a^d b^d\right)$.

Consequently, we have the following joint probability density function of the Generalized Skew t-Student IND PPCA model given N realisations of the random vectors in time $t = 1, \ldots, N$, that is

$$\begin{aligned}
&\pi_{\mathbf{Y}_{1:N},\mathbf{X}_{1:N},\mathbf{U}_{1:N},\mathbf{V}_{1:N}|\Psi}(\mathbf{y}_{1:N}, \mathbf{x}_{1:N}, \mathbf{u}_{1:N}, \mathbf{v}_{1:N}) \\
&= \pi_{\mathbf{Y}_{1:N}|\mathbf{X}_{1:N},\mathbf{U}_{1:N},\mathbf{V}_{1:N},\Psi}(\mathbf{y}_{1:N})\pi_{\mathbf{X}_{1:N}|\mathbf{U}_{1:N},\mathbf{V}_{1:N},\Psi}(\mathbf{x}_{1:N})\pi_{\mathbf{U}_{1:N}|\mathbf{V}_{1:N},\Psi}(\mathbf{u}_{1:N})\pi_{\mathbf{V}_{1:N}|\Psi}(\mathbf{v}_{1:N}) \\
&= \prod_{t=1}^{N}\left\{\pi_{\mathbf{Y}_t|\mathbf{X}_t,\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{y}_t)\cdot\pi_{\mathbf{X}_t|\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{x}_t)\cdot\pi_{\mathbf{U}_t|\Psi}(\mathbf{u}_t)\cdot\pi_{\mathbf{V}_t|\Psi}(\mathbf{v}_t)\right\},
\end{aligned} \tag{3.29}$$

for $\Psi = \left[\mathbf{W},\ \boldsymbol{\mu},\ \sigma^2,\ \delta_\epsilon,\ \delta_x,\ \nu_x,\ \nu_\epsilon\right]$ being a vector which consists of all static parameters in the model specified in Equation (2.7) under Generalized Skew t-Student IND PPCA assumptions. Recall that the distributions of $\mathbf{Y}_t$, $\mathbf{X}_t$ and $\boldsymbol{\epsilon}_t$ are conditionally multivariate Gaussian, such that

$$\begin{aligned}
\mathbf{Y}_t|\mathbf{X}_t,\mathbf{U}_t,\mathbf{V}_t,\Psi &\sim \mathcal{N}\left(\boldsymbol{\mu}+\delta_\epsilon\mathbf{D}_{\epsilon,t}^{-1}+\mathbf{X}_t\mathbf{W}^T,\sigma^2\mathbf{D}_{\epsilon,t}^{-1}\right), \\
\mathbf{X}_t|\mathbf{V}_t,\Psi &\sim \mathcal{N}\left(\delta_x\mathbf{D}_{x,t}^{-1},\mathbf{D}_{x,t}^{-1}\right), \\
\boldsymbol{\epsilon}_t|\mathbf{U}_t,\Psi &\sim \mathcal{N}\left(\delta_\epsilon\mathbf{D}_{\epsilon,t}^{-1},\sigma^2\mathbf{D}_{\epsilon,t}^{-1}\right),
\end{aligned}$$

where

$$\mathbf{D}_{\epsilon,t} = \begin{pmatrix} U_t^1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & U_t^d \end{pmatrix}_{d\times d}, \ \mathbf{D}_{x,t} = \begin{pmatrix} V_t^1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & V_t^k \end{pmatrix}_{k\times k}.$$

3.4.1.2. The model of Identical and Conditionally Independent Generalized Skew t Student Probabilistic Principal Component Analysis

The PPCA model defined in Subsection 3.4.1.1 assumes that the the realisation of the univariate uniform variables are independent in time. Consequently, the random vectors $\mathbf{X}_t$ and $\boldsymbol{\epsilon}_t$ are also independent in time and have a time-specific distributions. Therefore, each realisation of latent vectors is, in fact, a realisation of a variable with different probability function. Given the sample set which contains N observations of $\mathbf{Y}_t$, it represents one realisation of N independent in time random variables. In the following subsection we show how to relax this assumption and obtain the methodology with less degrees of freedom. We decrease the model's uncertainty by limiting the number of model's components which are free to vary. It is achieved by restraining the number of realisations of the two uniform variables to one per sample set. The random variables $S_\epsilon, S_x \sim \mathcal{U}(0,1)$ specify d and k dimensional random vectors $\mathbf{U}$ and $\mathbf{V}$ as follows

$$\begin{aligned}
\mathbf{U} &= \left(\frac{\chi^{-1}_{\nu_\epsilon^1}(S_\epsilon)}{\nu_\epsilon^1},\ldots,\frac{\chi^{-1}_{\nu_\epsilon^d}(S_\epsilon)}{\nu_\epsilon^d}\right)_{1\times d}, \\
\mathbf{V} &= \left(\frac{\chi^{-1}_{\nu_x^1}(S_x)}{\nu_x^1},\ldots,\frac{\chi^{-1}_{\nu_x^k}(S_x)}{\nu_x^k}\right)_{1\times k}.
\end{aligned} \tag{3.30}$$

Given this notation, the stochastic representation of the error term $\boldsymbol{\epsilon}_t$ and the latent vector $\mathbf{X}_t$ under **Identical and Conditionally Independent Generalized Skew t-Student Distribution** is formulated as follows

$$\begin{aligned}\mathbf{X}_t &= \mathbf{V}^{-1} \circ \boldsymbol{\delta}_x + \sqrt{\mathbf{V}^{-1}} \circ \mathbf{Z}_{x,t},\\ \boldsymbol{\epsilon}_t &= \mathbf{U}^{-1} \circ \boldsymbol{\delta}_\epsilon + \sqrt{\sigma^2\mathbf{U}^{-1}} \circ \mathbf{Z}_{\epsilon,t}.\end{aligned} \tag{3.31}$$

The definitions of these latent variables in Equation (3.31) are key differences between the Generalized Skew IND t-Student PPCA model and the model from this subsubsection, which we denote by Generalized Skew t-Student IID PPCA model (Generalized Skew t-Student IID PPCA). Recall that the random vectors $\boldsymbol{\epsilon}_t$ and $\mathbf{X}_t$ are still mutually independent but the new model does not impose their independence in time. Their distribution conditioned on random vectors $\mathbf{U}$ and $\mathbf{V}$ is time independent since $\mathbf{Z}_{\epsilon,t}$ and $\mathbf{Z}_{x,t}$ are independent in time. Instead, the new PPCA model imposes the identical distribution of their realisation in time as we use only one realisation of S_ϵ and S_x per sample set. Since the observation vector $\mathbf{Y}_t$ follows the model from Equation (2.7), its realisations are also only conditionally independent in time. Applying the Chain Rule of probabilities, the joint probability function of the all random components of the Identical and Conditionally Independent Generalized Skew t-Student model for N observations is given by

$$\begin{aligned}&\pi_{\mathbf{Y}_{1:N},\mathbf{X}_{1:N},\mathbf{U},\mathbf{V}|\Psi}(\mathbf{y}_{1:N},\mathbf{x}_{1:N},\mathbf{u},\mathbf{v}) = \pi_{\mathbf{Y}_{1:N}|\mathbf{X}_{1:N},\mathbf{U},\mathbf{V},\Psi}(\mathbf{y}_{1:N})\pi_{\mathbf{X}_{1:N}|\mathbf{U},\mathbf{V},\Psi}(\mathbf{x}_{1:N})\pi_{\mathbf{U}|\mathbf{V},\Psi}(\mathbf{u})\pi_{\mathbf{V}|\Psi}(\mathbf{v})\\ &= \pi_{\mathbf{U}|\Psi}(\mathbf{u})\ \pi_{\mathbf{V}|\Psi}(\mathbf{v}) \prod_{t=1}^{N} \left\{ \pi_{\mathbf{Y}_t|\mathbf{X}_t,\mathbf{U},\mathbf{V},\Psi}(\mathbf{y}_t) \cdot \pi_{\mathbf{X}_t|\mathbf{U},\mathbf{V},\Psi}(\mathbf{x}_t) \right\}.\end{aligned} \tag{3.32}$$

with corresponding conditional distributions which are given as follows

$$\begin{aligned}\mathbf{Y}_t|\mathbf{X}_t,\mathbf{U},\mathbf{V},\Psi &\sim \mathcal{N}\left(\boldsymbol{\mu} + \boldsymbol{\delta}_\epsilon\mathbf{D}_{\epsilon,t}^{-1} + \mathbf{X}_t\mathbf{W}^T, \sigma^2\mathbf{D}_\epsilon^{-1}\right),\\ \mathbf{X}_t|\mathbf{V},\Psi &\sim \mathcal{N}\left(\boldsymbol{\delta}_x\mathbf{D}_x^{-1}, \mathbf{D}_x^{-1}\right),\\ \boldsymbol{\epsilon}_t|\mathbf{U},\Psi &\sim \mathcal{N}\left(\boldsymbol{\delta}_\epsilon\mathbf{D}_\epsilon^{-1}, \sigma^2\mathbf{D}_\epsilon^{-1}\right),\end{aligned}$$

where

$$\mathbf{D}_\epsilon = \begin{pmatrix} U^1 & 0 & 0\\ 0 & \ddots & 0\\ 0 & 0 & U^d \end{pmatrix}_{d\times d}, \quad \mathbf{D}_x = \begin{pmatrix} V^1 & 0 & 0\\ 0 & \ddots & 0\\ 0 & 0 & V^k \end{pmatrix}_{k\times k}.$$

3.4.1.3. Special Cases: Probabilistic Principal Component Analysis with Skew t-Student Distribution and Grouped t-Student Distribution

Under appropriate assumptions on the deterministic parameters of the model, the Generalized Skew t-Student PPCA reduces to simpler special cases. We show how the special cases are defined using independent PPCA model assumptions from Subsection 3.4.2.1 as the identical

and conditionally distributed case proceeds analogously. We can straightforwardly obtain the following three models

Special Case 1: $\delta_\epsilon = \delta_x = \mathbf{0}$.

Given the skewness coefficients equal to zero, the generalized PPCA model reduces to **Grouped t-Student PPCA** case. The case-specific stochastic representation of the latent processes in the following

$$\begin{aligned} \mathbf{X}_t &= \sqrt{\mathbf{V}_t^{-1}} \circ \mathbf{Z}_{x,t}, \\ \boldsymbol{\epsilon}_t &= \sqrt{\sigma^2 \mathbf{U}_t^{-1}} \circ \mathbf{Z}_{\epsilon,t} \end{aligned} \tag{3.33}$$

which results in the conditional distributions given by

$$\begin{aligned} \mathbf{Y}_t|\mathbf{X}_t, \mathbf{U}_t, \mathbf{V}_t, \Psi &\sim \mathcal{N}\left(\boldsymbol{\mu} + \mathbf{X}_t\mathbf{W}^T, \sigma^2\mathbf{D}_{\epsilon,t}^{-1}\right), \\ \mathbf{X}_t|\mathbf{V}_t, \Psi &\sim \mathcal{N}\left(\mathbf{0}, \mathbf{D}_{x,t}^{-1}\right), \\ \boldsymbol{\epsilon}_t|\mathbf{U}_t, \Psi &\sim \mathcal{N}\left(\mathbf{0}, \sigma^2\mathbf{D}_{\epsilon,t}^{-1}\right). \end{aligned} \tag{3.34}$$

Special Case 2: $\nu_\epsilon^1 = \ldots = \nu_\epsilon^d = \nu_\epsilon$ and $\nu_x^1 = \ldots = \nu_x^k = \nu_x$.

We assume that the marginal distributions of $\mathbf{X}_t$ and $\boldsymbol{\epsilon}_t$ are characterized by the same tail behaviour. Consequently, the scaling variable $\mathbf{U}_t$ and $\mathbf{V}_t$ become one dimensional since and are identical across all components of the random vectors, $U_t = \frac{\chi_{\nu_\epsilon}^{-1}(S_{\epsilon,t})}{\nu_\epsilon}$ and $V_t = \frac{\chi_{\nu_x}^{-1}(S_{x,t})}{\nu_x}$. Therefore, the generalized PPCA model reduces to the **Skew t-Student PPCA** case, that is, for the independent case from Subsection 3.4.2.1, the representation of the latent processes then takes the following form

$$\begin{aligned} \mathbf{X}_t &= V_t^{-1}\boldsymbol{\delta}_x + \sqrt{V_t^{-1}}\mathbf{Z}_{x,t}, \\ \boldsymbol{\epsilon}_t &= U_t^{-1}\boldsymbol{\delta}_\epsilon + \sqrt{\sigma^2 U_t^{-1}}\mathbf{Z}_{\epsilon,t}, \end{aligned} \tag{3.35}$$

which results in the conditional distributions being

$$\begin{aligned} \mathbf{Y}_t|\mathbf{X}_t, U_t, V_t, \Psi &\sim \mathcal{N}\left(\boldsymbol{\mu} + U_t^{-1}\boldsymbol{\delta}_\epsilon + \mathbf{X}_t\mathbf{W}^T, \sigma^2 U_t^{-1}\mathbb{I}_d\right), \\ \mathbf{X}_t|V_t, \Psi &\sim \mathcal{N}\left(V_t^{-1}\boldsymbol{\delta}_x, V_t^{-1}\mathbb{I}_k\right), \\ \boldsymbol{\epsilon}_t|U_t, \Psi &\sim \mathcal{N}\left(U_t^{-1}\boldsymbol{\delta}_\epsilon, \sigma^2 U_t^{-1}\right). \end{aligned} \tag{3.36}$$

Special Case 3: $\delta_\epsilon = \delta_x = \mathbf{0}$, $\nu_\epsilon^1 = \ldots = \nu_\epsilon^d = \nu_\epsilon$ and $\nu_x^1 = \ldots = \nu_x^k = \nu_x$.

We assume that the marginal distributions of $\mathbf{X}_t$ and $\boldsymbol{\epsilon}_t$ are characterized by the same tail behaviour and the skewness coefficients are equal to zero. Then, the Generalized PPCA model reduces to the **t-Student PPCA** case which separates the tail effect of the error term $\boldsymbol{\epsilon}_t$ and

the latent process $\mathbf{X}_t$ on the random vector $\mathbf{Y}_t$. Given the independent case from Subsection 3.4.2.1, the representation of the latent processes in the following

$$\begin{aligned}\mathbf{X}_t &= \sqrt{V_t^{-1}}\mathbf{Z}_{x,t},\\ \boldsymbol{\epsilon}_t &= \sqrt{\sigma^2 U_t^{-1}}\mathbf{Z}_{\epsilon,t}\end{aligned} \tag{3.37}$$

which results in the conditional distributions being

$$\begin{aligned}\mathbf{Y}_t|\mathbf{X}_t, U_t, V_t, \Psi &\sim \mathcal{N}\left(\boldsymbol{\mu} + \mathbf{X}_t\mathbf{W}^T, \sigma^2 U_t^{-1}\mathbb{I}_d\right),\\ \mathbf{X}_t|V_t, \Psi &\sim \mathcal{N}\left(\mathbf{0}, V_t^{-1}\mathbb{I}_k\right),\\ \boldsymbol{\epsilon}_t|U_t, \Psi &\sim \mathcal{N}\left(\mathbf{0}, \sigma^2 U_t^{-1}\right).\end{aligned} \tag{3.38}$$

Recall that the above formulation is an additional variant to the t-Student PPCA model derived in (Toczydlowska and Peters 2018) and discussed in Section 3.2. The difference lies in the assumption that the random vectors $\boldsymbol{\epsilon}_t$ and $\mathbf{X}_t$ have individual scaling random variable, U_t and V_t, respectively. Consequently, they have a separate effect on the tail behaviour of the observation vector $\mathbf{Y}_t$.

3.4.2. Complete Data EM Algorithm for Generalized Skew t-Student Probabilistic Principal Component

The two iterative steps of the EM algorithm for Genrelized Skew t-Student PPCA which jointly maximize the expected log-likelihood of the observed and hidden variables, are given below using the formulation of the independent case

Step 1: Expectation Step (E-step)

We calculate the expectation of the the conditional distribution $\mathbf{X}_{1:N}, \mathbf{U}_{1:N}, \mathbf{V}_{1:N}|\mathbf{Y}_{1:N}, \Psi$ over the joint distribution likelihood function of the model (2.7). Given N realisations of the variables, the expectation is a function of two vectors with static parameters $\Psi = [\mathbf{W}, \boldsymbol{\mu}, \sigma^2, \delta_\epsilon, \delta_x]$ and $\Psi^* = [\mathbf{W}^*, \boldsymbol{\mu}^*, \sigma^{*2}, \delta_\epsilon^*, \delta_x^*]$, that is

$$Q_{GStS,ind}(\Psi, \Psi^*) = \mathbb{E}_{\mathbf{X}_{1:N},\mathbf{U}_{1:N},\mathbf{V}_{1:N}|\mathbf{Y}_{1:N},\Psi}\left[\log \pi_{\mathbf{Y}_{1:N},\mathbf{X}_{1:N},\mathbf{U}_{1:N},\mathbf{V}_{1:N}|\Psi^*}(\mathbf{Y}_{1:N}, \mathbf{X}_{1:N}, \mathbf{U}_{1:N}, \mathbf{V}_{1:N})\right]$$

Step 2: Maximisation Step (M-step)

We update the vector of static parameters Ψ via maximisation of the resulting $Q_{GStS,ind}$ function with respect to the vector Ψ^*, that is

$$\hat{\Psi}^* = \underset{\Psi^*}{\operatorname{argmax}}\, Q_{GStS,ind}(\Psi, \Psi^*)$$

The models introduced in Section 3.4.1 depend on the vectors with the non-negative parameters $\boldsymbol{\nu}_\epsilon$ and $\boldsymbol{\nu}_x$ which determine the tail behaviour of the random vectors and correspond to degrees of freedom. We show in Section 3.3 that the likelihood profile of the EM algorithms for t-Student

PPCA is very flat over the grid of degrees of freedom. Therefore, we recommended against using EM algorithm to specify v for this less complex model. The same argument applies to the family of Generalized t-Student PPCA models.

3.4.2.1. Expectation-Maximisation Algorithm for Probabilistic Principal Component Analysis with Independent Generalized Skew t-Student Distribution

The $Q_{GStS,ind}$ function of the E-step for the Generalized Skew t-Student IND PPCA is provided in Theorem 13 whereas the derivation of its maximizers is given in Theorem 14. Moreover, we derive the formulations of the E-step and the M-step of EM algorithms for special cases of the Generalized Skew t-Student IND PPCA specified in Subsubsection 3.4.1.3:

Special Case 1: Grouped t-Student IND PPCA in Corollary 1 and Corollary 5,
Special Case 2: Skew t-Student IND PPCA in Corollary 2 and Corollary 6,
Special Case 3: t-Student IND PPCA in Corollary 3 and Corollary 7.

The E-Step of The EM Algorithm

Theorem 13. *Let the observation process $\mathbf{Y}_t$ follows the Generalized Skew t-Student IND PPCA model defined in Subsubsection 3.4.1.1 in Equation (3.27) and Equation (3.28). Given N realisations of the observation vector $\mathbf{Y}_t$, denoted by $\mathbf{y}_{1:N} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$, the E-step of the corresponding EM algorithm is formulated as following*

$$\begin{aligned} Q_{GStS,ind}(\Psi, \Psi^*) &= \mathbb{E}_{\mathbf{X}_{1:N},\mathbf{U}_{1:N},\mathbf{V}_{1:N}|\mathbf{Y}_{1:N},\Psi}\left[\log \pi_{\mathbf{Y}_{1:N},\mathbf{X}_{1:N},\mathbf{U}_{1:N}\mathbf{Y}_{1:N}|\Psi^*}\left(\mathbf{Y}_{1:N}, \mathbf{X}_{1:N}, \mathbf{U}_{1:N}, \mathbf{V}_{1:N}\right)\right] \\ &= \frac{1}{\pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N})} \sum_{t=1}^{N} \left\{ I_1(\mathbf{y}_t; \Psi, \Psi^*) \prod_{s=1, s\neq t}^{N} I_2(\mathbf{y}_s; \Psi) \right\}, \end{aligned}$$

for the functions $I_1,\ I_2 : \mathbb{R}^d \to \mathbb{R}$ defined as the solutions to the integration problems on 2-dimensional unit cube given by

$$\begin{aligned} I_1(\mathbf{y}_t; \Psi, \Psi^*) &:= \int_0^1 \int_0^1 \tilde{w}\left(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi, \Psi^*\right) m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t}, \\ I_2(\mathbf{y}_t; \Psi) &:= \int_0^1 \int_0^1 m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t}, \end{aligned}$$

where $\tilde{w}\left(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi, \Psi^*\right) := w\Big(\mathbf{y}_t, T_\epsilon(s_{\epsilon,t}), T_x(s_{x,t}); \Psi, \Psi^*\Big)$ and the function $\tilde{w} : \mathbb{R}^d \times \mathbb{R}_+^d \times \mathbb{R}_+^k \longrightarrow \mathbb{R}$ is defined in Lemma A9 in Appendix A.4.5, the quantile transformations $T_\epsilon : [0,1] \longrightarrow \mathbb{R}_+^d$ and $T_x : [0,1] \longrightarrow \mathbb{R}_+^k$ are defined as

$$T_\epsilon(s) := \left(\frac{\chi^{-1}_{\nu_\epsilon^1}(s)}{\nu_\epsilon^1}, \ldots, \frac{\chi^{-1}_{\nu_\epsilon^d}(s)}{\nu_\epsilon^d}\right)_{1\times d} \text{ and } T_x(s) := \left(\frac{\chi^{-1}_{\nu_x^1}(s)}{\nu_x^1}, \ldots, \frac{\chi^{-1}_{\nu_x^k}(s)}{\nu_x^k}\right)_{1\times k},$$

and the function $m : \mathbb{R}^d \times [0,1]^2 \longrightarrow \mathbb{R}$ is given by

$$m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi) := \pi_{\mathbf{Y}_t|S_{\epsilon,t},S_{x,t},\Psi}(\mathbf{y}_t)\left(\sigma^2\right)^{\frac{k}{2}} \left|\mathbf{D}_{\epsilon,t}\right|^{\frac{1}{2}} \left|\mathbf{D}_{x,t}\right|^{\frac{1}{2}} \left|\mathbf{M}_t\right|^{-\frac{1}{2}} \left|\mathbf{N}_t\right|^{-\frac{1}{2}} e^{-\frac{1}{2}\delta_x \left(\mathbf{D}_{x,t}^{-1} - \sigma^2 \mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbf{N}_t^{-1}\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1} - \sigma^2 \mathbf{M}_t^{-1}\right.}$$

for probability density function $\pi_{\mathbf{Y}_t|S_{\epsilon,t},S_{x,t},\Psi}(\mathbf{y}_t^o)$ *which equals* $\pi_{\mathbf{Y}_t|T_\epsilon(S_{\epsilon,t}),T_x(S_{x,t}),\Psi}(\mathbf{y}_t^o)$ *specified in Lemma A8 in Appendix A.4.5, and matrices*

$$\mathbf{M}_t = \sigma^2\mathbf{D}_{x,t} + \mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbf{W} \text{ and } \mathbf{N}_t = \mathbf{D}_{\epsilon,t} - \mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t},$$

where

$$\mathbf{D}_{\epsilon,t} = \begin{pmatrix} \frac{\chi^{-1}_{\nu^1_\epsilon}(S_{\epsilon,t})}{\nu^1_\epsilon} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \frac{\chi^{-1}_{\nu^d_\epsilon}(S_\epsilon)}{\nu^d_\epsilon} \end{pmatrix}_{d\times d}, \mathbf{D}_{x,t} = \begin{pmatrix} \frac{\chi^{-1}_{\nu^1_x}(S_{x,t})}{\nu^1_x} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \frac{\chi^{-1}_{\nu^k_x}(S_{x,t})}{\nu^k_x} \end{pmatrix}_{k\times k}.$$

Proof. The proof of Theorem 13 is given in Subsection A.4.1.1 in Appendix A.4. □

Corollary 1. *Under the assumptions of Special Case 1: Grouped t-Student IND PPCA specified in Subsubsection 3.4.1.3, the E-step of the corresponding EM algorithm as a case-specific variant of Theorem 13, is given by*

$$Q_{GStS,ind,1}(\Psi,\Psi^*) = \mathbb{E}_{\mathbf{X}_{1:N},\mathbf{U}_{1:N},\mathbf{V}_{1:N}|\mathbf{Y}_{1:N},\Psi}\Big[\log \pi_{\mathbf{Y}_{1:N},\mathbf{X}_{1:N},\mathbf{U}_{1:N}\mathbf{V}_{1:N}|\Psi^*}(\mathbf{Y}_{1:N},\mathbf{X}_{1:N},\mathbf{U}_{1:N},\mathbf{V}_{1:N})\Big].$$

The E-step of the Grouped t-Student IND PPCA is formulated by using special cases of the functions $I_1,\ I_2 : \mathbb{R}^d \to \mathbb{R}$ *defined in Theorem 13. The case-specific variant of the function* $m : \mathbb{R}^d \times [0,1]^2 \longrightarrow \mathbb{R}$ *is given by*

$$m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t};\Psi) = \pi_{\mathbf{Y}_t|S_{\epsilon,t},S_{x,t},\Psi}(\mathbf{y}_t)\left(\sigma^2\right)^{\frac{k}{2}}\left|\mathbf{D}_{\epsilon,t}\right|^{\frac{1}{2}}\left|\mathbf{D}_{x,t}\right|^{\frac{1}{2}}\left|\mathbf{M}_t\right|^{-\frac{1}{2}}\left|\mathbf{N}_t\right|^{-\frac{1}{2}},$$

where, $\pi_{\mathbf{Y}_t|S_{\epsilon,t},S_{x,t},\Psi}(\mathbf{y}_t)$ *is a probability density function of the d-dimensional random vector* $\mathbf{Y}_t|S_{\epsilon,t},Sx,t,\Psi \sim \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma}_{y,t})$. *The case specific variant of the function* $\tilde{w} : \mathbb{R}^d \times [0,1]^2 \longrightarrow \mathbb{R}$ *is defined as following*

$$\begin{aligned}
\tilde{w}(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t};\Psi,\Psi^*) &= \log \pi_{S_{\epsilon,t}|\Psi^*}(s_{\epsilon,t}) + \log \pi_{S_{x,t}|\Psi^*}(s_{x,t}) - \frac{k+d}{2}\log 2\pi - \frac{d}{2}\log\sigma^{*2} \\
&+ \frac{1}{2}\sum_{i=1}^{d}\log\frac{\chi^{-1}_{\nu^i_\epsilon}(s_{\epsilon,t})}{\nu^i_\epsilon} + \frac{1}{2}\sum_{j=1}^{k}\log\frac{\chi^{-1}_{\nu^j_x}(s_{x,t})}{\nu^j_x} - \frac{1}{2\sigma^{*2}}Tr\left\{\mathbf{y}_t^T\mathbf{y}_t\mathbf{D}_{\epsilon,t}\right\} + \frac{1}{\sigma^{*2}}\mathbf{y}_t\mathbf{D}_{\epsilon,t}\boldsymbol{\mu}^{*T} \\
&- \frac{1}{2\sigma^{*2}}\boldsymbol{\mu}^*\mathbf{D}_{\epsilon,t}\boldsymbol{\mu}^{*T} + \frac{1}{\sigma^{*2}}Tr\left\{\mathbf{y}_t^T\mathbf{y}_t\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\right\} - \frac{\sigma^2}{2}Tr\left\{\mathbf{M}_t^{-1}\left(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\mathbf{W}^* + \mathbf{D}_{x,t}\right)\right\} \\
&- \frac{1}{\sigma^{*2}}(\mathbf{y}_t - \boldsymbol{\mu})\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\boldsymbol{\mu}^{*T} - \frac{1}{\sigma^{*2}}\mathbf{y}_t\mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\boldsymbol{\mu}^T \\
&- \frac{1}{2}Tr\left\{\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\left(\mathbf{y}_t^T\mathbf{y}_t - \mathbf{y}_t^T\boldsymbol{\mu} - \boldsymbol{\mu}^T\mathbf{y}_t + \boldsymbol{\mu}^T\boldsymbol{\mu}\right)\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\left(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\mathbf{W}^* + \mathbf{D}_{x,t}\right)\right\}.
\end{aligned}$$

Proof. The proof of Corollary 1 is analogous to the proof of Theorem 13 when one sets $\boldsymbol{\delta}_\epsilon = \boldsymbol{\delta}_x = \boldsymbol{\delta}^*_\epsilon = \boldsymbol{\delta}^*_x = \mathbf{0}$. □

Corollary 2. *Under the assumptions of Special Case 2: Skew t-Student IND PPCA specified in Subsubsection 3.4.1.3, the E-step of the corresponding EM algorithm as a case-specific variant of Theorem 13, is given by*

$$Q_{GStS,ind,2}(\Psi,\Psi^*)=\mathbb{E}_{\mathbf{X}_{1:N},U_{1:N},V_{1:N}|\mathbf{Y}_{1:N},\Psi}\Big[\log\pi_{\mathbf{Y}_{1:N},\mathbf{X}_{1:N},U_{1:N},V_{1:N}|\Psi^*}(\mathbf{Y}_{1:N},\mathbf{X}_{1:N},U_{1:N},V_{1:N})\Big].$$

The E-step of the Skew t-Student IND PPCA is formulated by using special cases of the functions $I_1,\ I_2:\mathbb{R}^d\to\mathbb{R}$ defined analogously as in Theorem 13. Let us denote the quantile transformations $T_\epsilon(s):=\frac{\chi^{-1}_{\nu_\epsilon}(s)}{\nu_\epsilon}$ and $T_x(s):=\frac{\chi^{-1}_{\nu_x}(s)}{\nu_x}$. The case-specific variant of the function $m:\mathbb{R}^d\times[0,1]^2\longrightarrow\mathbb{R}$ is given by

$$\begin{aligned}m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)&=\pi_{\mathbf{Y}_t|S_{\epsilon,t},S_{x,t},\Psi}(\mathbf{y}_t)\big(\sigma^2\big)^{\frac{k}{2}}T_\epsilon(s_{\epsilon,t})^{\frac{d}{2}}T_x(s_{x,t})^{\frac{k}{2}}\Big|\mathbf{M}_t\Big|^{-\frac{1}{2}}\Big|\mathbf{N}_t\Big|^{-\frac{1}{2}}\\&\times\exp\Big\{-\frac{1}{2}\boldsymbol{\delta}_x\Big(T_x(s_{x,t})^{-1}\mathbb{I}_k-\sigma^2T_\epsilon(s_{\epsilon,t})^2\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{N}_t^{-1}\mathbf{W}\mathbf{M}_t^{-1}-\sigma^2\mathbf{M}_t^{-1}\Big)\boldsymbol{\delta}_x^T\Big\},\end{aligned}$$

for $\pi_{\mathbf{Y}_t|S_{\epsilon,t},S_{x,t},\Psi}(\mathbf{y}_t)$ being a probability density function of the d-dimensional Gaussian random vector $\mathbf{Y}_t|S_{\epsilon,t},S_{x,t},\Psi$ with first and second central moments given by

$$\boldsymbol{\mu}_{y,t}=\boldsymbol{\mu}+T_\epsilon(s_{\epsilon,t})^{-1}\boldsymbol{\delta}_\epsilon+\sigma^2T_\epsilon(s_{\epsilon,t})\boldsymbol{\delta}_x\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{N}_t^{-1}\ \text{and}\ \boldsymbol{\Sigma}_{y,t}=\sigma^2\mathbf{N}_t^{-1},$$

for matrices

$$\mathbf{M}_t=\sigma^2T_x(s_{x,t})\mathbb{I}_k+T_\epsilon(s_{\epsilon,t})\mathbf{W}^T\mathbf{W}\ \text{and}\ \mathbf{N}_t=T_\epsilon(s_{\epsilon,t})\big(\mathbb{I}_d-T_\epsilon(s_{\epsilon,t})\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^T\big).$$

The case-specific variant of the function $\tilde{w}:\mathbb{R}^d\times[0,1]^2\longrightarrow\mathbb{R}$ is defined as

$$\begin{aligned}\tilde{w}(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi,\Psi^*)&=-\frac{k+d}{2}\log2\pi-\frac{d}{2}\log\sigma^{*2}+\frac{d}{2}\log T_\epsilon(s_{\epsilon,t})+\frac{k}{2}\log T_x(s_{x,t})\\&-\frac{T_x(s_{x,t})^{-1}}{2}\boldsymbol{\delta}_x^*\boldsymbol{\delta}_x^{*T}-\frac{T_\epsilon(s_{\epsilon,t})}{2\sigma^{*2}}Tr\Big\{\mathbf{y}_t^T\mathbf{y}_t\Big\}+\frac{1}{\sigma^{*2}}\mathbf{y}_t\Big(T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}^*+\boldsymbol{\delta}_\epsilon^*\Big)^T-\frac{T_\epsilon(s_{\epsilon,t})}{2\sigma^{*2}}\boldsymbol{\mu}^*\boldsymbol{\mu}^{*T}\\&-\frac{1}{\sigma^{*2}}\boldsymbol{\mu}^*\boldsymbol{\delta}_\epsilon^{*T}-\frac{T_\epsilon(s_{\epsilon,t})^{-1}}{2\sigma^{*2}}\boldsymbol{\delta}_\epsilon^*\boldsymbol{\delta}_\epsilon^{*T}+\frac{T_\epsilon(s_{\epsilon,t})^2}{\sigma^{*2}}Tr\Big\{\mathbf{y}_t^T\mathbf{y}_t\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}\Big\}\\&+T_\epsilon(s_{\epsilon,t})\mathbf{y}_t\mathbf{W}\mathbf{M}_t^{-1}\Big(-\frac{T_\epsilon(s_{\epsilon,t})}{\sigma^{*2}}\mathbf{W}^{*T}\boldsymbol{\mu}^{*T}-\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\boldsymbol{\delta}_\epsilon^{*T}+\boldsymbol{\delta}_x^{*T}\Big)\\&+\Big(\frac{T_\epsilon(s_{\epsilon,t})}{\sigma^{*2}}\mathbf{y}_t\mathbf{W}^*-\frac{T_\epsilon(s_{\epsilon,t})}{\sigma^{*2}}\boldsymbol{\mu}^*\mathbf{W}^*-\frac{1}{\sigma^{*2}}\boldsymbol{\delta}_\epsilon^*\mathbf{W}^*+\boldsymbol{\delta}_x^*\Big)\mathbf{M}_t^{-1}\Big(-T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W}-\boldsymbol{\delta}_\epsilon\mathbf{W}+\sigma^2\boldsymbol{\delta}_x\Big)^T\\&-\frac{1}{2}Tr\Big\{\mathbf{M}_t^{-1}T_\epsilon(s_{\epsilon,t})^2\mathbf{W}^T\mathbf{y}_t^T\mathbf{y}_t\mathbf{W}\mathbf{M}_t^{-1}\Big(\frac{T_\epsilon(s_{\epsilon,t})}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{W}^*+T_x(s_{x,t})\mathbb{I}_k\Big)\Big\}\\&-\frac{1}{2}T_\epsilon(s_{\epsilon,t})Tr\Big\{\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{y}_t^T\Big(-T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W}-\boldsymbol{\delta}_\epsilon\mathbf{W}+\sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}_t^{-1}\Big(\frac{T_\epsilon(s_{\epsilon,t})}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{W}^*+T_x(s_{x,t})\mathbb{I}_k\Big)\Big\}\\&-\frac{1}{2}T_\epsilon(s_{\epsilon,t})Tr\Big\{\mathbf{M}_t^{-1}\Big(-T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W}-\boldsymbol{\delta}_\epsilon\mathbf{W}+\sigma^2\boldsymbol{\delta}_x\Big)^T\mathbf{y}_t\mathbf{W}\mathbf{M}_t^{-1}\Big(\frac{T_\epsilon(s_{\epsilon,t})}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{W}^*+T_x(s_{x,t})\mathbb{I}_k\Big)\Big\}\\&-\frac{1}{2}Tr\Big\{\mathbf{M}_t^{-1}\Big(-T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W}-\boldsymbol{\delta}_\epsilon\mathbf{W}+\sigma^2\boldsymbol{\delta}_x\Big)^T\Big(-T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W}-\boldsymbol{\delta}_\epsilon\mathbf{W}+\sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}_t^{-1}\Big(\frac{T_\epsilon(s_{\epsilon,t})}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{W}^*+T_x(s_{x,t})\mathbb{I}_k\Big)\Big\}\\&-\frac{\sigma^2}{2}Tr\Big\{\mathbf{M}_t^{-1}\Big(\frac{T_\epsilon(s_{\epsilon,t})}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{W}^*+T_x(s_{x,t})\mathbb{I}_k\Big)\Big\}.\end{aligned}$$

Proof. The proof of Corollary 2 in given in Appendix A.4.1.2. □

Corollary 3. *Under the assumptions of Special Case 3: t-Student IND PPCA specified in Subsubsection 3.4.1.3, the E-step of the corresponding EM algorithm as a case-specific variant of Theorem 13, is given by*

$$Q_{GStS,ind,3}(\Psi,\Psi^*)=\mathbb{E}_{\mathbf{X}_{1:N},U_{1:N},V_{1:N}|\mathbf{Y}_{1:N},\Psi}\Big[\log\pi_{\mathbf{Y}_{1:N},\mathbf{X}_{1:N},U_{1:N},V_{1:N}|\Psi^*}(\mathbf{Y}_{1:N},\mathbf{X}_{1:N},U_{1:N},V_{1:N})\Big].$$

The E-step of the Generalized t-Student IND PPCA is formulated by using special cases of the functions $I_1,\ I_2:\mathbb{R}^d\to\mathbb{R}$ defined analogously as in Theorem 13. Let us denote the quantile transformations $T_\epsilon(s):=\frac{\chi^{-1}_{\nu_\epsilon}(s)}{\nu_\epsilon}$ and $T_x(s):=\frac{\chi^{-1}_{\nu_x}(s)}{\nu_x}$. The case-specific variant of the function $m:\mathbb{R}^d\times[0,1]^2\longrightarrow\mathbb{R}$ is given by

$$m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)=\pi_{\mathbf{Y}_t|S_{\epsilon,t},S_{x,t},\Psi}(\mathbf{y}_t)\left(\sigma^2\right)^{\frac{k}{2}}T_\epsilon(s_{\epsilon,t})^{\frac{d}{2}}T_x(s_{x,t})^{\frac{k}{2}}\left|\mathbf{M}_t\right|^{-\frac{1}{2}}\left|\mathbf{N}_t\right|^{-\frac{1}{2}},$$

for $\pi_{\mathbf{Y}_t|S_{\epsilon,t},S_{x,t},\Psi}(\mathbf{y}_t)$ being a probability density function of the d-dimensional Gaussian random vector $\mathbf{Y}_t|S_{\epsilon,t},S_{x,t},\Psi\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma}_{y,t})$, where

$$\begin{aligned}
\boldsymbol{\Sigma}_{y,t}&=\sigma^2\mathbf{N}_t^{-1},\\
\mathbf{M}_t&=\sigma^2T_x(s_{x,t})\mathbb{I}_k+T_\epsilon(s_{\epsilon,t})\mathbf{W}^T\mathbf{W},\\
\mathbf{N}_t&=T_\epsilon(s_{\epsilon,t})\left(\mathbb{I}_d-T_\epsilon(s_{\epsilon,t})\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^T\right).
\end{aligned}$$

The case-specific variant of the function $\tilde{w}:\mathbb{R}^d\times[0,1]^2\longrightarrow\mathbb{R}$ is defined as

$$\begin{aligned}
\tilde{w}(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi,\Psi^*)&=-\frac{k+d}{2}\log2\pi-\frac{d}{2}\log\sigma^{*2}+\frac{d}{2}\log T_\epsilon(s_{\epsilon,t})+\frac{k}{2}\log T_x(s_{x,t})\\
&-\frac{T_\epsilon(s_{\epsilon,t})}{2\sigma^{*2}}Tr\left\{\mathbf{y}_t^T\mathbf{y}_t\right\}+\frac{T_\epsilon(s_{\epsilon,t})}{\sigma^{*2}}\mathbf{y}_t\boldsymbol{\mu}^{*T}-\frac{T_\epsilon(s_{\epsilon,t})}{2\sigma^{*2}}\boldsymbol{\mu}^*\boldsymbol{\mu}^{*T}\\
&+\frac{T_\epsilon(s_{\epsilon,t})^2}{\sigma^{*2}}Tr\left\{\mathbf{y}_t^T\mathbf{y}_t\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}\right\}-\frac{T_\epsilon(s_{\epsilon,t})^2}{\sigma^{*2}}\mathbf{y}_t\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}\boldsymbol{\mu}^{*T}\\
&-\frac{T_\epsilon(s_{\epsilon,t})^2}{\sigma^{*2}}(\mathbf{y}_t-\boldsymbol{\mu}^*)\mathbf{W}^*\mathbf{M}_t^{-1}\mathbf{W}^T\boldsymbol{\mu}^T-\frac{1}{2}Tr\left\{\sigma^2\mathbf{M}_t^{-1}\Big(\frac{T_\epsilon(s_{\epsilon,t})}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{W}^*+T_x(s_{x,t})\mathbb{I}_k\Big)\right\}\\
&-\frac{T_\epsilon(s_{\epsilon,t})^2}{2}Tr\left\{\mathbf{M}_t^{-1}\mathbf{W}^T\left(\mathbf{y}_t^T\mathbf{y}_t-\mathbf{y}_t^T\boldsymbol{\mu}-\boldsymbol{\mu}^T\mathbf{y}_t+\boldsymbol{\mu}^T\boldsymbol{\mu}\right)\mathbf{W}\mathbf{M}_t^{-1}\Big(\frac{T_\epsilon(s_{\epsilon,t})}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{W}^*+T_x(s_{x,t})\mathbb{I}_k\Big)\right\}.
\end{aligned}$$

Proof. The proof of Corollary 3 is analogous to the proof of Corollary 2 given in Appendix A.4.1.2 when one sets $\delta_\epsilon=\delta_x=\delta_\epsilon^*=\delta_x^*=\mathbf{0}$. □

M-Step of The EM Algorithm

In order to calculate the maximizers of the function $Q_{GStS,ind}$ and proceed with the M-step of the EM algorithm, we need to introduce a new notation. Let I be an integral operator of a function $b:\mathbb{R}^{d_o}\times\mathbb{R}^d\times\mathbb{R}^k\to\mathbb{R}^{n\times n'}$ for some integer numbers n and n' defined as

$$I(b)(\mathbf{y}^o_{1:N})_{n\times n'}:=\sum_{t=1}^N\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\int_0^1\int_0^1 b\Big(\mathbf{y}_t,T_\epsilon(s_{\epsilon,t}),T_x(s_{x,t})\Big)\,m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\right\}. \tag{3.39}$$

The function m is specified as in Theorem 13 and the function $H : \mathbb{R}^{(N-1)\times d} \longrightarrow \mathbb{R}$ is a product of $N-1$ integration problems on $[0,1]^2$ defined by the function I_2 from Theorem 13, given by

$$H(\mathbf{y}_{1:N/t}; \Psi) := \prod_{s=1, s\neq t}^{N} I_2(\mathbf{y}_s; \Psi). \tag{3.40}$$

The set $\mathbf{y}_{1:N/t}$ is a $N-1$ sequence of d_o-dimensional vectors with observed entries of $\mathbf{y}_t$ without the vector $\mathbf{y}_t^o$, $\mathbf{y}_{1:N/t} = \left\{\mathbf{y}_1, \ldots, \mathbf{y}_{t-1}, \mathbf{y}_{t+1}, \ldots, \mathbf{y}_N\right\}$. let us remark that the co-domain of the operator I depends on the dimensionality of a co-domain of the function b. Given the definition of I, we can formulate the M-step of the EM Algorithm for the Generalized Skew t-Student PPCA as in Theorem 14.

Theorem 14. *Let the observation vector $\mathbf{Y}_t$ be modelled as in Equation (2.7) with the latent processes $\mathbf{X}_t$, $\boldsymbol{\epsilon}_t$, $\mathbf{U}_t$ and $\mathbf{V}_t$ following the assumptions of Independent Generalized Skew t-Student PPCA given in Subsection 3.4.2.1 in Equation (3.27) and Equation (3.28). The maximizers of the E-step of the Expectation-Maximisation algorithm for Independent Generalized Skew t-Student PPCA derived in Theorem 13 with respect to the vector of parameters Ψ^* are the solutions to the system of equations determined by $\nabla_{\Psi^*} Q_{GStS,ind}(\Psi, \Psi^*) = \mathbf{0}$ as following*

$$\begin{cases} I(\frac{\partial}{\partial \boldsymbol{\mu}^*} w)(\mathbf{y}_{1:N})_{1\times d} = \mathbf{0} \\ I(\frac{\partial}{\partial \mathbf{W}^*} w)(\mathbf{y}_{1:N})_{d\times k} = \mathbf{0} \\ I(\frac{\partial}{\partial \sigma^{*2}} w)(\mathbf{y}_{1:N})_{1\times 1} = 0 \\ I(\frac{\partial}{\partial \delta_\epsilon^*} w)(\mathbf{y}_{1:N})_{1\times d} = \mathbf{0} \\ I(\frac{\partial}{\partial \delta_x^*} w)(\mathbf{y}_{1:N})_{1\times k} = \mathbf{0} \end{cases}$$

for the integral operator I defined in Equation (3.39) and the gradient of the function w defined in Lemma A9 being the following

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\mu}^*} \tilde{w}(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi, \Psi^*) &= \frac{1}{\sigma^{*2}} \mathbf{y}_t \mathbf{D}_{\epsilon,t} - \frac{1}{\sigma^{*2}} \boldsymbol{\mu}^* \mathbf{D}_{\epsilon,t} - \frac{1}{\sigma^{*2}} \delta_\epsilon^* - \frac{1}{\sigma^{*2}} \mathbf{y}_t \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1} \mathbf{W}^{*T} \mathbf{D}_{\epsilon,t} \\ &+ \frac{1}{\sigma^{*2}} \boldsymbol{\mu} \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1} \mathbf{W}^{*T} \mathbf{D}_{\epsilon,t} + \frac{1}{\sigma^{*2}} \delta_\epsilon \mathbf{W} \mathbf{M}_t^{-1} \mathbf{W}^{*T} \mathbf{D}_{\epsilon,t} - \frac{\sigma^2}{\sigma^{*2}} \delta_x \mathbf{M}_t^{-1} \mathbf{W}^{*T} \mathbf{D}_{\epsilon,t}, \end{aligned}$$

$$\begin{aligned} \frac{\partial}{\partial \mathbf{W}^*} \tilde{w}(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi, \Psi^*) &= \frac{1}{\sigma^{*2}} \mathbf{D}_{\epsilon,t} \mathbf{y}_t^T \mathbf{y}_t \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1} - \frac{1}{\sigma^{*2}} \left(\mathbf{D}_{\epsilon,t} \boldsymbol{\mu}^{*T} + \delta_\epsilon^{*T}\right) \mathbf{y}_t \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1} \\ &+ \frac{1}{\sigma^{*2}} \left(\mathbf{y}_t \mathbf{D}_{\epsilon,t} - \boldsymbol{\mu}^* \mathbf{D}_{\epsilon,t} - \delta_\epsilon^*\right)^T \left(- \boldsymbol{\mu} \mathbf{D}_{\epsilon,t} \mathbf{W} - \delta_\epsilon \mathbf{W} + \sigma^2 \delta_x \right) \mathbf{M}_t^{-1} - \frac{1}{\sigma^{*2}} \mathbf{D}_{\epsilon,t} \mathbf{W}^* \mathbf{M}_t^{-1} \mathbf{W}^T \mathbf{D}_{\epsilon,t} \mathbf{y}_t^T \mathbf{y}_t \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1} \\ &+ \frac{1}{\sigma^{*2}} \mathbf{D}_{\epsilon,t} \mathbf{W}^* \mathbf{M}_t^{-1} \Big(\mathbf{W}^T \mathbf{D}_{\epsilon,t} \mathbf{y}_t^T \left(\boldsymbol{\mu} \mathbf{D}_{\epsilon,t} \mathbf{W} + \delta_\epsilon \mathbf{W} - \sigma^2 \delta_x\right) + \left(\boldsymbol{\mu} \mathbf{D}_{\epsilon,t} \mathbf{W} + \delta_\epsilon \mathbf{W} - \sigma^2 \delta_x\right)^T \mathbf{y}_t \mathbf{D}_{\epsilon,t} \mathbf{W} \Big) \mathbf{M}_t^{-1} \\ &- \frac{1}{\sigma^{*2}} \mathbf{D}_{\epsilon,t} \mathbf{W}^* \mathbf{M}_t^{-1} \left(\boldsymbol{\mu} \mathbf{D}_{\epsilon,t} \mathbf{W} + \delta_\epsilon \mathbf{W} - \sigma^2 \delta_x\right)^T \left(\boldsymbol{\mu} \mathbf{D}_{\epsilon,t} \mathbf{W} + \delta_\epsilon \mathbf{W} - \sigma^2 \delta_x\right) \mathbf{M}_t^{-1} - \frac{\sigma^2}{\sigma^{*2}} \mathbf{D}_{\epsilon,t} \mathbf{W}^* \mathbf{M}_t^{-1}, \end{aligned}$$

$$\begin{aligned} \frac{\partial}{\partial \sigma^{*2}} \tilde{w}(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi, \Psi^*) &= -\frac{d}{2\sigma^{*2}} - \frac{1}{2(\sigma^{*2})^2} \Big[- Tr\left\{\mathbf{y}_t^T \mathbf{y}_t \mathbf{D}_{\epsilon,t}\right\} + 2 \mathbf{y}_t \left(\mathbf{D}_{\epsilon,t} \boldsymbol{\mu}^{*T} + \delta_\epsilon^{*T}\right) \\ &- \boldsymbol{\mu}^* \mathbf{D}_{\epsilon,t} \boldsymbol{\mu}^{*T} - 2 \boldsymbol{\mu}^* \delta_\epsilon^{*T} - \delta_\epsilon^* \mathbf{D}_{\epsilon,t}^{-1} \delta_\epsilon^{*T} + 2\, Tr\left\{\mathbf{y}_t^T \mathbf{y}_t \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1} \mathbf{W}^{*T} \mathbf{D}_{\epsilon,t}\right\} \\ &- 2 \left(\mathbf{y}_t \mathbf{D}_{\epsilon,t} \mathbf{W} - \boldsymbol{\mu} \mathbf{D}_{\epsilon,t} \mathbf{W} - \delta_\epsilon \mathbf{W} + \sigma^2 \delta_x\right) \mathbf{M}_t^{-1} \mathbf{W}^{*T} \left(\boldsymbol{\mu}^* \mathbf{D}_{\epsilon,t} + \delta_\epsilon^*\right)^T \\ &+ 2 \mathbf{y}_t \mathbf{D}_{\epsilon,t} \mathbf{W}^* \mathbf{M}_t^{-1} \left(- \boldsymbol{\mu} \mathbf{D}_{\epsilon,t} \mathbf{W} - \delta_\epsilon \mathbf{W} + \sigma^2 \delta_x\right)^T - Tr\left\{\mathbf{M}_t^{-1} \mathbf{W}^T \mathbf{D}_{\epsilon,t} \mathbf{y}_t^T \mathbf{y}_t \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1} \mathbf{W}^{*T} \mathbf{D}_{\epsilon,t} \mathbf{W}^*\right\} \end{aligned}$$

$$+ Tr\left\{\mathbf{M}_t^{-1}\left(\mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbf{y}_t^T(\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}+\delta_\epsilon\mathbf{W}-\sigma^2\delta_x)+(\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}+\delta_\epsilon\mathbf{W}-\sigma^2\delta_x)^T\mathbf{y}_t\mathbf{D}_{\epsilon,t}\mathbf{W}\right)\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\mathbf{W}^*\right\}$$
$$- Tr\left\{\mathbf{M}_t^{-1}(\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}+\delta_\epsilon\mathbf{W}-\sigma^2\delta_x)^T(\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}+\delta_\epsilon\mathbf{W}\sigma^2\delta_x)\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\mathbf{W}^*\right\}-\sigma^2\,Tr\left\{\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\mathbf{W}^*\right\}\Big],$$

$$\frac{\partial}{\partial\delta_\epsilon^*}\tilde{w}(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi,\Psi^*)=\frac{1}{\sigma^{*2}}\mathbf{y}_t-\frac{1}{\sigma^{*2}}\boldsymbol{\mu}^*-\frac{1}{\sigma^{*2}}\delta_\epsilon^*\mathbf{D}_{\epsilon,t}^{-1}-\frac{1}{\sigma^{*2}}\mathbf{y}_t\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}$$
$$+\frac{1}{\sigma^{*2}}\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}+\frac{1}{\sigma^{*2}}\delta_\epsilon\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}-\frac{\sigma^2}{\sigma^{*2}}\delta_x\mathbf{M}_t^{-1}\mathbf{W}^{*T},$$

$$\frac{\partial}{\partial\delta_x^*}\tilde{w}(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi,\Psi^*)=-\delta_x^*\mathbf{D}_{x,t}^{-1}+\mathbf{y}_t\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}-\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}-\delta_\epsilon\mathbf{W}\mathbf{M}_t^{-1}+\sigma^2\delta_x\mathbf{M}_t^{-1}.$$

Proof. The proof of the Theorem 14 is given in Subsection A.4.1.3 in Appendix A.4. □

Corollary 4. *The solution to the system of equation from Theorem 14 which determines the maximizers of the function Q_{GStS} with respect to the parameters $\boldsymbol{\mu}^*, \delta_\epsilon^*, \delta_x^*, \sigma^2$ are given by explicit formulas defined by 2-dimensional integration problems on the hypercube $[0,1]^2$ as follows:*

$$\begin{cases}
\boldsymbol{\mu}^* = \Big[\mathbf{A}_6(\mathbf{y}_{1:N};\Psi)-\mathbf{A}_{10}(\mathbf{y}_{1:N};\Psi,\Psi^*)-\delta_\epsilon^*A_0(\mathbf{y}_{1:N};\Psi)+\boldsymbol{\mu}\mathbf{A}_{13}(\mathbf{y}_{1:N};\Psi,\Psi^*)+\left(\delta_\epsilon\mathbf{W}-\sigma^2\delta_x\right)\mathbf{A}_{12}(\mathbf{y}_{1:N};\Psi,\Psi^*)^T\Big]\mathbf{A}_1(\mathbf{y}_{1:N};\Psi)^{-1}\\
\delta_\epsilon^* = \Big[\mathbf{A}_5(\mathbf{y}_{1:N};\Psi)-\boldsymbol{\mu}^*A_0(\mathbf{y}_{1:N};\Psi)-\mathbf{A}_8(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}+\boldsymbol{\mu}\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}+(\delta_\epsilon\mathbf{W}-\sigma^2\delta_x)\mathbf{A}_4(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\Big]\mathbf{A}_2(\mathbf{y}_{1:N};\Psi)^{-1}\\
\delta_x^* = \Big[\mathbf{A}_8(\mathbf{y}_{1:N})-\boldsymbol{\mu}\mathbf{A}_{11}(\mathbf{y}_{1:N})-(\delta_\epsilon\mathbf{W}-\sigma^2\delta_x)\mathbf{A}_4(\mathbf{y}_{1:N})\Big]\mathbf{A}_3(\mathbf{y}_{1:N},\Psi)^{-1}\\
\sigma^{*2} = \frac{1}{dA_0(\mathbf{y}_{1:N};\Psi)}\Big[A_{20}(\mathbf{y}_{1:N};\Psi)+\boldsymbol{\mu}^*\mathbf{A}_1(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^{*T}+\delta_\epsilon^*\mathbf{A}_2(\mathbf{y}_{1:N};\Psi)\delta_\epsilon^{*T}+2\mathbf{A}_9(\mathbf{y}_{1:N};\Psi,\Psi^*)\boldsymbol{\mu}^T\\
\qquad+2\Big(\mathbf{A}_{10}(\mathbf{y}_{1:N};\Psi,\Psi^*)-\mathbf{A}_6(\mathbf{y}_{1:N};\Psi)-\boldsymbol{\mu}\mathbf{A}_{13}(\mathbf{y}_{1:N};\Psi,\Psi^*)\Big)\boldsymbol{\mu}^{*T}-2A_{21}(\mathbf{y}_{1:N};\Psi,\Psi^*)+\sigma^2\,Tr\left\{\mathbf{A}_{12}(\mathbf{y}_{1:N};\Psi)^T\mathbf{W}^*\right\}\\
\qquad+Tr\left\{\mathbf{A}_{17}(\mathbf{y}_{1:N};\Psi)^T\mathbf{W}^*\right\}-Tr\left\{\mathbf{A}_{18.1}(\mathbf{y}_{1:N};\Psi)^T\mathbf{W}^*\right\}-Tr\left\{\mathbf{A}_{18.2}(\mathbf{y}_{1:N};\Psi)^T\mathbf{W}^*\right\}+Tr\left\{\mathbf{A}_{19}(\mathbf{y}_{1:N};\Psi)^T\mathbf{W}^*\right\}\\
\qquad+2\Big(\boldsymbol{\mu}^*A_0(\mathbf{y}_{1:N};\Psi)-\mathbf{A}_5(\mathbf{y}_{1:N};\Psi)+\mathbf{A}_8(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}-\boldsymbol{\mu}\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\Big)\delta_\epsilon^{*T}\\
\qquad+2\Big(\mathbf{A}_7(\mathbf{y}_{1:N};\Psi,\Psi^*)-\boldsymbol{\mu}^*\mathbf{A}_{12}(\mathbf{y}_{1:N};\Psi,\Psi^*)-\delta_\epsilon^*\mathbf{W}^*\mathbf{A}_4(\mathbf{y}_{1:N};\Psi)\Big)(\delta_\epsilon\mathbf{W}-\sigma^2\delta_x)^T\Big].
\end{cases}$$

The solutions with respect to $\boldsymbol{\mu}^*$, δ_ϵ^* and σ^{*2} depend on parameter $\mathbf{W}^*$. The maximizer of the function $Q_{GStS,ind}$ with respect to $\mathbf{W}^*$ is determined by the equation

$$f(\mathbf{W}^*;\Psi,\Psi^*)=\mathbf{0} \tag{3.41}$$

for the function $f:\mathbb{R}^{d\times k}\to\mathbb{R}^{d\times k}$ such that $f(\mathbf{W}^*;\Psi,\Psi^*):=\frac{\partial Q_{GStS,ind}(\Psi,\Psi^*)}{\partial\mathbf{W}^*}$ being equal to

$$f(\mathbf{W}^*;\Psi,\Psi^*)=\mathbf{A}_{14}(\mathbf{y}_{1:N};\Psi)-\mathbf{A}_{15}(\mathbf{y}_{1:N};\Psi,\Psi^*)-\delta_\epsilon^{*T}\mathbf{A}_8(\mathbf{y}_{1:N};\Psi)-\mathbf{A}_{16}(\mathbf{y}_{1:N};\Psi,\Psi^*)+\delta_\epsilon^{*T}\boldsymbol{\mu}\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi)$$
$$+\delta_\epsilon^{*T}(\delta\mathbf{W}-\sigma^2\mathbf{W})\mathbf{A}_4(\mathbf{y}_{1:N};\Psi)-\mathbf{A}_{17}(\mathbf{y}_{1:N};\Psi,\Psi^*)+\mathbf{A}_{18.1}(\mathbf{y}_{1:N};\Psi,\Psi^*)+\mathbf{A}_{18.2}(\mathbf{y}_{1:N};\Psi,\Psi^*)$$
$$-\mathbf{A}_{19}(\mathbf{y}_{1:N};\Psi,\Psi^*)-\sigma^2\mathbf{A}_{12}(\mathbf{y}_{1:N};\Psi,\Psi^*),$$

and is linear with respect to the parameters $\boldsymbol{\mu}^*$ and δ_ϵ^*. The 2-dimensional integrals $\mathbf{A}_i$ on the hypercube $[0,1]^2$, for $i\in\{0,\ldots,22\}$, are defined in Subsubsection A.4.1.4 in Appendix A.4.

Proof. The proof of the Corollary 4 is given in Subection A.4.1.4 in Appendix A.4. □

Corollary 5. *Under the assumptions of Special Case 1: Grouped t-Student IND PPCA specified in Subsubsection 3.4.1.3, the M-step of the corresponding EM algorithm with the $Q_{GStS,ind,1}$ function determined in Corollary 1 is a case-specific variant of the solution to the system of equations defined in*

Theorem 14 under the problem formulation $\nabla_{\Psi^*} Q_{GStS,ind,1} = \mathbf{0}$ *for* $\Psi^* = \left[\mathbf{W}^*, \boldsymbol{\mu}^*, \sigma^{*2}\right]$. *The maximizers with respect to* Ψ^* *are given by explicit formulas defined by* 2*-dimensional integration problems on the hypercube* $[0,1]^2$ *as follows*

$$\begin{cases} \boldsymbol{\mu}^* = \left[\mathbf{A}_6(\mathbf{y}_{1:N};\Psi) - \mathbf{A}_{10}(\mathbf{y}_{1:N};\Psi) + \boldsymbol{\mu}\mathbf{A}_{13}(\mathbf{y}_{1:N};\Psi)\right]\mathbf{A}_1(\mathbf{y}_{1:N};\Psi)^{-1} \\ \sigma^{*2} = \frac{1}{dA_0(\mathbf{y}_{1:N};\Psi)}\Big[A_{20}(\mathbf{y}_{1:N};\Psi) - 2\mathbf{A}_6(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^{*T} + \boldsymbol{\mu}^*\mathbf{A}_1(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^{*T} - 2A_{21}(\mathbf{y}_{1:N};\Psi,\Psi^*) + 2\mathbf{A}_{10}(\mathbf{y}_{1:N};\Psi,\Psi^*)\boldsymbol{\mu}^{*T} \\ \quad +2\mathbf{A}_9(\mathbf{y}_{1:N};\Psi,\Psi^*)\boldsymbol{\mu}^T - 2\boldsymbol{\mu}\mathbf{A}_{13}(\mathbf{y}_{1:N};\Psi,\Psi^*)\boldsymbol{\mu}^{*T} + Tr\left\{\mathbf{A}_{17}(\mathbf{y}_{1:N};\Psi,\Psi^*)^T\mathbf{W}^*\right\} - Tr\left\{\mathbf{A}_{18.1}(\mathbf{y}_{1:N};\Psi,\Psi^*)^T\mathbf{W}^*\right\} \\ \quad - Tr\left\{\mathbf{A}_{18.2}(\mathbf{y}_{1:N};\Psi,\Psi^*)^T\mathbf{W}^*\right\} + Tr\left\{\mathbf{A}_{19}(\mathbf{y}_{1:N};\Psi,\Psi^*)^T\mathbf{W}^*\right\} + \sigma^2\, Tr\left\{\mathbf{A}_{12}(\mathbf{y}_{1:N};\Psi,\Psi^*)^T\mathbf{W}^*\right\}\Big]. \end{cases}$$

The solutions with respect to $\boldsymbol{\mu}^*$ *and* σ^{*2} *depend on the value of the parameter* $\mathbf{W}^*$. *The maximizer of the function* $Q_{GStS,ind,1}$ *with respect to* $\mathbf{W}^*$ *is determined by the equation*

$$f(\mathbf{W}^*;\Psi,\Psi^*) = \mathbf{0},$$

for the function $f : \mathbb{R}^{d\times k} \to \mathbb{R}^{d\times k}$ *such that* $f(\mathbf{W}^*;\Psi,\Psi^*) := \frac{\partial Q(\Psi,\Psi^*)}{\partial \mathbf{W}^*}$ *specified as*

$$\begin{aligned} f(\mathbf{W}^*;\Psi,\Psi^*) = \mathbf{A}_{14}(\mathbf{y}_{1:N};\Psi) - \mathbf{A}_{15}(\mathbf{y}_{1:N};\Psi,\Psi^*) - \mathbf{A}_{16}(\mathbf{y}_{1:N};\Psi,\Psi^*) - \mathbf{A}_{17}(\mathbf{y}_{1:N};\Psi,\Psi^*) + \mathbf{A}_{18.1}(\mathbf{y}_{1:N};\Psi,\Psi^*) \\ + \mathbf{A}_{18.2}(\mathbf{y}_{1:N};\Psi,\Psi^*) - \mathbf{A}_{19}(\mathbf{y}_{1:N};\Psi,\Psi^*) - \sigma^2\mathbf{A}_{12}(\mathbf{y}_{1:N};\Psi,\Psi^*), \end{aligned}$$

and is linear with respect to the parameters $\boldsymbol{\mu}^*$. *The* 2*-dimensional integrals* $\mathbf{A}_i$ *on the hypercube* $[0,1]^2$, *for* $i \in \{0,\ldots,22\}$, *are defined in Subsubsection A.4.1.4 in Appendix A.4 given the following simplifications*

$$\begin{aligned} \mathbf{A}_{16}(\mathbf{y}_{1:N};\Psi,\Psi^*)_{d\times k} &:= \sum_{t=1}^{N}\left\{H(\mathbf{y}_{1:N/t};\Psi)\int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}(\mathbf{y}_t - \boldsymbol{\mu}^*)^T\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\, m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t};\Psi)\, ds_{\epsilon,t}\, ds_{x,t}\right\}, \\ \mathbf{A}_{18.1}(\mathbf{y}_{1:N};\Psi,\Psi^*)_{k\times d} &:= \sum_{t=1}^{N}\left\{H(\mathbf{y}_{1:N/t};\Psi)\int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbf{y}_t^T\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\, m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t};\Psi)\, ds_{\epsilon,t}\, ds_{x,t}\right\}, \\ \mathbf{A}_{18.2}(\mathbf{y}_{1:N};\Psi,\Psi^*)_{k\times d} &:= \sum_{t=1}^{N}\left\{H(\mathbf{y}_{1:N/t};\Psi)\int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\boldsymbol{\mu}^T\mathbf{y}_t\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\, m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t};\Psi)\, ds_{\epsilon,t}\, ds_{x,t}\right\}, \\ \mathbf{A}_{19}(\mathbf{y}_{1:N};\Psi,\Psi^*)_{k\times d} &:= \sum_{t=1}^{N}\left\{H(\mathbf{y}_{1:N/t};\Psi)\int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\boldsymbol{\mu}^T\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\, m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t};\Psi)\, ds_{\epsilon,t}\, ds_{x,t}\right\}. \end{aligned}$$

Proof. The proof to Corollary 5 is analogous to the proof to Theorem 14 and Corollary 4 when one sets $\delta_\epsilon = \delta_x = \delta_\epsilon^* = \delta_x^* = \mathbf{0}$. □

Corollary 6. *Under the assumptions of Special Case 2: Skew t-Student IND PPCA specified in Subsubsection 3.4.1.3, the M-step of the corresponding EM algorithm with the* $Q_{GStS,ind,2}$ *function determined in Corollary 2 is a case-specific variant of the solution to the system of equations defined in Theorem 14 under the problem formulation* $\nabla_{\Psi^*} Q_{GStS,ind,2} = \mathbf{0}$ *for* $\Psi^* = \left[\mathbf{W}^*, \boldsymbol{\mu}^*, \boldsymbol{\delta}_\epsilon, \boldsymbol{\delta}_X, \sigma^{*2}\right]$. *The*

maximizers with respect to Ψ^* *are given by explicit formulas defined by* 2*-dimensional integration problems on the hypercube* $[0,1]^2$ *as follows*

$$
\begin{cases}
\boldsymbol{\mu}^* = A_1(\mathbf{y}_{1:N};\Psi)^{-1}\Big[\mathbf{A}_6(\mathbf{y}_{1:N};\Psi) - A_0(\mathbf{y}_{1:N};\Psi)\delta_\epsilon^* - \mathbf{A}_{10}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T} + \boldsymbol{\mu}\mathbf{W}\mathbf{A}_{13}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T} \\
\quad +(\delta_\epsilon\mathbf{W} + \sigma^2\delta_x)\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\Big] \\
\delta_\epsilon^* = A_2(\mathbf{y}_{1:N};\Psi)^{-1}\Big(\mathbf{A}_5(\mathbf{y}_{1:N};\Psi) - A_0(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^* - \mathbf{A}_8(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T} + \boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T} \\
\quad +\Big(\delta_\epsilon\mathbf{W} - \sigma^2\delta_x\Big)\mathbf{A}_4(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\Big) \\
\delta_x^* = A_3(\mathbf{y}_{1:N};\Psi)^{-1}\Big(\mathbf{A}_8(\mathbf{y}_{1:N};\Psi) - \boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi) - (\delta_\epsilon\mathbf{W} - \sigma^2\delta_x)\mathbf{A}_4(\mathbf{y}_{1:N};\Psi)\Big) \\
\mathbf{W}^* = \Big[\mathbf{A}_{14}(\mathbf{y}_{1:N};\Psi) - \boldsymbol{\mu}^{*T}\mathbf{A}_{10}(\mathbf{y}_{1:N};\Psi) - \delta_\epsilon^*\mathbf{A}_8(\mathbf{y}_{1:N};\Psi) - \mathbf{A}_{16}(\mathbf{y}_{1:N};\Psi) + \boldsymbol{\mu}^{*T}\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{13}(\mathbf{y}_{1:N};\Psi) + \boldsymbol{\mu}^{*T}(\delta_\epsilon - \sigma^2\delta_x)\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi) \\
+\delta_\epsilon^{*T}\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi) + \delta_\epsilon^{*T}(\delta_\epsilon\mathbf{W} - \sigma^2\delta_x)\mathbf{A}_4(\mathbf{y}_{1:N};\Psi)\Big] \\
\quad \times\Big[\mathbf{A}_{17}(\mathbf{y}_{1:N};\Psi) - \mathbf{A}_{18}(\mathbf{y}_{1:N};\Psi) - \mathbf{A}_{18}(\mathbf{y}_{1:N};\Psi)^T + \mathbf{A}_{19}(\mathbf{y}_{1:N};\Psi) + \sigma^2\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi)\Big]^{-1} \\
\sigma^{*2} = \frac{1}{dA_0(\mathbf{y}_{1:N};\Psi)}\Big[A_{20}(\mathbf{y}_{1:N};\Psi) - 2\mathbf{A}_6(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^{*T} - 2\mathbf{A}_5(\mathbf{y}_{1:N};\Psi)\delta_\epsilon^{*T} + A_1(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^*\boldsymbol{\mu}^{*T} + 2A_0(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^*\delta_\epsilon^{*T} \\
\quad +A_2(\mathbf{y}_{1:N};\Psi)\delta_\epsilon^*\delta_\epsilon^{*T} - 2\,Tr\Big\{\mathbf{A}_{14}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\Big\} + 2\mathbf{A}_9(\mathbf{y}_{1:N};\Psi,\Psi^*)\mathbf{W}^T\boldsymbol{\mu}^T \\
\quad +2\mathbf{A}_7(\mathbf{y}_{1:N};\Psi,\Psi)\mathbf{W}^T\delta_\epsilon^T - 2\sigma^2\mathbf{A}_7(\mathbf{y}_{1:N};\Psi,\Psi)\delta_x^T + 2\Big(\mathbf{A}_{10}(\mathbf{y}_{1:N};\Psi) \\
\quad -\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{13}(\mathbf{y}_{1:N};\Psi) - (\delta_\epsilon\mathbf{W} - \sigma^2\delta_x)\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi)\Big)\mathbf{W}^{*T}\boldsymbol{\mu}^{*T} + 2\Big(\mathbf{A}_8(\mathbf{y}_{1:N};\Psi) - \boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}_{1:N}^o;\Psi) \\
\quad -(\delta_\epsilon\mathbf{W} - \sigma^2\delta_x)\mathbf{A}_4(\mathbf{y}_{1:N};\Psi)\Big)\mathbf{W}^{*T}\delta_\epsilon^{*T} + Tr\Big\{\mathbf{A}_{17}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\mathbf{W}^*\Big\} \\
\quad - Tr\Big\{\Big(\mathbf{A}_{18}(\mathbf{y}_{1:N};\Psi) + \mathbf{A}_{18}(\mathbf{y}_{1:N};\Psi)^T\Big)\mathbf{W}^{*T}\mathbf{W}^*\Big\} + Tr\Big\{\mathbf{A}_{19}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\mathbf{W}^*\Big\} + \sigma^2\,Tr\Big\{\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\mathbf{W}^*\Big\}\Big].
\end{cases}
$$

The 2*-dimensional integrals* $\mathbf{A}_i$ *on the hypercube* $[0,1]^2$*, for* $i \in \Big\{0,\ldots,23\Big\}$*, are defined in Subsubsection A.4.1.5 in Appendix.*

Proof. The proof of the Corollary 6 is given in Subsection A.4.1.5 in Appendix A.4. □

Corollary 7. *Under the assumptions of Special Case 3: Generalized t-Student IND PPCA specified in Subsubsection 3.4.1.3, the M-step of the corresponding EM algorithm with the* $Q_{GStS,ind,3}$ *function determined in Corollary 3 is a case-specific variant of the solution to the system of equations defined in Theorem 14 under the problem formulation* $\nabla_{\Psi^*}Q_{GStS,ind,3} = \mathbf{0}$ *for* $\Psi^* = \big[\mathbf{W}^*, \boldsymbol{\mu}^*, \sigma^{*2}\big]$*. The maximizers with respect to* Ψ^* *are given by explicit formulas defined by* 2*-dimensional integration problems on the hypercube* $[0,1]^2$ *as follows*

$$
\begin{cases}
\boldsymbol{\mu}^* = \frac{1}{A_1(\mathbf{y}_{1:N};\Psi)}\Big(\mathbf{A}_6(\mathbf{y}_{1:N};\Psi) - \mathbf{A}_{10}(\mathbf{y}_{1:N};\Psi*)\mathbf{W}^{*T} + \boldsymbol{\mu}\mathbf{W}\mathbf{A}_{13}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\Big) \\
\mathbf{W}^* = \Big(\mathbf{A}_{14}(\mathbf{y}_{1:N};\Psi) + \boldsymbol{\mu}^{*T}\Big(\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{13}(\mathbf{y}_{1:N};\Psi) - \mathbf{A}_{10}(\mathbf{y}_{1:N};\Psi)\Big) - \mathbf{A}_{16}(\mathbf{y}_{1:N};\Psi)\Big) \\
\qquad \times\Big(\sigma^2\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi) + \mathbf{A}_{17}(\mathbf{y}_{1:N};\Psi) - \mathbf{A}_{18}(\mathbf{y}_{1:N};\Psi) - \mathbf{A}_{18}(\mathbf{y}_{1:N};\Psi)^T + \mathbf{A}_{19}(\mathbf{y}_{1:N};\Psi)\Big)^{-1} \\
\sigma^{*2} = \frac{1}{d}A_0(\mathbf{y}_{1:N};\Psi)^{-1}\Big[A_{20}(\mathbf{y}_{1:N};\Psi) - 2\mathbf{A}_6(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^{*T} + A_1(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^*\boldsymbol{\mu}^{*T} - 2\,Tr\Big\{\mathbf{A}_{14}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\Big\} + 2\mathbf{A}_{10}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\boldsymbol{\mu}^{*T} \\
\qquad +2\mathbf{A}_9(\mathbf{y}_{1:N};\Psi)\mathbf{W}^T\boldsymbol{\mu}^T - 2\boldsymbol{\mu}^*\mathbf{W}^*\mathbf{A}_{13}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^T\boldsymbol{\mu}^T + Tr\Big\{\Big(\mathbf{A}_{17}(\mathbf{y}_{1:N};\Psi) - 2\mathbf{A}_{18}(\mathbf{y}_{1:N};\Psi) + \mathbf{A}_{19}(\mathbf{y}_{1:N};\Psi) \\
\quad +\sigma^2\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi)\Big)\mathbf{W}^{*T}\mathbf{W}^*\Big\}\Big].
\end{cases}
$$

The 2*-dimensional integrals* $\mathbf{A}_i$ *on the hypercube* $[0,1]^2$*, for* $i \in \Big\{0,\ldots,22\Big\}$*, are defined in Subsubsection A.4.1.5 in Appendix A.4 with the simplifications of the following integrals*

$$
\mathbf{A}_{16}(\mathbf{y}_{1:N}^o;\Psi)_{d\times k} := \sum_{t=1}^{N}\Big\{H\big(\mathbf{y}_{1:N/t};\Psi\big)\ \mathbf{y}_t^T\boldsymbol{\mu}\mathbf{W}\ \int_0^1\int_0^1 T_\epsilon(s_{\epsilon,t})^2\mathbf{M}_t^{-1}\ m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\Big\},
$$

$$\mathbf{A}_{18}(\mathbf{y}_{1:N};\Psi)_{k\times k} := \sum_{t=1}^{N}\left\{H(\mathbf{y}^{o}_{1:N/t};\Psi)\int_0^1\int_0^1 T_\epsilon(s_{\epsilon,t})^3\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{y}_t^T\boldsymbol{\mu}\mathbf{W}\mathbf{M}_t^{-1}\, m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\right\},$$

$$\mathbf{A}_{19}(\mathbf{y}_{1:N};\Psi)_{k\times k} := \sum_{t=1}^{N}\left\{H(\mathbf{y}_{1:N/t};\Psi)\int_0^1\int_0^1 T_\epsilon(s_{\epsilon,t})^3\mathbf{M}_t^{-1}\mathbf{W}^T\boldsymbol{\mu}^T\boldsymbol{\mu}\mathbf{W}\mathbf{M}_t^{-1}\, m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\right\}.$$

Proof. The proof to Corollary 7 is analogous to the proof to Corollary 6 in Appendix A.4.1.5 when one sets $\mathbf{D}_{\epsilon,t} = T_\epsilon(s_{\epsilon,t})\mathbb{I}_d$ and $\mathbf{D}_{x,t} = T_x(s_{x,t})\mathbb{I}_k$ and $\delta_\epsilon = \delta_x = \mathbf{0}$. □

3.4.2.2. Expectation-Maximisation Algorithm for Probabilistic Principal Component Analysis with Identical and Conditionally Independent Generalized Skew t-Student Distribution

The $Q_{GStS,id}$ function of the E-step for the Generalized Skew t-Student IID PPCA is provided in Theorem 15 whereas the derivation of its maximizers is given in Theorem 16. Moreover, we derive the formulations of the E-step and the M-step of EM algorithms for special cases of the Generalized Skew t-Student IID PPCA specified in Subsubsection 3.4.1.3:

Special Case 1: Grouped t-Student IID PPCA in Corollary 8 and Corollary 12,
Special Case 2: Skew t-Student IID PPCA in Corollary 9 and Corollary 13,
Special Case 3: t-Student IID PPCA in Corollary 10 and Corollary 14.

The E-Step of The EM Algorithm

Theorem 15. *Let the observation process $\mathbf{Y}_t$ follows the Generalized Skew t-Student IID PPCA model defined in Subsubsection 3.4.1.2 in Equation (3.30) and Equation (3.31). Given N realisations of the observation vector $\mathbf{Y}_t$, denoted by $\mathbf{y}_{1:N} = \{\mathbf{y}_1,\ldots,\mathbf{y}_N\}$, the E-step of the corresponding EM algorithm is formulated as following*

$$Q_{GStS,id}(\Psi,\Psi^*) = I_1(\mathbf{y}_{1:N};\Psi,\Psi^*) + I_2(\mathbf{y}_{1:N};\Psi),$$

for the functions $I_1,\ I_2 : \mathbb{R}^{d\times N} \to \mathbb{R}$ defined as the solutions to the integration problems on 2-dimensional unit cube given by

$$I_1(\mathbf{y}_{1:N};\Psi,\Psi^*) := \frac{1}{\pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N})}\int_0^1\int_0^1\left\{\left(\sum_{t=1}^{N}\tilde{w}(\mathbf{y}_t,s_\epsilon,s_x;\Psi,\Psi^*)\right)m(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)\right\}ds_\epsilon\,ds_x,$$

$$I_2(\mathbf{y}_{1:N};\Psi) := -\frac{1}{\pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N})}\int_0^1\int_0^1\left\{\left(\log\left|\frac{\partial T_\epsilon(s_\epsilon)}{\partial s_\epsilon}\right| + \log\left|\frac{\partial T_x(s_x)}{\partial s_x}\right|\right)m(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)\right\}ds_\epsilon\,ds_x,$$

where $T_\epsilon : [0,1] \longrightarrow \mathbb{R}_+^d$ and $T_x : [0,1] \longrightarrow \mathbb{R}_+^k$ are quantile transformations such that

$$T_\epsilon(s) := \left(\frac{\chi^{-1}_{\nu_\epsilon^1}(s)}{\nu_\epsilon^1},\ldots,\frac{\chi^{-1}_{\nu_\epsilon^d}(s)}{\nu_\epsilon^d}\right)_{1\times d} \text{ and } T_x(s) := \left(\frac{\chi^{-1}_{\nu_x^1}(s)}{\nu_x^1},\ldots,\frac{\chi^{-1}_{\nu_x^k}(s)}{\nu_x^k}\right)_{1\times k}.$$

The function $\tilde{w}(\mathbf{y}_t, s_\epsilon, s_x; \Psi, \Psi^*) := w\Big(\mathbf{y}_t, T_\epsilon(s_\epsilon), T_x(s_x); \Psi, \Psi^*\Big)$ *is defined for* $w : \mathbb{R}^d \times \mathbb{R}^d_+ \times \mathbb{R}^k_+ \longrightarrow \mathbb{R}$ *being specified in Lemma A13 in Appendix A.4.6, and let the function* $m : \mathbb{R}^{d\times N} \times [0,1]^2 \longrightarrow \mathbb{R}$ *be defined as*

$$m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi) = \Bigg(\prod_{t=1}^{N} \pi_{\mathbf{Y}_t|S_\epsilon,S_x,\Psi}(\mathbf{y}_t)\Bigg)(\sigma^2)^{\frac{k}{2}}\Big|\mathbf{D}_\epsilon\Big|^{\frac{N}{2}}\Big|\mathbf{D}_x\Big|^{\frac{N}{2}}\Big|\mathbf{M}\Big|^{-\frac{N}{2}}\Big|\mathbf{N}\Big|^{-\frac{N}{2}}$$
$$\times \exp\Bigg\{-\frac{N}{2}\delta_x\Big(\mathbf{D}_x^{-1} - \sigma^2\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\mathbf{N}^{-1}\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}_t^{-1} - \sigma^2\mathbf{M}^{-1}\Big)\delta_x^T\Bigg\},$$

for the probability density function $\pi_{\mathbf{Y}_t|S_\epsilon,S_x,\Psi}(\mathbf{y}_t)$ *which equals* $\pi_{\mathbf{Y}_t|T_\epsilon(S_\epsilon),T_x(S_x),\Psi}(\mathbf{y}_t)$ *specified in Lemma A12 in Appendix A.4.6 and matrices*

$$\mathbf{M} = \sigma^2\mathbf{D}_x + \mathbf{W}^T\mathbf{D}_\epsilon\mathbf{W} \text{ and } \mathbf{N} = \mathbf{D}_\epsilon - \mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon$$

where

$$\mathbf{D}_\epsilon = \begin{pmatrix} \frac{\chi^{-1}_{\nu_\epsilon^1}(S_\epsilon)}{\nu_\epsilon^1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \frac{\chi^{-1}_{\nu_\epsilon^d}(S_\epsilon)}{\nu_\epsilon^d} \end{pmatrix}_{d\times d}, \mathbf{D}_{x,t} = \begin{pmatrix} \frac{\chi^{-1}_{\nu_x^1}(S_x)}{\nu_x^1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \frac{\chi^{-1}_{\nu_x^k}(S_x)}{\nu_x^k} \end{pmatrix}_{k\times k}.$$

Proof. The proof of Theorem 15 is given in Subsection A.4.3.1 in Appendix A.4. □

Corollary 8. *Under the assumptions of Special Case 1: Grouped t-Student PPCA specified in Subsubsection 3.4.1.3, the E-step of the corresponding EM algorithm as a case-specific variant of Theorem 15, is given by*

$$Q_{GStS,id,1}(\Psi, \Psi^*) = \mathbb{E}_{\mathbf{X}_{1:N},\mathbf{U},\mathbf{V}|\mathbf{Y}_{1:N},\Psi}\Big[\log \pi_{\mathbf{Y}_{1:N},\mathbf{X}_{1:N},\mathbf{UV}|\Psi^*}(\mathbf{Y}_{1:N}, \mathbf{X}_{1:N}, \mathbf{U}, \mathbf{V})\Big].$$

The E-step of Grouped t-Student IID PPCA is formulated by using special cases of the functions I_1, $I_2 : \mathbb{R}^{d\times N} \to \mathbb{R}$ *defined from Theorem 15. The case-specific variant of the function* $m : \mathbb{R}^{d\times N} \times [0,1]^2 \longrightarrow \mathbb{R}$ *is given by*

$$m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi) = \Bigg(\prod_{t=1}^{N} \pi_{\mathbf{Y}_t|S_\epsilon,S_x,\Psi}(\mathbf{y}_t)\Bigg)(\sigma^2)^{\frac{k}{2}}\Big|\mathbf{D}_\epsilon\Big|^{\frac{N}{2}}\Big|\mathbf{D}_x\Big|^{\frac{N}{2}}\Big|\mathbf{M}\Big|^{-\frac{N}{2}}\Big|\mathbf{N}\Big|^{-\frac{N}{2}},$$

where, $\pi_{\mathbf{Y}_t|S_\epsilon,S_x,\Psi}(\mathbf{y}_t)$ *is a probability density function of the d-dimensional random vector* $\mathbf{Y}_t|S_\epsilon, Sx, \Psi \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_y)$. *The case specific variant of the function* $\tilde{w} : \mathbb{R}^d \times [0,1]^2 \longrightarrow \mathbb{R}$ *is defined as following*

$$\tilde{w}(\mathbf{y}_t, s_\epsilon, s_x; \Psi, \Psi^*) = -\frac{k+d}{2}\log 2\pi - \frac{d}{2}\log\sigma^{*2} + \frac{1}{2}\log\Big|\mathbf{D}_\epsilon\Big| + \frac{1}{2}\log\Big|\mathbf{D}_x\Big| - \frac{1}{2\sigma^{*2}}(\mathbf{y}_t - \boldsymbol{\mu}^*)\mathbf{D}_\epsilon(\mathbf{y}_t - \boldsymbol{\mu}^*)^T$$
$$+ \frac{1}{\sigma^{*2}}(\mathbf{y}_t - \boldsymbol{\mu})\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon(\mathbf{y}_t - \boldsymbol{\mu}^*)^T - \frac{\sigma^2}{2}Tr\Big\{\mathbf{M}_t^{-1}\Big(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_\epsilon\mathbf{W}^* + \mathbf{D}_x\Big)\Big\}$$
$$- \frac{1}{2}Tr\Big\{\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon(\mathbf{y}_t - \boldsymbol{\mu})^T(\mathbf{y}_t - \boldsymbol{\mu})\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\Big(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_\epsilon\mathbf{W}^* + \mathbf{D}_x\Big)\Big\}.$$

Proof. The proof of Corollary 8 is analogous to the proof of Theorem 15 when one sets $\delta_\epsilon = \delta_x = \delta_\epsilon^* = \delta_x^* = \mathbf{0}$. □

Corollary 9. *Under the assumptions of Special Case 2: Skew t-Student PPCA specified in Subsubsection 3.4.1.3, the E-step of the corresponding EM algorithm as a case-specific variant of Theorem 15, is given by*

$$Q_{GStS,id,2}(\Psi,\Psi^*)=\mathbb{E}_{\mathbf{X}_{1:N},U,V|\mathbf{Y}_{1:N},\Psi}\Big[\log\pi_{\mathbf{Y}_{1:N},\mathbf{X}_{1:N},U,V|\Psi^*}(\mathbf{Y}_{1:N},\mathbf{X}_{1:N},U,V)\Big].$$

The E-step of Skew t-Student IID PPCA is formulated using special cases of the functions $I_1,\ I_2:\mathbb{R}^d\to\mathbb{R}$ *defined analogously as in Theorem 15.*

Let us denote the quantile transformations $T_\epsilon(s):=\frac{\chi^{-1}_{\nu_\epsilon}(s)}{\nu_\epsilon}$ *and* $T_x(s):=\frac{\chi^{-1}_{\nu_x}(s)}{\nu_x}$. *The case-specific variant of the function* $m:\mathbb{R}^d\times[0,1]^2\longrightarrow\mathbb{R}$ *is given by*

$$\begin{aligned}m(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)=&\left(\prod_{t=1}^N\pi_{\mathbf{Y}_t|S_\epsilon,S_x,\Psi}(\mathbf{y}_t)\right)(\sigma^2)^{\frac{k}{2}}\left(T_\epsilon(s_\epsilon)\right)^{\frac{dN}{2}}\left(T_x(s_x)\right)^{\frac{kN}{2}}\Big|\mathbf{M}\Big|^{-\frac{N}{2}}\Big|\mathbf{N}\Big|^{-\frac{N}{2}}\\&\times\exp\left\{-\frac{1}{2}\delta_x\Big(T_x(s_x)^{-1}\mathbb{I}_k-\sigma^2T_\epsilon(s_\epsilon)^2\mathbf{M}^{-1}\mathbf{W}^T\mathbf{N}^{-1}\mathbf{W}\mathbf{M}_t^{-1}-\sigma^2\mathbf{M}^{-1}\Big)\delta_x^T\right\},\end{aligned}$$

for $\pi_{\mathbf{Y}_t|S_\epsilon,S_x,\Psi}(\mathbf{y}_t)$ *being a probability density function of the d-dimensional Gaussian random vector* $\mathbf{Y}_t|S_\epsilon,S_x,\Psi$ *with first and second central moments given by*

$$\boldsymbol{\mu}_{y,t}=\boldsymbol{\mu}+T_\epsilon(s_\epsilon)^{-1}\boldsymbol{\delta}_\epsilon+\sigma^2T_\epsilon(s_\epsilon)\delta_x\mathbf{M}^{-1}\mathbf{W}^T\mathbf{N}^{-1}\ \textit{and}\ \boldsymbol{\Sigma}_y=\sigma^2\mathbf{N}^{-1}$$

for matrices matrices

$$\mathbf{M}=\sigma^2T_x(s_x)\mathbb{I}_k+T_\epsilon(s_\epsilon)\mathbf{W}^T\mathbf{W}\ \textit{and}\ \mathbf{N}=T_\epsilon(s_\epsilon)\mathbb{I}_d-T_\epsilon(s_\epsilon)^2\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^T,$$

where $\mathbf{D}_\epsilon=T_\epsilon(s_\epsilon)\mathbb{I}_d$ *and* $\mathbf{D}_x=T_x(s_x)\mathbb{I}_k$. *The case-specific variant of the function* $\tilde{w}:\mathbb{R}^d\times[0,1]^2\longrightarrow\mathbb{R}$ *is defined as*

$$\begin{aligned}\tilde{w}(\mathbf{y}_t,s_\epsilon,s_x;\Psi,\Psi^*)=&-\frac{k+d}{2}\log2\pi-\frac{d}{2}\log\sigma^{*2}+\frac{d}{2}\log T_\epsilon(s_\epsilon)+\frac{k}{2}\log T_x(s_x)-\frac{T_x(s_x)^{-1}}{2}\delta_x^*\delta_x^{*T}\\&-\frac{T_\epsilon(s_\epsilon)}{2\sigma^{*2}}\Big(\mathbf{y}_t-\boldsymbol{\mu}^*-T_\epsilon(s_\epsilon)^{-1}\boldsymbol{\delta}_\epsilon^*\Big)\Big(\mathbf{y}_t-\boldsymbol{\mu}^*-T_\epsilon(s_\epsilon)^{-1}\boldsymbol{\delta}_\epsilon^*\Big)^T-\frac{\sigma^2}{2}Tr\left\{\mathbf{M}_t^{-1}\Big(\frac{T_\epsilon(s_\epsilon)}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{W}^*+T_x(s_x)\mathbb{I}_k\Big)\right\}\\&+\left(T_\epsilon(s_\epsilon)(\mathbf{y}_t-\boldsymbol{\mu})\mathbf{W}-\delta_\epsilon\mathbf{W}+\sigma^2\delta_x\right)\mathbf{M}^{-1}\Big(\frac{T_\epsilon(s_\epsilon)}{\sigma^{*2}}\mathbf{W}^{*T}(\mathbf{y}_t-\boldsymbol{\mu}^*)^T-\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\boldsymbol{\delta}_\epsilon^{*T}+\boldsymbol{\delta}_x^{*T}\Big)\\&-\frac{T_\epsilon(s_\epsilon)^2}{2}Tr\left\{\mathbf{M}^{-1}\mathbf{W}^T\mathbf{y}_t^T\mathbf{y}_t\mathbf{W}\mathbf{M}^{-1}\Big(\frac{T_\epsilon(s_\epsilon)}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{W}^*+T_x(s_x)\mathbb{I}_k\Big)\right\}\\&+\frac{T_\epsilon(s_\epsilon)}{2}Tr\left\{\mathbf{M}^{-1}\mathbf{W}^T\mathbf{y}_t^T\Big(T_\epsilon(s_\epsilon)\boldsymbol{\mu}\mathbf{W}+\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}^{-1}\Big(\frac{T_\epsilon(s_\epsilon)}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{W}^*+T_x(s_x)\mathbb{I}_k\Big)\right\}\\&+\frac{T_\epsilon(s_\epsilon)}{2}Tr\left\{\mathbf{M}^{-1}\Big(T_\epsilon(s_\epsilon)\boldsymbol{\mu}\mathbf{W}+\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\Big)^T\mathbf{y}_t\mathbf{W}\mathbf{M}^{-1}\Big(\frac{T_\epsilon(s_\epsilon)}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{W}^*+T_x(s_x)\mathbb{I}_k\Big)\right\}\\&-\frac{1}{2}Tr\left\{\mathbf{M}^{-1}\Big(T_\epsilon(s_\epsilon)\boldsymbol{\mu}\mathbf{W}+\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\Big)^T\Big(T_\epsilon(s_\epsilon)\boldsymbol{\mu}\mathbf{W}+\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}^{-1}\Big(\frac{T_\epsilon(s_\epsilon)}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{W}^*+T_x(s_x)\mathbb{I}_k\Big)\right\}.\end{aligned}$$

Proof. The proof of Corollary 9 is analogous to the proof of Theorem 15 when one sets $\mathbf{D}_\epsilon=T_\epsilon(s_\epsilon)\mathbb{I}_d$ and $\mathbf{D}_x=T_x(s_x)\mathbb{I}_k$. □

Corollary 10. *Under the assumptions of Special Case 3: Generalized t-Student PPCA specified in Subsubsection 3.4.1.3, the E-step of the corresponding EM algorithm as a case-specific variant of Theorem 15, is given by*

$$Q_{GStS,id,3}(\Psi, \Psi^*) = \mathbb{E}_{\mathbf{X}_{1:N},U,V|\mathbf{Y}_{1:N},\Psi}\Big[\log \pi_{\mathbf{Y}_{1:N},\mathbf{X}_{1:N},U,V|\Psi^*}(\mathbf{Y}_{1:N}, \mathbf{X}_{1:N}, U, V)\Big].$$

The E-step of Generalized t-Student IID PPCA is formulated by using special cases of the functions $I_1,\ I_2 : \mathbb{R}^d \to \mathbb{R}$ defined analogously as in Theorem 15.

Let us denote the quantile transformations $T_\epsilon(s) := \frac{\chi^{-1}_{\nu_\epsilon}(s)}{\nu_\epsilon}$ and $T_x(s) := \frac{\chi^{-1}_{\nu_x}(s)}{\nu_x}$. The case-specific variant of the function $m : \mathbb{R}^d \times [0,1]^2 \longrightarrow \mathbb{R}$ is given by

$$m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi) = \left(\prod_{t=1}^{N} \pi_{\mathbf{Y}_t|S_\epsilon,S_x,\Psi}(\mathbf{y}_t)\right)(\sigma^2)^{\frac{k}{2}}\left(T_\epsilon(s_\epsilon)\right)^{\frac{dN}{2}}\left(T_x(s_x)\right)^{\frac{kN}{2}}\left|\mathbf{M}\right|^{-\frac{N}{2}}\left|\mathbf{N}\right|^{-\frac{N}{2}},$$

where $\pi_{\mathbf{Y}_t|S_\epsilon,S_x,\Psi}(\mathbf{y}_t)$ is a probability density function of the d-dimensional random vector $\mathbf{Y}_t|S_\epsilon, Sx, \Psi \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_y)$ for matrices

$$\mathbf{M} = \sigma^2 T_x(s_x)\mathbb{I}_k + T_\epsilon(s_\epsilon)\mathbf{W}^T\mathbf{W} \ \text{ and } \mathbf{N} = T_\epsilon(s_\epsilon)\mathbb{I}_d - T_\epsilon(s_\epsilon)^2\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^T$$

where $\mathbf{D}_\epsilon = T_\epsilon(s_\epsilon)\mathbb{I}_d$ and $\mathbf{D}_x = T_x(s_x)\mathbb{I}_k$. The case-specific variant of the function $\tilde{w} : \mathbb{R}^d \times [0,1]^2 \longrightarrow \mathbb{R}$ is defined as

$$\begin{aligned}\tilde{w}(\mathbf{y}_t, s_\epsilon, s_x; \Psi, \Psi^*) = &-\frac{k+d}{2}\log 2\pi - \frac{d}{2}\log\sigma^{*2} + \frac{d}{2}\log T_\epsilon(s_\epsilon) + \frac{k}{2}\log T_x(s_x) - \frac{T_\epsilon(s_\epsilon)}{2\sigma^{*2}}(\mathbf{y}_t - \boldsymbol{\mu}^*)(\mathbf{y}_t - \boldsymbol{\mu}^*)^T \\ &+ \frac{T_\epsilon(s_\epsilon)^2}{\sigma^{*2}}(\mathbf{y}_t - \boldsymbol{\mu})\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}(\mathbf{y}_t - \boldsymbol{\mu}^*)^T - \frac{\sigma^2}{2}Tr\left\{\mathbf{M}_t^{-1}\Big(\frac{T_\epsilon(s_\epsilon)}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{W}^* + T_x(s_x)\mathbb{I}_k\Big)\right\} \\ &- \frac{T_\epsilon(s_\epsilon)^2}{2}Tr\left\{\mathbf{M}^{-1}\mathbf{W}^T(\mathbf{y}_t - \boldsymbol{\mu})^T(\mathbf{y}_t - \boldsymbol{\mu})\mathbf{W}\mathbf{M}^{-1}\Big(\frac{T_\epsilon(s_\epsilon)}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{W}^* + T_x(s_x)\mathbb{I}_k\Big)\right\}.\end{aligned}$$

Proof. The proof of Corollary 10 is analogous to the proof of Theorem 15 when one sets $\mathbf{D}_\epsilon = T_\epsilon(s_\epsilon)\mathbb{I}_d$ and $\mathbf{D}_x = T_x(s_x)\mathbb{I}_k$ and $\boldsymbol{\delta}_\epsilon = \boldsymbol{\delta}_x = \boldsymbol{\delta}^*_\epsilon = \boldsymbol{\delta}^*_x = \mathbf{0}$. □

M-Step of The EM Algorithm

Let us define an integral operator I which is equivalent to the operator from Equation (3.39). Given a function $b : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^k \to \mathbb{R}^{n\times n'}$ for some integer numbers n and n', the integral operator I is the following

$$I(b)(\mathbf{y}_{1:N})_{n\times n'} := \int_0^1 \int_0^1 \sum_{t=1}^{N} b(\mathbf{y}_t, s_\epsilon, s_{x,t})\ m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\ ds_\epsilon\ ds_x \tag{3.42}$$

for the function m specified in Theorem 15. We remark that the co-domain of I depends on the dimensionality of the co-domain of the function b. The operator I is used in the derivations of the M-step of the EM Algorithm for Generalized Skew t-Student IID PPCA.

Theorem 16. *Let the observation vector* $\mathbf{Y}_t$ *be modelled as in Equation* (2.7) *with the latent processes* $\mathbf{X}_t$, $\boldsymbol{\epsilon}_t$, $\mathbf{U}$ *and* $\mathbf{V}$ *following the assumptions of Generalized Skew t-Student IID PPCA given in Subsection* 3.4.2.2 *in Equation* (3.30) *and Equation* (3.31). *The maximizers of the E-step of the Expectation-Maximisation algorithm for Generalized Skew t-Student IID PPCA derived in Theorem* 15 *with respect to the vector of parameters* Ψ^* *are the solutions to the system of equations determined by* $\nabla_{\Psi^*} Q_{GStS,id}(\Psi, \Psi^*) = \mathbf{0}$ *as following*

$$\begin{cases} I(\frac{\partial}{\partial \boldsymbol{\mu}^*}\tilde{w})(\mathbf{y}_{1:N})_{1\times d} = \mathbf{0} \\ I(\frac{\partial}{\partial \mathbf{W}^*}\tilde{w})(\mathbf{y}_{1:N})_{d\times k} = \mathbf{0} \\ I(\frac{\partial}{\partial \sigma^{*2}}\tilde{w})(\mathbf{y}_{1:N})_{1\times 1} = 0 \\ I(\frac{\partial}{\partial \boldsymbol{\delta}_\epsilon^*}\tilde{w})(\mathbf{y}_{1:N})_{1\times d} = \mathbf{0} \\ I(\frac{\partial}{\partial \boldsymbol{\delta}_x^*}\tilde{w})(\mathbf{y}_{1:N})_{1\times k} = \mathbf{0}, \end{cases}$$

for the integral operator I *defined in Equation* (3.42) *and the sum over of the gradient of the function*$\tilde{w}$ *defined in Theorem* 15 *being derived as*

$$\sum_{t=1}^{N} \frac{\partial}{\partial \boldsymbol{\mu}^*}\tilde{w}(\mathbf{y}_t, s_\epsilon, s_x; \Psi, \Psi^*) = \frac{N}{\sigma^{*2}}\Big[\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{D}_\epsilon - \boldsymbol{\mu}^*\mathbf{D}_\epsilon - \boldsymbol{\delta}_\epsilon^* - \bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon + \boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon + \boldsymbol{\delta}_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon - \sigma^2\boldsymbol{\delta}_x\mathbf{M}^{-1}$$

$$\begin{aligned}\sum_{t=1}^{N} \frac{\partial}{\partial \mathbf{W}^*}\tilde{w}(\mathbf{y}_t, s_\epsilon, s_x; \Psi, \Psi^*) = \frac{N}{\sigma^{*2}}\Big[&\mathbf{D}_\epsilon\bar{\mathbb{S}}(\mathbf{y}_{1:N})\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1} - \Big(\mathbf{D}_\epsilon\boldsymbol{\mu}^{*T} + \boldsymbol{\delta}_\epsilon^{*T}\Big)\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1} \\ &+ \Big(\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{D}_\epsilon - \boldsymbol{\mu}^*\mathbf{D}_\epsilon - \boldsymbol{\delta}_\epsilon^*\Big)^T\Big(-\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W} - \boldsymbol{\delta}_\epsilon\mathbf{W} + \sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}^{-1} - \mathbf{D}_\epsilon\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\bar{\mathbf{S}}(\mathbf{y}_{1:N})\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1} \\ &+ \mathbf{D}_\epsilon\mathbf{W}^*\mathbf{M}^{-1}\Big(\mathbf{W}^T\mathbf{D}_\epsilon\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})^T(\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W} + \boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x) + (\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W} + \boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x)^T\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{D}_\epsilon\mathbf{W}\Big)\mathbf{M}^{-1} \\ &- \mathbf{D}_\epsilon\mathbf{W}^*\mathbf{M}^{-1}\Big(\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W} + \boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\Big)^T(\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W} + \boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x)\mathbf{M}^{-1} - \sigma^2\mathbf{D}_\epsilon\mathbf{W}^*\mathbf{M}^{-1}\Big],\end{aligned}$$

$$\begin{aligned}\sum_{t=1}^{N} \frac{\partial}{\partial \sigma^{*2}}\tilde{w}(\mathbf{y}_t, s_\epsilon, s_x; \Psi, \Psi^*) = -\frac{dN}{2\sigma^{*2}} - \frac{N}{2(\sigma^{*2})^2}\Big[&- Tr\Big\{\bar{\mathbf{S}}(\mathbf{y}_{1:N})\mathbf{D}_\epsilon\Big\} + 2\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\Big(\mathbf{D}_\epsilon\boldsymbol{\mu}^{*T} + \boldsymbol{\delta}_\epsilon^{*T}\Big) \\ &- \boldsymbol{\mu}^*\mathbf{D}_\epsilon\boldsymbol{\mu}^{*T} - 2\boldsymbol{\mu}^*\boldsymbol{\delta}_\epsilon^{*T} - \boldsymbol{\delta}_\epsilon^*\mathbf{D}_\epsilon^{-1}\boldsymbol{\delta}_\epsilon^{*T} + 2\,Tr\Big\{\bar{\mathbf{S}}(\mathbf{y}_{1:N})\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon\Big\} \\ &- 2(\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{D}_\epsilon\mathbf{W} - \boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W} - \boldsymbol{\delta}_\epsilon\mathbf{W} + \sigma^2\boldsymbol{\delta}_x)\mathbf{M}^{-1}\mathbf{W}^{*T}(\boldsymbol{\mu}^*\mathbf{D}_\epsilon + \boldsymbol{\delta}_\epsilon^*)^T \\ &+ 2\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{D}_\epsilon\mathbf{W}^*\mathbf{M}^{-1}(-\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W} - \boldsymbol{\delta}_\epsilon\mathbf{W} + \sigma^2\boldsymbol{\delta}_x)^T - Tr\Big\{\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\bar{\mathbf{S}}(\mathbf{y}_{1:N})\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon\mathbf{W}^*\Big\} \\ &+ Tr\Big\{\mathbf{M}^{-1}\Big(\mathbf{W}^T\mathbf{D}_\epsilon\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})^T(\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W} + \boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x) + (\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W} + \boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x)^T\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{D}_\epsilon\mathbf{W}\Big)\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon\mathbf{W}^*\Big\} \\ &- Tr\Big\{\mathbf{M}^{-1}(\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W} + \boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x)^T(\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W} + \boldsymbol{\delta}_\epsilon\mathbf{W}\sigma^2\boldsymbol{\delta}_x)\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon\mathbf{W}^*\Big\} - \sigma^2\,Tr\Big\{\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon\mathbf{W}^*\Big\}\Big],\end{aligned}$$

$$\sum_{t=1}^{N} \frac{\partial}{\partial \boldsymbol{\delta}_\epsilon^*}\tilde{w}(\mathbf{y}_t, s_\epsilon, s_x; \Psi, \Psi^*) = \frac{N}{\sigma^{*2}}\Big[\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N}) - \boldsymbol{\mu}^* - \boldsymbol{\delta}_\epsilon^*\mathbf{D}_\epsilon^{-1} - \bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T} + \boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T} + \boldsymbol{\delta}_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T} - \sigma^2\boldsymbol{\delta}_x\mathbf{M}^{-1}\mathbf{W}^{*T}\Big],$$

$$\sum_{t=1}^{N} \frac{\partial}{\partial \boldsymbol{\delta}_x^*}\tilde{w}(\mathbf{y}_t, s_\epsilon, s_x; \Psi, \Psi^*) = N\Big[-\boldsymbol{\delta}_x^*\mathbf{D}_x^{-1} + \bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1} - \boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1} - \boldsymbol{\delta}_\epsilon\mathbf{W}\mathbf{M}^{-1} + \sigma^2\boldsymbol{\delta}_x\mathbf{M}^{-1}\Big],$$

where $\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})_{1\times d} = \frac{1}{N}\sum_{t=1}^{N}\mathbf{y}_t$ *and* $\bar{\mathbf{S}}(\mathbf{y}_{1:N})_{d\times d} = \frac{1}{N}\sum_{t=1}^{N}\mathbf{y}_t^T\mathbf{y}_t$.

Proof. The proof of Theorem 16 is given inSubsection A.4.3.2 in Appendix A.4. □

Corollary 11. *The solution to the system of equation from Theorem 16 which determines the maximizers of the function $Q_{GStS,id}$ with respect to the parameters $\boldsymbol{\mu}^*, \boldsymbol{\delta}_\epsilon^*, \boldsymbol{\delta}_x^*, \sigma^2$ are given by explicit formulas defined by 2-dimensional integration problems on the hypercube $[0,1]^2$ being the following*

$$
\begin{cases}
\boldsymbol{\mu}^* = \Big[\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\Big(\mathbf{A}_1(\mathbf{y}_{1:N};\Psi) - \mathbf{A}_7(\mathbf{y}_{1:N};\Psi,\Psi^*)\Big) - \boldsymbol{\delta}_\epsilon^* A_0(\mathbf{y}_{1:N};\Psi) + \boldsymbol{\mu}\mathbf{A}_7(\mathbf{y}_{1:N};\Psi,\Psi^*) \\
\quad + \Big(\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\Big)\mathbf{A}_6(\mathbf{y}_{1:N};\Psi,\Psi^*)^T\Big]\mathbf{A}_1(\mathbf{y}_{1:N};\Psi)^{-1}, \\
\boldsymbol{\delta}_\epsilon^* = \Big[\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\Big(A_0(\mathbf{y}_{1:N};\Psi) - \mathbf{A}_5(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\Big) - \boldsymbol{\mu}^* A_0(\mathbf{y}_{1:N};\Psi) + \boldsymbol{\mu}\mathbf{A}_5(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\Big) \\
\quad + (\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x)\mathbf{A}_4(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\Big]\mathbf{A}_2(\mathbf{y}_{1:N};\Psi)^{-1}, \\
\boldsymbol{\delta}_x^* = \Big[\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{A}_5(\mathbf{y}_{1:N}) - \boldsymbol{\mu}\mathbf{A}_5(\mathbf{y}_{1:N}) - (\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x)\mathbf{A}_4(\mathbf{y}_{1:N};\Psi)\Big]\mathbf{A}_3(\mathbf{y}_{1:N},\Psi)^{-1}, \\
\sigma^{*2} = \frac{1}{d\ A_0(\mathbf{y}_{1:N};\Psi)}\Big[Tr\Big\{\bar{\mathbf{S}}(\mathbf{y}_{1:N})\mathbf{A}_1(\mathbf{y}_{1:N};\Psi)\Big\} - 2\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{A}_1(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^* - 2\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})A_0(\mathbf{y}_{1:N};\Psi)\boldsymbol{\delta}_\epsilon^* + \boldsymbol{\mu}^*\mathbf{A}_1(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^{*T} \\
\quad + 2\boldsymbol{\mu}^*\boldsymbol{\delta}_\epsilon^{*T}A_0(\mathbf{y}_{1:N};\Psi) + \boldsymbol{\delta}_\epsilon^*\mathbf{A}_2(\mathbf{y}_{1:N};\Psi)\boldsymbol{\delta}_\epsilon^{*T} - 2\,Tr\Big\{\bar{\mathbf{S}}(\mathbf{y}_{1:N})\mathbf{A}_7(\mathbf{y}_{1:N};\Psi,\Psi^*)\Big\} + 2\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{A}_7(\mathbf{y}_{1:N};\Psi,\Psi^*)\boldsymbol{\mu}^{*T} \\
\quad - 2\boldsymbol{\mu}\mathbf{A}_7(\mathbf{y}_{1:N};\Psi,\Psi^*)\boldsymbol{\mu}^{*T} - 2(\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x)\mathbf{A}_6(\mathbf{y}_{1:N};\Psi)^T\boldsymbol{\mu}^* + 2\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{A}_5(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\boldsymbol{\delta}_\epsilon^{*T} \\
\quad - 2\boldsymbol{\mu}\mathbf{A}_5(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\boldsymbol{\delta}_\epsilon^{*T} - 2(\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x)\mathbf{A}_4(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\boldsymbol{\delta}_\epsilon^{*T} + 2\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{A}_7(\mathbf{y}_{1:N};\Psi,\Psi^*)^T\boldsymbol{\mu}^T \\
\quad + 2\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{A}_6(\mathbf{y}_{1:N};\Psi,\Psi^*)(\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x)^T + \sigma^2\,Tr\Big\{\mathbf{A}_6(\mathbf{y}_{1:N};\Psi,\Psi^*)\mathbf{W}^*\Big\} + Tr\Big\{\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi,\Psi^*)^T\mathbf{W}^*\Big\} \\
\quad - Tr\Big\{\mathbf{A}_{12}(\mathbf{y}_{1:N};\Psi,\Psi^*)^T\mathbf{W}^*\Big\} - Tr\Big\{\mathbf{A}_{13}(\mathbf{y}_{1:N};\Psi,\Psi^*)^T\mathbf{W}^*\Big\} + Tr\Big\{\mathbf{A}_{14}(\mathbf{y}_{1:N};\Psi,\Psi^*)^T\mathbf{W}^*\Big\}\Big].
\end{cases}
$$

The solutions with respect to $\boldsymbol{\mu}^*$, $\boldsymbol{\delta}_\epsilon^*$ and σ^{*2} depend on the parameter $\mathbf{W}^*$. The maximizer of the function $Q_{GStS,id}$ with respect to $\mathbf{W}^*$ is determined as follows

$$
f(\mathbf{W}^*;\Psi,\Psi^*) = \mathbf{0}, \tag{3.43}
$$

for the function $f : \mathbb{R}^{d\times k} \to \mathbb{R}^{d\times k}$ such that $f(\mathbf{W}^*;\Psi,\Psi^*) := \frac{\partial Q_{GStS,id}(\Psi,\Psi^*)}{\partial \mathbf{W}^*}$ being equal to

$$
\begin{aligned}
f(\mathbf{W}^*;\Psi,\Psi^*) = {} & \mathbf{A}_8(\mathbf{y}_{1:N};\Psi) - \mathbf{A}_9(\mathbf{y}_{1:N};\Psi,\Psi^*) - \boldsymbol{\delta}_\epsilon^{*T}\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{A}_5(\mathbf{y}_{1:N};\Psi) - \mathbf{A}_{10}(\mathbf{y}_{1:N};\Psi,\Psi^*) \\
& + \boldsymbol{\delta}_\epsilon^{*T}\boldsymbol{\mu}\mathbf{A}_5(\mathbf{y}_{1:N};\Psi) + \boldsymbol{\delta}_\epsilon^{*T}(\boldsymbol{\delta}\mathbf{W} - \sigma^2\mathbf{W})\mathbf{A}_4(\mathbf{y}_{1:N};\Psi) - \mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi,\Psi^*) + \mathbf{A}_{12}(\mathbf{y}_{1:N};\Psi,\Psi^*) \\
& + \mathbf{A}_{13}(\mathbf{y}_{1:N};\Psi,\Psi^*) - \mathbf{A}_{14}(\mathbf{y}_{1:N};\Psi,\Psi^*) - \sigma^2\mathbf{A}_6(\mathbf{y}_{1:N};\Psi,\Psi^*).
\end{aligned}
$$

The function f is linear with respect to the parameters $\boldsymbol{\mu}^*$ and $\boldsymbol{\delta}_\epsilon^*$. The 2-dimensional integrals $\mathbf{A}_i$ on the hypercube $[0,1]^2$, for $i \in \{0,\ldots,14\}$, are defined in Subsubsection A.4.3.3 in Appendix A.4.

Proof. The proof of the Corollary 11is given in Subection A.4.3.3 in Appendix A.4. □

Corollary 12. *Under the assumptions of Special Case 1: Grouped t-Student IID PPCA specified in Subsubsection 3.4.1.3, the M-step of the corresponding EM algorithm with the $Q_{GStS,id,1}$ function determined in Corollary 8 is a case-specific variant of the solution to the system of equations defined in Theorem 16 under the problem formulation $\nabla_{\Psi^*} Q_{GStS,id,1} = \mathbf{0}$ for $\Psi^* = \big[\mathbf{W}^*, \boldsymbol{\mu}^*, \sigma^{*2}\big]$. The maximizers with respect to Ψ^* are given by explicit formulas defined by 2-dimensional integration problems on the hypercube $[0,1]^2$ as follows*

$$
\begin{cases}
\boldsymbol{\mu}^* = \Big[\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\Big(\mathbf{A}_1(\mathbf{y}_{1:N};\Psi) - \mathbf{A}_7(\mathbf{y}_{1:N};\Psi,\Psi^*)\Big) + \boldsymbol{\mu}\mathbf{A}_7(\mathbf{y}_{1:N};\Psi,\Psi^*)\Big]\mathbf{A}_1(\mathbf{y}_{1:N};\Psi)^{-1} \\
\sigma^{*2} = \frac{1}{d\ A_0(\mathbf{y}_{1:N};\Psi)}\Big[Tr\Big\{\bar{\mathbf{S}}(\mathbf{y}_{1:N})\mathbf{A}_1(\mathbf{y}_{1:N};\Psi)\Big\} - 2\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{A}_1(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^* + \boldsymbol{\mu}^*\mathbf{A}_1(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^{*T} - 2\,Tr\Big\{\bar{\mathbf{S}}(\mathbf{y}_{1:N})\mathbf{A}_7(\mathbf{y}_{1:N};\Psi,\Psi^*)\Big\} \\
\quad + 2\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{A}_7(\mathbf{y}_{1:N};\Psi,\Psi^*)\boldsymbol{\mu}^{*T} - 2\boldsymbol{\mu}\mathbf{A}_7(\mathbf{y}_{1:N};\Psi,\Psi^*)\boldsymbol{\mu}^{*T} + 2\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{A}_7(\mathbf{y}_{1:N};\Psi,\Psi^*)^T\boldsymbol{\mu}^T + \sigma^2\,Tr\Big\{\mathbf{A}_6(\mathbf{y}_{1:N};\Psi,\Psi^*)\mathbf{W}^*\Big\} \\
\quad + Tr\Big\{\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi,\Psi^*)^T\mathbf{W}^*\Big\} - Tr\Big\{\mathbf{A}_{12}(\mathbf{y}_{1:N};\Psi,\Psi^*)^T\mathbf{W}^*\Big\} - Tr\Big\{\mathbf{A}_{13}(\mathbf{y}_{1:N};\Psi,\Psi^*)^T\mathbf{W}^*\Big\} + Tr\Big\{\mathbf{A}_{14}(\mathbf{y}_{1:N};\Psi,\Psi^*)^T\mathbf{W}^*\Big\}\Big].
\end{cases}
$$

The solutions with respect to $\boldsymbol{\mu}^*$ and σ^{*2} depend on the value of the parameter $\mathbf{W}^*$. The maximizer of the function $Q_{GStS,id,1}$ with respect to $\mathbf{W}^*$ is determined by the Equation 3.43 for the case-specific formulation of the function f given by

$$f(\mathbf{W}^*;\Psi,\Psi^*) = \mathbf{A}_8(\mathbf{y}_{1:N};\Psi) - \mathbf{A}_9(\mathbf{y}_{1:N};\Psi,\Psi^*) - \mathbf{A}_{10}(\mathbf{y}_{1:N};\Psi,\Psi^*) - \mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi,\Psi^*) + \mathbf{A}_{12}(\mathbf{y}_{1:N};\Psi,\Psi^*) + \mathbf{A}_{13}(\mathbf{y}_{1:N};\Psi,\Psi^*) - \mathbf{A}_{14}(\mathbf{y}_{1:N};\Psi,\Psi^*) - \sigma^2\mathbf{A}_6(\mathbf{y}_{1:N};\Psi,\Psi^*),$$

and is linear with respect to the parameters $\boldsymbol{\mu}^*$. The 2-dimensional integrals $\mathbf{A}_i$ on the hypercube $[0,1]^2$, for $i \in \{0,\ldots,22\}$, are defined in Subsubsection A.4.1.4 in Appendix A.4 given the following simplifications

$$\begin{aligned}
\mathbf{A}_{10}(\mathbf{y}_{1:N};\Psi,\Psi^*)_{d\times k} &:= \int_0^1\int_0^1 \mathbf{D}_\epsilon\left(\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi) - \boldsymbol{\mu}^*\right)^T\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\, m(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)\, ds_\epsilon\, ds_x,\\
\mathbf{A}_{12}(\mathbf{y}_{1:N};\Psi,\Psi^*)_{k\times d} &:= \int_0^1\int_0^1 \mathbf{D}_\epsilon\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)^T\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\, m(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)\, ds_\epsilon\, ds_x,\\
\mathbf{A}_{13}(\mathbf{y}_{1:N};\Psi,\Psi^*)_{k\times d} &:= \int_0^1\int_0^1 \mathbf{D}_\epsilon\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\boldsymbol{\mu}^T\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\, m(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)\, ds_\epsilon\, ds_x,\\
\mathbf{A}_{14}(\mathbf{y}_{1:N};\Psi,\Psi^*)_{k\times d} &:= \int_0^1\int_0^1 \mathbf{D}_\epsilon\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\boldsymbol{\mu}^T\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\, m(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)\, ds_\epsilon\, ds_x.
\end{aligned}$$

Proof. The proof of Corollary 12 is analogous to the proof to Theorem 16 and Corollary 11 when one sets $\delta_\epsilon = \delta_x = \delta_\epsilon^* = \delta_x^* = \mathbf{0}$. □

Corollary 13. *Under the assumptions of Special Case 2: Skew t-Student IID PPCA specified in Subsubsection 3.4.1.3, the M-step of the corresponding EM algorithm with the $Q_{GStS,id,2}$ function determined in Corollary 9 is a case-specific variant of the solution to the system of equations defined in Theorem 16 under the problem formulation $\nabla_{\Psi^*}Q_{GStS,id,2} = \mathbf{0}$ for $\Psi^* = \left[\mathbf{W}^*,\boldsymbol{\mu}^*,\boldsymbol{\delta}_\epsilon,\boldsymbol{\delta}_X,\sigma^{*2}\right]$. The maximizers with respect to Ψ^* are given by explicit formulas defined by 2-dimensional integration problems on the hypercube $[0,1]^2$ as follows*

$$\begin{cases}
\boldsymbol{\mu}^* = A_1(\mathbf{y}_{1:N};\Psi)^{-1}\Big[\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\Big(A_1(\mathbf{y}_{1:N};\Psi) - \mathbf{W}\mathbf{A}_6(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\Big) - \boldsymbol{\delta}_\epsilon^* A_0(\mathbf{y}_{1:N};\Psi) + \boldsymbol{\mu}\mathbf{W}\mathbf{A}_6(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\\
\quad +(\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x)\mathbf{A}_5(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\Big]\\
\boldsymbol{\delta}_\epsilon^* = A_2(\mathbf{y}_{1:N};\Psi)^{-1}\Big[\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\Big(A_0(\mathbf{y}_{1:N};\Psi) - \mathbf{W}\mathbf{A}_5(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\Big) - A_0(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^* + \boldsymbol{\mu}\mathbf{W}\mathbf{A}_5(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\\
\quad +(\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x)\mathbf{A}_4(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\Big]\\
\boldsymbol{\delta}_x^* = A_3(\mathbf{y}_{1:N};\Psi)^{-1}\Big[(\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N}) - \boldsymbol{\mu})\mathbf{A}_5(\mathbf{y}_{1:N};\Psi) - (\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x)\mathbf{A}_4(\mathbf{y}_{1:N};\Psi)\Big]\\
\mathbf{W}^* = \Big[\bar{\mathbf{S}}(\mathbf{y}_{1:N})\mathbf{W}\mathbf{A}_6(\mathbf{y}_{1:N};\Psi) - \boldsymbol{\mu}^{*T}\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{W}\mathbf{A}_6(\mathbf{y}_{1:N};\Psi) - \boldsymbol{\delta}_\epsilon^{*T}\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{W}\mathbf{A}_5(\mathbf{y}_{1:N};\Psi) - \bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})^T\boldsymbol{\mu}\mathbf{W}\mathbf{A}_6(\mathbf{y}_{1:N};\Psi)\\
\quad -\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})^T(\boldsymbol{\delta}_\epsilon - \sigma^2\boldsymbol{\delta}_x)\mathbf{A}_5(\mathbf{y}_{1:N};\Psi) + \boldsymbol{\mu}^{*T}\boldsymbol{\mu}\mathbf{W}\mathbf{A}_6(\mathbf{y}_{1:N};\Psi) + \boldsymbol{\mu}^{*T}(\boldsymbol{\delta}_\epsilon - \sigma^2\boldsymbol{\delta}_x)\mathbf{A}_5(\mathbf{y}_{1:N};\Psi) + \boldsymbol{\delta}_\epsilon^{*T}\boldsymbol{\mu}\mathbf{W}\mathbf{A}_5(\mathbf{y}_{1:N};\Psi)\\
\quad +\boldsymbol{\delta}_\epsilon^{*T}(\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x)\mathbf{A}_4(\mathbf{y}_{1:N};\Psi)\Big]\Big[\mathbf{A}_7(\mathbf{y}_{1:N};\Psi) - \mathbf{A}_8(\mathbf{y}_{1:N};\Psi) - \mathbf{A}_8(\mathbf{y}_{1:N};\Psi)^T + \mathbf{A}_9(\mathbf{y}_{1:N};\Psi) + \sigma^2\mathbf{A}_5(\mathbf{y}_{1:N};\Psi)\Big]^{-1}\\
\sigma^{*2} = \frac{1}{dA_0(\mathbf{y}_{1:N};\Psi)}\Big[A_1(\mathbf{y}_{1:N};\Psi)\,Tr\left\{\bar{\mathbf{S}}(\mathbf{y}_{1:N})\right\} - 2A_1(\mathbf{y}_{1:N};\Psi)\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\boldsymbol{\mu}^{*T} - 2A_0(\mathbf{y}_{1:N};\Psi)\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\boldsymbol{\delta}_\epsilon^{*T} + A_1(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^*\boldsymbol{\mu}^{*T}\\
\quad +2A_0(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^*\boldsymbol{\delta}_\epsilon^{*T} + A_2(\mathbf{y}_{1:N};\Psi)\boldsymbol{\delta}_\epsilon^*\boldsymbol{\delta}_\epsilon^{*T} - 2\,Tr\left\{\bar{\mathbf{S}}(\mathbf{y}_{1:N})\mathbf{W}\mathbf{A}_6(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\right\} + 2\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{W}\mathbf{A}_6(\mathbf{y}_{1:N})\mathbf{W}^{*T}\boldsymbol{\mu}^{*T}\\
\quad +2\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{W}\mathbf{A}_5(\mathbf{y}_{1:N})\mathbf{W}^{*T}\boldsymbol{\delta}_\epsilon^{*T} + 2\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{W}^*\mathbf{A}_6(\mathbf{y}_{1:N};\Psi)\mathbf{W}^T\boldsymbol{\mu}^T + 2\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{W}^*\mathbf{A}_5(\mathbf{y}_{1:N};\Psi)(\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x)^T\\
\quad -2\boldsymbol{\mu}\mathbf{W}\mathbf{A}_6(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\boldsymbol{\mu}^{*T} - 2(\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x)\mathbf{A}_5(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\boldsymbol{\mu}^{*T} - 2\boldsymbol{\mu}\mathbf{W}\mathbf{A}_5(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\boldsymbol{\delta}_\epsilon^{*T}\\
\quad -2(\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x)\mathbf{A}_4(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\boldsymbol{\delta}_\epsilon^{*T} + Tr\left\{\mathbf{A}_7(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\mathbf{W}^*\right\} + Tr\left\{\mathbf{A}_9(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\mathbf{W}^*\right\}\\
\quad - Tr\left\{\left(\mathbf{A}_8(\mathbf{y}_{1:N};\Psi) + \mathbf{A}_8(\mathbf{y}_{1:N};\Psi)^T\right)\mathbf{W}^{*T}\mathbf{W}^*\right\} + \sigma^2\,Tr\left\{\mathbf{A}_5(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\mathbf{W}^*\right\}\Big].
\end{cases}$$

The 2-dimensional integrals $\mathbf{A}_i$ on the hypercube $[0,1]^2$, for $i \in \{0,\ldots,9\}$, are defined in Subsubsection A.4.3.4 in Appendix A.4.

Proof. The proof of Corollary 13 is given in Subsection A.4.3.4 in Appendix A.4. □

Corollary 14. *Under the assumptions of Special Case 3: Generalized t-Student IID PPCA specified in Subsubsection 3.4.1.3, the M-step of the corresponding EM algorithm with the $Q_{GStS,id,3}$ function determined in Corollary 10 is a case-specific variant of the solution to the system of equations defined in Theorem 16 under the problem formulation $\nabla_{\Psi^*} Q_{GStS,id,3} = \mathbf{0}$ for $\Psi^* = \left[\mathbf{W}^*, \boldsymbol{\mu}^*, \sigma^{*2}\right]$. The maximizers with respect to Ψ^* are given by explicit formulas defined by 2-dimensional integration problems on the hypercube $[0,1]^2$ as follows*

$$
\begin{cases}
\boldsymbol{\mu}^* = A_1(\mathbf{y}_{1:N};\Psi)^{-1}\left[\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\left(A_1(\mathbf{y}_{1:N};\Psi) - \mathbf{W}\mathbf{A}_6(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\right) + \boldsymbol{\mu}\mathbf{W}\mathbf{A}_6(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\right] \\
\mathbf{W}^* = \left[\bar{\mathbf{S}}(\mathbf{y}_{1:N})\mathbf{W}\mathbf{A}_6(\mathbf{y}_{1:N};\Psi) - \boldsymbol{\mu}^{*T}\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{W}\mathbf{A}_6(\mathbf{y}_{1:N};\Psi) - \bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})^T\boldsymbol{\mu}\mathbf{W}\mathbf{A}_6(\mathbf{y}_{1:N};\Psi) + \boldsymbol{\mu}^{*T}\boldsymbol{\mu}\mathbf{W}\mathbf{A}_6(\mathbf{y}_{1:N};\Psi)\right] \\
\quad \times \left[\mathbf{A}_7(\mathbf{y}_{1:N};\Psi) - \mathbf{A}_8(\mathbf{y}_{1:N};\Psi) - \mathbf{A}_8(\mathbf{y}_{1:N};\Psi)^T + \mathbf{A}_9(\mathbf{y}_{1:N};\Psi) + \sigma^2\mathbf{A}_5(\mathbf{y}_{1:N};\Psi)\right]^{-1} \\
\sigma^{*2} = \frac{1}{dA_0(\mathbf{y}_{1:N};\Psi)}\left[A_1(\mathbf{y}_{1:N};\Psi)\,Tr\left\{\bar{\mathbf{S}}(\mathbf{y}_{1:N})\right\} - 2A_1(\mathbf{y}_{1:N};\Psi)\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\boldsymbol{\mu}^{*T} + A_1(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^*\boldsymbol{\mu}^{*T}\right. \\
\quad -2\,Tr\left\{\bar{\mathbf{S}}(\mathbf{y}_{1:N})\mathbf{W}\mathbf{A}_6(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\right\} + 2\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{W}\mathbf{A}_6(\mathbf{y}_{1:N})\mathbf{W}^{*T}\boldsymbol{\mu}^{*T} + 2\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{W}^*\mathbf{A}_6(\mathbf{y}_{1:N};\Psi)\mathbf{W}^T\boldsymbol{\mu}^T \\
\quad -2\boldsymbol{\mu}\mathbf{W}\mathbf{A}_6(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\boldsymbol{\mu}^{*T} + Tr\left\{\mathbf{A}_7(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\mathbf{W}^*\right\} + Tr\left\{\mathbf{A}_9(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\mathbf{W}^*\right\} \\
\quad \left. - Tr\left\{\left(\mathbf{A}_8(\mathbf{y}_{1:N};\Psi) + \mathbf{A}_8(\mathbf{y}_{1:N};\Psi)^T\right)\mathbf{W}^{*T}\mathbf{W}^*\right\} + \sigma^2\,Tr\left\{\mathbf{A}_5(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\mathbf{W}^*\right\}\right].
\end{cases}
$$

The 2-dimensional integrals $\mathbf{A}_i$ on the hypercube $[0,1]^2$, for $i \in \left\{0,\ldots,22\right\}$, are defined in Subsubsection A.4.1.5 in Appendix A.4 with the simplifications of the following integrals

$$
\begin{aligned}
\mathbf{A}_8(\mathbf{y}_{1:N};\Psi)_{k\times k} &:= \int_0^1\int_0^1 T_\epsilon(s_\epsilon)^3\mathbf{M}^{-1}\mathbf{W}^T\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})^T\boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}m(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x, \\
\mathbf{A}_9(\mathbf{y}_{1:N};\Psi)_{k\times k} &:= \int_0^1\int_0^1 T_\epsilon(s_\epsilon)^3\mathbf{M}^{-1}\mathbf{W}^T\boldsymbol{\mu}^T\boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}m(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x.
\end{aligned}
$$

Proof. The proof of Corollary 14 is analogous to the proof of Corollary 13 in Appendix A.4.3.4 when one sets $\mathbf{D}_\epsilon = T_\epsilon(s_\epsilon)\mathbb{I}_d$ and $\mathbf{D}_x = T_x(s_x)\mathbb{I}_k$ and $\boldsymbol{\delta}_\epsilon = \boldsymbol{\delta}_x = \mathbf{0}$. □

3.5. The Optimisation Steps of EM Algorithm for Generalized Skew t-Student Probabilistic Principal Component Analysis

In the following section we address two aspects of the EM algorithm for Generalized Skew t-Student PPCA which require additional steps and clarifications. Firstly, we show how to approximate the solutions to the E-step in the algorithms which are given by two dimensional integration problems. Next, we discuss the solutions to the normal equations which define the M-step of the algorithm and cannot be explicitly formulated.

3.5.1. Approximated Solutions to the Integrals in E-step of EM Algorithm for Generalized Skew t-Student Probabilistic Principal Component

The functions I_1 and I_2 defined in Theorem 13 and Theorem 15 or in their equivalents in the incomplete data setting, Theorem 25 and Theorem 27, can be efficiently approximated by a 2-dimensional quadrature rule on the hypercube $[-1,1]^2$. In the following subsection we show the steps one needs to employ in order to approximate the solutions to I_1 and I_2.

We show how to employ the product rule formed from the one dimensional Gauss-Legendre quadrature as discussed in or (Stroud and Secrest 1966), (Isaacson and Keller 1966) or (Davis and Rabinowitz 1984). The general Gauss quadrature rule enjoys many useful properties, which such

as high degrees of the precision and accuracy. We can distinguish many types of the Gaussian type quadrature rule which are related to various representations of weighting functions. These functions define the space metric. For the purposes of this work, we consider Gauss-Legendre quadrature rule which assumes the weighting function equal to one.

Since the problems defined by the functions I_1 and I_2 are two dimensional, we combine the product rule and one dimensional quadrature rule to approximate underlying integrals. Let us remark that the discussed methodology can be easily extended to the sparse grid rule introduced by (Smolyak 1963) which enjoys lower computational cost studied in (Novak and Schiller 1999),(Frank and Heinrich 1996) or (Wasilkowski and Wozniakowski 1995) among many other notable works. In the next few lines we revisit the concept of one dimensional Gaussian-Legendre quadrature rule and expand it to the two dimensional case by an application of the product rule.

Let us define the function $f(x)$ such that $f : [-1, 1] :\to \mathbb{R}$ is continuous on its domain. Under the Weierstrass' Approximation Theorem, see discussion in (Isaacson and Keller 1966), for a continuous function defined on closed interval, there exist a polynomial $p_n(x)$ of degree at most n which approximates the function $f(x)$ with some error term $r_n(x)$ dependent on the degree n. Hence the function $f(x)$ can be decomposed as follows

$$f(x) = p_n(x) + r_n(x), \tag{3.44}$$

where the interpolation polynomial $p_n(x)$ satisfies the following set of equations

$$f(x_i) = p_n(x_i) \text{ for } i = 0, \ldots, n, \tag{3.45}$$

for $n + 1$ distinct point $x_0 < x_1 < \ldots < x_n$ on $[-1, 1]$. Hence, the function f equals to a polynomial p_n on a set of distinct points x_i. A quadrature rule with equally spaced points x_i belongs to the class of Newton–Cotes quadrature formulas. The Gauss formulas relax the assumption on the equally spaced points in order to achieve improvements in the accuracy of the approximation.

The polynomial $p_n(x)$ is chosen to minimize a distance between the target function and its interpolation, that is

$$||f(x) - p_n(x)||_2 = \left\{ \int_{-1}^{1} \left[f(x) - p_n(x) \right]^2 dx \right\}^{\frac{1}{2}} = \left\{ \int_{-1}^{1} \left[r_n(x) \right]^2 dx \right\}^{\frac{1}{2}}$$

Please refer to (Isaacson and Keller 1966) for the proofs of the uniqueness and existence of the polynomial $p_n(x)$ as well as the properties of the reminder r_n.

The polynomial p_n can be represented using the monomial basis $1, x, x^2, \ldots, x^m$. If the function f belongs to the class $\mathbb{C}^{n-1}([-1, 1])$, the residual r_n is equal to zero and the interpolation is exact. However, we seek the quadrature rule for which the error vanishes as well when f is a polynomial of a higher order than n. By applying the Gram-Schmidt orthogonalization, the $p_n(x)$ can be determined in terms of the orthogonal polynomials called Legendre polynomials what leads to the Gauss-Legendre quadrature rule.

The use of orthogonal polynomials is proven to increase the accuracy of the approximation which equals to $2n-1$ and reduce computational costs as discussed in (Isaacson and Keller (1966), Chapter 7). The n points of the Gauss-Legendre quadrature rule exactly integrate the polynomials of degree $2n$. Without loss of generality, we define p_n as a linear combination of the orthogonal functions such that

$$p_n(x) = \sum_{i=0}^{n} c_i P_i(x) \text{ for } \int_{-1}^{1} P_i(x)P_j(x)\,dx = \delta_{i,j} \text{ and } i,j = 0,1,\ldots n, \tag{3.46}$$

where $\delta_{i,j}$ is the Dirac delta function which sets 1 when $i=j$, otherwise 0. The coefficients c_i are given by

$$c_i = \int_{-1}^{1} P_i(x)\,f(x)\,dx \text{ for } i = 0,1,\ldots,n. \tag{3.47}$$

The expansion of the polynomial $p_n(x)$ in terms of orthogonal functions enjoys many useful properties, see discussion in (Isaacson and Keller (1966), Chapters 5-7) such as convergence in $\mathbb{L}_2([-1,1])$, a bounded modulus of $r_n(x)$ and uniform convergence. The orthogonal polynomials which satisfy (3.46) are called Legendre polynomials and have the following representation

$$P_i(x) = \frac{1}{i!2^i}\frac{d^i}{dx^i}\left(x^2-1\right)^i. \tag{3.48}$$

The polynomials of this form satisfy the Bonnet's recursion formula

$$(i+1)P_{i+1}(x) := (2i+1)xP_i(x) - iP_{i-1}(x).$$

The resulting quadrature rule, called the Gauss-Legendre rule, defines the approximation of

$$I\{f\} := \int_{-1}^{1} f(x)\,dx,$$

by

$$I_n\{f\} := \sum_{i=0}^{n} \phi_i f(x_i),$$

where $f(x)$ is interpolated by the polynomial $p_n(x)$ being a linear combination of orthogonal polynomials $P_i(x)$. The quadrature rule which uses orthogonal polynomials enjoys the degree of precision $2n-1$. For the Legendre polynomials, the coefficients ϕ_i are given by

$$\phi_i = \frac{(2i)!}{2^i(i!)^2}\sqrt{\frac{2i+1}{2}}. \tag{3.49}$$

The distinct grid points $x_0,\ldots,x_n$ are $n+1$ roots of the polynomial $P_n(x)$. If $f(x)$ has a continuous derivative of order $2n$ in $[-1,1]$, the quadrature error term is expressed as follows

$$E_n\{f\} = I\{f\} - I^n\{f\} = \int_{-1}^{1} r_n(x)\,dx = \frac{2}{(2n+1)!}\left[\frac{2^n(n!)^2}{(2n)!}\right]^2 \frac{d^{2n}}{dx^{2n}}f(\zeta) \text{ for } -1 < \zeta < 1 \tag{3.50}$$

as shown in (Abramowitz and Stegun 1974) or (Isaacson and Keller 1966). There is an extensive literature on the assessment of error bonds for Gauss-type quadrature. Please refer to (Barrett 1961), (Chawla 1968), (Chawla and Jain 1968), (Tuan 1971) or (Lether 1980) as early examples of works as well as the books of (Isaacson and Keller 1966) or (Davis and Rabinowitz 1984).

Let us define the integration problem $I\{g\} = \int_{-1}^{1}\int_{-1}^{1} g(x,y)\ dx\ dy$ for the function $g : [-1,1]^2 \to \mathbb{R}$ continuous on its domain. For two positive natural numbers n and n', the approximation $I_{n,n'}\{g\}$ of $I\{g\}$ can be obtained using the product rule of the univariate Gauss-Legendre quadratures defined in (Davis and Rabinowitz 1984) as follows

$$I\{g\} =: \int_{-1}^{1} \underbrace{\int_{-1}^{1} g(x,y)dx}_{\tilde{g}(y)}\, dy = \int_{-1}^{1} \tilde{g}(y)dy \approx \sum_{j=0}^{n'} \phi_j \tilde{g}(y_j) \approx \sum_{i=0}^{n}\sum_{j=0}^{n'} \phi_i\phi_j g(x_i,y_j) =: I_{n,n'}\{g\}, \tag{3.51}$$

for the the distinct points $x_0, x_1, \ldots, x_n \in [-1,1]$ and $y_0, y_1, \ldots, y_{n'} \in [-1,1]$ being the zero root of the polynomials $P_n(x)$ and $P_{n'}(x)$,respectively, defined as in Equation (3.48). The error representation of the product cubature rules have been studied in (Lether 1970), (Lether 1971b) or (Lether 1971a) as one the the earliest works.

In order to apply the product rule of one dimensional Gauss-Lagrange quadratures to approximate the solution to the function I_1 and I_2, one needs to change the integration domains from $[0,1]^2$ to $[-1,1]^2$. By applying the change of variables given by the linear transformation $s \to 2s-1$ we obtain that

$$\begin{aligned}
I_1(\mathbf{y}_t;\Psi) &= \frac{1}{4}\int_{-1}^{1}\int_{-1}^{1} m\Big(\mathbf{y}_t, \frac{\tilde{s}_{\epsilon,t}+1}{2}, \frac{\tilde{s}_{x,t}+1}{2};\Psi\Big)\, d\tilde{s}_{\epsilon,t}\, d\tilde{s}_{x,t},\\
I_2(\mathbf{y}_t;\Psi,\Psi^*) &= \int_{0}^{1}\int_{0}^{1} \tilde{w}\Big(\mathbf{y}_t, \frac{\tilde{s}_{\epsilon,t}+1}{2}, \frac{\tilde{s}_{x,t}+1}{2};\Psi,\Psi^*\Big)\, m\Big(\mathbf{y}_t, \frac{\tilde{s}_{\epsilon,t}+1}{2}, \frac{\tilde{s}_{x,t}+1}{2};\Psi\Big)\, d\tilde{s}_{\epsilon,t}\, d\tilde{s}_{x,t}.
\end{aligned}$$

The Proposition 1 proves that the function I_1 and I_2 can be approximated by the product rule of univariate Gauss-Legendre rules as defined above.

Proposition 1. *The solutions to the integration problems*

$$\begin{aligned}
I_1(\mathbf{y}_t;\Psi) &= \int_{0}^{1}\int_{0}^{1} m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t};\Psi)\, ds_{\epsilon,t}\, ds_{x,t},\\
I_2(\mathbf{y}_t;\Psi,\Psi^*) &= \int_{0}^{1}\int_{0}^{1} \tilde{w}(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t};\Psi,\Psi^*)\, m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t};\Psi,\Psi^*) ds_{\epsilon,t}\, ds_{x,t},
\end{aligned}$$

for function m, $\tilde{w} : \mathbb{R}^d \times [0,1]^2 \to \mathbb{R}$ specified in Theorem 13 can be approximated by the 2-dimensional quadrature rule which combines Gauss-Legendre rules on $[0,1]$ and the product rule. The obtained approximation is formulated as following

$$\begin{aligned}
I_1(\mathbf{y}_t;\Psi) &\approx \frac{1}{4}\sum_{i=0}^{n}\sum_{j=0}^{n'} \phi_i\phi_j\, m\Big(\mathbf{y}_t, \frac{s_{n,i}+1}{2}, \frac{s_{n',j}+1}{2}\Big),\\
I_2(\mathbf{y}_t;\Psi,\Psi^*) &\approx \frac{1}{4}\sum_{i=0}^{n}\sum_{j=0}^{n'} \phi_i\phi_j\, h\Big(\mathbf{y}_t, \frac{s_{n,i}+1}{2}, \frac{s_{n',j}+1}{2}\Big),
\end{aligned}$$

where $h := m \cdot \tilde{w}$, *the coefficients* ϕ_i *are defined as in Equation* (3.49), *the distinct points* $s_{n,i} \in [-1,1]$ *for* $i = 0, \ldots, n$ *and* $s_{n',j} \in [-1,1]$ *for* $j = 0, \ldots, n'$ *are the zero roots of the Legendre polynomials* P_n *and* $P_{n'}$, *respectively, defined as in Equation* (3.48). *If* $n = \min\left[\frac{\nu_\epsilon}{4}\right]$ *and* $n' = \min\left[\frac{\nu_x}{4}\right]$, *the quadrature error terms denoted by* $E_{n,n'}\{m\}(\mathbf{y}_t)$ *and* $E_{n,n'}\{h\}(\mathbf{y}_t)$ *are obtained in closed forms given the definitions of the following operators*

$$
\begin{aligned}
E_{n,n'}\{f\}(\mathbf{y}_t) &= \frac{1}{4}\sum_{j=0}^{n'} \phi_j E_n\{f\}\Big(\mathbf{y}_t, \frac{s_{n',i}-1}{2}\Big) + \frac{1}{4}\sum_{i=0}^{n} \phi_i E_{n'}\{f\}\Big(\mathbf{y}_t, \frac{s_{n,i}+1}{2}\Big), \\
&+ \frac{1}{(2n+1)!(2n'+1)!}\left[\frac{2^{n'}(n'!)^2}{(2n')!}\,\frac{2^n(n!)^2}{(2n)!}\right]^2 \frac{\partial^{2n+2n'}}{\partial s_\epsilon^{2n}\partial s_x^{2n'}} f\Big(\mathbf{y}_t, \frac{\zeta+1}{2}, \frac{\varsigma+1}{2}\Big) \\
E_n\{f\}(\mathbf{y}_t, s_x) &= \frac{2}{(2n+1)!}\left[\frac{2^n(n!)^2}{(2n)!}\right]^2 \frac{\partial^{2n}}{\partial s_\epsilon^{2n}} f\Big(\mathbf{y}_t, \frac{\zeta+1}{2}, s_x; \Psi\Big), \\
E_{n'}\{f\}(\mathbf{y}_t, s_\epsilon) &= \frac{2}{(2n'+1)!}\left[\frac{2^{n'}(n'!)^2}{(2n')!}\right]^2 \frac{\partial^{2n'}}{\partial s_x^{2n'}} f\Big(\mathbf{y}_t, s_\epsilon, \frac{\varsigma+1}{2}; \Psi\Big),
\end{aligned}
$$

for an arbitrary function $f : \mathbb{R}^d \times [0,1]^2 \to \mathbb{R}$ *that belongs to* $\mathbb{C}^{2n,2n'}([0,1]^2)$ *with respect to* s_ϵ *and* s_x *and* $-1 < \zeta < 1$ *and* $-1 < \varsigma < 1$.

Proof. The proof of Proposition 1 is located in Section A.4.7 in Appendix A.4. The proof of the quadrature rule for the functions I_1 and I_2 are given in Lemma A18 and Lemma A20, respectively. □

Following analogous steps, we can derive the quadrature rules for functions I_1 and I_2 from Theorem 15 as well as extend the reasoning to formulate the approximations of the equivalent functions in Theorem 25 and Theorem 27 in the incomplete data setting.

3.5.2. Optimisation Framework for Iterative Root Search in M-step of EM Algorithm for Generalized Skew t-Student Probabilistic Principal Component Analysis

In the following subsection we describe the procedure of calculating the M-step of EM algorithm for Generalized Skew t-Student PPCA when the iterative updates of the the parameter $\mathbf{W}$ are not formulated explicitly. Let us remark that when $\boldsymbol{\nu}_\epsilon = \nu_\epsilon \mathbf{1}_d$ and $\boldsymbol{\nu}_x = \nu_x \mathbf{1}_k$, the iterations of the EM algorithm have closed forms defined by the two dimensional integration problems. Our methodology is explained for independent case in complete data setting but can be easily extended to other considered special cases of models and when the missing data appears.

The solution to the system of equations from Theorem 14 defines the M-steps of EM algorithms for Generalized Skew t-Student IND PPCA and consequently, the updates of the vector with static parameters, $\Psi = [\mathbf{W}, \boldsymbol{\mu}, \boldsymbol{\delta}_\epsilon, \boldsymbol{\delta}_x, \sigma^2]$. The estimates of the parameters are determined by maximising the expectation of the log likelihood of the PPCA model from Theorem 13, the function $Q_{GStS,ind}$. The maximizers with respect to $\boldsymbol{\mu}^*, \boldsymbol{\delta}_\epsilon^*, \boldsymbol{\delta}_x^*, \sigma^{*2}$ are given by explicit formulas whereas the maximizer with respect to $\mathbf{W}^*$ is determined by the equation

$$
f(\mathbf{W}^*; \Psi, \Psi^*) = \mathbf{0}
$$

for the function $f : \mathbb{R}^{d\times k} \to \mathbb{R}^{d\times k}$ defined in Theorem 14. The function f is linear with respect to the parameters $\boldsymbol{\mu}^*$ and δ_ϵ^*. Also, the explicit formulations of the maximizers with respect to $\boldsymbol{\mu}^*, \delta_\epsilon^*, \delta_x^*, \sigma^{*2}$ are dependent on the parameter $\mathbf{W}^*$.

We can use a gradient descent type algorithm to iteratively compute Ψ^* which maximizes the function $Q_{GStS,ind}$ in the M-step as both first order and second order partial derivatives of the function $Q_{GStS,ind}$ with respect to the vector Ψ^* are well defined and continuous. The procedure is called Generalized EM algorithm as defined in (Dempster et al. 1977). In fact, given the two subsequent updates of Ψ in iteration n and $n+1$, denoted by $\Psi^{(n)}$ and $\Psi^{(n+1)}$, for some integer number n, the E-step can be formulated as $Q_{GStS,ind}(\Psi^{(n)}, \Psi^{(n+1)})$. The $n+1$ update of the vector $\Psi^{(n+1)}$ is defined as

$$\Psi^{(n+1)} = \Psi^{(n)} + a^{(n)}\mathbf{D}^{(n)} \text{ for } \mathbf{D}^{(n)} = -\left[\left(\frac{\partial Q_{GStS,ind}(\Psi^{(n)}, \Psi^*)}{\partial \Psi^* \partial \Psi^{*T}}\right)^{-1} \frac{\partial Q_{GStS,ind}(\Psi^{(n)}, \Psi^*)}{\partial \Psi^*}\right]_{\Psi^*=\Psi^{(n)}}$$

and $a^{(n)}$ is chosen so that $Q_{GStS,ind}(\Psi^{(n)}, \Psi^{(n+1)}) \geq Q_{GStS,ind}(\Psi^{(n)}, \Psi^{(n)})$. When $a^{(n)} = 1$, we obtain the Newton-Raphson procedure to obtain the maximizers of the function $Q_{GStS,ind}$. Recall that each iteration of the algorithm requires solving various 2-dimensional integration problems on $[0,1]^2$.

In order to improve the accuracy of the procedure, we can use the fact that only the update of parameter $\mathbf{W}$ is not explicitly formulated and solve the lower dimension problem $f(\mathbf{W}^*; \Psi, \Psi^*) = 0$ which determines the maximizer of $Q_{GStS,ind}$ with respect to $\mathbf{W}^*$.

Since the iterative updates of $\boldsymbol{\mu}^*$, δ_ϵ^* and σ^{*2} are linear functions of the parameter $\mathbf{W}^*$ and the function f, which zero root determines the maximizer $\mathbf{W}^*$, depends on $\boldsymbol{\mu}^*$ and δ_ϵ^*, we need to solve the following system of equations

$$\begin{cases} \boldsymbol{\mu}^* = g(\mathbf{W}^*) \\ \delta_\epsilon^* = h(\mathbf{W}^*) \\ f(\mathbf{W}^*; \Psi, \Psi^*) = \mathbf{0}. \end{cases} \tag{3.52}$$

where function g and h are specified by simplifying the the following equations

$$\begin{cases} \boldsymbol{\mu}^* = \left[\mathbf{A}_6(\mathbf{y}_{1:N}; \Psi) - \mathbf{A}_{10}(\mathbf{y}_{1:N}; \Psi, \Psi^*) - \delta_\epsilon^* A_0(\mathbf{y}_{1:N}; \Psi) + \boldsymbol{\mu}\mathbf{A}_{13}(\mathbf{y}_{1:N}; \Psi, \Psi^*) + \left(\delta_\epsilon \mathbf{W} - \sigma^2\delta_x\right)\mathbf{A}_{12}(\mathbf{y}_{1:N}; \Psi, \Psi^*)^T\right]\mathbf{A}_1(\mathbf{y}_{1:N}; \Psi)^{-1} \\ \delta_\epsilon^* = \left[\mathbf{A}_5(\mathbf{y}_{1:N}; \Psi) - \boldsymbol{\mu}^* A_0(\mathbf{y}_{1:N}; \Psi) - \mathbf{A}_8(\mathbf{y}_{1:N}; \Psi)\mathbf{W}^{*T} + \boldsymbol{\mu}\mathbf{A}_{11}(\mathbf{y}_{1:N}; \Psi)\mathbf{W}^{*T} + \left(\delta_\epsilon \mathbf{W} - \sigma^2\delta_x\right)\mathbf{A}_4(\mathbf{y}_{1:N}; \Psi)\mathbf{W}^{*T}\right]\mathbf{A}_2(\mathbf{y}_{1:N}; \Psi)^{-1} \end{cases}$$

Hence, one iteration of the EM algorithm requires substituting the formulas corresponding to $\boldsymbol{\mu}^*$ and δ_ϵ^*, which are functions of $\mathbf{W}^*$, into the function f in order to find the solution for the parameter $\mathbf{W}^*$. Let us remark that the function f is a smooth function of $\mathbf{W}^*$ and hence its first and second derivatives exist and are well defined. Hence, we can employ the Newton-type gradient method to solve the non-linear system of equations (3.52). After the values of $\boldsymbol{\mu}^*$, $\mathbf{W}^*$ and δ_ϵ^* are calculated, one needs to update the parameters σ^{*2} and δ_x^* what ends an iteration of the EM algorithm.

3.6. Generalized Expectation-Maximisation Algorithm

Since the solution to the M-step of the EM algorithm for GStS PPCA is not obtained in the closed form, the update of Ψ is only an approximation to the solution of

$$\hat{\Psi}^* = \underset{\Psi^* \in \Omega}{\operatorname{argmax}}\, Q_{GStS,ind}(\Psi, \Psi^*)$$

which is calculated using the optimisation framework discussed in the previous section. Therefore, $\hat{\Psi}^*$ is not a global maximizers of $Q_{GStS,ind}(\Psi, \Psi^*)$ (equivalently, $Q_{GStS,iid}(\Psi, \Psi^*)$).

The author of (Dempster et al. 1977) uses the notion of the Generalized Expectation-Maximisation (GEM) algorithm, the procedure based on the same arguments as the standard EM algorithm however, which generates a sequence of parameters updates that only increases the likelihood function. Therefore, the sequence is not constructed from maximizers of the Q function, that it, the gradient of the Q function at the points which belong to the sequence is not necessary equal zero. The (GEM) algorithm requires that at i-th iteration of the algorithm, the update of the parameter Ψ is chosen such that

$$Q_{GStS,ind}(\Psi^{(i)}, \Psi^{(i+1)}) \geq Q_{GStS,ind}(\Psi^{(i)}, \Psi^{(i)}). \tag{3.53}$$

We remark that the EM algorithm is a special case of the GEM algorithm, which satisfies stronger condition that in Equation (3.53). Following the notation from Section 2.5, a sequence $\left\{\Psi^{(i)}\right\}_{i \in \mathbb{N}_0}$ denotes updates of the parameter Ψ which are specified by iterations of the GEM algorithm. The GEM algorithm defines implicitly the mapping $M : \Omega \to \Omega$ such that

$$\Psi^{(i+1)} = M(\Psi^{(i)}). \tag{3.54}$$

If M is continuous and the sequence $\left\{\Psi^{(i)}\right\}_{i \in \mathbb{N}_0}$ converges to a limit $\tilde{\Psi}$, then M must satisfy that

$$M(\tilde{\Psi}) = \tilde{\Psi}.$$

Hence, $\tilde{\Psi}$ acts as a fixed point of M. If $\tilde{\Psi}$ is the MLE estimate of Ψ then it is also a fixed point of M as shown in (Dempster et al. 1977). This property is ensured by the definition of the M-step which implies that

$$l\left(M(\tilde{\Psi}), \mathbf{y}_{1:N}\right) \geq l(\tilde{\Psi}, \mathbf{y}_{1:N}),$$

where $l(\Psi, \mathbf{y}_{1:N})$ is a likelihood function defined in Equation (2.8).

We denote $\left\{l^{(i)}\right\}_{i \in \mathbb{N}_0}$ as a sequence of real numbers which are obtained by applying the loglikelihood function in Equation (2.8) to the candidates of parameters in $\left\{\Psi^{(i)}\right\}_{i \in \mathbb{N}_0}$ obtained from the GEM algorithm, that is

$$l^{(i)} := l(\Psi^{(i)}; \mathbf{y}_{1:N}) \text{ for } i \in \mathbb{N}_0.$$

The condition in Equation (3.53) is sufficient to ensure that the sequence $\left\{l^{(i)}\right\}_{i\in\mathbb{N}_0}$ is non-decreasing – please refer to the steps of the proof to Theorem 1. Therefore, the GEM algorithm enjoys the same properties as defined in Theorem 1.

As mentioned in Section 2.5, (Dempster et al. 1977), (Wu 1983) and (Boyles 1983) study additional assumptions which need to hold in order to obtain the convergence of the GEM algorithm and provide generalizations of Theorem 2 and Theorem 3. Similarly to the EM algorithm, if the conditions (1)-(3) from Section 2.5 are satisfied, the GEM sequence $\left\{l^{(i)}\right\}_{i\in\mathbb{N}_0}$ converges to some limit what we summarize in Lemma 4.

Lemma 4. *(Wu 1983) Let the conditions (1)-(3) in Section 2.5 hold. It implies that the GEM sequence* $\left\{l^{(i)}\right\}_{i\in\mathbb{N}_0}$ *is bounded for any initial value* $\Psi^{(0)} \in \Omega$ *and converges to some limit, denoted by* $\tilde{l}$.

The outcome of Lemma 4 only guarantees the convergence of the GEM algorithm to some value of the loglikelihood function in Equation (2.8). The stronger convergence of the method requires additional conditions on the smoothness of the mapping M which we summarize in Theorem 17

Theorem 17. *(Wu 1983) Let* $\mathcal{S}$ *denote the set of all stationary points of the likelihood function in Equation* (2.8) *which belong to the interior of* Ω. *If the condition (1)-(3) hold and the following are satisfied*

(1): *the mapping* M *is closed over the complement of* $\mathcal{S}$,
(2): *the sequence* $\left\{l^{(i)}\right\}_{i\in\mathbb{N}_0}$ *is increasing for all* $\Psi^{(i)}$ *not* $\in \mathcal{S}$,

then all limit points of the sequence $\left\{\Psi^{(i)}\right\}_{i\in\mathbb{N}_0}$ *are the stationary points of the loglikelihood function in Equation* (2.8) *and the sequence* $\left\{\Psi^{(i)}\right\}_{i\in\mathbb{N}_0}$ *converges monotonically to* $\tilde{l}$ *such that* $\tilde{l} := l(\tilde{\Psi}; \mathbf{y}_{1:N})$ *for some stationary point* $\tilde{\Psi}$.

Similarly to the convergence of the EM algorithm proven in Theorem 2, the conditions of Theoream 17 guarantes only the convergence of $\left\{l^{(i)}\right\}_{i\in\mathbb{N}_0}$ to a stationary point of the loglikelihood. Again, we might use standard practices such run the EM with multiple initializations to obtain the convergence to local or global maximum. In the next lines we share the last results for the GEM algorithm which outlines the conditions which need to be satisfied that the GEM sequence $\left\{\Psi^{(i)}\right\}_{i\in\mathbb{N}_0}$ converges to the stationary point of the loglikelihood function in Equation (2.8).

Theorem 18. *(Wu 1983) Let the condition (1)-(3) in Section 2.5 hold. We denote* $\mathcal{S}_a$ *as a subset of all stationary points of the loglikelihood function in Equation* (2.8) *which belong to the interior of* Ω *and result in the loglikelihood function equal to* $a \in \mathbb{R}$, *that is*

$$l(\Psi; \mathbf{y}_{1:N}) = a \text{ for any } \Psi \in \Omega_a.$$

Suppose that either conditions in Theorem 17 are satisfied or $\left\{\Psi^{(i)}\right\}_{i\in\mathbb{N}_0}$ *is an instance of the GEM algorithm such that*

$$\frac{\partial Q(\Psi^{(i)}, \Psi^*)}{\partial \Psi^*} = 0 \text{ in } \Psi^* = \Psi^{(i+1)}.$$

In addition we assume that $\frac{\partial Q(\Psi,\Psi^*)}{\partial \Psi^*}$ *is continuous in* Ψ *and* Ψ^**. Then* $\left\{\Psi^{(i)}\right\}_{i\in\mathbb{N}_0}$ *converges to a stationary point of the loglikelihood in Equation (2.8), denoted by* $\tilde{\Psi} \in \mathcal{S}_{\tilde{l}}$*, such that* $\tilde{l} = l(\tilde{\Psi}; \mathbf{y}_{1:N})$ *is a limit of the corresponding sequence* $\left\{l^{(i)}\right\}_{i\in\mathbb{N}_0}$*, if either:*

(1): $\mathcal{S}_{\tilde{l}}$ *consists of a single point,*

(2): *or the elements of the sequence* $\left\{\Psi^{(i)}\right\}_{i\in\mathbb{N}_0}$ *satisfy that*

$$||\Psi^{(i+1)} - \Psi^{(i)}|| \to 0 \text{ for } i \to \infty,$$

and $\mathcal{S}_{\tilde{l}}$ *is discrete, where* $||\cdot||$ *denotes a norm on* Ω*.*

The results presented so far consider the broad range of GEM algorithms. In this thesis we study a spacial case of the algorithm which propose the approximation of the M-step by the Newton-Raphson procedure described in Subsection 3.5.2. This algorithm can be viewed as a gradient GEM algorithm introduced by (Lange 1995) when the mapping M is defined as a single iteration of the Newton-Raphson algorithm such that

$$\Psi^{(i+1)} = M(\Psi^{(i)}) \approx \underset{\Psi\in\Omega}{\operatorname{argmax}}\, Q(\Psi^{(i)}, \Psi)$$

where

$$M(\Psi^{(i)}) = \Psi^{(i)} + \mathbf{D}^{(i)} \text{ for } \mathbf{D}^{(i)} = -\left[\left(\frac{\partial Q_{GStS,ind}(\Psi^{(i)}, \Psi^*)}{\partial \Psi^* \partial \Psi^{*T}}\right)^{-1} \frac{\partial Q_{GStS,ind}(\Psi^{(i)}, \Psi^*)}{\partial \Psi^*}\right]_{\Psi^*=\Psi^{(i)}}$$

The study of (Lange 1995) defines this special case of the GEM algorithm as the gradient EM algorithm and proves its local convergence properties, under conditions (1)-(3) from Section 2.5 and in addition assuming that the second derivatives of $Q(\Psi, \Psi^*)$ with respect to Ψ^* are continuous in both Ψ and Ψ^* what implies that the matrix $\mathbf{D}^{(i)}$ always exists and the iterates of the gradient algorithm are well defined.

Lemma 5. *(Lange 1995) let the conditions (1)-(3) from Section 2.5 hold and the second derivative of* $Q(\Psi, \Psi^*)$ *with respect to* Ψ^* *is continuous in both* Ψ *and* Ψ^**. If* $\tilde{\Psi}$ *is a local maximum of the likelihood function in Equation (2.8), such that the second derivative is negative definite at* $\tilde{\Psi}$*, then the gradient EM algorithm converges the likelihood function in Equation (2.8) at this points.*

Under the conditions (1)-(3) in Section 2.5, the defined map is proven to be continuous. This property implies that the set of fixed points of the map M is a subset of the set $\mathcal{S}$ defined in Theorem 17. To prove the local convergence of the algorithm one should notice that the GEM algorithm is a the following fixed point equation

$$\tilde{\Psi} = M(\tilde{\Psi}) \approx \underset{\Psi\in\Omega}{\operatorname{argmax}}\, Q(\tilde{\Psi}, \Psi)$$

Since the mapping M, defined by the optimisation algorithm in Subsection 3.5.2, is shown in (Lange 1995) to be continuous, the limit of $\left\{\Psi^{(i)}\right\}_{i\in\mathbb{N}_0}$ is a fixed point of M and consequently

a stationary point of the likelihood function. (Lange 1995) shows that the mapping M satisfies conditions of Theorem 18 (See Propositions 3-6 in (Lange 1995)). We remark that the considered mapping is closed and satisfies the condition (2) from Theorem 17.

The discussion of (Lange 1995) only applies to the mapping which is a single iteration of a gradient descent algorithm. The convergence results for a mapping M which contains all iterations of the optimisation algorithm up to its convergence are not covered in this section. However, we can remark that the iterations of the corresponding GEM algorithm satisfy the conditions of Lemma 4. The generated sequence $\left\{l^{(i)}\right\}_{i\in\mathbb{N}_0}$ is non-decreasing since the gradient based optimisation methods search for ascent directions of the objective function Q_{GStS} which is initialized in the point that is updated in the previous iteration. Consequently, each iteration of the gradient-based optimisation algorithm increases the likelihood of the model or leaves it unchanged.

3.7. Case Study: Investigating the Imapct of Model Assumptions on Principal Components

In the following section we illustrate the impact of different assumptions about the distribution of the data on the resulting principal components of the covariance matrix. The undertaken analysis is based on synthetic data which we generate from different special cases of the model from Equation (2.7). In order to analyse the effects of various sources of heavy-tails we consider the five models denoted by:

M0: the standard Gaussian PPCA model from Section 2.4;

M1: the tStudent IND PPCA model from Subsection 3.2.2 with parameter ν;

M2: the special case of the PPCA model from Equation (2.7) when the latent vector $\mathbf{X}_t$ follows the t-Student distribution with zero mean, identity covariance matrix and degrees of freedom ν_x, the error term $\boldsymbol{\epsilon}_t$ remains Gaussian with zero mean and the homogeneous covariance matrix equal to $\sigma^2\mathbb{I}_d$; the observations are assumed to be independent in time;

M3: the special case of the PPCA model from Equation (2.7) when the latent vector $\mathbf{X}_t$ remains Gaussian with zero mean and identity covariance matrix, but the error term $\boldsymbol{\epsilon}_t$ follows t-Student distribution with zero mean, the homogeneous covariance matrix equal to $\sigma^2\mathbb{I}_d$ and degrees of freedom ν_ϵ; the observations are assumed to be independent in time;

M4: the Special Case 3 of the Generalized Skew tStudent IND PPCA from Subsubsection 3.4.1.3 with degrees of freedom degrees of freedom ν_x and ν_ϵ of $\mathbf{X}_t$ and $\boldsymbol{\epsilon}_t$, respectively.

We simulate 1000 realisations of each of the models **M0**-**M4** considering various input parameters: the dimensionality of a single observation of $\mathbf{Y}_t$ and $\boldsymbol{\epsilon}_t$ $d = 5,\ 10,\ 20$; the dimensionality of a single observation of the latent vector $\mathbf{X}_t$, $k = 2,\ 3,\ 4$; degrees of freedom $\nu,\ \nu_\epsilon,\ \nu_x \in \left\{1, 5, 10, 30, 100\right\}$ which specifies the tails of the corresponding t-Student distributions. All the models are zero mean ($\boldsymbol{\mu} = \mathbf{0}$) and share the same value of σ^2 and projection matrices $\mathbf{W}$ across different d and k.

3.7.1. The Misspecification of the Sample Covariance Matrix

In the first case study we show how different characteristic of the data, given by the models **M1**-**M4**, affect the estimation of the sample covariance matrix. Even though this exercise is

mathematically very simple, it perfectly summarizes the key motivation of this thesis: that the standard Principal Component Methodology do not provide reliable results if the underlying data set exhibits non-Gaussian dynamics. We treat the result from the model **M0** as the the closest to the true eigen decomposition of the covariance matrix. The Figures 3.12-3.16 illustrate how different assumption on the heavy tails of either $\mathbf{X}_t$ or $\boldsymbol{\epsilon}_t$ distort the information conveyed by the sample covariance matrix.

The lower triangles of sample estimators of the covariance matrix of $\mathbf{Y}_t$ under different model assumptions **M0**-**M4** is illustrated in Figure 3.12 and Figure 3.13. The columns wise order of the panels correspond to different dimensionality of the vectors $\boldsymbol{\epsilon}_t$ (d) and $\mathbb{X}_t$ (k) as well as the number of realisation of $\mathbf{Y}_t$ (N) used to estimate the covariance matrix. The row-wise order of the panels summerized the results for particular models **M0**-**M4** given different values of ν, ν_ϵ, ν_x. The highest despondencies values of the sample covariance matrices are obtained for the models **M1**, **M2** and **M4** where the degrees of freedom ν and ν_x are the smallest, especially when the dimensionality of the latent vector $\mathbf{X}_t$ increases. This observation give us an intuition that the heavy-tailed distribution of the latent vector $\mathbf{X}_t$ has more significant impact on the estimation of the covariance matrix that the distribution of the error term $\boldsymbol{\epsilon}_t$. it is additionally ephasise by the results presented in Figure 3.13 which refers to the model M4. Since the model's definition allows to control the heavy-tails of the distribution of $\mathbf{X}_t$ and $\boldsymbol{\epsilon}_t$, we can analyse the impact of combinations of different ν_ϵ, ν_x on the sample estimator of covariance matrices. The panels from the top four rows are characterised by the highest values of the sample covariance matrices and most different from the corresponding sample estimators for the model **M0** from Figure 3.12.

The Figures 3.14- 3.16 show that the intuition provided by Figures 3.12-3.13 might be misleading. By analysing the eigen structure of the investigated sample covariance matrices we obtain the information that the heavy-tailed distribution of $\boldsymbol{\epsilon}_t$ has more impact on the changes of the eigenvectors and the eigenvalues. The Figure 3.14 presents the distance between the first five eigenvectors of sample covariance matrices of models **M1**-**M3** and the corresponding eigenvectors given by the model **M0**. The measure of distance is defined by the Forbinious norm divided dimentionality of $\mathbf{Y}_t$, the parameter d. The x-axis of each panel refers the discrepancy between the eigenvectors per models with components which are assumed to follow the heavy tailed distribution. The colors correspond to the different values of the degree of freedom. The row-wise order corresponds to the sample sizes (N) and the number of an eigenvector (i).

We observe that the highest discrepancies between the first eigenvectors are obtained for the models **M3** and **M4** when the value of the parameter ν_ϵ are small. The discrepancies between the first eigenvectors for the other models are significantly lower, especially if the dimensionality of the observation is small and the number of samples is big. When the dimensionality of the observation is higher, the discrepancies decline. Interestingly, the impact of the heavy-tailed distribution of $\mathbf{X}_t$ is the most significant when both the sample size and the dimensionality of the data are the smallest (or, the proportion between these two numbers). The same comments apply to the analysis of the second eigenvectors however, with one exception. When the dimensionality of the observation is

small, the assumption of the heavy-tailed distribution of $\mathbf{X}_t$ has the highest, increasing importance, especially when k increases as well.

Surprisingly, the model **M1**, where both $\mathbf{X}_t$ and $\boldsymbol{\epsilon}_t$ follows the t-Student IND PPCA with common tail behaviour, does not provide the same results that the models **M3** and **M4** for $\nu = \nu_x$. One you expect that the model when two components have distribution with heavy tails, would break the estimation by sample covariance even more. It indicates that the sample estimator of the covariance matrix is especially sensitive to the different sources of heavy-tailless. Let us recall the stochastic representation of the vectors $\mathbf{X}_t$ and $\boldsymbol{\epsilon}_t$, under the model **M1**, from Subsection 3.2.2. Both random vectors share the same scaling variable, U_t, which decides the heavy tails of the particular realisation in time. In contrary, the model **M4** assumes independent scaling variables which specify the tails of the distribution of $\mathbf{X}_t$ and $\boldsymbol{\epsilon}_t$. One possible interpretation of this results is that the sample estimator of a covariance matrix is more robust towards the distant observations if both $\mathbf{X}_t$ and $\boldsymbol{\epsilon}_t$ are distant. The sample estimator of the covariance matrix is the most unreliable when the distant realisations of $\boldsymbol{\epsilon}_t$ impact on its diagonal elements.

The effect of the heavy-tailed distribution of $\mathbf{X}_t$ is the highest on the 3,4 and 5 eigenvectors especially when the dimensionality of the observation is small. However, the significance of these eigenvectors is negligible since their corresponding eigenvalues are very close to zero. Figure 3.15 and Figure 3.16 illustrate the decimal logarithms of the obtained eigenvalues, which order is marked on axis-x. The layout of panels is the same as for Figure 3.12 and Figure 3.13. The eigenvalues which are greater that the corresponding values of k should be very close to zero, and consequently their logarithms should be highly negative. We may observe that the majority of the plots shows the decline of the eigenvalues with the increase of their order, which become negligible after kth eigenvalue. The exemptions are panels corresponding to the the most heavy-tailed scenarios of the models **M1**, **M3** and **M4** for the smallest values of ν and ν_ϵ. The outcome additionally supports our reasoning that the sample estimator of the covariance matrix is very sensitive to the different sources of the heavy-tailed components, especially which impact on only a few elements of its eigenstructure.

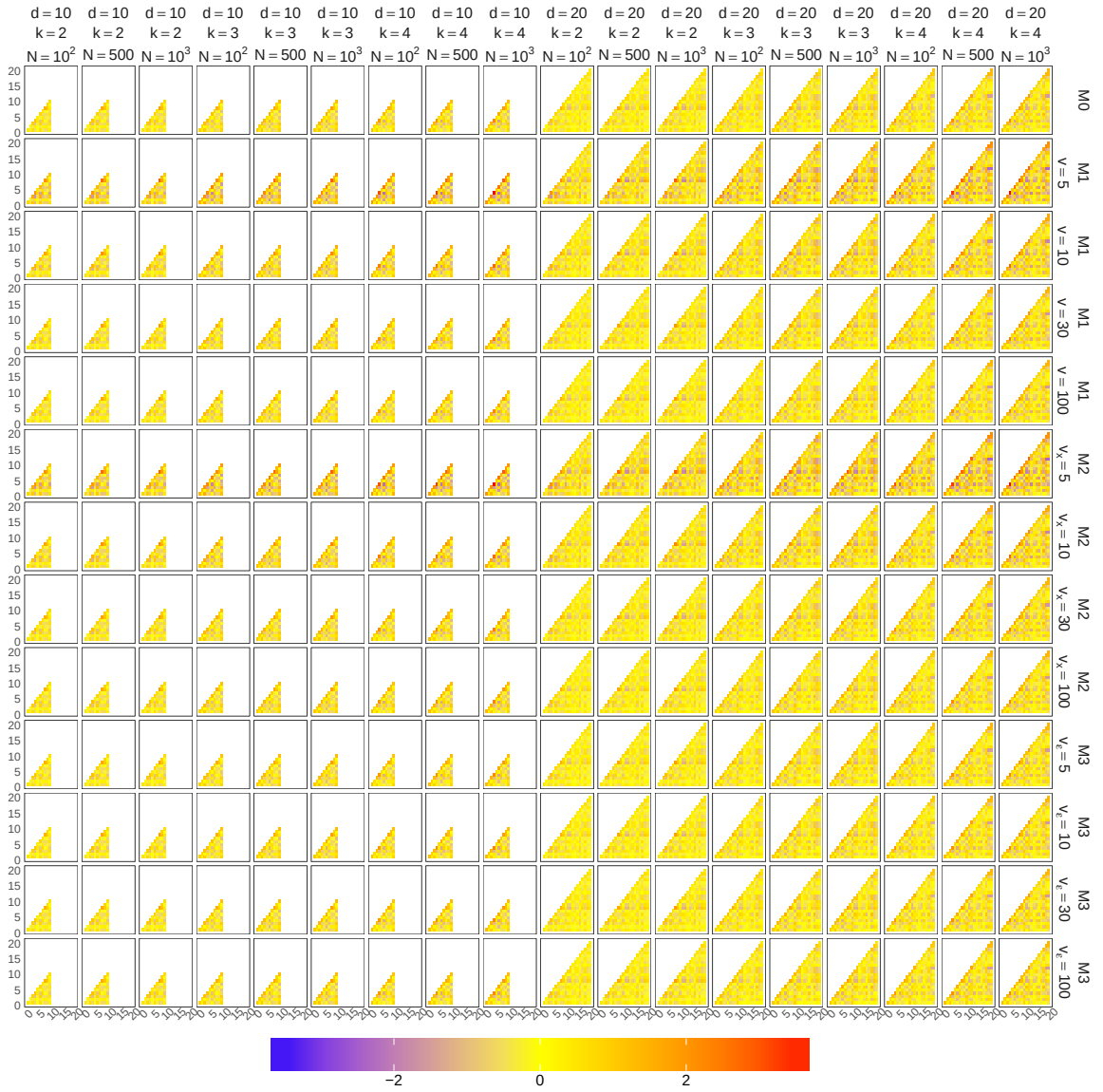


Figure 3.12. The lower triangle of the sample covariance matrix of the data which is simulated from the models **M0**-**M3** (row-wise) with assumed different degrees of freedom. The column-wise order of the panels correspond to results for different dimensionality of $\boldsymbol{\epsilon}_t$ (d) and $\mathbf{X}_t$ (k) and a sample size used to calculated the sample estimator of the covariance matrix of $\mathbf{Y}_t$. The colorscale correspond to the values of the sample covariance matrices.



Figure 3.13. The lower triangle of the sample covariance matrix of the data which is simulated from the model **M4**. The row-wise order of the panels correspond to the data which follows the model of Special Case 3 of the Generalized Skew tStudent IND PPCA across different value of degrees of freedom ν_ϵ and ν_x. The column-wise order of the panels correspond to results for different dimensionality of $\boldsymbol{\epsilon}_t$ (d) and $\mathbf{X}_t$ (k) and a sample size used to calculated the sample estimator of the covariance matrix of $\mathbf{Y}_t$. The colorscale correspond to the values of the sample covariance matrices.

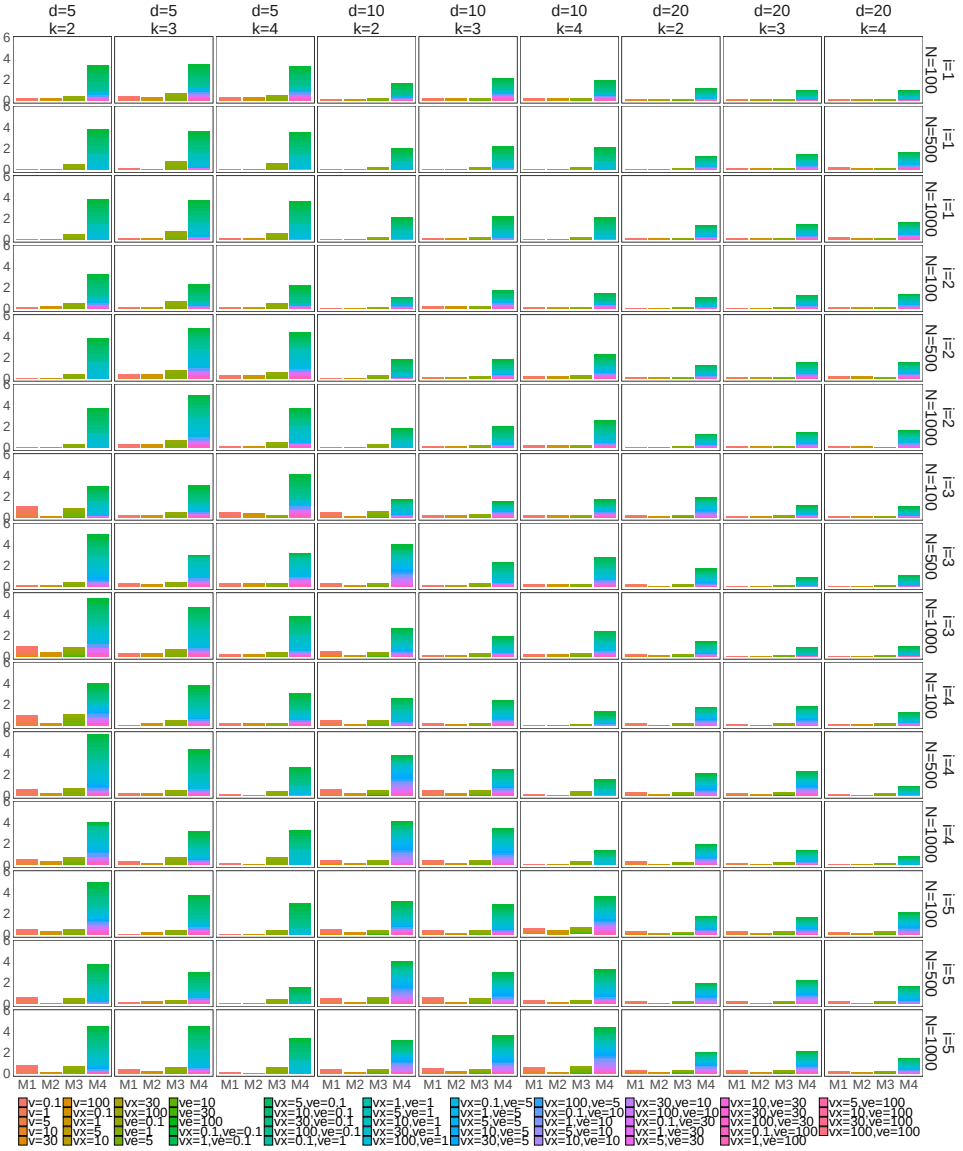


Figure 3.14. The discrepancies (by the Forbinius norm) between the first five eigenvectors of the sample covariance matrix of the models **M1**-**M3** and the corresponding eigenvectors given by the model **M0**. The colors correspond to the different values of the degrees of freedom. The column-wise order of the panels correspond to results for different dimensionality of $\boldsymbol{\epsilon}_t$ (d) and $\mathbf{X}_t$ (k). The row-wise order correspond to the different sample size (N) used to calculated the sample estimator of the covariance matrix of $\mathbf{Y}_t$ and the number of the eigenvector (i).

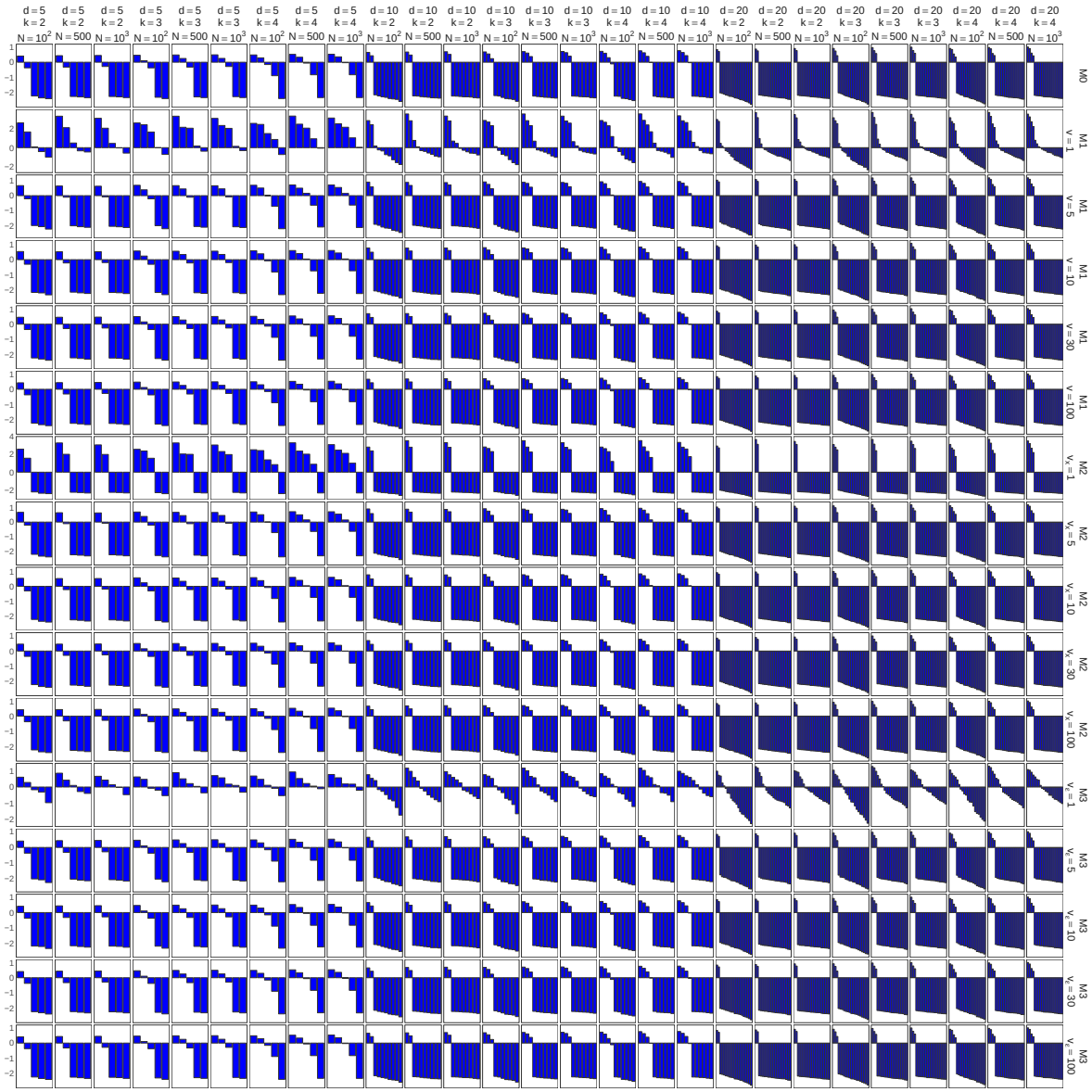


Figure 3.15. The the decimal logarithm ($\log_1 0$) of the subsequent eigenvalues (x-axis) of the sample covariance matrix provided by the models **M0**-**M3** (row-wise) with assumed different degrees of freedom. The column-wise order of the panels correspond to results for different dimensionality of $\boldsymbol{\epsilon}_t$ (d) and $\mathbf{X}_t$ (k) and a sample size used to calculated the sample estimator of the covariance matrix of $\mathbf{Y}_t$. The colorscale correspond to the values of the sample covariance matrices.

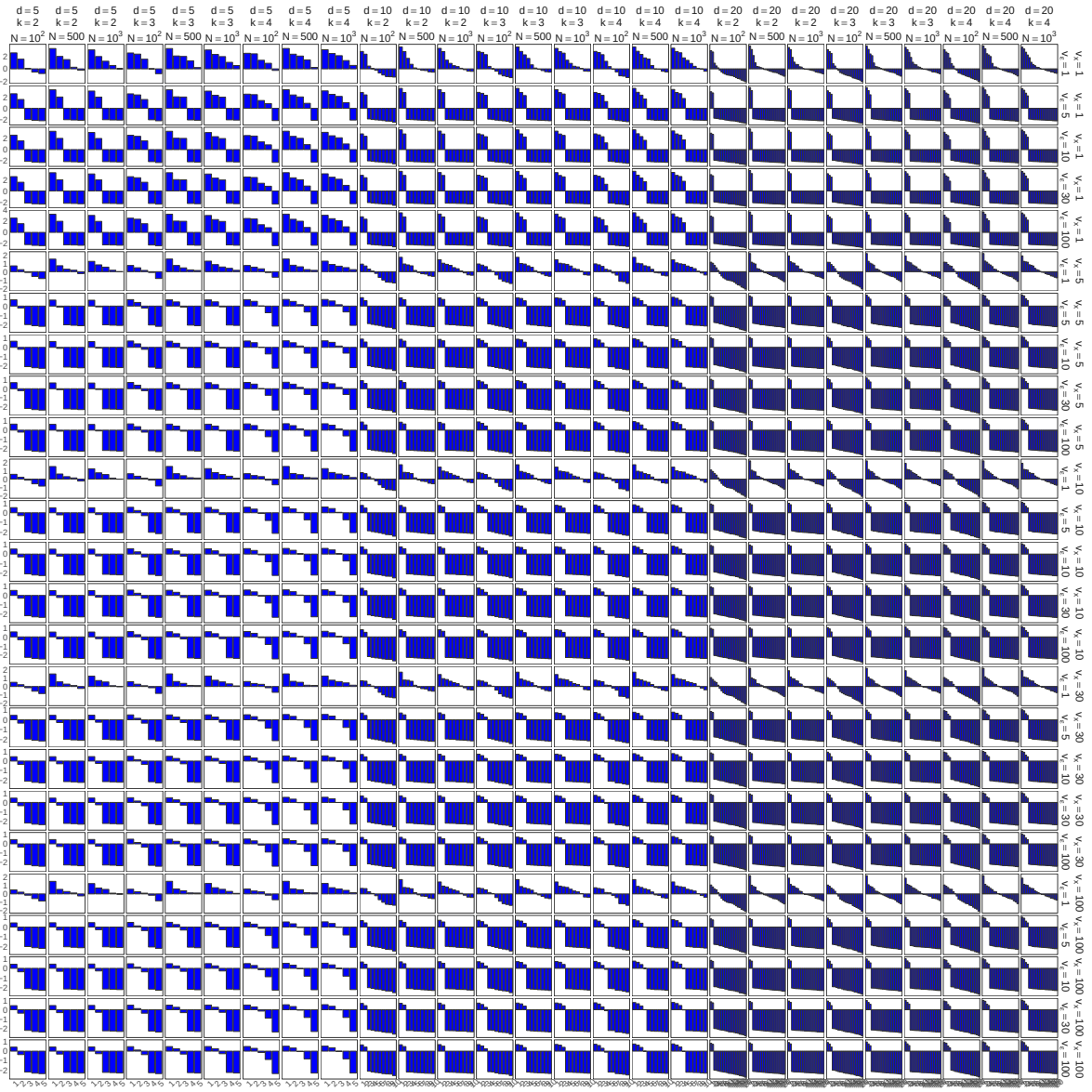


Figure 3.16. The the decimal logarithm ($\log_1 0$) of the subsequent eigenvalues (x-axis) of the sample covariance matrix provided by the models **M0**-**M3** (row-wise) with assumed different degrees of freedom. The column-wise order of the panels correspond to results for different dimensionality of $\boldsymbol{\epsilon}_t$ (d) and $\mathbf{X}_t$ (k) and a sample size used to calculated the sample estimator of the covariance matrix of $\mathbf{Y}_t$. The colorscale correspond to the values of the sample covariance matrices.

3.7.2. Performance of the PPCA frameworks

The following subsection provides the analysis of the estimation performance by various PPCA methods of the the static parameters from the model in Equation (2.7) given the data generated from the model **M0** and different variants of the models **M3** and **M4**. The degrees of freedom for the models **M3** and **M4**. are equal either2 or 100. We examine which of the PPCA frameworks, the Gaussian PPCA, the t-Student IND PPCA or the Special Case 3 of the GStS IND PPCA, provide the best quality of the estimation under various assumption of the tail behaviour of the dataset. Next, we discuss the model selection for the t-Student IND PPCA and the Special Case 3 of the GStS IND PPCA related to the chosen values of the degrees of freedom ν, ν_ϵ and ν_x from the set $\{2, 5, 10, 30, 50, 100\}$.

The set-up of the synthetic data experiment is the following: (1) we simulate once the true values of $\boldsymbol{\mu}$, $\mathbf{W}$ and σ^2 and (2) generate one each set of data for the model **M0** and different variants of the models **M3** and **M4** given different degrees of freedom from $\{2, 100\}$; we simulate the data set for different k, d and N as discussed in the introduction to this section; (3) we run 20 times the three PPCA framework on the generated dataset. Each run of an algorithm results in the estimates of $\boldsymbol{\mu}$, $\mathbf{W}$ and σ^2 and, if applicable, the values of the best values selected for the degrees of freedom from $\{2, 5, 10, 30, 50, 100\}$, which obtained the maximum likelihood of the PPCA framework for a given data set. The initialisation of the feature extraction algorithms is random.

3.7.2.1. The Estimation of the Static Parameters of the Model in Equation (2.7) Under Various Model Assumptions.

We analyse the estimation quality of the parameters $\boldsymbol{\mu}$, σ^2 and the singular vectors of the matrix $\mathbf{W}$ obtained from different PPCA frameworks. The following figures shows the logarithms of average the Frobenius norm of the difference between the true value of the parameter and its estimate. The red lines correspond to the 95% empirical quantiles. The results are given for 20 runs of the algorithms. The x-axis denoted the results for different PPCA frameworks: the Gaussian PPCA (1), t-Student IND PPCA (2) and the Special Case 3 of the GStS IND PPCA. The column-wise order of the panels correspond to results for different dimensionality of $\boldsymbol{\epsilon}_t$ (d) and $\mathbf{X}_t$ (k) and the sample size N. The row-wise order correspond to the variants of the models generating the sample set.

Figure 3.17 and Figure 3.18 illustrate the estimation quality for the vector of interceptions, $\boldsymbol{\mu}$. We notice that general for the small dimensionality and high sample size, especially when the parameters $k = 2$, the three PPCA methods provide comparable results for estimation of t$\boldsymbol{\mu}$ when the data is generated from the models **M0** and **M3**, regardless of the heavy tails of $\boldsymbol{\epsilon}_t$. It is shown, that in the case of high dimensionality of the observation and small sample size, the estimation provided by the Special Case 3 of the GStS IND PPCA and the Gaussian PPCA are the best for low values of ν_ϵ. For less heavy tailed data, t-Student IND PPCA performs better. When the dimensionality d is comparable to the number of the latent variables in the PPCA model and the sample size is small, the t-Student IND PPCA provides the highest accuracy. The quality of the estimation by the Gaussian PPCA drops when we look at the results presented in Figure 3.18 . When the profile of heavy tails of $\mathbf{X}_t$ is the highest (the lowest values of ν_x, the the Special Case 3 of the GStS IND PPCA is the most accurate, especially when the sample size is high. The t-Student IND PPCA performs better for lower sample sizes in high dimensional case.

Figures 3.19, 3.20, 3.21, 3.22 show the discrepancy between the first and second singular vector of the true projection matrix $\mathbf{W}$ and its estimate. We remark that the estimation of these vectors is more sensitive to the initialisation of the EM algorithm that the estimation of the interceptions. In majority of cases, the accuracy of the estimation is comparable in low dimensional case and $k = 2$. We notice that the Special Case 3 of the GStS IND PPCA performs better that the other algorithms when the sample size and the value of parameter$k$ increase for the data generated by the models **M0** and **M3**. For the model **M4**, the third PPCA framework is the most effective, when the $\boldsymbol{\epsilon}_t$ has

the most heavy-tailed distribution, the dimensionality of the data is the highest and for $k = 2$. Also, it is very accurate for less heavy-tailed data. The accuracy of the Special Case 3 of the GStS IND PPCA improves for the estimation of the second singular vector.

The results of the estimation of σ^2 are presented in Figures 3.23 and 3.24. It is fair to say that in majority of cases, the estimation provided by the Special Case 3 of the GStS IND PPCA is the most accurate, especially for the most heavy-tailed distributions of $\boldsymbol{\epsilon}_t$.

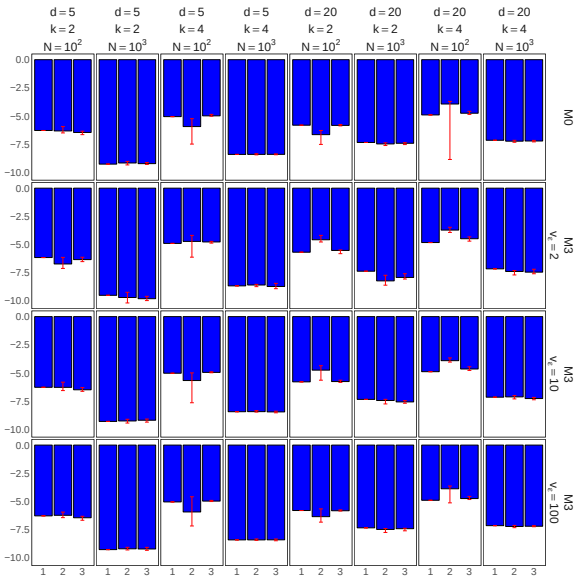


Figure 3.17. The average discrepancies (blue bars, the decimel logarithms of the Frobenius norm) with 95% empirical quantiles (the red lines) between the true $\boldsymbol{\mu}$ and its estimate obtained by three PPCA frameworks (x-axis): the Gaussian PPCA (1), t-Student IND PPCA (2) and the Special Case 3 of the GStS IND PPCA given the data generated from the models **M0** and **M3** (different values of ν_ϵ). The column-wise order of the panels correspond to results for different dimensionality of $\boldsymbol{\epsilon}_t$ (d) and $\mathbf{X}_t$ (k) and the sample size N. The row-wise order correspond to the variants of the models generating the sample set.

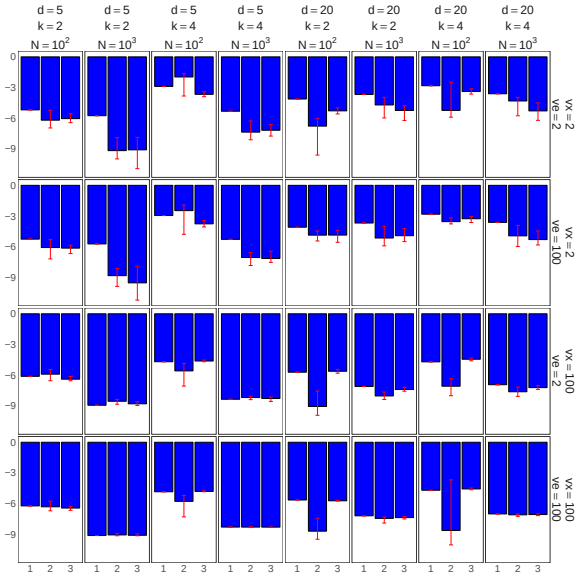


Figure 3.18. The average discrepancies (blue bars, the decimal logarithms of the Frobenius norm) with 95% empirical quantiles (the red lines) between the true $\boldsymbol{\mu}$ and its estimate obtained by three PPCA frameworks (x-axis): the Gaussian PPCA (1), t-Student IND PPCA (2) and the Special Case 3 of the GStS IND PPCA given the data generated from the model **M4**. The column-wise order of the panels correspond to results for different dimensionality of $\boldsymbol{\epsilon}_t$ (d) and $\mathbf{X}_t$ (k) and the sample size N. The row-wise order correspond to the variants of the model defined by different combinations of values for ν_ϵ and ν_x.

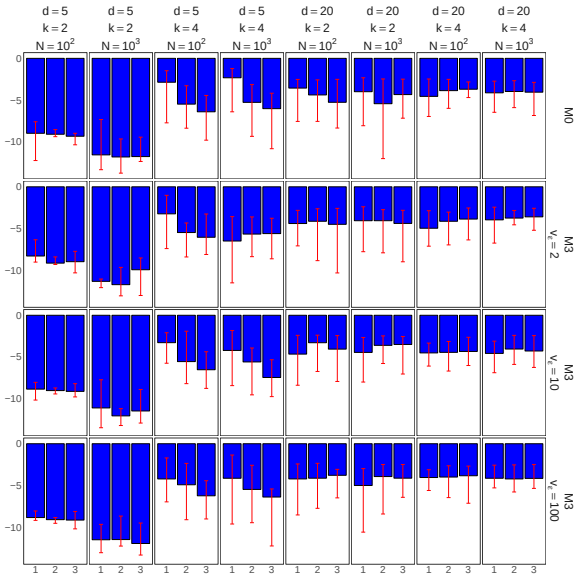


Figure 3.19. The average discrepancies (blue bars, the decimal logarithms of the Frobenius norm) with 95% empirical quantiles (the red lines) between the the first singular vector of the true matrix $\mathbf{W}$ and the first singular vector of its estimate obtained by three PPCA frameworks (x-axis): the Gaussian PPCA (1), t-Student IND PPCA (2) and the Special Case 3 of the GStS IND PPCA given the data generated from the models **M0** and **M3** (different values of ν_ϵ). The column-wise order of the panels correspond to results for different dimensionality of $\boldsymbol{\epsilon}_t$ (d) and $\mathbf{X}_t$ (k) and the sample size N. The row-wise order correspond to the variants of the models generating the sample set.

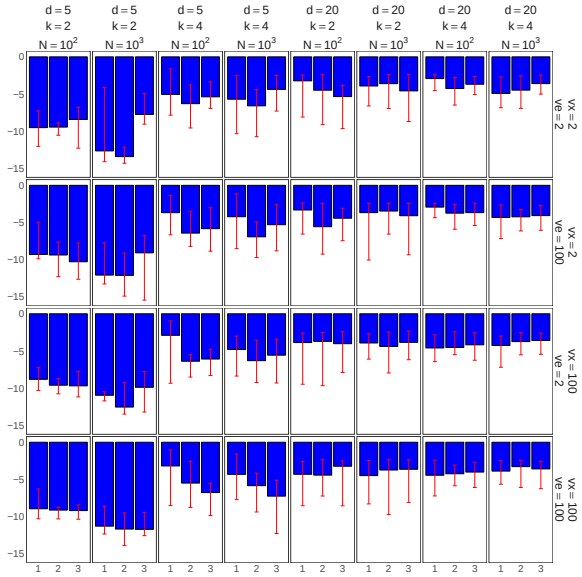


Figure 3.20. The average discrepancies (blue bars, the decimal logarithms of the Frobenius norm) with 95% empirical quantiles (the red lines) between the first singular vector of the true matrix **W** and the first singular vector of its estimate obtained by three PPCA frameworks (x-axis): the Gaussian PPCA (1), t-Student IND PPCA (2) and the Special Case 3 of the GStS IND PPCA given the data generated from the model **M4**. The column-wise order of the panels correspond to results for different dimensionality of $\boldsymbol{\epsilon}_t$ (d) and $\mathbf{X}_t$ (k) and the sample size N. The row-wise order correspond to the variants of the model defined by different combinations of values for ν_ϵ and ν_x.

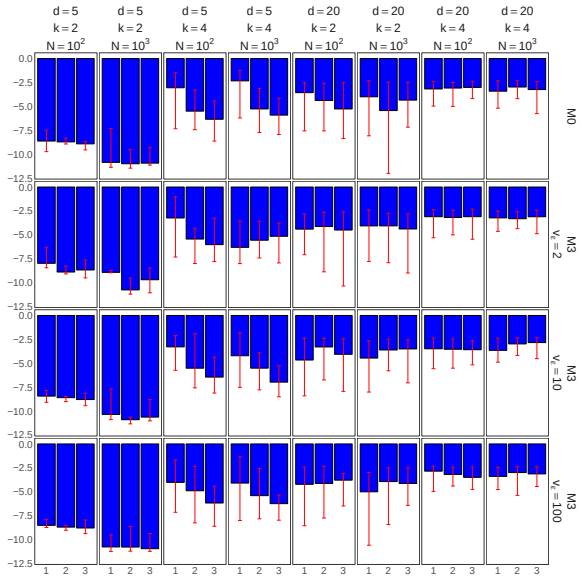


Figure 3.21. The average discrepancies (blue bars, the decimal logarithms of the Frobenius norm) with 95% empirical quantiles (the red lines) between the the second singular vector of the true matrix **W** and the second singular vector of its estimate obtained by three PPCA frameworks (x-axis): the Gaussian PPCA (1), t-Student IND PPCA (2) and the Special Case 3 of the GStS IND PPCA given the data generated from the models **M0** and **M3** (different values of ν_ϵ). The column-wise order of the panels correspond to results for different dimensionality of $\boldsymbol{\epsilon}_t$ (d) and $\mathbf{X}_t$ (k) and the sample size N. The row-wise order correspond to the variants of the models generating the sample set.

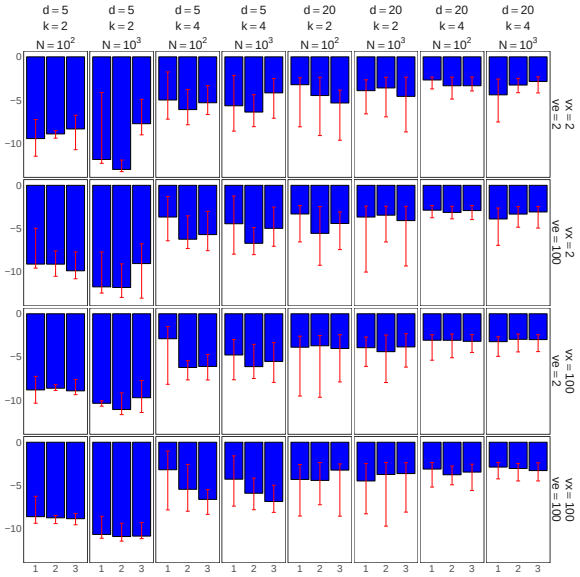


Figure 3.22. The average discrepancies (blue bars, the decimal logarithms of the Frobenius norm) with 95% empirical quantiles (the red lines) between the second singular vector of the true matrix **W** and the second singular vector of its estimate obtained by three PPCA frameworks (x-axis): the Gaussian PPCA (1), t-Student IND PPCA (2) and the Special Case 3 of the GStS IND PPCA given the data generated from the model **M4**. The column-wise order of the panels correspond to results for different dimensionality of ϵ_t (d) and $\mathbf{X}_t$ (k) and the sample size N. The row-wise order correspond to the variants of the model defined by different combinations of values for ν_ϵ and ν_x.

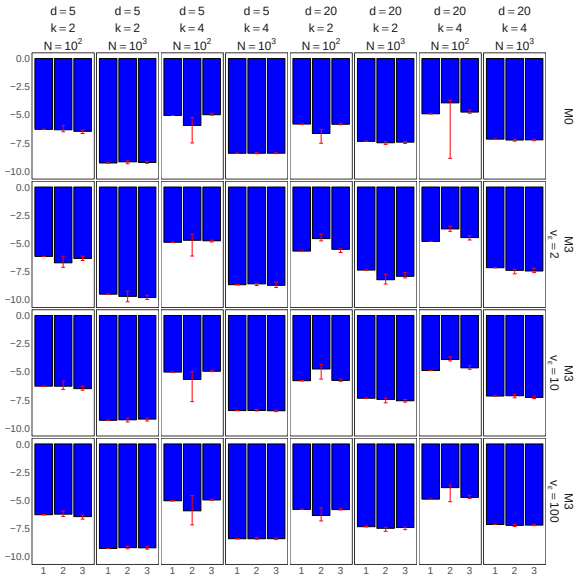


Figure 3.23. The average discrepancies (blue bars, the decimal logarithm of the mean square error) with 95% empirical quantiles (the red lines) between the true σ^2 and its estimate obtained by three PPCA frameworks (x-axis): the Gaussian PPCA (1), t-Student IND PPCA (2) and the Special Case 3 of the GStS IND PPCA given the data generated from the models **M0** and **M3** (different values of ν_ϵ). The column-wise order of the panels correspond to results for different dimensionality of ϵ_t (d) and $\mathbf{X}_t$ (k) and the sample size N. The row-wise order correspond to the variants of the models generating the sample set.

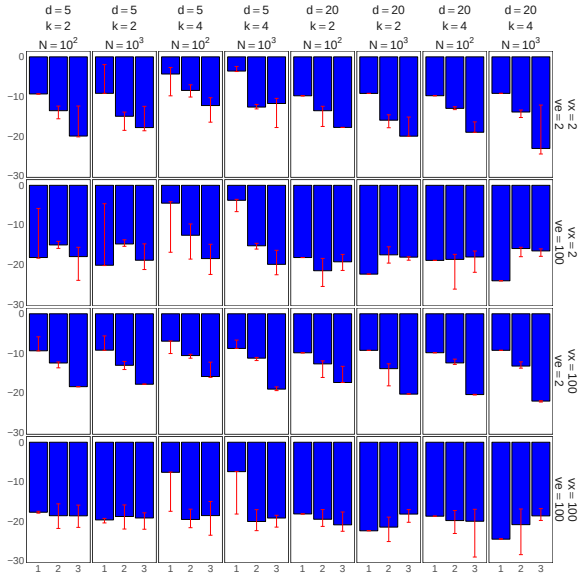


Figure 3.24. The average discrepancies (blue bars, the decimal logarithms of a mean square error) with 95% empirical quantiles (the red lines) between the true σ^2 and its estimate obtained by three PPCA frameworks (x-axis): the Gaussian PPCA (1), t-Student IND PPCA (2) and the Special Case 3 of the GStS IND PPCA given the data generated from the model **M4**. The column-wise order of the panels correspond to results for different dimensionality of $\boldsymbol{\epsilon}_t$ (d) and $\mathbf{X}_t$ (k) and the sample size N. The row-wise order correspond to the variants of the model defined by different combinations of values for ν_ϵ and ν_x.

3.7.2.2. Results of the Model Selection for the Degrees of Freedom: t-Student IND PPCA and Special Case 3 GStS IND PPCA

In the last part of this section, we show the results for the model selection of the two PPCA frameworks: the t-Student IND PPCA and the Special Case 3 of the GStS IND PPCA. Given variants of the tail structure of the the generated data sets, we select the best values of the degrees of freedom for these two feature extraction algorithms form the set from the set $\{2, 5, 10, 30, 50, 100\}$. Given 20 runs of the algorithms, for each variant of the data set, we collect 20 values of the selected degrees of freedom for the t-Student IND PPCA and the Special Case 3 of the GStS IND PPCA.

Figures 3.25 and 3.26 show the histograms of the selected values for ν. We see that the choice of the best ν reflects the heavy **M0** and **M3** of the dataset: for the data generated from **M0** and **M3** with high values of ν_ϵ, the values of ν are higher, however without strong mode of the choice. When **M3** is characterised by a small value of ν_ϵ, the best ν is in majority the lowest among possible values. The analysis of the choice of ν for the data sets generated by the variants of the model **M4** for majority of cases, except for the case when ν_x is small. We see that the choice for ν is mostly dictated by the heavy-tail of $\boldsymbol{\epsilon}_t$ but naturally when the dimensionality of the observation is close to the value of k or the dimensionality is low, the tail behaviour of $\mathbf{X}_t$ starts to dominate the choice of ν. We see that the choice of the ν for the t-Student IND PPCA is the most sensitive to the initialisation step when d is close to k.

Figures 3.27 and 3.28 show the choices for ν_ϵ and ν_x for the Special Case 3 of the GStS IND PPCA. We can notice that the the most accurately are chosen the degrees of freedom for the models with most heavy-tail distribution such as **M3** with $\nu_\epsilon = 2$ or **M4** where either $\nu_\epsilon = 2$ or $\nu_x = 2$. The

corresponding optimal degree so of freedom chosen for the Special Case 3 of the GStS IND PPCA match the same values.

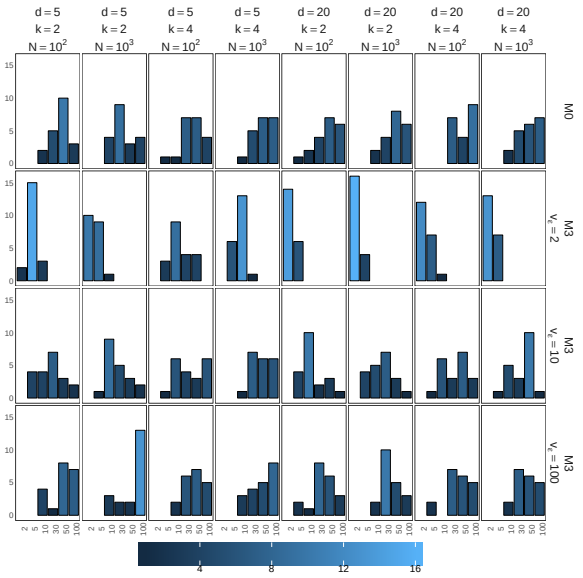


Figure 3.25. The histogram of chosen value for ν in t-Student IND PPCA given 20 simulations and the data generated from the model **M3** (different values of ν_ϵ). The column-wise order of the panels correspond to results for different dimensionality of $\boldsymbol{\epsilon}_t$ (d) and $\mathbf{X}_t$ (k) and the sample size N. The row-wise order correspond to the variants of the model model**M3** generating the sample set. The colors correspond to the frequency of the choice for the degrees of freedom.

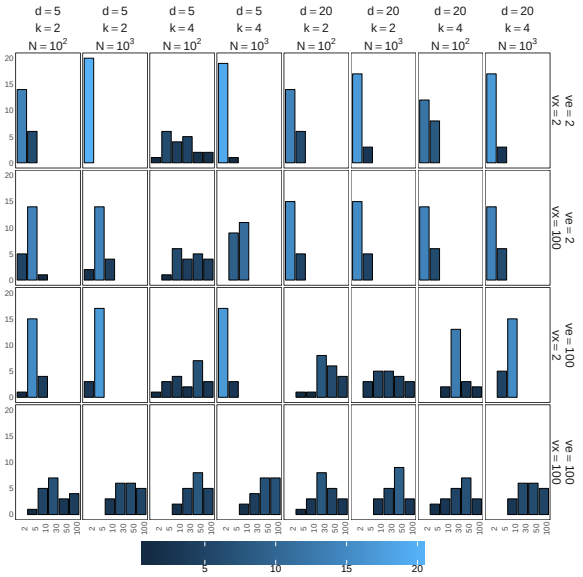


Figure 3.26. The histogram of chosen value for ν in t-Student IND PPCA given 20 simulations and the data generated from the different variants of the model **M4** (different values of ν_ϵ and ν_x). The column-wise order of the panels correspond to results for different dimensionality of $\boldsymbol{\epsilon}_t$ (d) and $\mathbf{X}_t$ (k) and the sample size N. The row-wise order correspond to the variants of the model model**M4** generating the sample set. The colors correspond to the frequency of the choice for the degrees of freedom.

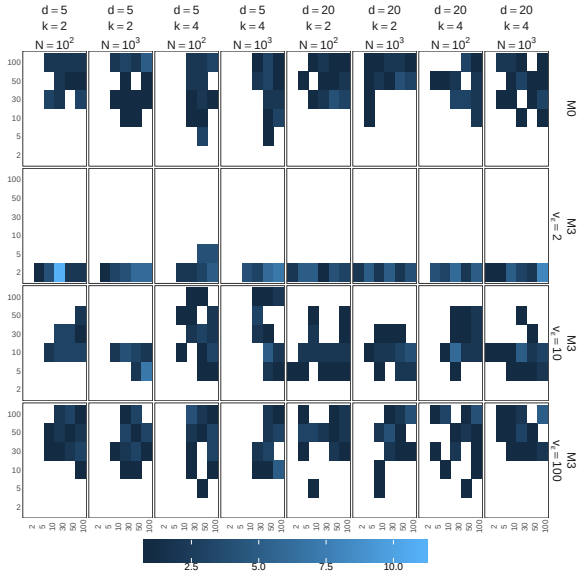


Figure 3.27. The heatmap of chosen values for ν_ϵ (y-axis) and ν_x (x-axis) in Special Case 3 of the GStS IND PPCA given 20 simulations and the data generated from the models **M3** (different values of ν_ϵ). The column-wise order of the panels correspond to results for different dimensionality of $\boldsymbol{\epsilon}_t$ (d) and $\mathbf{X}_t$ (k) and the sample size N. The row-wise order correspond to the variants of the model model**M3** generating the sample set. The colors correspond to the frequency of the choice for the degrees of freedom.

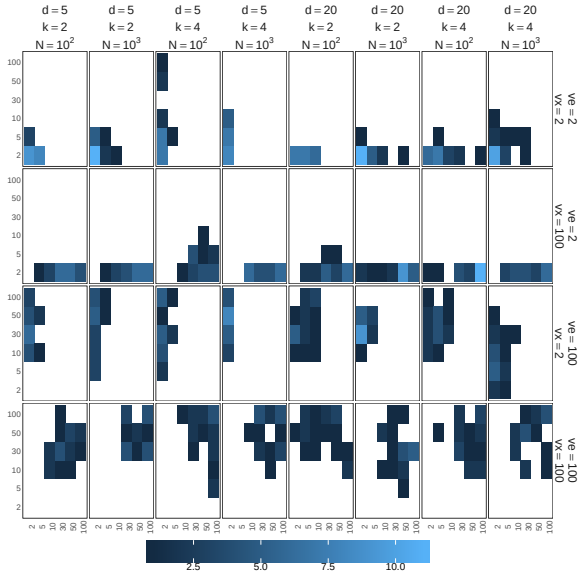


Figure 3.28. The heatmap of chosen values for ν_ϵ (y-axis) and ν_x (x-axis) in Special Case 3 of the GStS IND PPCA given 20 simulations and the data generated from the different variants of the model **M4** (different values of ν_ϵ and ν_x). The column-wise order of the panels correspond to results for different dimensionality of $\boldsymbol{\epsilon}_t$ (d) and $\mathbf{X}_t$ (k) and the sample size N. The row-wise order correspond to the variants of the model model**M4** generating the sample set. The colors correspond to the frequency of the choice for the degrees of freedom.

4. Robust Probabilistic Principal Component Analysis in Presence of Missing Data

Until now, we assumed that the data set, which we analyse, does not contain any missing observations. However, in many dataset there are numerous types of missing data. This is, therefore, an important aspect to address in the feature extraction. In the following chapter we discuss our important extension not previously explored in the existing literature, that allows to deal with general missingness structures in the response vectors $\mathbf{Y}_t$ over time. The Probabilistic Principal Component Analysis is especially efficient in a presence of missing data. We show how to model the incompleteness of the observations in developed frameworks.

We start with discussing the concept of missigness, and address the questions of how this characteristics of the data can be incorporate into the EM algorithms of our models and show how one can extend the related assumptions. Each of the steps of the EM algorithms for PPCA models introduced in Chapter 3 are redefined with consideration of the missing observations. We provide equivalent theorems specifying the E-step and M-step for Gaussian PPCA, t-Student PPCA and Generalized Skew t-Student PPCA in the incomplete data setting in Section 4.2, Section 4.3 and Section 4.4, respectively.

To achieve this we introduce the new notation. Let us assume that the random vector $\mathbf{Y}_t$ is partitioned into two subvectors, one which contains observed values $\mathbf{Y}_t^o$ and the second which indicates missing entries $\mathbf{Y}_t^m$, such that $\mathbf{Y}_t = \left[\mathbf{Y}_t^o, \mathbf{Y}_t^m\right]$. We denote d_o as the number of observed elements of the vector $\mathbf{Y}_t$ and $d_m = d - d_o$ the number of missing entries in time t. The numbers d_o and d_m can vary over time.

In the incomplete-data case related sections we denote by $\mathbf{W}_o$ and $\mathbf{W}_m$ the $d_o \times k$ and $d_m \times k$ non-squared submatrices of $\mathbf{W}$ with corresponding rows to the elements of the vector $\mathbf{Y}_t$ which are observed and missing, respectively. In general, by lower index o and m, we further refer to the elements of some objects corresponding to observed and missing values of $\mathbf{Y}_t$, respectively.

4.1. Novel Extension of Probabilistic Principal Component Analysis for Missing Data Setting

Let us define the indicator random variable $\mathbf{R}_t$ which decides which entries of $\mathbf{Y}_t$ are missing denoting them by 1, otherwise 0. Recall, that a single observation consists of the pair $[\mathbf{Y}_t^o, \mathbf{R}_t]$ with distribution parameters $[\Psi, \Theta]$ respectively. We assume the parameters to be distinct. The likelihood of parameters is proportional to the conditional probability $\mathbf{Y}_t^o, \mathbf{R}_t | \Psi, \Theta$ that is

$$\pi_{\mathbf{Y}_t^o, \mathbf{R}_t | \Psi, \Theta}\left(\mathbf{y}_t^o, \mathbf{r}_t\right) = \int \pi_{\mathbf{Y}_t^o, \mathbf{Y}_t^m, \mathbf{R}_t | \Psi, \Theta}\left(\mathbf{y}_t^o, \mathbf{y}_t^m, \mathbf{r}_t\right) d\mathbf{y}_t^m = \int \pi_{\mathbf{R}_t | \mathbf{Y}_t, \Psi, \Theta}\left(\mathbf{r}_t\right) \pi_{\mathbf{Y}_t | \Psi, \Theta}\left(\mathbf{y}_t\right) d\mathbf{y}_t^m \tag{4.1}$$

In our study, we assume the pattern of missing data to be MAR - missing at random as defined in (Little and Rubin 2002). The assumptions imposes the indicator variable $\mathbf{R}_t$ to be independent of of the value of missing data. Then the vector $\mathbf{Y}_t$ which is MAR satisfies

$$\pi_{\mathbf{R}_t|\mathbf{Y}_t,\Psi}(\mathbf{r}_t) = \pi_{\mathbf{R}_t|\mathbf{Y}_t^o,\Psi}(\mathbf{r}_t)$$

what results in

$$\pi_{\mathbf{Y}_t^o,\mathbf{R}_t|\Psi,\Theta}\left(\mathbf{y}_t^o\right) = \pi_{\mathbf{R}_t|\mathbf{Y}_t^o,\Theta}\left(\mathbf{r}_t\right) \int \pi_{\mathbf{Y}_t|\Psi}\left(\mathbf{y}_t\right) d\mathbf{y}_t^m = \pi_{\mathbf{R}_t|\mathbf{Y}_t^o,\Theta}\left(\mathbf{r}_t\right) \pi_{\mathbf{Y}_t^o|\Psi}\left(\mathbf{y}_t^o\right).$$

Under the MAR assumption, the estimation of Ψ via maximum likelihood of the joint distribution $\mathbf{Y}_t^o, \mathbf{R}_t|\Psi,\Theta$ is equivalent to the maximisation of the likelihood of the marginal distribution $\mathbf{Y}_t^o|\Psi$. Hence, we do not worry about the distribution of the indicator random variable $\mathbf{R}_t$ and the joint distribution of $\mathbf{Y}_t^o$ and $\mathbf{R}_t$. If the assumption about MAR does not hold, one needs to solve the integral from Equation (4.1) in order to maximize the joint likelihood in the corresponding EM algorithm.

4.2. Extending EM Algorithm for Gaussian Probabilistic Principal Component Analysis to Incomplete Data Setting

As an illustration, we begin with extending the standard EM algorithm for Gaussian PPCA discussed in Section 2.4 to incomplete data setting. Theorem 19 and Theorem 20 define the E-step and M-step of the algorithm in a presence of missing data which are equivalent to Theorem 4 and Theorem 5 when the response is fully process is observed.

Theorem 19. *Given the partition of the observation vector* $\mathbf{Y}_t = \left[\mathbf{Y}_t^o, \mathbf{Y}_t^m\right]$ *modelled by Gaussian PPCA model specified in Section 2.4, the E-step of the EM algorithm for Gaussian PPCA in the presence of missing values given N realisations of the observation vector* $\mathbf{Y}_t^o$ *denoted by* $\mathbf{y}_{1:N}^o = \left\{\mathbf{y}_1^o, \ldots, \mathbf{y}_N^o\right\}$ *is defined as*

$$\begin{aligned}
\tilde{Q}\left(\Psi, \Psi^*\right) &= \mathbb{E}_{\mathbf{Y}_{1:N}^m, \mathbf{X}_{1:N}|\mathbf{Y}_{1:N}^o,\Psi}\left[\log\left(\pi_{\mathbf{Y}_{1:N},\mathbf{X}_{1:N}|\Psi^*}(\mathbf{Y}_{1:N},\mathbf{X}_{1:N})\right)\right] \\
&= -\frac{1}{2}\sum_{t=1}^{N}\left\{(d+k)\log 2\pi + d\log\sigma^{*2} + \frac{1}{\sigma^{*2}}\left(tr\left\{\mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t|\mathbf{Y}_t^o,\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right]\right\} - 2\mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t|\mathbf{Y}_t^o,\Psi}\left[\mathbf{Y}_t\right]\boldsymbol{\mu}^{*T}\right.\right. \\
&\quad + \boldsymbol{\mu}^*\boldsymbol{\mu}^{*T} - 2tr\left\{\mathbf{W}^*\mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t|\mathbf{Y}_t^o,\Psi}\left[\mathbf{X}_t^T\mathbf{Y}_t\right]\right\} + 2\mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t|\mathbf{Y}_t^o,\Psi}\left[\mathbf{X}_t\right]\mathbf{W}^{*T}\boldsymbol{\mu}^{*T}\left.\right) \\
&\quad + Tr\left\{\left(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{W}^* + \mathbb{I}_k\right)\mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t|\mathbf{Y}_t^o,\Psi}\left[\mathbf{X}_t^T\mathbf{X}_t\right]\right\}\right\},
\end{aligned}$$

for the corresponding moments of the conditional distribution $\mathbf{Y}_t^m, \mathbf{X}_t | \mathbf{Y}_t^o, \Psi$

$$
\begin{aligned}
\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\Psi}\left[\mathbf{Y}_t\right]_{1\times d} &= \begin{bmatrix} \mathbf{Y}_t^o \\ \boldsymbol{\mu}_m + \left(\mathbf{Y}_t^o - \boldsymbol{\mu}_o\right)\left(\mathbf{W}_o\mathbf{W}_o^T + \sigma^2\mathbb{I}_{d_o}\right)^{-1}\mathbf{W}_o\mathbf{W}_m^T \end{bmatrix}, \\
\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right]_{d\times d} &= \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{mm} - \mathbf{W}_m\mathbf{W}_o^T\mathbf{C}_{oo}^{-1}\mathbf{W}_o\mathbf{W}_m^T \end{bmatrix} + \mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\Psi}\left[\mathbf{Y}_t\right]^T\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\Psi}\left[\mathbf{Y}_t\right], \\
\mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t|\mathbf{Y}_t^o,\Psi}\left[\mathbf{X}_t\right]_{1\times k} &= \left(\mathbb{E}_{\mathbf{Y}_t|\mathbf{Y}_t^o,\Psi}\left[\mathbf{Y}_t\right] - \boldsymbol{\mu}\right)\mathbf{W}\mathbf{M}^{-1}, \\
\mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t|\mathbf{Y}_t^o,\Psi}\left[\mathbf{X}_t^T\mathbf{X}_t\right]_{k\times k} &= \sigma^2\mathbf{M}^{-1} + \mathbf{M}^{-1}\mathbf{W}^T\left(\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\Psi}\left[\left(\mathbf{Y}_t - \boldsymbol{\mu}\right)^T\left(\mathbf{Y}_t - \boldsymbol{\mu}\right)\right]\mathbf{W}\mathbf{M}^{-1}, \right. \\
\mathbb{E}_{\mathbf{X}_t,\mathbf{Y}_t^m|\mathbf{Y}_t^o,\Psi}\left[\mathbf{X}_t^T\mathbf{Y}_t\right]_{k\times d} &= \left(\mathbf{W}^T\mathbf{W} + \sigma^2\mathbb{I}_k\right)^{-1}\mathbf{W}^T\left(\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right] - \boldsymbol{\mu}^T\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\Psi}\left[\mathbf{Y}_t\right]\right).
\end{aligned}
$$

where $\mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbb{I}_d$ *and* $\mathbf{M} = \mathbf{W}^T\mathbf{W} + \sigma^2\mathbb{I}_k$.

Proof. The proof of the Theorem 19 is given in Appendix A.2 in Subsection A.2.3. □

Theorem 20. *The maximizers of the function* $\tilde{Q}\left(\Psi, \Psi^*\right)$ *defined in Theorem 19 are the solution to the following system of the problems* $\nabla_{\Psi^*}\tilde{Q} = \mathbf{0}$ *and are given by*

$$
\begin{aligned}
\boldsymbol{\mu}^* &= \bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o;\Psi)\left(\mathbb{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\right) + \boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}, \\
\mathbf{W}^* &= \bar{\mathbf{C}}_{\boldsymbol{\mu},\boldsymbol{\mu}^*}(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}\mathbf{M}^{-1}\left(\sigma^2\mathbf{M}^{-1} + \mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\boldsymbol{\mu}}(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}\mathbf{M}^{-1}\right)^{-1}, \\
\sigma^{*2} &= \frac{1}{d}\,Tr\left\{\bar{\mathbf{C}}_{\boldsymbol{\mu}^*}(\mathbf{y}_{1:N}^o;\Psi) - 2\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\boldsymbol{\mu},\boldsymbol{\mu}^*}(\mathbf{y}_{1:N}^o;\Psi) + \mathbf{W}^{*T}\mathbf{W}^*\left(\sigma^2\mathbf{M}^{-1} + \mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\boldsymbol{\mu}}(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}\mathbf{M}^{-1}\right)\right\},
\end{aligned}
$$

where

$$
\begin{aligned}
\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o;\Psi) &= \frac{1}{N}\sum_{t=1}^{N}\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\Psi}\left[\mathbf{Y}_t\right], \\
\bar{\mathbf{S}}^m(\mathbf{y}_{1:N}^o;\Psi) &= \frac{1}{N}\sum_{t=1}^{N}\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\Psi}\left[\mathbf{Y}_t\right]^T\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\Psi}\left[\mathbf{Y}_t\right], \\
\bar{\mathbf{Q}} &= \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_m\mathbf{W}_m^T + \sigma^2\mathbb{I}_{d_m} - \mathbf{W}_m\mathbf{W}_o^T\left(\mathbf{W}_o\mathbf{W}_o^T + \sigma^2\mathbb{I}_{d_o}\right)^{-1}\mathbf{W}_o\mathbf{W}_m^T \end{bmatrix}, \\
\bar{\mathbf{C}}_{\boldsymbol{\mu}}(\mathbf{y}_{1:N}^o;\Psi) &= \bar{\mathbf{S}}^m(\mathbf{y}_{1:N}^o;\Psi) + \bar{\mathbf{Q}} - \boldsymbol{\mu}^T\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o;\Psi) - \bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o;\Psi)^T\boldsymbol{\mu} + \boldsymbol{\mu}^T\boldsymbol{\mu}, \\
\bar{\mathbf{C}}_{\boldsymbol{\mu}^*}(\mathbf{y}_{1:N}^o;\Psi) &= \bar{\mathbf{S}}^m(\mathbf{y}_{1:N}^o;\Psi) + \bar{\mathbf{Q}} - \boldsymbol{\mu}^{*T}\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o;\Psi) - \bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o;\Psi)^T\boldsymbol{\mu}^* + \boldsymbol{\mu}^{*T}\boldsymbol{\mu}^*, \\
\bar{\mathbf{C}}_{\boldsymbol{\mu},\boldsymbol{\mu}^*}(\mathbf{y}_{1:N}^o;\Psi) &= \bar{\mathbf{S}}^m(\mathbf{y}_{1:N}^o;\Psi) + \bar{\mathbf{Q}} - \boldsymbol{\mu}^{*T}\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o;\Psi) - \bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o;\Psi)^T\boldsymbol{\mu} + \boldsymbol{\mu}^{*T}\boldsymbol{\mu}.
\end{aligned}
$$

4.3. Incomplete Data EM Algorithm for t-Student Probabilistic Principal Component Analysis

In the following section, we define steps of the EM algorithms for t-Student PPCA model from Subsection 3.2.1 in the incomplete data setting which are equivalent to the formulations of the algorithm described in Subsection 3.2.2. The steps corresponding to the EM algorithm for

the t-Student IND PPCA model are given in Subsection 4.3.1 and the steps for Generelized Skew t-Student IID PPCA model are given in Subsection 4.3.2.

4.3.1. EM Algorithm for Probabilistic Principal Component with Independent t-Student Distribution in the Presence of Missing Data

We provide the E-Step and M-step for t-Student IND PPCA in Theorem 21 and Theorem 22 as equivalent steps of EM Algorithm specified in Subsubsection 3.2.2.1 for the PPCA model t-Student IND in presence of missing elements of observation process. In this framework we assume the independent in time probability model of $\mathbf{X}_t$ and $\boldsymbol{\epsilon}_t$. The steps of the corresponding EM algorithm are obtained in the closed forms.

Theorem 21. *Let the observation vector $\mathbf{Y}_t$ be determined by the t-Student IND model's assumption given in Subsection 3.2.1 in Equation (3.19). The E-step of the EM algorithm for Independent t-Student PPCA in the presence of missing values for realisations of N observation vectors $\mathbf{Y}_t = \left[\mathbf{Y}_t^o, \mathbf{Y}_t^m\right]$, denoted by $\mathbf{y}_{1:N}^o = \left\{\mathbf{y}_1^o, \ldots, \mathbf{y}_N^o\right\}$ is given by*

$$
\begin{aligned}
&\tilde{Q}_{TS,ind}(\Psi, \Psi^*) = \mathbb{E}_{\mathbf{Y}_{1:N}^m, \mathbf{X}_{1:N}, U_{1:N} | \mathbf{Y}_{1:N}^o, \Psi}\left[\log \pi_{\mathbf{Y}_{1:N}, \mathbf{X}_{1:N}, U_{1:N} | \Psi^*}(\mathbf{Y}_{1:N}, \mathbf{X}_{1:N}, U_{1:N})\right] \\
&= -\frac{1}{2}\sum_{t=1}^{N}\left[\left(d + k - \frac{v^*}{2}\right)\log 2\pi + d\log\sigma^{*2} + \frac{1}{\sigma^{*2}}\, Tr\left\{\mathbb{E}_{\mathbf{Y}_t^m, U_t | \mathbf{Y}_t^o, \Psi}\left[U_t\left(\mathbf{Y}_t - \boldsymbol{\mu}^*\right)^T\left(\mathbf{Y}_t - \boldsymbol{\mu}^*\right)\right]\right\}\right. \\
&\quad - (d + k + v^* - 2)\mathbb{E}_{U_t | \mathbf{Y}_t^o, \Psi}\left[\log U_t\right] - \frac{2}{\sigma^{*2}}\, Tr\left\{\mathbf{W}^*\left(\mathbb{E}_{\mathbf{Y}_t^m, \mathbf{X}_t, U_t | \mathbf{Y}_t^o, \Psi}\left[U_t \mathbf{X}_t^T \mathbf{Y}_t\right] - \mathbb{E}_{\mathbf{Y}_t^m, \mathbf{X}_t, U_t | \mathbf{Y}_t^o, \Psi}\left[U_t \mathbf{X}_t\right]^T \boldsymbol{\mu}^*\right)\right\} \\
&\quad \left. + Tr\left\{\left(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{W}^* + \mathbb{I}_k\right)\mathbb{E}_{\mathbf{Y}_t^m, \mathbf{X}_t, U_t | \mathbf{Y}_t^o, \Psi}\left[U_t \mathbf{X}_t^T \mathbf{X}_t\right]\right\} + v^*\mathbb{E}_{U_t | \mathbf{Y}_t^o, \Psi}\left[U_t\right] - 2\log\left(\frac{\left(\frac{v^*}{2}\right)^{\frac{v^*}{2}}}{\Gamma\left(\frac{v^*}{2}\right)}\right)\right],
\end{aligned}
$$

for the corresponding moments of the conditional distribution $\mathbf{Y}_t^m, \mathbf{X}_t, U_t | \mathbf{Y}_t^o, \Psi$

$$
\begin{aligned}
&\mathbb{E}_{U_t | \mathbf{Y}_t^o, \Psi}\left[U_t\right] = \frac{v + d_o}{v + D_t^o}, \\
&\mathbb{E}_{U_t | \mathbf{Y}_t^o, \Psi}\left[\log U_t\right] = \psi\left(\frac{v + d_o}{2}\right) - \log\left(\frac{v + D_t^o}{2}\right), \\
&\mathbb{E}_{\mathbf{Y}_t^m | \mathbf{Y}_t^o, U_t \Psi}\left[\mathbf{Y}_t\right]_{1\times d} = \begin{bmatrix} \mathbf{y}_t^o \\ \boldsymbol{\mu}_m + \left(\mathbf{y}_t^o - \boldsymbol{\mu}_o\right)\left(\mathbf{W}_o\mathbf{W}_o^T + \sigma^2\mathbb{I}_{d_o}\right)^{-1}\mathbf{W}_o\mathbf{W}_m^T \end{bmatrix}, \\
&\mathbb{E}_{\mathbf{Y}_t^m, U_t | \mathbf{Y}_t^o, \Psi}\left[U_t\mathbf{Y}_t\right]_{1\times d} = \mathbb{E}_{U_t | \mathbf{Y}_t^o, \Psi}\left[U_t\right]\mathbb{E}_{\mathbf{Y}_t^m | \mathbf{Y}_t^o, U_t \Psi}(\mathbf{Y}_t), \\
&\mathbb{E}_{\mathbf{Y}_t^m, U_t | \mathbf{Y}_t^o, \Psi}\left[U_t\mathbf{Y}_t^T\mathbf{Y}_t\right]_{d\times d} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{mm} - \mathbf{W}_m\mathbf{W}_o^T\mathbf{C}_{oo}^{-1}\mathbf{W}_o\mathbf{W}_m^T \end{bmatrix} + \mathbb{E}_{U_t | \mathbf{Y}_t^o, \Psi}\left[U_t\right]\mathbb{E}_{\mathbf{Y}_t^m | \mathbf{Y}_t^o, U_t, \Psi}\left[\mathbf{Y}_t\right]^T\mathbb{E}_{\mathbf{Y}_t^m | \mathbf{Y}_t^o, U_t, \Psi}\left[\mathbf{Y}t\right], \\
&\mathbb{E}_{\mathbf{Y}_t^m, \mathbf{X}_t, U_t | \mathbf{Y}_t^o, \Psi}\left[U_t\mathbf{X}_t\right]_{1\times k} = \mathbb{E}_{U_t | \mathbf{Y}_t^o, \Psi}\left[U_t\right]\mathbb{E}_{\mathbf{Y}_t^m | \mathbf{Y}_t^o, U_t \Psi}\left[\mathbf{Y}_t\right]\mathbf{W}\mathbf{M}^{-1}, \\
&\mathbb{E}_{\mathbf{Y}_t^m, \mathbf{X}_t, U_t | \mathbf{Y}_t^o, \Psi}\left[U_t\mathbf{X}_t^T\mathbf{X}_t\right]_{k\times k} = \sigma^2\mathbf{M}^{-1} + \mathbf{M}^{-1}\mathbf{W}^T\mathbb{E}_{\mathbf{Y}_t^m, U_t | \mathbf{Y}_t^o, \Psi}\left[U_t\left(\mathbf{Y}_t - \boldsymbol{\mu}\right)^T\left(\mathbf{Y}_t - \boldsymbol{\mu}\right)\right]\mathbf{W}\mathbf{M}^{-1}, \\
&\mathbb{E}_{\mathbf{Y}_t^m, U_t | \mathbf{Y}_t^o, \Psi}\left[U_t\left(\mathbf{Y}_t - \boldsymbol{\mu}\right)^T\left(\mathbf{Y}_t - \boldsymbol{\mu}\right)\right] = \mathbb{E}_{\mathbf{Y}_t^m, U_t | \mathbf{Y}_t^o, \Psi}\left[U_t\mathbf{Y}_t^T\mathbf{Y}_t\right] - 2\boldsymbol{\mu}^T\mathbb{E}_{\mathbf{Y}_t^m, U_t | \mathbf{Y}_t^o, \Psi}\left[U_t\mathbf{Y}_t\right] + \mathbb{E}_{U_t | \mathbf{Y}_t^o, \Psi}\left[U_t\right]\boldsymbol{\mu}^T\boldsymbol{\mu}, \\
&\mathbb{E}_{\mathbf{Y}_t^m, U_t | \mathbf{Y}_t^o, \Psi}\left[U_t\mathbf{Y}_t^T\mathbf{X}_t\right]_{d\times k} = \left(\mathbb{E}_{\mathbf{Y}_t^m, U_t | \mathbf{Y}_t^o, \Psi}\left[U_t\mathbf{Y}_t^T\mathbf{Y}_t\right] - \mathbb{E}_{\mathbf{Y}_t^m, U_t | \mathbf{Y}_t^o, \Psi}\left[U_t\mathbf{Y}_t\right]^T\boldsymbol{\mu}\right)\mathbf{W}\mathbf{M}^{-1},
\end{aligned}
$$

where $\mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbb{I}_d$, $\mathbf{M}_{k\times k} = \mathbf{W}^T\mathbf{W} + \sigma^2\mathbb{I}_k$, $D_t^o(\mathbf{y}_{1:N}^o; \Psi) = \left(\mathbf{y}_t^o - \boldsymbol{\mu}_o\right)\mathbf{C}_{oo}^{-1}\left(\mathbf{y}_t^o - \boldsymbol{\mu}_o\right)^T$ and $\psi(\cdot)$ the digamma function.

Proof. The proof of Theorem 21 is given in Appendix A.3 in Subsection A.3.5. □

Theorem 22. *The maximizers of $\tilde{Q}_{TS,ind}(\Psi, \Psi^*)$ determined in Theorem 21 are the solution to the system of equations given by $\nabla_{\Psi^*}\tilde{Q}_{TS,ind} = \mathbf{0}$. The maximizers with respect to $\boldsymbol{\mu}^*$, $\mathbf{W}^*$ and σ^{*2} can be obtained in closed form according to the following solutions:*

$$\boldsymbol{\mu}^* = \frac{1}{\bar{u}(\mathbf{y}^o_{1:N};\Psi)}\left(\bar{\boldsymbol{\mu}}_{ts}(\mathbf{y}^o_{1:N};\Psi)\left(\mathbb{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\right) + \bar{u}(\mathbf{y}^o_{1:N};\Psi)\boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\right),$$

$$\mathbf{W}^* = \bar{\mathbf{C}}^{ts}_{\mu,\mu^*}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\mathbf{W}\mathbf{M}^{-1}\left(\sigma^2\mathbf{M}^{-1} + \mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}^{ts}_{\mu}(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}\mathbf{M}^{-1}\right)^{-1},$$

$$\sigma^{*2} = \frac{1}{d}\,Tr\left\{\bar{\mathbf{C}}^{ts}_{\mu^*}(\mathbf{y}^o_{1:N};\Psi,\Psi^*) - 2\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}^{ts}_{\mu,\mu^*}(\mathbf{y}^o_{1:N};\Psi,\Psi^*) + \mathbf{W}^{*T}\mathbf{W}^*\left(\sigma^2\mathbf{M}^{-1} + \mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}^{ts}_{\mu}(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}\mathbf{M}^{-1}\right)\right\}$$

where

$$\bar{u}(\mathbf{y}^o_{1:N};\Psi) = \frac{1}{N}\sum_{t=1}^{N}\frac{v+d_o}{v+D^o_t},$$

$$\bar{\boldsymbol{\mu}}_{ts}(\mathbf{y}^o_{1:N};\Psi) = \frac{1}{N}\sum_{t=1}^{N}\frac{v+d_o}{v+D^o_t}\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,U_t,\Psi}(\mathbf{Y}_t),$$

$$\bar{\mathbf{S}}_{ts}(\mathbf{y}^o_{1:N};\Psi) = \frac{1}{N}\sum_{t=1}^{N}\frac{v+d_o}{v+D^o_t(\mathbf{y}^o_{1:N};\Psi)}\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,U_t,\Psi}(\mathbf{Y}_t)^T\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,U_t,\Psi}(\mathbf{Y}_t),$$

$$\bar{\mathbf{Q}} = \begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{W}_m\mathbf{W}_m^T + \sigma^2\mathbb{I}_{d_m} - \mathbf{W}_m\mathbf{W}_o^T\left(\mathbf{W}_o\mathbf{W}_o^T + \sigma^2\mathbb{I}_{d_o}\right)^{-1}\mathbf{W}_o\mathbf{W}_m^T\end{bmatrix},$$

$$\bar{\mathbf{C}}^{ts}_{\mu}(\mathbf{y}^o_{1:N};\Psi) = \bar{\mathbf{S}}_{ts}(\mathbf{y}^o_{1:N};\Psi) + \bar{\mathbf{Q}} - \boldsymbol{\mu}^T\bar{\boldsymbol{\mu}}_{ts}(\mathbf{y}^o_{1:N};\Psi) - \bar{\boldsymbol{\mu}}_{ts}(\mathbf{y}^o_{1:N};\Psi)^T\boldsymbol{\mu} + \bar{u}(\mathbf{y}^o_{1:N};\Psi)\boldsymbol{\mu}^T\boldsymbol{\mu},$$

$$\bar{\mathbf{C}}^{ts}_{\mu^*}(\mathbf{y}^o_{1:N};\Psi,\Psi^*) = \bar{\mathbf{S}}_{ts}(\mathbf{y}^o_{1:N};\Psi) + \bar{\mathbf{Q}} - \boldsymbol{\mu}^{*T}\bar{\boldsymbol{\mu}}_{ts}(\mathbf{y}^o_{1:N};\Psi) - \bar{\boldsymbol{\mu}}_{ts}(\mathbf{y}^o_{1:N};\Psi)^T\boldsymbol{\mu}^* + \bar{u}(\mathbf{y}^o_{1:N};\Psi)\boldsymbol{\mu}^{*T}\boldsymbol{\mu}^*,$$

$$\bar{\mathbf{C}}^{ts}_{\mu,\mu^*}(\mathbf{y}^o_{1:N};\Psi,\Psi^*) = \bar{\mathbf{S}}_{ts}(\mathbf{y}^o_{1:N};\Psi) + \bar{\mathbf{Q}} - \boldsymbol{\mu}^{*T}\bar{\boldsymbol{\mu}}_{ts}(\mathbf{y}^o_{1:N};\Psi) - \bar{\boldsymbol{\mu}}_{ts}(\mathbf{y}^o_{1:N};\Psi)^T\boldsymbol{\mu} + \bar{u}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\boldsymbol{\mu}^{*T}\boldsymbol{\mu},$$

for $\mathbf{M}_{k\times k} = \mathbf{W}^T\mathbf{W} + \sigma^2\mathbb{I}_k$, $D^o_t(\mathbf{y}^o_{1:N};\Psi) = (\mathbf{y}^o_t - \boldsymbol{\mu}_o)\,\mathbf{C}^{-1}_{oo}\,(\mathbf{y}^o_t - \boldsymbol{\mu}_o)^T$ *where* $\mathbf{C}_{oo} = \mathbf{W}_o\mathbf{W}_o^T + \sigma^2\mathbb{I}_{d_o}$.

Proof. The proof of Theorem 22 is given in Appendix A.3 in Subsection A.3.6. □

4.3.2. EM Algorithm for Probabilistic Principal Component with Identical and Conditionally Independent t-Student Distribution in the Presence of Missing Data

Analogously to the previous Subsection, we detail the E-Step and M-step of the EM algorithm for t-Student IID PPCA, given in Theorem 23 and Theorem 24 as equivalent steps of EM Algorithm specified in Subsubsection 3.2.2.1 for the PPCA model t-Student IID in presence of missing data. In this framework we assume the identical probability model of $\mathbf{X}_t$ and $\boldsymbol{\epsilon}_t$ over time which results in conditional independence over time. This restriction still admits closed form solutions.

Theorem 23. *The E-step of the EM Algorithm for the t-Student IID PPCA given N realisations of the observation vector $\mathbf{Y}^o_t$, denoted by $\mathbf{y}^o_{1:N} = \{\mathbf{y}^o_1,\ldots,\mathbf{y}^o_N\}$, is given by*

$$\tilde{Q}_{TS,id}(\Psi,\Psi^*) = \mathbb{E}_{\mathbf{X}_{1:N},\mathbf{Y}^m_{1:N},U|\mathbf{Y}^o_{1:N},\Psi}\left[\log\pi_{\mathbf{Y}_{1:N},\mathbf{X}_{1:N},U|\Psi^*}(\mathbf{Y}_{1:N},\mathbf{X}_{1:N},U)\right]$$
$$= \frac{C_H(\Psi)C_\beta(\mathbf{y}^o_{1:N},\Psi)}{\pi_{\mathbf{Y}^o_{1:N}|\Psi}(\mathbf{y}^o_{1:N})}\left\{\int_{\mathbb{R}_+}\tilde{\pi}_{U|\Psi}(u)\log\pi_{U|\Psi^*}(u)du + \int_{\mathbb{R}_+}\tilde{\pi}_{U|\Psi}(u)\sum_{t=1}^{N}\tilde{w}(\mathbf{y}^o_t,u;\Psi,\Psi^*)\,du\right\},$$

where the scalar C_H is defined as in Lemma A5, the function $\tilde{w} : \mathbb{R}^{d_o \times N} \times \mathbb{R}_+ \to \mathbb{R}$ in Lemma A6, and, the functions $C_\beta : \mathbb{R}^{d_o \times N} \to \mathbb{R}$ and the density $\tilde{\pi}_{U|\Psi}(u)$, are derived in Lemma A7. Each of these ancillary lemmas and function definitions are contained in Appendix A.3.9.

Proof. The proof of Theorem 23 is contained in Appendix A.3.7. □

Theorem 24. *The maximizers of $\tilde{Q}_{TS,id}(\Psi, \Psi^*)$ defined in Theorem 23 are the solution to the system of equations $\nabla_{\Psi^*} \tilde{Q}_{TS,id} = \mathbf{0}$. The maximizers with respect to $\boldsymbol{\mu}^*$, $\mathbf{W}^*$ and σ^{*2} can be obtained in closed form according to the following solutions:*

$$\boldsymbol{\mu}^* = \bar{\boldsymbol{\mu}}(\mathbf{y}^o_{1:N}; \Psi)\left(\mathbf{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\right) + \boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T},$$

$$\mathbf{W}^* = \left(\sigma^2\bar{\mathbf{Q}} + \frac{\alpha}{\beta(\mathbf{y}^o_{1:N}; \Psi)}\bar{\mathbf{C}}_{\mu,\mu^*}(\mathbf{y}^o_{1:N}; \Psi, \Psi^*)\right)\mathbf{W}\mathbf{M}^{-1}\left(\mathbf{M}^{-1}\mathbf{W}^T\left(\sigma^2\bar{\mathbf{Q}} + \frac{\alpha}{\beta(\mathbf{y}^o_{1:N}; \Psi)}\bar{\mathbf{C}}_{\mu}(\mathbf{y}^o_{1:N}; \Psi)\right)\mathbf{W}\mathbf{M}^{-1} + \sigma^2\mathbf{M}^{-1}\right)^{-1},$$

$$\begin{aligned}\sigma^{*2} = \frac{1}{d}\Bigg(&\sigma^2\, Tr\left\{\mathbf{M}^{-1}\mathbf{W}^{*\,T}\mathbf{W}^*\right\} + Tr\left\{\sigma^2\bar{\mathbf{Q}} + \frac{\alpha}{\beta(\mathbf{y}^o_{1:N}; \Psi)}\bar{\mathbf{C}}_{\mu^*}(\mathbf{y}^o_{1:N}; \Psi, \Psi^*)\right\}\\ &- 2\, Tr\left\{\left(\sigma^2\bar{\mathbf{Q}} + \frac{\alpha}{\beta(\mathbf{y}^o_{1:N}; \Psi)}\bar{\mathbf{C}}_{\mu,\mu^*}(\mathbf{y}^o_{1:N}; \Psi, \Psi^*)\right)\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\right\}\\ &+ Tr\left\{\mathbf{M}^{-1}\mathbf{W}^T\left(\sigma^2\bar{\mathbf{Q}} + \frac{\alpha}{\beta(\mathbf{y}^o_{1:N}; \Psi, \Psi^*)}\bar{\mathbf{C}}_{\mu}(\mathbf{y}^o_{1:N}; \Psi)\right)\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{W}^*\right\},\end{aligned}$$

for the scalar $\alpha = \frac{v}{2} + \frac{d_o N}{2}$ and the function $\beta : \mathbb{R}^{d_o \times N} \to \mathbb{R}$

$$\beta(\mathbf{y}^o_{1:N}; \Psi) = \frac{v}{2} + \frac{1}{2\sigma^2}\sum_{t=1}^{N}\left(\mathbf{y}^o_t - \boldsymbol{\mu}_o\right)\mathbf{V}_{oo}\left(\mathbf{y}^o_t - \boldsymbol{\mu}_o\right)^T.$$

defined as in Lemma A7 in Appendix A.3.9, the matrices $\mathbf{V}_{d\times d} = \mathbb{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^T$ and $\mathbf{M}_{k\times k} = \mathbf{W}^T\mathbf{W} + \sigma^2\mathbb{I}_k$, and the following function of $\mathbf{y}^o_{1:N}$,

$$\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,U,\Psi}\left[\mathbf{Y}_t\right] = \begin{bmatrix}\mathbf{y}^o_t\\ \boldsymbol{\mu}_m + \left(\mathbf{y}^o_t - \boldsymbol{\mu}_o\right)\mathbf{V}_{oo}\mathbf{V}^{-1}_{om}\end{bmatrix},$$

$$\bar{\boldsymbol{\mu}}(\mathbf{y}^o_{1:N}; \Psi) = \frac{1}{N}\sum_{t=1}^{N}\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,U,\Psi}\left[\mathbf{Y}_t\right],$$

$$\bar{\mathbf{S}}(\mathbf{y}^o_{1:N}; \Psi) = \frac{1}{N}\sum_{t=1}^{N}\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,U,\Psi}\left[\mathbf{Y}_t\right]^T\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,U,\Psi}\left[\mathbf{Y}_t\right],$$

$$\bar{\mathbf{Q}} = \frac{1}{N}\sum_{t=1}^{N}\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{V}^{-1}_{mm} - \mathbf{V}^{-1}_{mo}\mathbf{V}_{oo}\mathbf{V}^{-1}_{om}\end{bmatrix},$$

$$\bar{\mathbf{C}}_{\mu,\mu^*}(\mathbf{y}^o_{1:N}; \Psi, \Psi^*) = \bar{\mathbf{S}}(\mathbf{y}^o_{1:N}; \Psi) - \boldsymbol{\mu}^{*T}\bar{\boldsymbol{\mu}}(\mathbf{y}^o_{1:N}; \Psi) - \bar{\boldsymbol{\mu}}(\mathbf{y}^o_{1:N}; \Psi)^T\boldsymbol{\mu} + \boldsymbol{\mu}^{*T}\boldsymbol{\mu},$$

$$\bar{\mathbf{C}}_{\mu}(\mathbf{y}^o_{1:N}; \Psi) = \bar{\mathbf{S}}(\mathbf{y}^o_{1:N}; \Psi) - \bar{\boldsymbol{\mu}}(\mathbf{y}^o_{1:N}; \Psi)^T\boldsymbol{\mu} - \boldsymbol{\mu}^T\bar{\boldsymbol{\mu}}(\mathbf{y}^o_{1:N}; \Psi) + \boldsymbol{\mu}^T\boldsymbol{\mu},$$

$$\bar{\mathbf{C}}_{\mu^*}(\mathbf{y}^o_{1:N}; \Psi, \Psi^*) = \bar{\mathbf{S}}(\mathbf{y}^o_{1:N}; \Psi) - \boldsymbol{\mu}^{*T}\bar{\boldsymbol{\mu}}(\mathbf{y}^o_{1:N}; \Psi) - \bar{\boldsymbol{\mu}}(\mathbf{y}^o_{1:N}; \Psi)^T\boldsymbol{\mu}^* + \boldsymbol{\mu}^{*T}\boldsymbol{\mu}^*.$$

Proof. The proof of Theorem 24 is given in Appendix A.3.8. □

4.4. Incomplete Data EM Algorithm for Generalized Skew t-Student Probabilistic principal Component Analysis

In the following section, we define the steps of the EM algorithms for Generalized Skew t-Student PPCA model from Subsection 3.4.1and its special cases specified in Subsubsection 3.4.1.3, in the incomplete data setting. The obtained formulations are equivalent to the algorithm described in Subsection 3.4.2. The steps corresponding to the EM algorithm for the Generalized Skew t-Student IND PPCA model defined in Subsection 3.4.1.1 are given in Subsection 4.4.1 and the steps for Generalized Skew t-Student IID PPCA model defined in Subsection 3.4.1.2 are given in Subsection 4.4.2.

4.4.1. EM Algorithm for Probabilistic Principal Component Analysis with Independent Generalized Skew t-Student Distribution in the Presence of Missing Data

We provide the E-Step and M-step for Generalized Skew t-Student IND PPCA in Theorem 25 and Theorem 26 as equivalent steps of the EM Algorithm specified in Subsubsection 3.4.2.1 for the Generalized Skew t-Student IND PPCA model in the presence of missing elements of observation process. In this framework we assume the independent in time probability model of $\mathbf{X}_t$ and $\boldsymbol{\epsilon}_t$.

Theorem 25. *Let the observation vector $\mathbf{Y}_t$ be modelled as in Equation (2.7) with the latent processes $\boldsymbol{\epsilon}_t$, $\mathbf{X}_t$, $\mathbf{U}_t$ and $\mathbf{V}_t$ following the assumptions of Generalized Skew t-Student IND PPCA given in Subsection 3.4.1 in Equation (3.27) and Equation (3.28). Given N observations of the observed entries of the random vector $\mathbf{Y}_t^o$, denoted by $\mathbf{y}_{1:N}^o := \left[\mathbf{y}_1^o, \ldots, \mathbf{y}_N^o\right]$, the E-step of the Expectation-Maximisation algorithm for Generalized Skew t-Student IND PPCA in the incomplete data setting is specified as follows*

$$\begin{aligned}\tilde{Q}_{GStS,ind}(\Psi, \Psi^*) &= \mathbb{E}_{\mathbf{Y}_{1:N}^m, \mathbf{X}_{1:N}, \mathbf{U}_{1:N}, \mathbf{V}_{1:N} | \mathbf{Y}_{1:N}^o, \Psi}\left[\log \pi_{\mathbf{Y}_{1:N}, \mathbf{X}_{1:N}, \mathbf{U}_{1:N} \mathbf{Y}_{1:N} | \Psi^*}\left(\mathbf{Y}_{1:N}, \mathbf{X}_{1:N}, \mathbf{U}_{1:N}, \mathbf{V}_{1:N}\right)\right] \\ &= \frac{1}{\pi_{\mathbf{Y}_{1:N}^o | \Psi}(\mathbf{y}_{1:N}^o)} \sum_{t=1}^{N} \left\{ I_1(\mathbf{y}_t^o; \Psi, \Psi^*) \prod_{s=1, s \neq t}^{N} I_2(\mathbf{y}_s^o; \Psi) \right\}.\end{aligned}$$

The functions $I_1,\ I_2 : \mathbb{R}^{d_o} \to \mathbb{R}$ are defined as

$$\begin{aligned}I_1(\mathbf{y}_t^o; \Psi, \Psi^*) &:= \int_0^1 \int_0^1 \tilde{v}(\mathbf{y}_t^o, s_{\epsilon,t}, s_{x,t}; \Psi, \Psi^*)\, m(\mathbf{y}_t^o, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t}, \\ I_2(\mathbf{y}_t^o; \Psi) &:= \int_0^1 \int_0^1 m(\mathbf{y}_t^o, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t}.\end{aligned}$$

where $\tilde{v}\left(\mathbf{y}_t^o, s_{\epsilon,t}, s_{x,t}; \Psi, \Psi^*\right) := v\left(\mathbf{y}_t^o, T_\epsilon(s_{\epsilon,t}), T_x(s_{x,t}); \Psi, \Psi^*\right)$ specified in Lemma A11 in Appendix A.4 in Section A.4.5 and the function $m : \mathbb{R}^{d_o} \times [0,1]^2 \longrightarrow \mathbb{R}$ is given by

$$\begin{aligned}m(\mathbf{y}_t^o, s_{\epsilon,t}, s_{x,t}; \Psi) &= e^{-\frac{1}{2}\delta_x \left(\mathbf{D}_{x,t}^{-1} - \sigma^2 \mathbf{M}_t^{-1} \mathbf{W}^T \mathbf{D}_{\epsilon,t} \mathbf{N}_t^{-1} \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1} - \sigma^2 \mathbf{M}_t^{-1}\right) \delta_x^T} \pi_{\mathbf{Y}_t^o | S_{\epsilon,t}, S_{x,t}, \Psi}\left(\mathbf{y}_t^o\right) \left(\sigma^2\right)^{\frac{k}{2}} \\ &\times \left|\mathbf{D}_{\epsilon,t}\right|^{\frac{1}{2}} \left|\mathbf{D}_{x,t}\right|^{\frac{1}{2}} \left|\mathbf{M}_t\right|^{-\frac{1}{2}} \left|\mathbf{N}_t\right|^{-\frac{1}{2}}.\end{aligned}$$

The probability density function $\pi_{\mathbf{Y}_t^o|S_{\epsilon,t},S_{x,t},\Psi}(\mathbf{y}_t^o)$ *equals to the density function* $\pi_{\mathbf{Y}_t^o|T_\epsilon(S_{\epsilon,t}),T_x(S_{x,t}),\Psi}(\mathbf{y}_t^o)$ *specified in Lemma A10 in Appendix A.4 in Section A.4.5 and matrices* $\mathbf{D}_{\epsilon,t}$, $\mathbf{D}_{x,t}$, $\mathbf{M}_t$ *and* $\mathbf{N}_t$ *are defined as in Theorem 13.*

Proof. The proof of Theorem 25 is given in Subsection A.4.2.1 in Appendix A.4. □

Theorem 26. *The maximizers of the function* $\tilde{Q}_{GStS,ind}$ *derived in Theorem 25 with respect to the vector of parameters* Ψ^* *are the solutions to the system of equations determined by* $\nabla_{\Psi^*}\tilde{Q}_{GStS,ind}(\Psi,\Psi^*)=\mathbf{0}$ *as following*

$$\begin{cases} I(\frac{\partial}{\partial\boldsymbol{\mu}^*}\tilde{v})(\mathbf{y}_{1:N}^o)_{1\times d}=\mathbf{0} \\ I(\frac{\partial}{\partial\mathbf{W}^*}\tilde{v})(\mathbf{y}_{1:N}^o)_{d\times k}=\mathbf{0} \\ I(\frac{\partial}{\partial\sigma^{*2}}\tilde{v})(\mathbf{y}_{1:N}^o)_{1\times 1}=0 \\ I(\frac{\partial}{\partial\delta_\epsilon^*}\tilde{v})(\mathbf{y}_{1:N}^o)_{1\times d}=\mathbf{0} \\ I(\frac{\partial}{\partial\delta_x^*}\tilde{v})(\mathbf{y}_{1:N}^o)_{1\times k}=\mathbf{0}, \end{cases}$$

for the integral operator I *defined in Equation (3.39) and the gradient of the function* $\tilde{v}$ *derived as follows*

$$\begin{aligned}\frac{\partial}{\partial\boldsymbol{\mu}^*}\tilde{v}(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi,\Psi^*)&=\frac{1}{\sigma^{*2}}\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}[\mathbf{Y}_t]\mathbf{D}_{\epsilon,t}-\frac{1}{\sigma^{*2}}\boldsymbol{\mu}^*\mathbf{D}_{\epsilon,t}-\frac{1}{\sigma^{*2}}\delta_\epsilon^*-\frac{1}{\sigma^{*2}}\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}[\mathbf{Y}_t]\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\\&+\frac{1}{\sigma^{*2}}\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}+\frac{1}{\sigma^{*2}}\delta_\epsilon\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}-\frac{\sigma^2}{\sigma^{*2}}\delta_x\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t},\end{aligned}$$

$$\begin{aligned}\frac{\partial}{\partial\mathbf{W}^*}\tilde{v}(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi,\Psi^*)&=\frac{1}{\sigma^{*2}}\mathbf{D}_{\epsilon,t}\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}[\mathbf{Y}_t^T\mathbf{Y}_t]\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}-\frac{1}{\sigma^{*2}}\Big(\mathbf{D}_{\epsilon,t}\boldsymbol{\mu}^{*T}+\delta_\epsilon^{*T}\Big)\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}[\mathbf{Y}_t]\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\\&+\frac{1}{\sigma^{*2}}\Big(\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}[\mathbf{Y}_t]\mathbf{D}_{\epsilon,t}-\boldsymbol{\mu}^*\mathbf{D}_{\epsilon,t}-\delta_\epsilon^*\Big)^T\Big(-\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}-\delta_\epsilon\mathbf{W}+\sigma^2\delta_x\Big)\mathbf{M}_t^{-1}\\&-\frac{1}{\sigma^{*2}}\mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}[\mathbf{Y}_t^T\mathbf{Y}_t]\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\\&-\frac{1}{\sigma^{*2}}\mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}[\mathbf{Y}_t]^T\Big(-\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}-\delta_\epsilon\mathbf{W}+\sigma^2\delta_x\Big)\mathbf{M}_t^{-1}\\&-\frac{1}{\sigma^{*2}}\mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}_t^{-1}\Big(-\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}-\delta_\epsilon\mathbf{W}+\sigma^2\delta_x\Big)^T\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}[\mathbf{Y}_t]\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\\&-\frac{1}{\sigma^{*2}}\mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}_t^{-1}\Big(-\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}-\delta_\epsilon\mathbf{W}+\sigma^2\delta_x\Big)^T\Big(-\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}-\delta_\epsilon\mathbf{W}+\sigma^2\delta_x\Big)\mathbf{M}_t^{-1}-\frac{\sigma^2}{\sigma^{*2}}\mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}_t^{-1},\end{aligned}$$

$$\begin{aligned}\frac{\partial}{\partial\sigma^{*2}}\tilde{v}(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi,\Psi^*)&=-\frac{d}{2\sigma^{*2}}-\frac{1}{2(\sigma^{*2})^2}\Bigg[-Tr\Big\{\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}[\mathbf{Y}_t^T\mathbf{Y}_t]\mathbf{D}_{\epsilon,t}\Big\}+2\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}[\mathbf{Y}_t]\Big(\mathbf{D}_{\epsilon,t}\boldsymbol{\mu}^{*T}+\delta_\epsilon^{*T}\Big)\\&-\boldsymbol{\mu}^*\mathbf{D}_{\epsilon,t}\boldsymbol{\mu}^{*T}-2\boldsymbol{\mu}^*\delta_\epsilon^{*T}-\delta_\epsilon^*\mathbf{D}_{\epsilon,t}^{-1}\delta_\epsilon^{*T}+2\,Tr\Big\{\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}[\mathbf{Y}_t^T\mathbf{Y}_t]\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\Big\}\\&-2\Big(\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}[\mathbf{Y}_t]\mathbf{D}_{\epsilon,t}\mathbf{W}-\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}-\delta_\epsilon\mathbf{W}+\sigma^2\delta_x\Big)\mathbf{M}_t^{-1}\mathbf{W}^{*T}\Big(\boldsymbol{\mu}^*\mathbf{D}_{\epsilon,t}+\delta_\epsilon^*\Big)^T\\&+2\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}[\mathbf{Y}_t]\mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}_t^{-1}\Big(-\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}-\delta_\epsilon\mathbf{W}+\sigma^2\delta_x\Big)^T-\sigma^2\,Tr\Big\{\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\mathbf{W}^*\Big\}\\&-Tr\Big\{\mathbf{M}_t^{-1}\Big(\mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}[\mathbf{Y}_t^T\mathbf{Y}_t]\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\mathbf{W}^*\Big\}\\&-Tr\Big\{\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}[\mathbf{Y}_t]^T\Big(-\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}-\delta_\epsilon\mathbf{W}+\sigma^2\delta_x\Big)\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\mathbf{W}^*\Big\}\\&-Tr\Big\{\mathbf{M}_t^{-1}\Big(-\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}-\delta_\epsilon\mathbf{W}+\sigma^2\delta_x\Big)^T\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}[\mathbf{Y}_t]\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\mathbf{W}^*\Big\}\\&-Tr\Big\{\mathbf{M}_t^{-1}\Big(-\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}-\delta_\epsilon\mathbf{W}+\sigma^2\delta_x\Big)^T\Big(-\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}-\delta_\epsilon\mathbf{W}+\sigma^2\delta_x\Big)\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\mathbf{W}^*\Big\}\Bigg],\end{aligned}$$

$$\frac{\partial}{\partial \delta_\epsilon^*}\tilde{v}\left(\mathbf{y}_t^o, s_{\epsilon,t}, s_{x,t}; \Psi, \Psi^*\right) = \frac{1}{\sigma^{*2}}\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right] - \frac{1}{\sigma^{*2}}\boldsymbol{\mu}^* - \frac{1}{\sigma^{*2}}\delta_\epsilon^*\mathbf{D}_{\epsilon,t}^{-1} - \frac{1}{\sigma^{*2}}\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}$$
$$+ \frac{1}{\sigma^{*2}}\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T} + \frac{1}{\sigma^{*2}}\delta_\epsilon\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T} - \frac{\sigma^2}{\sigma^{*2}}\delta_x\mathbf{M}_t^{-1}\mathbf{W}^{*T},$$

$$\frac{\partial}{\partial \delta_x^*}\tilde{v}(\mathbf{y}_t^o, s_{\epsilon,t}, s_{x,t}; \Psi, \Psi^*) = -\delta_x^*\mathbf{D}_{x,t}^{-1} + \mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1} - \boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1} - \delta_\epsilon\mathbf{W}\mathbf{M}_t^{-1} + \sigma^2\delta_x\mathbf{M}_t^{-1}.$$

Proof. The proof of Theorem 26 is given in Subsection A.4.2.3 in Appendix A.4. □

Corollary 15. *The solutions to the system of equation from Theorem 26 which specify the maximizers of the function $\tilde{Q}_{GStS,ind}$ with respect to the parameters $\boldsymbol{\mu}^*, \delta_\epsilon^*, \delta_x^*, \sigma^2$ are given by explicit formulas defined by 2-dimensional integration problems on the hypercube $[0,1]^2$ as follows*

$$\begin{cases}
\boldsymbol{\mu}^* = \left[\mathbf{A}_6(\mathbf{y}_{1:N}^o;\Psi) - \mathbf{A}_{10}(\mathbf{y}_{1:N}^o;\Psi,\Psi^*) - \delta_\epsilon^* A_0(\mathbf{y}_{1:N}^o;\Psi) + \boldsymbol{\mu}\mathbf{A}_{13}(\mathbf{y}_{1:N}^o;\Psi,\Psi^*) + \left(\delta_\epsilon\mathbf{W} - \sigma^2\delta_x\right)\mathbf{A}_{12}(\mathbf{y}_{1:N}^o;\Psi,\Psi^*)^T\right]\mathbf{A}_1(\mathbf{y}_{1:N}^o;\Psi)^{-1} \\
\delta_\epsilon^* = \left[\mathbf{A}_5(\mathbf{y}_{1:N}^o;\Psi) - \boldsymbol{\mu}^* A_0(\mathbf{y}_{1:N}^o;\Psi) - \mathbf{A}_8(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T} + \boldsymbol{\mu}\mathbf{A}_{11}(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T} + \left(\delta_\epsilon\mathbf{W} - \sigma^2\delta_x\right)\mathbf{A}_4(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}\right]\mathbf{A}_2(\mathbf{y}_{1:N}^o;\Psi)^{-1} \\
\delta_x^* = \left[\mathbf{A}_8(\mathbf{y}_{1:N}^o) - \boldsymbol{\mu}\mathbf{A}_{11}(\mathbf{y}_{1:N}^o) - \left(\delta_\epsilon\mathbf{W} - \sigma^2\delta_x\right)\mathbf{A}_4(\mathbf{y}_{1:N}^o)\right]\mathbf{A}_3(\mathbf{y}_{1:N},\Psi)^{-1} \\
\sigma^{*2} = \frac{1}{dA_0(\mathbf{y}_{1:N}^o;\Psi)}\Big[A_{20}(\mathbf{y}_{1:N}^o;\Psi) + \boldsymbol{\mu}^*\mathbf{A}_1(\mathbf{y}_{1:N}^o;\Psi)\boldsymbol{\mu}^{*T} + \delta_\epsilon^*\mathbf{A}_2(\mathbf{y}_{1:N}^o;\Psi)\delta_\epsilon^{*T} + 2\mathbf{A}_9(\mathbf{y}_{1:N}^o;\Psi,\Psi^*)\boldsymbol{\mu}^T \\
\qquad +2\left(\mathbf{A}_{10}(\mathbf{y}_{1:N}^o;\Psi,\Psi^*) - \mathbf{A}_6(\mathbf{y}_{1:N}^o;\Psi) - \boldsymbol{\mu}\mathbf{A}_{13}(\mathbf{y}_{1:N}^o;\Psi,\Psi^*)\right)\boldsymbol{\mu}^{*T} - 2A_{21}(\mathbf{y}_{1:N}^o;\Psi,\Psi^*) + \sigma^2\, Tr\left\{\mathbf{A}_{12}(\mathbf{y}_{1:N}^o;\Psi)^T\mathbf{W}^*\right\} \\
\qquad + Tr\left\{\mathbf{A}_{17}(\mathbf{y}_{1:N}^o;\Psi)^T\mathbf{W}^*\right\} - Tr\left\{\mathbf{A}_{18.1}(\mathbf{y}_{1:N}^o;\Psi)^T\mathbf{W}^*\right\} - Tr\left\{\mathbf{A}_{18.2}(\mathbf{y}_{1:N}^o;\Psi)^T\mathbf{W}^*\right\} + Tr\left\{\mathbf{A}_{19}(\mathbf{y}_{1:N}^o;\Psi)^T\mathbf{W}^*\right\} \\
\qquad +2\left(\boldsymbol{\mu}^* A_0(\mathbf{y}_{1:N}^o;\Psi) - \mathbf{A}_5(\mathbf{y}_{1:N}^o;\Psi) + \mathbf{A}_8(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T} - \boldsymbol{\mu}\mathbf{A}_{11}(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}\right)\delta_\epsilon^{*T} \\
\qquad +2\left(\mathbf{A}_7(\mathbf{y}_{1:N}^o;\Psi,\Psi^*) - \boldsymbol{\mu}^*\mathbf{A}_{12}(\mathbf{y}_{1:N}^o;\Psi,\Psi^*) - \delta_\epsilon^*\mathbf{W}^*\mathbf{A}_4(\mathbf{y}_{1:N}^o;\Psi)\right)\left(\delta_\epsilon\mathbf{W} - \sigma^2\delta_x\right)^T\Big] \\
f(\mathbf{W}^*;\Psi,\Psi^*) = \mathbf{0}.
\end{cases}$$

The solutions with respect to $\boldsymbol{\mu}^*$, δ_ϵ^* and σ^{*2} are linear function of the parameter $\mathbf{W}^*$. The maximizer of the function $\tilde{Q}_{GStS,ind}$ with respect to $\mathbf{W}^*$ is determined by the equation

$$f(\mathbf{W}^*;\Psi,\Psi^*) = \mathbf{0},$$

for the function $f : \mathbb{R}^{d\times k} \to \mathbb{R}^{d\times k}$ such that $f(\mathbf{W}^*;\Psi,\Psi^*) := \frac{\partial \tilde{Q}_{GStS,ind}(\Psi,\Psi^*)}{\partial \mathbf{W}^*}$ being equal to

$$f(\mathbf{W}^*;\Psi,\Psi^*) = \mathbf{A}_{14}(\mathbf{y}_{1:N}^o;\Psi) - \mathbf{A}_{15}(\mathbf{y}_{1:N}^o;\Psi,\Psi^*) - \delta_\epsilon^{*T}\mathbf{A}_8(\mathbf{y}_{1:N}^o;\Psi) - \mathbf{A}_{16}(\mathbf{y}_{1:N}^o;\Psi,\Psi^*) + \delta_\epsilon^{*T}\boldsymbol{\mu}\mathbf{A}_{11}(\mathbf{y}_{1:N}^o;\Psi)$$
$$+ \delta_\epsilon^{*T}\left(\delta\mathbf{W} - \sigma^2\mathbf{W}\right)\mathbf{A}_4(\mathbf{y}_{1:N}^o;\Psi) - \mathbf{A}_{17}(\mathbf{y}_{1:N}^o;\Psi,\Psi^*) + \mathbf{A}_{18.1}(\mathbf{y}_{1:N}^o;\Psi,\Psi^*) + \mathbf{A}_{18.2}(\mathbf{y}_{1:N}^o;\Psi,\Psi^*)$$
$$- \mathbf{A}_{19}(\mathbf{y}_{1:N}^o;\Psi,\Psi^*) - \sigma^2\mathbf{A}_{12}(\mathbf{y}_{1:N}^o;\Psi,\Psi^*)^T.$$

The function f is linear with respect to the parameters $\boldsymbol{\mu}^*$ and δ_ϵ^*. The 2-dimensional integrals $\mathbf{A}_i$ on the hypercube $[0,1]^2$, for $i \in \left\{0,\ldots,21\right\}$, are defined in Subsubsection A.4.2.4 in Appendix A.4.

Proof. The proof of Corollary 15 is located in Subsubsection A.4.2.4 in Appendix A.4. □

4.4.1.1. EM Algorithm for Special Cases of Probabilistic Principal Component with Independent Generalized Skew t-Student Distribution in the Presence of Missing Data

In the following part we derive the formulations of the E-step and the M-step of EM algorithms for special cases of the Generalized Skew t-Student IND PPCA specified in Subsubsection 3.4.1.3. The reader may find the derivations under following notation

Special Case 1: Grouped t-Student IND PPCA in Corollary 16 and Corollary 17,
Special Case 2: Skew t-Student IND PPCA in Corollary 18 and Corollary 19,
Special Case 3: t-Student IND PPCA in Corollary 20 and Corollary 21.

Corollary 16. *Under the assumptions of Special Case 1: Grouped t-Student IND PPCA specified in Subsubsection 3.4.1.3, the E-step of the corresponding EM algorithm in incomplete data setting as a case-specific variant of Theorem 25, is given by*

$$\tilde{Q}_{GStS,ind,1}(\Psi,\Psi^*) = \mathbb{E}_{\mathbf{Y}^m_{1:N},\mathbf{X}_{1:N},\mathbf{U}_{1:N},\mathbf{V}_{1:N}|\mathbf{Y}^o_{1:N},\Psi}\left[\log\left(\pi_{\mathbf{Y}_{1:N},\mathbf{X}_{1:N},\mathbf{U}_{1:N}\mathbf{Y}_{1:N}|\Psi^*}\left(\mathbf{Y}_{1:N},\mathbf{X}_{1:N},\mathbf{U}_{1:N},\mathbf{V}_{1:N}\right)\right)\right].$$

The function $\tilde{Q}_{GStS,ind,1}(\Psi,\Psi^*)$ is represented by the functions $I_1,\ I_2:\mathbb{R}^{d_o}\to\mathbb{R}$ as in Theorem 25 given the case-specific formulation of the function $m:\mathbb{R}^{d_o}\times[0,1]^2\longrightarrow\mathbb{R}$ defined as

$$m(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi) = \pi_{\mathbf{Y}^o_t|S_{\epsilon,t},S_{x,t},\Psi}(\mathbf{y}^o_t)\left(\sigma^2\right)^{\frac{k}{2}}\left|\mathbf{D}_{\epsilon,t}\right|^{\frac{1}{2}}\left|\mathbf{D}_{x,t}\right|^{\frac{1}{2}}\left|\mathbf{M}_t\right|^{-\frac{1}{2}}\left|\mathbf{N}_t\right|^{-\frac{1}{2}},$$

the probability function $\pi_{\mathbf{Y}^o_t|S_{\epsilon,t},S_{x,t},\Psi}(\mathbf{y}^o_t)$ of the d_o-dimensional random vector $\mathbf{Y}^o_t|S_{\epsilon,t},Sx,t,\Psi\ \sim\ \mathcal{N}(\boldsymbol{\mu}^o,\boldsymbol{\Sigma}^{oo}_{y,t})$, where $\boldsymbol{\mu}^o$ contains the elements of the vector $\boldsymbol{\mu}$ which correspond to the observed entries of the vector $\mathbf{Y}_t$ and $\boldsymbol{\Sigma}^{oo}_{y,t}$ is a submatrix of $\boldsymbol{\Sigma}_{y,t}=\sigma^2\mathbf{N}_t^{-1}$ which correspond by rows and columns to the observed elements of $\mathbf{Y}_t$. The matrices $\mathbf{M}_t,\ \mathbf{N}_t,\mathbf{D}_{\epsilon,t}$ and $\mathbf{D}_{x,t}$ are defined as in Theorem 25 and the function $\tilde{v}:\mathbb{R}^{d_o}\times[0,1]^2\to\mathbb{R}$ is defined as

$$\begin{aligned}
\tilde{v}(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi,\Psi^*) &= \log\pi_{T_\epsilon(S_{\epsilon,t})|\Psi^*}(T_\epsilon(s_{\epsilon,t}) + \log\pi_{T_x(S_{x,t}|\Psi^*}(T_x(s_{x,t}) - \frac{k+d}{2}\log 2\pi - \frac{d}{2}\log\sigma^{*2} + \frac{1}{2}\log\left|\mathbf{D}_{\epsilon,t}\right| \\
&+ \frac{1}{2}\log\left|\mathbf{D}_{x,t}\right| - \frac{1}{2\sigma^{*2}}\,Tr\left\{\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}^T_t\mathbf{Y}_t\right]\mathbf{D}_{\epsilon,t}\right\} + \frac{1}{\sigma^{*2}}\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]\mathbf{D}_{\epsilon,t}\boldsymbol{\mu}^{*T} - \frac{1}{2\sigma^{*2}}\boldsymbol{\mu}^*\mathbf{D}_{\epsilon,t}\boldsymbol{\mu}^{*T} \\
&+ \frac{1}{\sigma^{*2}}\,Tr\left\{\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}^T_t\mathbf{Y}_t\right]\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\right\} - \frac{1}{\sigma^{*2}}\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\boldsymbol{\mu}^{*T} \\
&- \frac{1}{\sigma^{*2}}\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]\mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\boldsymbol{\mu}^T + \frac{1}{\sigma^{*2}}\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\boldsymbol{\mu}^{*T} \\
&- \frac{1}{2}\,Tr\left\{\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}^T_t\mathbf{Y}_t\right]\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\left(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\mathbf{W}^* + \mathbf{D}_{x,t}\right)\right\} \\
&+ \frac{1}{2}\,Tr\left\{\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]^T\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\left(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\mathbf{W}^* + \mathbf{D}_{x,t}\right)\right\} \\
&+ \frac{1}{2}\,Tr\left\{\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\boldsymbol{\mu}^T\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]\mathbf{M}_t^{-1}\left(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\mathbf{W}^* + \mathbf{D}_{x,t}\right)\right\} \\
&- \frac{1}{2}\,Tr\left\{\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\boldsymbol{\mu}^T\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\left(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\mathbf{W}^* + \mathbf{D}_{x,t}\right)\right\} \\
&- \frac{\sigma^2}{2}\,Tr\left\{\mathbf{M}_t^{-1}\left(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\mathbf{W}^* + \mathbf{D}_{x,t}\right)\right\},
\end{aligned}$$

for the corresponding non-central moments of $\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi$ specified as

$$\begin{aligned}
\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right] &= \left[\mathbf{Y}^o_t,\ \boldsymbol{\mu}^m + \left(\mathbf{Y}^o_t-\boldsymbol{\mu}^o\right)\boldsymbol{\Sigma}^{oo\ -1}_{y,t}\boldsymbol{\Sigma}^{om}_{y,t}\right], \\
\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}^T_t\mathbf{Y}_t\right] &= \begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Sigma}^{mm}_{y,t} - \boldsymbol{\Sigma}^{mo}_{y,t}\boldsymbol{\Sigma}^{oo\ -1}_{y,t}\boldsymbol{\Sigma}^{om}_{y,t}\end{bmatrix} + \mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t}\Psi}\left[\mathbf{Y}_t\right]^T\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right],
\end{aligned}$$

where $\boldsymbol{\mu}^m$ contains elements of the vector $\boldsymbol{\mu}$ which correspond to the non-observed entries of the vector $\mathbf{Y}_t$ and $\boldsymbol{\Sigma}^{mm}_{y,t}$ is a square submatrix of $\boldsymbol{\Sigma}_{y,t}$ which correspond by rows and columns to missing elements of $\mathbf{Y}_t$. The elements of the $d_m\times d_o$ submatrix $\boldsymbol{\Sigma}^{mo}_{y,t}=\boldsymbol{\Sigma}^{om\ T}_{y,t}$ correspond by rows to missing entries of the vector $\mathbf{Y}_t$ and by columns to its observed elements.

Proof. The proof of Corollary 16 is analogous to the proof of Theorem 25 when one sets $\delta_\epsilon = \delta_x = \delta_\epsilon^* = \delta_x^* = \mathbf{0}$. □

Corollary 17. *Under the assumptions of Special Case 1: Grouped t-Student IND PPCA specified in Subsubsection 3.4.1.3, the M-step of the corresponding EM algorithm in incomplete data setting as a case-specific variant of Theorem 26 given the problem formulation* $\nabla_{\Psi^*}\tilde{Q}_{GStS,ind,1} = \mathbf{0}$ *for* $\Psi^* = \left[\mathbf{W}^*, \boldsymbol{\mu}^*, \sigma^{*2}\right]$ *is defined by 2-dimensional integration problems on the hypercube* $[0,1]^2$ *as following*

$$\begin{cases} \boldsymbol{\mu}^* = \left[\mathbf{A}_6(\mathbf{y}^o_{1:N};\Psi) - \mathbf{A}_{10}(\mathbf{y}^o_{1:N};\Psi) + \boldsymbol{\mu}\mathbf{A}_{13}(\mathbf{y}^o_{1:N};\Psi)\right]\mathbf{A}_1(\mathbf{y}^o_{1:N};\Psi)^{-1} \\ \sigma^{*2} = \frac{1}{dA_0(\mathbf{y}^o_{1:N};\Psi)}\Big[A_{20}(\mathbf{y}^o_{1:N};\Psi) - 2\mathbf{A}_6(\mathbf{y}^o_{1:N};\Psi)\boldsymbol{\mu}^{*T} + \boldsymbol{\mu}^*\mathbf{A}_1(\mathbf{y}^o_{1:N};\Psi)\boldsymbol{\mu}^{*T} - 2A_{21}(\mathbf{y}^o_{1:N};\Psi,\Psi^*) + 2\mathbf{A}_{10}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\boldsymbol{\mu}^{*T} \\ \quad +2\mathbf{A}_9(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\boldsymbol{\mu}^T - 2\boldsymbol{\mu}\mathbf{A}_{13}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\boldsymbol{\mu}^{*T} + Tr\left\{\mathbf{A}_{17}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)^T\mathbf{W}^*\right\} - Tr\left\{\mathbf{A}_{18.1}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)^T\mathbf{W}^*\right\} \\ \quad - Tr\left\{\mathbf{A}_{18.2}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)^T\mathbf{W}^*\right\} + Tr\left\{\mathbf{A}_{19}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)^T\mathbf{W}^*\right\} + \sigma^2\, Tr\left\{\mathbf{A}_{12}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)^T\mathbf{W}^*\right\}\Big]. \end{cases}$$

The solutions with respect to $\boldsymbol{\mu}^*$ *and* σ^{*2} *are linear functions of the parameter* $\mathbf{W}^*$*. The maximizer of the function* $\tilde{Q}_{GStS,ind,1}$ *with respect to* $\mathbf{W}^*$ *is determined as a zero root of a case specific function* $f : \mathbb{R}^{d\times k} \to \mathbb{R}^{d\times k}$ *such that* $f(\mathbf{W}^*;\Psi,\Psi^*) := \frac{\partial\tilde{Q}_{GStS,ind,1}(\Psi,\Psi^*)}{\partial\mathbf{W}^*}$ *is specified as*

$$\begin{aligned} f(\mathbf{W}^*;\Psi,\Psi^*) = \mathbf{A}_{14}(\mathbf{y}^o_{1:N};\Psi) - \mathbf{A}_{15}(\mathbf{y}^o_{1:N};\Psi,\Psi^*) - \mathbf{A}_{16}(\mathbf{y}^o_{1:N};\Psi,\Psi^*) - \mathbf{A}_{17}(\mathbf{y}^o_{1:N};\Psi,\Psi^*) + \mathbf{A}_{18.1}(\mathbf{y}^o_{1:N};\Psi,\Psi^*) \\ + \mathbf{A}_{18.2}(\mathbf{y}^o_{1:N};\Psi,\Psi^*) - \mathbf{A}_{19}(\mathbf{y}^o_{1:N};\Psi,\Psi^*) - \sigma^2\mathbf{A}_{12}(\mathbf{y}^o_{1:N};\Psi,\Psi^*), \end{aligned}$$

and is linear with respect to the parameters $\boldsymbol{\mu}^*$*. The 2-dimensional integrals* $\mathbf{A}_i$ *on the hypercube* $[0,1]^2$*, for* $i \in \left\{0,\ldots,21\right\}$*, are defined in Subsubsection A.4.2.4 in Appendix A.4 given the following simplifications*

$$\mathbf{A}_{16}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)_{d\times k} := \sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}\left(\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right] - \boldsymbol{\mu}^*\right)^T\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\, m(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi)\, ds_{\epsilon,t}\, ds_{x,t}\right\},$$

$$\mathbf{A}_{18.1}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)_{k\times d} := \sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]^T\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\, m(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi)\, ds_{\epsilon,t}\, ds_{x,t}\right\},$$

$$\mathbf{A}_{18.2}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)_{k\times d} := \sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\boldsymbol{\mu}^T\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\, m(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi)\, ds_{\epsilon,t}\, ds_{x,t}\right\},$$

$$\mathbf{A}_{19}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)_{k\times d} := \sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\boldsymbol{\mu}^T\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\, m(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi)\, ds_{\epsilon,t}\, ds_{x,t}\right\}$$

and the definition of the following moments of $\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi$ *are given in Corollary 16.*

Proof. The proof of Corollary 17 is analogous to the proof of Theorem 25 and Corollary 15 when one sets $\delta_\epsilon = \delta_x = \delta_\epsilon^* = \delta_x^* = \mathbf{0}$. □

Corollary 18. *Under the assumptions of Special Case 2: Skew t-Student IND PPCA specified in Subsubsection 3.4.1.3, the E-step of the corresponding EM algorithm in incomplete data setting as a case-specific variant of Theorem 25, is given by*

$$\tilde{Q}_{GStS,ind,2}(\Psi,\Psi^*) = \mathbb{E}_{\mathbf{Y}^m_{1:N},\mathbf{X}_{1:N},U_{1:N},V_{1:N}|\mathbf{Y}^o_{1:N},\Psi}\left[\log\left(\pi_{\mathbf{Y}_{1:N},\mathbf{X}_{1:N},U_{1:N}V_{1:N}|\Psi^*}\left(\mathbf{Y}_{1:N},\mathbf{X}_{1:N},U_{1:N},V_{1:N}\right)\right)\right],$$

defined by I_1*,* $I_2 : \mathbb{R}^{d_o} \to \mathbb{R}$ *from Theorem 25 given the case-specific formulation of the function* $m : \mathbb{R}^{d_o}\times[0,1]^2 \longrightarrow \mathbb{R}$ *defined as*

$$m(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi) = \pi_{\mathbf{Y}_t^o|S_{\epsilon,t},S_{x,t},\Psi}(\mathbf{y}^o_t)\left(\sigma^2\right)^{\frac{k}{2}}T_\epsilon(s_{\epsilon,t})^{\frac{d}{2}}T_x(s_{x,t})^{\frac{k}{2}}\left|\mathbf{M}_t\right|^{-\frac{1}{2}}\left|\mathbf{N}_t\right|^{-\frac{1}{2}}$$

$$\times\, e^{-\frac{1}{2}\boldsymbol{\delta}_x\left(T_x(s_{x,t})^{-1}\mathbb{I}_k-\sigma^2T_\epsilon(s_{\epsilon,t})^2\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{N}_t^{-1}\mathbf{W}\mathbf{M}_t^{-1}-\sigma^2\mathbf{M}_t^{-1}\right)\boldsymbol{\delta}_x^T},$$

for the probability function $\pi_{\mathbf{Y}_t^o|S_{\epsilon,t},S_{x,t},\Psi}(\mathbf{y}_t^o)$ *of the* d_o*-dimensional random vector* $\mathbf{Y}_t^o|S_{\epsilon,t},S_{x,t},\Psi \sim \mathcal{N}\left(\boldsymbol{\mu}_{y,t}^o,\boldsymbol{\Sigma}_{y,t}^{oo}\right)$*, where* $\boldsymbol{\mu}_{y,t}^o$ *contains elements of the vector* $\boldsymbol{\mu}_{y,t}$ *which correspond to the observed entries of the vector* $\mathbf{Y}_t$ *and* $\boldsymbol{\Sigma}_{y,t}^{oo}$ *is a submatrix of* $\boldsymbol{\Sigma}_{y,t}$ *which correspond by rows and columns to observed elements of* $\mathbf{Y}_t$*, for* $\boldsymbol{\mu}_{y,t}$ *and* $\boldsymbol{\Sigma}_{y,t}$ *specified as*

$$\boldsymbol{\mu}_{y,t}=\boldsymbol{\mu}+T_\epsilon(s_{\epsilon,t})^{-1}\boldsymbol{\delta}_\epsilon+\sigma^2T_\epsilon(s_{\epsilon,t})\boldsymbol{\delta}_x\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{N}_t^{-1} \text{ and } \boldsymbol{\Sigma}_{y,t}=\sigma^2\mathbf{N}_t^{-1}$$

and matrices $\mathbf{M}_t$ *and* $\mathbf{N}_t$ *are defined as in Corollary 2. The function* $\tilde{v}:\mathbb{R}^{d_o}\times[0,1]^2\longrightarrow\mathbb{R}$ *is defined as*

$$\begin{aligned}
\tilde{v}(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi,\Psi^*) &= -\frac{k+d}{2}\log 2\pi-\frac{d}{2}\log\sigma^{*2}+\frac{d}{2}T_\epsilon(s_{\epsilon,t})+\frac{k}{2}T_x(s_{x,t})-\frac{1}{2}T_x(s_{x,t})^{-1}\boldsymbol{\delta}_x^*\boldsymbol{\delta}_x^{*T}\\
&-\frac{1}{2\sigma^{*2}}T_\epsilon(s_{\epsilon,t})\,Tr\left\{\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right]\right\}+\frac{1}{\sigma^{*2}}\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]\left(T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}^{*T}+\boldsymbol{\delta}_\epsilon^{*T}\right)\\
&-\frac{1}{2\sigma^{*2}}T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}^*\boldsymbol{\mu}^{*T}-\frac{1}{\sigma^{*2}}\boldsymbol{\mu}^*\boldsymbol{\delta}_\epsilon^{*T}-\frac{1}{2\sigma^{*2}}T_\epsilon(s_{\epsilon,t})^{-1}\boldsymbol{\delta}_\epsilon^*\boldsymbol{\delta}_\epsilon^{*T}+\frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})^2\,Tr\left\{\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U}_t,\mathbf{V}_t,\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right]\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}\right\}\\
&+T_\epsilon(s_{\epsilon,t})\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]\mathbf{W}\mathbf{M}_t^{-1}\left(-\frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})\mathbf{W}^{*T}\boldsymbol{\mu}^{*T}-\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\boldsymbol{\delta}_\epsilon^{*T}+\boldsymbol{\delta}_x^{*T}\right)\\
&+\frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]\mathbf{W}^*\mathbf{M}_t^{-1}\left(-\boldsymbol{\mu}T_\epsilon(s_{\epsilon,t})\mathbf{W}-\boldsymbol{\delta}_\epsilon\mathbf{W}+\sigma^2\boldsymbol{\delta}_x\right)^T\\
&+\left(-T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W}-\boldsymbol{\delta}_\epsilon\mathbf{W}+\sigma^2\boldsymbol{\delta}_x\right)\mathbf{M}_t^{-1}\left(-\frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})\mathbf{W}^{*T}\boldsymbol{\mu}^{*T}-\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\boldsymbol{\delta}_\epsilon^{*T}+\boldsymbol{\delta}_x^{*T}\right)\\
&-\frac{1}{2}T_\epsilon(s_{\epsilon,t})^2\,Tr\left\{\mathbf{M}_t^{-1}\mathbf{W}^T\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right]\mathbf{W}\mathbf{M}_t^{-1}\left(\frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})\mathbf{W}^{*T}\mathbf{W}^*+T_x(s_{x,t})\mathbb{I}_k\right)\right\}\\
&-\frac{1}{2}T_\epsilon(s_{\epsilon,t})\,Tr\left\{\mathbf{M}_t^{-1}\mathbf{W}^T\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]^T\left(-T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W}-\boldsymbol{\delta}_\epsilon\mathbf{W}+\sigma^2\boldsymbol{\delta}_x\right)\mathbf{M}_t^{-1}\left(\frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})\mathbf{W}^{*T}\mathbf{W}^*+T_x(s_{x,t})\mathbb{I}_k\right)\right\}\\
&-\frac{1}{2}T_\epsilon(s_{\epsilon,t})\,Tr\left\{\mathbf{M}_t^{-1}\left(-T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W}-\boldsymbol{\delta}_\epsilon\mathbf{W}+\sigma^2\boldsymbol{\delta}_x\right)^T\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]\mathbf{W}^T\mathbf{M}_t^{-1}\left(\frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})\mathbf{W}^{*T}\mathbf{W}^*+T_x(s_{x,t})\mathbb{I}_k\right)\right\}\\
&-\frac{1}{2}\,Tr\left\{\mathbf{M}_t^{-1}\left(-T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W}-\boldsymbol{\delta}_\epsilon\mathbf{W}+\sigma^2\boldsymbol{\delta}_x\right)^T\left(-T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W}-\boldsymbol{\delta}_\epsilon\mathbf{W}+\sigma^2\boldsymbol{\delta}_x\right)\mathbf{M}_t^{-1}\left(\frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})\mathbf{W}^{*T}\mathbf{W}^*+T_x(s_{x,t})\mathbb{I}_k\right)\right\}\\
&-\frac{\sigma^2}{2}\,Tr\left\{\mathbf{M}_t^{-1}\left(\frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})\mathbf{W}^{*T}\mathbf{W}^*+T_x(s_{x,t})\mathbb{I}_k\right)\right\},
\end{aligned}$$

for the corresponding non-central moments of $\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi$ *specified as*

$$\begin{aligned}
\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right] &= \left[\mathbf{Y}_t^o,\ \boldsymbol{\mu}_{y,t}^m+\left(\mathbf{Y}_t^o-\boldsymbol{\mu}_{y,t}^o\right)\boldsymbol{\Sigma}_{y,t}^{oo\ -1}\boldsymbol{\Sigma}_{y,t}^{om}\right],\\
\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right] &= \begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Sigma}_{y,t}^{mm}-\boldsymbol{\Sigma}_{y,t}^{mo}\boldsymbol{\Sigma}_{y,t}^{oo\ -1}\boldsymbol{\Sigma}_{y,t}^{om}\end{bmatrix}+\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]^T\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right],
\end{aligned}$$

where $\boldsymbol{\mu}_{y,t}^m$ *contains elements of the vector* $\boldsymbol{\mu}_{y,t}$ *which correspond to the non-observed entries of the vector* $\mathbf{Y}_t$ *and* $\boldsymbol{\Sigma}_{y,t}^{mm}$ *is a square submatrix of* $\boldsymbol{\Sigma}_{y,t}$ *which correspond by rows and columns to missing elements of* $\mathbf{Y}_t$*. The elements of the* $d_m\times d_o$ *submatrix* $\boldsymbol{\Sigma}_{y,t}^{mo}=\boldsymbol{\Sigma}_{y,t}^{om\ T}$ *correspond by rows to missing entries of the vector* $\mathbf{Y}_t$ *and by columns to its observed elements.*

Proof. The proof of Corollary 18 is analogous to the proof of Theorem 25 and Corollary 15 when one sets $\mathbf{D}_{\epsilon,t}=T_\epsilon(s_{\epsilon,t})\mathbb{I}_d$ and $\mathbf{D}_{x,t}=T_x(s_{x,t})\mathbb{I}_d$. □

Corollary 19. *Under the assumptions of Special Case 2: Skew t-Student IND PPCA specified in Subsubsection 3.4.1.3, the M-step of the corresponding EM algorithm in incomplete data setting as*

a case-specific variant of Theorem 26 given the problem formulation $\nabla_{\Psi^*}\tilde{Q}_{GStS,ind,2} = \mathbf{0}$ *for* $\Psi^* = \left[\mathbf{W}^*, \boldsymbol{\mu}^*, \delta_\epsilon^*, \delta_x^*, \sigma^{*2}\right]$ *is defined by 2-dimensional integration problems on the hypercube* $[0,1]^2$ *as following*

$$
\begin{cases}
\boldsymbol{\mu}^* = A_1(\mathbf{y}^o_{1:N};\Psi)^{-1}\Big[\mathbf{A}_6(\mathbf{y}^o_{1:N};\Psi) - A_0(\mathbf{y}^o_{1:N};\Psi)\delta_\epsilon^* - \mathbf{A}_{10}(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}^{*T} + \boldsymbol{\mu}\mathbf{W}\mathbf{A}_{13}(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}^{*T} + (\delta_\epsilon\mathbf{W} + \sigma^2\delta_x)\mathbf{A}_{11}(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}^{*T}\Big] \\
\delta_\epsilon^* = A_2(\mathbf{y}^o_{1:N};\Psi)^{-1}\Big[\mathbf{A}_5(\mathbf{y}_{1:N};\Psi) - A_0(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^* - \mathbf{A}_8(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}^{*T} + \boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}^{*T} + \Big(\delta_\epsilon\mathbf{W} - \sigma^2\delta_x\Big)\mathbf{A}_4(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}^{*T}\Big] \\
\delta_x^* = A_3(\mathbf{y}^o_{1:N};\Psi)^{-1}\Big(\mathbf{A}_8(\mathbf{y}_{1:N};\Psi) - \boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}^o_{1:N};\Psi) - (\delta_\epsilon\mathbf{W} - \sigma^2\delta_x)\mathbf{A}_4(\mathbf{y}^o_{1:N};\Psi)\Big) \\
\mathbf{W}^* = \Big[\mathbf{A}_{14}(\mathbf{y}^o_{1:N};\Psi) - \boldsymbol{\mu}^{*T}\mathbf{A}_{10}(\mathbf{y}^o_{1:N};\Psi) - \delta_\epsilon^*\mathbf{A}_8(\mathbf{y}^o_{1:N};\Psi) - \mathbf{A}_{16}(\mathbf{y}^o_{1:N};\Psi) + \boldsymbol{\mu}^{*T}\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{13}(\mathbf{y}^o_{1:N};\Psi) + \boldsymbol{\mu}^{*T}(\delta_\epsilon - \sigma^2\delta_x)\mathbf{A}_{11}(\mathbf{y}^o_{1:N};\Psi) \\
\quad +\delta_\epsilon^{*T}\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}^o_{1:N};\Psi) + \delta_\epsilon^{*T}(\delta_\epsilon\mathbf{W} - \sigma^2\delta_x)\mathbf{A}_4(\mathbf{y}^o_{1:N};\Psi)\Big]\Big[\mathbf{A}_{17}(\mathbf{y}^o_{1:N};\Psi) - \mathbf{A}_{18}(\mathbf{y}^o_{1:N};\Psi) - \mathbf{A}_{18}(\mathbf{y}^o_{1:N};\Psi)^T + \mathbf{A}_{19}(\mathbf{y}^o_{1:N};\Psi) \\
\quad +\sigma^2\mathbf{A}_{11}(\mathbf{y}^o_{1:N};\Psi)\Big]^{-1} \\
\sigma^{*2} = \frac{1}{dA_0(\mathbf{y}^o_{1:N};\Psi)}\Big[A_{20}(\mathbf{y}^o_{1:N};\Psi) - 2\mathbf{A}_6(\mathbf{y}^o_{1:N};\Psi)\boldsymbol{\mu}^{*T} - 2\mathbf{A}_5(\mathbf{y}^o_{1:N};\Psi)\delta_\epsilon^{*T} + A_1(\mathbf{y}^o_{1:N};\Psi)\boldsymbol{\mu}^*\boldsymbol{\mu}^{*T} + 2A_0(\mathbf{y}^o_{1:N};\Psi)\boldsymbol{\mu}^*\delta_\epsilon^{*T} \\
\quad +A_2(\mathbf{y}^o_{1:N};\Psi)\delta_\epsilon^*\delta_\epsilon^{*T} - 2\,Tr\left\{\mathbf{A}_{14}(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}^{*T}\right\} + 2\mathbf{A}_9(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\mathbf{W}^T\boldsymbol{\mu}^T + 2\mathbf{A}_7(\mathbf{y}^o_{1:N};\Psi,\Psi)\mathbf{W}^T\delta_\epsilon^T \\
\quad -2\sigma^2\mathbf{A}_7(\mathbf{y}^o_{1:N};\Psi,\Psi)\delta_x^T + 2\Big(\mathbf{A}_{10}(\mathbf{y}^o_{1:N};\Psi) - \boldsymbol{\mu}\mathbf{W}\mathbf{A}_{13}(\mathbf{y}^o_{1:N};\Psi) - (\delta_\epsilon\mathbf{W} - \sigma^2\delta_x)\mathbf{A}_{11}(\mathbf{y}^o_{1:N};\Psi)\Big)\mathbf{W}^{*T}\boldsymbol{\mu}^{*T} \\
\quad +2\Big(\mathbf{A}_8(\mathbf{y}^o_{1:N};\Psi) - \boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}^o_{1:N};\Psi) - (\delta_\epsilon\mathbf{W} - \sigma^2\delta_x)\mathbf{A}_4(\mathbf{y}^o_{1:N};\Psi)\Big)\mathbf{W}^{*T}\delta_\epsilon^{*T} + Tr\left\{\mathbf{A}_{17}(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}^{*T}\mathbf{W}^*\right\} \\
\quad - Tr\left\{\Big(\mathbf{A}_{18}(\mathbf{y}^o_{1:N};\Psi) + \mathbf{A}_{18}(\mathbf{y}^o_{1:N};\Psi)^T\Big)\mathbf{W}^{*T}\mathbf{W}^*\right\} + Tr\left\{\mathbf{A}_{19}(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}^{*T}\mathbf{W}^*\right\} + \sigma^2\,Tr\left\{\mathbf{A}_{11}(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}^{*T}\mathbf{W}^*\right\}\Big].
\end{cases}
$$

The 2-dimensional integrals $\mathbf{A}_i$ *on the hypercube* $[0,1]^2$*, for* $i \in \Big\{0,\ldots,2\Big\}$*, are defined in Subsubsection A.4.2.5 in Appendix A.4 and the moments of* $\mathbf{Y}_t^m|\mathbf{Y}_t^o, S_{\epsilon,t}, S_{x,t}, \Psi$ *are given in Corollary 18.*

Proof. The proof of Corollary 19 is given in in Subsubsection A.4.2.5 in Appendix A.4. □

Corollary 20. *Under the assumptions of Special Case 3: t-Student IND PPCA specified in Subsubsection 3.4.1.3, the E-step of the corresponding EM algorithm in the incomplete data setting as a case-specific variant of Theorem 25, is given by*

$$
\begin{aligned}
\tilde{Q}_{GStS,ind,3}(\Psi,\Psi^*) &= \mathbb{E}_{\mathbf{Y}^m_{1:N},\mathbf{X}_{1:N},U_{1:N},V_{1:N}|\mathbf{Y}^o_{1:N},\Psi}\left[\log\left(\pi_{\mathbf{Y}_{1:N},\mathbf{X}_{1:N},U_{1:N}V_{1:N}|\Psi^*}\left(\mathbf{Y}_{1:N},\mathbf{X}_{1:N},u_{1:N},v_{1:N}\right)\right)\right] \\
&= \frac{1}{\pi_{\mathbf{Y}^o_{1:N}|\Psi}(\mathbf{y}^o_{1:N})}\sum_{t=1}^{N}\Big\{I_1(\mathbf{y}^o_t;\Psi,\Psi^*)\prod_{s=1,s\neq t}^{N} I_2(\mathbf{y}^o_s;\Psi)\Big\},
\end{aligned}
$$

for the functions $I_1,\ I_2 : \mathbb{R}^{d_o} \to \mathbb{R}$ *are defined as is Theorem 25 given the case-specific formulation of the function* $m : \mathbb{R}^{d_o} \times [0,1]^2 \longrightarrow \mathbb{R}$ *defined as*

$$
m(\mathbf{y}^o_t, s_{\epsilon,t}, s_{x,t};\Psi) = \pi_{\mathbf{Y}^o_t|S_{\epsilon,t},S_{x,t},\Psi}(\mathbf{y}^o_t)\left(\sigma^2\right)^{\frac{k}{2}}T_\epsilon(s_{\epsilon,t})^{\frac{d}{2}}T_x(s_{x,t})^{\frac{k}{2}}\left|\mathbf{M}_t\right|^{-\frac{1}{2}}\left|\mathbf{N}_t\right|^{-\frac{1}{2}},
$$

for the probability function $\pi_{\mathbf{Y}^o_t|S_{\epsilon,t},S_{x,t},\Psi}(\mathbf{y}^o_t)$ *of the* d_o*-dimensional random vector* $\mathbf{Y}^o_t|S_{\epsilon,t},S_{x,t},\Psi \sim \mathcal{N}\left(\boldsymbol{\mu}^o, \boldsymbol{\Sigma}^{oo}_{y,t}\right)$*, where* $\boldsymbol{\mu}^o_{y,t}$ *contains elements of the vector* $\boldsymbol{\mu}$ *which correspond to the observed entries of the vector* $\mathbf{Y}_t$ *and* $\boldsymbol{\Sigma}^{oo}_{y,t}$ *is a submatrix of* $\boldsymbol{\Sigma}_{y,t} = \sigma^2\mathbf{N}_t^{-1}$ *which correspond by rows and columns to observed elements of* $\mathbf{Y}_t$*. The and matrices* $\mathbf{M}_t$ *and* $\mathbf{N}_t$ *are defined as in Corollary 3. The function* $\tilde{v} : \mathbb{R}^{d_o} \times [0,1]^2 \longrightarrow \mathbb{R}$ *is defined as*

$$
\begin{aligned}
\tilde{v}(\mathbf{y}^o_t, s_{\epsilon,t}, s_{x,t};\Psi,\Psi^*) &= -\frac{k+d}{2}\log 2\pi - \frac{d}{2}\log\sigma^{*2} + \frac{d}{2}\log T_\epsilon(s_{\epsilon,t}) + \frac{k}{2}\log T_x(s_{x,t}) \\
&- \frac{T_\epsilon(s_{\epsilon,t})}{2\sigma^{*2}}\,Tr\left\{\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right]\right\} + \frac{T_\epsilon(s_{\epsilon,t})}{\sigma^{*2}}\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]\boldsymbol{\mu}^{*T} - \frac{T_\epsilon(s_{\epsilon,t})}{2\sigma^{*2}}\boldsymbol{\mu}^*\boldsymbol{\mu}^{*T}
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{T_\epsilon(s_{\epsilon,t})^2}{\sigma^{*2}}\,Tr\left\{\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right]\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}\right\}-\frac{T_\epsilon(s_{\epsilon,t})^2}{\sigma^{*2}}\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}\boldsymbol{\mu}^{*T}\\
&-\frac{T_\epsilon(s_{\epsilon,t})^2}{\sigma^{*2}}\left(\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]-\boldsymbol{\mu}^*\right)\mathbf{W}^*\mathbf{M}_t^{-1}\mathbf{W}^T\boldsymbol{\mu}^T\\
&-\frac{T_\epsilon(s_{\epsilon,t})^2}{2}\,Tr\left\{\mathbf{M}_t^{-1}\mathbf{W}^T\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right]\mathbf{W}\mathbf{M}_t^{-1}\left(\frac{T_\epsilon(s_{\epsilon,t})}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{W}^*+T_x(s_{x,t})\mathbb{I}_k\right)\right\}\\
&+\frac{T_\epsilon(s_{\epsilon,t})^2}{2}\,Tr\left\{\mathbf{M}_t^{-1}\mathbf{W}^T\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]^T\boldsymbol{\mu}\mathbf{W}\mathbf{M}_t^{-1}\left(\frac{T_\epsilon(s_{\epsilon,t})}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{W}^*+T_x(s_{x,t})\mathbb{I}_k\right)\right\}\\
&+\frac{T_\epsilon(s_{\epsilon,t})^2}{2}\,Tr\left\{\mathbf{M}_t^{-1}\mathbf{W}^T\boldsymbol{\mu}^T\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]\mathbf{W}\mathbf{M}_t^{-1}\left(\frac{T_\epsilon(s_{\epsilon,t})}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{W}^*+T_x(s_{x,t})\mathbb{I}_k\right)\right\}\\
&-\frac{T_\epsilon(s_{\epsilon,t})^2}{2}\,Tr\left\{\mathbf{M}_t^{-1}\mathbf{W}^T\boldsymbol{\mu}^T\boldsymbol{\mu}\mathbf{W}\mathbf{M}_t^{-1}\left(\frac{T_\epsilon(s_{\epsilon,t})}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{W}^*+T_x(s_{x,t})\mathbb{I}_k\right)\right\}\\
&-\frac{\sigma^2}{2}\,Tr\left\{\mathbf{M}_t^{-1}\left(\frac{T_\epsilon(s_{\epsilon,t})}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{W}^*+T_x(s_{x,t})\mathbb{I}_k\right)\right\},
\end{aligned}
$$

for the corresponding non-central moments of $\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi$ *specified as*

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]=\left[\mathbf{Y}_t^o,\,\boldsymbol{\mu}^m+\left(\mathbf{Y}_t^o-\boldsymbol{\mu}^o\right)\boldsymbol{\Sigma}_{y,t}^{oo\;-1}\boldsymbol{\Sigma}_{y,t}^{om}\right],\\
&\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right]=\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Sigma}_{y,t}^{mm}-\boldsymbol{\Sigma}_{y,t}^{mo}\boldsymbol{\Sigma}_{y,t}^{oo\;-1}\boldsymbol{\Sigma}_{y,t}^{om}\end{bmatrix}+\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]^T\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right],
\end{aligned}
$$

where $\boldsymbol{\mu}^m$ *contains elements of the vector* $\boldsymbol{\mu}$ *which correspond to the non-observed entries of the vector* $\mathbf{Y}_t$ *and* $\boldsymbol{\Sigma}_{y,t}^{mm}$ *is a square submatrix of* $\boldsymbol{\Sigma}_{y,t}$ *which correspond by rows and columns to missing elements of* $\mathbf{Y}_t$*. The elements of the* $d_m\times d_o$ *submatrix* $\boldsymbol{\Sigma}_{y,t}^{mo}=\boldsymbol{\Sigma}_{y,t}^{om\;T}$ *correspond by rows to missing entries of the vector* $\mathbf{Y}_t$ *and by columns to its observed elements.*

Proof. The proof of Corollary 2 is analogous to the proof of Theorem 25 and Corollary 15 when one sets $\mathbf{D}_{\epsilon,t}=T_\epsilon(s_{\epsilon,t}\mathbb{I}_d$, $\mathbf{D}_{x,t}=T_x(s_{x,t}\mathbb{I}_d$ and $\boldsymbol{\delta}_\epsilon=\boldsymbol{\delta}_x=\boldsymbol{\delta}_\epsilon^*=\boldsymbol{\delta}_x^*=\mathbf{0}$. □

Corollary 21. *Under the assumptions of Special Case 3: t-Student IND PPCA specified in Subsubsection 3.4.1.3, the M-step of the corresponding EM algorithm in incomplete data setting as a case-specific variant of Theorem 26 given the problem formulation* $\nabla_{\Psi^*}\tilde{Q}_{GStS,ind,3}=\mathbf{0}$ *for* $\Psi^*=\left[\mathbf{W}^*,\boldsymbol{\mu}^*,\sigma^{*2}\right]$ *is defined by* 2*-dimensional integration problems on the hypercube* $[0,1]^2$ *as following*

$$
\begin{cases}
\boldsymbol{\mu}^*=A_1(\mathbf{y}_{1:N}^o;\Psi)^{-1}\left[\mathbf{A}_6(\mathbf{y}_{1:N}^o;\Psi)-\mathbf{A}_{10}(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}+\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{13}(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}\right]\\
\mathbf{W}^*=\left[\mathbf{A}_{14}(\mathbf{y}_{1:N}^o;\Psi)-\boldsymbol{\mu}^{*T}\mathbf{A}_{10}(\mathbf{y}_{1:N}^o;\Psi)-\mathbf{A}_{16}(\mathbf{y}_{1:N}^o;\Psi)+\boldsymbol{\mu}^{*T}\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{13}(\mathbf{y}_{1:N}^o;\Psi)\right]\\
\quad\times\left[\mathbf{A}_{17}(\mathbf{y}_{1:N}^o;\Psi)-\mathbf{A}_{18}(\mathbf{y}_{1:N}^o;\Psi)-\mathbf{A}_{18}(\mathbf{y}_{1:N}^o;\Psi)^T+\mathbf{A}_{19}(\mathbf{y}_{1:N}^o;\Psi)+\sigma^2\mathbf{A}_{11}(\mathbf{y}_{1:N}^o;\Psi)\right]^{-1}\\
\sigma^{*2}=\frac{1}{dA_0(\mathbf{y}_{1:N}^o;\Psi)}\left[A_{20}(\mathbf{y}_{1:N}^o;\Psi)-2\mathbf{A}_6(\mathbf{y}_{1:N}^o;\Psi)\boldsymbol{\mu}^{*T}+A_1(\mathbf{y}_{1:N}^o;\Psi)\boldsymbol{\mu}^*\boldsymbol{\mu}^{*T}-2\,Tr\left\{\mathbf{A}_{14}(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}\right\}+2\mathbf{A}_9(\mathbf{y}_{1:N}^o;\Psi,\Psi^*)\mathbf{W}^T\boldsymbol{\mu}^T\right.\\
\quad+2\left(\mathbf{A}_{10}(\mathbf{y}_{1:N}^o;\Psi)-\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{13}(\mathbf{y}_{1:N}^o;\Psi)\right)\mathbf{W}^{*T}\boldsymbol{\mu}^{*T}+Tr\left\{\mathbf{A}_{17}(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}\mathbf{W}^*\right\}-Tr\left\{\left(\mathbf{A}_{18}(\mathbf{y}_{1:N}^o;\Psi)+\mathbf{A}_{18}(\mathbf{y}_{1:N}^o;\Psi)^T\right)\mathbf{W}^{*T}\mathbf{W}^*\right\}\\
\quad\left.+Tr\left\{\mathbf{A}_{19}(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}\mathbf{W}^*\right\}+\sigma^2\,Tr\left\{\mathbf{A}_{11}(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}\mathbf{W}^*\right\}\right].
\end{cases}
$$

The 2*-dimensional integrals* $\mathbf{A}_i$ *on the hypercube* $[0,1]^2$*, for* $i\in\left\{0,\ldots,21\right\}$*, are defined in Subsubsection A.4.2.5 in Appendix A.4 given the following simplifications*

$$
\mathbf{A}_{16}(\mathbf{y}_{1:N}^o;\Psi,\Psi^*)_{d\times k}:=\sum_{t=1}^N\left\{H\left(\mathbf{y}_{1:N/t}^o;\Psi\right)\int_0^1\int_0^1 T_\epsilon(s_{\epsilon,t})^2\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]^T\boldsymbol{\mu}\mathbf{W}\mathbf{M}_t^{-1}m(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\right\},
$$

$$\mathbf{A}_{18}(\mathbf{y}^o_{1:N};\Psi)_{k\times k} := \sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\int_0^1\int_0^1 T_\epsilon(s_{\epsilon,t})^3\mathbf{M}_t^{-1}\mathbf{W}^T\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]^T\boldsymbol{\mu}\mathbf{W}\mathbf{M}_t^{-1}m(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\right\},$$
$$\mathbf{A}_{19}(\mathbf{y}^o_{1:N};\Psi)_{k\times k} := \sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\int_0^1\int_0^1 T_\epsilon(s_{\epsilon,t})^3\mathbf{M}_t^{-1}\mathbf{W}^T\boldsymbol{\mu}^T\boldsymbol{\mu}\mathbf{W}\mathbf{M}_t^{-1}m(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\right\},$$

and the definition of the following moments of $\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi$ *are given in Corollary 20.*

Proof. The proof of Corollary 21 is analogous to the proof of Corollary 19 when one sets $\delta_\epsilon=\delta_x=\delta_\epsilon^*=\delta_x^*=\mathbf{0}$. □

4.4.2. EM Algorithm for Probabilistic Principal Component Analysis with Identical and Conditionally Independent Generalized Skew t-Student Distribution in the Presence of Missing Data

We provide the E-Step and M-step for Generalized Skew t-Student IID PPCA in Theorem 27 and Theorem 28 as equivalent steps of the EM Algorithm specified in Subsubsection 3.4.2.2 for the Generalized Skew t-Student IID PPCA model in the presence of missing elements of the observation process. In this framework we assume the identical and, consequently, the conditionally independent in time probability model of $\mathbf{X}_t$ and $\boldsymbol{\epsilon}_t$.

Theorem 27. *Let the observation vector* $\mathbf{Y}_t$ *be modelled as in Equation (2.7) with the latent processes* $\mathbf{X}_t$, $\boldsymbol{\epsilon}_t$, $\mathbf{U}_t$ *and* $\mathbf{V}_t$ *following the assumptions of Generalized Skew t-Student IID PPCA given in Subsection 3.4.1 in Equation (3.30) and Equation (3.31). Given N observations of the observed entries of the random vector* $\mathbf{Y}_t^o$, *denoted by* $\mathbf{y}^o_{1:N}:=\left[\mathbf{y}_1^o,\ldots,\mathbf{y}_N^o\right]$, *the E-step of the Expectation-Maximisation algorithm for Generalized Skew t-Student IID PPCA in incomplete data setting is specified as follows*

$$\tilde{Q}_{GStS,id}(\Psi,\Psi^*) = I_1(\mathbf{y}^o_{1:N};\Psi,\Psi^*) + I_2(\mathbf{y}^o_{1:N};\Psi)$$

for the functions $I_1,\ I_2:\mathbb{R}^{d_o\times N}\to\mathbb{R}$ *defined as the solutions to the integration problems on 2-dimensional unit cube given by*

$$I_1(\mathbf{y}^o_{1:N};\Psi,\Psi^*) := \frac{1}{\pi_{\mathbf{Y}^o_{1:N}|\Psi}(\mathbf{y}^o_{1:N})}\int_0^1\int_0^1\left\{\left(\sum_{t=1}^{N}\tilde{v}(\mathbf{y}_t^o,s_\epsilon,s_x;\Psi,\Psi^*)\right)m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\right\}ds_\epsilon\,ds_x,$$
$$I_2(\mathbf{y}^o_{1:N};\Psi) := -\frac{1}{\pi_{\mathbf{Y}^o_{1:N}|\Psi}(\mathbf{y}^o_{1:N})}\int_0^1\int_0^1\left\{\left(\log\left|\frac{\partial T_\epsilon(s_\epsilon)}{\partial s_\epsilon}\right|+\log\left|\frac{\partial T_x(s_x)}{\partial s_x}\right|\right)m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\right\}ds_\epsilon\,ds_x,$$

for function $\tilde{v}\left(\mathbf{y}_t^o,s_\epsilon,s_x;\Psi,\Psi^*\right) := v\left(\mathbf{y}_t^o,T_\epsilon(s_\epsilon),T_x(s_x);\Psi,\Psi^*\right)$ *specified in Lemma A15 in Appendix A.4 in Section A.4.6 and the function* $m:\mathbb{R}^{d_o\times N}\times[0,1]^2\longrightarrow\mathbb{R}$ *given by*

$$m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi) = \left(\prod_{t=1}^{N}\pi_{\mathbf{Y}_t^o|S_\epsilon,S_x,\Psi}(\mathbf{y}_t^o)\right)(\sigma^2)^{\frac{N}{2}}\left|\mathbf{D}_\epsilon\right|^{\frac{N}{2}}\left|\mathbf{D}_x\right|^{\frac{N}{2}}\left|\mathbf{M}\right|^{-\frac{N}{2}}\left|\mathbf{N}\right|^{-\frac{N}{2}}$$
$$\times\exp\left\{-\frac{N}{2}\delta_x\left(\mathbf{D}_x^{-1}-\sigma^2\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\mathbf{N}^{-1}\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}_t^{-1}-\sigma^2\mathbf{M}^{-1}\right)\delta_x^T\right\}.$$

The probability density function $\pi_{\mathbf{Y}_t^o|S_\epsilon,S_x,\Psi}(\mathbf{y}_t^o)$ *equals to the density function* $\pi_{\mathbf{Y}_t^o|T_\epsilon(S_\epsilon),T_x(S_x),\Psi}(\mathbf{y}_t^o)$ *specified in Lemma A14 in Appendix A.4 in Section A.4.6 and matrices* $\mathbf{D}_\epsilon$, $\mathbf{D}_x$, $\mathbf{M}$ *and* $\mathbf{N}$ *are defined as in Theorem 15.*

Proof. The proof of Theorem 27 is given in Subsection A.4.4.1 in Appendix A.4. □

Theorem 28. *The maximizers of the function* $\tilde{Q}_{GStS,id}$ *derived in Theorem 27 with respect to the vector of parameters* Ψ^* *are the solutions to the system of equations determined by* $\nabla_{\Psi^*}\tilde{Q}_{GStS,id}(\Psi,\Psi^*)=\mathbf{0}$ *as following*

$$\begin{cases} I(\frac{\partial}{\partial\boldsymbol{\mu}^*}\tilde{v})(\mathbf{y}_{1:N}^o)_{1\times d}=\mathbf{0} \\ I(\frac{\partial}{\partial\mathbf{W}^*}\tilde{v})(\mathbf{y}_{1:N})_{d\times k}=\mathbf{0} \\ I(\frac{\partial}{\partial\sigma^{*2}}\tilde{w})(\mathbf{y}_{1:N}^o)_{1\times 1}=0 \\ I(\frac{\partial}{\partial\boldsymbol{\delta}_\epsilon^*}\tilde{v})(\mathbf{y}_{1:N}^o)_{1\times d}=\mathbf{0} \\ I(\frac{\partial}{\partial\boldsymbol{\delta}_x^*}\tilde{v})(\mathbf{y}_{1:N}^o)_{1\times k}=\mathbf{0}, \end{cases}$$

for the integral operator I *defined in Equation (3.42). Denoting* $\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)_{1\times d}:=\frac{1}{N}\sum_{t=1}^N\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_\epsilon,S_x,\Psi}[\mathbf{Y}_t]$ *and* $\bar{\mathbf{S}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)_{d\times d}:=\frac{1}{N}\sum_{t=1}^N\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_\epsilon,S_x,\Psi}[\mathbf{Y}_t^T\mathbf{Y}_t]$, *the sum of the gradients of the function* $\tilde{v}$ *over the observed realisation of the random vector* $\mathbf{Y}_t^o$ *derived as following*

$$\begin{aligned}
&\sum_{t=1}^N\frac{\partial}{\partial\boldsymbol{\mu}^*}\tilde{v}(\mathbf{y}_t^o,s_\epsilon,s_x;\Psi,\Psi^*)=\frac{N}{\sigma^{*2}}\Big[\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon-\boldsymbol{\mu}^*\mathbf{D}_\epsilon-\boldsymbol{\delta}_\epsilon^*-\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon\\
&\quad+\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon+\boldsymbol{\delta}_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon-\sigma^2\boldsymbol{\delta}_x\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon\Big],\\
&\sum_{t=1}^N\frac{\partial}{\partial\mathbf{W}^*}\tilde{v}(\mathbf{y}_t^o,s_\epsilon,s_x;\Psi,\Psi^*)=\frac{N}{\sigma^{*2}}\Big[\mathbf{D}_\epsilon\bar{\mathbf{S}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}-(\mathbf{D}_\epsilon\boldsymbol{\mu}^{*T}+\boldsymbol{\delta}_\epsilon^{*T})\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\\
&\quad+(\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon-\boldsymbol{\mu}^*\mathbf{D}_\epsilon-\boldsymbol{\delta}_\epsilon^*)^T(-\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}-\boldsymbol{\delta}_\epsilon\mathbf{W}+\sigma^2\boldsymbol{\delta}_x)\mathbf{M}^{-1}-\mathbf{D}_\epsilon\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\bar{\mathbf{S}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\\
&\quad-\mathbf{D}_\epsilon\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)^T(-\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}-\boldsymbol{\delta}_\epsilon\mathbf{W}+\sigma^2\boldsymbol{\delta}_x)\mathbf{M}^{-1}-\mathbf{D}_\epsilon\mathbf{W}^*\mathbf{M}^{-1}(-\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}-\boldsymbol{\delta}_\epsilon\mathbf{W}+\sigma^2\boldsymbol{\delta}_x)^T\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\\
&\quad-\mathbf{D}_\epsilon\mathbf{W}^*\mathbf{M}^{-1}\Big(-\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}-\boldsymbol{\delta}_\epsilon\mathbf{W}+\sigma^2\boldsymbol{\delta}_x\Big)^T\Big(-\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}-\boldsymbol{\delta}_\epsilon\mathbf{W}+\sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}^{-1}-\sigma^2\mathbf{D}_\epsilon\mathbf{W}^*\mathbf{M}^{-1}\Big],\\
&\sum_{t=1}^N\frac{\partial}{\partial\sigma^{*2}}\tilde{v}(\mathbf{y}_t^o,s_\epsilon,s_x;\Psi,\Psi^*)=-\frac{dN}{2\sigma^{*2}}-\frac{N}{2(\sigma^{*2})^2}\Big[-Tr\Big\{\bar{\mathbf{S}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\Big\}+2\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)(\mathbf{D}_\epsilon\boldsymbol{\mu}^{*T}+\boldsymbol{\delta}_\epsilon^{*T})\\
&\quad-\boldsymbol{\mu}^*\mathbf{D}_\epsilon\boldsymbol{\mu}^{*T}-2\boldsymbol{\mu}^*\boldsymbol{\delta}_\epsilon^{*T}-\boldsymbol{\delta}_\epsilon^*\mathbf{D}_\epsilon^{-1}\boldsymbol{\delta}_\epsilon^{*T}+2\,Tr\Big\{\bar{\mathbf{S}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon\Big\}\\
&\quad-2\Big(\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}-\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}-\boldsymbol{\delta}_\epsilon\mathbf{W}+\sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}^{-1}\mathbf{W}^{*T}(\boldsymbol{\mu}^*\mathbf{D}_\epsilon+\boldsymbol{\delta}_\epsilon^*)^T\\
&\quad+2\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}^*\mathbf{M}^{-1}(-\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}-\boldsymbol{\delta}_\epsilon\mathbf{W}+\sigma^2\boldsymbol{\delta}_x)^T-\sigma^2\,Tr\Big\{\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon\mathbf{W}^*\Big\}\\
&\quad-Tr\Big\{\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\bar{\mathbf{S}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon\mathbf{W}^*\Big\}\\
&\quad-Tr\Big\{\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)^T\Big(-\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}-\boldsymbol{\delta}_\epsilon\mathbf{W}+\sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon\mathbf{W}^*\Big\}\\
&\quad-Tr\Big\{\mathbf{M}^{-1}\Big(-\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}-\boldsymbol{\delta}_\epsilon\mathbf{W}+\sigma^2\boldsymbol{\delta}_x\Big)^T\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon\mathbf{W}^*\Big\}\\
&\quad-Tr\Big\{\mathbf{M}^{-1}\Big(-\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}-\boldsymbol{\delta}_\epsilon\mathbf{W}+\sigma^2\boldsymbol{\delta}_x\Big)^T\Big(-\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}-\boldsymbol{\delta}_\epsilon\mathbf{W}+\sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon\mathbf{W}^*\Big\}\Big]\\
&\sum_{t=1}^N\frac{\partial}{\partial\boldsymbol{\delta}_\epsilon^*}\tilde{v}\Big(\mathbf{y}_t^o,s_\epsilon,s_x;\Psi,\Psi^*\Big)=\frac{N}{\sigma^{*2}}\Big[\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)-\boldsymbol{\mu}^*-\boldsymbol{\delta}_\epsilon^*\mathbf{D}_\epsilon^{-1}-\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\\
&\quad+\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}+\boldsymbol{\delta}_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}-\sigma^2\boldsymbol{\delta}_x\mathbf{M}^{-1}\mathbf{W}^{*T}\Big],
\end{aligned}$$

$$\sum_{t=1}^{N}\frac{\partial}{\partial\delta_x^*}\bar{v}(\mathbf{y}_t^o,s_\epsilon,s_x;\Psi,\Psi^*)=N\Big[-\delta_x^*\mathbf{D}_x^{-1}+\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}-\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}-\delta_\epsilon\mathbf{W}\mathbf{M}^{-1}+\sigma^2\delta_x\mathbf{M}^{-1}\Big].$$

Proof. The proof of the Theorem 28 is given in Appendix A.4 Subsection A.4.4.2. □

Corollary 22. *The solutions to the system of equation from Theorem 28 which specify the maximizers of the function $\tilde{Q}_{GStS,id}$ with respect to the parameters $\boldsymbol{\mu}^*,\delta_\epsilon^*,\delta_x^*,\sigma^2$ are given by explicit formulas defined by 2-dimensional integration problems on the hypercube $[0,1]^2$ as follows*

$$\begin{cases}
\boldsymbol{\mu}^* = \Big[\mathbf{A}_6(\mathbf{y}_{1:N}^o;\Psi)-\mathbf{A}_{10}(\mathbf{y}_{1:N}^o;\Psi,\Psi^*)-\delta_\epsilon^*A_0(\mathbf{y}_{1:N}^o;\Psi)+\boldsymbol{\mu}\mathbf{A}_{13}(\mathbf{y}_{1:N}^o;\Psi,\Psi^*)+\Big(\delta_\epsilon\mathbf{W}-\sigma^2\delta_x\Big)\mathbf{A}_{12}(\mathbf{y}_{1:N}^o;\Psi,\Psi^*)^T\Big]\mathbf{A}_1(\mathbf{y}_{1:N}^o;\Psi)^{-1}\\
\delta_\epsilon^* = \Big[\mathbf{A}_5(\mathbf{y}_{1:N}^o;\Psi)-\boldsymbol{\mu}^*A_0(\mathbf{y}_{1:N}^o;\Psi)-\mathbf{A}_8(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}+\boldsymbol{\mu}\mathbf{A}_{11}(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}+(\delta_\epsilon\mathbf{W}-\sigma^2\delta_x)\mathbf{A}_4(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}\Big]\mathbf{A}_2(\mathbf{y}_{1:N}^o;\Psi)^{-1}\\
\delta_x^* = \Big[\mathbf{A}_8(\mathbf{y}_{1:N}^o;\Psi)-\boldsymbol{\mu}\mathbf{A}_{11}(\mathbf{y}_{1:N}^o;\Psi)-(\delta_\epsilon\mathbf{W}-\sigma^2\delta_x)\mathbf{A}_4(\mathbf{y}_{1:N}^o;\Psi)\Big]\mathbf{A}_3(\mathbf{y}_{1:N}^o,\Psi)^{-1}\\
\sigma^{*2} = \frac{1}{dA_0(\mathbf{y}_{1:N}^o;\Psi)}\Big[A_{21}(\mathbf{y}_{1:N}^o;\Psi)+\boldsymbol{\mu}^*\mathbf{A}_1(\mathbf{y}_{1:N}^o;\Psi)\boldsymbol{\mu}^{*T}+\delta_\epsilon^*\mathbf{A}_2(\mathbf{y}_{1:N}^o;\Psi)\delta_\epsilon^{*T}+2\mathbf{A}_9(\mathbf{y}_{1:N}^o;\Psi,\Psi^*)\boldsymbol{\mu}^T\\
\qquad+2\Big(\mathbf{A}_{10}(\mathbf{y}_{1:N}^o;\Psi,\Psi^*)-\mathbf{A}_6(\mathbf{y}_{1:N}^o;\Psi)-\boldsymbol{\mu}\mathbf{A}_{13}(\mathbf{y}_{1:N}^o;\Psi,\Psi^*)\Big)\boldsymbol{\mu}^{*T}-2\,Tr\Big\{A_{14}(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}\Big\}+\sigma^2\,Tr\Big\{\mathbf{A}_{12}(\mathbf{y}_{1:N}^o;\Psi)^T\mathbf{W}^*\Big\}\\
\qquad+Tr\Big\{\mathbf{A}_{17}(\mathbf{y}_{1:N}^o;\Psi)^T\mathbf{W}^*\Big\}-Tr\Big\{\mathbf{A}_{18}(\mathbf{y}_{1:N}^o;\Psi)^T\mathbf{W}^*\Big\}-Tr\Big\{\mathbf{A}_{19}(\mathbf{y}_{1:N}^o;\Psi)^T\mathbf{W}^*\Big\}+Tr\Big\{\mathbf{A}_{20}(\mathbf{y}_{1:N}^o;\Psi)^T\mathbf{W}^*\Big\}\\
\qquad+2\Big(\boldsymbol{\mu}^*A_0(\mathbf{y}_{1:N}^o;\Psi)-\mathbf{A}_5(\mathbf{y}_{1:N}^o;\Psi)+\mathbf{A}_8(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}-\boldsymbol{\mu}\mathbf{A}_{11}(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}\Big)\delta_\epsilon^{*T}\\
\qquad+2\Big(\mathbf{A}_7(\mathbf{y}_{1:N}^o;\Psi,\Psi^*)-\boldsymbol{\mu}^*\mathbf{A}_{12}(\mathbf{y}_{1:N}^o;\Psi,\Psi^*)-\delta_\epsilon^*\mathbf{W}^*\mathbf{A}_4(\mathbf{y}_{1:N}^o;\Psi)\Big)(\delta_\epsilon\mathbf{W}-\sigma^2\delta_x)^T\Big].
\end{cases}$$

The solutions with respect to $\boldsymbol{\mu}^*$, δ_ϵ^* and σ^{*2} are linear function of the parameter $\mathbf{W}^*$. The maximizer of the function $\tilde{Q}_{GStS,id}$ with respect to $\mathbf{W}^*$ is determined by the equation

$$f(\mathbf{W}^*;\Psi,\Psi^*)=\mathbf{0},$$

for the function $f:\mathbb{R}^{d\times k}\to\mathbb{R}^{d\times k}$ such that $f(\mathbf{W}^*;\Psi,\Psi^*):=\frac{\partial\tilde{Q}_{GStS,id}(\Psi,\Psi^*)}{\partial\mathbf{W}^*}$ being equal to

$$\begin{aligned}
f(\mathbf{W}^*;\Psi,\Psi^*) &= \mathbf{A}_{14}(\mathbf{y}_{1:N};\Psi)-\mathbf{A}_{15}(\mathbf{y}_{1:N};\Psi,\Psi^*)-\delta_\epsilon^{*T}\mathbf{A}_8(\mathbf{y}_{1:N};\Psi)-\mathbf{A}_{16}(\mathbf{y}_{1:N};\Psi,\Psi^*)+\delta_\epsilon^{*T}\boldsymbol{\mu}\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi)\\
&\quad+\delta_\epsilon^{*T}(\delta\mathbf{W}-\sigma^2\mathbf{W})\mathbf{A}_4(\mathbf{y}_{1:N};\Psi)-\mathbf{A}_{17}(\mathbf{y}_{1:N};\Psi,\Psi^*)+\mathbf{A}_{18}(\mathbf{y}_{1:N};\Psi,\Psi^*)+\mathbf{A}_{19}(\mathbf{y}_{1:N};\Psi,\Psi^*)-\mathbf{A}_{20}(\mathbf{y}_{1:N};\Psi,\Psi^*)\\
&\quad-\sigma^2\mathbf{A}_{12}(\mathbf{y}_{1:N};\Psi,\Psi^*),
\end{aligned}$$

The function f is linear with respect to the parameters $\boldsymbol{\mu}^*$ and δ_ϵ^*. The 2-dimensional integrals $\mathbf{A}_i$ on the hypercube $[0,1]^2$, for $i\in\Big\{0,\ldots,21\Big\}$, are defined in Subsubsection A.4.4.3 in Appendix A.4.

Proof. The proof of Corollary 22 is located in Subsubsection A.4.4.3 in Appendix A.4. □

4.4.2.1. EM Algorithm for Special Cases of Probabilistic Principal Component with Identical and Conditionally Independent Generalized Skew t-Student Distribution in the Presence of Missing Data

In the following part we derive the formulations of the E-step and the M-step of EM algorithms for special cases of the Generalized Skew t-Student IID PPCA specified in Subsubsection 3.4.1.3. The reader may find the derivations under following notation

Special Case 1: Grouped t-Student IID PPCA in Corollary 23 and Corollary 24,
Special Case 2: Skew t-Student IID PPCA in Corollary 25 and Corollary 26,
Special Case 3: t-Student IID PPCA in Corollary 27 and Corollary 28.

Corollary 23. *Under the assumptions of Special Case 1: Grouped t-Student IID PPCA specified in Subsubsection 3.4.1.3, the E-step of the corresponding EM algorithm in incomplete data setting as a case-specific variant of Theorem 27, is given by*

$$\tilde{Q}_{GStS,id,1}(\Psi,\Psi^*)=\mathbb{E}_{\mathbf{Y}^m_{1:N},\mathbf{X}_{1:N},\mathbf{U},\mathbf{V}|\mathbf{Y}^o_{1:N},\Psi}\left[\log\left(\pi_{\mathbf{Y}_{1:N},\mathbf{X}_{1:N},\mathbf{U},\mathbf{V}|\Psi^*}\left(\mathbf{Y}_{1:N},\mathbf{X}_{1:N},\mathbf{U},\mathbf{V}\right)\right)\right],$$

defined the functions $I_1,\ I_2:\mathbb{R}^{d_o\times N}\to\mathbb{R}$ *from Theorem 27 given the case-specific formulation of the function* $m:\mathbb{R}^{d_o\times N}\times[0,1]^2\longrightarrow\mathbb{R}$ *defined as*

$$m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)=\left(\prod_{t=1}^{N}\pi_{\mathbf{Y}^o_t|S_\epsilon,S_x,\Psi}(\mathbf{y}^o_t)\right)(\sigma^2)^{\frac{N}{2}}\left|\mathbf{D}_\epsilon\right|^{\frac{N}{2}}\left|\mathbf{D}_x\right|^{\frac{N}{2}}\left|\mathbf{M}\right|^{-\frac{N}{2}}\left|\mathbf{N}\right|^{-\frac{N}{2}},$$

the probability function $\pi_{\mathbf{Y}^o_t|S_\epsilon,S_x,\Psi}(\mathbf{y}^o_t)$ *of the* d_o*-dimensional random vector* $\mathbf{Y}^o_t|S_\epsilon,Sx,\Psi\sim\mathcal{N}(\boldsymbol{\mu}^o,\boldsymbol{\Sigma}^{oo}_y)$*, where* $\boldsymbol{\mu}^o$ *contains the elements of the vector* $\boldsymbol{\mu}$ *which correspond to the observed entries of the vector* $\mathbf{Y}_t$ *and* $\boldsymbol{\Sigma}^{oo}_y$ *is a submatrix of* $\boldsymbol{\Sigma}_y=\sigma^2\mathbf{N}^{-1}$ *which correspond by rows and columns to the observed elements of* $\mathbf{Y}_t$*, and matrices* $\mathbf{M},\ \mathbf{N},\mathbf{D}_\epsilon$ *and* $\mathbf{D}_x$ *defined as in Theorem 27 and the function* $\tilde{v}:\mathbb{R}^{d_o}\times[0,1]^2\longrightarrow\mathbb{R}$ *is defined as*

$$\begin{aligned}
\tilde{v}(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi,\Psi^*)&=-\frac{k+d}{2}\log 2\pi-\frac{d}{2}\log\sigma^{*2}+\frac{1}{2}\log\left|\mathbf{D}_\epsilon\right|+\frac{1}{2}\log\left|\mathbf{D}_x\right|\\
&-\frac{1}{2\sigma^{*2}}Tr\left\{\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_\epsilon,S_x,\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right]\mathbf{D}_\epsilon\right\}+\frac{1}{\sigma^{*2}}\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_\epsilon,S_x,\Psi}\left[\mathbf{Y}_t\right]\mathbf{D}_\epsilon\boldsymbol{\mu}^{*T}-\frac{1}{2\sigma^{*2}}\boldsymbol{\mu}^*\mathbf{D}_\epsilon\boldsymbol{\mu}^{*T}\\
&+\frac{1}{\sigma^{*2}}Tr\left\{\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_\epsilon,S_x,\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right]\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon\right\}-\frac{1}{\sigma^{*2}}\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_\epsilon,S_x,\Psi}\left[\mathbf{Y}_t\right]\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon\boldsymbol{\mu}^{*T}\\
&-\frac{1}{\sigma^{*2}}\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_\epsilon,S_x,\Psi}\left[\mathbf{Y}_t\right]\mathbf{D}_\epsilon\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\boldsymbol{\mu}^T+\frac{1}{\sigma^{*2}}\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon\boldsymbol{\mu}^{*T}\\
&-\frac{1}{2}Tr\left\{\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_\epsilon,S_x,\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right]\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\left(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_\epsilon\mathbf{W}^*+\mathbf{D}_x\right)\right\}\\
&+\frac{1}{2}Tr\left\{\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_\epsilon,S_x,\Psi}\left[\mathbf{Y}_t\right]^T\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\left(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_\epsilon\mathbf{W}^*+\mathbf{D}_x\right)\right\}\\
&+\frac{1}{2}Tr\left\{\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\mathbf{W}^T\mathbf{D}_\epsilon\boldsymbol{\mu}^T\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_\epsilon,S_x,\Psi}\left[\mathbf{Y}_t\right]\mathbf{M}^{-1}\left(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_\epsilon\mathbf{W}^*+\mathbf{D}_x\right)\right\}\\
&-\frac{1}{2}Tr\left\{\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\boldsymbol{\mu}^T\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\left(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_\epsilon\mathbf{W}^*+\mathbf{D}_x\right)\right\}\\
&-\frac{\sigma^2}{2}Tr\left\{\mathbf{M}^{-1}\left(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_\epsilon\mathbf{W}^*+\mathbf{D}_x\right)\right\},
\end{aligned}$$

for the corresponding non-central moments of $\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_\epsilon,S_x,\Psi$ *specified as*

$$\begin{aligned}
\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_\epsilon,S_x,\Psi}\left[\mathbf{Y}_t\right]&=\left[\mathbf{Y}^o_t,\ \boldsymbol{\mu}^m+(\mathbf{Y}^o_t-\boldsymbol{\mu}^o)\boldsymbol{\Sigma}^{oo\ -1}_y\boldsymbol{\Sigma}^{om}_y\right],\\
\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_\epsilon,S_x,\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right]&=\begin{bmatrix}\mathbf{0}&\mathbf{0}\\\mathbf{0}&\boldsymbol{\Sigma}^{mm}_y-\boldsymbol{\Sigma}^{mo}_y\boldsymbol{\Sigma}^{oo\ -1}_y\boldsymbol{\Sigma}^{om}_y\end{bmatrix}+\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_\epsilon,S_x,\Psi}\left[\mathbf{Y}_t\right]^T\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_\epsilon,S_x,\Psi}\left[\mathbf{Y}_t\right],
\end{aligned}$$

where $\boldsymbol{\mu}^m$ *contains elements of the vector* $\boldsymbol{\mu}$ *which correspond to the non-observed entries of the vector* $\mathbf{Y}_t$ *and* $\boldsymbol{\Sigma}^{mm}_y$ *is a square submatrix of* $\boldsymbol{\Sigma}_y$ *which correspond by rows and columns to missing elements of* $\mathbf{Y}_t$*. The elements of the* $d_m\times d_o$ *submatrix* $\boldsymbol{\Sigma}^{mo}_y=\boldsymbol{\Sigma}^{om\ T}_{y,t}$ *correspond by rows to missing entries of the vector* $\mathbf{Y}_t$ *and by columns to its observed elements.*

Proof. The proof of Corollary 23 is analogous to the proof of Theorem 27 when one sets $\delta_\epsilon = \delta_x = \delta_\epsilon^* = \delta_x^* = \mathbf{0}$. □

Corollary 24. *Under the assumptions of Special Case 1: Grouped t-Student IID PPCA specified in Subsubsection 3.4.1.3, the M-step of the corresponding EM algorithm in incomplete data setting as a case-specific variant of Theorem 28 given the problem formulation* $\nabla_{\Psi^*}\tilde{Q}_{GStS,id,1} = \mathbf{0}$ *for* $\Psi^* = \left[\mathbf{W}^*, \boldsymbol{\mu}^*, \sigma^{*2}\right]$ *is defined by 2-dimensional integration problems on the hypercube* $[0,1]^2$ *as following*

$$\begin{cases} \boldsymbol{\mu}^* = \left[\mathbf{A}_6(\mathbf{y}^o_{1:N};\Psi) - \mathbf{A}_{10}(\mathbf{y}^o_{1:N};\Psi,\Psi^*) + \boldsymbol{\mu}\mathbf{A}_{13}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\right]\mathbf{A}_1(\mathbf{y}^o_{1:N};\Psi)^{-1} \\ \sigma^{*2} = \frac{1}{dA_0(\mathbf{y}^o_{1:N};\Psi)}\Big[A_{21}(\mathbf{y}^o_{1:N};\Psi) + \boldsymbol{\mu}^*\mathbf{A}_1(\mathbf{y}^o_{1:N};\Psi)\boldsymbol{\mu}^{*T} + 2\mathbf{A}_9(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\boldsymbol{\mu}^T - 2\,Tr\left\{A_{14}(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}^{*T}\right\} \\ \quad +2\Big(\mathbf{A}_{10}(\mathbf{y}^o_{1:N};\Psi,\Psi^*) - \mathbf{A}_6(\mathbf{y}^o_{1:N};\Psi) - \boldsymbol{\mu}\mathbf{A}_{13}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\Big)\boldsymbol{\mu}^{*T} + \sigma^2\,Tr\left\{\mathbf{A}_{12}(\mathbf{y}^o_{1:N};\Psi)^T\mathbf{W}^*\right\} \\ \quad + Tr\left\{\mathbf{A}_{17}(\mathbf{y}^o_{1:N};\Psi)^T\mathbf{W}^*\right\} - Tr\left\{\mathbf{A}_{18}(\mathbf{y}^o_{1:N};\Psi)^T\mathbf{W}^*\right\} - Tr\left\{\mathbf{A}_{19}(\mathbf{y}^o_{1:N};\Psi)^T\mathbf{W}^*\right\} + Tr\left\{\mathbf{A}_{20}(\mathbf{y}^o_{1:N};\Psi)^T\mathbf{W}^*\right\}\Big]. \end{cases}$$

The solutions with respect to $\boldsymbol{\mu}^*$ *and* σ^{*2} *are linear function of the parameter* $\mathbf{W}^*$*. The maximizer of the function* $\tilde{Q}_{GStS,id,1}$ *with respect to* $\mathbf{W}^*$ *zero root of the case specific formulation of the function* $f : \mathbb{R}^{d\times k} \to \mathbb{R}^{d\times k}$ *such that* $f(\mathbf{W}^*;\Psi,\Psi^*) := \frac{\partial\tilde{Q}_{GStS,id,1}(\Psi,\Psi^*)}{\partial\mathbf{W}^*}$ *being equal to*

$$\begin{aligned} f(\mathbf{W}^*;\Psi,\Psi^*) = {} & \mathbf{A}_{14}(\mathbf{y}_{1:N};\Psi) - \mathbf{A}_{15}(\mathbf{y}_{1:N};\Psi,\Psi^*) - \mathbf{A}_{16}(\mathbf{y}_{1:N};\Psi,\Psi^*) - \mathbf{A}_{17}(\mathbf{y}_{1:N};\Psi,\Psi^*) + \mathbf{A}_{18}(\mathbf{y}_{1:N};\Psi,\Psi^*) + \mathbf{A}_{19}(\mathbf{y}_{1:N};\Psi,\Psi^*) \\ & - \mathbf{A}_{20}(\mathbf{y}_{1:N};\Psi,\Psi^*) - \sigma^2\mathbf{A}_{12}(\mathbf{y}_{1:N};\Psi,\Psi^*), \end{aligned}$$

and is linear with respect to the parameters $\boldsymbol{\mu}^*$*. The 2-dimensional integrals* $\mathbf{A}_i$ *on the hypercube* $[0,1]^2$*, for* $i \in \left\{0,\ldots,21\right\}$*, are defined in Subsubsection A.4.4.3 in Appendix A.4 given the following simplifications*

$$\begin{aligned} \mathbf{A}_{16}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)_{d\times k} &:= \int_0^1\int_0^1 \mathbf{D}_\epsilon\left(\bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi) - \boldsymbol{\mu}^*\right)^T\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\, m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\, ds_\epsilon\, ds_x, \\ \mathbf{A}_{18}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)_{k\times d} &:= \int_0^1\int_0^1 \mathbf{D}_\epsilon\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)^T\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\, m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\, ds_\epsilon\, ds_x, \\ \mathbf{A}_{19}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)_{k\times d} &:= \int_0^1\int_0^1 \mathbf{D}_\epsilon\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\boldsymbol{\mu}^T\bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\, m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\, ds_\epsilon\, ds_x, \\ \mathbf{A}_{20}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)_{k\times d} &:= \int_0^1\int_0^1 \mathbf{D}_\epsilon\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\boldsymbol{\mu}^T\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\, m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\, ds_\epsilon\, ds_x. \end{aligned}$$

Proof. The proof of Corollary 24 is analogous to the proof of Theorem 27 and Corollary 22 when one sets $\delta_\epsilon = \delta_x = \delta_\epsilon^* = \delta_x^* = \mathbf{0}$. □

Corollary 25. *Under the assumptions of Special Case 2: Skew t-Student IID PPCA specified in Subsubsection 3.4.1.3, the E-step of the corresponding EM algorithm in incomplete data setting as a case-specific variant of Theorem 27, is given by*

$$\tilde{Q}_{GStS,id,2}(\Psi,\Psi^*) = \mathbb{E}_{\mathbf{Y}^m_{1:N},\mathbf{X}_{1:N},U,V|\mathbf{Y}^o_{1:N},\Psi}\left[\log\left(\pi_{\mathbf{Y}_{1:N},\mathbf{X}_{1:N},UV|\Psi^*}\left(\mathbf{Y}_{1:N},\mathbf{X}_{1:N},U,V\right)\right)\right],$$

defined by the functions $I_1,\ I_2 : \mathbb{R}^{d_o\times N} \to \mathbb{R}$ *from Theorem 27 given the case-specific formulation of the function* $m : \mathbb{R}^{d_o\times N}\times[0,1]^2 \longrightarrow \mathbb{R}$ *given by*

$$\begin{aligned} m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi) = {} & \left(\prod_{t=1}^N \pi_{\mathbf{Y}^o_t|S_\epsilon,S_x,\Psi}(\mathbf{y}^o_t)\right)(\sigma^2)^{\frac{kN}{2}}T_\epsilon(s_{\epsilon,t})^{\frac{dN}{2}}T_x(s_{x,t})^{\frac{kN}{2}}\left|\mathbf{M}\right|^{-\frac{1}{2}}\left|\mathbf{N}\right|^{-\frac{1}{2}} \\ & \times e^{-\frac{1}{2}\delta_x\left(T_x(s_x)^{-1}\mathbb{I}_k - \sigma^2T_\epsilon(s_\epsilon)^2\mathbf{M}^{-1}\mathbf{W}^T\mathbf{N}^{-1}\mathbf{W}\mathbf{M}^{-1} - \sigma^2\mathbf{M}^{-1}\right)\delta_x^T}, \end{aligned}$$

for the probability function $\pi_{\mathbf{Y}_t^o|S_\epsilon,S_x,\Psi}(\mathbf{y}_t^o)$ *of the* d_o*-dimensional random vector* $\mathbf{Y}_t^o|S_\epsilon,S_x,\Psi \sim \mathcal{N}(\boldsymbol{\mu}_y^o,\boldsymbol{\Sigma}_y^{oo})$*, where* $\boldsymbol{\mu}_y^o$ *contains elements of the vector* $\boldsymbol{\mu}_y$ *which correspond to the observed entries of the vector* $\mathbf{Y}_t$ *and* $\boldsymbol{\Sigma}_y^{oo}$ *is a submatrix of* $\boldsymbol{\Sigma}_y$ *which correspond by rows and columns to observed elements of* $\mathbf{Y}_t$*, for* $\boldsymbol{\mu}_y$ *and* $\boldsymbol{\Sigma}_y$ *specified as*

$$\boldsymbol{\mu}_y = \boldsymbol{\mu} + T_\epsilon(s_\epsilon)^{-1}\delta_\epsilon + \sigma^2 T_\epsilon(s_{\epsilon,t})\delta_x \mathbf{M}^{-1}\mathbf{W}^T\mathbf{N}^{-1} \text{ and } \boldsymbol{\Sigma}_y = \sigma^2\mathbf{N}^{-1}$$

and matrices $\mathbf{M}$ *and* $\mathbf{N}$ *are defined as in Corollary 9. The case-specific formulation of the function* $\tilde{v}$: $\mathbb{R}^{d_o}\times[0,1]^2 \longrightarrow \mathbb{R}$ *is given by*

$$\begin{aligned}
\tilde{v}(\mathbf{y}_t^o, s_\epsilon, s_x; \Psi, \Psi^*) = &-\frac{k+d}{2}\log 2\pi - \frac{d}{2}\log\sigma^{*2} + \frac{d}{2}T_\epsilon(s_\epsilon) + \frac{k}{2}T_x(s_x) - \frac{1}{2}T_x(s_x)^{-1}\delta_x^*\delta_x^{*T} \\
&- \frac{1}{2\sigma^{*2}}T_\epsilon(s_\epsilon)\,Tr\left\{\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_\epsilon,S_x,\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right]\right\} + \frac{1}{\sigma^{*2}}\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_\epsilon,S_x,\Psi}\left[\mathbf{Y}_t\right]\left(T_\epsilon(s_\epsilon)\boldsymbol{\mu}^{*T} + \delta_\epsilon^{*T}\right) \\
&- \frac{1}{2\sigma^{*2}}T_\epsilon(s_\epsilon)\boldsymbol{\mu}^*\boldsymbol{\mu}^{*T} - \frac{1}{\sigma^{*2}}\boldsymbol{\mu}^*\delta_\epsilon^{*T} - \frac{1}{2\sigma^{*2}}T_\epsilon(s_\epsilon)^{-1}\delta_\epsilon^*\delta_\epsilon^{*T} + \frac{1}{\sigma^{*2}}T_\epsilon(s_\epsilon)^2\,Tr\left\{\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_\epsilon,S_x,\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right]\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\right\} \\
&+ T_\epsilon(s_\epsilon)\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_\epsilon,S_x,\Psi}\left[\mathbf{Y}_t\right]\mathbf{W}\mathbf{M}^{-1}\left(-\frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})\mathbf{W}^{*T}\boldsymbol{\mu}^{*T} - \frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\delta_\epsilon^{*T} + \delta_x^{*T}\right) \\
&+ \frac{1}{\sigma^{*2}}T_\epsilon(s_\epsilon)\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_\epsilon,S_x,\Psi}\left[\mathbf{Y}_t\right]\mathbf{W}^*\mathbf{M}^{-1}\left(-\boldsymbol{\mu}T_\epsilon(s_\epsilon)\mathbf{W} - \delta_\epsilon\mathbf{W} + \sigma^2\delta_x\right)^T \\
&+ \left(-T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W} - \delta_\epsilon\mathbf{W} + \sigma^2\delta_x\right)\mathbf{M}^{-1}\left(-\frac{1}{\sigma^{*2}}T_\epsilon(s_\epsilon)\mathbf{W}^{*T}\boldsymbol{\mu}^{*T} - \frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\delta_\epsilon^{*T} + \delta_x^{*T}\right) \\
&- \frac{1}{2}T_\epsilon(s_\epsilon)^2\,Tr\left\{\mathbf{M}^{-1}\mathbf{W}^T\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_\epsilon,S_x,\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right]\mathbf{W}\mathbf{M}^{-1}\left(\frac{1}{\sigma^{*2}}T_\epsilon(s_\epsilon)\mathbf{W}^{*T}\mathbf{W}^* + T_x(s_x)\mathbb{I}_k\right)\right\} \\
&- \frac{1}{2}T_\epsilon(s_\epsilon)\,Tr\left\{\mathbf{M}^{-1}\mathbf{W}^T\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_\epsilon,S_x,\Psi}\left[\mathbf{Y}_t\right]^T\left(-T_\epsilon(s_\epsilon)\boldsymbol{\mu}\mathbf{W} - \delta_\epsilon\mathbf{W} + \sigma^2\delta_x\right)\mathbf{M}^{-1}\left(\frac{1}{\sigma^{*2}}T_\epsilon(s_\epsilon)\mathbf{W}^{*T}\mathbf{W}^* + T_x(s_x)\mathbb{I}_k\right)\right\} \\
&- \frac{1}{2}T_\epsilon(s_\epsilon)\,Tr\left\{\mathbf{M}^{-1}\left(-T_\epsilon(s_\epsilon)\boldsymbol{\mu}\mathbf{W} - \delta_\epsilon\mathbf{W} + \sigma^2\delta_x\right)^T\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_\epsilon,S_x,\Psi}\left[\mathbf{Y}_t\right]\mathbf{W}^T\mathbf{M}^{-1}\left(\frac{1}{\sigma^{*2}}T_\epsilon(s_\epsilon)\mathbf{W}^{*T}\mathbf{W}^* + T_x(s_x)\mathbb{I}_k\right)\right\} \\
&- \frac{1}{2}\,Tr\left\{\mathbf{M}^{-1}\left(-T_\epsilon(s_\epsilon)\boldsymbol{\mu}\mathbf{W} - \delta_\epsilon\mathbf{W} + \sigma^2\delta_x\right)^T\left(-T_\epsilon(s_\epsilon)\boldsymbol{\mu}\mathbf{W} - \delta_\epsilon\mathbf{W} + \sigma^2\delta_x\right)\mathbf{M}^{-1}\left(\frac{1}{\sigma^{*2}}T_\epsilon(s_\epsilon)\mathbf{W}^{*T}\mathbf{W}^* + T_x(s_x)\mathbb{I}_k\right)\right\} \\
&- \frac{\sigma^2}{2}\,Tr\left\{\mathbf{M}_t^{-1}\left(\frac{1}{\sigma^{*2}}T_\epsilon(s_\epsilon)\mathbf{W}^{*T}\mathbf{W}^* + T_x(s_x)\mathbb{I}_k\right)\right\},
\end{aligned}$$

for the corresponding non-central moments of $\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_\epsilon,S_x,\Psi$ *specified as*

$$\begin{aligned}
\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_\epsilon,S_x,\Psi}\left[\mathbf{Y}_t\right] &= \left[\mathbf{Y}_t^o,\ \boldsymbol{\mu}_y^m + \left(\mathbf{Y}_t^o - \boldsymbol{\mu}_y^o\right)\boldsymbol{\Sigma}_y^{oo\ -1}\boldsymbol{\Sigma}_y^{om}\right], \\
\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_\epsilon,S_x,\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right] &= \begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Sigma}_y^{mm} - \boldsymbol{\Sigma}_y^{mo}\boldsymbol{\Sigma}_y^{oo\ -1}\boldsymbol{\Sigma}_y^{om}\end{bmatrix} + \mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_\epsilon,S_x,\Psi}\left[\mathbf{Y}_t\right]^T\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_\epsilon,S_x,\Psi}\left[\mathbf{Y}_t\right],
\end{aligned}$$

where $\boldsymbol{\mu}_y^m$ *contains elements of the vector* $\boldsymbol{\mu}_y$ *which correspond to the non-observed entries of the vector* $\mathbf{Y}_t$ *and* $\boldsymbol{\Sigma}_y^{mm}$ *is a square submatrix of* $\boldsymbol{\Sigma}_y$ *which correspond by rows and columns to missing elements of* $\mathbf{Y}_t$*. The elements of the* $d_m \times d_o$ *submatrix* $\boldsymbol{\Sigma}_y^{mo} = \boldsymbol{\Sigma}_y^{om\ T}$ *correspond by rows to missing entries of the vector* $\mathbf{Y}_t$ *and by columns to its observed elements.*

Proof. The proof of Corollary 25 is analogous to the proof of Theorem 27 and Corollary 22 when one sets $\mathbf{D}_\epsilon = T_\epsilon(s_\epsilon)\mathbb{I}_d$ and $\mathbf{D}_x = T_x n(s_x)\mathbb{I}_k$ □

Corollary 26. *Under the assumptions of Special Case 2: Skew t-Student IID PPCA specified in Subsubsection 3.4.1.3, the M-step of the corresponding EM algorithm in incomplete data setting as*

a case-specific variant of Theorem 28 given the problem formulation $\nabla_{\Psi^*}\tilde{Q}_{GStS,id,2} = \mathbf{0}$ *for* $\Psi^* = \left[\mathbf{W}^*, \boldsymbol{\mu}^*, \delta_\epsilon^*, \boldsymbol{\delta}_x^*, \sigma^{*2}\right]$ *is defined by 2-dimensional integration problems on the hypercube* $[0,1]^2$ *as following*

$$
\begin{cases}
\boldsymbol{\mu}^* = A_1(\mathbf{y}_{1:N}^o;\Psi)^{-1}\Big[\mathbf{A}_6(\mathbf{y}_{1:N}^o;\Psi) - A_0(\mathbf{y}_{1:N}^o;\Psi)\delta_\epsilon^* - \mathbf{A}_{10}(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T} + \boldsymbol{\mu}\mathbf{W}\mathbf{A}_{13}(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T} + (\delta_\epsilon\mathbf{W} + \sigma^2\delta_x)\mathbf{A}_{11}(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}\Big] \\
\delta_\epsilon^* = A_2(\mathbf{y}_{1:N}^o;\Psi)^{-1}\Big[\mathbf{A}_5(\mathbf{y}_{1:N};\Psi) - A_0(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^* - \mathbf{A}_8(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T} + \boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T} + \Big(\delta_\epsilon\mathbf{W} - \sigma^2\delta_x\Big)\mathbf{A}_4(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}\Big] \\
\boldsymbol{\delta}_x^* = A_3(\mathbf{y}_{1:N}^o;\Psi)^{-1}\Big(\mathbf{A}_8(\mathbf{y}_{1:N};\Psi) - \boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}_{1:N}^o;\Psi) - (\delta_\epsilon\mathbf{W} - \sigma^2\delta_x)\mathbf{A}_4(\mathbf{y}_{1:N}^o;\Psi)\Big) \\
\mathbf{W}^* = \Big[\mathbf{A}_{14}(\mathbf{y}_{1:N}^o;\Psi) - \boldsymbol{\mu}^{*T}\mathbf{A}_{10}(\mathbf{y}_{1:N}^o;\Psi) - \delta_\epsilon^*\mathbf{A}_8(\mathbf{y}_{1:N}^o;\Psi) - \mathbf{A}_{16}(\mathbf{y}_{1:N}^o;\Psi) + \boldsymbol{\mu}^{*T}\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{13}(\mathbf{y}_{1:N}^o;\Psi) + \boldsymbol{\mu}^{*T}(\delta_\epsilon - \sigma^2\delta_x)\mathbf{A}_{11}(\mathbf{y}_{1:N}^o;\Psi) \\
\quad +\delta_\epsilon^{*T}\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}_{1:N}^o;\Psi) + \delta_\epsilon^{*T}(\delta_\epsilon\mathbf{W} - \sigma^2\delta_x)\mathbf{A}_4(\mathbf{y}_{1:N}^o;\Psi)\Big]\Big[\mathbf{A}_{17}(\mathbf{y}_{1:N}^o;\Psi) - \mathbf{A}_{18}(\mathbf{y}_{1:N}^o;\Psi) - \mathbf{A}_{18}(\mathbf{y}_{1:N}^o;\Psi)^T + \mathbf{A}_{19}(\mathbf{y}_{1:N}^o;\Psi) \\
\quad +\sigma^2\mathbf{A}_{11}(\mathbf{y}_{1:N}^o;\Psi)\Big]^{-1} \\
\sigma^{*2} = \frac{1}{dA_0(\mathbf{y}_{1:N}^o;\Psi)}\Big[A_{20}(\mathbf{y}_{1:N}^o;\Psi) - 2\mathbf{A}_6(\mathbf{y}_{1:N}^o;\Psi)\boldsymbol{\mu}^{*T} - 2\mathbf{A}_5(\mathbf{y}_{1:N}^o;\Psi)\delta_\epsilon^{*T} + A_1(\mathbf{y}_{1:N}^o;\Psi)\boldsymbol{\mu}^*\boldsymbol{\mu}^{*T} + 2A_0(\mathbf{y}_{1:N}^o;\Psi)\boldsymbol{\mu}^*\delta_\epsilon^{*T} \\
\quad +A_2(\mathbf{y}_{1:N}^o;\Psi)\delta_\epsilon^*\delta_\epsilon^{*T} - 2\,Tr\left\{\mathbf{A}_{14}(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}\right\} + 2\mathbf{A}_9(\mathbf{y}_{1:N}^o;\Psi,\Psi^*)\mathbf{W}^T\boldsymbol{\mu}^T \\
\quad +2\mathbf{A}_7(\mathbf{y}_{1:N}^o;\Psi,\Psi)(\delta_\epsilon\mathbf{W} - \sigma^2\delta_x)^T + 2\Big(\mathbf{A}_{10}(\mathbf{y}_{1:N}^o;\Psi) - \boldsymbol{\mu}\mathbf{W}\mathbf{A}_{13}(\mathbf{y}_{1:N}^o;\Psi) - (\delta_\epsilon\mathbf{W} - \sigma^2\delta_x)\mathbf{A}_{11}(\mathbf{y}_{1:N}^o;\Psi)\Big)\mathbf{W}^{*T}\boldsymbol{\mu}^{*T} \\
\quad +2\Big(\mathbf{A}_8(\mathbf{y}_{1:N}^o;\Psi) - \boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}_{1:N}^o;\Psi) - (\delta_\epsilon\mathbf{W} - \sigma^2\delta_x)\mathbf{A}_4(\mathbf{y}_{1:N}^o;\Psi)\Big)\mathbf{W}^{*T}\delta_\epsilon^{*T} + Tr\left\{\mathbf{A}_{17}(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}\mathbf{W}^*\right\} \\
\quad - Tr\left\{\Big(\mathbf{A}_{18}(\mathbf{y}_{1:N}^o;\Psi) + \mathbf{A}_{18}(\mathbf{y}_{1:N}^o;\Psi)^T\Big)\mathbf{W}^{*T}\mathbf{W}^*\right\} + Tr\left\{\mathbf{A}_{19}(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}\mathbf{W}^*\right\} + \sigma^2\,Tr\left\{\mathbf{A}_{11}(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}\mathbf{W}^*\right\}\Big].
\end{cases}
$$

The 2-dimensional integrals $\mathbf{A}_i$ *on the hypercube* $[0,1]^2$*, for* $i \in \Big\{0,\ldots,20\Big\}$*, are defined in Subsubsection A.4.4.4 in AppendixA.4.*

Proof. The proof of Corollary 26 is in Subsubsection A.4.4.4 in AppendixA.4. □

Corollary 27. *Under the assumptions of Special Case 3: t-Student IID PPCA specified in Subsubsection 3.4.1.3, the E-step of the corresponding EM algorithm in incomplete data setting as a case-specific variant of Theorem 27, is given by*

$$
\begin{aligned}
\tilde{Q}_{GStS,id,3}(\Psi,\Psi^*) &= \mathbb{E}_{\mathbf{Y}_{1:N}^m,\mathbf{X}_{1:N},U,V|\mathbf{Y}_{1:N}^o,\Psi}\left[\log\left(\pi_{\mathbf{Y}_{1:N},\mathbf{X}_{1:N},U,V|\Psi^*}\left(\mathbf{Y}_{1:N},\mathbf{X}_{1:N},U,V\right)\right)\right] \\
&= \frac{1}{\pi_{\mathbf{Y}_{1:N}^o|\Psi}(\mathbf{y}_{1:N}^o)}\sum_{t=1}^{N}\Big\{I_1(\mathbf{y}_t^o;\Psi,\Psi^*)\prod_{s=1,s\neq t}^{N}I_2(\mathbf{y}_s^o;\Psi)\Big\},
\end{aligned}
$$

defined by the functions $I_1,\ I_2 : \mathbb{R}^{d_o\times N} \to \mathbb{R}$ *from Theorem 27 given the case-specific formulation of the function* $m : \mathbb{R}^{d_o\times N}\times[0,1]^2 \longrightarrow \mathbb{R}$ *given by*

$$
m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi) = \left(\prod_{t=1}^{N}\pi_{\mathbf{Y}_t^o|S_\epsilon,S_x,\Psi}(\mathbf{y}_t^o)\right)(\sigma^2)^{\frac{kN}{2}}T_\epsilon(s_{\epsilon,t})^{\frac{dN}{2}}T_x(s_{x,t})^{\frac{kN}{2}}\Big|\mathbf{M}\Big|^{-\frac{1}{2}}\Big|\mathbf{N}\Big|^{-\frac{1}{2}},
$$

for the probability function $\pi_{\mathbf{Y}_t^o|S_\epsilon,S_x,\Psi}(\mathbf{y}_t^o)$ *of the* d_o*-dimensional random vector* $\mathbf{Y}_t^o|S_\epsilon,S_x,\Psi \sim \mathcal{N}\left(\boldsymbol{\mu}^o,\boldsymbol{\Sigma}_y^{oo}\right)$*, where* $\boldsymbol{\mu}^o$ *contains elements of the vector* $\boldsymbol{\mu}$ *which correspond to the observed entries of the vector* $\mathbf{Y}_t$ *and* $\boldsymbol{\Sigma}_y^{oo}$ *is a submatrix of* $\boldsymbol{\Sigma}_y = \sigma^2\mathbf{N}^{-1}$ *which correspond by rows and columns to observed elements of* $\mathbf{Y}_t$*. The and matrices* $\mathbf{M}$ *and* $\mathbf{N}$ *are defined as in Corollary 20. The case-specific formulation of the function* $\tilde{v} : \mathbb{R}^{d_o}\times[0,1]^2 \longrightarrow \mathbb{R}$ *is given by*

$$
\begin{aligned}
\tilde{v}(\mathbf{y}_t^o,s_\epsilon,s_x;\Psi,\Psi^*) &= -\frac{k+d}{2}\log 2\pi - \frac{d}{2}\log\sigma^{*2} + \frac{d}{2}T_\epsilon(s_\epsilon) + \frac{k}{2}T_x(s_x) - \frac{1}{2\sigma^{*2}}T_\epsilon(s_\epsilon)\,Tr\left\{\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_\epsilon,S_x,\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right]\right\} \\
&+ \frac{1}{\sigma^{*2}}T_\epsilon(s_\epsilon)\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_\epsilon,S_x,\Psi}\left[\mathbf{Y}_t\right]\boldsymbol{\mu}^{*T} - \frac{1}{2\sigma^{*2}}T_\epsilon(s_\epsilon)\boldsymbol{\mu}^*\boldsymbol{\mu}^{*T} + \frac{1}{\sigma^{*2}}T_\epsilon(s_\epsilon)^2\,Tr\left\{\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_\epsilon,S_x,\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right]\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\right\} \\
&- \frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})^2\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_\epsilon,S_x,\Psi}\left[\mathbf{Y}_t\right]\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\boldsymbol{\mu}^{*T} - \frac{1}{\sigma^{*2}}T_\epsilon(s_\epsilon)^2\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_\epsilon,S_x,\Psi}\left[\mathbf{Y}_t\right]\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\boldsymbol{\mu}^T
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})^2\boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\boldsymbol{\mu}^{*T}-\frac{1}{2}T_\epsilon(s_\epsilon)^2\,Tr\left\{\mathbf{M}^{-1}\mathbf{W}^T\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_\epsilon,S_x,\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right]\mathbf{W}\mathbf{M}^{-1}\Big(\frac{1}{\sigma^{*2}}T_\epsilon(s_\epsilon)\mathbf{W}^{*T}\mathbf{W}^*+T_x(s_x)\mathbb{I}_k\Big)\right\}\\
&+\frac{1}{2}T_\epsilon(s_\epsilon)^2\,Tr\left\{\mathbf{M}^{-1}\mathbf{W}^T\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_\epsilon,S_x,\Psi}\left[\mathbf{Y}_t\right]^T\boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}\Big(\frac{1}{\sigma^{*2}}T_\epsilon(s_\epsilon)\mathbf{W}^{*T}\mathbf{W}^*+T_x(s_x)\mathbb{I}_k\Big)\right\}\\
&+\frac{1}{2}T_\epsilon(s_\epsilon)^2\,Tr\left\{\mathbf{M}^{-1}\mathbf{W}\boldsymbol{\mu}^T\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_\epsilon,S_x,\Psi}\left[\mathbf{Y}_t\right]\mathbf{W}^T\mathbf{M}^{-1}\Big(\frac{1}{\sigma^{*2}}T_\epsilon(s_\epsilon)\mathbf{W}^{*T}\mathbf{W}^*+T_x(s_x)\mathbb{I}_k\Big)\right\}\\
&-\frac{1}{2}T_\epsilon(s_\epsilon)^2\,Tr\left\{\mathbf{M}^{-1}\mathbf{W}^T\boldsymbol{\mu}^T\boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}\Big(\frac{1}{\sigma^{*2}}T_\epsilon(s_\epsilon)\mathbf{W}^{*T}\mathbf{W}^*+T_x(s_x)\mathbb{I}_k\Big)\right\}-\frac{\sigma^2}{2}\,Tr\left\{\mathbf{M}_t^{-1}\Big(\frac{1}{\sigma^{*2}}T_\epsilon(s_\epsilon)\mathbf{W}^{*T}\mathbf{W}^*+T_x(s_x)\mathbb{I}_k\Big)\right\},
\end{aligned}
$$

for the corresponding non-central moments of $\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_\epsilon,S_x,\Psi$ *specified as*

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_\epsilon,S_x,\Psi}\left[\mathbf{Y}_t\right]=\left[\mathbf{Y}_t^o,\,\boldsymbol{\mu}^m+\left(\mathbf{Y}_t^o-\boldsymbol{\mu}^o\right){\boldsymbol{\Sigma}_y^{oo}}^{-1}\boldsymbol{\Sigma}_y^{om}\right],\\
&\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_\epsilon,S_x,\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right]=\begin{bmatrix}\mathbf{0}&\mathbf{0}\\ \mathbf{0}&\boldsymbol{\Sigma}_y^{mm}-\boldsymbol{\Sigma}_y^{mo}{\boldsymbol{\Sigma}_y^{oo}}^{-1}\boldsymbol{\Sigma}_y^{om}\end{bmatrix}+\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_\epsilon,S_x,\Psi}\left[\mathbf{Y}_t\right]^T\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_\epsilon,S_x,\Psi}\left[\mathbf{Y}_t\right],
\end{aligned}
$$

where $\boldsymbol{\mu}^m$ *contains elements of the vector* $\boldsymbol{\mu}$ *which correspond to the non-observed entries of the vector* $\mathbf{Y}_t$ *and* $\boldsymbol{\Sigma}_y^{mm}$ *is a square submatrix of* $\boldsymbol{\Sigma}_y$ *which correspond by rows and columns to missing elements of* $\mathbf{Y}_t$*. The elements of the* $d_m\times d_o$ *submatrix* $\boldsymbol{\Sigma}_y^{mo}=\boldsymbol{\Sigma}_y^{om\,T}$ *correspond by rows to missing entries of the vector* $\mathbf{Y}_t$ *and by columns to its observed elements.*

Proof. The proof of Corollary 27 is analogous to the proof of Corollary 25 when one sets $\boldsymbol{\delta}_\epsilon=\boldsymbol{\delta}_x=\boldsymbol{\delta}_\epsilon^*=\boldsymbol{\delta}_x^*=\mathbf{0}$. □

Corollary 28. *Under the assumptions of Special Case 3: t-Student IID PPCA specified in Subsubsection 3.4.1.3, the M-step of the corresponding EM algorithm in incomplete data setting as a case-specific variant of Theorem 28 given the problem formulation* $\nabla_{\Psi^*}\tilde{Q}_{GStS,id,3}=\mathbf{0}$ *for* $\Psi^*=\left[\mathbf{W}^*,\boldsymbol{\mu}^*,\sigma^{*2}\right]$ *is defined by* 2*-dimensional integration problems on the hypercube* $[0,1]^2$ *as following*

$$
\begin{cases}
\boldsymbol{\mu}^*=A_1(\mathbf{y}_{1:N}^o;\Psi)^{-1}\Big[\mathbf{A}_6(\mathbf{y}_{1:N}^o;\Psi)-\mathbf{A}_{10}(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}+\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{13}(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}\Big]\\
\mathbf{W}^*=\Big[\mathbf{A}_{14}(\mathbf{y}_{1:N}^o;\Psi)-\boldsymbol{\mu}^{*T}\mathbf{A}_{10}(\mathbf{y}_{1:N}^o;\Psi)-\mathbf{A}_{16}(\mathbf{y}_{1:N}^o;\Psi)+\boldsymbol{\mu}^{*T}\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{13}(\mathbf{y}_{1:N}^o;\Psi)\Big]\\
\quad\times\Big[\mathbf{A}_{17}(\mathbf{y}_{1:N}^o;\Psi)-\mathbf{A}_{18}(\mathbf{y}_{1:N}^o;\Psi)-\mathbf{A}_{18}(\mathbf{y}_{1:N}^o;\Psi)^T+\mathbf{A}_{19}(\mathbf{y}_{1:N}^o;\Psi)+\sigma^2\mathbf{A}_{11}(\mathbf{y}_{1:N}^o;\Psi)\Big]^{-1}\\
\sigma^{*2}=\frac{1}{dA_0(\mathbf{y}_{1:N}^o;\Psi)}\Big[A_{20}(\mathbf{y}_{1:N}^o;\Psi)-2\mathbf{A}_6(\mathbf{y}_{1:N}^o;\Psi)\boldsymbol{\mu}^{*T}+A_1(\mathbf{y}_{1:N}^o;\Psi)\boldsymbol{\mu}^*\boldsymbol{\mu}^{*T}+2A_0(\mathbf{y}_{1:N}^o;\Psi)\boldsymbol{\mu}^*\boldsymbol{\delta}_\epsilon^{*T}\\
\quad+A_2(\mathbf{y}_{1:N}^o;\Psi)\boldsymbol{\delta}_\epsilon^*\boldsymbol{\delta}_\epsilon^{*T}-2\,Tr\left\{\mathbf{A}_{14}(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}\right\}+2\mathbf{A}_9(\mathbf{y}_{1:N}^o;\Psi,\Psi^*)\mathbf{W}^T\boldsymbol{\mu}^T\\
\quad+2\Big(\mathbf{A}_{10}(\mathbf{y}_{1:N}^o;\Psi)-\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{13}(\mathbf{y}_{1:N}^o;\Psi)\Big)\mathbf{W}^{*T}\boldsymbol{\mu}^{*T}+Tr\left\{\mathbf{A}_{17}(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}\mathbf{W}^*\right\}\\
\quad-Tr\left\{\Big(\mathbf{A}_{18}(\mathbf{y}_{1:N}^o;\Psi)+\mathbf{A}_{18}(\mathbf{y}_{1:N}^o;\Psi)^T\Big)\mathbf{W}^{*T}\mathbf{W}^*\right\}+Tr\left\{\mathbf{A}_{19}(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}\mathbf{W}^*\right\}+\sigma^2\,Tr\left\{\mathbf{A}_{11}(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}\mathbf{W}^*\right\}\Big].
\end{cases}
$$

The 2*-dimensional integrals* $\mathbf{A}_i$ *on the hypercube* $[0,1]^2$*, for* $i\in\Big\{0,\ldots,19\Big\}$*, are defined in Subsubsection A.4.4.4 in Appendix A.4 given the following simplifications*

$$
\begin{aligned}
\mathbf{A}_{16}(\mathbf{y}_{1:N}^o;\Psi,\Psi^*)_{d\times k}&:=\int_0^1\int_0^1 T_\epsilon(s_\epsilon)^2\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)^T\boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}\,m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x,\\
\mathbf{A}_{18}(\mathbf{y}_{1:N}^o;\Psi,\Psi^*)_{k\times d}&:=\int_0^1\int_0^1 T_\epsilon(s_\epsilon)^3\mathbf{M}^{-1}\mathbf{W}^T\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)^T\boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}\,m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x,\\
\mathbf{A}_{19}(\mathbf{y}_{1:N}^o;\Psi,\Psi^*)_{k\times d}&:=\int_0^1\int_0^1 T_\epsilon(s_\epsilon)^3\mathbf{M}^{-1}\mathbf{W}^T\boldsymbol{\mu}^T\boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}\,m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x.
\end{aligned}
$$

Proof. The proof of Corollary 28 is analogous to the proof of Corollary 25 when one sets $\boldsymbol{\delta}_\epsilon=\boldsymbol{\delta}_x=\boldsymbol{\delta}_\epsilon^*=\boldsymbol{\delta}_x^*=\mathbf{0}$. □

Part II

State Space Models

Summary

The remainder of this thesis discusses a modelling framework that allows one to use the information, extracted by the dimensionality reduction methods, derived in Part I, in various statistical inference problems. We restrict our discussion to considering only discrete-time processes, as these reflect the modelling context that was undertaken in our research.

We start by introducing the modelling framework proposed in this study and show various structural approaches one can adopt to incorporate exogenous variables and investigate their influence. The applicability of the modelling framework is further examined in the real-data studies summarised in Part III. We advice how to evaluate the relevance of studied models by appropriate model selection techniques in Chapter 9.

The reminder of Part II revisits several methodologies which apply to the inference of models that are of relevance to future extensions to model formulation and estimation methods developed in this thesis. We explore the estimation techniques for the studied class of models given two paradigms of statistical inference: the Bayesian and frequentist approach.

We begin by revisiting the background on the class of state space models which facilitate the estimation of hidden dynamics. Chapter 6 introduces the concept of a general state space representation and briefly reviews key ideas behind its optimal estimation approach. It is followed by the discussion on the optimal filtering algorithm for linear Gaussian state space models in Chapter 7, known as the Kalman Filter. We explain the derivation of the filter, its most relevant properties as well as its application to the model selection and prediction.

Chapter 8 further generalises the filtering framework to Bayesian formulations which addresses the uncertainty of high dimensional static parameter estimation within the state space modelling context. We introduce the notion of Bayesian inference and detail the steps of suitable estimation procedures for models considered in this thesis. Chapter 10 briefly summarises stochastic approximation methods that can be considered when the filtering formulations do not admit analytical expressions.

5. Hybrid Factor Model

We demonstrate a general approach one may adopt to incorporate additional observable factors with extracted features from arbitrary datasets in statistical modelling framework. The proposed class of models provides not only a good quality of calibrating an observation process but also informs about interactions between explanatory components or latent states of a model over time. The discussed methodology allows for both, efficient and optimal estimation of studied models as well as their interpretation.

We start by introducing and justifying the construction of the considered models in Section 5.1. It is followed by the discussion on the model selection techniques one may adopt to asses models' in-sample performance.

5.1. Model Assumptions

We begin our discussion by briefly revisiting the concept of linear dynamic systems with innovations that follows a Gaussian distribution, which we adopt in our studies. The more insightful description of state-space models and review of their inference frameworks is provided in the subsequent chapters of this part. The class of models, which is relevant to the problems studied in this thesis, is formulated by the following dynamic system of equations

$$\begin{aligned} \mathbf{y}_t &= \boldsymbol{\mu} + \boldsymbol{\beta}_t \boldsymbol{\Lambda} + \boldsymbol{\epsilon}_t \\ \boldsymbol{\beta}_t &= \mathbf{c} + \boldsymbol{\beta}_{t-1} \mathbf{A} + \boldsymbol{\eta}_t \end{aligned} \tag{5.1}$$

for the $k \times d$ and $k \times k$ transition matrices $\boldsymbol{\Lambda}$ and $\mathbf{A}$. The error terms $\boldsymbol{\epsilon}_t$ and $\boldsymbol{\eta}_t$ are assumed to have specified first and second moments, that is

$$\begin{aligned} &\mathbb{E}\left[\boldsymbol{\epsilon}_t\right]_{1\times d} = \mathbf{0},\ \mathbb{E}\left[\boldsymbol{\epsilon}_t^T \boldsymbol{\epsilon}_t\right]_{d\times d} = \mathbf{Q} \text{ and } \mathbb{E}\left[\boldsymbol{\eta}_t\right]_{1\times k} = \mathbf{0},\ \mathbb{E}\left[\boldsymbol{\eta}_t^T \boldsymbol{\eta}_t\right]_{k\times k} = \mathbf{R}, \\ &\mathbb{E}\left[\boldsymbol{\epsilon}_t^T \boldsymbol{\eta}_t\right]_{d\times k} = \mathbf{0},\ \mathbb{E}\left[\boldsymbol{\epsilon}_t^T \boldsymbol{\epsilon}_{t-s}\right]_{d\times d} = \mathbf{0} \text{ and } \mathbb{E}\left[\boldsymbol{\eta}_t^T \boldsymbol{\eta}_{t-s}\right]_{k\times k} = \mathbf{0}, \end{aligned}$$

for $s = \{1, \ldots, t\}$, the $d \times d$ and $k \times k$ covariance matrices $\mathbf{Q}$ and $\mathbf{R}$. The matrices $\boldsymbol{\Lambda}$, $\mathbf{A}$, $\mathbf{Q}$, $\mathbf{R}$ and vectors $\boldsymbol{\mu}$ and $\mathbf{c}$ are called the static parameters of the model in Equation (5.1) and the generic collection of all such static model parameters will be denoted by $\Psi = \left[\boldsymbol{\Lambda},\ \mathbf{A},\ \mathbf{Q}, \mathbf{R}, \boldsymbol{\mu},\ \mathbf{c}\right]$. We assume $\boldsymbol{\beta}_t$ to be first- order Markov what can be easily extended.

In the contexts of the problems studied in this work, we add the features, extracted from various data sets, to the linear dynamic model in a static form but with dynamic coefficient on factor loadings, the latent state processes. Thus we can assess the impact of the factors on a response variable over time. This approach has this advantage that we do not have to model explicitly the analysed datasets which may have a complex structure.

Let us define $\mathbf{F}_t$ as a $d \times m$ dimensional matrix with various features. We can obtain $\mathbf{F}_t$ as a matrix with leading eigenvectors from datasets which are assumed to be an exogenous observable

input that is believed to have a potential influence on the studied dynamics. Each feature vector regressor, extracted from the exogenous data, we add to the state-space model in Equation (5.1).

There are numerous structural ways to achieve this in a state-space model. For instance, the factors may either influence the whole observation vector by entering the factor into the state equation, or the individual components of $\mathbf{Y}_t$ differently by adding them to the observation equation at different tenors on the term structure. We can also try a combination of such approaches, dependent on which feature was extracted from the context of the model construction.

We assume that the influence of features vary over time. Therefore we specify a time dynamic for the processes which are coefficient factor loadings,a new latent variable, a $d \cdot m$ dimensional vector $\boldsymbol{\varrho}_t$ which is modelled by VAR(1) process given by

$$\boldsymbol{\varrho}_t = \boldsymbol{\Psi} + \boldsymbol{\Omega}\boldsymbol{\varrho}_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \overset{iid}{\sim} \mathcal{N}(0, \mathbf{H}_{dm\times dm}).$$

The model assumption can be easily changed but for the convenience of the notation in this thesis, we keep the vector autoregressive structure with a single lag. The random vector $\boldsymbol{\varrho}_t$ is a dynamic regression parameter which specifies the impact of an i-th features corresponding to the i-th column of $[\mathbf{F}_t]$. Depending on the interpretation of the desired model, one may incorporate $\mathbf{F}_t$ into observation, the top equation of Equation (5.1) (**Case 1**), the latent dynamic of $\boldsymbol{\beta}_t$ (**Case 2**), that is the bottom equation of Equation (5.1) or both (**Case 3**). The new class of hybrid models, which combines the standard state-space model with exogenous features, named Hybrid Factor Model (HFM) class, is defined as follows

$$\mathbf{Y}_t = \tilde{\boldsymbol{\Lambda}}_t\tilde{\boldsymbol{\beta}}_t + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{Q}), \tag{5.2a}$$

$$\tilde{\boldsymbol{\beta}}_t = \tilde{\mathbf{c}} + \tilde{\mathbf{A}}_t\tilde{\boldsymbol{\beta}}_{t-1} + \tilde{\boldsymbol{\eta}}_t, \quad \tilde{\boldsymbol{\eta}}_t \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \tilde{\mathbf{R}}) \tag{5.2b}$$

where $\tilde{\boldsymbol{\beta}}_t = [\boldsymbol{\beta}_t, \boldsymbol{\varrho}_t]$ is a $(k + dm)$-dimensional latent process vector and

$$\tilde{\mathbf{c}} = \begin{pmatrix} \mathbf{c}_{k\times 1} \\ \boldsymbol{\Psi}_{dm\times 1} \end{pmatrix}_{(k+dm)\times 1}$$

is an interception term of the state equation. The covariance matrix of the error terms of state equation, $\tilde{\boldsymbol{\eta}}_t$, is given by

$$\tilde{\mathbf{R}} = \left(\begin{array}{c|c} \mathbf{R} & \mathbf{Y} \\ \hline \mathbf{Y} & \mathbf{H} \end{array}\right)_{(k+dm)\times(k+dm)}$$

where $\mathbf{Y}$ defines the correlation structure between latent processes $\boldsymbol{\beta}_t$ and $\boldsymbol{\varrho}_t$. Let us specify the following two objects, $\tilde{\mathbf{F}}_t = \bigoplus_{j=1}^{k}[\mathbf{F}_t]_{j,\cdot}$, for $\bigoplus$ being a direct sum operator, and $\tilde{\mathbf{f}}_t = vec\left(\mathbf{F}_t^T\right)$, that is

$$\tilde{\mathbf{F}}_t = \begin{pmatrix} [\mathbf{F}_t]_{1,\cdot} & 0 & 0 & \cdots & 0 \\ 0 & [\mathbf{F}_t]_{2,\cdot} & 0 & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & \cdots & & & [\mathbf{F}_t]_{p,\cdot} \end{pmatrix}_{dm\times dm} \quad \text{and } \tilde{\mathbf{f}}_t = \begin{pmatrix} [\mathbf{F}_t]_{1,1} \\ [\mathbf{F}_t]_{1,2} \\ \vdots \\ [\mathbf{F}_t]_{d,m} \end{pmatrix}_{dm\times 1}$$

where $[\mathbf{F}_t]_{j,\cdot}$ and $[\mathbf{F}_t]_{j,i}$ represent the vector of the j-th row of the matrix $\mathbf{F}_t$ and the element corresponding to jth row and ith column, respectively. The case-specific formulations of the transition matrices of the observation and state equations of the extended model in Equation (5.2), is the following

$$\tilde{\mathbf{\Lambda}}_{t\ d\times(k+dm)} = \begin{cases} \left(\begin{array}{c|c} \mathbf{\Lambda}_{d\times k} & \tilde{\mathbf{F}}_t \end{array}\right) & \text{for } \textbf{Case 1}, \\ \left(\begin{array}{c|c} \mathbf{\Lambda}_{d\times k} & \mathbf{0}_{d\times dm} \end{array}\right) & \text{otherwise}, \end{cases}$$

$$\tilde{\mathbf{A}}_{t,\ (k+dm)\times(k+dm)} = \begin{cases} \left(\begin{array}{c|c} \mathbf{A}_{k\times k} & \mathbf{0}_{k\times dm} \\ \hline \mathbf{0}_{dm\times k} & \mathbf{\Omega}_{dm\times dm} \end{array}\right) & \text{for } \textbf{Case 1}, \\ \left(\begin{array}{c|c} & \tilde{\mathbf{f}}_t^T \\ \mathbf{A}_{k\times k} & \tilde{\mathbf{f}}_t^T \\ & \tilde{\mathbf{f}}_t^T \\ \hline \mathbf{0}_{dm\times k} & \mathbf{\Omega}_{dm\times dm} \end{array}\right) & \text{for } \textbf{Case 2}. \end{cases}$$

The filtering of the state variables and the estimation of the static parameters of the models can be carried out by the procedures discussed in Chapter 7 or Chapter 8 – depending on the inference setting.

6. Optimal Filtering Problem

A general state space model is composed of two equations. The first equation, called the measurement equation, links a d-dimensional observable random variable, denoted by $\mathbf{Y}_t$, to a k-dimensional latent state (unobserved) random vector, denoted by $\boldsymbol{\beta}_t$. The relation between observation and state variables is assumed to be not directly observed but instead it is observed in the presence of observation noise - typically additive in nature, such that the observation equation then admits an additive random vector for the observation error, denoted by $\boldsymbol{\epsilon}_t$. The second equation specifies a Markov dynamic of the latent process $\boldsymbol{\beta}_t$ as a function of its past realisations. Again, we assume the k-dimensional error term which stochastically drives the state process, denoted by $\boldsymbol{\eta}_t$. The formulation of state space models is the following

$$\begin{aligned} \mathbf{Y}_t &= f(\boldsymbol{\beta}_t, \boldsymbol{\epsilon}_t) \\ \boldsymbol{\beta}_t &= g(\boldsymbol{\beta}_{0:t-1}, \boldsymbol{\eta}_t) \end{aligned} \tag{6.1}$$

for the functions $f : \mathbb{R}^k \times \mathbb{R}^d \to \mathbb{R}^d$ and $g : \mathbb{R}^{t\times k} \times \mathbb{R}^k \to \mathbb{R}^k$ which specify the iterations between $\mathbf{Y}_t$ and $\boldsymbol{\beta}_t$ and $\boldsymbol{\beta}_t$ and its past realisations, respectively. The class of state spaces models is defined by various assumptions on the function f and g and the distribution and the dependency structures between $\boldsymbol{\epsilon}_t$ and $\boldsymbol{\eta}_t$ and their components. Let us remark that it is not limited to time invariant static parameters of the model, however, such extensions are not studied in this thesis.

6.1. Optimal Filtering Problem Statement

An insightful introduction to the optimal filtering problem can be found in (Jazwinski 1970) (Anderson and Moore 1979) or (Kallianpur 1980) from a stochastic processes perspective, among other relevant references. In the following section we briefly revisit the concept of the optimal estimation of the unobserved states $\boldsymbol{\beta}_t$ in the general model in Equation (6.1).

In the famous work of (Kalman 1960), the author defines the problem of filtering and the class of measures which determine the optimality of its solution. The general filtering problem specifies an estimate of the unobserved process $\boldsymbol{\beta}_t$ given the information available up to time point t, which we denote by $\boldsymbol{\beta}_{t|t}$. In the remainder of this part, we assume that the available information is represented fully by the realisations of the vector $\mathbf{Y}_t$, the set of random observation vectors, denoted by $\mathbf{Y}_{1:t} = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_t\}$. In addition, we assume that there exists a distribution of $\boldsymbol{\beta}_t$ which is conditional on $\mathbf{Y}_{1:t}$ characterized by the probability density function $\pi_{\boldsymbol{\beta}_t|\mathbf{Y}_{1:t},\Psi}(\boldsymbol{\beta}_t)$ where Ψ is a vector with all static parameters of the model in Equation (6.1).

The optimality of a solution to the filtering problem is assessed according to some measure of accuracy of the estimates $\boldsymbol{\beta}_{t|t}$. Let us denote $\mathbf{e}_t := \boldsymbol{\beta}_t - \boldsymbol{\beta}_{t|t}$ as an estimation error. Following (Kalman 1960), the quality of estimation is measured by a loss function which is a non-decreasing and symmetric function of the estimation error, denoted by $\lambda : \mathbb{R}^k \to \mathbb{R}_+$, such that $\lambda(\mathbf{0}) = 0$.

One way to estimate $\boldsymbol{\beta}_t$ given $\mathbf{Y}_{1:t}$ is to minimize the expected loss of estimation. The specified estimator is then given by the expectation of $\boldsymbol{\beta}_t$ conditional on $\mathbf{Y}_{1:t}$ as proven in Theorem 29.

Theorem 29. *(Kalman 1960) If the probability density function $\pi_{\boldsymbol{\beta}_t|\mathbf{Y}_{1:t},\Psi}(\boldsymbol{\beta}_t)$ defined by the model in Equation (6.1) is symmetric about its mean and convex, the estimator of $\boldsymbol{\beta}_t$, which is a solution to the minimisation problem*

$$\boldsymbol{\beta}_{t|t} = \underset{\mathbf{b}\in\mathbb{R}^k}{\operatorname{argmin}}\, \mathbb{E}_{\boldsymbol{\beta}_t|\mathbf{Y}_{1:t},\Psi}\big[\lambda(\boldsymbol{\beta}_t - \mathbf{b})\big],$$

is given by

$$\boldsymbol{\beta}_{t|t} = \mathbb{E}_{\boldsymbol{\beta}_t|\mathbf{Y}_{1:t},\Psi}\big[\boldsymbol{\beta}_t\big].$$

Proof. Please refer to Theorem 1 in (Kalman 1960) for the proof of the results. □

The author of (Kalman 1960) shows that for the quadratic loss function, such that $\lambda(\mathbf{e}) = \mathbf{e}\mathbf{e}^T$, the formulation of the optimal estimator of $\boldsymbol{\beta}_t$ from Theorem 29 holds even when the conditions imposed on the probability density function $\pi_{\boldsymbol{\beta}_t|\mathbf{Y}_{1:t},\Psi}(\boldsymbol{\beta}_t)$ are not satisfied, as summarised in Theorem 30.

Theorem 30. *(Kalman 1960) Let $\pi_{\boldsymbol{\beta}_t|\mathbf{Y}_{1:t},\Psi}(\boldsymbol{\beta}_t)$ be any probability density function defined by the model in Equation (6.1). If the loss function λ is quadratic such that $\lambda(\mathbf{e}) = \mathbf{e}\mathbf{e}^T$, the estimator of $\boldsymbol{\beta}_t$ which is a solution to the minimisation problem*

$$\boldsymbol{\beta}_{t|t} = \underset{\mathbf{b}\in\mathbb{R}^k}{\operatorname{argmin}}\, \mathbb{E}_{\boldsymbol{\beta}_t|\mathbf{Y}_{1:t},\Psi}\big[(\boldsymbol{\beta}_t - \mathbf{b})(\boldsymbol{\beta}_t - \mathbf{b})^T\big],$$

is given by

$$\boldsymbol{\beta}_{t|t} = \mathbb{E}_{\boldsymbol{\beta}_t|\mathbf{Y}_{1:t},\Psi}\big[\boldsymbol{\beta}_t\big].$$

Proof. Please refer to Theorem 1a in (Kalman 1960) for the proof of the results. □

In the reminder of Part II, we seek an estimator of the state vectors $\boldsymbol{\beta}_t$ conditional on available information $\mathbf{Y}_{1:t}$, which is optimal in a sense that it minimizes the variance of the estimation error $\mathbf{e}_t$. Under these assumptions, the optimal filtering of the states $\boldsymbol{\beta}_t$ requires the knowledge about the conditional distribution of the model in Equation (6.1). In majority of cases this distribution is not feasible to specify in closed form, due to intractability of the normalizing constant. Due to this fact, it is common to reformulate the distribution of $\boldsymbol{\beta}_t|\mathbf{Y}_{1:t},\Psi$ by the application of the Bayes' rule as follows

$$\pi_{\boldsymbol{\beta}_t|\mathbf{Y}_{1:t},\Psi}(\boldsymbol{\beta}_t) = \pi_{\boldsymbol{\beta}_t|\mathbf{Y}_t,\mathbf{Y}_{1:(t-1)},\Psi}(\boldsymbol{\beta}_t) = \frac{\pi_{\boldsymbol{\beta}_t,\mathbf{Y}_t|\mathbf{Y}_{1:(t-1)},\Psi}(\boldsymbol{\beta}_t,\mathbf{y}_t)}{\pi_{\mathbf{Y}_t|\mathbf{Y}_{1:(t-1)},\Psi}(\mathbf{y}_t)} = \frac{\pi_{\mathbf{Y}_t|\boldsymbol{\beta}_t,\Psi}(\mathbf{y}_t)\ \pi_{\boldsymbol{\beta}_t|\mathbf{Y}_{1:(t-1)},\Psi}(\boldsymbol{\beta}_t)}{\int_{\mathbb{R}^k} \pi_{\mathbf{Y}_t|\boldsymbol{\beta}_t,\Psi}(\mathbf{y}_t)\ \pi_{\boldsymbol{\beta}_t|\mathbf{Y}_{1:(t-1)},\Psi}(\boldsymbol{\beta}_t)\, d\boldsymbol{\beta}_t} \tag{6.2}$$

where

$$\pi_{\boldsymbol{\beta}_t|\mathbf{Y}_{1:(t-1)},\Psi}(\boldsymbol{\beta}_t) = \int_{\mathbb{R}^k} \pi_{\boldsymbol{\beta}_t,\boldsymbol{\beta}_{t-1}|\mathbf{Y}_{1:(t-1)},\Psi}(\boldsymbol{\beta}_t,\boldsymbol{\beta}_{t-1})\, d\boldsymbol{\beta}_{t-1} = \int_{\mathbb{R}^k} \pi_{\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t-1},\Psi}(\boldsymbol{\beta}_t)\ \pi_{\boldsymbol{\beta}_{t-1}|\mathbf{Y}_{1:(t-1)},\Psi}(\boldsymbol{\beta}_{t-1})\, d\boldsymbol{\beta}_{t-1}. \tag{6.3}$$

In filtering problems, Equation (6.2) is referred as an update step of the filter in time t whereas Equation (6.3) is termed as a prediction step.

The $\pi_{\mathbf{Y}_t|\mathbf{Y}_{1:(t-1)},\Psi}(\mathbf{y}_t)$ is known as the evidence obtained from the measurements. The quantity is broadly used as a component of model selection procedures and hypothesis testing for different values of Ψ. The evidence serves as a summary of the model structure which is contained within an observation set.

We advise the reader to refer to (Kalman 1960), (Jazwinski 1970) or (Anderson and Moore 1979) for more detailed discussion on the filtering of the state-space models under various loss functions and their properties.

7. Kalman Filter and Gaussian Linear State Space Models

The linear state-space model in Equation (5.1) assumes $\boldsymbol{\beta}_t$ to be first- order Markov what can be easily extended and still provides explicit recursions of the filter, as shown in (Harvey 1990).

We assume that the initial state random vector follows $\boldsymbol{\beta}_0 \sim \mathcal{N}(\mathbf{b}_0, \mathbf{P}_0)$ for a real values k-dimensional vector $\mathbf{b}_0$ and a $k \times k$ positive-definite symmetric matrix $\mathbf{P}_0$.

The conditional distribution defined by the model in Equation (5.1) is Gaussian, and therefore, the optimal estimator of the state vector is given by the conditional mean $\mathbb{E}_{\boldsymbol{\beta}_t|\mathbf{Y}_{1:t},\Psi}[\boldsymbol{\beta}_t]$. The optimality is ensured for a wide range of loss functions as defined in the body of Theorem 29 and Chapter 6.1. The density of a Gaussian distribution is convex and symmetric around its mean, and hence the conditions of the theorem are satisfied. In the remainder of this chapter, we derive the optimal estimation framework for the class of models in Equation (5.1) and highlight its properties which are of most relevance to the research summarised in this thesis.

7.1. Recursions of Kalman Filter for Gaussian Linear State-Space Models

(Kalman 1960) and (Kalman and Bucy 1961) introduce the optimal filtering procedure, known as the Kalman Filter, which is derived for linear Gaussian state space models in Equation (5.1). The framework can be derived in many ways as shown (Harvey 1990), (Hamilton 1994) or (Tanizaki 1993). In the following section, we show only the steps of one derivation approach.

Recalling the assumptions about the model in Equation (5.1) and that the initial state of the latent vector is assumed be Gaussian, $\boldsymbol{\beta}_0 \sim \mathcal{N}(b_0, \mathbf{P}_0)$ with an initial mean $\mathbf{b}_0$ and an initial covariance matrix $\mathbf{P}_0$. Then, the joint distribution of $\boldsymbol{\beta}_0$ and $\mathbf{Y}_0$ is the following

$$\begin{bmatrix} \boldsymbol{\beta}_0 \\ \mathbf{Y}_0 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{b}_0 \\ \boldsymbol{\mu} + \mathbf{b}_0\boldsymbol{\Lambda} \end{bmatrix}, \begin{bmatrix} \mathbf{P}_0 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \right).$$

Applying a well known formulations of the conditional distribution of two Gaussian vectors, given in Lemma A1, the first and second central moments of the random vector $\boldsymbol{\beta}_0|\mathbf{Y}_0$ are given by

$$\begin{aligned} \boldsymbol{\beta}_{0|0} &= \mathbb{E}_{\boldsymbol{\beta}_0|\mathbf{Y}_0,\Psi}[\boldsymbol{\beta}_0] = \mathbf{b}_0 \\ \mathbf{P}_{0|0} &= \mathbb{E}_{\boldsymbol{\beta}_0|\mathbf{Y}_0,\Psi}\left[(\boldsymbol{\beta}_0 - \boldsymbol{\beta}_{0|0})^T (\boldsymbol{\beta}_0 - \boldsymbol{\beta}_{0|0}) \right] = \mathbf{P}_0. \end{aligned}$$

The predictive joint distribution of $\boldsymbol{\beta}_1$ and $\mathbf{Y}_1$ conditional on the recent measurement $\mathbf{Y}_0$ is also Gaussian, that is

$$\boldsymbol{\beta}_1,\ \mathbf{Y}_1|\mathbf{Y}_0 \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\beta}_{1|0} \\ \mathbf{y}_{1|0} \end{bmatrix}, \begin{bmatrix} \mathbf{P}_{1|0} & \mathbf{P}_{1|0}\boldsymbol{\Lambda} \\ \boldsymbol{\Lambda}^T\mathbf{P}_{1|0} & \boldsymbol{\Sigma}_{1|0} \end{bmatrix} \right),$$

with first and second moments specified as

$$\begin{aligned}
\boldsymbol{\beta}_{1|0} &= \mathbb{E}_{\boldsymbol{\beta}_1|\mathbf{Y}_0,\Psi}\left[\boldsymbol{\beta}_1\right] = \mathbb{E}_{\boldsymbol{\eta}_1,\boldsymbol{\beta}_0|\mathbf{Y}_0,\Psi}\left[\mathbf{c}+\boldsymbol{\beta}_0\mathbf{A}+\boldsymbol{\eta}_1\right] \\
&= \mathbf{c}+\boldsymbol{\beta}_{0|0}\mathbf{A}, \\
\mathbf{P}_{1|0} &= \mathbb{E}_{\boldsymbol{\beta}_1|\mathbf{Y}_0,\Psi}\left[\left(\boldsymbol{\beta}_1-\boldsymbol{\beta}_{1|0}\right)^T\left(\boldsymbol{\beta}_1-\boldsymbol{\beta}_{1|0}\right)\right] = \mathbb{E}_{\boldsymbol{\eta}_1,\boldsymbol{\beta}_0|\mathbf{Y}_0,\Psi}\left[\left(\boldsymbol{\beta}_0\mathbf{A}-\boldsymbol{\beta}_{0|0}\mathbf{A}+\boldsymbol{\eta}_1\right)^T\left(\boldsymbol{\beta}_0\mathbf{A}-\boldsymbol{\beta}_{0|0}\mathbf{A}+\boldsymbol{\eta}_1\right)\right] \\
&= \mathbf{A}^T\mathbf{P}_{0|0}\mathbf{A}+\mathbf{R},
\end{aligned}$$

and

$$\begin{aligned}
&\mathbf{y}_{1|0} = \mathbb{E}_{\mathbf{Y}_1|\mathbf{Y}_0,\Psi}\left[\mathbf{Y}_1\right] = \mathbb{E}_{\boldsymbol{\epsilon}_1,\boldsymbol{\beta}_1|\mathbf{Y}_0,\Psi}\left[\boldsymbol{\mu}+\boldsymbol{\beta}_1\boldsymbol{\Lambda}+\boldsymbol{\epsilon}_1\right] = \boldsymbol{\mu}+\boldsymbol{\beta}_{1|0}\boldsymbol{\Lambda}, \\
&\boldsymbol{\Sigma}_{1|0} = \mathbb{E}_{\mathbf{Y}_1|\mathbf{Y}_0,\Psi}\left[\left(\mathbf{Y}_1-\mathbf{y}_{1|0}\right)^T\left(\mathbf{Y}_1-\mathbf{y}_{1|0}\right)\right] = \mathbb{E}_{\boldsymbol{\epsilon}_1,\boldsymbol{\beta}_1|\mathbf{Y}_0,\Psi}\left[\left(\boldsymbol{\beta}_1\boldsymbol{\Lambda}-\boldsymbol{\beta}_{1|0}\boldsymbol{\Lambda}+\boldsymbol{\epsilon}_1\right)^T\left(\boldsymbol{\beta}_1\boldsymbol{\Lambda}-\boldsymbol{\beta}_{1|0}\boldsymbol{\Lambda}+\boldsymbol{\epsilon}_1\right)\right] = \boldsymbol{\Lambda}^T\mathbf{P}_{1|0}\boldsymbol{\Lambda}+\mathbf{Q}.
\end{aligned}$$

Again, by applying Lemma A1, the expectation of $\boldsymbol{\beta}_1|\mathbf{Y}_1,\mathbf{Y}_0,\Psi$ can be formulated as follows

$$\begin{aligned}
\boldsymbol{\beta}_{1|1} &= \mathbb{E}_{\boldsymbol{\beta}_1|\mathbf{Y}_1,\mathbf{Y}_0,\Psi}\left[\boldsymbol{\beta}_1\right] = \boldsymbol{\beta}_{1|0}+\left(\mathbf{y}_1-\boldsymbol{\mu}-\mathbf{b}_{1|0}\boldsymbol{\Lambda}\right)\boldsymbol{\Sigma}_{1|0}^{-1}\boldsymbol{\Lambda}^T\mathbf{P}_{1|0}, \\
\mathbf{P}_{1|1} &= \mathbb{E}_{\boldsymbol{\beta}_1|\mathbf{Y}_1,\mathbf{Y}_0,\Psi}\left[\left(\boldsymbol{\beta}_1-\boldsymbol{\beta}_{1|0}\right)^T\left(\boldsymbol{\beta}_1-\boldsymbol{\beta}_{1|0}\right)\right] = \mathbf{P}_{1|0}+\mathbf{P}_{1|0}\boldsymbol{\Lambda}\boldsymbol{\Sigma}_{1|0}^{-1}\boldsymbol{\Lambda}^T\mathbf{P}_{1|0}.
\end{aligned}$$

Following the same reasoning as above, we can specify two recursive steps required to find the optimal estimators of $\boldsymbol{\beta}_t|\mathbf{Y}_{0:t},\Psi$. Recall that these steps are the expectations given is the the closed forms by the steps in Equation (6.2) and Equation (6.3) for Gaussian linear state space models. First we carry out the prediction step, which specifies the moments of the prediction $\boldsymbol{\beta}_t|\mathbf{Y}_{1:(t-1)},\Psi$. After the current realisation of $\mathbf{Y}_t$, the vector $\mathbf{y}_t$, arrives, we update the expectation of the random vector and determine the estimator of $\boldsymbol{\beta}_t|\mathbf{Y}_{0:t},\Psi$. The described procedure can be summarize by the conditional distributions involved in the multivariate Kalman Filter recursions as follows

$$\begin{aligned}
&\boldsymbol{\beta}_{t-1}|\mathbf{Y}_{0:t-1} \sim \mathcal{N}(\boldsymbol{\beta}_{t-1|t-1},\mathbf{P}_{t-1|t-1}), \\
&\boldsymbol{\beta}_t|\mathbf{Y}_{0:t-1} \sim \mathcal{N}(\boldsymbol{\beta}_{t|t-1},\mathbf{P}_{t|t-1}), \\
&\mathbf{Y}_t|\mathbf{Y}_{0:t-1} \sim \mathcal{N}(\mathbf{y}_{t|t-1},\boldsymbol{\Sigma}_{t|t-1}),
\end{aligned} \tag{7.1}$$

where

$$\begin{aligned}
\boldsymbol{\beta}_{t|t-1} &= \mathbf{c}+\boldsymbol{\beta}_{t-1|t-1}\mathbf{A}, \\
\mathbf{P}_{t|t-1} &= \mathbf{A}^T\mathbf{P}_{t-1|t-1}\mathbf{A}+\mathbf{R}, \\
\mathbf{y}_{t|t-1} &= \boldsymbol{\mu}+\boldsymbol{\beta}_{t|t-1}\boldsymbol{\Lambda} \\
\boldsymbol{\Sigma}_{t|t-1} &= \boldsymbol{\Lambda}^T\mathbf{P}_{t|t-1}\boldsymbol{\Lambda}+\mathbf{Q}, \\
\boldsymbol{\beta}_{t|t} &= \boldsymbol{\beta}_{t|t-1}+\left(\mathbf{y}_t-\boldsymbol{\mu}-\boldsymbol{\beta}_{t|t-1}\boldsymbol{\Lambda}\right)\boldsymbol{\Sigma}_{t|t-1}^{-1}\boldsymbol{\Lambda}^T\mathbf{P}_{t|t-1}, \\
\mathbf{K}_t &= \boldsymbol{\Sigma}_{t|t-1}^{-1}\boldsymbol{\Lambda}^T\mathbf{P}_{t|t-1}, \qquad (7.2) \\
\mathbf{P}_{t|t} &= \mathbf{P}_{t|t-1}-\mathbf{P}_{t|t-1}\boldsymbol{\Lambda}\mathbf{K}_t \qquad (7.3)
\end{aligned}$$

for $t = 1, \ldots, N$.

We can find many variants of the Kalman Filter in the existing literature each with the general objective to address arising computational difficulties of such filtering recursions in different modelling contexts, beyond the linear Gaussian state-space model setting. For instance, the information filter of (Fraser 1967), which yields a set of recursions for the inverse matrix $\boldsymbol{P}_{t|t}^{-1}$, is especially applicable to the situation when $\mathbf{P}_0$ is infinite.

Information Filter.

The Information Filter framework defines the vector $\mathbf{z}_{t|t-1} = \boldsymbol{\beta}_{t|t-1}\mathbf{P}_{t|t-1}^{-1}$ and the matrices $\mathbf{Z}_{t|t-1} = \mathbf{P}_{t|t-1}^{-1}$ and $\mathbf{Z}_{t|t} = \mathbf{P}_{t|t}^{-1}$ as the information state vector and the information state matrix. The core idea is to use the matrix inversion lemma to to obtained $\mathbf{P}_{t|t}^{-1}$. By incorporating these transformations into the classical Kalman Filter recursions, the formulation of the Information Filter are the following

$$\mathbf{Z}_{t|t} = \mathbf{P}_{t|t}^{-1} = \mathbf{Z}_{t|t-1} + \boldsymbol{\Lambda}^T\mathbf{Q}^{-1}\boldsymbol{\Lambda}, \qquad \mathbf{Z}_{t|t-1} = \mathbf{P}_{t|t-1}^{-1} = \Big(\mathbb{I}_k - \mathbf{B}_t\Big)\mathbf{A}_t,$$
$$\mathbf{A}_t = \big(\mathbf{A}^{-1}\big)^T\mathbf{Z}_{t|t}\mathbf{A}^{-1}, \qquad \mathbf{B}_t = \mathbf{A}_t\big(\mathbf{A}_t + \mathbf{R}^{-1}\big)^{-1},$$
$$\mathbf{z}_{t|t-1} = \mathbf{c}\mathbf{Z}_{t|t-1} + \mathbf{z}_{t|t-1}\mathbf{A}^{-1}\big(\mathbf{I}_k - \mathbf{B}_t\big), \qquad \mathbf{z}_{t|t} = \mathbf{z}_{t|t} + \big(\mathbf{y}_t - \boldsymbol{\mu}\big)\mathbf{Q}^{-1}\boldsymbol{\Lambda}.$$

Square Root Filter.

The second widely used formulation of the Kalman Filter is known as the square root filter. The formulations of the filter ensure that the covariance matrix $\boldsymbol{P}_{t+1|t}$ and $\boldsymbol{P}_{t|t}$ will always remain positive semi-definite by using their square roots in the recursions of the Kalman Filter. This change ensure the improvement to the stability of the Kalman filter in comparison to its conventional implementation .

The square root Kalman Filter exploits the square roots of the covariance matrices $\mathbf{P}_{t|t-1}$ and $\mathbf{P}_{t|t}$ in the formulations of the filtering recursions. Since square root of a matrix is not unique, one ma consider approaches to the factorisation of the covariance matrices such as the following Cholesky factorisation or the LDL factorisation. In addition, the square root matrices can be chosen to be upper or lower triangular and this flexibility is further exploited in various implementations of the filter in the literature. The review of concepts on the computational efficiency of Kalman Filter can be found in (Chapter 6, Anderson and Moore (1979)), (Sayed and Kailath 1994) or (Bar-Shalom et al. 2001) .

Let us remark that the standard recursions of Kalman Filter discussed in this section are useful in the situation when some elements of the state-space vector have variance equal to zero. Given the objectives of the information of the square root filter, it would not be feasible to carry on the recursion with block diagonal parts of the matrices $\boldsymbol{P}_{t|t-1}$ or $\boldsymbol{P}_{t|t}$ assumed to be zero. The standard formulations of the steps of Kalman Filter allow for such assumption since $\boldsymbol{\Sigma}_{t|t-1}$ will be still positive definite.

7.1.1. Irregular dimensionality of the observation vector

An appealing feature of the dynamic model in Equation (5.1) is its efficiency in handling missing data within the observation set, i.e. irregular time series. Let us suppose that the observation vector at time t, $\mathbf{Y}_t$, can be partitioned into two subvectors which correspond to its observed and unobserved entries, $\mathbf{Y}_t = [\mathbf{Y}_t^o, \mathbf{Y}_t^m]$ that are d_0 and d_m-dimensional for $d_o + d_m = d$. Such settings arise in numerous practical applications, such as those studied in this thesis. For instance, consider lack of information about a yield curve at some maturities for a given period of time due to illiquidity on the market. The general formulation of the state space models in Equation (6.1) sets the following distribution of $\mathbf{Y}_t^m$ conditioned on the random vector of observed entries $\mathbf{Y}_t^o$ and the past information $\mathbf{Y}_{1:(t-1)}$

$$\pi_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{Y}_{0:(t-1)},\Psi}\left(\mathbf{y}_t^m\right) = \int_{\mathbb{R}^k} \pi_{\mathbf{Y}_t^m,\boldsymbol{\beta}_t|\mathbf{Y}_t^o,\mathbf{Y}_{0:(t-1)},\Psi}\left(\mathbf{y}_t^m,\boldsymbol{\beta}_t\right) d\boldsymbol{\beta}_t = \int_{\mathbb{R}^k} \pi_{\mathbf{Y}_t^m|\boldsymbol{\beta}_t,\Psi}\left(\mathbf{y}_t^m\right) \pi_{\boldsymbol{\beta}_t|\mathbf{Y}_t^o,\mathbf{Y}_{0:(t-1)},\Psi}\left(\boldsymbol{\beta}_t\right) d\boldsymbol{\beta}_t \tag{7.4}$$

Given Theorem 30, the optimal estimation of the random vector $\mathbf{Y}_t^m$ given the observation $\mathbf{Y}_t^o$ and $\mathbf{Y}_{1:(t-1)}$, which minimizes the covariance of the estimation error, is the conditional expectation formulated as follows

$$\begin{aligned}\mathbf{y}_{t|t}^m &= \mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{Y}_{0:(t-1)},\Psi}\left[\mathbf{Y}_t^m\right] = \int_{\mathbb{R}^{d_m}} \mathbf{y}_t^m \; \pi_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{Y}_{0:(t-1)},\Psi}\left(\mathbf{y}_t^m\right) d\mathbf{y}_t^m \\ &= \int_{\mathbb{R}^{d_m}} \int_{\mathbb{R}^k} \mathbf{y}_t^m \pi_{\mathbf{Y}_t^m|\boldsymbol{\beta}_t,\Psi}\left(\mathbf{y}_t^m\right) \pi_{\boldsymbol{\beta}_t|\mathbf{Y}_t^o,\mathbf{Y}_{0:(t-1)},\Psi}\left(\boldsymbol{\beta}_t\right) d\boldsymbol{\beta}_t \, d\mathbf{y}_t^m = \mathbb{E}_{\boldsymbol{\beta}_t|\mathbf{Y}_t^o,\mathbf{Y}_{0:(t-1)},\Psi}\left[\mathbb{E}_{\mathbf{Y}_t^m|\boldsymbol{\beta}_t,\Psi}\left[\mathbf{Y}_t^m\right]\right]\end{aligned} \tag{7.5}$$

with the covariance of the estimation error specifies as

$$\boldsymbol{\Sigma}_{t|t}^m = \mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{Y}_{0:(t-1)},\Psi}\left[\left(\mathbf{Y}_t^m - \mathbf{y}_{t|t}^m\right)^T\left(\mathbf{Y}_t^m - \mathbf{y}_{t|t}^m\right)\right] = \mathbb{E}_{\boldsymbol{\beta}_t|\mathbf{Y}_t^o,\mathbf{Y}_{1:(t-1)},\Psi}\left[\mathbb{E}_{\mathbf{Y}_t^m|\boldsymbol{\beta}_t,\Psi}\left[\mathbf{Y}_t^{mT}\mathbf{Y}_t^m\right]\right] + \left(\mathbf{y}_{t|t}^m\right)^T\mathbf{y}_{t|t}^m \tag{7.6}$$

Since the observation vector $\mathbf{Y}_t$ is Gaussian, the moments in Equation (7.5) and Equation (7.6) are formulated explicitly. The measurement equation of the model in Equation (5.1) can be re-expressed in terms of the subvectors $\mathbf{Y}_t^o$ and $\mathbf{Y}_t^m$ as follows

$$\begin{pmatrix}\mathbf{Y}_t^o \\ \mathbf{Y}_t^m\end{pmatrix} = \begin{pmatrix}\boldsymbol{\mu}_o \\ \boldsymbol{\mu}_m\end{pmatrix} + \boldsymbol{\beta}_t \begin{pmatrix}\boldsymbol{\Lambda}_o \\ \boldsymbol{\Lambda}_m\end{pmatrix} + \begin{pmatrix}\boldsymbol{\epsilon}_t^o \\ \boldsymbol{\epsilon}_t^m\end{pmatrix}, \quad \begin{pmatrix}\boldsymbol{\epsilon}_t^o \\ \boldsymbol{\epsilon}_t^m\end{pmatrix} \overset{iid}{\sim} \mathcal{N}\left(\mathbf{0}, \begin{pmatrix}\mathbf{Q}_{oo} & \mathbf{Q}_{om} \\ \mathbf{Q}_{mo} & \mathbf{Q}_{mm}\end{pmatrix}\right)$$

where $\boldsymbol{\mu}_o$ and $\boldsymbol{\mu}_m$ are d_o- and d_m-dimensional subvectors of $\boldsymbol{\mu}$ with elements corresponding to observed and missing entries of the observation process $\mathbf{Y}_t$, $\boldsymbol{\Lambda}_o$ and $\boldsymbol{\Lambda}_m$ are $d_o \times (d+k)$ and $d_m \times k$ submatrices of $\boldsymbol{\Lambda}$ with rows corresponding to observed and missing entries of the observation process $\mathbf{Y}_t$, respectively, the square $d_o \times d_o$ and $d_m \times d_m$ matrices $\mathbf{Q}_{oo}$ and $\mathbf{Q}_{mm}$ are a submatrices of $\mathbf{Q}$ corresponding by rows and columns to observed and missing entries of the observation vector $\mathbf{y}_t$.

The Kalman Filter framework proceeds exactly as in the standard case, provided that $\mathbf{Y}_t$, $\mathbf{\Lambda}$ and $\mathbf{Q}$ are replaced by $\mathbf{Y}_t^o$, $\mathbf{\Lambda}^o$ and $\mathbf{Q}_{oo}$ respectively at relevant time points what does not affect the validity of the filtering recursion. When the realisation $\mathbf{y}_t^o$ of the random subvector $\mathbf{Y}_t^o$ is available, we obtain the following distribution of the random vector $\mathbf{Y}_t^m$

$$\begin{pmatrix} \mathbf{Y}_t^o \\ \mathbf{Y}_t^m \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{y}_t^o \\ \boldsymbol{\mu}_m + \boldsymbol{\beta}_t \mathbf{\Lambda}_m \end{bmatrix}, \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Lambda}_m^T \mathbf{P}_t \mathbf{\Lambda}_m + \mathbf{Q}_m m \end{bmatrix} \right)$$

The moments of $\mathbf{Y}_t^m$ conditional on the observed entries $\mathbf{Y}_t^o$ one obtains as

$$\begin{aligned}
\mathbf{y}_{t|t}^m &= \mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{Y}_{0:(t-1)},\Psi}\left[\mathbf{Y}_t^m\right] = \mathbb{E}_{\boldsymbol{\epsilon}_t,\boldsymbol{\beta}_t|\mathbf{Y}_t^o,\mathbf{Y}_{0:(t-1)},\Psi}\left[\boldsymbol{\mu}_m + \boldsymbol{\beta}_t \mathbf{\Lambda}_m + \boldsymbol{\epsilon}_t^m\right] = \boldsymbol{\mu}_m + \boldsymbol{\beta}_{t|t} \mathbf{\Lambda}_m, \\
\mathbf{\Sigma}_{t|t}^m &= \mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{Y}_{0:(t-1)},\Psi}\left[\left(\mathbf{Y}_t^m - \mathbf{y}_{t|t}^m\right)^T \left(\mathbf{Y}_t^m - \mathbf{y}_{t|t}^m\right)\right] = \mathbb{E}_{\boldsymbol{\epsilon}_t,\boldsymbol{\beta}_t|\mathbf{Y}_t^o,\mathbf{Y}_{0:(t-1)},\Psi}\left[\left(\boldsymbol{\beta}_t \mathbf{\Lambda}_m - \boldsymbol{\beta}_{t|t} \mathbf{\Lambda}_m + \boldsymbol{\epsilon}_t^m\right)^T \left(\boldsymbol{\beta}_t \mathbf{\Lambda}_m - \boldsymbol{\beta}_{t|t} \mathbf{\Lambda}_m + \right.\right. \\
&= \mathbf{\Lambda}_m^T \mathbf{P}_{t|t} \mathbf{\Lambda}_m + \mathbf{Q}_{mm}.
\end{aligned}$$

The missing entries of the observation vector at time t can be estimated without a bias by $\mathbf{y}_{t|t}^m$ with the covariance of the estimation error equal to $\mathbf{\Sigma}_{t|t}^m$.

7.2. Statistical Properties & Key Interpretations

In the following subsection, we show the most important properties of the Kalman Filter assuming that the error terms are mutually independent and independent in time.

7.2.1. Minimum Mean Square Error Estimation by Kalman Filter for Gaussian Linear Models

As mentioned in the previous section, the conditional expectations, defined by the Gaussian linear state-space model, are an optimal estimators for a wide range of loss functions. That includes the minimum mean square errors (MSE)of the estimations what we show next. Let us start with the estimator $\boldsymbol{\beta}_{t|t}$. Given the recursion of the Kalman Filter from Equation (7.1), $\boldsymbol{\beta}_{t|t} = \mathbb{E}_{\boldsymbol{\beta}_t|\mathbf{Y}_{0:t},\Psi}\left[\boldsymbol{\beta}_t\right]$, the estimator $\boldsymbol{\beta}_{t|t}$ is unbiased since

$$\mathbb{E}_{\boldsymbol{\beta}_t|\mathbf{Y}_{0:t},\Psi}\left[\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t|t}\right] = 0. \tag{7.7}$$

Given the following decomposition of the estimation error $\mathbf{e}_t := \boldsymbol{\beta}_t - \boldsymbol{\beta}_{t|t}$, that is

$$\mathbf{e}_t = \boldsymbol{\beta}_t - \mathbb{E}_{\boldsymbol{\beta}_t|\mathbf{Y}_{0:t},\Psi}\left[\boldsymbol{\beta}_t\right] + \mathbb{E}_{\boldsymbol{\beta}_t|\mathbf{Y}_{0:t},\Psi}\left[\boldsymbol{\beta}_t\right] - \boldsymbol{\beta}_{t|t},$$

the MSE of the estimation is specified as

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\beta}_t|\mathbf{Y}_{0:t},\Psi}\left[\mathbf{e}_t \mathbf{e}_t^T\right] &= \mathbb{E}_{\boldsymbol{\beta}_t|\mathbf{Y}_{0:t},\Psi}\left[\left(\boldsymbol{\beta}_t - \mathbb{E}_{\boldsymbol{\beta}_t|\mathbf{Y}_{0:t},\Psi}\left[\boldsymbol{\beta}_t\right]\right)\left(\boldsymbol{\beta}_t - \mathbb{E}_{\boldsymbol{\beta}_t|\mathbf{Y}_{0:t},\Psi}\left[\boldsymbol{\beta}_t\right]\right)^T\right] \\
&\quad + 2\mathbb{E}_{\boldsymbol{\beta}_t|\mathbf{Y}_{0:t},\Psi}\left[\left(\boldsymbol{\beta}_t - \mathbb{E}_{\boldsymbol{\beta}_t|\mathbf{Y}_{0:t},\Psi}\left[\boldsymbol{\beta}_t\right]\right)\right]\left(\mathbb{E}_{\boldsymbol{\beta}_t|\mathbf{Y}_{0:t},\Psi}\left[\boldsymbol{\beta}_t\right] - \boldsymbol{\beta}_{t|t}\right)^T \\
&\quad + \left(\mathbb{E}_{\boldsymbol{\beta}_t|\mathbf{Y}_{0:t},\Psi}\left[\boldsymbol{\beta}_t\right] - \boldsymbol{\beta}_{t|t}\right)\left(\mathbb{E}_{\boldsymbol{\beta}_t|\mathbf{Y}_{0:t},\Psi}\left[\boldsymbol{\beta}_t\right] - \boldsymbol{\beta}_{t|t}\right)^T
\end{aligned}$$

$$= \text{Tr}\left\{\mathbf{Var}_{\boldsymbol{\beta}_t|\mathbf{Y}_{0:t},\Psi}\left[\boldsymbol{\beta}_t\right]\right\} + \left(\mathbb{E}_{\boldsymbol{\beta}_t|\mathbf{Y}_{0:t},\Psi}\left[\boldsymbol{\beta}_t\right] - \boldsymbol{\beta}_{t|t}\right)\left(\mathbb{E}_{\boldsymbol{\beta}_t|\mathbf{Y}_{0:t},\Psi}\left[\boldsymbol{\beta}_t\right] - \boldsymbol{\beta}_{t|t}\right)^T$$

where the first term $\mathbf{Var}_{\boldsymbol{\beta}_t|\mathbf{Y}_{0:t},\Psi}\left[\boldsymbol{\beta}_t\right]$ represents the variance of $\boldsymbol{\beta}_t$ conditioned on the available information and the second term is a bias of the estimator $\boldsymbol{\beta}_{t|t}$. Recall that in order to minimize the mean square error, one needs to minimize the terms which represents the variance of the estimation and select where possible, the zero bias class of estimators. Since $\boldsymbol{\beta}_{t|t} = \mathbb{E}_{\boldsymbol{\beta}_t|\mathbf{Y}_{0:t},\Psi}\left[\boldsymbol{\beta}_t\right]$, the MSE of the estimation is minimal as the term, which reflects the strength of the estimation bias, is equal to zero, that is

$$\mathbb{E}_{\boldsymbol{\beta}_t|\mathbf{Y}_{0:t},\Psi}\left[\mathbf{e}_t\mathbf{e}_t^T\right] = \text{Tr}\left\{\mathbf{Var}_{\boldsymbol{\beta}_t|\mathbf{Y}_{0:t},\Psi}\left[\boldsymbol{\beta}_t\right]\right\}$$

Therefore, the estimator $\boldsymbol{\beta}_{t|t}$ is the minimum MSE estimator of the state vector $\boldsymbol{\beta}_t$ over the whole set of observations.

The proofs related to the optimality of the other estimators which are present in the recursions of the Kalman Filter follow the same reasoning. Hence, the Kalman Filter provides the optimal estimation for Gaussian linear state space models.

7.2.2. Kalman Filter as Minimum Mean Square Linear Filter for Non-Gaussian Data

When the data generative process does not follow Gaussian distributions, the recursions of the Kalman Filter do not correspond to the various conditional expectations of the observation process and latent states. Therefore, the obtained estimators of the state vector are not optimal in a sense of Optimal Filtering Theory discussed in Section 6.1.

However, due to the construction of the Kalman Filter, the obtained estimators of the states still provide unbiased and minimum mean square estimates among all linear projections . The properties stem from the fact that the matrix $\mathbf{K}_t$, known as the Kalman gain, is derived in a way which ensure unbiased estimation of states and minimal covariance matrix $\mathbf{P}_{t|t-1}$.

What is more, given the specification of $\mathbf{K}_t$, the linear estimator given by Kalman Filter recursions belongs to the class of the best linear unbiased estimators, known as BLU estimator, which posses additional statistical properties such as consistency and efficiency. If instead of using linear projection given by the Kalman Filter formulation, one derives the optimal expectations, the obtained estimators might loose these statistical properties. The obtained estimators by conditional expectations are optimal but might not be linear. We remark that the above discussion on the linearity or non-linearity of estimators is separate from the linearity or non-linearity of the underlying state-space model. We refer the reader for an illustrative discussion in (Chapter 5, Anderson and Moore (1979)) or (Chapter 3,5, Bar-Shalom et al. (2001)).

7.2.3. The Interpretation of the Kalman Gain

The Kalman gain matrix $\mathbf{K}_t := \boldsymbol{\Sigma}_{t|t-1}^{-1}\boldsymbol{\Lambda}^T\mathbf{P}_{t|t-1}$ controls the relation between the prediction of the state vector and the observation vector and its impact on the estimation of the state process: the magnitude of its error and its volatility specified by $\mathbf{P}_{t|t}$. The gain is is proportional to the

state prediction variance $\mathbf{P}_{t|t-1}$ and inversely proportional to the innovation variance $\mathbf{\Sigma}_{t|t-1}^{-1}$ what implies that its magnitude relies on the proportion between accuracy of the filtering of states and prediction of the measurements. If $\mathbf{K}_t$ equals zero, the estimator

$$\boldsymbol{\beta}_{t|t} = \boldsymbol{\beta}_{t|t-1} + (\mathbf{y}_t - \mathbf{y}_{t|t-1})\mathbf{K}_t$$

and its covariance matrix

$$\mathbf{P}_{t|t} = \mathbf{P}_{t|t-1} - \mathbf{K}_t^T \mathbf{\Lambda} \mathbf{P}_{t|t-1}$$

do not account for the prediction error of the observation, $\mathbf{v}_t = \mathbf{y}_t - \mathbf{y}_{t|t-1}$. There are two main scenario when the matrix $\mathbf{K}_t$ becomes zero. The first one reflects the situation when the prediction of the state vector is itself very accurate. Then the magnitude of $\mathbf{P}_{t|t-1}$ is small and the covariance matrix of the prediction error approaches zero, what is presented by the following element-wise limit

$$\lim_{\mathbf{P}_{t|t-1} \to \mathbf{0}} \mathbf{K}_t = \mathbf{0}$$

On the other had, if the covariance matrix of the observation error term is large, the noise dominates arriving measurements which are not reliable to estimate the state vectors. The situation is illustrated by the second limiting case given by

$$\lim_{\mathbf{Q} \to \infty} \mathbf{K}_t = \mathbf{0}$$

We should remark on the other important limiting case of the Kalman Gain, which appears when the measurements are accurate and the prediction error of the observation vector is negligible, that is

$$\lim_{\mathbf{Q} \to \mathbf{0}} \mathbf{K}_t = \mathbf{\Lambda}^{-T}.$$

Consequently, $\mathbf{P}_{t|t}$ equals zero and $\boldsymbol{\beta}_{t|t} = (\mathbf{y}_t - \boldsymbol{\mu})\mathbf{\Lambda}^{-T}$ since the information provided by the measurement about the state vector is itself very accurate.

7.2.4. Main Characteristic of Dynamic Systems

The dynamic systems, which are an equivalent name of the state space models in control engineering works, can be examined given the properties such as controllability or observability which notion was firstly studied by (Kalman 1962).

Intuitively, the dynamic system from Equation (5.1) is controllable if the state vector $\boldsymbol{\beta}_t$ can be transferred to the desired state $\boldsymbol{\beta}^*$ in finite time m, that is $\boldsymbol{\beta}_{t+m} = \boldsymbol{\beta}^*$, by choosing the "control inputs" $\boldsymbol{\eta}_{t+1}, \ldots, \boldsymbol{\eta}_{t+m}$.

Definition 10. *The system from Equation* (5.1) *is controllable when*

$$Rank \left[\mathbf{R}, \mathbf{AR}, \ldots, \mathbf{A}^{m-1}\mathbf{R} \right] = m.$$

If Kalman Filer of the model in Equation (5.1) is not controllable, it means that the error term in the state equation does not affect the estimation of every element of the state vector $\boldsymbol{\beta}_t$ and the estimates are uncorrupted by this noise vector. The corresponding elements of the covariance matrix $\mathbf{P}_{t|t-1}$ will approach zero which lead to the fixed estimates of the states $\boldsymbol{\beta}_{t|t}$ in time.

Another characteristic of the dynamic system is observability which determines whether at any point of time t, there exists a finite time $t+m$ such that the state vector $\boldsymbol{\beta}_t$ can be specified given the observations $\mathbf{y}_t, \ldots, \mathbf{y}_{t+m}$, given known $\boldsymbol{\eta}_t, \ldots, \boldsymbol{\eta}_{t+m}$.

Definition 11. *The system from Equation* (5.1) *is observable when*

$$Rank\left[\boldsymbol{\Lambda}^T, \mathbf{A}^T\boldsymbol{\Lambda}^T, \ldots, \mathbf{A}^{T\,m-1}\boldsymbol{\Lambda}^T\right] = m.$$

When the uncontrollable system can be sometimes a desired outcome for Kalman Filter, the unobservable system informs that the filter does not work. If the system is unobservable it indicates that the measurements do not contain enough information to determine some elements of the state vector, regardless of the number of observations taken.

The dynamic system which is controllable and observable for all $t \to \infty$ and have bounded matrices $\mathbf{P}_{t|t}$, $\mathbf{P}_{t|t-1}$ and $\mathbf{K}_t$, is shown by (Kalman 1963) or (Jazwinski 1970) to be asymptotically stable. It implies that given sufficiently large number of measurements, the estimates provided by the filter become insensitive to the initial assumptions.

Such model-based stability defined by observability and controllability is highly desirable in the class of dynamic models studied in this thesis as they directly concern the identification of studied systems. One of the reasons for the lack of defined stability in a dynamic model is a presence of redundant state variables. It would suggest that the subset of the state vector is affected neither by the measurement process nor the reminding elements of the latent vector. In the context of the studied models with incorporated extracted features, this would indicate that the model uses more explanatory variables that are needed to characterise the observation process. Therefore, some of the features add unnecessary complexity to the filtering procedure and possible bias.

7.3. Forecasting via Kalman Filter

The Kalman Filter yields a straightforward procedure to forecast linear Gaussian state-space models. Let us assume that the information about the system from Equation (6.1) is available for $t = 1, \ldots, N$. We seek an m-step-ahead prediction of the observation vector $\mathbf{Y}_{N+m}$ given the information provided up to time $t = N$. This predictive distribution of the random vector $\mathbf{Y}_{N+m}$ is expresses as

$$\begin{aligned}\pi_{\mathbf{Y}_{N+m}|\mathbf{Y}_{1:N},\Psi}\left(\mathbf{y}_{N+m}\right) &= \int_{\mathbb{R}^{k\times m}} \pi_{\mathbf{Y}_{N+m},\boldsymbol{\beta}_{N:(N+m)}|\mathbf{Y}_{1:N},\Psi}\left(\mathbf{y}_{N+m}, \boldsymbol{\beta}_{N:(N+m)}\right)\, d\boldsymbol{\beta}_{N:(N+m)} \\ &= \int_{\mathbb{R}^{k\times m}} \pi_{\mathbf{Y}_{N+m}|\boldsymbol{\beta}_{N+m},\Psi}\left(\mathbf{y}_{N+m}\right)\pi_{\boldsymbol{\beta}_{N:(N+m)}|\mathbf{Y}_{1:N},\Psi}\left(\boldsymbol{\beta}_{N:(N+m)}\right)\, d\boldsymbol{\beta}_{N:(N+m)} \quad (7.8)\end{aligned}$$

where

$$\pi_{\boldsymbol{\beta}_{N:(N+m)}|\mathbf{Y}_{1:N},\Psi}(\boldsymbol{\beta}_{N:(N+m)}) = \pi_{\boldsymbol{\beta}_N|\mathbf{Y}_{1:N},\Psi}(\boldsymbol{\beta}_N)\prod_{t=N}^{N+m}\pi_{\boldsymbol{\beta}_{t+1}|\boldsymbol{\beta}_t,\Psi}(\boldsymbol{\beta}_{t+1}).$$

Under Theorem 30, the optimal predictor of the random vector $\mathbf{Y}_{N+m}|\mathbf{Y}_{1:N}$, which minimizes the covariance of the prediction error, is the conditional expectation given by

$$\begin{aligned}\mathbf{y}_{N+m|N} &= \mathbb{E}_{\mathbf{Y}_{N+m}|\mathbf{Y}_{1:N},\Psi}\left[\mathbf{Y}_{N+m}\right] = \int_{\mathbb{R}^d}\mathbf{y}_{N+m}\,\pi_{\mathbf{Y}_{N+m}|\mathbf{Y}_{1:N},\Psi}(\mathbf{y}_{N+m})\,d\mathbf{y}_{N+m}\\ &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^{k\times m}}\mathbf{y}_{N+m}\pi_{\mathbf{Y}_{N+m}|\boldsymbol{\beta}_{N+m},\Psi}(\mathbf{y}_{N+m})\pi_{\boldsymbol{\beta}_{N:(N+m)}|\mathbf{Y}_{1:N},\Psi}(\boldsymbol{\beta}_{N:(N+m)})\,d\boldsymbol{\beta}_{N:(N+m)}\\ &= \mathbb{E}_{\boldsymbol{\beta}_{N:(N+m)}|\mathbf{Y}_{1:N},\Psi}\left[\mathbb{E}_{\mathbf{Y}_{N+m}|\boldsymbol{\beta}_{N+m},\Psi}\left[\mathbf{Y}_{N+m}\right]\right]. \qquad (7.9)\end{aligned}$$

The variance of the prediction error is specified by

$$\begin{aligned}\boldsymbol{\Sigma}_{N+m|N} &= \mathbb{E}_{\mathbf{Y}_{N+m}|\mathbf{Y}_{1:N},\Psi}\left[\left(\mathbf{Y}_{N+m}-\mathbf{y}_{N+m|N}\right)^T\left(\mathbf{Y}_{N+m}-\mathbf{y}_{N+m|N}\right)\right]\\ &= \mathbb{E}_{\boldsymbol{\beta}_{N:(N+m)}|\mathbf{Y}_{1:N},\Psi}\left[\mathbb{E}_{\mathbf{Y}_{N+m}|\boldsymbol{\beta}_{N+m},\Psi}\left[\mathbf{Y}_{N+m}^T\mathbf{Y}_{N+m}\right]\right]+\mathbf{y}_{N+m|N}^T\mathbf{y}_{N+m|N} \qquad (7.10)\end{aligned}$$

Let us consider the one-step-ahead joint predictive distribution of $\boldsymbol{\beta}_{N+1}$ and $\mathbf{Y}_{N+1}$, which is specified by the dynamic model in Equation (5.1). Under the model assumptions, we note that the distribution in Equation (7.8) is Gaussian with first and second moments given in closed forms. Hence, we can obtain the following predictors of the latent and measurement vectors

$$\boldsymbol{\beta}_{N+1}|\mathbf{Y}_{1:N}\sim\mathcal{N}\left(\mathbf{c}+\boldsymbol{\beta}_{N|N}\mathbf{A},\ \mathbf{A}^T\mathbf{P}_{N|N}\mathbf{A}+\mathbf{R}\right),\quad \mathbf{Y}_{N+1}|\mathbf{Y}_{1:N}\sim\mathcal{N}\left(\boldsymbol{\mu}+\boldsymbol{\beta}_{N+1|N}\boldsymbol{\Lambda},\ \boldsymbol{\Sigma}_{N+1|N}\right).$$

Then, the m-step-ahead predictions of the vectors $\boldsymbol{\beta}_{N+1}$ and $\mathbf{Y}_{N+1}$ are obtained recursively by

$$\boldsymbol{\beta}_{N+m}|\mathbf{Y}_{1:N}\sim\mathcal{N}\left(\boldsymbol{\beta}_{N+m|N},\ \mathbf{P}_{N+m|N}\right),\qquad \mathbf{Y}_{N+m}|\mathbf{Y}_{1:N}\sim\mathcal{N}\left(\mathbf{y}_{N+m|N},\ \boldsymbol{\Sigma}_{N+m|N}\right)$$

for

$$\begin{aligned}&\boldsymbol{\beta}_{N+m|N}=\boldsymbol{\beta}_{N|N}(\mathbf{A})^m+\mathbf{c}\left(\mathbb{I}_k+\sum_{i=1}^{m-1}(\mathbf{A})^i\right),\quad \mathbf{P}_{N+m|N}=(\mathbf{A}^T)^m\mathbf{P}_{N|N}(\mathbf{A})^m+\sum_{i=0}^{m-1}(\mathbf{A}^T)^i\mathbf{R}(\mathbf{A})^i,\\ &\mathbf{y}_{N+m|N}=\boldsymbol{\mu}+\boldsymbol{\beta}_{N+m|N}\boldsymbol{\Lambda},\qquad \boldsymbol{\Sigma}_{N+m|N}=\boldsymbol{\Lambda}^T\mathbf{P}_{N+m|N}\boldsymbol{\Lambda}+\mathbf{Q}.\end{aligned}$$

The derived quantities are a special case of distribution defined in Equation (7.8), Equation (7.9) and Equation (7.10) under the assumption about the Gaussian linear state-space models. In order to access the accuracy of the prediction, we define the mean square error of prediction (MSEP) as

$$\text{MSEP}_N(m):=\mathbb{E}_{\mathbf{Y}_{N+m}|\mathbf{Y}_{1:N},\Psi}\left[\left(\mathbf{Y}_{N+m}-\mathbb{E}_{\mathbf{Y}_{N+m}|\mathbf{Y}_{1:N},\bar{\Psi}}\left[\mathbf{Y}_{N+m}\right]\right)\left(\mathbf{Y}_{N+m}-\mathbb{E}_{\mathbf{Y}_{N+m}|\mathbf{Y}_{1:N},\Psi}\left[\mathbf{Y}_{N+m}\right]\right)^T\right].$$

7.4. Smoothing via Kalman Filter

Whilst the objective of the filtering is to estimate the state $\boldsymbol{\beta}_t$ conditioned on the current information, the goal of smoothing is to compute the conditional distribution after receiving all

measurements. The objective of smoothing is to revise estimates specify by the filtering problem at some prior time given all information available. We can distinguish various types of smoothing

- a fixed interval smoother, the off-line procedure, which yields smoothed estimates of a finite set of state vectors $\mathbf{b}_t|\mathbf{Y}_{1:s}$ for all $t \leq N$ and fixed $t < s \leq N$;
- a fixed point smoother, which calculates a smoothed estimates of $\boldsymbol{\beta}_t$ conditional on $\mathbf{Y}_{t:(t+s)}$ for all $N - t > s > 0$;
- or a fixed-lag smoother, being an online procedure, which is useful if we observe a fixed delay between an measurement reception and the estimate of $\boldsymbol{\beta}_t$ conditioned on the measurement, $\mathbf{Y}_{t:(t+s)}$ for all t.

Please refer to (Anderson and Moore 1979) or more recently, (Doucet and Johansen 2009) or Briers et al. (2010), for more details on the smoothing procedures for linear Gaussian state-space models.

In the following section we discuss the concept of the fixed-interval smoothing which is used in application studies of this thesis. Let us assume that the fixed interval reflects the whole available information given by the measurements of our model, the observation set $\mathbf{Y}_{1:N}$. The fixed-interval smoothing is characterized by the following probability density functions specified in (Kitagawa 1983) or (Anderson and Moore 1979)

$$\begin{aligned}\pi_{\boldsymbol{\beta}_t|\boldsymbol{Y}_{1:N},\Psi}(\boldsymbol{\beta}_t) &= \int_{\mathbb{R}^k} \pi_{\boldsymbol{\beta}_t,\boldsymbol{\beta}_{t+1}|\boldsymbol{Y}_{1:N},\Psi}(\boldsymbol{\beta}_t,\boldsymbol{\beta}_{t+1})\, d\boldsymbol{\beta}_{t+1} \\ &= \pi_{\boldsymbol{\beta}_t|\boldsymbol{Y}_{1:t},\Psi}(\boldsymbol{\beta}_t) \int_{\mathbb{R}^k} \frac{\pi_{\boldsymbol{\beta}_{t+1}|\boldsymbol{\beta}_t,\Psi}(\boldsymbol{\beta}_t)\ \pi_{\boldsymbol{\beta}_{t+1}|\boldsymbol{Y}_{1:N},\Psi}(\boldsymbol{\beta}_{t+1})}{\pi_{\boldsymbol{\beta}_{t+1}|\boldsymbol{Y}_{1:t},\Psi}(\boldsymbol{\beta}_t)}\, d\boldsymbol{\beta}_{t+1} \end{aligned} \tag{7.11}$$

Given the similar reasoning as in Theorem 29 and Theorem 30, we can prove that the optimal smoothing estimator of the random variable $\boldsymbol{\beta}_t|\boldsymbol{Y}_{1:N},\Psi$ equals a conditional expectation $\boldsymbol{\beta}_{t|N} = \mathbb{E}_{\boldsymbol{\beta}_t|\boldsymbol{Y}_{1:N},\Psi}[\boldsymbol{\beta}_t]$

The probability density function from Equation (7.11) yields the forward-backward recursive procedure. The procedure is very efficient in calculating the smoothed estimates of $\boldsymbol{\beta}_{t|N}$ for Gaussian linear state-space models using forward Kalman Filter and recursive backward iteration steps. The, so called, Kalman smoother was introduced by (Rauch et al. 1965). Let us remark that given the Kalman Filter recursions in Section 7.1 in Equation (7.1) and Equation (7.2), we have that

$$\begin{aligned}\boldsymbol{\beta}_{t+1}|\boldsymbol{Y}_{1:t},\Psi &\sim \mathcal{N}(\boldsymbol{\beta}_{t+1|t},\mathbf{P}_{t+1|t}) \\ \boldsymbol{\beta}_t|\boldsymbol{Y}_{1:t},\Psi &\sim \mathcal{N}(\boldsymbol{\beta}_{t|t},\mathbf{P}_{t|t}) \\ \boldsymbol{\beta}_{t+1}|\boldsymbol{\beta}_t,\Psi &\sim \mathcal{N}(c+\boldsymbol{\beta}_t\mathbf{A},\mathbf{R})\end{aligned}$$

which results in the distribution of $\boldsymbol{\beta}_t|\boldsymbol{Y}_{1:N},\Psi$ for linear Gaussian state-space models such that

$$\boldsymbol{\beta}_t|\boldsymbol{Y}_{1:N},\Psi \sim \mathcal{N}(\boldsymbol{\beta}_{t|N},\mathbf{P}_{t|N}) \tag{7.12}$$

The smoothing algorithm consists of a set of recursions which start with the final quantities given by the Kalman Filter, $\boldsymbol{\beta}_{N|N}$ and $\mathbf{P}_{N|N}$ and moves backward given the steps

$$\begin{aligned}\boldsymbol{\beta}_{t|N} &= \boldsymbol{\beta}_{t|t} + (\boldsymbol{\beta}_{t+1|N} - \boldsymbol{\beta}_{t+1|t})\mathbf{P}_{t+1|t}^{-1}\mathbf{A}^T\mathbf{P}_{t|t} \\ \mathbf{P}_{t|N} &= \mathbf{P}_{t|t} - \mathbf{P}_{t|t}\mathbf{A}^T\mathbf{P}_{t+1|t}^{-1}(\mathbf{P}_{t+1|N} - \mathbf{P}_{t+1|t})\mathbf{P}_{t+1|t}^{-1}\mathbf{A}\mathbf{P}_{t|t}.\end{aligned} \tag{7.13}$$

The trajectory of the estimates $\boldsymbol{\beta}_{t|N}$ obtained by smoothing methods tend to be smoother than those obtained by filtering. Consequently, the variance of the smoothed estimator is not greater than of the filtered estimator $\boldsymbol{\beta}_{t|t}$, that is

$$\mathbf{P}_{t|t+1} \leq \mathbf{P}_{t|t}$$

The reader may refer to (Doucet and Johansen 2009) or (Briers et al. 2010) for a detailed discussion on the properties of smoothing algorithms and their advantages.

7.5. Estimation of Static Parameters via Marginalized Likelihood for Linear Gaussian State-Space Models

The classical statistical method for the estimation of the parameters in the frequentists setting is Maximum Likelihood Estimation (MLE). The method finds an estimator of Ψ, the static parameters of the model by maximising the joint probability function of the measurements, $\pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N})$. Therefore, we chose the value of the static parameters which maximize the probability of observing given measurements. The MLE estimation of Ψ is not optimal for the finite sample but it poses appealing limiting cases such as consistency, unbiasedness, efficiency, sufficiency - please refer to (Schervish 1995) for broader discussion, among many other references. Since the realisations of the observation vector $\mathbf{Y}_t$, defined by the model in Equation (5.1), are not independent in time, we recurrently apply the Chain Rule of probabilities to obtain the following factorisation of the likelihood

$$L(\Psi; \mathbf{y}_{1:N}) := \pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N}) = \pi_{\mathbf{Y}_N|\mathbf{Y}_{1:(N-1)}|\Psi}(\mathbf{y}_N)\pi_{\mathbf{Y}_{1:(N-1)}|\Psi}(\mathbf{y}_{1:(N-1)}) = \prod_{t=1}^{N}\pi_{\mathbf{Y}_t|\mathbf{Y}_{1:(t-1)},\Psi}(\mathbf{y}_t) \tag{7.14}$$

where $\pi_{\mathbf{Y}_t|\mathbf{Y}_{1:(t-1)}|\Psi}(\mathbf{y}_t)$ denotes the distribution of the random vector $\mathbf{Y}_t$ conditioned on observed values up to time $t-1$. When the error terms are Gaussian, the Kalman Filter yields the marginal distribution of $\mathbf{Y}_t|\mathbf{Y}_{1:(t-1)}$ and provides its moments specified in Equation (7.1). Then, the vector of static parameters Ψ is specified by the decomposition of the prediction error. Let us define $\mathbf{v}_t := \mathbf{y}_t - \mathbf{y}_{t|t-1}$ as a vector of residuals at time t, being a difference between the observed data and its in-sample one-step-ahead prediction. The estimation of Ψ is carried out by the optimisation of the following likelihood function

$$l(\Psi; \mathbf{y}_{1:N}) := \log L(\Psi; \mathbf{y}_{1:N}) = -\frac{dN}{2}\log\pi - \frac{1}{2}\sum_{t=1}^{N}\log|\boldsymbol{\Sigma}_{t|t-1}| - \frac{1}{2}\sum_{t=1}^{N}\mathbf{v}_t^T\boldsymbol{\Sigma}_{t|t-1}^{-1}\mathbf{v}_t. \tag{7.15}$$

(Harvey 1990) shows that for the state space model defined in Equation (5.1), the maximizers of the loglikelihood function from Equation (7.15) admit closed form solutions defined by the gradient

$$\left[\nabla_{\Psi} l(\Psi; \mathbf{y}_{1:N})\right]_i := \frac{\partial}{\partial \Psi_i} l(\Psi; \mathbf{y}_{1:N}) = \frac{1}{2} \sum_{t=1}^{N} \left\{ \mathrm{Tr}\left\{ \left(\boldsymbol{\Sigma}_{t|t-1} \frac{\partial \boldsymbol{\Sigma}_{t|t-1}}{\partial \Psi_i} \right) \left(\mathbf{1}_d - \boldsymbol{\Sigma}_{t|t-1} \mathbf{v}_t^T \mathbf{v}_t \right) \right\} + 2 \frac{\partial \mathbf{v}_t}{\partial \Psi_i} \boldsymbol{\Sigma}_{t|t-1} \mathbf{v}_t^T \right\}, \tag{7.16}$$

and the information matrix

$$-\left[\nabla^2_{\Psi} l(\Psi; \mathbf{y}_{1:N})\right]_{i,j} := -\frac{\partial^2}{\partial \Psi_i \partial \Psi_j} l(\Psi; \mathbf{y}_{1:N}) = \frac{1}{2} \sum_{t=1}^{N} \left[\mathrm{Tr}\left\{ \boldsymbol{\Sigma}^{-1}_{t|t-1} \frac{\partial \boldsymbol{\Sigma}_{t|t-1}}{\partial \Psi_i} \boldsymbol{\Sigma}^{-1}_{t|t-1} \frac{\partial \boldsymbol{\Sigma}_{t|t-1}}{\partial \Psi_j} \right\} + \frac{\partial \mathbf{v}_t^T}{\partial \Psi_i} \boldsymbol{\Sigma}^{-1}_{t|t-1} \frac{\partial \mathbf{v}_t}{\partial \Psi_j} \right], \tag{7.17}$$

for $i, j = 1, \ldots, |\Psi|$. In order to evaluate the gradient and the information matrix, we need to specify the following partial derivatives

$$\begin{aligned}
&\frac{\partial \mathbf{v}_t}{\partial \Psi_i} = -\frac{\partial \boldsymbol{\mu}}{\partial \Psi_i} - \boldsymbol{\beta}_{t|t-1} \frac{\partial \boldsymbol{\Lambda}}{\partial \Psi_i} - \frac{\partial \boldsymbol{\beta}_{t|t-1}}{\partial \Psi_i} \boldsymbol{\Lambda}, \\
&\frac{\partial \boldsymbol{\Sigma}_{t|t-1}}{\partial \Psi_i} = \frac{\partial \boldsymbol{\Lambda}^T}{\Psi_i} \mathbf{P}_{t|t-1} \boldsymbol{\Lambda} + \boldsymbol{\Lambda}^T \frac{\partial \mathbf{P}_{t|t-1}}{\Psi_i} \boldsymbol{\Lambda} + \boldsymbol{\Lambda}^T \mathbf{P}_{t|t-1} \frac{\partial \boldsymbol{\Lambda}}{\Psi_i} + \frac{\partial \mathbf{Q}}{\partial \Psi_i}, \\
&\frac{\partial \boldsymbol{\beta}_{t|t-1}}{\partial \Psi_i} = \frac{\partial \mathbf{c}}{\partial \Psi_i} - \boldsymbol{\beta}_{t|t} \frac{\partial \mathbf{A}}{\partial \Psi_i} - \frac{\partial \boldsymbol{\beta}_{t|t}}{\partial \Psi_i} \mathbf{A}, \\
&\frac{\partial \mathbf{P}_{t|t-1}}{\partial \Psi_i} = \frac{\partial \mathbf{A}^T}{\Psi_i} \mathbf{P}_{t-1|t-1} \mathbf{A} + \mathbf{A}^T \frac{\partial \mathbf{P}_{t-1|t-1}}{\Psi_i} \mathbf{A} + \mathbf{A}^T \mathbf{P}_{t-1|t-1} \frac{\partial \mathbf{A}}{\Psi_i} + \frac{\partial \mathbf{R}}{\partial \Psi_i}, \\
&\frac{\partial \boldsymbol{\beta}_{t-1|t-1}}{\partial \Psi_i} = \frac{\partial \boldsymbol{\beta}_{t-1|t-2}}{\partial \Psi_i} - \left(\frac{\partial \boldsymbol{\mu}}{\partial \Psi_i} + \frac{\partial \boldsymbol{\beta}_{t-1|t-2}}{\partial \Psi_i} \boldsymbol{\Lambda} - \boldsymbol{\beta}_{t-1|t-2} \frac{\partial \boldsymbol{\Lambda}}{\partial \Psi_i} \right) \mathbf{K}_{t-1} + \left(\mathbf{y}_{t-1} - \boldsymbol{\mu} - \boldsymbol{\beta}_{t-1|t-2} \boldsymbol{\Lambda} \right) \frac{\partial \mathbf{K}_{t-1}}{\partial \Psi_i}, \\
&\frac{\partial \mathbf{K}_{t-1}}{\partial \Psi_i} = -\boldsymbol{\Sigma}^{-1}_{t-1|t-2} \frac{\partial \boldsymbol{\Sigma}_{t-1|t-2}}{\partial \Psi_i} \boldsymbol{\Sigma}^{-1}_{t-1|t-2} \boldsymbol{\Lambda}^T \mathbf{P}_{t-1|t-2} + \boldsymbol{\Sigma}^{-1}_{t-1|t-2} \left(\frac{\partial \boldsymbol{\Lambda}^T}{\partial \Psi_i} \mathbf{P}_{t-1|t-2} + \boldsymbol{\Lambda}^T \frac{\partial \mathbf{P}_{t-1|t-2}}{\partial \Psi_i} \right), \\
&\frac{\partial \mathbf{P}_{t-1|t-1}}{\partial \Psi_i} = \frac{\partial \mathbf{P}_{t-1|t-2}}{\partial \Psi_i} - \frac{\partial \mathbf{K}^T_{t-1}}{\partial \Psi_i} \boldsymbol{\Sigma}_{t|t-1} \mathbf{K}_{t-1} - \mathbf{K}^T_{t-1} \left(\frac{\partial \boldsymbol{\Sigma}_{t|t-1}}{\partial \Psi_i} \mathbf{K}_{t-1} + \boldsymbol{\Sigma}_{t|t-1} \frac{\partial \mathbf{K}_{t-1}}{\partial \Psi_i} \right),
\end{aligned} \tag{7.18}$$

for $\mathbf{K}_t = \boldsymbol{\Sigma}^{-1}_{t|t-1} \boldsymbol{\Lambda}^T \mathbf{P}_{t|t-1}$. Notice that $\frac{\partial \mathbf{b}_0}{\partial \Psi_i} = \mathbf{0}$ and $\frac{\partial \mathbf{P}_0}{\partial \Psi_i} = \mathbf{0}$ for $i = 1, \ldots, |\Psi|$. Even though the gradient equations can be calculated explicitly, the procedure requires to find a solution to multiple non-linear equations and can become computationally intensive. For this reasons, a common practice is to adopt a gradient-based approach - a recursive estimation which uses the gradient in Equation (7.16) and the information matrix in Equation (7.17). Then, the optimal parameters are obtained by the iterative scheme with an mth update step being the following

$$\Psi^{(m+1)} = \Psi^{(m)} - \left[\nabla^2_{\Psi} l(\Psi^{(m)}; \mathbf{y}_{1:N}) \right]^{-1} \nabla_{\Psi} l(\Psi^{(m)}; \mathbf{y}_{1:N}). \tag{7.19}$$

The procedure stops when predefined convergence criteria are satisfied. The defined gradient descent algorithm belongs to the class of batch estimation methods.

Let us remark that maximisation of the loglikelihood function in Equation (7.15) with respect to Ψ is equivalent to the recursive least square problem. The book of (Ljung and Soderstrom 1983) provides the descriptions of the wide range of approaches to the recursive identification and details the Recursive Maximum Likelihood (RML) algorithm - the numerical optimisation routine which is used to determined the estimates of Ψ when the solutions to the multiple gradient equations in Equation (7.17) are not obtained analytically. The other set of procedures discussed in (Ljung and

Soderstrom 1983), called stochastic gradient descent, updates the parameters of the loglikelihood function in Equation (7.15) by a stochastic approximation. This online estimation framework has the following form at time t

$$\Psi_{t+1} = \Psi_t + a_t \nabla_{\Psi_{1:t}} \log \pi_{\mathbf{Y}_{1:t}|\Psi_{1:t}}(\mathbf{y}_{1:t}), \tag{7.20}$$

where Ψ_{t+1} and Ψ_t are the parameters' estimates at time $t+1$ and t, $\Psi_{1:t}$ is a sequence of estimates of the vector of parameters over time and $\pi_{\mathbf{Y}_{1:t}|\Psi_{1:t}}(\mathbf{y}_{1:t})$ is equivalent to the likelihood function in Equation (7.14) but with a time varying vector of parameters $\Psi_{1:t}$, which is updated over time steps. The sequence a_t, which determines the step size of the algorithm, is assumed to be a non-increasing such that $\sum_t a_t = \infty$ and $\sum_t a_t^2 < \infty$.

Under the class of linear Gaussian state-space models, the corresponding likelihood function is obtained via the Kalman Filter recursions as in Equation (7.15). If the assumption is relaxed, the distribution of more complex state space models may need to be specified via the Monte Carlo techniques, especially its most efficient group of online frameworks known as Sequential Monte Carlo (SMC). Instead of calculating the likelihood function in Equation (7.14) or the function $\pi_{\mathbf{Y}_{1:t}|\Psi_{1:t}}(\mathbf{y}_{1:t})$, one may adopt the approximation approach studied in (Poyiadjis et al. 2005) and (Poyiadjis et al. 2011) for both, a batch and recursive ML estimation of the parameters. The authors propose the numerical sampling procedure, based on SMC, which approximates optimal filtering integrals and their derivatives. They suggest to use the Fisher identity for the gradient in Equation (7.16)

$$\nabla_\Psi \log \pi_{\mathbf{Y}_{1:t}|\Psi}(\mathbf{y}_{1:t}) = \int \nabla_\Psi \log \pi_{\mathbf{Y}_{1:t},\mathbf{X}_{0:t}|\Psi}(\mathbf{y}_{1:t}) \pi_{\mathbf{X}_{0:t}|\mathbf{Y}_{1:t},\Psi}(\mathbf{x}_{0:t})\, d\mathbf{x}_{0:t}$$

and Louis' identity for the information matrix in Equation (7.17)

$$-\nabla^2_\Psi \log \pi_{\mathbf{Y}_{1:t}|\Psi}(\mathbf{y}_{1:t}) = \left[\nabla_\Psi \log \pi_{\mathbf{Y}_{1:t}|\Psi}(\mathbf{y}_{1:t})\right]^T \nabla_\Psi \log \pi_{\mathbf{Y}_{1:t}|\Psi}(\mathbf{y}_{1:t}) - \frac{\nabla^2_\Psi \pi_{\mathbf{Y}_{1:t}|\Psi}(\mathbf{y}_{1:t})}{\pi_{\mathbf{Y}_{1:t}|\Psi}(\mathbf{y}_{1:t})}$$

where

$$\begin{aligned}\frac{\nabla^2_\Psi \pi_{\mathbf{Y}_{1:t}|\Psi}(\mathbf{y}_{1:t})}{\pi_{\mathbf{Y}_{1:t}|\Psi}(\mathbf{y}_{1:t})} &= \int \left[\nabla_\Psi \log \pi_{\mathbf{Y}_{1:t},\mathbf{X}_{0:t}|\Psi}(\mathbf{y}_{1:t})\right]^T \nabla_\Psi \log \pi_{\mathbf{Y}_{1:t},\mathbf{X}_{0:t}|\Psi}(\mathbf{y}_{1:t}) \pi_{\mathbf{X}_{0:t}|\mathbf{Y}_{1:t},\Psi}(\mathbf{x}_{0:t})\, d\mathbf{x}_{0:t} \\ &+ \int \nabla^2_\Psi \log \pi_{\mathbf{Y}_{1:t},\mathbf{X}_{0:t}|\Psi}(\mathbf{y}_{1:t}) \pi_{\mathbf{X}_{0:t}|\mathbf{Y}_{1:t},\Psi}(\mathbf{x}_{0:t})\, d\mathbf{x}_{0:t}\end{aligned}$$

The proofs to the identities can be found in (Cappe et al. 2005). The gradient and information matrix of the loglikelihood function, expressed in terms of the integrals in Fishers and Louis' identities, are dependent on the path-space of the filtering distribution, $\pi_{\mathbf{X}_{0:t}|\mathbf{Y}_{1:t},\Psi}(\mathbf{x}_{0:t})$ and can be approximated by generating the particles from the conditional distribution using efficient SMC techniques.

8. Bayesian Inference for Linear Gaussian State-Space Models

So far we have considered the state estimation problem in the frequentists paradigm. Now, we generalise the filtering framework to the joint estimation of static model parameters and state variables. In general, this can be achieved from either a frequentist or Bayesian perspective. In this chapter, we focus on Bayesian formulations of such estimation problems which we use in the further studies to reduce the uncertainty of the inference in dynamic models setting. It is followed by the discussion on efficient learning methodologies used in this work – especially relevant to the estimation of latent variables and static parameters in the class of models defined in Equation (5.1).

8.1. Introduction to Bayesian Inference

A Bayesian approach to the inference of state space models starts with an initial knowledge or assumptions about a model structure. We assume prior distribution on the model's parameters. The distribution represents the model's structure before any measurement arrives. Therefore, a priori distribution reflects our beliefs about the parameters and is subjective among individuals who incorporate them into the model. In the light of new measurements, the knowledge is updated to obtain a posterior distribution over the model's unobserved variables and its static parameters. Let us remark that in the frequentists approach, the information about the static parameters is fully contained within a likelihood function. The frequentists inference is based only on the sample set and hence is purely objective without incorporating any assumptions on the model's parameters.

Since the Bayesian setting formally considers Ψ as a random vector with an assigned prior distribution, the posterior density, $\pi_{\Psi|\mathbf{Y}_{1:N}}(\boldsymbol{\psi})$, which can be derived from the likelihood of a model and its prior distribution, serves as entire information available on the vector of static parameters. Let us denote by $\pi_{\Psi}(\boldsymbol{\psi})$ the joint prior distribution. From the Bayes' Theorem, we note that the posterior distribution, the knowledge about the parameters after the observation set arrives, satisfies the following

$$\pi_{\Psi|\mathbf{Y}_{1:N}}(\psi) \;\propto\; \pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N})\; \pi_{\Psi}(\boldsymbol{\psi})$$

where $\pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N})$ is equivalent to the likelihood function of a model. The posterior distribution reflects our uncertainty about the parameters, after observing the measurement process, and allows to quantify our confidence.

In Bayesian setting, the inference of Ψ relies entirely on the posterior distribution of the model and its available summaries such as a posterior mean, a posterior median or a posterior mode. Similarly to the definition of the optimal filtering problem in Section 6.1, the best estimator of Ψ is determined by a choice of a loss criterion based on the expected loss of the estimation defined as

$$\hat{\Psi} = \underset{\Psi^* \in \mathbb{D} \subseteq \Theta}{\operatorname{argmin}}\, \mathbb{E}_{\Psi|\mathbf{Y}_{1:N}}\left[\lambda(\Psi, \Psi^*)\right] = \underset{\Psi^* \in \mathbb{D} \subseteq \Theta}{\operatorname{argmin}} \int_{\Theta} \lambda(\boldsymbol{\psi}, \Psi^*) \pi_{\Psi|\mathbf{Y}_{1:N}}(\boldsymbol{\psi})\, d\boldsymbol{\psi}$$

where $\lambda : \Theta \times \mathbb{D} \subseteq \Theta \rightarrow \mathbb{R}_+$ is a loss function being a measure of proximity between the true parameter vector and its proposed estimator. The loss function is often assumed to be even, differentiable and convex – please refer to (Robert 2007) or Section 6.1 for broaden discussion on its characteristics and properties. We can apply the reasoning of Theorem 29 from Section 6.1 as the justification of the estimation optimality. For instance, the posterior mean corresponds to the quadratic loss function between Ψ and its estimate, whereas the posterior median to the absolute difference. The posterior mode, known as maximum a posteriori (MAP) estimator, assumes the zero-positive loss function given by

$$\lambda(\Psi, \Psi^*) = \lim_{\epsilon \rightarrow \mathbf{0}} \begin{cases} 0 & \text{for } |\Psi - \Psi^*| < \epsilon, \\ 1 & \text{for } |\Psi - \Psi^*| \geq \epsilon. \end{cases}$$

The MAP estimator of Ψ is a Bayesian equivalent of the classical MLE – the optimal estimator is obtained by maximizing $\pi_{\Psi|\mathbf{Y}_{1:N}}(\psi)$. This posterior distribution can be expressed as a penalized likelihood function as defined in (Akaike 1978) or (Akaike 1983). When $\pi_\Psi(\boldsymbol{\psi})$ is a probability density function of the uniform random vector, the two approaches to the estimation, MLE and MAP, provide the same results. Let us highlight that the Bayesian estimators of Ψ are optimal in the finite sample setting on the contrary to the MLE, which optimality is based on asymptotic limits to the infinite sample as argued in (Schervish 1995) or (Robert 2007), among other relevant references on the statistical inference.

The books of (Berger 1985), (O'Hagan et al. 2006) or (Robert 2007), among many other excellent references, provide insightful discussions into the intuition behind the design of the Bayesian inference framework which core element is a choice of a prior distribution on parameters. There are many schools when it comes to selecting the priors – we will exemplify briefly only the main ones. The subjective priors attempt to capture the maximal prior knowledge about the parameters. When no prior information about the parameters is available, or the inference is based only on an available data, one wants to guarantee minimal influence of proposed prior densities on the posterior distribution. Such class of prior distributions is called uninformative or objective prior. The inferences performed using uninformative prior prioritise the information conveyed by data which remains unaffected by information external to the set of observations. The other favourable class of priors are conjugate priors as defined in (Raiffa and Schlaifer 1961), which allows both posterior distribution and prior distribution to remain in the same probability distribution family and follow the same parametric form. The use of conjugate priors is computationally convenient. The method has many practical advantages such as analytical Bayesian integrals in many inference problems, tractability of the subsequent posterior distributions and intuitive interpretations. The conjugate priors are an especially popular choice when studied models belong to the exponential family. However, if the available information about the distribution of the parameters contradicts the choice of the priors, it might be necessary to use a different family of distributions.

8.2. Bayesian Model Development and Inference for Linear Gaussian State-Space Models

The estimation, filtering and prediction of the model in Equation (5.1) via the Bayesian inference allows one to consider the static parameter uncertainty. Recalling the discussion in Section 6.1, the optimal estimation of the latent states $\boldsymbol{\beta}_t$ in the frequentists setting, relies on specification of conditional distributions given by the density function

$$\pi_{\boldsymbol{\beta}_t|\mathbf{Y}_{1:N},\Psi}(\boldsymbol{\beta}_t).$$

When prior assumptions about Ψ are incorporated in a state space model, the target density changes to

$$\pi_{\boldsymbol{\beta}_{0:N},\Psi|\mathbf{Y}_{1:N}}(\boldsymbol{\beta}_{0:N},\boldsymbol{\psi}). \tag{8.1}$$

since both, the vector of latent states $\boldsymbol{\beta}_{0:N}$ and Ψ are unobserved random variables which are inferred from the measurements.

In the following section, we detail the step of an efficient estimation framework of $\boldsymbol{\beta}_{0:N}$ and Ψ which is relevant to the statistical models studied in this thesis. The procedure consists of two stages: the block sampling for the latent state via, so-called, the forward-filtering-backwards-sampling (FFBS) algorithm of (Carter and Kohn 1994b), and obtaining the posterior samples of the static parameters via conjugate priors. For the computational convenience, we focus on the family of conjugate prior distributions as the studies probability density functions of static parameters are often known in linear Gaussian state space models. The sampling procedure is described in Algorithm 2, where M is the number of MCMC iterations performed. Let us remark that the FFBS has the same ability to handle missing observations as classical Kalman Filter.

8.2.1. Combining Gibbs Sampler and Kalman Filter: Forward-Backward Filtering for Latent State Dynamics

The FFBS filter explores the time ordering of a state space model to calculate the posterior mean given by

$$\mathbb{E}_{\boldsymbol{\beta}_{0:N},\Psi|\mathbf{Y}_{1:N}}\Big[(\boldsymbol{\beta}_{0:N},\Psi)\Big] = \mathbb{E}_{\Psi|\mathbf{Y}_{1:N}}\Big[\big(\mathbb{E}_{\boldsymbol{\beta}_{0:N}|\mathbf{Y}_{1:N},\Psi}\big[\boldsymbol{\beta}_{0:N}\big],\Psi\big)\Big] \tag{8.2}$$

which serves as an optimal estimator of the vector $\boldsymbol{\beta}_{0:N},\Psi$ under the quadratic lose criteria as justified in Section 6.1. The procedure obtains the samples from a sequence of distributions which are first constructed recursively by the Kalman Filter. For the state space model in Equations (5.1), the filtering distributions defined by the Kalman Filter recursions are listed in Equation (7.1) and Equation (7.2). The core idea behind the algorithm is to use the following factorization of the density in Equation (8.1), that is

$$\pi_{\boldsymbol{\beta}_{0:N},\Psi|\mathbf{Y}_{1:N}}(\boldsymbol{\beta}_{0:N},\boldsymbol{\psi}) = \pi_{\Psi|\mathbf{Y}_{1:N}}(\boldsymbol{\psi})\pi_{\boldsymbol{\beta}_N|\mathbf{Y}_{1:N},\Psi}(\boldsymbol{\beta}_N)\prod_{t=0}^{N}\pi_{\boldsymbol{\beta}_t|\mathbf{Y}_{1:t},\boldsymbol{\beta}_{t+1},\Psi}(\boldsymbol{\beta}_t). \tag{8.3}$$

This representation of the full posterior probability density function yields an alternative smoothing procedure based on the Kalman Filter, which originates from (Kohn and Ansley 1989) and (de Jong 1989). Given the definition of the model in Equations (5.1), the first step of the FFBS filter is to sample the realisation of $\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t+1}, \mathbf{Y}_{1:t}, \Psi \sim \mathcal{N}(\boldsymbol{\beta}_{t|t+1}, \mathbf{P}_{t|t+1})$ backwardly for the smoothed estimates of $\boldsymbol{\beta}_t$ and its covariance matrix specified as

$$\begin{aligned}\boldsymbol{\beta}_{t|t+1} &= \boldsymbol{\beta}_{t|t} + \mathbf{P}_{t|t}\mathbf{A}^T\mathbf{P}^{-1}_{t+1|t}(\boldsymbol{\beta}_{t+1} - \boldsymbol{\beta}_{t+1|t}) \\ \mathbf{P}_{t|t+1} &= \mathbf{P}_{t|t} - \mathbf{P}_{t|t}\mathbf{A}^T\mathbf{P}^{-1}_{t+1|t}\mathbf{A}\mathbf{P}_{t|t},\end{aligned}$$

Next, the posteriori distribution of Ψ is determined by the realisation of $\boldsymbol{\beta}_t|\boldsymbol{\beta}_{t+1}, \mathbf{Y}_{1:t}, \Psi$ and assumed prior distributions.

The main difficulty of the considered problem is specifying the posteriori density $\pi_{\Psi|\mathbf{Y}_{1:N}}(\boldsymbol{\psi})$. The study of (Carter and Kohn 1994b) proposes a Gibbs sampler procedure which converges to the expectation in Equation (8.2) for sufficiently large number of samples. Instead of sampling from $\pi_{\boldsymbol{\beta}_{0:N},\Psi|\mathbf{Y}_{1:N}}(\boldsymbol{\psi})$, the iterative procedure of Gibbs sampler generates the samples from the conditional densities $\pi_{\boldsymbol{\beta}_{0:N}|\mathbf{Y}_{1:N},\Psi}(\boldsymbol{\beta}_{0:N})$ and $\pi_{\Psi_i|\mathbf{Y}_{1:N},\boldsymbol{\beta}_{0:N},\Psi_{-i}}(\psi_i)$, for $i = 1, \ldots, |\Psi|$ where Ψ_{-i} denotes the random vector of static parameters of the model without ith element. The authors of (Carter and Kohn 1994b) argue that the proposed forward-backward smoothing algorithm, based on the Gibbs sampler and the smoothing distribution in Equation (8.3), is more efficient and computational stable than the equivalent procedure with the smoothing distribution reviewed in Section 7.4. The FFBS filter of (Carter and Kohn 1994b) minimises the autocorrelation between the elements of a Markov Chain by generating realisations from the full posterior density of the model in Equations (5.1). Consequently, the filter enjoys faster convergence to the true posterior distribution. The steps of FFBS are illustrated in Algorithm 2 whereas the framework of the Gibbs sampler is discussed in details in Subsection 10.2.2.

Algorithm 2 MCMC sampling for $\pi_{\boldsymbol{\beta}_{0:N},\Psi|\mathbf{Y}_{1:N}}(\boldsymbol{\beta}_{0:N}, \psi)$

1: Initialise: $\psi = \psi^{(0)}$.
2: **for** $i = 1, \ldots, M$ **do**
3: Sample $\boldsymbol{\beta}^{(i)}_{0:N}$ from $\pi_{\boldsymbol{\beta}_{0:N}|\mathbf{Y}_{1:N},\Psi^{(i-1)}}(\boldsymbol{\beta}_{0:N})$ via FFBS
4: **for** $h = 1, \ldots, p$ **do**
5: Sample $\psi^{(i)}_h$ from $\pi_{\Psi_h|\boldsymbol{\beta}^{(i)}_{0:N},\boldsymbol{\psi}^{(i)}_{-h},\mathbf{Y}_{1:N}}(\psi_h)$
6: where $\boldsymbol{\psi}^{(i)}_{-h} = (\psi^{(i)}_1, \ldots, \psi^{(i)}_{h-1}, \psi^{(i-1)}_{h+1}, \ldots, \psi^{(i-1)}_p)$.
7: **end for**
8: **end for**

8.2.2. Forecasting via FFBS

The Bayesian state-space framework allows us to obtain the forecasting distributions using MCMC samples given by

$$\begin{aligned}\pi_{\mathbf{Y}_{N+m}|\mathbf{Y}_{1:N}}(\mathbf{y}_{N+m}) &= \int_{\Omega}\int_{\mathbb{R}^{k\times m}} \pi_{\mathbf{Y}_{N+m},\boldsymbol{\beta}_{N:(N+m)},\Psi|\mathbf{Y}_{1:N}}(\mathbf{y}_{N+m},\boldsymbol{\beta}_{N:(N+m)},\boldsymbol{\psi})\, d\boldsymbol{\beta}_{N:(N+m)}\, d\boldsymbol{\psi} \\ &= \int_{\Omega}\int_{\mathbb{R}^{k\times m}} \pi_{\mathbf{Y}_{N+m}|\boldsymbol{\beta}_{N+m},\Psi}(\mathbf{y}_{N+m})\,\pi_{\boldsymbol{\beta}_{N:(N+m)}|\mathbf{Y}_{1:N},\Psi}(\boldsymbol{\beta}_{N:(N+m)})\,\pi_{\Psi|\mathbf{Y}_{1:N}}(\boldsymbol{\psi})\, d\boldsymbol{\beta}_{N:(N+m)}\, d\boldsymbol{\psi} \end{aligned} \tag{8.4}$$

where

$$\pi_{\boldsymbol{\beta}_{N:(N+m)}|\mathbf{Y}_{1:N},\Psi}(\boldsymbol{\beta}_{N:(N+m)}) = \pi_{\boldsymbol{\beta}_N|\mathbf{Y}_{1:N},\Psi}(\boldsymbol{\beta}_N)\prod_{t=N}^{N+m}\pi_{\boldsymbol{\beta}_{t+1}|\boldsymbol{\beta}_t,\Psi}(\boldsymbol{\beta}_t+1).$$

The optimal predictor of the random vector $\mathbf{Y}_{N+m}|\mathbf{Y}_{1:N}$, which minimizes the covariance of the prediction error, is the conditional expectation formulated as follows

$$\begin{aligned}\mathbf{y}_{N+m|N} &= \mathbb{E}_{\mathbf{Y}_{N+m}|\mathbf{Y}_{1:N}}\left[\mathbf{Y}_{N+m}\right] = \int_{\mathbb{R}^d}\mathbf{y}_{N+m}\,\pi_{\mathbf{Y}_{N+m}|\mathbf{Y}_{1:N}}(\mathbf{y}_{N+m})\, d\mathbf{y}_{N+m} \\ &= \int_{\mathbb{R}^d}\mathbf{y}_{N+m}\int_{\Omega}\int_{\mathbb{R}^{k\times m}}\pi_{\mathbf{Y}_{N+m},\boldsymbol{\beta}_{N:(N+m)},\Psi|\mathbf{Y}_{1:N}}(\mathbf{y}_{N+m},\boldsymbol{\beta}_{N:(N+m)},\boldsymbol{\psi})\, d\boldsymbol{\beta}_{N:(N+m)}\, d\boldsymbol{\psi}\, d\boldsymbol{\beta}_{N:(N+m)} \\ &= \mathbb{E}_{\Psi|\mathbf{Y}_{1:N}}\left[\mathbb{E}_{\boldsymbol{\beta}_{N:(N+m}|\mathbf{Y}_{1:N},\Psi}\left[\mathbb{E}_{\mathbf{Y}_{N+m}|\boldsymbol{\beta}_{N+m},\Psi}\left[\mathbf{Y}_{N+m}\right]\right]\right] \end{aligned} \tag{8.5}$$

with the variance of the estimation error given by

$$\boldsymbol{\Sigma}_{N+m|N} = \mathbb{E}_{\Psi|\mathbf{Y}_{1:N}}\left[\mathbb{E}_{\boldsymbol{\beta}_{N:(N+m}|\mathbf{Y}_{1:N},\Psi}\left[\mathbb{E}_{\mathbf{Y}_{N+m}|\boldsymbol{\beta}_{N+m},\Psi}\left[\mathbf{Y}_{N+m}^T\mathbf{Y}_{N+m}\right]\right]\right] + \mathbf{y}_{N+m|N}^T\mathbf{y}_{N+m|N} \tag{8.6}$$

By sampling recursively, we obtain the following forecasting distributions, when (i) is the number of the iteration

$$\begin{aligned}\boldsymbol{\beta}_{N+s}^{(i)} &\sim \mathcal{N}\left(\mathbf{c}^{(i)} + \boldsymbol{\beta}_{N+s-1|N}^{(i)}\mathbf{A}^{(i)}, \mathbf{R}^{(i)}\right), \\ \mathbf{Y}_{N+s}^{(i)} &\sim \mathcal{N}\left(\boldsymbol{\mu}^{(i)} + \boldsymbol{\beta}_{N+s}^{(i)}\boldsymbol{\Lambda}^{(i)}, \mathbf{Q}^{(i)}\right).\end{aligned} \tag{8.7}$$

for $s = 1,\dots m$ where $\boldsymbol{\beta}_{N+1}^{(i)} \sim \mathcal{N}\left(\mathbf{c}^{(i)} + \boldsymbol{\beta}_{N|N}^{(i)}\mathbf{A}^{(i)}, \mathbf{R}^{(i)}\right)$. Alternatively, we can use the forecasting distribution given by the Kalman Filter given in Section 7.3 and replace the static parameters by their estimates given by FFBS and realisation provided by the Gibbs sampler. Hence, we define the two types of mean square errors of prediction (MSEP) which can be used to access the accuracy of the forecasting

$$\begin{aligned}\text{MSEP}_{N,MCMC}(m) &:= \mathbb{E}_{\mathbf{Y}_{N+m}|\mathbf{Y}_{1:N}}\left[\left(\mathbf{Y}_{N+m} - \mathbb{E}_{\mathbf{Y}_{N+m}|\mathbf{Y}_{1:N}}\left[\mathbf{Y}_{N+m}\right]\right)\left(\mathbf{Y}_{N+m} - \mathbb{E}_{\mathbf{Y}_{N+m}|\mathbf{Y}_{1:N}}\left[\mathbf{Y}_{N+m}\right]\right)^T\right], \\ \text{MSEP}_{N,KF}(m) &:= \mathbb{E}_{\mathbf{Y}_{N+m}|\mathbf{Y}_{1:N},\bar{\Psi}}\left[\left(\mathbf{Y}_{N+m} - \mathbb{E}_{\mathbf{Y}_{N+m}|\mathbf{Y}_{1:N},\bar{\Psi}}\left[\mathbf{Y}_{N+m}\right]\right)\left(\mathbf{Y}_{N+m} - \mathbb{E}_{\mathbf{Y}_{N+m}|\mathbf{Y}_{1:N},\bar{\Psi}}\left[\mathbf{Y}_{N+m}\right]\right)^T\right].\end{aligned} \tag{8.8}$$

Therefore, the mean square error of prediction using MCMC distribution is denoted as $\text{MSEP}_{N,MCMC}$ whereas the mean square error of prediction using the distribution provided by

Kalman Filter is denoted by $\text{MSEP}_{N,KF}(m)$ where $\bar{\Psi}$ denotes the posterior mean of the vector of parameters Ψ.

9. In-sample Model Assessment of the Hybrid Factor Model

Given the state-space representation of the HFM model in Equation (5.2), we have multiple choices for the model calibration. We assume the matrix $\mathbf{Q}$ to be diagonal whereas the matrix $\tilde{\mathbf{R}}$ to be unrestricted, that is non-diagonal. The assumption of a diagonal $\mathbf{Q}$ matrix, what implies mutually uncorrelated deviations of elements of $\mathbf{Y}_t$, is quite common as one may expect the measurement errors to be independent across components of the response variable. The HFM models allow accounting for serial correlation between components of the latent vector. The form of the transition matrix $\tilde{\mathbf{A}}$ and the covariance matrix $\tilde{\mathbf{R}}$ specifies the strength of the linear dependencies between the elements of $\tilde{\boldsymbol{\beta}}_t$. Therefore, in this study we examine the following collection of models

Model 1: diagonal $\mathbf{Q}, \tilde{\mathbf{A}}, \tilde{\mathbf{R}}$
Model 2: diagonal $\mathbf{Q}, \tilde{\mathbf{A}}$,
Model 3: diagonal $\mathbf{Q}, \tilde{\mathbf{R}}$,
Model 4: diagonal $\mathbf{Q}$.

In addition, we examine two covariance structures: heterogeneous ('hete') and homogeneous ('homo'). In addition we assume that the subvectors of $\tilde{\boldsymbol{\beta}}_t$, the latent vectors $\boldsymbol{\beta}_t$ and $\boldsymbol{\varrho}_t$, are orthogonal and hence the matrices $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{R}}$ are block-diagonal.

We use as a model selection measures, the log likelihood function of the Kalman Filter in Equation (7.15) , $l(\Psi; \mathbf{y}_{1:N})$, for Ψ being the estimators of the static parameters. However, given various model assumptions on the linear state-space model in Equation (5.2), the vector of static parameters might vary in length and the likelihood function stops being a reliable measure of comparison between model structures.

Therefore, in order to avoid the possibility of selecting a model which overfits the data, we use a criterion which penalises the number of parameters. In frequentist setting, a natural candidate for such a measure is Akaike Information Criterion (AIC) introduced in Akaike (1973) and (Akaike 1974), a popular estimator of the quality of statistical models founded on Information Theory and having the Bayesian interpretation

$$AIC(\Psi) = 2|\Psi| - 2l(\Psi; \mathbf{y}_{1:N}), \tag{9.1}$$

where $|\Psi|$ denotes the number of free parameters in the model to estimate. The preferred model is selected as the one which minimises AIC.

In Bayesian setting, an equivalent measure to AIC is the deviance of the model (DIC). The device information criterion is a popular measure in Bayesian setting which trades off model fit against its complexity (the effective of parameters) as introduced in (Spiegelhalter et al. 2002). Among the various versions of DIC we decided to use so-called conditional DIC which treats the latent states as parameters when calculating the conditional loglikelihood, for details please refer

to (Celeux et al. 2006). The conditional log-likelihood is given by the formulation of the of marginal likelihood function of the Kalman Filter, $l(\Psi, \boldsymbol{\beta}_{0:N}; \mathbf{y}_{1:N})$. We define the deviance of the model as follows

$$D(\Psi, \boldsymbol{\beta}_{0:N}) := -2l(\Psi; \mathbf{y}_{1:N}).$$

and the effective number of parameters as

$$p_D := \bar{D}(\Psi, \boldsymbol{\beta}_{0:N}) - D(\bar{\Psi}, \bar{\boldsymbol{\beta}}_{0:N})$$

where $\bar{D}(\Psi, \boldsymbol{\beta}_{0:N})$ is a mean of deviance over different samples of the vector $\Psi, \boldsymbol{\beta}_{0:N}$ from FFBS algorithm and $D(\bar{\Psi}, \bar{\boldsymbol{\beta}}_{0:N})$ is a deviance of the posteriori mean of the vector of parameters $\Psi, \boldsymbol{\beta}_{0:N}$. The DIC then is defined as

$$DIC := \bar{D}(\Psi, \boldsymbol{\beta}_{0:N}) + p_D = 2\bar{D}(\Psi, \boldsymbol{\beta}_{0:N}) - D(\bar{\Psi}, \bar{\boldsymbol{\beta}}_{0:N}) \tag{9.2}$$

which can be calculated using the MCMC samples.

In addition to the mentioned indicators of the models' in-sample performance, we calculate the mean square errors (MSE) defined as the difference between the observed data, $\mathbf{y}_t$, and its in-sample one-step ahead prediction given by Kalman Filter, $\mathbf{y}_{t|t-1} = \mathbb{E}_{\mathbf{Y}_t|\mathbf{Y}_{1:(t-1)},\Psi}[\mathbf{y}_t]$ defined in Equation (7.2). Therefore, we define $\mathbf{e}_t := \mathbf{y}_t - \mathbf{y}_{t|t-1}$ and

$$\text{MSE}(\bar{\Psi}) := \frac{1}{N}\sum_{t=1}^{N} \mathbb{E}_{\mathbf{Y}_t|\mathbf{Y}_{1:(t-1)},\bar{\Psi}}\left[\mathbf{e}_t\mathbf{e}_t^T\right] \tag{9.3}$$

where $\bar{\Psi}$ is an estimator of Ψ.

10. Monte Carlo Methods for State-Space Models

The following chapter intends to provide the reader with a brief introduction to the class of approximation methods that can be used when no analytical expression of filtering distributions is available. The majority of estimation problems for state space models does not admit a closed-form solution as it is achieved for the linear Gaussian models in Chapter 6. The estimation becomes even more difficult when the Bayesian inference is considered, and a modelling framework accounts for the uncertainty of static parameters. The evaluation of the filtering recursions from Equations (6.2)-(6.3), which are formulated in terms of integrals, can be addressed by a wide range of numerical integration techniques. In the next sections, we review a few approaches which are especially relevant to the estimation problems considered in our research.

The methods, which are discussed in the next section find an approximation to a integration problem defined as

$$\mathbb{E}_{\mathbf{Y}|\Psi}\left[\phi(\mathbf{Y})\right] = \int_{\mathcal{Y}} \phi(\mathbf{y})\ \pi_{\mathbf{Y}|\Psi}(\mathbf{y})\ d\mathbf{y}$$

where $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^k$, for $k \in \mathbb{N}$, $\pi_{\mathbf{Y}|\Psi}$ is assumed to be a probability density function of the d-dimensional random vector $\mathbf{Y} \in \mathcal{Y} \subseteq \mathbb{R}^d$, and Ψ represents the set of static parameters.

We begin with an introduction to the classical concept of numerical approximation via sampling techniques and discuss the class of Monte Carlo methods which reformulate a generic integration problem to the probabilistic setting. The stochastic simulation methods use samples from a posterior distribution to summarise information contained in a density. Next, our attention is moved to the class of Monte Carlo Markov Chain (MCMC) methods that apply to the filtering problems when the realisations form a posterior distribution cannot be directly simulated due to its complexity or intractability. In order to obtain in-depth understating of the discussed approaches, the reader is referred to the studies exemplified in Chapter 1 and the books of (Tanizaki 1993), (Rober and Casella 2004), (Gamerman and Lopes 2006) (Brooks et al. 2011) or (Durbin and Koopman 2012), among many other insightful works. We remark that we review the following sampling techniques considering only continuous distributions of random vectors.

10.1. Classical Monte Carlo Problem

The simple MC methodology is based on the idea of generating identical and independent samples from a target distribution which are then used to approximate required expectations defined by integrals. Let us denote $\mathbf{y}_{1:N}$ as N i.i.d samples from the target distribution $\pi_{\mathbf{Y}|\Psi}$ which represent the N realisations of the random vector $\mathbf{Y}$. Provided that the expectation $\mathbb{E}_{\mathbf{Y}|\Psi}\left[\phi(\mathbf{Y})\right]$

exists and is finite, under the Strong Law of Large Numbers, see Shiryaev (1996), we obtain the following convergence

$$\lim_{N\to\infty_+} \frac{1}{N}\sum_{i=1}^{N} \phi(\mathbf{y}_i) \xrightarrow{a.s} \mathbb{E}_{\mathbf{Y}|\Psi}\big[\phi(\mathbf{Y})\big] \tag{10.1}$$

In addition, if the variance of $\phi(\mathbf{Y})$ under the target distribution is finite, the Central Limit Theorem (CLT, Shiryaev (1996)) ensures that the approximation $\frac{1}{N}\sum_{i=1}^{N}\phi(\mathbf{y}_i)$ converges in distribution to the true expectation, that is

$$\lim_{N\to\infty_+} \sqrt{N}\Big(\frac{1}{N}\sum_{i=1}^{N}\phi(\mathbf{y}_i) - \mathbb{E}_{\mathbf{Y}|\Psi}\big[\phi(\mathbf{Y})\big]\Big) \xrightarrow{d} \mathcal{N}\Big(\mathbf{0},\ \mathbf{Var}\big[\phi(\mathbf{Y})\big]\Big). \tag{10.2}$$

10.1.1. Importance Sampling

If it is difficult or too costly to sample from the target distribution $\pi_{\mathbf{Y}|\Psi}$, the approximation of $\mathbb{E}_{\mathbf{Y}|\Psi}\big[\phi(\mathbf{Y})\big]$ can be achieved by the importance sampling as introduced in (Marshall 1956). The method reformulates the integration problem by deriving the formula which includes the auxiliary function $\tilde{\pi}_{\mathbf{Y}|\tilde{\Psi}}(\mathbf{y})$. The proposal function acts as a density which generates the samples $\mathbf{y}_i$, that is

$$\mathbb{E}_{\mathbf{Y}|\Psi}\big[\phi(\mathbf{Y})\big] = \int_{\mathcal{Y}} \phi(\mathbf{y})\ \pi_{\mathbf{Y}|\Psi}(\mathbf{y})\ d\mathbf{y} = \int_{\mathcal{Y}} \phi(\mathbf{y})\ \frac{\pi_{\mathbf{Y}|\Psi}(\mathbf{y})}{\tilde{\pi}_{\mathbf{Y}|\tilde{\Psi}}(\mathbf{y})}\ \tilde{\pi}_{\mathbf{Y}|\tilde{\Psi}}(\mathbf{y})\ d\mathbf{y} = \mathbb{E}_{\tilde{\mathbf{Y}}|\tilde{\Psi}}\big[w(\tilde{\mathbf{Y}})\phi(\tilde{\mathbf{Y}})\big]$$

where $\tilde{\mathbf{Y}}$ follows distribution defined by the density function $\tilde{\pi}_{\mathbf{Y}|\Psi}(\mathbf{y})$ and $w(\mathbf{y}) = \frac{\pi_{\mathbf{Y}|\Psi}(\mathbf{y})}{\tilde{\pi}_{\mathbf{Y}|\tilde{\Psi}}(\mathbf{y})}$ is defined as an importance weight. Therefore, importance sampling use the weights w in order to compensate for the fact that the sample are generated from the auxiliary density function ie. to correct for the bias that would otherwise arise in the expectation. If the importance distribution is not renormalized, that is, it is known only up to proportionality, the obtained weights do not sum to one and require re-normalisation.

The obtained estimator $\frac{1}{N}\sum_{i=1}^{N}\frac{\psi(\mathbf{y}_i)\pi_{\mathbf{Y}|\Psi}(\mathbf{y}_i)}{\tilde{\pi}_{\mathbf{Y}|\tilde{\Psi}}(\mathbf{y}_i)}$ enjoys the properties of the simple MC method: it is unbiased and can obtain the convergence according the Strong Law of Large Number and Central Limit Theorem. The main condition of the convergence in distribution is a finite variance of $\frac{\psi(\mathbf{Y}_i)\pi_{\mathbf{Y}|\Psi}(\mathbf{Y})}{\tilde{\pi}_{\mathbf{Y}|\tilde{\Psi}}(\mathbf{Y})}$ under the distribution defined by density $\tilde{\pi}_{\mathbf{Y}|\tilde{\Psi}}(\mathbf{y})$. Therefore, the choice of the proposal function is crucial to ensure that the variance, formulated as

$$\mathbf{Var}\left(\frac{\psi(\mathbf{Y}_i)\pi_{\mathbf{Y}|\Psi}(\mathbf{Y})}{\tilde{\pi}_{\mathbf{Y}|\tilde{\Psi}}(\mathbf{Y})}\right) = \int_{\mathcal{Y}} \phi(\mathbf{y})^2\ \frac{\pi_{\mathbf{Y}|\Psi}(\mathbf{y})^2}{\tilde{\pi}_{\mathbf{Y}|\tilde{\Psi}}(\mathbf{y})}\ d\mathbf{y} - \left(\int_{\mathcal{Y}} \phi(\mathbf{y})\ \pi_{\mathbf{Y}|\Psi}(\mathbf{y})\ d\mathbf{y}\right)^2,$$

is finite.

The discussed in Chapter 1 remarks on the main difficulty of the importance sampling, known as the sample degeneracy or sample impoverishment. Especially in high dimensional problems, the importance sampling becomes more expensive since it samples from the whole domain without learning to target the region of importance. It becomes difficult to to find an importance sampling distribution in multiple dimensions which ensures that the ratio of the weights is not degenerated. As commented, a small distortion in mass with respect to the true distribution in the ratio causes

discrepancy in the weights and results in higher variance of the estimates when one integrates those samples.

10.1.2. Rejection Sampling

Another stochastic approximation approach which addresses the difficulty of sampling from the true distribution, is rejection sampling. The basic idea behind this methodology is to use an auxiliary distribution $\tilde{\pi}_{\mathbf{Y}|\Psi}(\mathbf{y})$ which is convenient for sampling and satisfies $\pi_{\mathbf{Y}|\Psi}(\mathbf{y}) \leq M\, \tilde{\pi}_{\mathbf{Y}|\tilde{\Psi}}(\mathbf{y})$ for $1 < M < +\infty$. The procedure rejects the samples from $\tilde{\pi}_{\mathbf{Y}|\Psi}(\mathbf{y})$ which are unlikely under $\pi_{\mathbf{Y}|\Psi}(\mathbf{y})$ with the acceptance probability equal to $\frac{\pi_{\mathbf{Y}|\Psi}(\mathbf{y})}{M\tilde{\pi}_{\mathbf{Y}|\tilde{\Psi}}(\mathbf{y})}$. Hence, in order to correct for the fact that we sample from an auxiliary distribution, we reject some of the values proposed by $\tilde{\pi}_{\mathbf{Y}|\tilde{\Psi}}(\mathbf{y})$. Given the threshold determined by M, one can consider various choices of the density function $\tilde{\pi}_{\mathbf{Y}|\Psi}(\mathbf{y})$.

The method is justified by the Fundamental Theorem of Simulation which states that the simulation of $\mathbf{Y}$ under the density function $\pi_{\mathbf{Y}|\Psi}(\mathbf{y})$ is equivalent to generating uniformly the pair

$$(\mathbf{Y}, S) \sim \mathcal{U}\Big((\mathbf{y}, s) : 0 < s < \pi_{\mathbf{Y}|\Psi}(\mathbf{y})\Big).$$

The equivalence can be shown by remarking that the probability of simulating $\mathbf{Y} \leq \mathbf{y}$ under $\tilde{\pi}_{\mathbf{Y}|\tilde{\Psi}}(\mathbf{y})$ conditional on the the rejection rate $\frac{\pi_{\mathbf{Y}|\Psi}(\mathbf{y})}{M\tilde{\pi}_{\mathbf{Y}|\tilde{\Psi}}(\mathbf{y})}$ is the same as the probability of $\mathbf{Y} \leq \mathbf{y}$ when $\mathbf{Y}$ follows $\pi_{\mathbf{Y}|\Psi}(\mathbf{y})$, since

$$\begin{aligned}\mathbb{P}_{\tilde{\pi}}\left(\mathbf{Y} \leq \mathbf{y} |\, S \leq \frac{\pi_{\mathbf{Y}|\Psi}(\mathbf{y})}{M\tilde{\pi}_{\mathbf{Y}|\tilde{\Psi}}(\mathbf{y})}\right) &= \frac{\mathbb{P}_{\tilde{\pi}}\left(\mathbf{Y} \leq \mathbf{y},\, S \leq \frac{\pi_{\mathbf{Y}|\Psi}(\mathbf{Y})}{M\tilde{\pi}_{\mathbf{Y}|\tilde{\Psi}}(\mathbf{Y})}\right)}{\mathbb{P}\left(S \leq \frac{\pi_{\mathbf{Y}|\Psi}(\mathbf{Y})}{M\tilde{\pi}_{\mathbf{Y}|\tilde{\Psi}}(\mathbf{Y})}\right)} = M \int_{-\infty}^{\mathbf{y}} \int_{0}^{\frac{\pi_{\mathbf{Y}|\Psi}(\mathbf{y})}{M\tilde{\pi}_{\mathbf{Y}|\tilde{\Psi}}(\mathbf{y})}} du\; \tilde{\pi}_{\mathbf{Y}|\tilde{\Psi}}(\mathbf{y})\, d\mathbf{y} \\ &= \int_{-\infty}^{\mathbf{y}} \pi_{\mathbf{Y}|\Psi}(\mathbf{y})\, d\mathbf{y} = \mathbb{P}_{\pi}(\mathbf{Y} \leq \mathbf{y})\end{aligned}$$

where we used the fact that the probability of acceptance conditioned on a sample's value equals to

$$\mathbb{P}\left(S \leq \frac{\pi_{\mathbf{Y}|\Psi}(\mathbf{Y})}{M\tilde{\pi}_{\mathbf{Y}|\tilde{\Psi}}(\mathbf{Y})} \Big| \mathbf{Y} \leq \mathbf{y}\right) = \frac{\pi_{\mathbf{Y}|\Psi}(\mathbf{y})}{M\tilde{\pi}_{\mathbf{Y}|\tilde{\Psi}}(\mathbf{y})},$$

and the unconditional acceptance probability depends only on the value of M, since

$$\mathbb{P}\left(S \leq \frac{\pi_{\mathbf{Y}|\Psi}(\mathbf{Y})}{M\tilde{\pi}_{\mathbf{Y}|\tilde{\Psi}}(\mathbf{Y})}\right) = \int_{\mathbb{R}^d} \int_{0}^{\frac{\pi_{\mathbf{Y}|\Psi}(\mathbf{y})}{M\tilde{\pi}_{\mathbf{Y}|\tilde{\Psi}}(\mathbf{y})}} du\; \tilde{\pi}_{\mathbf{Y}|\Psi}(\mathbf{y})\, d\mathbf{y} = \mathbb{E}_{\tilde{\pi}}\left[\frac{\pi_{\mathbf{Y}|\Psi}(\mathbf{Y})}{M\tilde{\pi}_{\mathbf{Y}|\tilde{\Psi}}(\mathbf{Y})}\right] = \frac{1}{M},$$

where $\mathbb{E}_{\tilde{\pi}}$ denotes an expectation operator with respect to the probability density function $\tilde{\pi}$. The last identity is true since $\tilde{\pi}_{\mathbf{Y}|\tilde{\Psi}}(\mathbf{y})$ includes support of $\pi_{\mathbf{Y}|\Psi}(\mathbf{y})$.

Therefore, rejection sampling is a generic method to simulate from any density which is known up to multiplicative factor since its normalizing constant does not need to be known. This is a particularly appealing property of the sampling methods for any Bayesian calculations as they rely on proportionality to the a posteriori density. This link is discussed in more details in the next part of the chapter. The algorithm of the sampling method is given below

Algorithm 3 Rejection sampling algorithm for $\pi_{\mathbf{Y}|\Psi}(\mathbf{y})$

1: Initialise: specify $\tilde{\pi}_{\mathbf{Y}|\tilde{\Psi}}(\mathbf{y})$ such that $\pi_{\mathbf{Y}|\Psi}(\mathbf{y}) \leq M\tilde{\pi}_{\mathbf{Y}|\tilde{\Psi}}(\mathbf{y})$.
2: **for** $i = 1, \ldots, N$ **do**
3: Sample $\mathbf{Y}_i$ from the proposal distribution $\tilde{\pi}_{\mathbf{Y}|\tilde{\Psi}}(\mathbf{y})$
4: Sample $S \sim \mathcal{U}[0,1]$
5: Accept $\mathbf{y}_i$ if $s \leq \frac{\pi_{\mathbf{Y}|\Psi}(\mathbf{y}_i)}{M\tilde{\pi}_{\mathbf{Y}|\tilde{\Psi}}(\mathbf{y}_i)}$, otherwise $\mathbf{0}$
6: **end for**

As remarked in the Chapter 1, the classical rejection sampling algorithm might turn out to be computationally expensive even for one dimensional problems since the procedure requires calculating the ratio between the proposal and the true density at each iteration. Therefore, it is usually not recommended to use this sampling technique for multiple dimensional problems.

10.2. Markov Chain Monte Carlo

Class of Markov Chain Monte Carlo, which originates from (Metropolis and Ulam 1949), addresses the difficulty of sampling from a density which is too complex, computationally demanding or intractable. On the contrary to the simulation approaches discussed before, the methods do not generate identical and independent realisations from the underlying distribution but construct Markov Chains which converge a true distribution with a sufficient number of steps. The reader is advised to refer to (Fishman 1996), (Rober and Casella 2004),(Gamerman and Lopes 2006), (Särkkä 2013) or (Brooks et al. 2011), among many other, for broader discussion on sophisticated technicalities of MCMC.

Let us denote $\mathbf{Y}_{0:N}$ as a sequence of random variables with realisation $\mathbf{y}_{1:N}$ that is an element of a Markov Chain characterized by the transition kernel $K(\mathbf{y}_{n+1}, \mathbf{y}_n)$, which defines the distribution of $\mathbf{Y}_{n+1}$ conditioned on $\mathbf{Y}_n$. We consider only the chains which are time homogeneous, what implies that the distribution of $(\mathbf{Y}_{n_1}, \ldots \mathbf{Y}_{n_k})$ given the initial state $\mathbf{y}_{n_0}$ is the same as the distribution of $(\mathbf{Y}_{n_1-n_0}, \ldots \mathbf{Y}_{n_k-n_0})$ given $\mathbf{y}_0$ for every $(k+1)$-uplet $n_0 \leq n_1 \leq \ldots \leq n_k$. Consequently, if an initial distribution of state is known, the construction of the Markov chain is entirely defined by the transition (conditional) probability from one state to the another, given the kernel K.

Also, we assume that the considered chains are ergodic. The ergodicity of the chain ensures that irrespectively to the choice of initial distribution, the marginal distribution of a random variable, which belongs the Markov Chain, converges to the invariant distribution, called the stationary distribution. In our discussion, the invariant distribution of the chain is denoted by $\pi_{\mathbf{Y}|\Psi}(\mathbf{y})$.

If the kernel K allows for free moves over all possible states of $\mathbf{Y}_n$, it enjoys strong stability given by existence of the stationary probability distribution $\pi_{\mathbf{Y}|\Psi}(\mathbf{y})$. Also, it ensures that the chain is recurrent, what means that the average number of time a Markov chain revisits an arbitrary state, is infinite, or even Harris recurrent - the chain reaches its stationary distribution regardless of the starting point. The latter is an important characteristic for MCMC.

The convergence property of Markov chains to their stationary distribution implies that the estimator provided by MCMC coverages to the expectation of interests almost surely, as in

Equation (10.1). When Markov Chain is reversible, that is, its kernel and the invariant distribution satisfies the following detailed balance equation

$$\pi_{\mathbf{Y}|\Psi}(\mathbf{y}_n)K(\mathbf{y}_{n+1},\mathbf{y}_n) = \pi_{\mathbf{Y}|\Psi}(\mathbf{y}_{n+1})K(\mathbf{y}_n,\mathbf{y}_{n+1}), \tag{10.3}$$

and consequently,

$$\int_{\mathcal{Y}\subseteq\mathbb{R}^d} \pi_{\mathbf{Y}|\Psi}(\mathbf{y}_n)\, d\mathbf{y}_n = \int_{\mathbb{R}^d} \pi_{\mathbf{Y}|\Psi}(\mathbf{y}_{n+1})K(\mathbf{y}_n,\mathbf{y}_{n+1}\in\mathcal{Y})d\mathbf{y}_{n+1}. \tag{10.4}$$

The balanced equalities implies that the Central Limit Theorem holds and the estimation enjoys the converges in distribution as in Equation (10.2).

10.2.1. Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm of (Metropolis et al. 1953) and (Hastings 1970) is one of the most popular MCMC frameworks. We briefly discuss the main idea behind the framework following the insightful explanation which is given (Chib and Greenberg 1995).

Let us denote by $\mathbf{x}$ and $\mathbf{y}$ the realisations of the random vectors $\mathbf{Y}_n$ and $\mathbf{Y}_{n+1}$, respectively. The main idea behind the Metropolis-Hastings algorithms in the formulation of a transition kernel of the underlying Markov chain. We would like to impose the following parametric form of a transition kernel K

$$K(\mathbf{x},\mathbf{y}) = p(\mathbf{x},\mathbf{y}) + \mathbb{I}_{\{\mathbf{y}=\mathbf{x}\}} r(\mathbf{x}). \tag{10.5}$$

Then, the function $p(\mathbf{x},\mathbf{y})$ would reflect the probability of changing the state from $\mathbf{x}$ to $\mathbf{y}$, whereas $r(\mathbf{x}) = 1 - \int p(\mathbf{x},\mathbf{y})\, d\mathbf{z}$ would correspond to the probability of staying in the current state. As shown in (Tierney 1994), if the function p satisfies the balance condition in Equation (10.3), that is,

$$\pi_{\mathbf{Y}|\Psi}(\mathbf{x})p(\mathbf{x},\mathbf{y}) = \pi_{\mathbf{Y}|\Psi}(\mathbf{y})p(\mathbf{y},\mathbf{x}), \tag{10.6}$$

it implies that $\pi_{\mathbf{Y}|\Psi}$ is an invariant density of the kernel $K(\mathbf{x},\mathbf{y})$. Therefore, we can simplify the problem of finding the kernel $K(\mathbf{x},\mathbf{y})$ to searching for the function $p(\mathbf{x},\mathbf{y})$.

However, since $p(\mathbf{x},\mathbf{y})$ is unknown, the Metropolis-Hastings algorithms uses a candidate function, denoted by $q(\mathbf{x},\mathbf{y})$, such that $\int_{\mathbb{R}^d} q(\mathbf{x},\mathbf{y})\, d\mathbf{y} = 1$. If the chosen function already satisfies the condition from Equation (10.6), we have a perfect candidate for $p(\mathbf{x},\mathbf{y})$. However, in majority of cases, the moves of the process characterised by the function $q(\mathbf{x},\mathbf{y})$ are not balanced, that is

$$\pi_{\mathbf{Y}|\Psi}(\mathbf{x})q(\mathbf{x},\mathbf{y}) > \pi_{\mathbf{Y}|\Psi}(\mathbf{y})q(\mathbf{y},\mathbf{x}), \tag{10.7}$$

In order to correct the arising inequality, the algorithm introduces the probability density function, denoted by $\rho(\mathbf{x},\mathbf{y}) < 1$, which decides if the Markov Chain makes a transition to the new state or it remains in the current state. When the sample in the nth iteration conditioned on the $n-1$ state is drawn, the algorithm assumes additional rejection step of moving to the new state similarly to

the rejection sampling. However, this acceptance-rejection procedure is different from the rejection sampling discussed before, where the rejected states are removed from the sample set.

As a result, the final form of the transition probability of the Metropolis-Hastings algorithm is given by $q(\mathbf{x}, \mathbf{y})\rho(\mathbf{x}, \mathbf{y})$ and the inequality in Equation (10.7), is reformulated as

$$\pi_{\mathbf{Y}|\Psi}(\mathbf{x})q(\mathbf{x}, \mathbf{y})\rho(\mathbf{x}, \mathbf{y}) > \pi_{\mathbf{Y}|\Psi}(\mathbf{y})q(\mathbf{y}, \mathbf{x})\rho(\mathbf{y}, \mathbf{x}).$$

In order to satisfy the reversibility condition, and balance the above inequality, the probability $\rho(\mathbf{y}, \mathbf{x})$ should be as large as possible. By setting it to 1, we obtain the formula for the probability of move

$$\rho(\mathbf{x}, \mathbf{y}) = \min\left\{\frac{\pi_{\mathbf{Y}|\Psi}(\mathbf{y})}{\pi_{\mathbf{Y}|\Psi}(\mathbf{x})}\frac{K(\mathbf{x}, \mathbf{y})}{K(\mathbf{y}, \mathbf{x})}, 1\right\} \tag{10.8}$$

Therefore, the rejection probability is a crucial element of the algorithm which impacts on the construction of the stationary distribution. The transition kernel K in the Metropolis-Hastings algorithm is formulated as

$$K(\mathbf{x}, \mathbf{y}) = q(\mathbf{x}, \mathbf{y})\rho(\mathbf{x}, \mathbf{y}) + \mathbb{I}_{\{\mathbf{y}=\mathbf{x}\}}\left[1 - \int q(\mathbf{x}, \mathbf{z})\rho(\mathbf{x}, \mathbf{z})\, d\mathbf{z}\right]. \tag{10.9}$$

The proposal value of the chain in nth step condition on the value in the previous step, is drawn using the conditional probability $q(\mathbf{Y}^{(n-1)}, \mathbf{Y}^{(n)})$. After sufficiently large length of the chain, the sequence of samples, $\mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(N)}$ converges to its stationary distribution $\pi_{\mathbf{Y}|\Psi}(\mathbf{y})$ since the transition kernel K satisfies the condition in Equation (10.3), that is

$$\begin{aligned}
&\pi_{\mathbf{Y}|\Psi}(\mathbf{x})q(\mathbf{x}, \mathbf{y})\rho(\mathbf{x}, \mathbf{y}) = \pi_{\mathbf{Y}|\Psi}(\mathbf{y})q(\mathbf{y}, \mathbf{x})\rho(\mathbf{y}, \mathbf{x}) \\
&\pi_{\mathbf{Y}|\Psi}(\mathbf{y})\mathbb{I}_{\{\mathbf{x}=\mathbf{y}\}}\left[1 - \int q(\mathbf{y}, \mathbf{z})\rho(\mathbf{y}, \mathbf{z})\, d\mathbf{z}\right] = \pi_{\mathbf{Y}|\Psi}(\mathbf{x})\mathbb{I}_{\{\mathbf{y}=\mathbf{x}\}}\left[1 - \int q(\mathbf{x}, \mathbf{z})\rho(\mathbf{x}, \mathbf{z})\, d\mathbf{z}\right]
\end{aligned}$$

and the detailed balance condition in Equation (10.3) holds. The Metropolis-Hastings algorithm is given below

Algorithm 4 Metropolis-Hastings algorithm sampling for $\pi_{\mathbf{Y}|\Psi}(\mathbf{y})$

1: Initialise: $\mathbf{Y}^{(0)} = \mathbf{y}^{(0)}$.
2: **for** $n = 1, \ldots, N$ **do**
3: Generate $\mathbf{Y}^{(n)}$ from the proposition distribution $q(\mathbf{Y}^{(n-1)} = \mathbf{y}^{(n-1)}, \mathbf{Y}^{(n)})$
4: Generate $S \sim \mathcal{U}(0, 1)$
5: Take

$$\mathbf{x}^{(n)} = \begin{cases} \mathbf{y}^{(n)} & \text{if } U \leq \rho(\mathbf{y}^{(n-1)}, \mathbf{y}^{(n)}) \\ \mathbf{y}^{(n-1)} & \text{otherwise} \end{cases}$$

6: Set $\mathbf{y}^{(n)} = \mathbf{x}^{(n)}$
7: **end for**

10.2.2. Gibbs Sampler Algorithm

A special case of the Metropolis-Hastings algorithm is the Gibbs sampler introduced by (Geman and German 1984) which is by construction a multidimensional framework. Its interpretation regarding the Metropolis-Hastings algorithms is studied in (Chapter 7.1, Casella and Robert (1996)). We recommend the reader to follow the work of (Casella and George 1992) or (Chapter 7, Casella and Robert (1996)) for an insightful explanation of the algorithm and its properties.

The distinguishing feature of the Gibbs Sampler is that the underlying Markov Chain is constructed from a sequence of conditional densities defined by the target distribution. Therefore, the sampler is especially applicable when the joint distribution is not known, but the formulations of the considered model allow one to construct conditional probabilities. The Gibbs sampler may be viewed as a particular Metropolis-Hastings algorithm that uses the conditional distributions as proposal distributions. The main benefit of using the Gibbs sampler is that it does not require specification of the proposal distribution since known conditional distributions express it. Hence, the framework does not rely on any free parameters and a need of the optimisation or tunning.

Each iteration of the Gibbs sampler is a cycle which goes through the subvectors of random vector drawing each subset of its element conditioned on the value of all the others. Given the d-dimensional random vector $\mathbf{Y}$ and its realisation $\mathbf{y}$, let us denote $\mathbf{Y}_{-i}$ and $\mathbf{y}_{-i}$ as $(d-1)$-dimensional subvectors of $\mathbf{Y}$ and $\mathbf{y}$, respectively, which do not contain the ith element, that is $\mathbf{Y}_{-i} = \left[Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_d\right]$ and $\mathbf{y}_{-i} = \left[y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_d\right]$. The Gibbs sampler is based on the idea of using, so called full conditionals, the distribution of $Y_i|\mathbf{Y}_{-i}, \Psi$ with the density function $\pi_{Y_i|\mathbf{Y}_{-i},\Psi}(y)$. Given the fact that the target density $\pi_{\mathbf{Y}|\Psi}(\mathbf{y})$ can be formulated as follows

$$
\begin{aligned}
\pi_{\mathbf{Y}|\Psi}(\mathbf{y}) &= \pi_{Y_1|\mathbf{Y}_{-1},\Psi}(y_1)\pi_{\mathbf{Y}_{-1}|\Psi}(\mathbf{y}_{-1}) = \pi_{Y_1|\mathbf{Y}_{-1},\Psi}(y_1)\int_{\mathbb{R}} \pi_{\mathbf{Y}|\Psi}(\tilde{y}_1, \mathbf{y}_{2:d})\, d\tilde{y}_1 \\
&= \pi_{Y_1|\mathbf{Y}_{-1};\Psi}(y_1)\int_{\mathbb{R}} \pi_{Y_2|\mathbf{Y}_{-2};Y_1=\tilde{y}_1,\Psi}(y_2)\pi_{\mathbf{Y}_{-2}|\Psi}(\tilde{y}_1, \mathbf{y}_{3:d})\, d\tilde{y}_1 \\
&= \pi_{Y_1|\mathbf{Y}_{-1};\Psi}(y_1)\int_{\mathbb{R}} \pi_{Y_2|\mathbf{Y}_{-2};Y_1=\tilde{y}_1,\Psi}(y_2)\int_{\mathbb{R}} \pi_{Y_3|\mathbf{Y}_{-3};\mathbf{Y}_{1:2}=\tilde{\mathbf{y}}_{1:2},\Psi}(y_3)\pi_{\mathbf{Y}_{-3}|\Psi}(\tilde{\mathbf{y}}_{1:2}, \mathbf{y}_{4:d})\, d\tilde{\mathbf{y}}_{1:2} \\
&\vdots \\
&= \pi_{Y_1|\mathbf{Y}_{-1};\Psi}(y_1)\int_{\mathbb{R}^d} \prod_{i=2}^{d} \pi_{Y_i|\mathbf{Y}_{-i};\mathbf{Y}_{1:i-1}=\tilde{\mathbf{y}}_{1:(i-1)},\Psi}(y_i)\, d\tilde{\mathbf{y}}_{1:d} \qquad (10.10)
\end{aligned}
$$

it is easy to see that the expectation $\mathbb{E}_{\mathbf{Y}|\Psi}\left[\phi(\mathbf{Y})\right]$ can be specified using full conditionals of $Y_i|\mathbf{Y}_{-i}, \Psi$. Given above justification, the transition kernel of the Gibbs sampler has the following form

$$
K(\mathbf{x}, \mathbf{y}) = \pi_{Y_j|\mathbf{Y}_{(j+1):d}=\mathbf{x}_{(j+1):d},\mathbf{Y}_{1:j-1}=\mathbf{x}_{1:(j-1)},\Psi}(y_j)\prod_{i\in\mathcal{I}_j} \pi_{Y_i|\mathbf{Y}_{j:i-1}=\mathbf{y}_{j:(i-1)},\mathbf{Y}_{(i+1):d}=\mathbf{x}_{(i+1):d},\mathbf{Y}_{1:(j-1)}=\mathbf{x}_{1:(j-1)},\Psi}(y_i)
$$

where the set $\mathcal{I}_j = \{j+1, \ldots, d, 1, \ldots, j-1\}$ specifies the order of the Gibbs sampler steps which start from the element jth of the vector $\mathbf{Y}$. The above decomposition which defines the Gibbs sampler kernel does not have to start necessary for the first element of the vector. The kernel K

satisfies the balance equation under the Hammersley-Clifford conditions, see (Theorem 7.1.20, Rober and Casella (2004)).

The Gibbs sampler requires a priori knowledge about properties of the joint distribution function $\pi_{\mathbf{Y}|\Psi}(\mathbf{y})$ and is especially applicable to the problems which allow sampling from conditional distributions easily. Also, due to the construction of the kernel, the acceptance probability of the algorithm is equal to 1. It impacts on the length of the chain, its convergence and consequently the properties of the estimates. The steps of the Gibbs sampler are given in Algorithm 5.

Algorithm 5 Gibbs Sampler algorithm sampling for $\pi_{\mathbf{Y}|\Psi}(\mathbf{y})$

1: Initialise: $\mathbf{Y}^{(0)} = \mathbf{y}^{(0)}$.
2: **for** $n = 1, \ldots, N$ **do**
3: Sample $Y_j^{(n)} | Y_{j+1}^{(n-1)}, \ldots, Y_d^{(t-1)}, Y_d^{(n-1)}, Y_1^{(n-1)}, \ldots, Y_{j-1}^{(n-1)}, \Psi$
4: Sample $Y_{j+1}^{(n)} | Y_j^{(n)}, Y_{j+2}^{(n-1)}, \ldots, Y_d^{(n-1)}, Y_1^{(n-1)}, \ldots, Y_{j-1}^{(n-1)} \Psi$
5: $\vdots$
6: Sample $Y_i^{(n)} | Y_j^{(n)}, \ldots, Y_{i-1}^{(n)}, Y_{i+1}^{(n-1)}, \ldots, Y_d^{(n-1)}, Y_1^{(n-1)}, \ldots, Y_{j-1}^{(n-1)} \Psi$
7: $\vdots$
8: Sample $Y_{j-1}^{(n)} | Y_j^{(n)}, \ldots, Y_d^{(n)}, Y_1^{(n)}, \ldots, Y_{j-2}^{(n)}, \Psi$
9: **end for**

The convergence of the Gibbs sampler is more difficult to asses than with the standard Metropolis Hastings algorithm due to the nature of the chain movements and its advantages of no rejection. It affects the needed length of the chain to reach the stationary distribution as well as the length of the burnninig samples. The reader is advised to see the discussion on required conditions, theoretical and informal, in (Gamerman and Lopes 2006) or (Brooks et al. 2011).

One way to improve the convergence of the Gibbs sampler, which is relevant to the studies conducted in this thesis, is known asthe block Gibbs sampler. The general framework of the Gibbs sampler is not limited to only single elements of the vector $\mathbf{Y}$ but also it is feasible to allow for block or group sampling. As mentioned in Chapter 1, the grouping methods of sampling by the Gibbs Sampler improve the speed of convergence and decrease autocorrelation of the Markov chain elements. The authors of (Liu et al. 1994) and (Liu et al. 1995) show the improve the rate of convergence of the chain to its invariant distribution.

We will present this idea by defining 2 blocks of the elements of the vector $\mathbf{Y}$ such that $\mathbf{Y} = \left[\tilde{\mathbf{Y}}_1, \tilde{\mathbf{Y}}_2\right]$, which sum of number of elements is equal to the dimensionality of $\mathbf{Y}$. Given this notation, the transition kernel of the Gibbs sampler for two blocks can be formulated as follows

$$K_B(\mathbf{x}, \mathbf{y}) = \pi_{\tilde{\mathbf{Y}}_1|\tilde{\mathbf{Y}}_2=\tilde{\mathbf{x}}_2,\Psi}(\tilde{\mathbf{y}}_1)\pi_{\tilde{\mathbf{Y}}_2|\tilde{\mathbf{Y}}_1=\tilde{\mathbf{y}}_1,\Psi}(\tilde{\mathbf{y}}_2).$$

The formula can be generalized to different number of block with various orders of elements of $\mathbf{Y}$.

Part III

Applications

11. Financial Big Data Solutions for State Space Panel Regression in Interest Rates Dynamics

A novel class of dimension reduction methods is combined with a stochastic multi-factor panel regression-based state-space model from Chapter 5 in order to model the dynamic of yield curves whilst incorporating regression factors. This is achieved via Probabilistic Principal Component Analysis (PPCA) as defined in defined in Chapter 3 and Chapter 4 in which new statistically robust variants are derived also treating missing data. We embed the rank reduced feature extractions into a stochastic representation for state-space models for yield curve dynamics and compare the results to classical multi-factor dynamic Nelson-Siegel state-space models. This leads to important new representations of yield curve models that can be practically important for addressing questions of financial stress testing and monetary policy interventions which can incorporate efficiently financial big data. We illustrate our results on various financial and macroeconomic data sets from the Euro Zone and international market.

Our contribution is extending the state space structure of the Nelson-Siegel Yield Curve model to efficiently incorporate additional factors. We discuss the methodology of incorporating meaningful market information into an econometric model of a yield curve. We show how to extend the model which calibrates and forecast Euro Libor yield curve in order to utilize available market information. Such a model can serve as a useful tool for parsimonious stress testing exercise. The introduction of the methodology is followed by the discussion on an optimal estimation technique based on Kalman Filter studied by (Kalman and Bucy 1961) and discussed in Chapter 6. We perform comprehensive state-space models selection and structuring and the in-sample and out-of-sample analysis confirms that our methodology outperforms the standard method.

11.1. Introduction

The increased collection and accessibility to complex financial and macroeconomic data is transforming the way in which the financial services operate. Therefore the methods which study such data for financial applications should acknowledge these progressions in methodological developments and utilise increasingly big financial data sets to understand and reinterpret key financial applications such as interest rates dynamics. Furthermore, the increasing volume of market data poses both opportunity and challenge for financial institutions to deepen their understanding of market-wide and country-specific sources of risk present in financial markets and their implications on modelling and forecasting markets dynamics. Such information is of key importance for efficient investment decision making, understanding the influence of unconventional monetary policies and risk management procedures such as stress testing.

In this application study, we focus on providing a coherent methodology which utilises global macroeconomic and financial market data sets in modelling one of the most popular indicators of economic activity, the London Interbank Offered Rate (Libor). Libors are broadly used as benchmark rates for a wide collection of financial products. In order to use the information available in a range of financial data sets in a meaningful and parsimonious way to help explain the dynamics of Libor, we apply feature extraction methodology with special tailoring for handling irregular time series (missing data) and outliers.

11.1.1. Multifactor Models For Yield Curve Dynamics

Before the financial crisis of 2007-2008, it was widely believed that interest rate dynamics were well specified by single interest rate models such as short-rate models, rational pricing kernels, forward single rate models (Heath-Jarrow-Morton or Libor market models) with a leading assumption that there were no arbitrage relations between the instruments associated to different tenors and hence the rates were feasibly modelled separately. The financial crisis changed this perception uncovering not only the significant spreads between the rates of different tenors which are no longer negligible but also the arising discrepancies between the rates which were very close before the crisis, such as Libor rates and Overnight-Index swaps. This resulted in the emergence of multi-curve interest rate models which aim to capture the simultaneous presence of various interest rates and to reflect the implications of liquidity and credit risks resulting from lending in the interbank money market. For discussion on the theory of interest models and the post-crisis challenges the reader may refer to (Grbac and Runggaldier 2015) or (Cuchiero et al. 2016b). The study of (Filipovic and Trolle 2013) discusses multiple factors which are not captured by single rates models but are proven to have a significant impact on the term structure of Libor. The authors indicate the influence of the default risk which increases with a rate's maturity and is a dominant driver of long maturities. The study highlights the importance of modelling the term structure of Libor rates.

The existing literature on modelling the term structure of interest rates can be divided into two sets. The first group of methodologies aim to perform the exact replication of observed market rates via stochastic interpolations. These models are useful for daily calibration on the term structure but they are often not consistent on intraday dynamics. These term structure models are used in pricing or hedging applications. For details of methodological developments in the first class of models the reader may refer to recent studies of (Chib and Ergashev 2009), (Cuchiero et al. 2016b) or (Cuchiero et al. 2016a) and notably the consistent prediction framework of (Teichmann and Wuthrich 2016). The second class of approaches perform regression as opposed to interpolation. Such methods produce models more applicable for risk management, monetary policy, econometrics application such as and stress testing as they can be used as consistent in-sample estimations and forecasting applications.

The present study belongs to the second group of methods. We aim to develop under a regression state-space setting the incorporation of features extracted from macroeconomic and financial information present in big data sets which explain Euro Libor yield curve over time and

are not associated with traditional yield curve factors. The current literature on incorporating macroeconomic and financial big data into financial modelling mostly focus on the prediction of portfolio returns. The reader may refer to the well known works of (Cao and Tay 2003), (Armano et al. 2005) or (Kwon and Moon 2007) or more recent studies of (Geva and Zahavi 2014) (Kercheval and Zhang 2015), (Sirignano 2016) and (Borovykh et al. 2017) for applications of advanced learning algorithms in finance. These approaches have one significant disadvantage: they leave the interpretation of the results behind without addressing the question of the individual influence of financial and macroeconomic features. Therefore, they are not applicable for policy making and stress testing purposes when one wants to source the information on interactions between market components.

The existing literature on incorporating market-specific information into yield curve modelling mostly focus on using raw macroeconomic time series as exogenous, explanatory variables and incorporating them into a yield curve model by tracking the dynamics of their contributions to yield curve calibration over the time. Other methods use compressed information obtained using feature extraction techniques from financial data. For instance, the authors of (Diebold et al. 2006) used various economy condition proxies in addition to the components of Nelson-Siegel model, introduced by (Nelson and Siegel 1987), to show their significant link to US Treasury Yield and improve the forecast of the yield. The interactions between UK government bond yields and inflation, monetary rates and unemployment related proxies are combined in a stochastic volatility Nelson-Siegel model setting in (Bianchi et al. 2009). Particular attention is focused on the contribution to the yield curve volatility under monetary policy shocks that introduces developments of a coherent stress testing methodology. The study of (Joslin et al. 2014) further confirms the substantial effects of inflation and economic activity on the US Treasury yield curve by introducing affine, arbitrage-free methodology which incorporates the relevant macroeconomic factors. The approach to model simultaneously yield curve and filter unobserved macroeconomic variables using macroeconomic factors not spanned by classic yield curve components of the Nelson-Siegel model class is discussed in (Coroneo et al. 2016) where authors apply a dynamic factor methodology to the U.S. Treasury yields. More recently, (Karimalis et al. 2017) propose a multiple-country stress-testing framework to efficiently capture the effects on the set of European government yield curves from shocks in country-specific liquidity and credit-related variables by employing two procedures: modelling of dynamics of country-specific yield curves by employing dynamic macro-finance Nelson-Siegel model of (Diebold et al. 2006) and a robust covariance regression model applied to yield curves, both with inclusion of macro-finance factors.

11.1.2. Feature Extraction Methods For Financial Data

The modelling of the yield curve in the presence of a data-rich environment requires an application of feature extraction methodologies in order to reduce required model parameters and focus only on the incorporation of meaningful information into the model. A common practice in yield curve modelling is to use Principal Component Analysis in order to reduce the dimensionality of the data and obtain features which explain the highest portion of the

variance. Beginning with the study of (Ang and Piazzesi 2003) or (Emanuel Moench 2008), the authors use reduced information from inflation-linked and real-economic activity linked sets of US financial markets by extracting the first principal component from each group of data sets and then adopting the factor-augmented Vector Autoregressive (FAVAR) model to calibrate the US Treasury yield. The Factor Augmented Regression with static factors obtained by Principal Component Analysis is introduced in (Ludvigson and Ng 2010) to study the relation between bond excess returns and the macroeconomy. The use of international sovereign yield curves from leading economies in modelling, so-called, global yield curves, the sets of rates from leading countries, is investigated in (Abbritti et al. 2013). The authors utilise the FAVAR framework for combining traditional determinants of yield curve level, slope and curvature and three global factors. The other applications of incorporating features extracted from the sets of international sovereign yield curves in a term structure calibration by are the work of (Joslin et al. 2010), for the U.S. Treasury bond yields, or the study (Wright 2011) that investigates the decomposition of the forward rates into expected future short-term interest rates and term premiums.

Hence, the following study contributes to the latter group of works in combining the financial market data feature extraction and modelling of the structure of yield curves. However, the novelty is largely in the attention to the efficient and statistically robust feature extraction methodologies. We use international macroeconomic and financial big data sets to study the effects of global economies on the dynamic of the EUR Libor market. In order to handle the presence of irregularities in the real data time series and the possibility of outliers, we use developed novel statistically robust class of methods for feature extraction based on Probabilistic Principal Component Analysis introduced by (Tipping and Bishop 1999b) introduced in Part I. The extension which employ in the following study, use a reference distributions based on two versions of multivariate t-Student discussed in Section 3.2 in general framework and extended to the missing data setting in Section 4.3. The extracted features are used as predictors and are incorporated into a novel dynamic state-space formulation of Nelson Siegel model of (Nelson and Siegel 1987) following the work of (Diebold et al. 2006), with application to the EUR Libor yield curve.

11.1.3. Contributions & Structure

The following study can be split into two main parts corresponding to our two main contributions to feature extraction and dynamic yield curve modelling.

Our first contribution is illustration of the t-Student IID PPCA methodology and its application for econometric studies on financial big data sets. We compare the new methodology with the Gaussian PPCA and t-Student IND PPCA.

Our second contribution is extending the state space structure of the Nelson-Siegel Yield Curve model to efficiently incorporate additional factors. We discuss the methodology of incorporating meaningful market information into an econometric model of a yield curve. We show how to extend the model which calibrates and forecast Euro Libor yield curve in order to utilize available market information. Such a model can serve as a useful tool for parsimonious stress testing exercise. The introduction of the methodology is followed by the discussion on an optimal

estimation technique based on Kalman Filter studied by (Kalman and Bucy 1961). We perform comprehensive state-space models selection and structuring and the in-sample and out-of-sample analysis confirm that our methodology outperforms the standard method.

The utilisation of the market information in yield curve modelling is well documented in the current literature. In the study of (Diebold et al. 2006), the authors extended the dynamic Nelson-Siegel of (Nelson and Siegel 1987) model by including observable macroeconomic variables in order to investigate the interaction between the macroeconomy and a yield curve. Nevertheless, given current availability of the data, we find this approach disadvantageous due to two aforementioned reasons: the high possibility of the model over parametrization and hidden links between various data sets what may spoil the modelling quality. Therefore, instead of incorporating raw data we discuss how to use the derived market features.

The rest of this chapter is organised as follows. The list of data sets used in this study is detailed in Section 11.2 and is discussed in details in Section B.1 in Appendix B. The Section 11.3 introduces the concept of yield curve modelling as well as extends the framework by showing how to incorporate the market specific features in an econometric manner. The subsequent parts of the section overview the employed methodology of the model estimation and selection. The outcomes of the derived methodologies of feature extraction applied to real datasets are discussed in Section 11.4. The framework for the calibration and forecasting Euro Libor yield curve is examined on real data studies presented in Section 11.5.

11.2. International Macroeconomic and Financial Big Data Sets

In the following section, we briefly overview various international macroeconomic and financial data sets which are investigated in this study, summarized in Table 11.1. We advise the reader to refer to Section B.1 in Appendix B for a detailed discussion and insight into investigated time series.

Our attention is focused on investigating features which measure Euro-Zone market conditions and their decomposition into their influences from the domestic and international markets. The Euro-zone market is represented by the Euro Libor yield curve. We collect the data from leading domestic and global economies, in a total of 12 countries, categorized as follows:

Domestic economies: Germany (DE), France (FR), Portugal (PO), Spain (ES), Italy (IT), Ireland (IR) and United Kingdom (GB);

International economies Japan (JP), United States (US), Australia (AU), Brazil (BR) and South Africa (SA).

The subsets of data which serve to investigate various domestic and global influences are categorized as follows:

GOV the panel valued time series which measure the interactions between country-specific sovereign yield curves;

INF the panel valued time series which measure the influence of country-specific inflation risks represented by inflation-linked yield curves and Consumer Price Indexes;

PRD the time series which measure the influence of country-specific productivity proxies represented by Gross Domestic Product (GDP), Unemployment Rates, and Labour Productivity;

FX the time series which measure the interactions between leading foreign exchange markets and Euro-Zone market given by exchange rates Euro (EUR), and Japanese Yen (JPY), United States dollar (USD) and Australian dollar (AUD);

LIQ the time series which serves as proxies of Euro-Zone liquidity represented by 2W Euro Repo rate and German Bund Open Interest;

CR the time series which are used as proxies of the credit quality of Euro-Zone given by 5Y Markit Intraxx Index.

The instruments described by three first row of Table 11.1 are country-specific yield curves which have a term structure. Therefore, their one observation is a multidimensional vector with elements corresponding to the rates with specific maturities belonging to the set $\tau \in \{1 \text{ months } (M),\ 3 \text{ months } (M), \ldots, 20 \text{ years } (Y)\}$.

Table 11.1. The list of collected financial and macroeconomic time series investigated in the study. The first three columns of the table correspond to the name, sector category and frequency of the quotes of an instrument. The rest of the columns indicated country-specific availabilities of the collected data. The Bloomberg identifiers of the data are listed in Table A2 and Table A3 in Section B.1.

Instrument	Cat	Freq.	Availability											
			EU	DE	FR	PO	ES	IT	IR	GB	JP	US	AU	BR	SA
Libor Curve		daily	✓												
Sovereign Curve	GOV	daily		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Inflation Curve	INF	daily		✓	✓		✓	✓		✓	✓	✓	✓		
FX Rates	FX	daily								✓	✓	✓	✓		
CPI	INF	monthly		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
GDP	PRD	yearly		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Unemployment Rates	PRD	yearly		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Labour Productivity	PRD	yearly		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
German Bund OI	LIQ	daily													
Eur Repo and Eurobor 3M	LIQ	daily													
Markit iTraxx 5Y	CR	daily													

11.3. Multifactor Model with Macroeconomic Factors for EUR Libor Yield Curve

In this section, we show how to utilize the features extracted using PPCA to incorporate into statistical models as exogenous factors in state or observation equations. In particular, we consider latent factor term structure yield curve dynamic regression models which we use to calibrate and forecast EUR Libor yields over time. The Nelson-Siegel model was initially proposed by (Nelson and Siegel 1987), and subsequently extended, by (Diebold et al. 2006), to a dynamic latent factor model to allow for time-varying parameters. The dynamic approach provides not only a good quality of the calibration of the yield curves but in addition supplies information about interactions between the explanatory components of the level, slope and curvature over time allowing both efficient and optimal estimation as well as direct model interpretation. Furthermore, the class of models we develop improves the forecast performance and interpretation of the model. We start by discussing the classic version of the model to later show the extension to its dynamic version which utilizes exogenous factors. The section is finished with a brief overview of the estimation and model selection methodology.

11.3.1. The dynamic Nelson-Siegel model

The dynamic version of Nelson-Siegel model is an extension to classic, regression-based model introduced in (Nelson and Siegel 1987). Let $\mathbf{y}_{1:N} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]$ be a set of N observations of an yield curve. Each observation is a d-dimensional vector of yields corresponds to rates at different maturities $\tau_1 \leq \ldots \leq \tau_d$, that is, $\mathbf{y}_t = [\mathbf{y}_t^1, \ldots, \mathbf{y}_t^d]$ where $\mathbf{y}_t^i$ is a rate at maturity τ_i available in calendar time t. Following the work of (Diebold et al. 2006), the term structure of yields at maturities $\tau_1 \leq \ldots \leq \tau_d$ in time t is modelled by the set of equations

$$\begin{pmatrix} y_t^{\tau_1} \\ y_t^{\tau_2} \\ \vdots \\ y_t^{\tau_d} \end{pmatrix} = \begin{pmatrix} 1 & \frac{1-e^{-\lambda\tau_1}}{\lambda\tau_1} & \frac{1-e^{-\lambda\tau_1}}{\lambda\tau_1} - e^{-\lambda\tau_1} \\ 1 & \frac{1-e^{-\lambda\tau_2}}{\lambda\tau_2} & \frac{1-e^{-\lambda\tau_2}}{\lambda\tau_2} - e^{-\lambda\tau_2} \\ \vdots & & \\ 1 & \frac{1-e^{-\lambda\tau_d}}{\lambda\tau_d} & \frac{1-e^{-\lambda\tau_d}}{\lambda\tau_d} - e^{-\lambda\tau_d} \end{pmatrix} \begin{pmatrix} L_t \\ S_t \\ C_t \end{pmatrix} + \begin{pmatrix} \epsilon_t^{\tau_1} \\ \epsilon_t^{\tau_2} \\ \vdots \\ \epsilon_t^{\tau_d} \end{pmatrix},$$

$$\begin{pmatrix} L_t \\ S_t \\ C_t \end{pmatrix} = \begin{pmatrix} c^L \\ c^S \\ c^C \end{pmatrix} + \begin{pmatrix} a^{L,L} & a^{L,S} & a^{L,C} \\ a^{S,L} & a^{S,S} & a^{S,C} \\ a^{C,L} & a^{C,S} & a^{C,C} \end{pmatrix} \begin{pmatrix} L_{t-1} \\ S_{t-1} \\ C_{t-1} \end{pmatrix} + \begin{pmatrix} \eta_t^L \\ \eta_t^S \\ \eta_t^C \end{pmatrix}. \tag{11.1}$$

which we summarize as follows

$$\begin{aligned} \mathbf{y}_t &= \boldsymbol{\Lambda}\boldsymbol{\beta}_t + \boldsymbol{\epsilon}_t \\ \boldsymbol{\beta}_t &= \mathbf{c}_{3\times 1} + \mathbf{A}_{3\times 3}\boldsymbol{\beta}_{t-1} + \boldsymbol{\eta}_t \end{aligned} \tag{11.2}$$

for the matrix $\boldsymbol{\Lambda}_{d\times 3} := \boldsymbol{\Lambda}(\lambda)$ with Nelson-Siegel factor loading being a function of the static parameter λ. The (i, j)th element is given by

$$\boldsymbol{\Lambda}_{ij}(\lambda) = \begin{cases} 1, & \text{for} \quad j = 1, \\ (1 - e^{-\lambda\cdot\tau_i})/\lambda\cdot\tau_i, & \text{for} \quad j = 2, \\ (1 - e^{-\lambda\cdot\tau_i} - \lambda\cdot\tau_i\, e^{-\lambda\cdot\tau_i})/\lambda\cdot\tau_i, & \text{for} \quad j = 3. \end{cases}$$

The measurement (top) equation in Equation (11.1) or Equation (11.2) relates observed variables to the unobserved vector $\boldsymbol{\beta}_t$, that is, a set of d yields of bonds to the three latent factors L_t, S_t and C_t. The vector of unobserved variables is modelled by the state (bottom) equation in (11.1) or (11.2). The error terms of the observation and state equations, $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}_{d\times 1}, \mathbf{Q}_{d\times d})$ and $\boldsymbol{\eta}_t \sim \mathcal{N}(\mathbf{0}_{3\times 1}, \mathbf{R}_{3\times 3})$, respectively, are assumed to be mutually independent and independent over time. In this study, the vector of latent process $\boldsymbol{\beta}_t = [L_t, S_t, C_t]$ follows a vector autoregressive process of first order, VAR(1), but this assumption can be easily extended.

We assume that the matrix with factors loading, $\boldsymbol{\Lambda}(\lambda)$ does not vary over time, i.e we assume one value of λ for the whole period of the examination per a given yield. The rationale behind this assumption is the interpretability of the Nelson Siegel's polynomials which can be seen as a level, slope and curvature of a yield curve. The loading on the first factor from the matrix $\boldsymbol{\Lambda}$

equals 1 and is interpreted as the level of the yield curve since it affects all components of the yield equally. The second loading is called slope because, as a function that starts at 1 and converges quickly and monotonically to 0 as τ_i increases, it affects short rates more heavily than long rates and as a consequence changes the slope of the yield curve. The last factor from the matrix $\mathbf{\Lambda}$ loads medium rates more heavily since it rapidly start from zero and, after a hump, decays towards zero. Therefore, the third factor changes the yield curve curvature.

Given the fixed λ, the magnitude of the impact of each of the factors on the shape of a yield is given by the values of coefficients L_t, S_t and C_t over time. In order to capture the interactions between L_t, S_t and C_t and construct the forecasting distribution, the vector of coefficients $\boldsymbol{\beta}_t = [L_t, S_t, C_t]$ is assumed to have a time series structure.

11.3.2. Extending the Dynamic Nelson-Siegel to Macroeconomic Factor Model

Chapter 5 demonstrates a general approach one may adopt to extend state-space model formulations from Equation (5.1) to incorporate additional observable factors which contain compressed information about the given dataset. In this section we illustrate how to utilize the features, extracted from the various data set as macroeconomic, financial or demographic data sets, to the dynamic Nelson-Siegel model in a static form but with dynamic coefficient on factor loadings, the latent state processes. Thus we are able to assess the impact of the factors on a yield over time. This approach has this advantage, that we do not have to model explicitly the macroeconomic data which may have a complex structure.

Let us define $\mathbf{F}_t$ as a $d \times k$ dimensional matrix with various market features, where k represents a number of the maturity-related factors. We can obtain $\mathbf{F}_t$ as a matrix with leading eigenvectors from datasets which are assumed to be an exogenous observable input that is believed to have a potential influence on the term structure of yield curve dynamics under study in the responses. Each feature vector regressor, extracted from the exogenous data, we add to the state-space model (11.2).

There are numerous structural ways to achieve this in a state-space model. For instance, the factors may either influence the whole term structure by entering the factor into the state equation, or it may influence components of the yield curve differently by adding them to the observation equation at different tenors on the term structure. We can also try a combination of such approaches, dependent on which macroeconomic data the feature was extracted from the context of the model construction.

We assume that the influence of features vary over time. Therefore we specify a time dynamic for the processes which are coefficients factor loadings, the latent variable, a $d \cdot k$ dimensional vector $\boldsymbol{\varrho}_t$ which is modelled by VAR(1) process given by

$$\boldsymbol{\varrho}_t = \mathbf{\Psi} + \mathbf{\Omega}\boldsymbol{\varrho}_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \overset{iid}{\sim} \mathcal{N}(0, \mathbf{H}_{dk \times dk})$$

with homogeneous covariance matrix of the error term $\boldsymbol{\omega}_t$. $\boldsymbol{\varrho}_t$ is a dynamic regression parameter which specifies the impact of an i-th market features corresponding to the i-th column of $[\mathbf{F}_t]$.

Depending on the interpretation of the desired model, one may incorporate $\mathbf{F}_t$ into observation, the top equation of Equation (11.2) (**Case 1**), the latent dynamic of $\boldsymbol{\beta}_t$ (**Case 2**), that is the bottom equation of Equation (11.2) or both (**Case 3**). The new class of the hybrid models, which combines the standard Nelson-Siegel model with exogenous features, names Extended Nelson-Siegel class, is defined as follows

$$\boldsymbol{y}_t = \tilde{\boldsymbol{\Lambda}}_t \tilde{\boldsymbol{\beta}}_t + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{Q}), \tag{11.3a}$$

$$\tilde{\boldsymbol{\beta}}_t = \tilde{\mathbf{c}} + \tilde{\mathbf{A}}_t \tilde{\boldsymbol{\beta}}_{t-1} + \tilde{\boldsymbol{\eta}}_t, \quad \tilde{\boldsymbol{\eta}}_t \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \tilde{\mathbf{R}}) \tag{11.3b}$$

where $\tilde{\boldsymbol{\beta}}_t = [\boldsymbol{\beta}_t, \boldsymbol{\varrho}_t]$ is a $(3 + dk)$-dimensional latent process vector and

$$\tilde{\mathbf{c}} = \begin{pmatrix} \mathbf{c}_{3\times 1} \\ \boldsymbol{\Psi}_{pk\times 1} \end{pmatrix}_{(3+dk)\times 1}$$

is an interception term of the state equation. The covariance matrix of the error terms of state equation, $\tilde{\boldsymbol{\eta}}_t$, is given by

$$\tilde{\mathbf{R}} = \left(\begin{array}{c|c} \mathbf{R} & \mathbf{Y} \\ \hline \mathbf{Y} & \mathbf{H} \end{array}\right)_{(3+dk)\times(3+dk)}$$

where $\mathbf{Y}$ defines the correlation structure between latent processes $\boldsymbol{\beta}_t$ and $\boldsymbol{\varrho}_t$. Let us specify the following two objects, $\tilde{\mathbf{F}}_t = \bigoplus_{j=1}^{k} [\mathbf{F}_t]_{j,\cdot}$, for $\bigoplus$ being a direct sum operator, and $\tilde{\mathbf{f}}_t = vec\left(\mathbf{F}_t^T\right)$, that is

$$\tilde{\mathbf{F}}_t = \begin{pmatrix} [\mathbf{F}_t]_{1,\cdot} & 0 & 0 & \cdots & 0 \\ 0 & [\mathbf{F}_t]_{2,\cdot} & 0 & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & \cdots & & & [\mathbf{F}_t]_{p,\cdot} \end{pmatrix}_{dk\times dk} \quad \text{and } \tilde{\mathbf{f}}_t = \begin{pmatrix} [\mathbf{F}_t]_{1,1} \\ [\mathbf{F}_t]_{1,2} \\ \vdots \\ [\mathbf{F}_t]_{d,k} \end{pmatrix}_{dk\times 1}$$

where $[\mathbf{F}_t]_{j,\cdot}$ and $[\mathbf{F}_t]_{j,m}$ represent the vector of the j-th row of the matrix $\mathbf{F}_t$ and the element corresponding to jth row and mth column, respectively. The case-specific formulations of the

transition matrices of the observation and state equations of the extended model (11.3), is the following

$$\tilde{\mathbf{\Lambda}}_{t\ d\times(3+dk)} = \begin{cases} \left(\begin{array}{c|c} \mathbf{\Lambda}_{d\times 3} & \tilde{\mathbf{F}}_t \end{array} \right) & \text{for } \textbf{Case 1}, \\ \left(\begin{array}{c|c} \mathbf{\Lambda}_{d\times 3} & \mathbf{0}_{d\times dk} \end{array} \right) & \text{otherwise}, \end{cases}$$

$$\tilde{\mathbf{A}}_{t,\ (3+dk)\times(3+dk)} = \begin{cases} \left(\begin{array}{c|c} \mathbf{A}_{3\times 3} & \mathbf{0}_{3\times dk} \\ \hline \mathbf{0}_{dk\times 3} & \mathbf{\Omega}_{dk\times dk} \end{array} \right) & \text{for } \textbf{Case 1}, \\ \left(\begin{array}{c|c} & \tilde{\mathbf{f}}_t^T \\ \mathbf{A}_{3\times 3} & \tilde{\mathbf{f}}_t^T \\ & \tilde{\mathbf{f}}_t^T \\ \hline \mathbf{0}_{dk\times 3} & \mathbf{\Omega}_{dk\times dk} \end{array} \right) & \text{for } \textbf{Case 2}. \end{cases}$$

11.3.3. Estimation based on the Kalman Filter

Given the fact that the transition matrix $\mathbf{\Lambda}$ of the observation equation (11.1), as well as the version incorporating exogenous factors (11.3), is a non-linear function of the static parameter λ, both the dynamic Nelson-Siegel model and its extensions are non-linear Gaussian models. However, as pointed out in (Cairns and Pritchard 2001), the non-linear estimators are extremely sensitive to the initialisation values of an estimation process and the probability of getting local optima is high. The existence of more than one maximum can lead to miscalibration of a yield curve. To overcome this problem while working with Nelson-Siegel class of models, a common approach is to work with conditionally linear model (11.2) (equivalently the model in Equations (11.3)), that is, to fix the parameters λ in the estimation process of the other static parameters and specify it separately. The parameter λ is treated a part of a model selection stage, rather than the estimation one, and its selection is carried out a λ-based grid search given some score function of the estimation, usually the likelihood function of a model.

When we treat λ as a model selection criterion, the dynamic Nelson-Siegel model (11.2) and its extension (11.3) is a linear Gaussian state-space model. Hence, the vector of latent variable $\boldsymbol{\beta}_t$ is optimally estimated by means of Kalman Filter introduced by (Kalman and Bucy 1961) and discussed in Chapter 6 given the historical and current observations. The conditional distributions involved in the multivariate Kalman Filter recursions are given by

$$\begin{aligned} \tilde{\boldsymbol{\beta}}_{t-1}|\mathbf{y}_{1:t-1} &\sim \mathcal{N}(\tilde{\boldsymbol{\beta}}_{t-1|t-1}, \mathbf{P}_{t-1|t-1}), \\ \tilde{\boldsymbol{\beta}}_{t}|\mathbf{y}_{1:t-1} &\sim \mathcal{N}(\tilde{\boldsymbol{\beta}}_{t|t-1}, \mathbf{P}_{t|t-1}), \\ \mathbf{y}_{t}|\mathbf{y}_{1:t-1} &\sim \mathcal{N}(\mathbf{f}_t, \mathbf{F}_t) \end{aligned} \tag{11.4}$$

where

$$\tilde{\boldsymbol{\beta}}_{t-1|t} = \tilde{\mathbf{c}} + \tilde{\mathbf{A}}\tilde{\boldsymbol{\beta}}_{t-1|t-1}$$
$$\mathbf{P}_{t|t-1} = \tilde{\mathbf{A}}\mathbf{P}_{t-1|t-1}\tilde{\mathbf{A}}^T + \tilde{\mathbf{R}}$$
$$\mathbf{f}_t = \tilde{\boldsymbol{\Lambda}}_t\tilde{\boldsymbol{\beta}}_{t|t-1}$$
$$\mathbf{F}_t = \tilde{\boldsymbol{\Lambda}}_t\mathbf{P}_{t|t-1}\tilde{\boldsymbol{\Lambda}}_t^T + \mathbf{Q}$$
$$\tilde{\boldsymbol{\beta}}_{t|t} = \tilde{\boldsymbol{\beta}}_{t-1|t} + \mathbf{P}_{t|t-1}\tilde{\boldsymbol{\Lambda}}_t^T\mathbf{F}_t^{-1}(\mathbf{y}_t - \mathbf{f}_t)$$
$$\mathbf{P}_{t|t} = \mathbf{P}_{t|t-1} - \mathbf{P}_{t|t-1}\tilde{\boldsymbol{\Lambda}}_t^T\mathbf{F}_t\tilde{\boldsymbol{\Lambda}}_t\mathbf{P}_{t|t-1}$$

for $t = 1, \ldots, T$. The estimation of the vector of static parameters $\Psi = [\mathbf{Q}, \mathbf{c}, \boldsymbol{\Psi}, \mathbf{A}, \boldsymbol{\Omega}, \mathbf{R}, \mathbf{H}]$ is carried out by maximising the loglikelihood function of Kalman Filter based on the conditional distribution of the prediction errors $\mathbf{e}_t := \mathbf{y}_t - \mathbf{f}_t$ determined in Equation (7.15). The value of λ which obtains the highest value of the likelihood function from Equation (7.15) of a given model (the same assumption about $\tilde{\mathbf{Q}}$, $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{R}}$), is then chosen as an estimator of the shape parameter.

11.3.4. Model Selection Methodology For Nelson-Siegel Class of Models

Given the state-space representation of the Extended Nelson-Siegel model in equations 11.3, we have multiple choices for the model calibration. In the work of (Diebold et al. 2006), the authors assumed the matrix $\mathbf{Q}$ to be diagonal whereas the matrix $\tilde{\boldsymbol{R}}$ to be unrestricted, that is non-diagonal. The assumption of a diagonal Q matrix, what implies mutually uncorrelated deviations of yield elements at various maturities, is quite common as one may expect the measurement errors to be independent across a term structure. The dynamic Nelson-Siegel models allow accounting for serial correlation between components of the latent vector. The form of the transition matrix $\tilde{\mathbf{A}}$ and the covariance matrix $\tilde{\mathbf{R}}$ specifies the strength of the linear dependencies between the elements of $\tilde{\boldsymbol{\beta}}_t$. Therefore, in this study we examine the following collection of models

Model 1: diagonal $\mathbf{Q}, \tilde{\mathbf{A}}, \tilde{\mathbf{R}}$
Model 2: diagonal $\mathbf{Q}, \tilde{\mathbf{A}}$,
Model 3: diagonal $\mathbf{Q}, \tilde{\mathbf{R}}$,
Model 4: diagonal $\mathbf{Q}$.

In addition, we examine two covariance structures: heterogeneous ('hete') and homogeneous ('homo'). In addition we assume that the subvectors of $\tilde{\boldsymbol{\beta}}_t$, the latent vectors $\boldsymbol{\beta}_t$ and $\boldsymbol{\varrho}_t$, are orthogonal and hence the matrices $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{R}}$ are block-diagonal. The preferred model is selected as the one which minimizes AIC defined in Equation (9.1) .

11.4. Feature Extraction using Cross-Country Macroeconomic Data Sets

We undertook excessive synthetic data case studies on setting up the PPCA EM algorithm detailed in Chapter 3.3. From this we learned that the robust initialisation step and the robust estimation of the data's moments increase the sensitivity of the algorithms to the initialisation step

without significantly improving the quality of the estimation (recovery) of PPCA models' static parameters (such as the matrix **W** in the model in Equation (2.7)).

In addition, the studies aim to identify which methodology is the most robust in the presence of two different perturbation patterns: when the whole observation is perturbed or its individual elements are perturbed. We examine the methodology under different input parameters: length of the sample, the dimensionality of an observation, the proportion of missing values and the proportion of the corrupted sample. These synthetic studies performed taught us that in the presence of the element-wise perturbation, the best methodology which estimates the mean vector $\boldsymbol{\mu}$ is t-Student IND PPCA regardless of the input parameters. The three first eigenvectors of the estimated covariance matrix are the most accurately estimated either t-Student IND or Gaussian PPCA in most of the cases except for the high dimensions and highest perturbation when the t-Student IID outperforms the other methodologies. When we perturb the data row-wise, the t-Student IID PPCA is most accurate in estimating the three eigenvectors when the dimensionality of the data is high.

In this section, we detail the feature extraction, using described PPCA methodologies, which is conducted yearly on the real datasets outlined in Section 11.2, summarized in Table 11.1. As a consequence of the synthetic data studies, we decided to exclude the robust initialisation and the robust Gaussian PPCA from the examined methodologies.

The features extracted using PPCA represent the relevant directions of most meaningful variation of a data set, described by the spectral information in the form of eigenvectors of the robust factor model estimated covariance matrix via PPCA frameworks. The eigenvectors determine the directions along which the data set can be orthogonally decomposed in decreasing degree of variability. The first eigenvector explains the projection direction of the data with most variability.

The three first subsections discuss the features extracted from country-specific sovereign yield curves from Subsection B.1.2, the various inflation-linked yield curves from Subsection B.1.3.2 and the data set which consists of proxies to the Euro-Zone market condition: the liquidity, credit and foreign exchange risk from Subsection B.1.5. The last subsection examines the correlation between the features among various data sets per year and the PPCA framework. We use the following acronyms for the analysed datasets:

D2 the country-specific government yield curves for countries listed in Table A2 in Section B.1;

D3 the set of country-specific inflation-linked sovereign yield curves for German, France, United Kingdom, Italy, Japan and the United States of America, mixed with inflation-linked swap curves for countries in which inflation-linked government instruments are not available, that is, Australia and Spain;

D3b the set of country-specific inflation-linked sovereign yield curves;

D3s the set of country-specific inflation-linked swap curves;

D4 the set of the Euro-zone activity proxies: the four FX rates against EUR: AUDEUR, JPYEUR, GBPEUR, USDEUR; the proxies for Euro-zone liquidity risk: the open interest of the front

future contract on Euro-Bund and the spread between 3 month Euribor and 3 month Euro repo rate; and proxy for the credit risk of Euro-zone: 5Y Markit Itraxx Index.

We test and compare each of the PPCA methods on each data set **D2-D4**, treating the presence of missing values which are illustrated in the corresponding figures in Section 11.2. Some results are matrix values time series, (such as sets **D2-D3b**) as containing information across various maturities and countries, others are vector valued. We assess the discrepancies between leading eigenvectors obtained from considered PPCA frameworks.

The feature extraction for the sets of time series from Table 11.1 which observations arrive with yearly or monthly frequencies (such as CPI, GDP, unemployment rates and labour productivity), is not discussed in this section. For these data sets, we employ a different feature extraction procedure which is done by carrying forward the step-function of the observed signal until the next observed value of the variable is recorded. Therefore, we replicate less frequent values in order to match daily frequency of Libor yield curve, that is, if a CPI is recorded monthly, all daily records in a given month related to this CPI are equal to the monthly value. The procedure results in the data matrix of daily variables corresponding to the blocks of yearly or monthly values of GDP, Labour Productivity, Unemployment Rates and CPIs.

11.4.1. Features Extraction for Country-Specific Sovereign Yield Curves

The daily observations of the instruments from the data set **D2** are multidimensional, that is, their entries correspond to the points on the term structure of a country-specific sovereign yield. Consequently, the eigenvectors of the data's covariance matrix, which have the same dimensionality, can be partitioned into the country-specific sub-vectors. The dimensionality of each such sub-eigenvector equals to the number of tenors of the corresponding term structures for a given country.

The plots in Figure 11.1 show the yearly logarithms of three eigenvalues for the data set **D2** (the left panel) and the optimal degrees of freedom for t-Student PPCA methodologies (the right panel) across different years. When we compare the difference in magnitude between the logarithms of the three eigenvalues we notice that the eigenvalues obtained by t-Student IID PPCA are the most consistent, especially, the distance between the second and third eigenvalue. What is more, the t-Student IID PPCA provides the highest three first eigenvalues among considered methodologies.

The degrees of freedom chosen for t-Student IND PPCA are more volatile across years than for t-Student IID PPCA which most frequently equal to 1 - the lowest possible value on the grid. Let us recall that the t-Student IID PPCA down-weights the whole set of the observations in order to estimate the moments of the distribution, whereas the t-Student IND PPCA treats each observation individually and calculates the moments using observation-specific weights. Therefore the independent methodology leans toward the lighter tails, for the same sample set. Intuitively, when an individual point of a sample is extreme, we would expect that one needs larger degrees of freedom to capture it as a distant point since observation-specific weight diminishes the information about the tails than if we consider the aggregated sample with one weight as in

the case of t-Student IID PPCA. The synthetic data study in Section 3.3.2 confirms this reasoning, especially for the higher dimensionality of the data and the presence of missing entries.

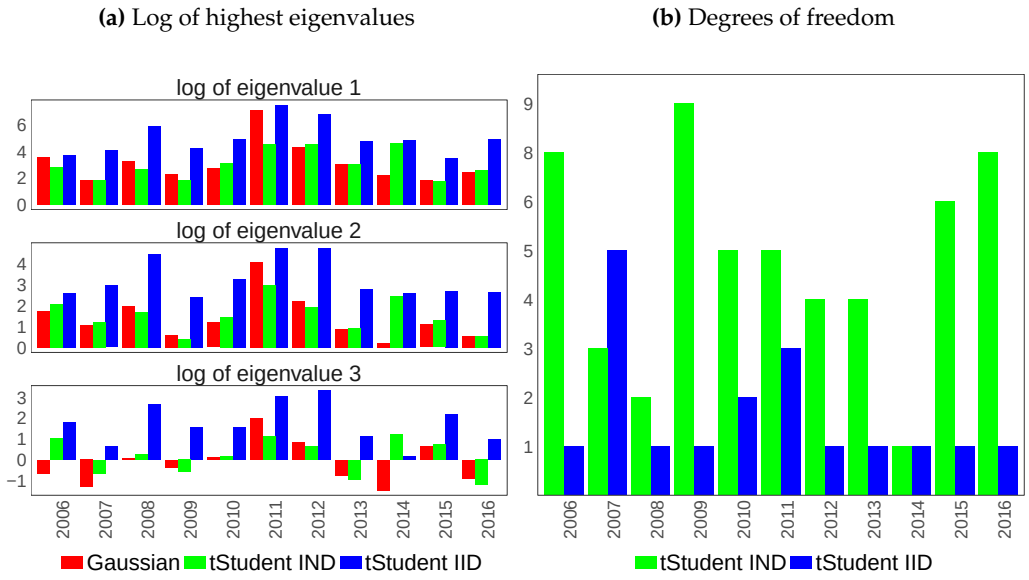


Figure 11.1. The left panel (a): The logarithms of the three highest eigenvalues of the covariance matrix. The right panel (b): the optimal values of the degrees of freedom for t-Student PPCA frameworks over years (x axis) for the data sets **D2** which consists of country-specific sovereign yields.

The relative difference between logarithms of the first and second eigenvalue in Figure 11.1a indicates that the first eigenvectors of **D2** data set explain at least ten times higher portion of the variance than the second eigenvectors and remain significantly dominant for the whole period 2006-2016.

Figure 11.2 illustrates the first eigenvectors which are obtained on yearly non-overlapping subsets of **D2** data corresponding to period 2006-2016. The column-wise order of the panels illustrate the results for subsequent years, whereas the rows-wise order corresponds to the country-specific sub-vectors. The term structure of the eigenvectors is given on the x-axis of the plots. We show the results for different frameworks: Gaussian PPCA (red line), t-Student IND PPCA (green line) and t-Student IND PPCA (blue line).

It is worth to recall, that the eigenvectors obtained by different PPCA methodologies can be rotated, that is, they may have the opposite signs (+/-1) but still provide similar directions. As an example, let us consider the plot of South Africa's sub-eigenvector for the year 2007. The green and blue lines have the maximum point in 5Y maturity whereas the red line indicates the minimum is the same point. If we multiply the red line by '-1', all three lines would be consistently pointing towards maximum in 5Y but would indicate different magnitudes. Therefore, when analysing the resulting eigenvectors, we must distinguish between the difference in a rotation and a difference in the information given by the eigenvectors obtained by various PPCA methodologies.

The eigenvectors for years 2008, 2011, 2016 are very similar among all PPCA frameworks. The Gaussian PPCA and t-Student IID PPCA provide consistent features for years 2009, 2012 and 2014 whereas the eigenvectors extracted using t-Student IND PPCA are more smooth and shifted

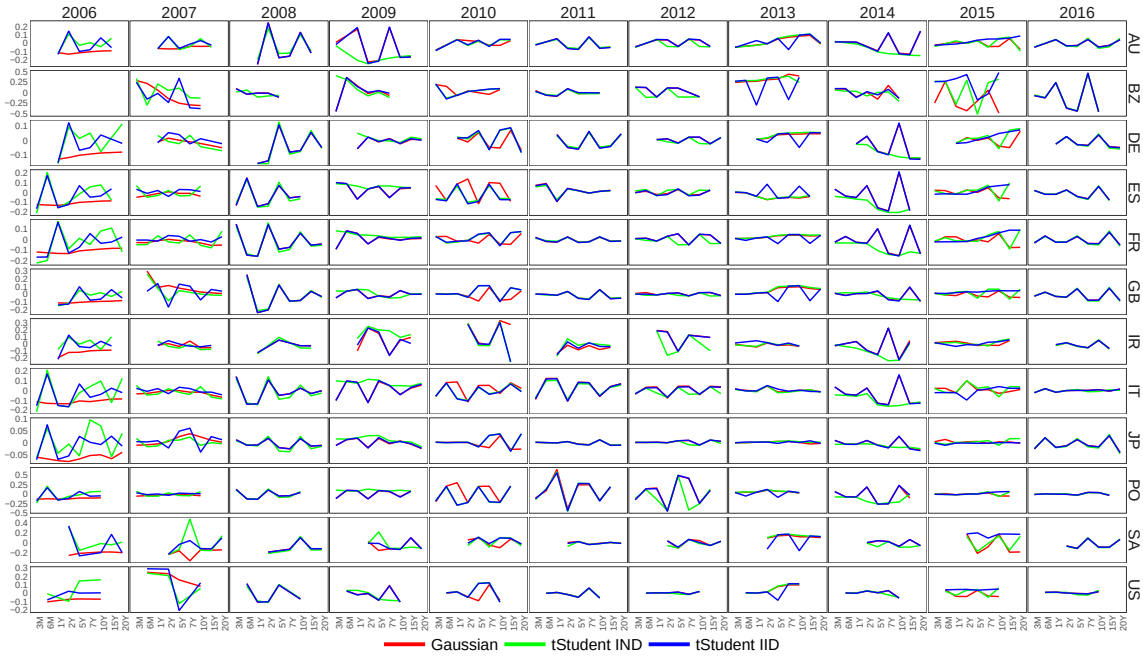


Figure 11.2. The first eigenvector of a covariance matrix for the **D2** set of country-specific sovereign yield curves across different years (column-wise) partitioned into country-specific subvectors (row-wise) and their term structure (x-axis). The colours of lines correspond to different PPCA frameworks: Gaussian PPCA (red line), t-Student IND PPCA (green line) and t-Student IID PPCA (blue line).

one term structure point backwards. The shift indicates that t-Student IND PPCA indicates the heavier loadings on the shorter terms. For the year 2013, the features extracted using Gaussian and t-Student IND PPCA are both smoother across the term structure of a given country. The two t-Student methodologies agree in 2010 with the terms at maturities which have the heaviest loadings. The vectors obtained by Gaussian PPCA highlighted longer points of maturities having the highest values.

With regards to the pre-crisis period 2006-2007, we notice strong differences between t-Student and Gaussian PPCA frameworks. The values of the vectors provided by the non-robust methodology are consistent across the term structures and indicate non-varying decomposition of the features without pointing towards any subset of maturities. The contradicting information is given by t-Student PPCA methodologies which highlight particular points on the term structures. **It is an interesting outcome which illustrates the discrepancies between information provided by robust and non-robust frameworks and highlighting the importance on real data of adopting our proposed new methodology in such feature extraction in practice.**

When analysing the country-specific panels year by year, one may remark that each methodology provides the consistent set of maturities for all examined countries, which have the highest loadings. The outcome gives us the intuition that the variability of the country-specific yield curves is determined by a common set of maturities across countries. The results differ on the yearly basis. In 2006 the eigenvectors point out in the direction of the South African curve as the main source of the variability which is complemented with most of the European countries. The source of the dominant variability of the data set **D2** in 2007, corresponds to the middle maturities from Brazil, United Kingdom, South Africa and the United States.

In addition, one of the exceptional findings is the changed order of the degrees of freedom for the two t-Student PPCA frameworks in 2007. The optimal degrees of freedom chosen for t-Student IID PPCA in 2007 is higher than for the Student IND PPCA. One may expect that the reason for the increase in the degrees of freedom for the former methodology reflects the smaller variability of the yields. The first column of the panels in Figure A11 in Section B.1 in Appendix B illustrates the sovereign yields for 2006 and 2007 across various countries (row-wise) and maturities (the colours of the lines). The dynamic of the yields of European countries and the United States does not support this reasoning as the yields have started to decline in the mid of 2007 (the middle components of the United States yield even earlier). In addition, the optimal degrees of freedom for t-Student IND PPCA decreased in 2007 and consequently, the individual weights of the observations are lower in 2007 than in 2006. The assumed distribution in 2007 is more heavy-tailed than in 2006. The eigenvectors obtained by the two t-Student PPCA frameworks illustrated in Figure 11.2 are similar for most of the considered countries. The two methodologies agree that one of the main sources of the data's variability is the middle tenors of United States (US) and United Kingdom (GB) yields. The strongest discrepancies between the results obtained by the two PPCA frameworks relate to the eigenvector components corresponding to Brazil (BZ) and South African (SA) yields. The t-Student IID PPCA assumes the greater impact of the Brazilian yield on the variability of the data set and loads more heavily on the middle terms of the yield whereas t-Student IND PPCA selects the South African yield.

The Brazilian and South African yields are characterized by the moderate and strong missingness, respectively, as shown in the corresponding panels in Figure A5. The projections which are used to handle the presence of missing values in Brazilian and South African yields are determined by the conditional means of the corresponding distributions. We suspect that the source of this discrepancy is the difference in how the missing values present in the Brasilia and South African yields are projected. To verify this explanation we check the eigenvectors and optimal degrees of freedom of the two t-Student frameworks for 2007 when the South African or Brazilian yields are removed from the dataset. The panels in Figure 11.3 show the obtained eigenvectors across the term structures (x-axis) and examined countries (row-wise). When either of the time series is removed from the dataset, the order of degrees of freedom reverts. In addition, when the South African yield is removed, the two PPCA methodologies provide consistent eigenvectors. Therefore, the South African yield source of the disagreement between the two t-Student frameworks in 2007.

(a) D2 without South Africa **(b) D2** without Brazil

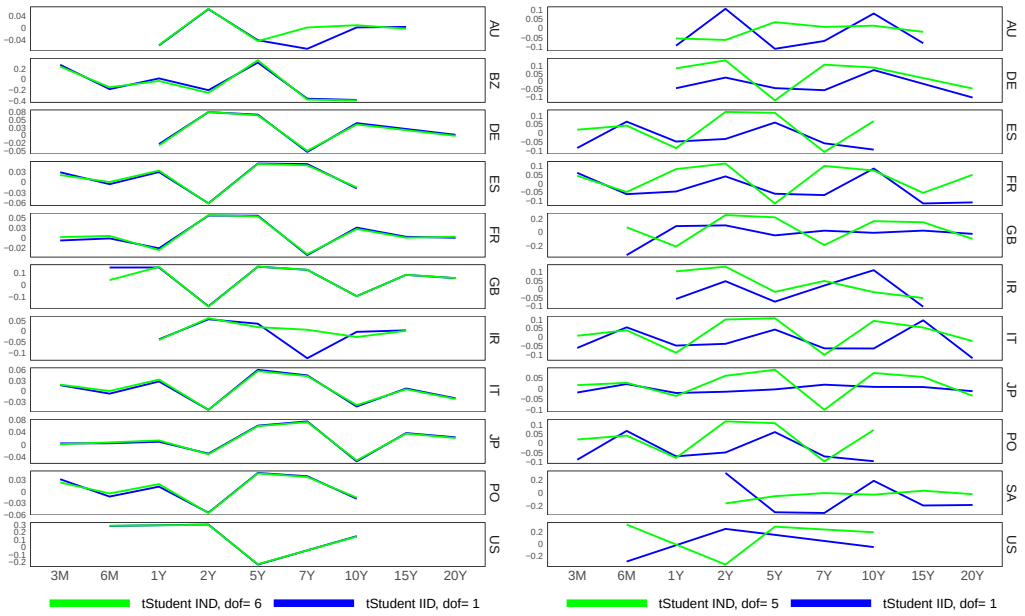


Figure 11.3. The first eigenvector of a covariance matrix for the **D2** data set of country-specific soveregin yield curves without South African or Brazilian yields in 2007 partitioned into country-specific subvectors (row-wise) and their term structure (x axis). The colours of lines correspond to diffrent PPCA frameworks: t-Student IND PPCA (green line) and t-Student IID PPCA (blue line). The optimal degrees of freedom chosen for t-Student IND PPCA and t-Student IID PPCA are 6 and 1, respectively.

11.4.2. Features Extraction for Country-Specific Inflation-Linked Yield Curves

The eigenvectors of the datasets **D3**, **D3b** and **D3s** correspond to the country specific inflation-linked yield curves and their term structures' points. The first eigenvectors over different periods of time and across countries are illustrated in Figure 11.4, 11.5 and 11.6. The corresponding eigenvalues over years are also displayed in Figure 11.7a.

Interestingly, the eigenvectors of the sets **D3** and **D3b** provide different information even though **D3** contains all-time series included in **D3b** in addition to two inflation-indexed swap yields, from Australia (AU) and Spain (ES). Surprisingly, the eigenvectors for swap curves in **D3** are significantly different from their equivalents in the set **D3s** for Australia (AU) and Spain (ES). The eigenvectors extracted from data sets **D3** and **D3b** differ more significantly across the three examined PPCA methodologies than the corresponding results of **D2** data set. If the choice of the fixed maturities was the same across these data sets, we would expect the information contained in eigenvectors of inflation-linked sovereign bond yield curves to be more consistent with the direction calculated from the conventional sovereign bonds. We notice smaller discrepancies when we compare the features between **D3s** and **D2**, especially when we analyse the yearly and country-specific patterns highlighting maturities with highest loadings.

The plots of eigenvalues are displayed in Figure 11.7a. The magnitude of the values is similar among the sets with various inflation-linked instruments. The yearly sequences of the second and thirds eigenvalue are consistent for the sets **D3** and **D3b**, whereas the results for **D3s** are greater.

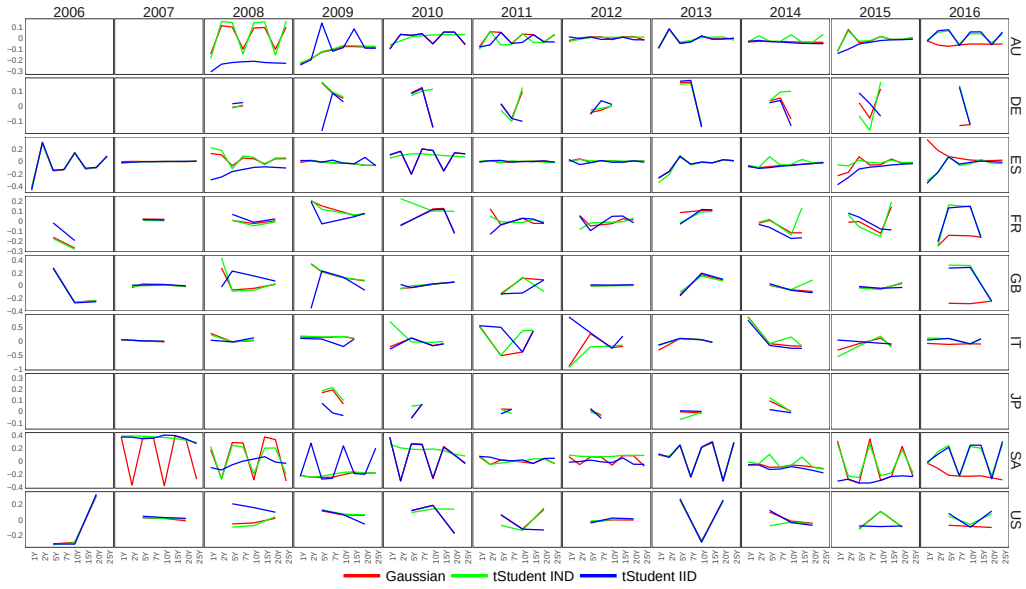


Figure 11.4. The first eigenvector of a covariance matrix for the **D3** set of inflation-linked mixed yield curves across different years (column-wise) partitioned into country-specific subvectors (row-wise) and their term structure (x-axis). The colours of lines correspond to different PPCA frameworks: Gaussian PPCA (red line), t-Student IND PPCA (green line) and t-Student IID PPCA (blue line).

The choices of degrees of freedom for two t-Student frameworks for three sets of inflation-linked rates are illustrated in Figure 11.8a. The IND t-Student PPCA exhibits more variability of the sequences of selected degrees of freedom but it fluctuates around small values indicating robust weights are beneficial

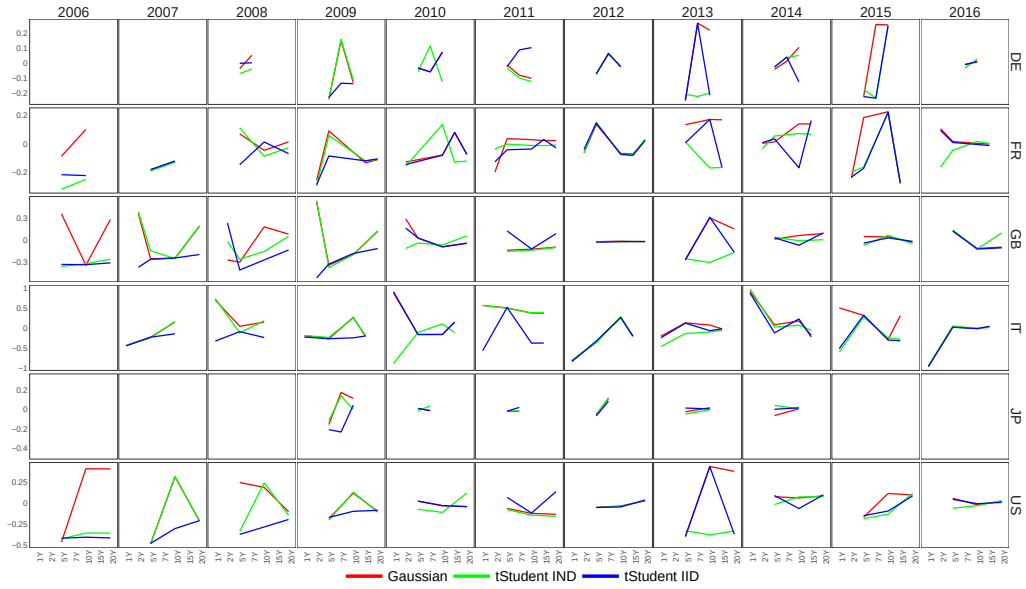


Figure 11.5. The first eigenvector of a covariance matrix for the **D3b** set of inflation-linked bond-only yield curves across different years (column-wise) partitioned into country-specific subvectors (row-wise) and their term structure (x-axis). The colours of lines correspond to different PPCA frameworks: Gaussian PPCA (red line), t-Student IND PPCA (green line) and t-Student IID PPCA (blue line).

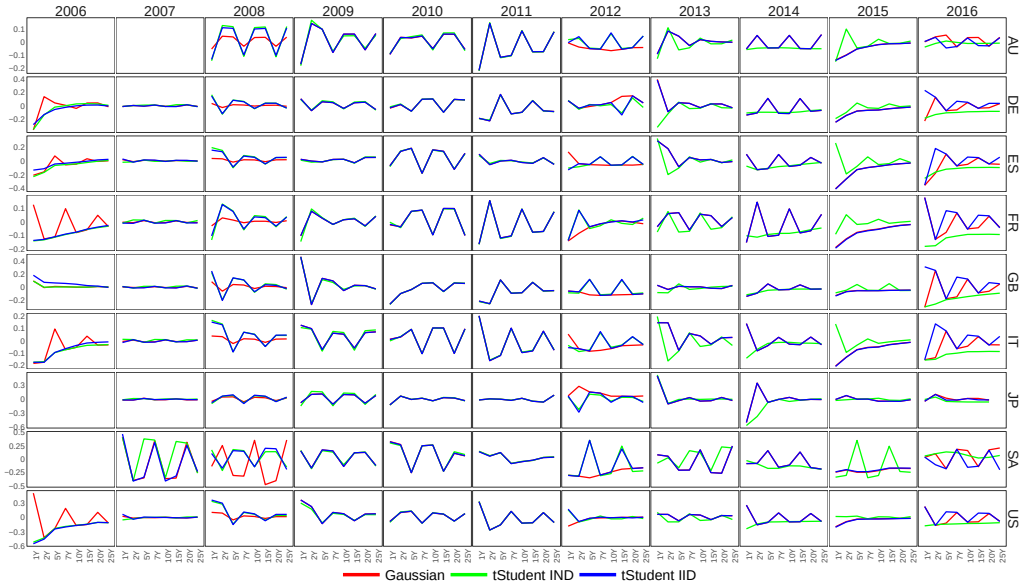


Figure 11.6. The first eigenvector of a covariance matrix for the **D3s** set of inflation-linked swap-only yield curves across different years (column-wise) partitioned into country-specific subvectors (row-wise) and their term structure (x axis). The colours of lines correspond to diffrent PPCA frameworks: Gaussian PPCA (red line), t-Student IND PPCA (green line) and t-Student IID PPCA (blue line).

(a) D3 **(b)** D3b **(c)** D3s

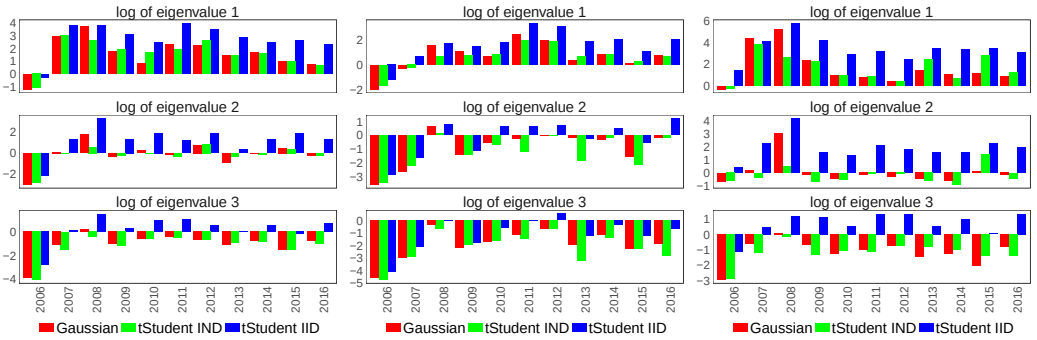


Figure 11.7. The evolution of three highest eigenvalues of covariance matrix for **D3**, **D3b** and **D3s** data sets of country-specific inflation-linked yield curves over years (x axis).

(a) D3 **(b)** D3b **(c)** D3s

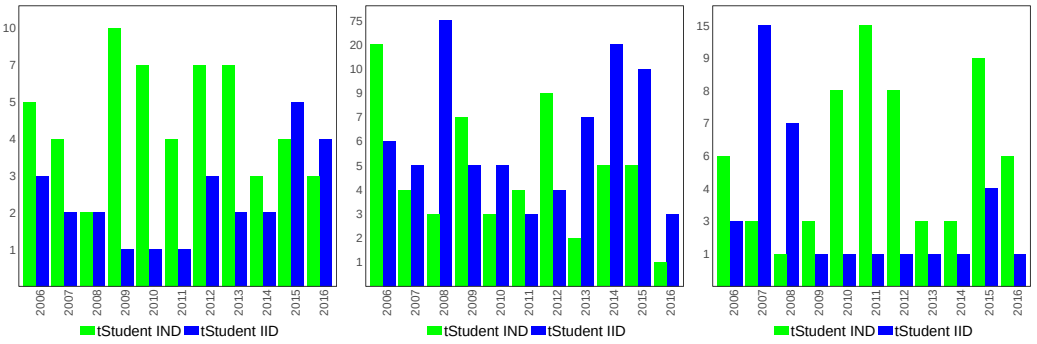


Figure 11.8. The evolution of degrees of freedom for **D3**, **D3b** and **D3c** data sets of country-specific inflation-linked yield curves over years (x axis).

11.4.3. Features Extraction for Macroeconomic Proxies of Euro-Zone Activity

The features extracted from the data set **D4** are considered in this section. Since the instruments included into the data set are vector-valued rather than matrix-valued, the resulting eigenvectors have the length corresponding to the number of instruments. Figure 11.9 shows the eigenvectors which are extracted on yearly non-overlapping windows. The row-wise order of the panels illustrates the yearly sequences of the instrument-related elements of the three most significant eigenvectors whereas the corresponding eigenvalues and selected degrees of freedom are shown in the panels of Figure 11.10b.

The two t-Student methodologies select the parameters of degrees of freedom which differ significantly across years except for 2010-2012 and 2014-2016. The discrepancies are the highest among all examined data set. Interestingly, the extracted features are very consistent across three PPCA frameworks and exhibit stronger differences only in the third eigenvector which explains the insignificant portion of the data's variability as shown in Figure 11.10a.

The first eigenvector is indifferent across years and resembles straight line for most of the instruments except for the element which corresponds to the open interests of Euro-bund features. The open interests are indicated as the most influential direction of the data's variability.

Starting with 2011 we notice the increase in the importance of the second eigenvector what is related to the presence of Markit Itraxx Index which starts to be available from 2011. The presence of the index changes the yearly structure of the second eigenvector. Until 2011, its highest values correspond to the exchange rates USDEUR, AUDEUR and GDPEUR, whereas after 2011 the only non-zero elements relate to the index.

The dominance the open interests over years in the variability of the set **D4** given both by the values of corresponding elements of the eigenvector and a significance in the disproportion between first and second eigenvalue, we can freely use only this data set as a predictor of Euro-zone economic activity. The other time series contributes more negligible to the shape of proxies indicators.

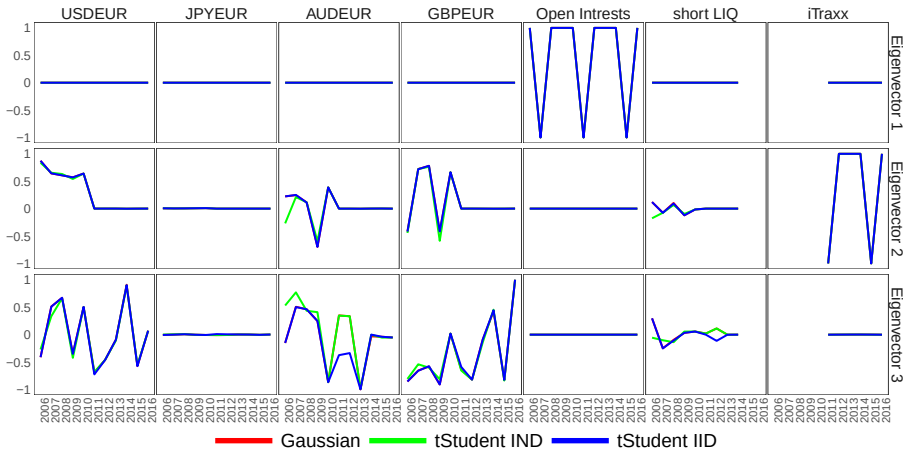


Figure 11.9. The first eigenvector of a covariance matrix for of **D4** data ser of Euro-zone activity proxies across different years (row-wise) partitioned into country-specific subvectors (column-wise) and their term structure (x-axis). The colours of lines correspond to different PPCA frameworks: Gaussian PPCA (red line), t-Student IND PPCA (green line) and t-Student IID PPCA (blue line).

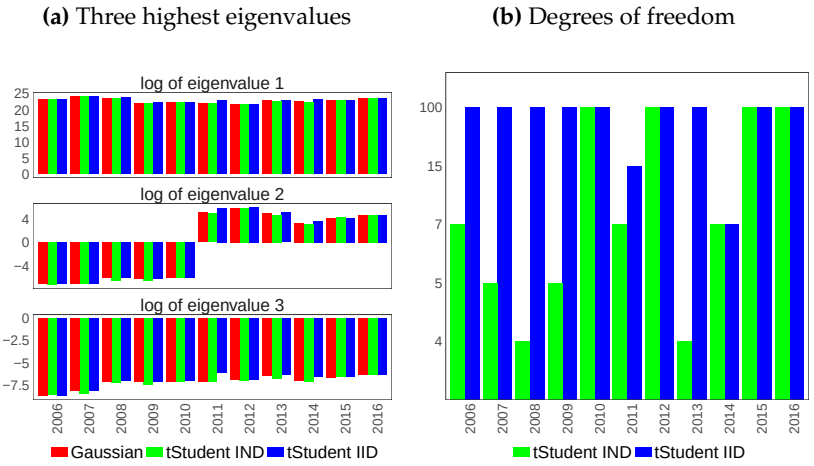


Figure 11.10. The evolution of three highest eigenvalues of covariance matrix (a) and the evolution of degrees of freedom of t-Student PPCA framework (b) over years (x axis) for the data sets which consists of country-specific soveregin yield.

11.4.4. Simmilarity of Extracted Features Across Various Financial Data Sets

Having extracted spectral features via diffrent PPCA methods, for a range of different financial datasets, we seek to study statistical similarities between the resulting features. To achieve this we analyse the strength of matrix correlation between yearly features extracted from data sets **D2**-**D4** discussed in Subsection 11.4.1- 11.4.3. This is performed using a multivariate generalization of the squared Pearson correlation coefficient (Robert and Escoufier 1976) taken between two real $m \times n$ and $m \times l$ matrices $\mathbf{A}$ and $\mathbf{B}$ according to

$$\mathrm{RV}(\mathbf{A}, \mathbf{B}) = \frac{\mathrm{Tr}\left\{\mathbf{A}\mathbf{A}^T\mathbf{B}\mathbf{B}^T\right\}}{\mathrm{Tr}\left\{\mathbf{A}\mathbf{A}^T\right\}^2 \mathrm{Tr}\left\{\mathbf{B}\mathbf{B}^T\right\}^2}.$$

Since the matrices with features are incorporated as exogenous factors to the Extended Nelson-Siegel model for Euro Libor Yield Curve, we need to ensure that the rows of matrices with features correspond to the term structure of Euro Libor curve. Consequently, we obtain the matrices with features which have consistent number of rows and are ready to calculate the correlation between them.

The last set of features extracted from data sets consists of proxies for Euro-Zone credit, liquidity and foreign exchange risks, does not have a term structure, that is, an extracted feature for yearly data is represented as a d_{euro} - dimensional vector, where d_{euro} denotes the number of scalar financial instruments in the data set **D4**. Therefore, we replicate the vector with features across the terms structure in order to obtain a new $d \times d^{euro}$ matrix, denoted by $\mathbf{F}_t^4$. The new matrix has identical values across the term structure corresponding to Euro Libor Yield Curve.

Figure 11.11 illustrates RV coefficients across years (x axis) between features extracted yearly from different examined data sets and using different PPCA frameworks. The bars correspond to pairwise RV coefficients for a given year between factors obtained using different methodologies and given different data set. The row-wise and column-wise order specifies the coefficient over

years for matrices and PPCA methodologies given by the row's and column's labels. The strength of similarity reaches values between 0 and 1 for 1 being identical matrices. Both, the height of the bars as well as the colours correspond to the values of the RV coefficient.

The indicators of strong correlations between matrices are highlighted by values above 0.75 and support the previous discussion related to the eigenvectors of data sets **M4** in Figure 11.9 which are consistent among examined PPCA methodologies. One may notice the strong yearly autocorrelation between features extracted from the data set **D4** within different PPCA methodologies highlighted by dark blue pars in the bottom right corner of the plots which correspond by rows and columns to factors from the **D4** set. In general, we see resulting higher values of RV coefficient for factors from the same data sets. The strength of similarity between the factors across different sets is mostly weak which indicates that our different classes of extracted exogenous factors may act as distinct sets of possible predictors to enter into our dynamic Nelson Siegel term structure multi-factor models. As expected, the yearly factors from **D3** and **D3b** data sets usually exhibit high correlation due to their common attributes. Interestingly, we see strong autocorrelation between yearly factors from **D2** and **D3s** datasets which is higher than between the factors from the country-specific sovereign yield curves dataset and other examined inflation-linked data sets.

Fig. The yearly RV coefficients between different sets of features

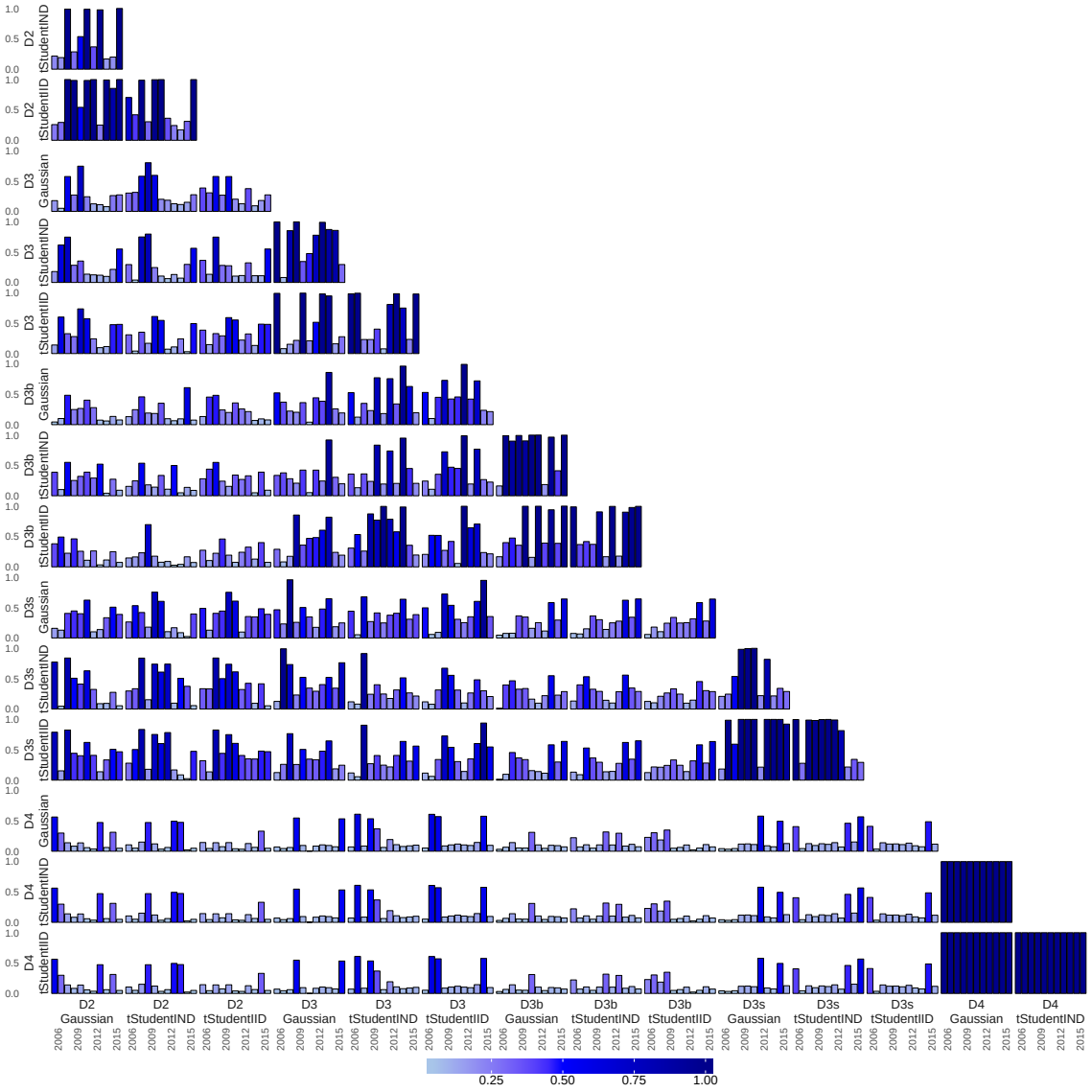


Figure 11.11. The RV coefficients between various factors over different years (x-axis). The row-wise and column-wise orders of the panels correspond to the RV between different the features extracted from sets of data using different PPCA methodologies in the same calendar year. The colours of bars refer to the value of the coefficient and the strength of similarity.

11.5. Hybrid Multi-Factor Dynamic Nelson-Siegel Yield Curve Model of Euro Libor

In the following section, we examine the predictive and explanatory ability of extracted features from macroeconomic and financial data sets. Our goal is to verify whether the yearly market features provide us with additional explanatory power in presence of the Nelson-Siegel yield regression state-space model. Given the extended model of Nelson-Siegel from Section 11.3 we use the yearly extracted features from data sets **D2-D4**, discussed in Section 11.4, as exogenous factors. The features consist of the first eigenvectors of the sets' covariance matrix calculated yearly.

Since the analysis of eigenvectors from the dataset **D4** highlights that the Euro-bund Open Interests dominate the variance of the data set, we utilise the daily time series of Open Interests as an exogenous variable in the observation equation. In addition, we examine the information contained in macroeconomic data sets with monthly or yearly observations, the raw time series of GDP, Labour Productivity, Unemployment Rates and CPI discussed in Subsection B.1.5 and Subsubsection B.1.3.1.

The analysed Euro Libor yield consists of two points at maturity 1Y which correspond to either the 12 month Euro Libor rate provided by ICE or the 12 month Euro Libor swap rate. In the following section, we verify which of these two rates is the most informative in predicting and in-sample estimation of the Euro Libor yield model.

11.5.1. The Specification of the Optimal Models for Eur Libor Yield Dynamic Analysis

The selection of the set of assumptions for the models of the Extended Nelson-Siegel class is a multiple step procedure. Therefore, before presenting the calibration results for the Euro Libor yield, we describe the four steps of which we follow in order to find an optimal model per year.

Step 1: The choice of λ.

The parameter λ specifies the shape of the Nelson-Siegel basis regression functions which affect concavity and convexity of the fitted yield curve model. The first step of the model selection is to select the best values of parameter per year for all possible variants of the standard Nelson-Siegel model (all possible combinations of variants for matrices $\mathbf{A}$, $\mathbf{Q}$, $\mathbf{R}$: diagonal or non-diagonal, heterogeneous or homogeneous described in Subsection 11.3.4).

A grid of λ is fixed for every year and is specified by obtaining daily estimates of λ from the static daily fits of the static Nelson-Siegel model using non-linear regression for the period of time 2006-2016. The grid consists of 20 equally spaced points between 97.5% and 2.5% quantiles of the daily estimated of λ. Next, we calculate the yearly estimates of the matrices $\mathbf{A}$, $\mathbf{Q}, \mathbf{R}$ for all values of λ, which belong to the grid, given different assumptions on the structure of the matrices. The estimates are obtained by maximising the likelihood of the corresponding Kalman Filter estimations. This step provides us with $20 \times 2^4 = 320$ different Nelson-Siegel models per year: (the number of points on the λ-grid $\times$ (diagonal or non-diagonal matrix $\mathbf{A}$) $\times$ (diagonal or

non-diagonal matrix **Q**) $\times$ (diagonal or non-diagonal matrix **R**) $\times$ (heterogeneous or homogeneous matrix **Q**) $\times$ (heterogeneous or homogeneous matrix **R**). Hence, we have 20 possible different values of λ for each of the 2^4 variants of the Nelson-Siegel model. Within 20 values of λ we select the one which provides the highest value of the marginalized likelihood obtained from the Kalman Filter for each variant of the Nelson-Siegel model.

Step 2: The specification of the baseline model: the assumptions on matrices A, Q, R per year.

In the previous step, we selected the optimal values of λ for each of 2^4 examined variants of the Nelson-Siegel model per year. We use Akaike Information Criterion (AIC) in order to compare between 2^4 different Nelson-Siegel models per year penalising for model complexity. The goal of this study is to examine the supplementary information provided by the various market features to the factors of standard Nelson-Siegel model: the levels, slope and curvature of a yield. Hence, we first chose a baseline model fro Nelson-Siegel class of models and as a next step, we plug in additional factors, the yearly market features, to the model under the baseline's model assumptions. We follow the two selection criteria for determining the Nelson-Siegel baseline model per year:

1. minAIC

We select two standard Nelson-Siegel models with different restrictions on the static parameters which provided the lowest AIC;

2. diagAll

We choose two models with a low complexity (the number of parameters to estimate) per year: the matrices of the model (11.2) are all diagonal, the covariance matrix **Q** is heterogeneous, the covariance matrix **R** is heterogeneous or homogeneous.

Step 2 left us with at most 4 models per year: the two models with smallest AIC and the two models with low complexity.

Step 3: The choice of the yearly market features.

The number of yearly country-specific market features is equal to $3 \times 4 = 12$. The number 4 reflects the number of the data sets **D2-D3s** discussed in Section 11.4 and 3 is the number of different PPCA frameworks used to obtain the features: Gaussian PPCA, t-Student IND PPCA and t-Student IID PPCA. The analysis of the pairwise correlation discussed in Subsection 11.4.4, indicates the presence of strong correlations between some of the features. Therefore, we proceed with preliminary feature selection methodology in order to remove the features which provide duplicated information. We choose as a yearly benchmark feature the first eigenvector from the data set **D2** obtained by the Gaussian PPCA framework. Every set of features contains at least the benchmark feature. Any other vector with features, which correlation with the benchmark feature, or any other feature already included to the analysis, is lower than 0.75, is also added to the set of yearly market features. Hence, the sets of feature might be of different sizes across years (maximum 9) and consist of at least one element: the benchmark feature. The yearly choices of features are listed in Table A4 in Section B.3 in Appendix B. We denote the sets of models with different features by **M2- M3s** where the digits correspond to the data sets **D2- D3s**.

We subset the obtained matrices of features from country specific sovereign and inflation-linked yield curves in order to take only these elements which correspond to Euro Libor term structure. If the data sets does not provide the feature for particular maturity, we set the corresponding elements of the new factor matrices to zero. Hence, the new $\mathbf{F}_t^2$, $\mathbf{F}_t^3$, $\mathbf{F}_t^{3b}$ and $\mathbf{F}_t^{3s}$ matrices with features extracted as first eigenvectors of data sets **D2 - D3s** are $d \times d^{govt}$, $d \times d^{inf}$, $d \times d^{infb}$, $d \times d^{infs}$ dimensional, respectively, where d denotes the number of elements in the term structure of Euro Libor curve, d^{govt} denotes number of country-specific Sovereign yield curve, d^{inf} - number of country-specific bond-base filled with swaps inflation-linked yield curves, d^{infb} - number of country-specific inflation-linked government bond yield curves, and d^{infs} - number of country-specific inflation-linked swap yield curves.

Furthermore, we separately examine the forecasting and calibration performance of the daily Euro-bund Open Interests since the time series of the instrument significantly dominate the variability of the data set **D4** as has been discussed in Subsection 11.4.3. The daily observations are added to the standard Nelson-Siegel models as the exogenous variable. The model which utilizes daily values of Euro-bund Open Interests is denoted by **M4**.

The daily features of monthly and yearly time series of GDP, Labour Productivity, Unemployment enter the base-line Nelson-Siegel model into the observation equation with the same magnitude across all maturities. Before the features are used for modelling, we verify that the columns of the loading matrix are collinear on yearly basis. We detect the collinearity between the time series using Variance Inflation Factors test. The test discards all the positions in the data sets except the monthly provided CPIs. We denote the model with the raw values of CPI, which are added to standard Nelson-Siegel models as an exogenous variable, by **M1**.

The standard Nelson-Siegel model is denoted by **M0**.

Step 4: The choice of the optimal model per year.

After adding the additional explanatory variables (features from the datasets **D2**-**D3**, daily observations of Open Interests or monthly observations of country-specific CPIs) to the standard Nelson-Siegel model, we use the Kalman Filter to re-estimates the matrices $\mathbf{A}$, $\mathbf{Q}$, $\mathbf{R}$ under the reference model's assumptions for all selected features. We compare the values of AIC in order to select the optimal model per year from the considered set of models: **M0**-**M4**.

11.5.2. The Comparison of the Baseline Model Selection Methodologies

Table A6 and Table A5 in Section B.3 in Appendix B compare the two baseline model selection methodologies described in **Step 2** from the previous subsection within the considered two data sets of Euro Libor rates: with **1** year ICE Euro Libor rate or with **1** year swap Euro Libor rate, respectively. The table consists of the columns with yearly model's specifications, its in-sample and out-of-sample performance. If the model belongs to the class of the Extended Nelson-Siegel models, the last two columns describe which PPCA framework and the market-related dataset were utilized in the optimal model in particular year.

The green colours of cells correspond to the years when the two baseline model selection methodologies agree on the best standard Nelson-Siegel. The blue color of cells correspond to the years when the Extended Nelson-Siegel models chosen using **minAIC** methodology obtained better in-sample performance than the Extended Nelson-Siegel models with the lower complexity, whereas the yellow color of cells correspond to the years when it achieved better out-of-sample performance measured by mean square errors of 1 day, 1 week or 1 month predictions over the next calendar year. The upper index '*' correspond to the year models belonging the Extended-Nelson-Siegel family which obtained poorer in-sample performance than their reference models from the standard Nelson-Siegel class.

Table A6 shows that the models which provide the best in-sample and out-of-sample fit for Euro Libor rates with **1** year ICE Euro Libor rate according the **minAIC** criterion have mostly either non-diagonal transition matrix $\tilde{\mathbf{A}}$ or the covariance matrix $\tilde{\mathbf{R}}$. The models chosen according to the selection methodology **diagAll** outperformed these models according to the in-sample calibration only in 2007 and according to the out-of-sample performance (1 day prediction) in 2008, 2009, 2011, 2012 and 2015.

The best models which are selected for Euro Libor rates with **1** year swap rate have mostly all matrices diagonal. The exceptions are the years 2006-2007 and 2014-2015. It is observed that for this data set we see particular years when the models with market features achieved poorer in-sample performance than their reference models. It is a case for the models selected according to **minAIC** in years 2007, 2009 and 2010. What is more, we observe only three years when the models selected according to the **minAIC** methodology performed better than the models with lower complexity, the years 2006 and 2014-2015 and the improvement of the in-sample performance did not ensure better forecasting results.

In general, the differences between the models selected according to the two baseline model selection methodologies which are highlighted by blue and yellow colours, are not significant in terms of their in-sample and out-of-sample performance. Considering the fact that the Extended-Nelson-Siegel models specified according to the **minAIC** methodology are more prone to calibrate the curve poorer than their simpler reference model, we drop this group of results from the following analysis.

11.5.3. In-Sample Fit of the Best Models for Euro Libor Yield Curve

The calibration of Euro Libor yield curve with either 1Y ICE Libor rate or 1Y swap rate is the most accurate for the hybrid models from Extended Nelson-Siegel family which significantly outperform the standard Nelson-Siegel model fits. We show that only for the short rates in the crisis period, the estimation of the yields carried by the standard Nelson-Siegel models is as good as by more complex models.

The Table 11.2 provides the description of the Extended Nelson-Siegel and the standard Nelson-Siegel models which are selected on a yearly basis (the first column) as the best models for two Euro Libor yields as well as their in-sample performance results. The column **Model** lists the acronyms of the best models corresponding to **Step 3** in Subsection 11.5.1 (in case of

Extended Nelson-Siegel family it refers to a specific dataset of features). If applicable, the column **PPCA** indicates the PPCA framework used to obtain the market feature, the column **R** describes the structure of the covariance matrix $\tilde{\mathbf{R}}$ and, lastly, the column $\log\lambda$ informs about the value of the shape parameter of the Nelson-Siegel model. The models across the years have diagonal transition and covariance matrices and heterogeneous matrix **Q** as imposed by the baseline model selection methodology **diagQ**. The repeatedly chosen structure of $\tilde{\mathbf{R}}$ is heterogeneous allowing for differences in the magnitude of the latent vector's elements variability. The smallest values of λ for the Euro Libor dataset with **1Y swap rate** are selected at the beginning and the end of the sample set, years 2006 and 2007 indicating that the shape of the yield across the term structure is the flattest among examined periods. The highest values of the shape parameter correspond to the crisis period, 2008-2009 when the relative differences between the Euro Libor rates are the biggest. The shapes of the Euro Libor yield with **1Y ICE Euro Libor rate** do not follow this pattern, indicating rapid decline of the parameter's value in the middle of the sample period, in 2010.

Table 11.2. The model specification and in-sample statistics for the best standard Nelson-Siegel model and the Extended Nelson-Siegel models chosen by **diagAll** per calendar year (the first column) for the two sets of Euro Libor yield: with 1Y ICE Euro Libor rate or 1Y swap rate. The column **Model** lists the acronyms of the best models corresponding to **Step 3** in Subsection 11.5.1 (in case of Extended Nelson-Siegel family it refers to a specific data set of features). The column **PPCA** indicates the PPCA framework used to obtained the market feature (1 -Gaussian PPCA, 2-t-Student IND PPCA and 3- t-Student IID PPCA), the column **R** describes the structure of the covariance matrix $\tilde{\mathbf{R}}$ and the column $\log\lambda$ informs about the value of the shape parameter of the Nelson-Siegel model. The columns l, AIC and logMSE provide the yearly values of measures of the estimation accurancy: loglikelihood of the Kalman Filter, Akaike Information Criterion statistics from Equation (9.1) and the logarithm of the mean square error defined in Chapter 9.

Year	1Y ICE Euro Libor rate							1Y swap rate						
	Model	PPCA	R	$\log\lambda$	l	AIC	logMSE	Model	PPCA	R	$\log\lambda$	L	AIC	logMSE
2006	M0		hete	-1.42	5922.73	-11843.46	-3.99	M0		hete	-1.42	6164.91	-12327.83	-4.07
	M2	2	hete	-1.42	7898.12	-15794.25	-4.30	M2	2	hete	-1.42	8101.40	-16200.80	-4.36
2007	M0		hete	-0.76	5965.95	-11929.90	-3.16	M0		hete	-0.96	6256.79	-12511.59	-3.23
	M2	1	hete	-0.76	7305.62	-14609.24	-3.67	M2	1	hete	-0.96	7114.59	-14227.19	-3.99
2008	M0		hete	-0.52	3331.25	-6660.50	-3.03	M0		hete	-0.48	3581.24	-7160.47	-2.64
	M3	3	hete	-0.52	4841.08	-9680.15	-3.98	M3	3	hete	-0.48	5077.73	-10153.47	-3.81
2009	M0		hete	-0.52	5184.72	-10367.44	-4.27	M0		hete	-0.48	5679.58	-11357.15	-4.47
	M2	2	hete	-0.52	6764.61	-13527.22	-5.64	M2	2	hete	-0.48	7108.11	-14214.22	-5.77
2010	M0		hete	-2.18	5422.09	-10842.19	-3.66	M0		hete	-0.64	6027.31	-12052.63	-4.31
	M2	2	hete	-2.18	7650.23	-15298.46	-6.20	M2	2	hete	-0.64	7450.24	-14898.48	-5.81
2011	M0		hete	-0.52	4357.33	-8712.66	-3.40	M0		hete	-0.56	5191.04	-10380.08	-4.26
	M3	1	hete	-0.52	6502.81	-13003.63	-5.19	M3	1	hete	-0.56	6513.74	-13025.47	-5.07
2012	M0		hete	-0.52	4368.44	-8734.87	-2.90	M0		hete	-0.74	4377.26	-8752.52	-4.66
	M2	2	hete	-0.52	7773.09	-15544.17	-6.23	M2	2	hete	-0.74	7535.36	-15068.73	-5.76
2013	M0		hete	-0.83	5738.36	-11474.72	-4.69	M0		hete	-0.74	5390.17	-10778.33	-5.29
	M2	3	hete	-0.83	7854.03	-15706.06	-6.55	M2	3	hete	-0.74	8128.64	-16255.27	-6.72
2014	M0		hete	-0.91	5225.34	-10448.68	-4.43	M0		hete	-0.96	5638.23	-11274.46	-5.40
	M2	2	hete	-0.91	8462.48	-16922.96	-6.42	M2	2	hete	-0.96	8525.73	-17049.46	-6.47
2015	M0		hete	-0.99	6243.56	-12485.13	-4.93	M0		hete	-0.96	6639.60	-13277.21	-5.87
	M2	1	hete	-0.99	8886.02	-17770.05	-6.63	M2	1	hete	-0.96	8927.51	-17853.02	-6.67
2016	M0		hete	-0.99	5150.36	-10298.73	-4.69	M0		hete	-1.09	6108.33	-12214.66	-5.54
	M2	1	hete	-0.99	8595.78	-17189.57	-6.98	M2	1	hete	-1.09	8908.23	-17814.46	-7.15

Regardless of the different 1Y rates, we observe the agreement in the choice of the sets of assumption on the static parameters for Nelson-Siegel and Extended Nelson-Siegel models per year. Recall that the estimates of the parameters may still differ. The utilized market features mostly belong to the data set **D2** (the model **M2**) with the country-specific sovereign yields discussed

in Subsection 11.4.1, except for the year 2008 and 2011 when the inflation-linked features from the data set **D3** are chosen to be more relevant. Most of the features selected for the Extended Nelson-Siegel models are extracted using robust frameworks highlighting the importance of the information one may omit when does not account for outliers in the data. The robust features are confirmed to provide more explanatory power in most of the cases.

The features from the data set **D2** in the year 2006 extracted by the two t-Student PPCA frameworks are characterized by strong correlation reaching 0.7. However, the Independent t-Student methodology provides better in-sample performance. The reader may recall the discussion of the features from the data set **D2** in 2007 in Subsection 11.4.1 which are extracted using different PPCA frameworks. In 2007 all the features from the data **D2** set are very distant but the smoother eigenvectors obtained by Gaussian PPCA are shown to obtain the explanatory power of the Euro Libor dynamics. In 2009 the most informative features are obtained using Independent t-Student PPCA framework whereas the features from the same set next year obtained by two t-Student frameworks reveal high correlation and are similarly informative. The best features selected for 2011 which contain inflation-related information are extracted using non-robust methodology. In 2012, the Gaussian and t-Student IID PPCA provide the similar eigenvectors which have less explanatory power than the vectors obtained by t-Student IND PPCA. The opposite relation is present in 2013 when the results of the t-Student IND PPCA agrees with the Gaussian PPCA but capture fewer dynamics of corresponding Euro Libor yields. The features selected to model the yields in 2014 and 2013 are distant from the other choices of factors from the data set **D2** whereas the 3 vectors with features align in 2016.

11.5.3.1. The Calibration of the Euro Libor Yield Curve using 1Y ICE Euro Libor or 1Y swap rates

We observe that the difference of the time series of the rates at 1Y maturity impacts on resulting values of AIC, especially related to the standard Nelson-Siegel models fit. Before and during the crisis period, 2006-2010, the optimal models for the set with **1Y swap rate** obtained lower AIC values whereas after 2012 this trend reverts. We want to verify if the information contained in different 1Y rates influences the calibration of the other terms of the Euro Libor yield. The Figure 11.12 illustrates the yearly mean square errors of in-sample fit on the log scale across the term structure for different classes of the models and data sets. The means square error for the Kalman Filter is defined in Chapter 9.

Recall that the only difference between the two data sets is a time series of the rate at 1Y maturity. Hence all the results illustrated in Figure 11.12 except the ones corresponding to the 1Y rate, give us the intuition which of the time series of the 1Y rate performs better in the calibration of the other tenors of the yield according to MSE. The column-wise order of the panels in Figure 11.12 corresponds to the calendar years whereas the row-wise order of the panels refers to the results for different maturities. Every panel shows the MSE statistics for a particular year and rate. The grey bars presents results for the Nelson-Siegel models and the green ones for the Extended Nelson-Siegel models. The colour of borders corresponds to the different data sets: red - with

the 1Y ICE Euro Libor rate, blue - with the 1Y swap rate. Comparing between the two model classes we observe that in all maturities and all years the Extended Nelson-Siegel family of models outperformed the classical standard Nelso-Siegel model.. Hence, we can conclude that using market features benefited all tenors of the yield. The results do not provide the clear information

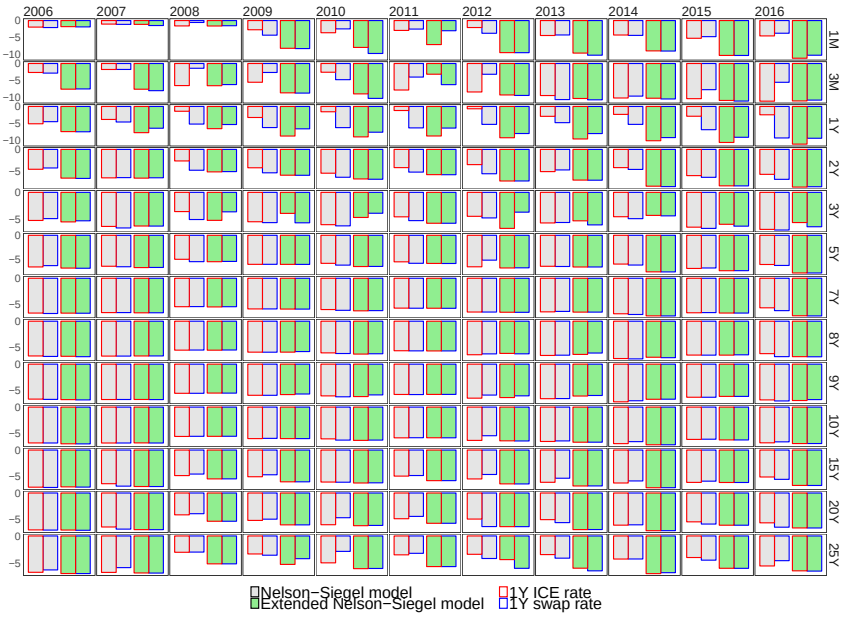


Figure 11.12. The logarithms of mean square errors (MSE) per year (column-wise) and tenor (row-wise) of best Nelson-Siegel (grey colour of bars) and Extended Nelson-Siegel (green colour of bars) models. The color of bar correspond to different data sets, Euro Libor yield with 1 **year ICE Euro Libor rate** (red) or 1 **year swap Euro Libor rate** (blue).

which rate is more informative for calibration of the yield in considered period. The Nelson-Siegel model calibrated using 1Y ICE Euro Libor rate more accurately estimate the rates at 1M and 3M rate than the model fitted on the dataset containing the swap rate, with the exceptions in 2009 and 2012 for the 1M rate and, 2010 and 2013 for the 3M rate. In general, the corresponding Extended Nelson-Siegel model performs better in the calibration of the short end of the curve, when the 1-year swap rate is present in the sample set. Only in 2011 the hybrid model, which is fitted to the Euro Libor data with 1Y ICE rate, outperforms all other examined model. Recall that for the hybrid model estimated this year we use the market features extracted from the data set **D3** with national inflation-linked yields.

The impact of the data sets on the calibration of the rates at 2Y to 5Y maturity is the highest for 2006-2013 when the choice of the rate at 1Y maturity is year-specific. Th middle terms, 7Y-20Y, are calibrated with the consistent accuracy regardless of the data set. Also, the rates at the maturities between 5Y and 10Y are similarly estimated by both classes of models. The rates at longest maturities are more successfully calibrated using the set containing 1Y swap rate and market-related features.

The panels in Figure 11.13 shows the resulting estimates of the Euro Libor rates at various maturities against their actual values over time. The column-wise order of the panels corresponds to different model classes and the data sets. The red colour of lines and shading correspond to the 1 step ahead in-sample predictions given by $\mathbf{y}_{t|t-1}$ with corresponding 97.5% confidence

intervals which are carried out on the dataset containing the 1Y ICE Euro Libor rate. The blue colour of lines and shading refers to the results for the data set with 1Y swap rate. Again, the displayed time series of the actual observations differ only for the rate at 1Y maturity. Hence, the reference observation over time in the panels corresponding to 1Y maturity is different for the results highlighted by red and blue colour. We can observe the aforementioned lower accurately of the estimation given by Extended-Nelson Siegel model for the rates at short maturities, 1M and 3M for a crisis period. The gap between the actual observations and their predictions is visible for the hybrid models and the corresponding confidence intervals are wide. On the other hand, the estimation given by the standard Nelson-Siegel model is very volatile (it jumps around the actual curve) what results in the comparable values of MSE as discussed above. In 2011, we see that the red Extended Nelson-Siegel model is most accurate in estimating the 1M rate and is the worst for 3M rate whereas the pattern reveres for the Extended Nelson-Siegel model fitted to the data set with the swap rate. In addition, the panels which refer to the estimation of the rates at the longer maturities confirm that the Extended Nelson-Siegel models should be used in calibration of these rates.

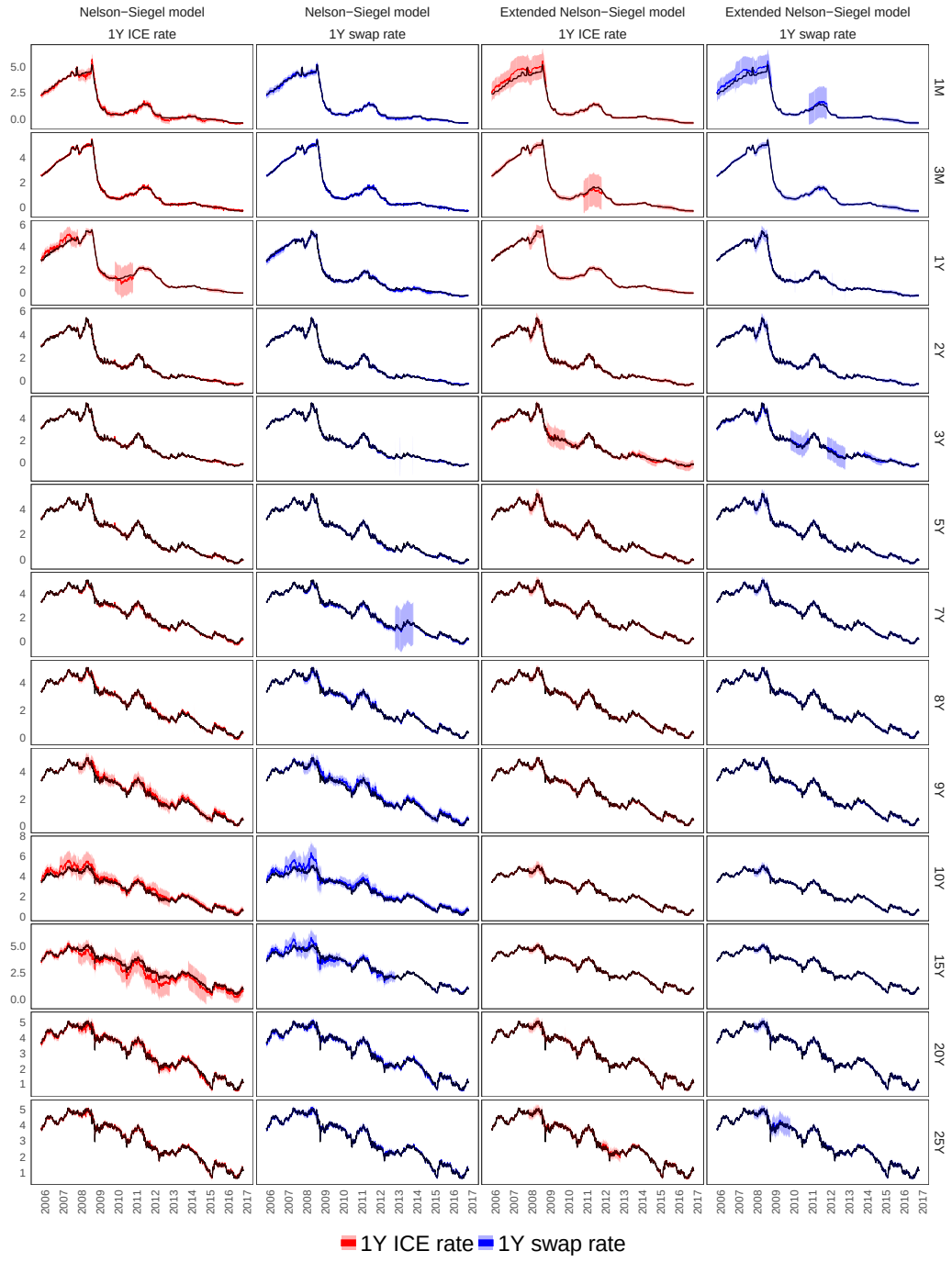


Figure 11.13. The market observed yields of the Euro Libor rates at various maturities overtime against the conditional mean of in-sample 1 step ahead prediction $\mathbf{y}_{t|t-1}$ with corresponding predictive 97.5% confidence intervals. The column-wise order of the panels corresponds to the results of the models which belongs to different model classes: the standard Nelson-Siegel or the Extended Nelson-Siegel family, and to the different set of the data Euro Libor yield with 1 **year ICE Euro Libor rate** (red lines and shading) or 1 **year swap Euro Libor rate** (blue lines and shading). The row-wise order of the panels corresponds to different terms of the yield.

11.5.3.2. Filtering of Latent Variables

Since the market features utilized in the Extended Nelson-Siegel model are country-specific we can analyse the results in terms of the strength of the influence of each examined countries on Euro Libor yield. The panels in Figure 11.14 and 11.15 shows the filtered values of the coefficient

loadings, the latent state vector, over time given by the conditional mean $\tilde{\beta}_{t|t}$ from Equation 11.4 with corresponding 97.5% confidence intervals, for the Nelson-Siegel and the Extended Nelson-Siegel models over time, respectively. The illustrated time series are continuous only within a one year period since the models are optimized every year. The blue colour of lines corresponds to the results associated with the data set containing **1Y swap rate**. The dynamics of different components of the latent process are ordered by rows. The latent processes corresponding to level, slope and curvature have the consistent dynamics for the two datasets. The only exception is the year 2010 when the levels of the latent processes shifted upwards when the **1Y ICE Euro Libor rate** is present in the data set. The reader may recall the model's specification given in Table 11.2 and the rapid decrease of the shape parameter λ for this year. Hence, the reason for the shift is the significant change in the shape of the polynomials of the baseline Nelson-Siegel model for the data set with **1Y ICE Euro Libor rate**. When considering the other country-specific components of

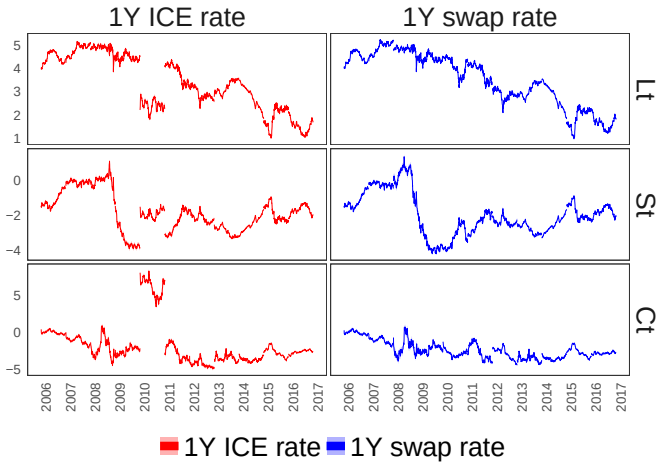


Figure 11.14. The conditional mean of filtered latent states $\beta_{t|t}$ of the coefficient loadings with corresponding confidence intervals of the Nelson-Siegel models over time which are optimal in term of **the AIC criterion**. The column-wise order of the panels corresponds to the models for Euro Libor yield with 1 **year ICE Euro Libor rate** (the left column) or 1 **year swap Euro Libor rate** (the right column). The row-wise order of the panels corresponds to different terms of the yield.

the latent vectors of the Extended Nelson-Siegel model, we may specify the list of countries which have the biggest impact of the Euro Libor yields across the examined period of time. Recall that the majority of the Extended Nelson-Siegel models utilized the market features extracted from the data sets **D2** with international sovereign yields. Only in 2008 and 2011, we use the features from the data set **D3** with inflation-linked yield curves.

The country-specific time series which fluctuate around zero informs us that this specific country does not have an impact on Euro Libor yields. With the beginning of the crisis, we can observe that all the latent vectors corresponding to the European countries are characterized by non-zero dynamics with increased volatility, especially Portugal, Ireland, Spain and France. In addition, we see the increased activity of the processes corresponding to international markets, Australia, Japan and Brazil. The influences of other countries start to grow in 2009. However, for this year we see the discrepancies between the information provided by the two different Euro Libor data set. The dataset with 1Y swap rate indicates the lack of influence of the French

features whereas the other set in the opposite. In the recent years, we observe that most of the country-specific latent processes fluctuate around zero what agrees with the previous remark that the fits provided by the standard Nelson-Siegel are closely informative to the fits of the models with the market features. We observe the higher shifts of the levels of the processes and the increase of their volatility in 2016

In the majority of the considered periods, the difference between the inclusion either the original 1Y Euro Libor rate or the swap rate affects only the magnitude of the movements of the country-specific processes what is also linked to the different estimated of the models' statics parameters. Mostly, the two sets with different 1Y time series indicate similar set of countries having an impact on Euro Libor curve per year.

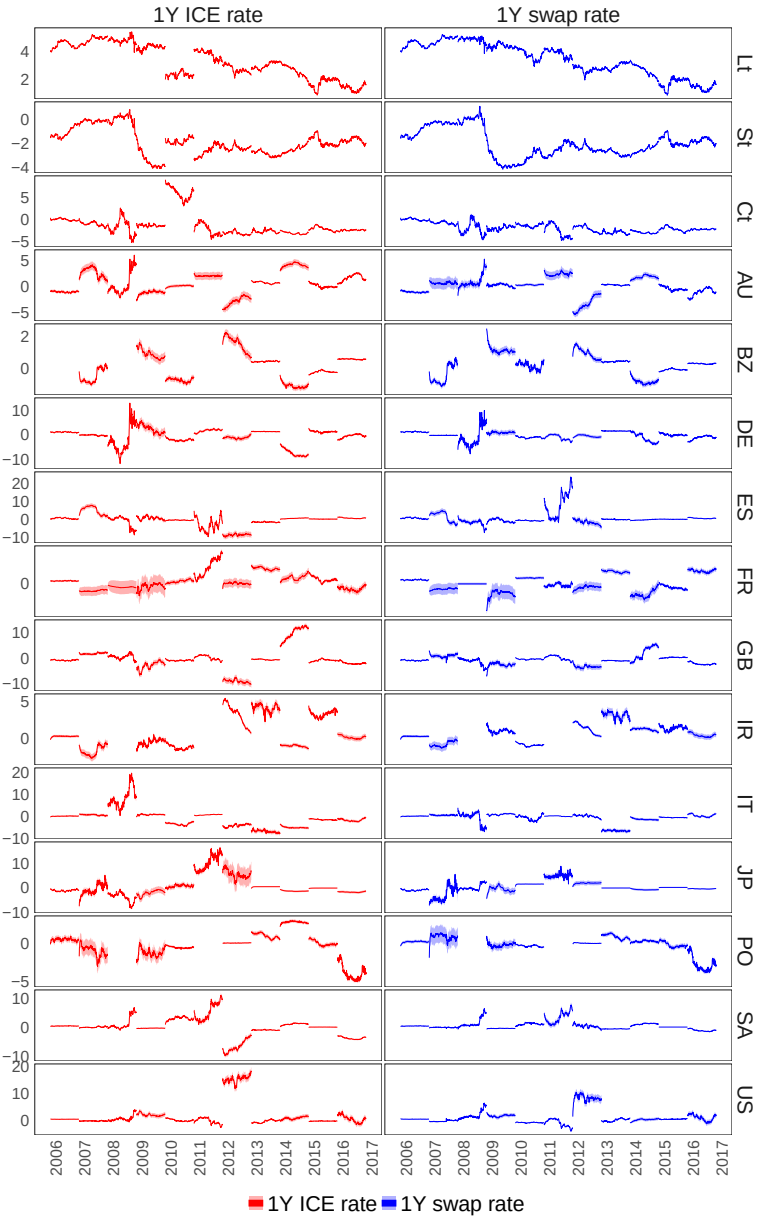


Figure 11.15. The conditional mean of filtered latent states $\tilde{\boldsymbol{\beta}}_{t|t}$ with corresponding confidence intervals of the Extended Nelson-Siegel models over times which are optimal in term of **the AIC criterion**. The column-wise order of the panels corresponds to the models for Euro Libor yield with 1 **year ICE Euro Libor rate** (the left column) or 1 **year swap Euro Libor rate** (the right column). The row-wise order of the panels corresponds to different terms of the yield.

11.5.4. Forecasting Performance of the Extended Nelson-Siegel Models with Macroeconomic Factors

We use our models to produce term-structure forecasts at 1-day, 1=week and 1-month horizons, with encouraging results. The out-of-sample prediction of the Euro Libor yield with 1 and 5 steps ahead using Extended Nelson-Siegel models outperforms the results provided by the standard models. When we increase the forecast horizon to 1 month, the discrepancies between the models from different families decrease and the Nelson-Siegel models provide a more accurate out-of-sample prediction, especially for shorter rates before 2011. What is more, the two model

classes provide mostly consistent forecasts for the rates at middle terms maturities, between 5Y and 20Y.

The prediction is performed on the 1 day sliding window. Let us denote k a forecasting horizon. We fit the models in the calendar year Y_0 (T observations) and take as an out-of-sample period the next calendar year, Y_1 (T_1 observation points). Then, we predict the value of a rate in time $T+t+k$ where $t = 1, \ldots, T_1 - T$, using the window of T consecutive observations from $t, \ldots, T+t$. Recall that the model is previously fitted to the data in time $1, \ldots, T$.

We use the Kalman Filter is order to obtain forecasting distribution. The Bayesian state-space framework allows us to obtain the forecasting distributions given by the Kalman Filter discussed in Section 7.3, that is

$$\begin{aligned}\tilde{\boldsymbol{\beta}}_{T+t+k|T+t+k-1} &\sim \mathcal{N}\left(\tilde{\mathbf{c}} + \tilde{\mathbf{A}}\tilde{\boldsymbol{\beta}}_{T+t+k-1}, \mathbf{P}_{T+t+k|T+t+k-1}\right)\\ \mathbf{y}_{T+t+k|T+t+k-1} &\sim \mathcal{N}\left(\tilde{\boldsymbol{\Lambda}}_T\tilde{\boldsymbol{\beta}}_{T+t+k}, \mathbf{F}_{T+s+k}\right).\end{aligned}$$

where $\tilde{\boldsymbol{\Lambda}}_T$ is a coefficient matrix with feature up to the end of the in-sample period, the time T. The Table 11.3 compares the forecasting performance of the examined models for different sets of sample data, with **1Y swap rate** and **1Y ICE Euro Libor rate**, respectively. The measure which we use to asses the out-of-sample performance of the models is MSEP defined in Section 7.3. The MSEP-related columns in Table 11.3 show the logarithms of the mean square errors of 1 day, 1 week and 1 month ahead predictions of the Euro Libor yield given the data from the next calendar year using the following conditional means: $\mathbf{y}_{t|t-1}$, $\mathbf{y}_{t|t-5}$, $\mathbf{y}_{t|t-22}$. The yellow shading of the cells highlights the examples when the standard Nelson-Siegel model outperforms the hybrid model. The reader can observe, that for the majority of years and the prediction horizons 1 day and 1 week, the Extended Nelson-Siegel models provide better out-of-sample forecasting performance and hence the information contained in the market features is meaningful for the out-of-sample prediction of Euro Libor yield in addition to the in-sample calibration properties. The discrepancies between the the two classes of models decline when we move to one month ahead prediction, especially for the calendar years belonging to the pre- and crisis period. The Nelson-Siegel model for 1 month ahead prediction obtains better results for 2008 and 2011 when the 1Y swap rate is present in the data set, or only in 2011 when we use the corresponding swap rate.

Similarly to the in-sample analysis, we can study the tenor-specific out-of-sample forecasting performance of the models and verify if the quality of prediction is consistent across the tenors. We calculate the maturity and calendar year specific MSEP for two data sets with different 1Y rates and two families of models. The Figure 11.16, Figure 11.17 and Figure 11.18 illustrate the MSEP of 1 day, 1 week and 1 month predictions across different years and maturities. Regardless of the prediction horizon, the dependencies between the forecasting performance between the two classes of models, the standard and hybrid one, are the smallest for the rates at maturities between 2Y and 20Y. It is worth to remark that the prediction of the models calibrated on the data set with 1Y ICE Euro Libor rate provides in majority the best results for the rates with maturities longer than 5Y for all prediction horizons. The exceptions is year 2012 for the rate at 25Y maturity for

Table 11.3. The forecasting performance of the the best Extended Nelson-Siegel and Nelson-Siegiel models for the Euro Libor yield curve with **1 year ICE Euro Libor rate** and **1 year swap rate** per year.

Year	1Y ICE Euro Libor rate	log MSEP			1Y swap rate	log MSEP		
	Model	1D	1W	1M	Model	1D	1W	1M
2006	M0	-3.93	-3.78	-2.94	M0	-4.01	-3.85	-3.26
	M2	-6.35	-4.93	-3.41	M2	-6.20	-4.96	-3.59
2007	M0	-1.95	-2.09	-1.88	M0	-2.24	-2.36	-1.75
	M2	-4.90	-3.33	-2.01	M2	-4.92	-3.42	-1.94
2008	M0	-4.10	-3.74	-2.81	M0	-4.43	-3.89	-2.65
	M3	-5.70	-3.96	-2.57	M3	-5.46	-3.56	-2.05
2009	M0	-3.77	-2.77	-1.26	M0	-3.80	-2.76	-1.38
	M2	-5.60	-3.07	-1.45	M2	-5.66	-3.16	-1.53
2010	M0	-2.29	-2.29	-1.78	M0	-4.18	-3.62	-2.34
	M2	-5.52	-3.69	-2.22	M2	-6.02	-4.51	-2.68
2011	M0	-2.90	-2.90	-2.23	M0	-3.49	-3.18	-2.12
	M3	-6.16	-4.23	-1.99	M3	-6.02	-3.98	-1.70
2012	M0	-3.86	-3.69	-3.03	M0	-4.79	-4.43	-3.52
	M2	-7.31	-5.79	-4.13	M2	-7.26	-5.82	-4.17
2013	M0	-4.30	-4.24	-2.97	M0	-4.91	-4.81	-2.85
	M2	-7.18	-5.74	-3.73	M2	-7.19	-5.88	-3.74
2014	M0	-4.86	-4.42	-2.92	M0	-5.83	-4.93	-2.92
	M2	-7.05	-5.25	-3.17	M2	-7.01	-5.21	-3.01
2015	M0	-4.34	-4.17	-3.55	M0	-5.32	-5.05	-3.97
	M2	-7.77	-6.31	-4.49	M2	-7.69	-6.29	-4.38

prediction 1 day and 1 week ahead, when the Extended Nelson-Siegel model calibrated on the data set with 1Y ICE rate obtained the poorest results among all examined models. Except for 2006 and 2012, the hybrid models exhibit significantly better forecasting properties of the rate at highest maturity.

With regards to the rates at short maturities, there is no clear pattern which indicates a model with the best prediction across years for particular maturity. The rates at shortest maturities, 1M and 3M, are more successfully predicted one step ahead by Extended Nelson-Siegel models, except for 3M rates in 2008 and 2011 when the standard Nelson-Siegel model calibrated on the data set with 1Y ICE rate obtained the smallest MSEP. The similar pattern refers to the 1-week prediction of these rates. However, when we move the forecasting horizon to 1 month the less complex standard Nelson-Siegel models predict the short end of the Euro Libor curve at least as successfully as the hybrid models with the country-specific market features, especially for years 2007-2011. Hence, the better performance 1 month ahead prediction by the Nelson-Siegel models in 2008 and 2011 noted in Table 11.3 is a consequence of their more accurate forecast of the short end of the yields. In addition, the results of 1 month ahead prediction are characterized by the biggest discrepancies between the models calibrated on different Euro Libor data sets. It is not a surprising outcome since the inaccuracy of the prediction starts to dominate the results due to the long forecast horizon.

The panels on Figure A12 in Section B.3 in Appendix B and Figure 11.19 show the one step ahead out-of-sample prediction with corresponding 97.5% confidence intervals across different maturities and for different models. The Figure 11.19 illustrates the portions of the forecast when the biggest discrepancies are present. The blue colour of lines highlights the results for models calibrated on the data set with 1Y swap rate. Hence, the panels which show the results for 1Y rate, consist of the different set of data for models stressed by blue and red lines and shading. Recall that the out-of-sample prediction is conducted on the next calendar year to the year when a

model was estimated. Hence the results plotted on the panel of Figure A12 for the year Y_0 show the forecasting ability of a model calibrated in the year Y_1. The most remarkable discrepancies of the various model's forecast are noted for short maturities, 1M and 3M and the longest maturity 25Y, especially for the year before 2013. We see substantial differences between two classes of models predicting 1Y ICE Euro Libor rate. The Extended Nelson-Siegel model more accurately capture the future movement of the rate. What is more, both hybrid models are characterized by the higher confidence of the results for 1Y rates as in majority the confidence intervals of reference Nelson-Siegel are wider.

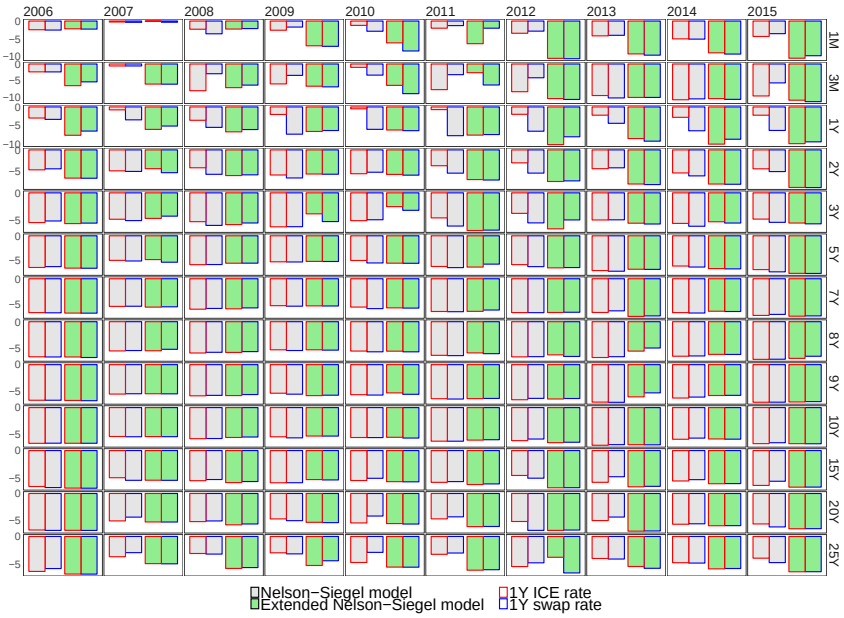


Figure 11.16. The logarithms of Mean Square Errors of 1 day prediction per year (column-wise) and tenor (row-wise) of optimal Nelson-Siegel(NS) and Extended Nelson-Siegel (ENS) models (x-axis). The color of bar correspond to different data sets, Euro Libor yield with 1 **year ICE Euro Libor rate** (red) or 1 **year swap Euro Libor rate** (blue).

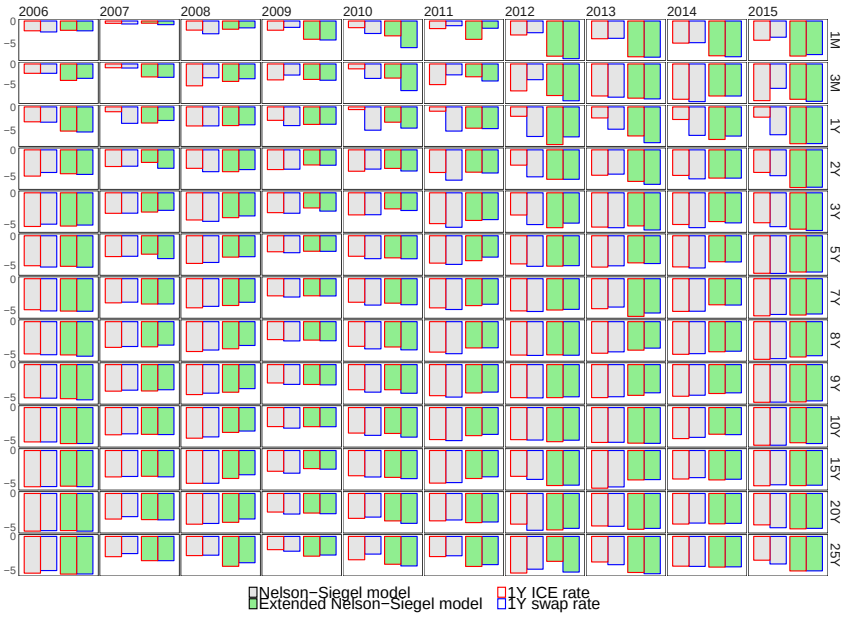


Figure 11.17. The logarithms of Mean Square Errors of 1 week prediction per year (column-wise) and tenor (row-wise) of optimal Nelson-Siegel(NS) and Extended Nelson-Siegel (ENS) models (x-axis). The color of bar correspond to different data sets, Euro Libor yield with 1 **year ICE Euro Libor rate** (red) or 1 **year swap Euro Libor rate** (blue).

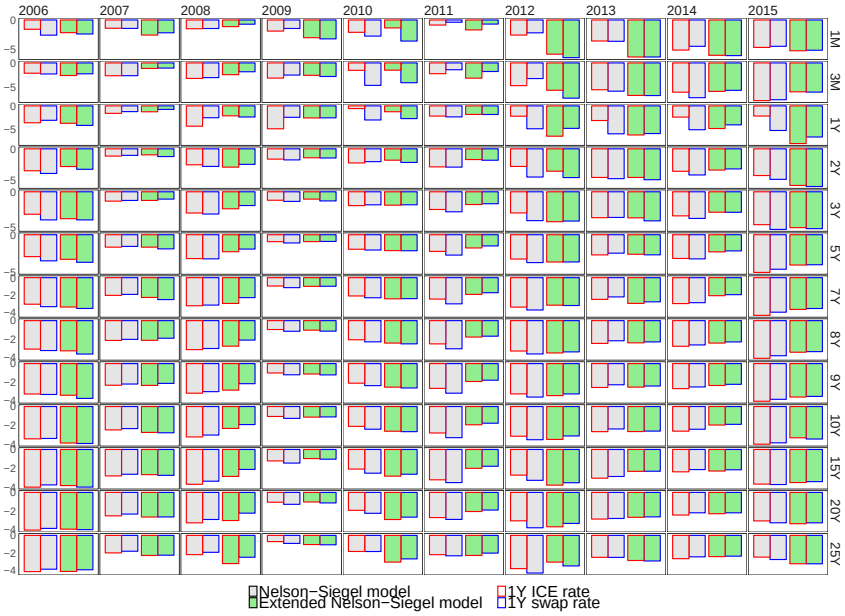


Figure 11.18. The logarithms of Mean Square Errors of 1 month prediction per year (column-wise) and tenor (row-wise) of optimal Nelson-Siegel(NS) and Extended Nelson-Siegel (ENS) models (x-axis). The color of bar correspond to different data sets, Euro Libor yield with 1 **year ICE Euro Libor rate** (red) or 1 **year swap Euro Libor rate** (blue).

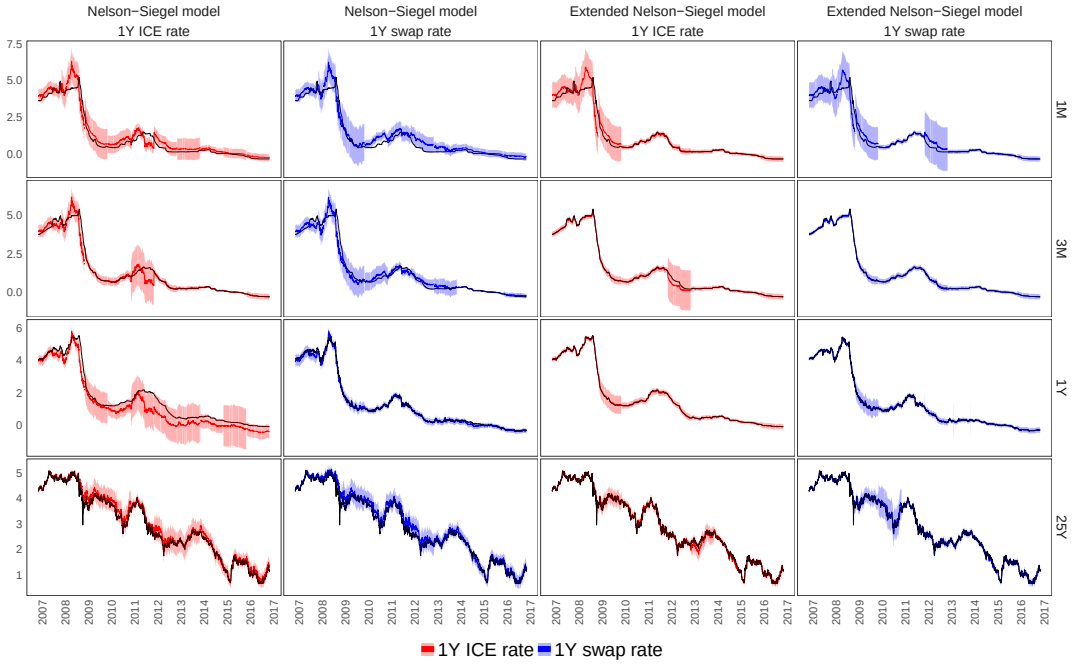


Figure 11.19. The Euro Libor rates at 1M, 3M, 1Y and 25Y maturities (black line) with the one step ahead out-of-sample prediction given by the conditional mean $\mathbf{y}_{T+s+1|T+s}$ with the corresponding predictive 97.5% confidence intervals over considered period. The column-wise order of panels corresponds to different classes of models, the Nelson-Siegel or Extended Nelson-Siegel model, calibration on two data sets with different 1Y rate (colours of lines and shading). The row-wise order corresponds to the evolution of rates at different maturities over time.

12. Stochastic Period and Cohort Effect State-Space Mortality Models Incorporating Demographic Factors

In this study we develop a multi-factor extension of the family of Lee-Carter stochastic mortality models. We build upon the time, period and cohort stochastic model structure to extend it to include exogenous observable demographic features that can be used as additional factors to improve model fit and forecasting accuracy. We develop a dimension reduction feature extraction framework which (a) employs projection based techniques of dimensionality reduction; in doing this we also develop (b) a robust feature extraction framework that is amenable to different structures of demographic data; (c) we analyse demographic data sets from the patterns of missingness and the impact of such missingness on the feature extraction, and (d) introduce a class of multi-factor stochastic mortality models incorporating time, period, cohort and demographic features, which we develop within a Bayesian state-space estimation framework; finally (e) we develop an efficient combined Markov chain and filtering framework for sampling the posterior and forecasting. We undertake a detailed case study on the Human Mortality Database demographic data from European countries and we use the extracted features to better explain the term structure of mortality in the UK over time for male and female populations when compared to a pure Lee-Carter stochastic mortality model, demonstrating our feature extraction framework and consequent multi-factor mortality model improves both in sample fit and importantly out-off sample mortality forecasts by a non-trivial gain in performance.

12.1. Motivations and Related Works

Modelling the "term-structure" of age specific mortality rates by gender and country has enjoyed a growing resurgence in the actuarial and statistics literature. This is primarily driven by the importance of better understanding and forecasting age specific mortality rates for purposes of understanding longevity risk, pension design, annuities pricing and population studies.

The most widely utilised class of stochastic mortality models in actuarial science and statistics arise from the class of regression or state-space models that incorporate explanatory factors which correspond to stylised latent stochastic factors representing structural features in the evolution of the age specific mortality rates. Typically these latent stochastic features are interpreted as either temporal effects, period effects and cohort effects. The most famous class of such models is the Lee-Carter type models, see a summary recently in (Fung et al. 2018) and references therein.

In this paper we aim to combine these classes of stochastic mortality model with other observable exogenous features obtained from a range of demographic data sets. The purpose being that they offer two advantages to standard Lee-Carter models, firstly they may improve predictive

power of the models, secondly they may improve the interpretation of behaviour of the dynamic of the “term-structure" of age specific mortality rates.

We expect the mortality experience and demographic data to be characterised by a strong causal and time-varying interaction. There is an existing literature on incorporation of demographic data in stochastic mortality models. However, unlike the state space mortality age-term structure dynamic factor model approach we develop in this manuscript, the existing works have been primarily focused on regression type structures that consider single age group models. Furthermore, there is limited work on feature extraction methods in this space. We highlight a few related approaches that have considered demographic data to study single age group mortality. We comment on some of the widely used exogenous factors in such studies, which include for instance macroeconomic variables, as well as demographic variables. In (Hanewald 2011) and (Niu and Melenberg 2014), the authors investigate the links between the economic growth and morality trends through a class of single age group regression models which are estimated in a frequentist estimation framework. In addition to the period effect in the standard Lee-Carter mode, the authors incorporate gross domestic product (GDP) as an observable factor what improves the in-sample and out-of-sample performance of the model.

Other classes of factors that have been explored in such settings also include cause-of-death categorical variables, what has been also partly investigated in (Hanewald 2011). The relation between causes of death and their influence on mortality has started to be more detailed explore since the accessibility of the data improved. In (Murray and Lopez 1997), the authors develop the scenarios of future mortality based on a multi factor linear regression model where the logarithm of the rate of mortality per age group, sex and clustered cause of death is regressed against the socio-economic, educational, technological and cause-of-death related predictors. The Bayesian inference has been adopted in (Girosi and King 2008) to build a regression framework for forecasting mortality rates which are age, sex, country and case of death specific. The work is mostly focused on the methodological side of the forecasting but uses as examples the applications of demography and macro-epidemiology data as explanatory variables for the regression-type model of mortality. Moreover, the dependency structures between cause-specific death rates are studied in (Arnold (-Gaille) and Sherris 2015). The authors use Vector Error Correction Models to examine such causal relations within the countries.

The usage of principal components of the mortality curves as linear regressors has been examined in (Hyndman and Yasmeen 2012). The authors explored the common features of the data applying the functional version of Principal Component Analysis. The concept is further developed in (Erbas et al. 2010), where the cause-of-death-specific smoothed mortality curves are treated as functional data. The obtained principal components serve as basis functions in functional data analysis.

In the following study, we aim to broaden these concepts and investigate the impact of global mortality trends given by various sets of international demographic data, and their potential influence on the mortality experience in one country, in our case study the United Kingdom. To

achieve this in a manner suitable for incorporation in multi-age stochastic mortality models we need to perform a parsimonious feature extraction method in order to reduce the large dimensional sets of data to a form suitable for inclusion in such a mortality model. Therefore, we introduce a methodology which is not exclusive to one type of demographic data and is capable of handling the analysis jointly over many different exogenous variables.

To achieve this we must undertake several tasks: the first is to explain a canonical and principled approach to combining of such demographic time series data into the stochastic mortality models, for which there is a number of structural approaches we develop and present.

The second aspect is that large demographic data sets are now available, but a naive incorporation of such features into a stochastic mortality model would result in far too many parameters to perform estimation, the models would be overfit and would not provide good generalisation properties for out-off-sample forecast performance. Therefore, we introduce a class of probabilistic, statistically robust feature extraction approaches to reduce dimensionality and capture core information present in the demographic data that can be more parsimoniously included in the stochastic mortality models. The standard concept of robust Principal Component Analysis by means of M- and S- estimators cannot be easily utilized since the demographic data is not equal length and contains missing values across different age groups. Hence, we adopt a standard probabilistic formulation of Principal Component Analysis introduced in Part I in Section 2.4 and which additionally allows to model the hidden process of missing values in Section 4.2.

Another challenge is the issue of parameters uncertainty in mortality modelling which we address by adopting the Bayesian Inference framework based on efficient Markov Chain Monte Carlo as in (Fung et al. 2018). The estimation of the model is achieved via a Rao-Blackwellised Gibbs sampler. We sample the static parameters via conjugate Gibbs sampling steps which are followed by Forward Backward filtering sampler for state variables to inference from the resulting posteriors.

The contributions of each part of the study are as follows. Firstly, we briefly overview the concept of the mortality modelling and discuss the state-space formulation of the Lee-Carter model with cohort effects and impose identification contains. Section 12.3 provides with several illustrations of how to incorporate observable factors into Lee- Carter models. The discussion an overview of the data in Section 12.4 whereas numerical illustrations of empirical studies are presented in Section 12.5.

12.2. Period and Cohort Effect Stochastic Mortality Models: State-Space Formulation

We begin this section by briefly recalling the classical two factor period effect and cohort effect models that have been proposed in (Renshaw and Haberman 2006); (Pedroza 2006) and (Kogure and Kurachi 2010). This includes in particular the state-space formulation of such models which was developed in (Fung et al. 2017) and (Fung et al. 2018).

Extension of Lee-Carter model to cohort features proposed in (Renshaw and Haberman 2006) introduces the concept of the stochastic cohort factor, denoted by γ_{t-x}, is incorporated into the one

factor stochastic Lee-Carter (Lee and Carter 1992) stochastic period effect, denoted κ_t to produce a two factor stochastic model. This second cohort factor, like the period effect factor, can also have an age-modulating coefficient, denoted by β_x^γ. In this work we adopt the recommendations discussed in (Cairns et al. 2009); (Haberman and Renshaw 2011) and (Hunt and Villegas 2015), where it is proposed to simplify this feature to be a constant age-modulating coefficient accross all age groups, given by:

$$\log m_{x,t} = \alpha_x + \beta_x \kappa_t + \gamma_{t-x}, \tag{12.1}$$

where $x \in \{x_1, \ldots, x_p\}$ and $t \in \{1, \ldots, T\}$ represent age and year respectively. $m_{x,t}$ denotes the mortality rate in age group x and time t. α_x and β_x are the age specific static parameters of the model. This simplifying assumption that $\beta_x^\gamma = 1$ (or generally any constant other than one) is supposed to improve estimation performance when fitting the model in practice. Furthermore, there is discussion in the literature to argue that it may also be justified based on empirical findings. By studying the mortality experience of England and Wales males, the study of (Willets 2004) finds that the cohort effect is not "wearing off" with increasing ages and the mortality improvement rates, defined as $1 - m_{x,t}/m_{x,t-1}$, of different cohorts seem to be rather stable. Together with the consideration of the convergence problem, one may argue that it is indeed reasonable to assume that $\beta_x^\gamma = 1$ to ensure estimation can be successfully performed for a range of mortality data while the explanatory power of the simplified model is comparable to the full model.

Next we recall the two-factor state-space formulation of the Lee-Carter type period-cohort models for stochastic mortality, see derivations and properties in (Fung et al. 2017) and (Fung et al. 2018). Note, we adopt the same standard notation as proposed in these papers to present the models in this manuscript.

The formulation of stochastic period-cohort models in state-space form is given by specification of both an observation equation and a state equation. Let $y_{x,t} = \log \hat{m}_{x,t}$ where $x = x_1, \ldots, x_p$ and $t = 1, \ldots, T$. The general form of the observation equation (when β_x^γ is flexible) of the cohort model Equation (12.1) is given in matrix form by (recall that $\gamma_t^x := \gamma_{t-x}$):

$$\begin{pmatrix} y_{x_1,t} \\ y_{x_2,t} \\ \vdots \\ y_{x_p,t} \end{pmatrix} = \begin{pmatrix} \alpha_{x_1} \\ \alpha_{x_2} \\ \vdots \\ \alpha_{x_p} \end{pmatrix} + \begin{pmatrix} \beta_{x_1} & \beta_{x_1}^\gamma & 0 & \cdots & 0 \\ \beta_{x_2} & 0 & \beta_{x_2}^\gamma & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \beta_{x_p} & 0 & 0 & \cdots & \beta_{x_p}^\gamma \end{pmatrix} \begin{pmatrix} \kappa_t \\ \gamma_t^{x_1} \\ \gamma_t^{x_2} \\ \vdots \\ \gamma_t^{x_p} \end{pmatrix} + \begin{pmatrix} \varepsilon_{x_1,t} \\ \varepsilon_{x_2,t} \\ \vdots \\ \varepsilon_{x_p,t} \end{pmatrix}, \tag{12.2}$$

where iid noise terms $\varepsilon_{x,t}$ are included as we aim to model the crude death rates. In a given year t, we identify in this model the state vector as $(\kappa_t, \gamma_t^{x_1}, \ldots, \gamma_t^{x_p})^\top$ which represents the $p+1$ dimensional latent unobserved stochastic factor driving the observed log-mortality rates. The

dynamic of this stochastic state vector is then specified in the state equation. In this work, we consider the state-space model given in matrix form as follows:

$$\begin{pmatrix} \kappa_t \\ \gamma_t^{x_1} \\ \gamma_t^{x_2} \\ \vdots \\ \gamma_t^{x_{p-1}} \\ \gamma_t^{x_p} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \lambda & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} \kappa_{t-1} \\ \gamma_{t-1}^{x_1} \\ \gamma_{t-1}^{x_2} \\ \vdots \\ \gamma_{t-1}^{x_{p-1}} \\ \gamma_{t-1}^{x_p} \end{pmatrix} + \begin{pmatrix} \theta \\ \eta \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} \omega_t^{\kappa} \\ \omega_t^{\gamma} \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}. \tag{12.3}$$

In this particular instance, we assume κ_t is a random walk with drift process (ARIMA(0,1,0) with a constant) and the dynamics of $\gamma_t^{x_1}$ are described by a stationary AR(1) process with drift (ARIMA(1,0,0) with a constant) where $|\lambda| < 1$. One may consider other dynamics for $\gamma_t^{x_1}$ by specifying the second row of the $p+1$ by $p+1$ matrix in Equation (12.3). For example, an ARIMA(2,0,0) process for $\gamma_t^{x_1}$ can be assumed if one fixes

$$\begin{pmatrix} \kappa_t \\ \gamma_t^{x_1} \\ \gamma_t^{x_2} \\ \vdots \\ \gamma_t^{x_{p-1}} \\ \gamma_t^{x_p} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \lambda_1 & \lambda_2 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} \kappa_{t-1} \\ \gamma_{t-1}^{x_1} \\ \gamma_{t-1}^{x_2} \\ \vdots \\ \gamma_{t-1}^{x_{p-1}} \\ \gamma_{t-1}^{x_p} \end{pmatrix} + \begin{pmatrix} \theta \\ \eta \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} \omega_t^{\kappa} \\ \omega_t^{\gamma} \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \tag{12.4}$$

The matrix form of Equations (12.2) and (12.3) can be expressed succinctly as

$$\mathbf{y}_t = \boldsymbol{\alpha} + \mathbf{B}\boldsymbol{\varphi}_t + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \overset{iid}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2 \mathbb{I}_p), \tag{12.5a}$$

$$\boldsymbol{\varphi}_t = \boldsymbol{\Lambda}\boldsymbol{\varphi}_{t-1} + \boldsymbol{\Theta} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \overset{iid}{\sim} \mathcal{N}(0, \Upsilon), \tag{12.5b}$$

where $\boldsymbol{\varphi}_t = (\kappa_t, \gamma_t^{x_1}, \ldots, \gamma_t^{x_p})^\top$, $\mathbb{I}_p$ the p-dimensional identity matrix and Υ is a $p+1$ by $p+1$ diagonal matrix with diagonal $(\sigma_\kappa^2, \sigma_\gamma^2, 0, \ldots, 0)$. The matrices $\boldsymbol{\alpha}$, $\mathbf{B}$, $\boldsymbol{\Lambda}$ and $\boldsymbol{\Theta}$ can be easily identified in Equations (12.2) and (12.3). For simplicity we assume homoscedasticity in the observation equation; heteroscedasticity can be incorporated straightforwardly as developed in (Fung et al. 2017).

We adopt the identification constraints which are based on (Hunt and Villegas 2015) and are broadly discussed and examined in (Fung et al. 2018). These are given by

$$\sum_x \beta_x = 1, \ \sum_x \beta_x^{\gamma} = 1, \ \sum_t \kappa_t = 0, \ \sum_{c=t_1-x_p}^{t_N-x_1} \gamma_c = 0. \tag{12.6}$$

12.3. Demographic Factor Model Extension to the Period-Cohort Stochastic Mortality State-Space Models

In this section we demonstrate several approaches one may adopt to extend the state-space model formulations presented previously for the Period-Cohort stochastic mortality models to allow for incorporation of additional observable covariate factors. The form of factor model we develop here is generic and in future sections we will develop a framework for factor extraction from demographic data that can be used in the models developed in this section.

We note that there are two fundamental ways to develop a factor time series based regression structure for incorporation of demographic data to a stochastic mortality model. We advocate in this paper an approach which is specifically developed to work with data which may be high dimensional in nature, structured but be represented by short time series lengths. This type of data is particularly prevalent in demographic studies. The main concept here is that the feature extraction is performed over the entire available time series of observable demographic data. The resultant features extracted are then added to the stochastic mortality model in a static form but with dynamic latent state processes for the factor loadings over time. That is the effect of the factor incorporated will be allowed to time vary through the factor loading. This approach has the advantage of not having to model explicitly the demographic data which may have a complex structure and furthermore, only requires forecasting components of the latent factor loading process. This is often significantly easier to perform since one may use for instance a standard parametric time series model such as a VAR model for their temporal evolution. Note, such approaches as this are also utilised in other financial term-structure state-space models for instance in the context of yield curve modelling, such as the dynamic Nelson-Siegel model of (Diebold et al. 2006) as discussed in Chapter 11. However, we believe to the best of our knowledge we are the first to propose this type of factor model framework for incorporating demographic factors into stochastic mortality models.

We assume that we have an available set of factors that can be country, population, gender or age specific features. We wish to incorporate these age-specific and country-specific demographic/population information into the cohort state-space model described by Equations (12.5a) and (12.5b).

We will denote by $\mathbf{F}_t$ the $p \times k$ factors matrix where p may represent number of age groups and k may represent number of age specific factors. As in any feature based regression factor analysis such as PCA Regression ((Jolliffe 2002)), we treat $\mathbf{F}_t$ as extracted via feature extraction methods from an exogenous observable input that is believed to have potential influence on the age specific mortality rates under study in the responses, over time. Then for each feature vector regressor, extracted from the exogenous demographic data, we will add this feature to the state-space model.

There are numerous structural ways to achieve this in a state-space model as noted in Chapter 5. For instance, the factor may either influence all age groups equally by entering the factor into the state equation, or it may influence each age specific mortality rate differently by adding it in the

observation equation. Of course, there may also be a combination of such approaches, dependent on which demographic data the feature was extracted from the context of the model construction.

The influence of the feature on the log mortality is reasonable to assume it varies over time, so to achieve this we will specify a time dynamic for the regression factor loading. This requires that we specify an additional latent variable, a pk dimensional vector $\boldsymbol{\varrho}_t$ which denotes the vector of factor loading for year t. We assume $\boldsymbol{\varrho}_t$ to be modelled by VAR(1) process given by

$$\boldsymbol{\varrho}_t = \boldsymbol{\Omega}\boldsymbol{\varrho}_{t-1} + \boldsymbol{\Psi} + \boldsymbol{\omega}_t^{\varrho}, \quad \boldsymbol{\omega}_t^{\varrho} \overset{iid}{\sim} \mathcal{N}(0, \sigma_{\varrho}^2 \mathbb{I}_{pk}) \tag{12.7}$$

with homogeneous variant for covariance matrix of error term $\boldsymbol{\omega}_t^{\varrho}$. $\boldsymbol{\varrho}_t$ is a dynamic regression parameter for the factors matrix $\mathbf{F}_t$ which specifies the impact of $x_i \in \{x_1, \ldots, x_p\}$ age group and $m \in \{1, \ldots, k\}$ component corresponding to $[\mathbf{F}_t]_{i,m}$ by $\varrho_t^{i,m}$ element.

As noted, depending on the interpretability of the desired model, one may incorporate $\mathbf{F}_t$ into observation Equation (12.5a) (**Case 1**) or into the latent dynamic of either calendar year factor period effect κ_t (**Case 2**) or cohort factors vector $\boldsymbol{\gamma}_t$ (**Case 3**) from the state Equation (12.5b).

Next we develop the extended model of Equations (12.5a) and (12.5b), which incorporates information $\mathbf{F}_t$. The general notation of the model is as follows

$$\boldsymbol{y}_t = \boldsymbol{\alpha} + \tilde{\mathbf{B}}_t\tilde{\boldsymbol{\varphi}}_t + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \overset{iid}{\sim} \mathcal{N}(0, \sigma_{\varepsilon}^2 \mathbb{I}_p), \tag{12.8a}$$

$$\tilde{\boldsymbol{\varphi}}_t = \tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\varphi}}_{t-1} + \tilde{\boldsymbol{\Theta}} + \tilde{\boldsymbol{\omega}}_t, \quad \tilde{\boldsymbol{\kappa}}_t \overset{iid}{\sim} \mathcal{N}(0, \tilde{\Upsilon}) \tag{12.8b}$$

where $\tilde{\boldsymbol{\varphi}}_t = (\boldsymbol{\varphi}_t, \boldsymbol{\varrho}_t)$ is a $(p + pk + 1) \times 1$ latent process vector and

$$\tilde{\boldsymbol{\Theta}} = \begin{pmatrix} \boldsymbol{\Theta}_{(p+1)\times 1} \\ \boldsymbol{\Psi}_{pk\times 1} \end{pmatrix}_{(p+pk+1)\times 1} \tag{12.9}$$

is a vector of drift parameters for state equations, where $\boldsymbol{\Psi}$ corresponds to the model Equation (12.7) of $\boldsymbol{\varrho}_t$. We assume independence of error terms in latent variables what gives the following structure of a covariance matrix for the state equation error term $\tilde{\boldsymbol{\omega}}_t$

$$\tilde{\Upsilon} = \left(\begin{array}{c|c} \Upsilon_{(p+1)\times(p+1)} & 0 \\ \hline 0 & \sigma_{\varrho}^2 \mathbb{I}_{pk} \end{array}\right)_{(p+pk+1)\times(p+pk+1)} \tag{12.10}$$

Let us specify the following two objects, $\tilde{\mathbf{F}}_t = \bigoplus_{j=1}^{k} [\mathbf{F}_t]_{j,\cdot}$, for $\bigoplus$ being a direct sum operator, and $\tilde{\mathbf{f}}_t = vec\left(\mathbf{F}_t^T\right)$, that is

$$\tilde{\mathbf{F}}_t = \begin{pmatrix} [\mathbf{F}_t]_{1,\cdot} & 0 & 0 & \cdots & 0 \\ 0 & [\mathbf{F}_t]_{2,\cdot} & 0 & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & \cdots & & & [\mathbf{F}_t]_{p,\cdot} \end{pmatrix}_{p\times pk} \quad \text{and} \quad \tilde{\mathbf{f}}_t = \begin{pmatrix} [\mathbf{F}_t]_{1,1} \\ [\mathbf{F}_t]_{1,2} \\ \vdots \\ [\mathbf{F}_t]_{p,k} \end{pmatrix}_{pk\times 1} \tag{12.11}$$

where $[\mathbf{F}_t]_{j,\cdot}$ and $[\mathbf{F}_t]_{j,m}$ represent the vector of the jth row of the matrix $\mathbf{F}_t$ and the element corresponding to jth row and mth column, respectively. The structures of the other matrices and vectors for extended model Equations (12.8a) and (12.8b) depend on the introduced cases, that is

$$\tilde{\mathbf{B}}_{t\ p\times(p+pk+1)} = \begin{cases} \left(\mathbf{B}_{p\times(p+1)} \,\middle|\, \tilde{\mathbf{F}}_t \right) & \text{for } \textbf{Case 1}, \\ \left(\mathbf{B}_{p\times(p+1)} \,\middle|\, \mathbf{0}_{p\times pk} \right) & \text{otherwise}, \end{cases}$$

$$\tilde{\mathbf{\Lambda}}_{(p+pk+1)\times(p+pk+1)} = \begin{cases} \left(\begin{array}{c|c} \mathbf{\Lambda}_{(p+1)\times(p+1)} & \mathbf{0}_{(p+1)\times pk} \\ \hline \mathbf{0}_{pk\times(p+1)} & \mathbf{\Omega}_{pk\times pk} \end{array} \right) & \text{for } \textbf{Case 1}, \\ \left(\begin{array}{c|c} \mathbf{\Lambda}_{(p+1)\times(p+1)} & \tilde{\mathbf{f}}_t^T \\ & \mathbf{0}_{p\times pk} \\ \hline \mathbf{0}_{pk\times(p+1)} & \mathbf{\Omega}_{pk\times pk} \end{array} \right) & \text{for } \textbf{Case 2}, \\ \left(\begin{array}{c|c} \mathbf{\Lambda}_{(p+1)\times(p+1)} & \mathbf{0}_{1\times pk} \\ & \tilde{\mathbf{F}}_t \\ \hline \mathbf{0}_{pk\times(p+1)} & \mathbf{\Omega}_{pk\times pk} \end{array} \right) & \text{for } \textbf{Case 3}. \end{cases} \tag{12.12}$$

Given the formulated demographic factor model of the period-cohort stochastic mortality state-space models, we will denote this class of model by notation (DFM-PC). The estimation of this model will be achieved through a Bayesian model formulation and a specialised Markov chain Monte Carlo sampling framework based on Forward-Backward sampler for the latent state components and block-Gibbs conjugate sampling for the static model parameters. This follows closely the detailed framework developed extensively in (Fung et al. 2017) and (Fung et al. 2018), therefore, we repeat this only in relevant details in the Appendix 12.6.

Remark 9. *The matrix $\tilde{\mathbf{F}}_t$ contains the observed (feature extracted) with exogenous factors and is a component of the model which is conditioned. Therefore it does not require to be estimated and, at this point, represents a known deterministic constant. As such $\tilde{\mathbf{F}}_t$ is fixed and held constant for the time period which the factor model is run. Hence, it is not a parameter but a covariate which is observed and deterministic. Therefore, the identifications constraints given by Equation (12.6) are valid for the new model.*

Remark 10. *Time series in the study undertaken are typically of the length less than 100. Compared to the number of available samples, the classical Lee-Carter model with cohort effect requires large number of parameters and has been already reported to be prone to overfitting as noted in (Cairns et al. 2009); (Haberman and Renshaw 2011) and (Hunt and Villegas 2015). The new model with the latent process introduced in Equation (12.7) provides with additional number of parameters. Therefore to decrease the risk of overfitting we choose to stay with the assumption of VAR(1) process for $\boldsymbol{\varrho}_t$. We believe that the parsimony argument to keep the autoregressive structure without too many additional parameters was more important that adding more lags in this study.*

12.4. Data

The examined data consists of male and female mortality and demographic data obtained from Human Mortality Database (http://www.mortality.org) for European countries. The Table 12.1 summarizes the availability of the data for all the countries included into the study.

We use four different sets of mortality data, raw data: Birth counts and Death counts, and life tables: Life Expectancy at Birth and Death Rates. We conduct separate analysis for female and male populations.

The time series vary in terms of the number of available observations. The longest time series is provided by Swedish and French mortality data, starting from 1751 and 1816, respectively. The shortest time series are given for Greece and Slovenia, 1983–2014 and 1981–2013, respectively.

With regards to Birth counts and Life expectancy at Birth, the information per country in time point is one dimensional, i.e., annual counts of live births by sex in year t (Birth counts) or the expected life span of a person born in year t (Life Expectancy at Birth). Hence, a single observation in these cases consists of the number of entries equal to the number of countries included into the study, that is 31 listed in Table 12.1, per gender.

Table 12.1. The availability of the demographic data per country (Human Mortality Database).

Country	Life Expectancy ($E0$)	No. Births	Death Rate (m_x)	No. Deaths
Austria	1947–2014	1871–2014	1947–2014	1947–2014
Belarus	1959–2014	1959–2014	1959–2014	1959–2014
Belgium	1841–2015	1840–2015	1841–2015	1841–2015
Czech Republic	1950–2010	1947–2014	1950–2014	1950–2014
Denmark	1835–2014	1835–2014	1835–2014	1835–2014
Estonia	1959–2013	1959–2013	1959–2013	1959–2013
Finland	1878–2012	1865–2012	1878–2012	1878–2012
France	1816–2014	1806–2014	1816–2014	1816–2014
East Germany	1956–2013	1946–2013	1956–2013	1956–2013
West Germany	1956–2013	1946–2013	1956–2013	1956–2013
Greece	1981–2013	1981–2013	1981–2013	1981–2013
Estonia	1959–2013	1959–2013	1959–2013	1959–2013
Hungary	1950–2014	1950–2014	1950–2014	1950–2014
Iceland	1838–2013	1838–2013	1838–2013	1838–2013
Ireland	1950–2014	1950–2014	1950–2014	1950–2014
Italy	1872–2012	1862–2012	1872–2012	1872–2012
Latvia	1959–2013	1959–2013	1959–2013	1959–2013
Lithuania	1959–2013	1959–2013	1959–2013	1959–2013
Luxembourg	1960–2014	1950–2014	1960–2014	1960–2014
Netherlands	1850–2012	1850–2012	1850–2012	1850–2012
Norway	1846–2014	1846–2014	1846–2014	1846–2014
Poland	1958–2014	1958–2014	1958–2014	1958–2014
Portugal	1940–2012	1886–2012	1940–2012	1940–2012
Russia	1959–2014	1959–2014	1959–2014	1959–2014
Slovakia	1950–2014	1950–2014	1950–2014	1950–2014
Slovenia	1983–2014	1983–2014	1983–2014	1983–2014
Spain	1908–2014	1908–2014	1908–2014	1908–2014
Sweden	1751–2014	1747–2014	1751–2014	1751–2014
Switzerland	1876–2014	1871–2014	1876–2014	1876–2014
United Kingdom	1922–2013	1922–2013	1922–2013	1922–2013
Ukraine	1959–2013	1946–2013	1959–2013	1959–2013

The age specific information is provided for Death counts and Death Rates. A single observation per country in time t describes a number of deaths of people with ages from 0 to 110+ (Death counts) or number of deaths for ages from 0 to 110+ scaled to the size of that population, per unit of time (Death Rates). The availability of time series is different among age groups. Usually, the shortest time series are collected for age groups above 100 years.

Since the Lee-Carter model with the cohort effect has been already reported to be prone to over-fitting when fitted to short time series (the currently available mortality data is classified as a short time series), we decide to work with a data aggregated in the format "5×1", i.e., by

5-year age group per calendar year. The ages are grouped into following stratifications: 0, 1–4, 10–14, 15–19, 20–24, 25–29, 30–34, 35–39, 40–44, 45–49, 5–9, 50–54, 55–59, 60–64, 65–69, 70–74, 75–79, 80–84, 85–89, 90–94, 95–99, 100–104, 105–110, 110+ as additing additional latent processes increases the number of parameters to estimate.

A single observation in Deaths or Death Rates consist of the number of entries equal to the number of countries included into the study (i.e., 31) times the number of age groups, that is 24. It accounts for the information in time t available for all 31 countries among all 24 age groups.

12.4.1. Preprocessing of Data

Human Mortality Database (HMD) team applies several preprocessing procedures that aim to "clean" Death counts and population sizes before using them in order to calculate and distribute death rates and life tables. The subsequent steps are discussed in the technical report (Wilmoth et al. 2007). The adjustments are applied in order to distribute people of unknown age across age groups and splitting data into age categories, i.e., from age stratification "5×1" to "1×1" and from "1×1" to Lexis triangles. The common practice is to use a regression model for splitting deaths counts in format "1×1" to Lexis triangles and apply cubic splines to split "5×1" to "1×1". Additionally to the adjustments applied to Deaths counts, the age specific population size is estimated using four methods: linear interpolation, intercensal survival, extinct cohorts and survivor ratios.

The life tables are calculated using Lexis triangles and population sizes. Before death rates are converted into the probabilities of death, the rates at older ages (80 and above) are smoothed using logistic regression. The "abridged" life tables are calculated based on the Lexis triangles tables rather than the raw data. It ensures the both sets of tables to contain identical values of life expectancy and other quantities.

Recall that the smoothing applied to the mortality data can influence the feature extraction and the diminish the effect of robust versus non-robust versions of feature extraction methodology which we study in this paper. The topic is further discussed in Section 12.5.

Additionally to the briefly discussed procedures which have been already applied to the data in HMD, we needed to adjust the data to provide reliable information about missing values. We notice the ambiguity in labelling unavailable data which is either denoted by "NA" value or "0". Death Rates, Birth counts and Life Expectancy at Birth are unlikely to produce values equal to "0". Hence we replace all "0" which appeared in these data sets by "NA". The zeros which appear in Deaths counts in older age groups are more difficult to handle as there is no certainty whether there was no person in particular age group who died or the record has missing values. Due to this fact we decided to limit our analysis to age groups up to 90 and again replace any "0" by "NA".

12.4.2. Missing Data

A summary of different types of missingness across countries, age groups and sexes which occurs in the demographic data set analysed in our study can be found in Section C.1 in Appendix C. Our findings are the following: among the Birth related data where a total observation is of vector-type (one dimensional information per country), the incompleteness of data is due to a

general unavailability of the information per country. However, for Death related data where a total observation is of matrix-type (a age specific vector of information per country), we notice a patter of single missing values in time point which fits definition of MAR. In the subsequent section with the empirical analysis of the data set under the derived framework of Probabilistic Principal Component Analysis, we assumed MAR type of missingness. Extending this assumption requires calculating the integral from Equation (4.1) and incorporating it to PPCA framework.

12.5. Feature Extraction from European Demographic Data via Probabilistic Principal Component Analysis

The following section provides results for the feature extraction using Gaussian PPCA introduced in Section 2.4 and extended to the missing data setting in Section 4.2. The framework is applied to different types of European demographic data sets. The attention is drawn to the effect of simple and straightforward robustification overviewed in the previous sections which application we demonstrate on overviewed data sets. The main observation from this study is the difference in the consistency of the features over time and proportion of missingness for two frameworks, robust and non-robust. The features calculated using the means of robust estimators are more consistent over time and over different proportions of missingness that their non robust alternatives. It is especially visible for the features extracted from data that has not been previously preprocessed, e.g., Birth counts. The effect of robustification is smaller if a data set has been smoothed as Life Expectancy at Birth.

12.5.1. The Assessment of the Methodology

We conduct a comparison between robust and non-robust Probabilistic Principal Component Analysis (PPCA) in a stochastic setting. We undertake this exercise in order to incorporate the most meaningful eigenvectors as exogenous factors to the model in Equation (12.8). Each of the datasets discussed in the section is treated separately, that is, we compute the eigenvectors for Births, Life Expectancy at Birth, Deaths and Death Rates. Recall, an observation in time t, $\mathbf{y}_t$, conveys the information about a given data set in calendar year t from the 31 countries listed in Table 12.1. Hence, an observation of Births or Life Expectancy is 31 dimensional. Since Deaths and Death Rates carry the information which is age group specific, a single observation in these data sets is equal to the number of countries times the number of age groups.

The following section summarizes the results of PCA according to

- Population: Females and Males;
- Subsamples referring to maximum allowed proportion of missing values per an observation: 0% (no missing), 25%, 50% and 75% as discussed in Section 12.4.2 in Appendix C;
- Type of the standardisation procedure: robust and non-robust, which are used for the estimators of location and covariance in the PPCA framework.

We use M-estimators of the covariance and mean as a robust alternative to the sample estimators. As discussed in Section 3.1.1.2, the class of M-estimator has a very good performance

in small dimensions as its robustness is a function of the dimensionality and not the sample size. Since the data we use is not a long time series and we standardize every variable marginally (Algorithm 1), this simple estimator should provide us with reliable outcomes. In particular, we consider the Huber type M-estimators as discussed in Remark 3 which are characterized by normal asymptotic distribution with convergence rate of $\frac{1}{2}$ and several characteristics ensuring both uniqueness and optimality of the estimation Additionally, it is the estimator of covariance which has the minimal asymptotic variance among all estimators for Gaussian data. Thus, the choice is consistent with our assumption of normally distributed data for the treatment of missingness in the PPCA framework which we outlined in Section 4.2.

Among the objects which we analyse are eigenvectors, eigenvalues, scores and Mahalanobias distances which use estimated covariance matrix $\hat{\mathbf{C}} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbb{I}_d$ calculated by iteratively evaluated σ^2 and the projection matrix $\mathbf{W}$. The Mahalanobias distance is measured around vector $\mathbf{0}$ (the data has been centred), and is therefore given by

$$d_n^2 := d^2\left(\mathbf{y}_n, \hat{\mathbf{C}}\right) = \mathbf{y}_n\hat{\mathbf{C}}^{-1}\mathbf{y}_n^T. \tag{12.13}$$

The results show how distant from the assumed long term mean is a single observation.

The EM algorithm described in Section 4.2 provides the eigenvectors corresponding to the k largest eigenvalues. Due to the specifics of the data (small number of observations), it would be difficult to incorporate many factors into the model from Equation (12.8) and achieve reliable estimation results. This is primarily a result of the curse of dimensionality in the parameter space that would lead in this case to diffusivity in the resulting likelihood utilised in the estimation. The latent states and static parameters will become difficult to filter and estimate. Thus, we limit our analysis to $k = 3$ main eigenvectors which explain most of the variability as shows the standard PCA with non missing data conducted separately for each country showed.

12.5.2. Births

The results of PPCA for Births counts among Females and Males are similar. It is the outcome which follows general intuition as there is no external factor which influences differently the births of woman and man in European countries. Also, recall that the Birth counts are the least pre-processed data set in out analysis.

Figure 12.1 shows the Mahalanobias distances (x axis) of Number of Births over the time (y axis) for female (a) and male (b) population. Each sub-panel consists of two plots, which present results for data being standardized by robust (lower plots) and non-robust (upper plots) estimators of the mean and covariance matrix. Different colours of lines depict distances for subsamples where maximum allowed proportion of missigness per an observation is 0%, 25%, 50% and 75%. Recall, that the subsamples starts in different times and therefore the corresponding results are the outcome of the estimation on different data with different impacts of distribution assumptions.

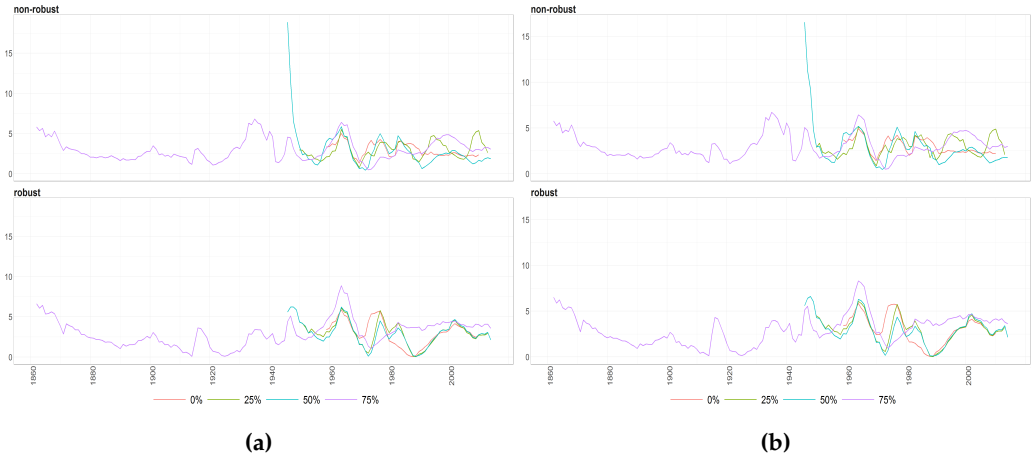


Figure 12.1. The Mahalanobias distances obtained using Probabilistic Principal Component Analysis (PPCA) for Females (**a**) and Males (**b**) Births over time (x axis). Different colours of lines correspond to the cases of different percentages of maximal missing values in a signle observation (light blue (75%), dark brown (50%), dark blue (25%), light brown (0%)). Every subfigure is divided into two subplots corresponding to robust estimation of standard divinations (upper plot) and sample one (bottom plot).

As expected, the effect of robust standardisation is substantial since the framework produces features by down-weights outliers in data. The distances which correspond to the robust standardisation, are more aligned historically, that is, are more 'robust' when the missingness increases. It indicates that robust standardisation of data captures more efficiently the characteristics of the population distribution. Recall the earliest non-robust Mahalanobias distance of 50% case which is very distant from the statistic in the same calendar year but for the 75% missingness case. For the robust case, the corresponding distances are more aligned what demonstrates the effect of robustness.

Since the subsample of the 75% missingness case is substantially longer, we expect the PPCA results to be different as the sample captures more regimes present in demographic data. Also, the sample has higher number of missing values which are estimated using the projection based on assumption of normal distribution. It also impacts the outcomes of PPCA.

The eigenvalues of estimated covariance matrices are shown in Figure 12.2. Different colours of lines highlight eigenvalues which correspond to different cases of missingness. Upper panels show the results for non-robust framework whereas bottom plots for the robust one. The magnitude of eigenvalues as well as the spreads between them over different levels of accepted missingness are higher for robust case.

The 75% case of missingness results in smaller discrepancies between eigenvalues for robust and non-robust frameworks. The corresponding eigenvectors exhibit similar behaviour what is shown in Figure 12.3. The discrepancies between the robust and non-robust eigenvectors are more significant for the cases with smaller proportions of missing values per observation. This outcome can be justified by the fact that the case of 75% is more affected by the priori assumptions of the normal distributions. The discrepancies between two methods of standardisation got smaller as

the projection of missing values starts to dominate the estimation of the principal components. The robust and non-robust estimators similarly capture the information about the normal distribution.

The blue dotted vertical lines on the plots of eigenvectors disaggregate the outcomes into developed and developing countries listed in Table 12.1. Regardless of the case of missingness (except 75% case) and type of the standardisation, we notice resembling features for countries from each of the groups.

We would expect the alignment of the robust scores as it has been observed for the corresponding Mahalanobias distances. However, the described PPCA methodology does not re-estimate the mean. The data is centred one during the initialisation. In the presence of missing values, their projection changes the mean and the data is centred only at the start of the procedure. This, the mean is only static when there is no missing data. This simplification of the framework results in different levels of the scores in Figure 12.4.

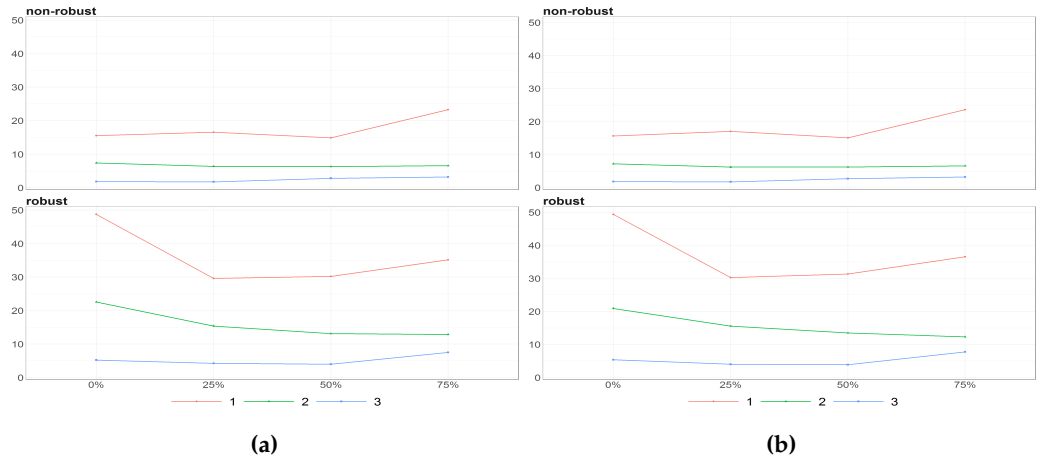


Figure 12.2. The eigenvalues obtained using PPCA for Females (**a**) and Males (**b**) population of Number of Births for different percentages of maximal missing entries in rows (x axis). Colours of lines corresponds to different eigenvalues, first (light brown), second (dark blue) and third (dark brown) highest. Every subfigure is divided into two subplots corresponding to robust estimation of standard divinations (upper plot) and sample one (bottom plot).

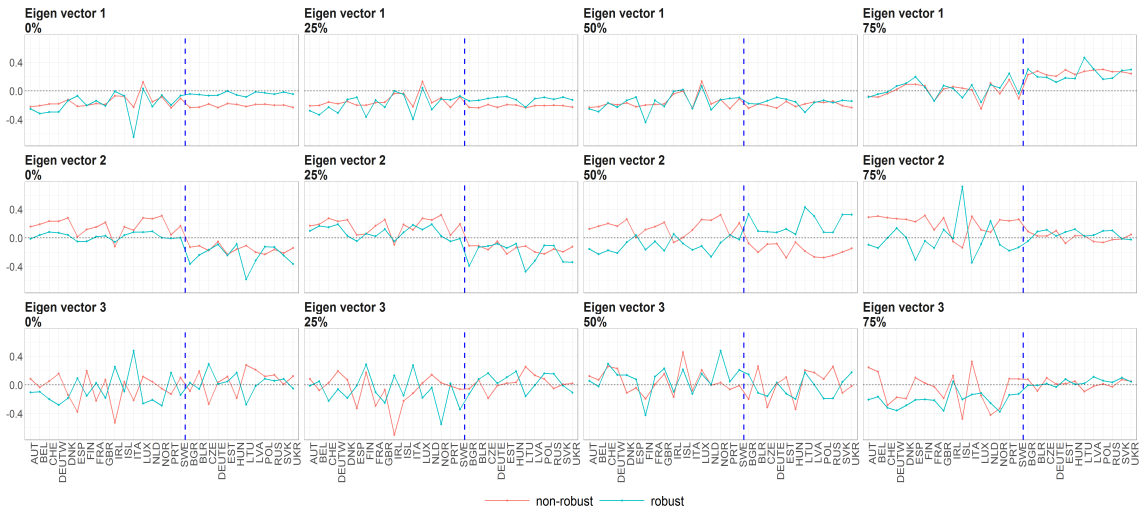


Figure 12.3. The eigenvectors (y axis) over the joint distribution of countries (x axis) obtained using PPCA for Females in Births. Every row of subfigure corresponds to a different eigenvector. Every column corresponds to different case of missing values (0%, 25%, 50% and 75%). The blue line corresponds to robust standardisation whereas red line to non-robust standardisation of data.

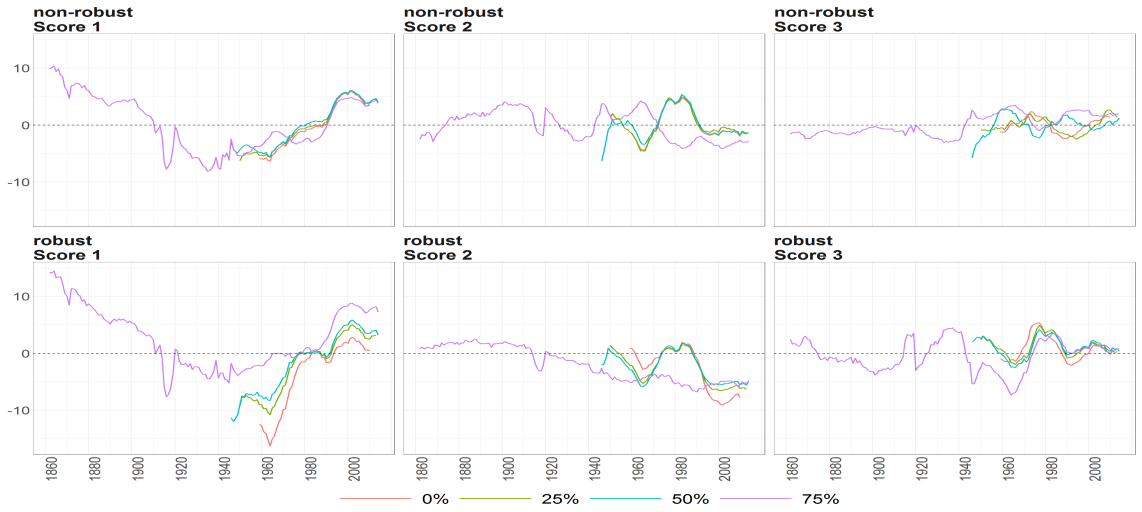


Figure 12.4. The scores (y axis) over time (x axis) obtained using PPCA for Females in Births. Colours of lines correspond to the scores calculated on subsample different cases of missing values (0%, 25%, 50% and 75%, refer to legend). The plots placed in the first row correspond to the results using non- robust standardization of entry data, where the second row correspond to robust standardisation. The plots scores, first, second and thirds are ordered by columns.

12.5.3. Life Expectancy at Birth

As mentioned in Section 12.4.1, Birth counts are the only data set in our analysis, which has been not modified or preprocessed before being available in HMD. The stages of several adjustments which are applied to the Death counts and population sizes result in a smaller number of outliers in Life Expectancy at Birth. The outcomes of PPCA for the data which is standardized using robust and non-robust estimators of mean and covariance matrix do not vary so significantly as in Birth counts over different proportions of missingness.

The robust Mahalanobias distances in Figure 12.5 are slightly more distant than their non-robust equivalents. However, the distances for different cases of missing values are similarly aligned in both cases in contrast to Birth counts. We may observe the same pattern among the scores of three principal components in Figures 12.9 and 12.10.

Figures 12.6, 12.7 and 12.8 show eigenvalues and eigenvectors for Females and Males respectively. Only the results for the 75% case of missingness exhibit more variation among the standardisation procedures. Since the subsample corresponding to 75% case is significantly longer, the discrepancies can be again rationalized by the two reasons: effect of the prior assumption on distributions and higher number of captured regimes in the data. Moreover, analysing the outcomes from the aggregation among the European countries, we notice that the second eigenvector has opposite signs for the countries of the two countries groups. Recall that its values are more volatile for the male population of developing countries (except for the case of 75% of missing values).

The obtained scores for Life Expectancy at Birth are shown in Figures 12.9 and 12.10. The levels of scores are very close to zero. It is an expected outcome as we differenced the data since it exhibits polynomial trend. The distribution of differences is expected to have zero mean. Hence, the scores are aligned for all the cases of missingness and distributed around zero what is negligibly affected by the simplification of the discussed framework. However, the two methodologies of the standardisation result in different magnitudes of scores. Since the corresponding eigenvectors are very similar for two standardisations, the magnitude is a reliable indicator whether an observation is outlying.

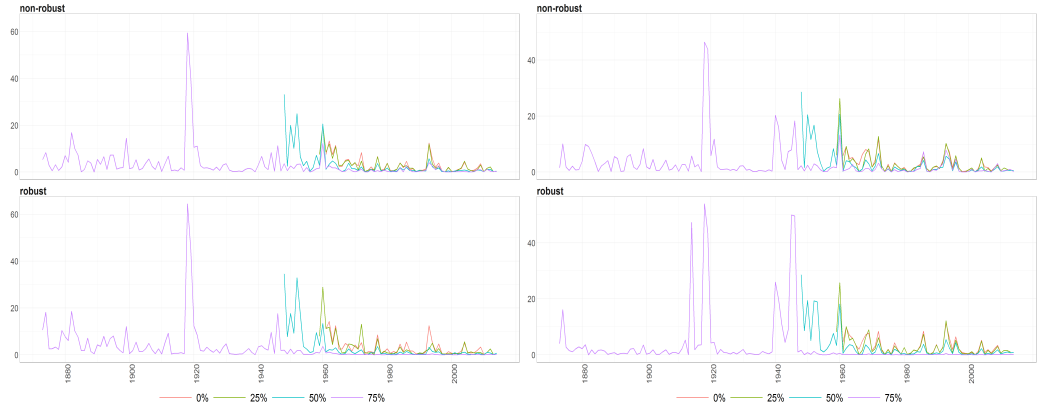
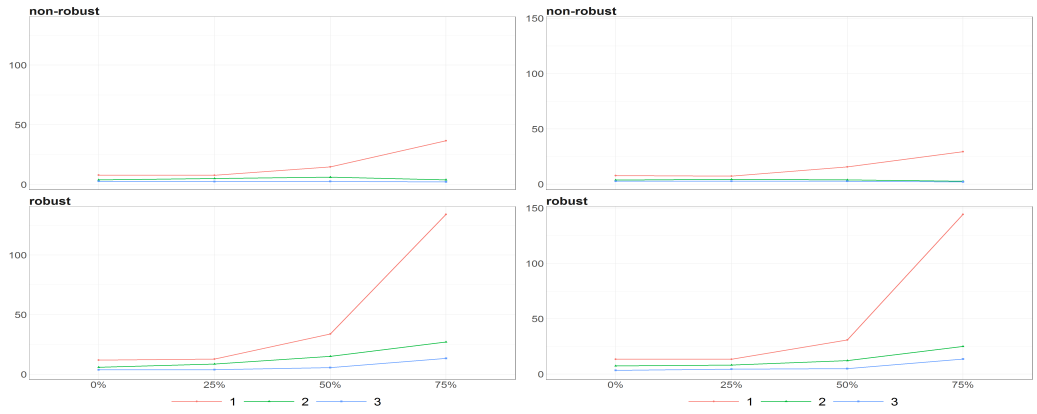


Figure 12.5. The Mahalanobias distances obtained using PPCA for Females (**a**) and Males (**b**) of Life Expectancy at Birth over time (x axis). Different colours of lines correspond to the cases of different percentages of maximal missing entries in rows (light blue (75%), dark brown (50%), dark blue (25%), light brown (0%)). Every subfigure is divided into two subplots corresponding to the robust estimation of standard divinations (upper plot) and sample one (bottom plot).



(a) **(b)**

Figure 12.6. The eigenvalues obtained using PPCA for Females (**a**) and Males (**b**) of Life Expectancy at Birth data for different percentages of maximal missing entries in rows (x axis). Colours of lines corresponds to different eigenvalues, first (light brown), second (dark blue) and third (dark brown) highest. Every subfigure is divided into two subplots corresponding to robust estimation of standard divinations (upper plot) and sample one (bottom plot).

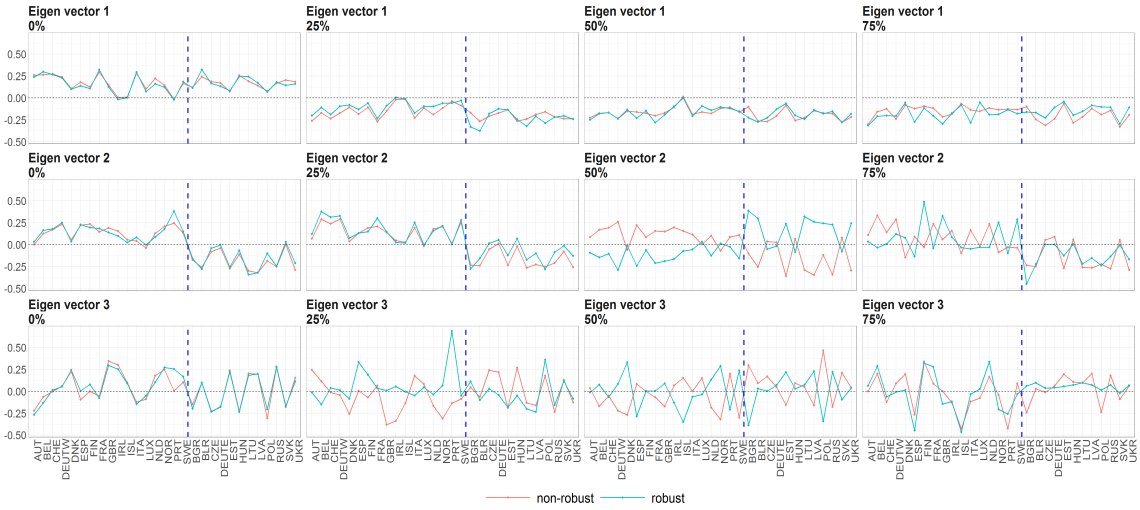


Figure 12.7. The eigenvectors (y axis) over the joint distribution of countries (x axis) obtained using PPCA for female population of Life Expectancy at Birth. Every row of subfigure corresponds to different eigenvector. Every column corresponds to different level of maximum missing values per observation (0%, 25%, 50% and 75%). The blue line corresponds to robust standardisation whereas red line to non-robust standardisation of data.

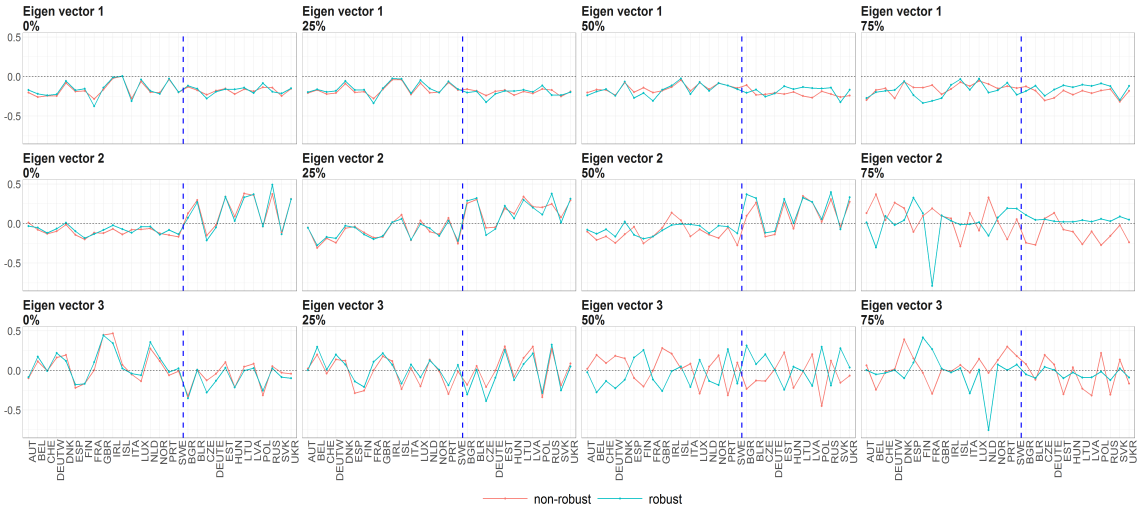


Figure 12.8. The eigenvectors (y axis) over the joint distribution of countries (x axis) obtained using PPCA for male population of Life Expectancy at Birth. Every row of subfigure corresponds to different eigenvector. Every column corresponds to different level of maximum missing values per observation (0%, 25%, 50% and 75%). The blue line corresponds to robust standardisation whereas red line to non-robust standardisation of data.

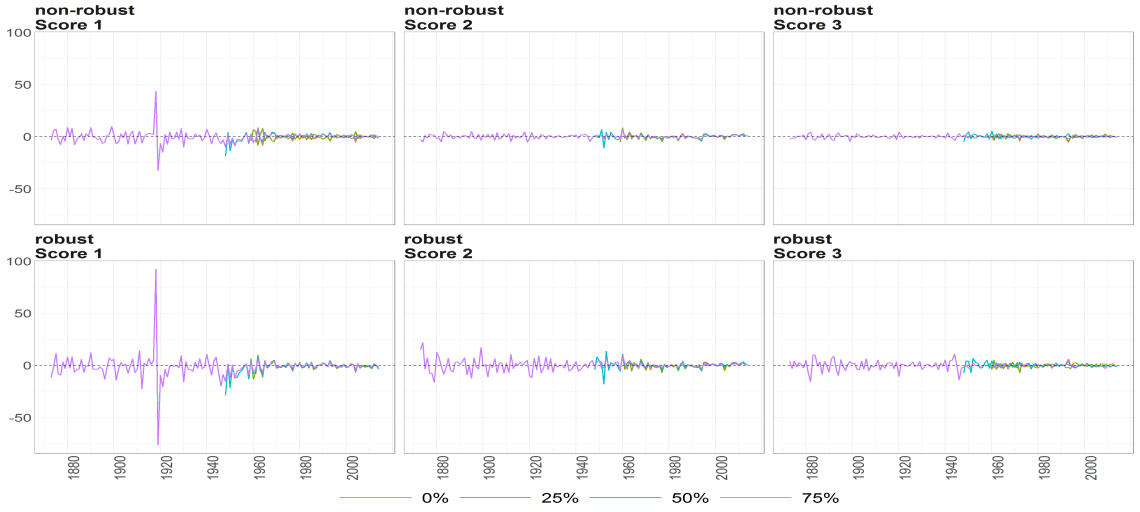


Figure 12.9. The scores (y axis) over time (x axis) obtained using PPCA for female population of Life Expectancy at Birth. Colours of lines correspond to the scores calculated on subsample where are different levels of maximum missing values per observation (0%, 25%, 50% and 75%, refer to legend). The plots placed in the first row correspond to the results using non- robust standardization of entry data, where the second row correspond to robust standardisation. The plots of scores are ordered by columns.



Figure 12.10. The scores (y axis) over the time (x axis) obtained using PPCA for Males of Life Expectancy at Birth. Colours of lines correspond to the scores calculated on subsample where are different levels of maximum missing values per observation (0%, 25%, 50% and 75%, refer to legend). The plots placed in the first row correspond to the results using non- robust standardization of entry data, where the second row correspond to robust standardisation. The plots of scores are ordered by columns.

12.5.4. Deaths

The Mahalanobias distances for the Deaths counts exhibit resembling behaviour for Females regardless of the standardisation procedure. The discrepancies between the statistics are more substantial for the male population what is highlighted in Figure 12.11.

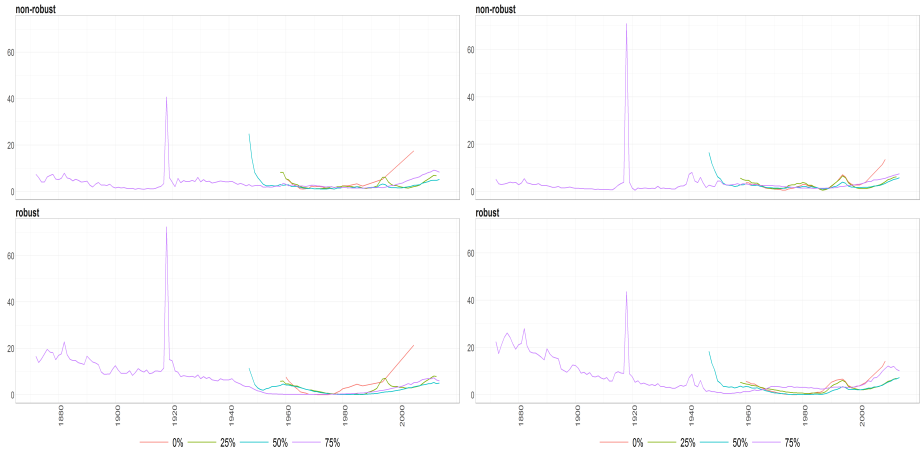
The eigenvalues of the non-robust estimator of the covariance matrix does not vary between the cases of missigness up to 50%. The plots are shown in Figure 12.12. For the case 75% the first eigenvalue starts to dominate more significantly than for other missigness cases, especially for Females. On the other hand, the robust eigenvalues are more volatile. Especially the results for Females provide unexpected outcomes for the case 50% in comparison to the case 25% even though the subsamples for these cases differ only by a few calendar years in mid 1940. With regards to the Males, we observe that the dominance of first eigenvalue increases with the number of missing values.

The corresponding eigenvectors are shown in Figures 12.14 and 12.15 for Females and Males respectively. The colours of the heatmaps correspond to the magnitude of components of eigenvector which are country (x axis) and age group (y axis) specific. The non-robust estimation results in the eigenvectors with smaller magnitude and smoother within the age groups and countries. Recall, that the distribution of colours for the robust case has bigger spreads between values (so called "bumps") what is highlighted by more intense colours of blue and red. The exception is made for the case of 75%.

The vertical black dotted lines on the heatmaps divides the countries listed on x axis as developed (left side) and developing (right side). Again, this order of results stresses the differences between the eigenvectors within these two groups of countries. The first eigenvector for developing countries has the break point around age group of 40 for all cases of missingness for the male

population and all cases except 50% for the female population. The eigenvector for developed countries has a break point for age group in 80 for female population and 75 for male population with additional break in 35 for 0% and 25% cases.

The case of 75 is analysed separately. The first eigenvectors do not differ within two types of standardisation but exhibit the structure which is country group specific. The developed countries are characterized by the uniformed values around zero for all age groups. The vectors of developing countries are more volatile with breaks around 20–30 for Males and even more volatile for Females. The second eigenvector differs within both types of standardisation and among two groups of countries. It is almost constant around zero for Males in developing countries and more volatile for Males in developed countries with breaks in age groups between 50 and 60. Second eigenvector for Females resembles third eigenvector of Males .



(a) **(b)**

Figure 12.11. The Mahalanobias distances obtained using PPCA for Females (**a**) and Males (**b**) of Number of Death over time (x axis). Different colours of lines correspond to the cases of different percentages of maximal missing entries in rows (light blue (75%), dark brown (50%), dark blue (25%), light brown (0%)). Every subfigure is divided into two subplots corresponding to robust estimation of standard divinations (upper plot) and sample one (bottom plot).

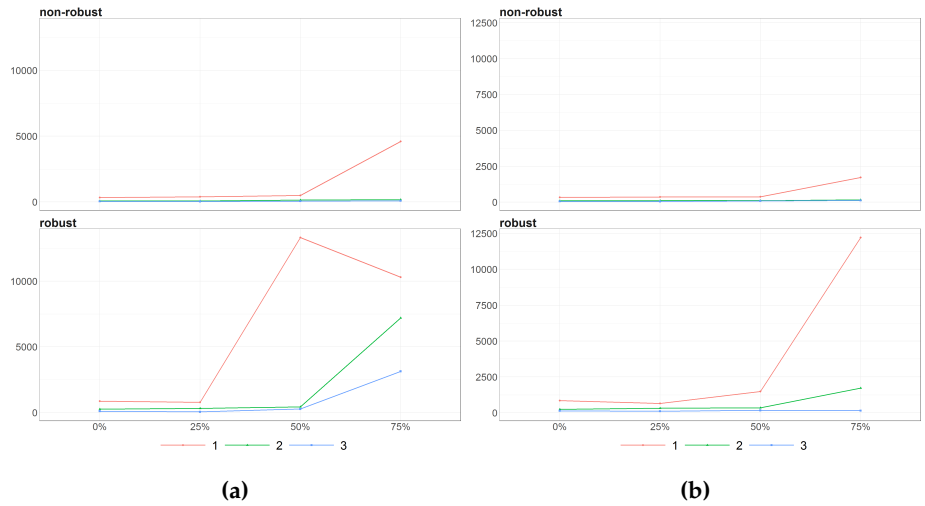


Figure 12.12. The eigenvalues of Deaths counts obtained using PPCA for Females(**a**) and Males (**b**) over different cases of missing entries (x axis). Colours of lines corresponds to different eigenvalues, first (light brown), second (dark blue) and third (dark brown) highest. Every subfigure is divided into two subplots corresponding to robust estimation of standard deviations (upper plot) and sample one (bottom plot).

The scores are presented in Figures 12.13 and 12.16. The second and third scores of Males are smoother in contrary to results for Females. The exception is made for the case 75% which results vary both among sexes and standardization procedures.

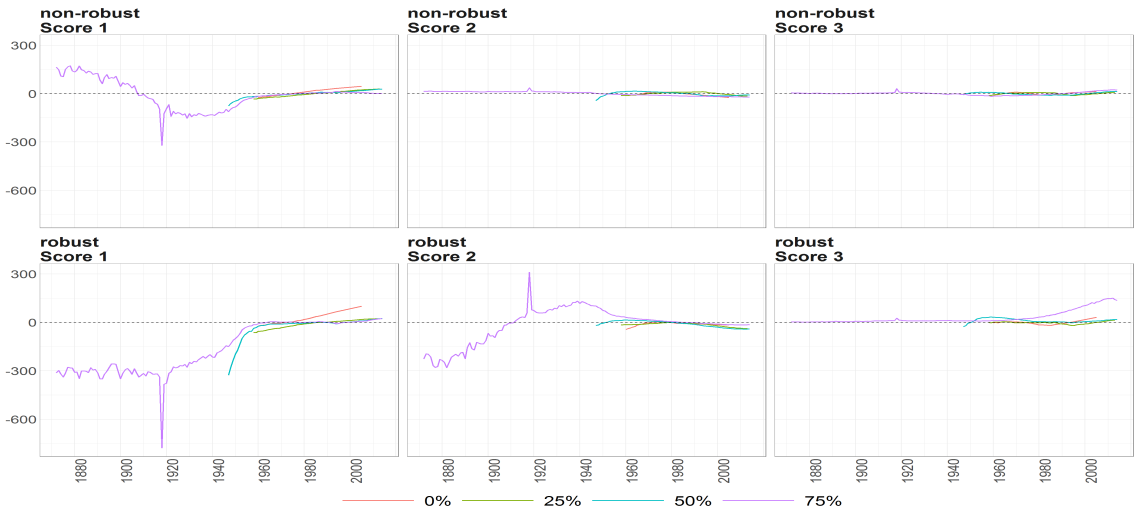


Figure 12.13. The scores (y axis) over time (x axis) obtained using PPCA for female population of Number of Deaths. Colours of lines correspond to the scores calculated on subsample where are different levels of maximum missing values per observation (0%, 25%, 50% and 75%, refer to legend). The plots placed in the first row correspond to the results using non- robust standardization of entry data, where the second row correspond to robust standardisation. The plots of scores are ordered by columns.

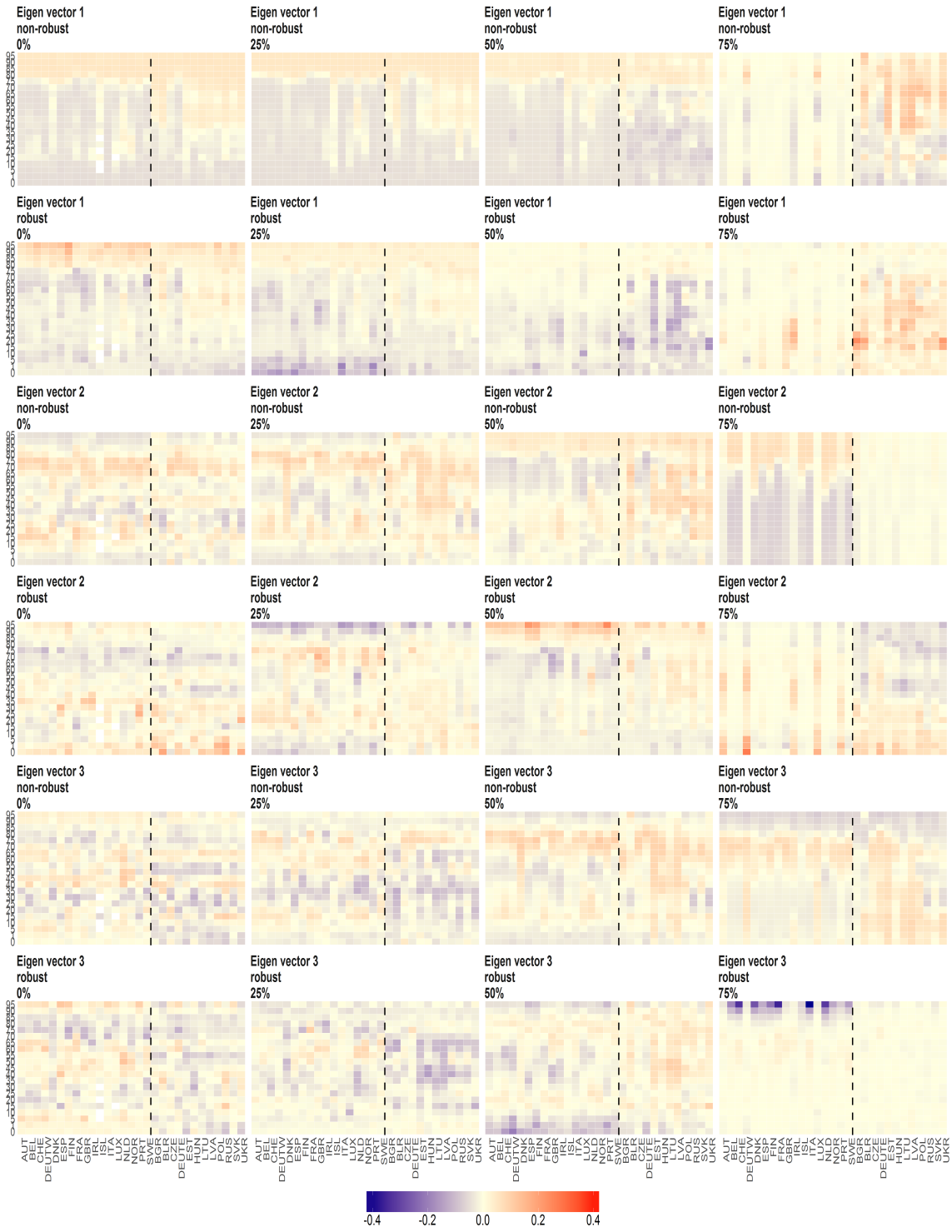


Figure 12.14. The eigenvectors of Death counts (y axis) over age groups (y axis) and countries (x axis) obtained using PPCA for Females. Every row of subfigure corresponds to different eigenvector. Every column corresponds to different level of maximum missing values per observation (0%, 25%, 50% and 75%). The blue line corresponds to robust standardisation whereas red line to non-robust standardisation of data.

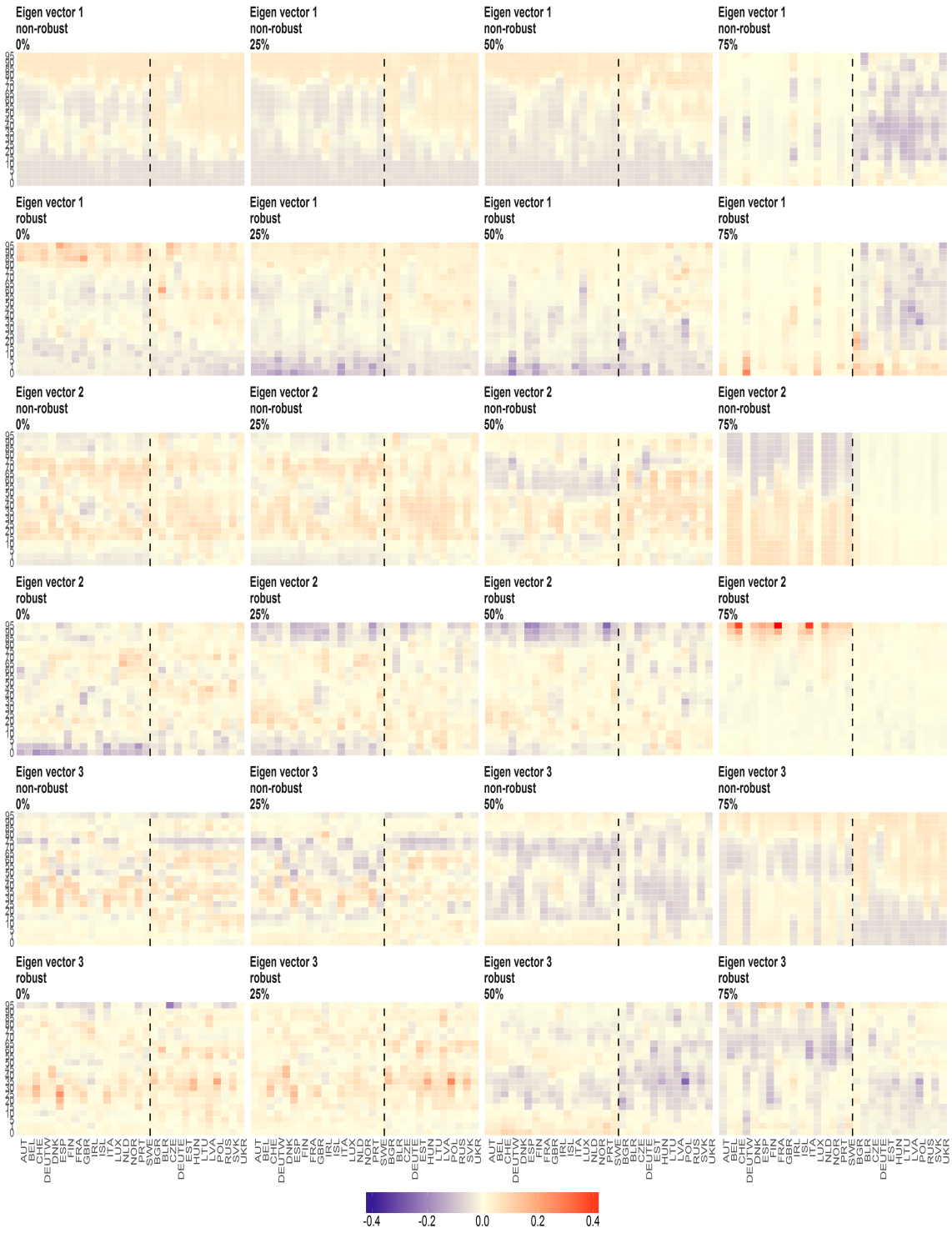


Figure 12.15. The eigenvectors of Death counts over age groups (y axis) and countries (x axis) obtained using PPCA for Males . Every row of subfigure corresponds to different eigenvector. Every column corresponds to different level of maximum missing values per observation (0%, 25%, 50% and 75%). The blue line corresponds to robust standardisation whereas red line to non-robust standardisation of data.

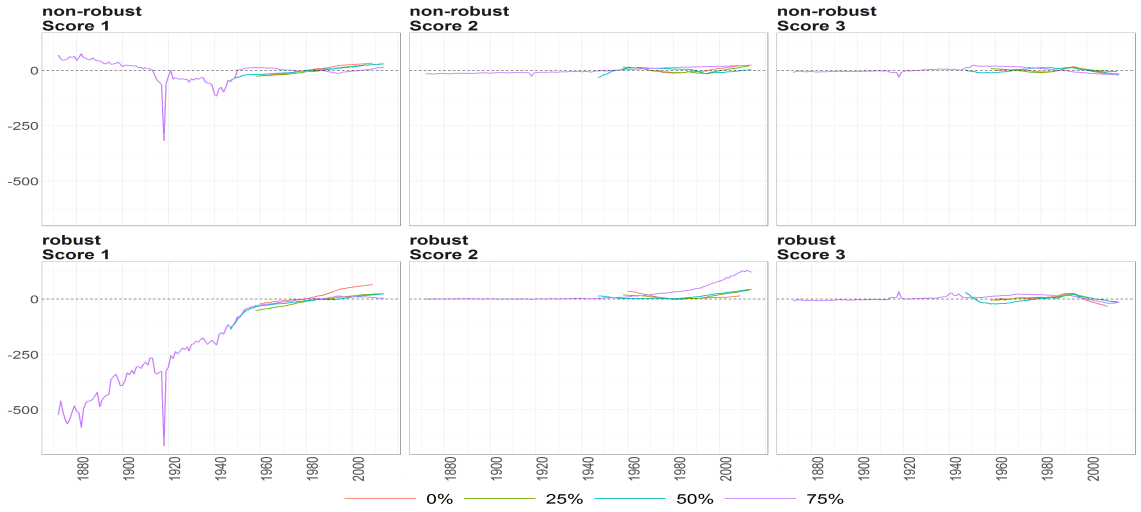


Figure 12.16. The scores (y axis) over time (x axis) obtained using PPCA for male population of Number of Deaths. Colours of lines correspond to the scores calculated on subsample where are different levels of maximum missing values per observation (0%, 25%, 50% and 75%, refer to legend). The plots placed in the first row correspond to the results using non- robust standardization of entry data, where the second row correspond to robust standardisation. The plots of scores are ordered by columns.

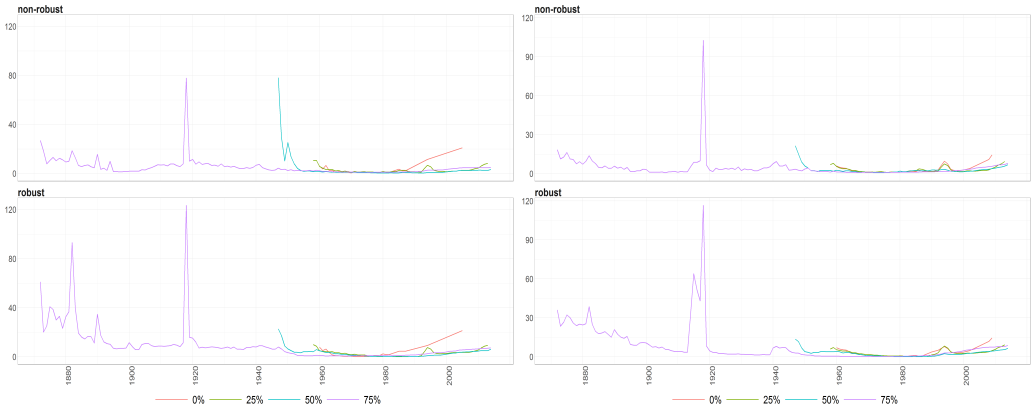
12.5.5. Death Rates

The analysis for Death Rates provides with similar conclusions as for Death counts. The corresponding Figures are 12.17, 12.18, 12.21 and 12.22 respectively. The only discrepancies are exhibited by the eigenvalues for Males for the 75% case as they are more aligned and higher in terms of magnitude from other cases than the corresponding eigenvalues in the Deaths counts analysis.

As shown in Figures 12.19 and 12.20 the robustification does not influenced greatly the estimation of eigenvectors. As the Deaths Rates are preprocessed and smoothed before being distributed by HMD, we again conclude that the preprocessing decreased number of outlying data points. In this particular case, the robust standardisation is similarly informative about the the true distribution as the non robust one.

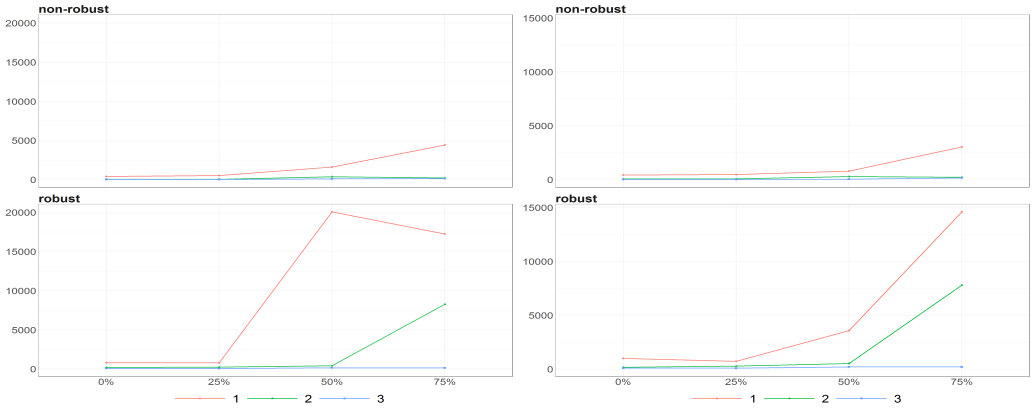
Also, recall the similarity of results among different cases of missingness, especially for first eigenvector. The second and third ones are smoother for high levels of missigness. The colour map is affected by the scaling parameter $(1, -1)$ which may cause red to become blue, but except this fact, we see resembling outcomes.

The three most meaningful eigenvectors differ among two groups of analysed countries: developed and developing. The corresponding results are divided by the vertical black dotted lines. The first eigenvectors of developed countries is very smooth among age groups. For developing countries we observe U shape structure with peak in 45–60 age groups for Males and similarly for Females in the cases with small proportion of missigness. When we allow more missing values, the eigenvectors for Females in developed countries are closer to zero and flat among age groups whereas for developing countries are more distant and volatile.



(a) **(b)**

Figure 12.17. The Mahalanobias distances obtained using PPCA for female (**a**) and male (**b**) population of Death Rates over time (x axis). Different colours of lines correspond to the cases of different percentages of maximal missing entries in rows (light blue (75%), dark brown (50%), dark blue (25%), light brown (0%)). Every subfigure is divided into two subplots corresponding to robust estimation of standard divinations (upper plot) and sample one (bottom plot).



(a) **(b)**

Figure 12.18. The eigenvalues obtained using PPCA for female (**a**) and male (**b**) population of Death Rates for different percentages of maximal missing entries in rows (x axis). Colours of lines corresponds to different eigenvalues, first (light brown), second (dark blue) and third (dark brown) highest. Every subfigure is divided into two subplots corresponding to robust estimation of standard divinations (upper plot) and sample one (bottom plot).

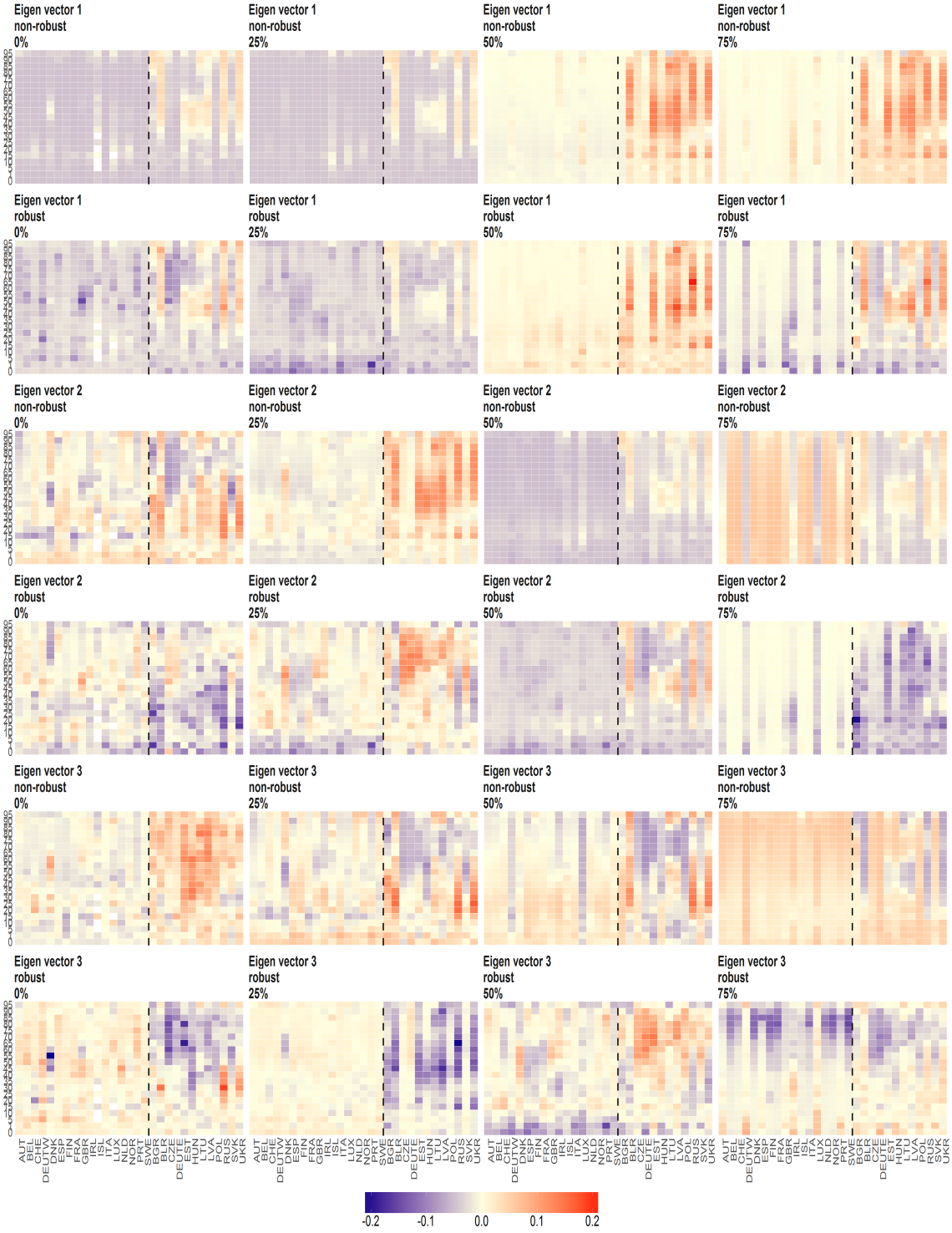


Figure 12.19. The eigenvectors (y axis) over the joint distribution of countries (x axis) obtained using PPCA for female population of Death Rates. Every row of subfigure corresponds to different eigenvector. Every column corresponds to different level of maximum missing values per observation (0%, 25%, 50% and 75%). The blue line corresponds to robust standardisation whereas red line to non-robust standardisation of data.

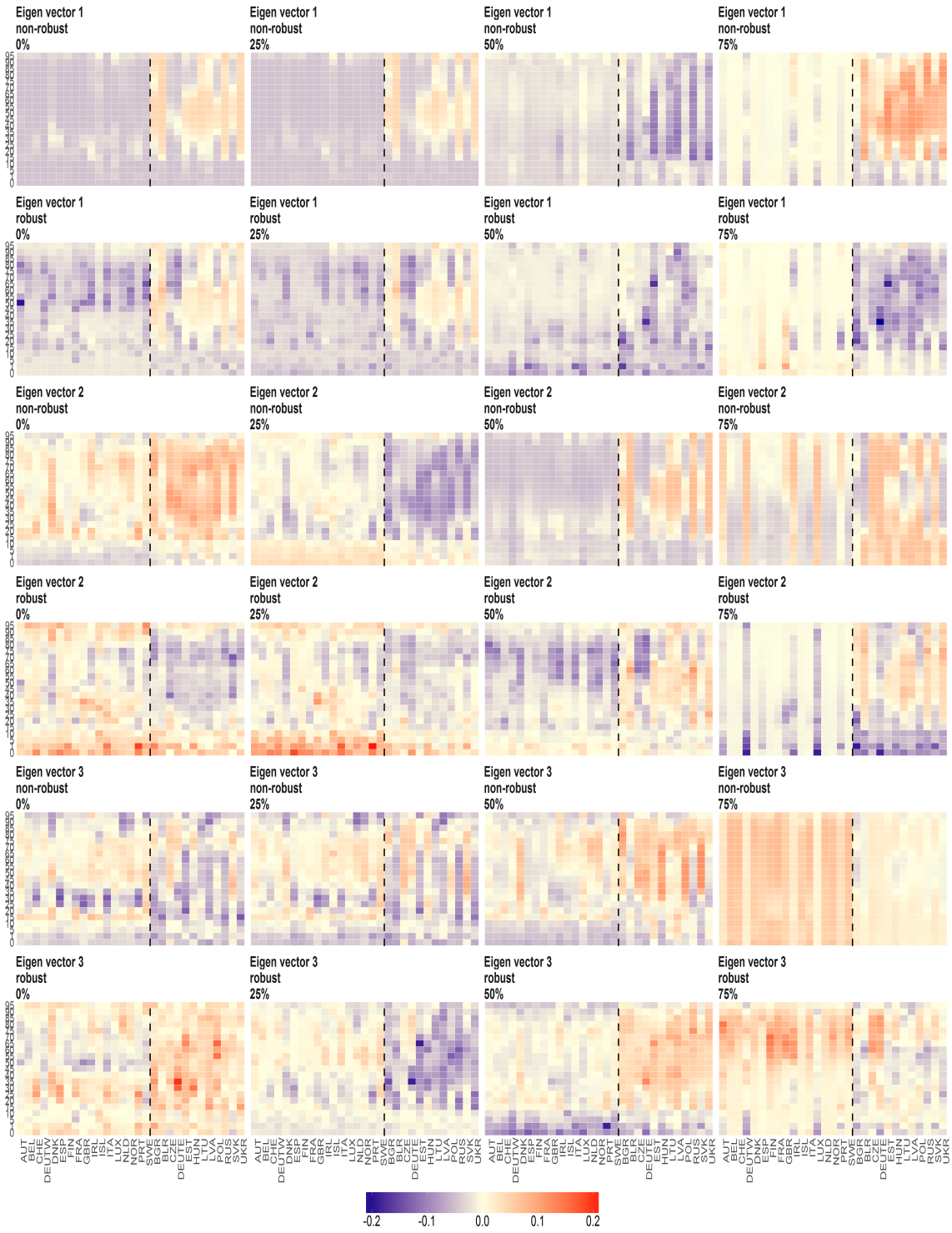


Figure 12.20. The eigenvectors (y axis) over the joint distribution of countries (x axis) obtained using PPCA for male population of Death Rates. Every row of subfigure corresponds to different eigenvector. Every column corresponds to different level of maximum missing values per observation (0%, 25%, 50% and 75%). The blue line corresponds to robust standardisation whereas red line to non-robust standardisation of data.

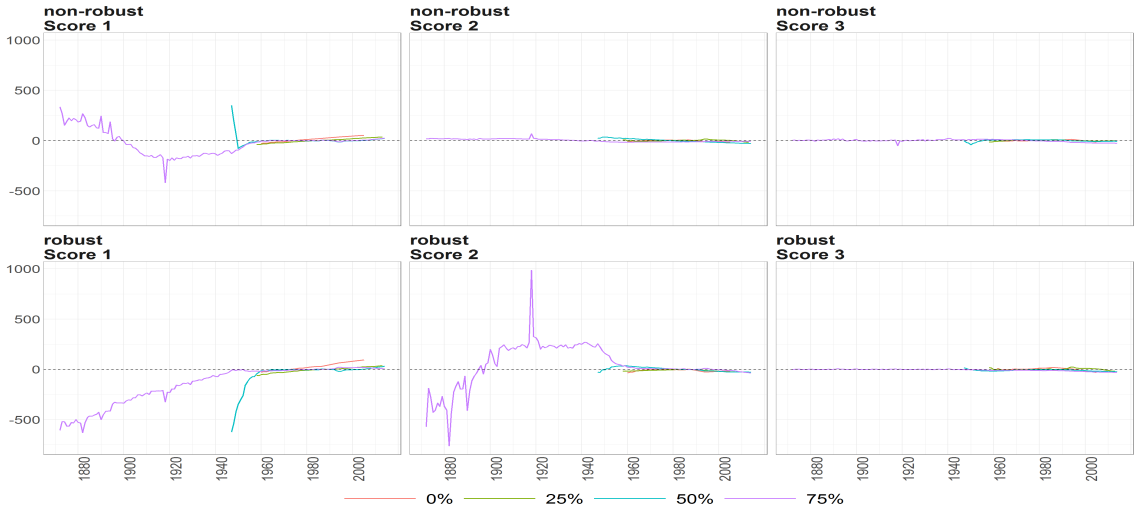


Figure 12.21. The scores (y axis) over time (x axis) obtained using PPCA for female population of Death Rates. Colours of lines correspond to the scores calculated on subsample where are different levels of maximum missing values per observation (0%, 25%, 50% and 75%, refer to legend). The plots placed in the first row correspond to the results using non- robust standardization of entry data, where the second row correspond to robust standardisation. The plots of scores are ordered by columns.

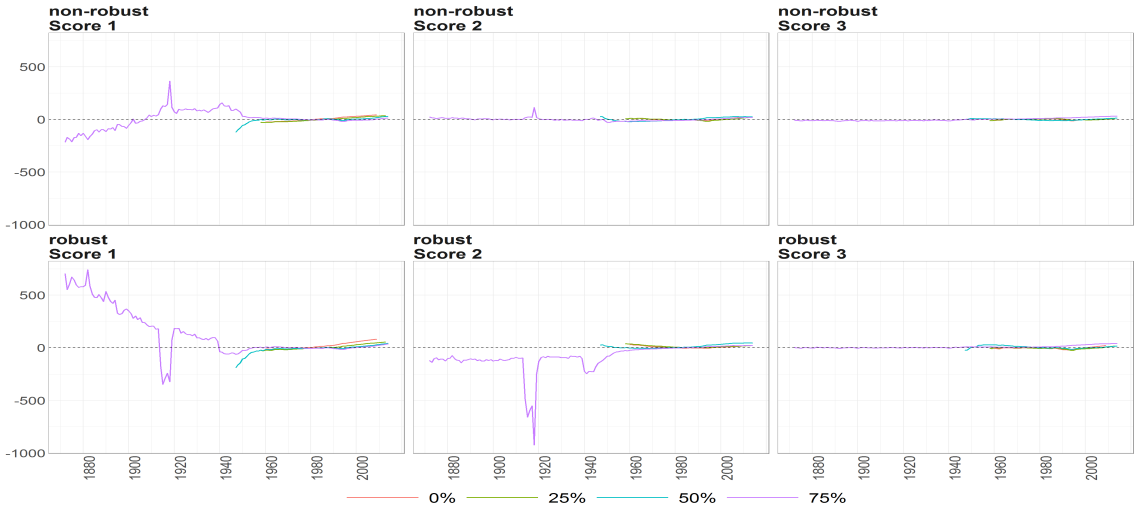


Figure 12.22. The scores (y axis) over time (x axis) obtained using PPCA for male population of Death Rates. Colours of lines correspond to the scores calculated on subsample where are different levels of maximum missing values per observation (0%, 25%, 50% and 75%, refer to legend). The plots placed in the first row correspond to the results using non- robust standardization of entry data, where the second row correspond to robust standardisation. The plots of scores are ordered by columns.

12.6. Bayesian Modelling and Sampling of Demographic Factor Model Extension to the Period-Cohort Stochastic Mortality State-Space Models

In this appendix section we explain the Bayesian models developed for the estimation of the demographic factor model extension to the period-cohort stochastic mortality state-space models

that are applied in this paper. These are based on the frameworks detailed and developed in (Fung et al. 2017) and (Fung et al. 2018).

12.6.1. Bayesian Model Development and Inference for Stochastic Mortality Models

The observation and state equations Equations (12.5a) and (12.5b) imply that the cohort model that we have formulated here belongs to the linear-Gaussian class of state-space models. As a result one can perform efficient maximum-likelihood or Bayesian estimation on fitting the model to data as discussed in detail in (Fung et al. 2017). In this study we focus on Bayesian inference so that the forecasting distribution can take into account parameter uncertainty. The general framework for Bayesian filtering and prediction is provided in Chapter 8.

In this section we detail the Bayesian estimation of the cohort model Equations (12.5a) and (12.5b) and its extensions incorporating population information to state-space formulation Equations (12.8a) and (12.8b). We firstly note that the models belong to the linear and Gaussian class of state-space models. As a result one can apply an efficient MCMC estimation algorithm based on Gibbs sampling with conjugate priors combined with forward-backward filtering as described in Section 8.2. We describe procedures using the notation of general cohort model Equations (12.5a) and (12.5b) and indicate the differences to considered extensions.

We borrow the notation from the LC cohort model Equations (12.5a) and (12.5b) and indicate similarities and differences between Equations (12.5a) and (12.5b) and Equations (12.8a) and (12.8b) while developing the estimation algorithm for the target density, a joint posterior density of the states $\boldsymbol{\varrho}_{0:T}$ and the vector of static parameters $\boldsymbol{\Psi}$, is the following

$$\pi_{\boldsymbol{\varphi}_{0:T},\boldsymbol{\Psi}|\boldsymbol{Y}_{1:T}}\left(\boldsymbol{\varphi}_{0:T},\boldsymbol{\psi}\right). \tag{12.14}$$

In the case of the cohort LC model $\boldsymbol{\varphi}_{0:T} := (\kappa_{0:T}, \gamma_{0:T}^{x_1}, \ldots, \gamma_{0:T}^{x_p})$ is the $p+1$ dimensional (for each t) latent state vector and $\boldsymbol{\Psi} := (\alpha_{x_3:x_p}, \beta_{x_2:x_p}, \theta, \eta, \lambda, \sigma_\varepsilon^2, \sigma_\kappa^2, \sigma_\gamma^2)$ is the $2p-1$ dimensional static parameter vector. For extended models, we add vectors of global factors latent variable $\boldsymbol{\varrho}_t$ and its static model parameters. Recall that our proposed identification constraint is given by Equation (12.6); therefore only $\alpha_{x_3:x_p}$ and $\beta_{x_2:x_p}$ are required to be estimated. We perform block sampling for the latent state via the so-called forward-filtering-backward-sampling (FFBS) algorithm (Carter and Kohn 1994b) and the posterior samples of the static parameters are obtained via conjugate priors. The general sampling procedure is described in Subsection 8.2.

12.6.2. Forward-Backward Filtering for Latent State Dynamics

The FFBS procedure is discussed in Subsection 8.2.1. For the cohort model Equations (12.2) and (12.3) (or Equations (12.5a) and (12.5b)), the conditional distributions involved in the multivariate Kalman filtering recursions are given by

$$\boldsymbol{\varphi}_{t-1}|\boldsymbol{y}_{1:t-1} \sim \mathcal{N}(\boldsymbol{m}_{t-1}, C_{t-1}),$$

$$\begin{aligned}\boldsymbol{\varphi}_t|\boldsymbol{y}_{1:t-1} &\sim \mathcal{N}(\boldsymbol{a}_t, R_t),\\ \boldsymbol{y}_t|\boldsymbol{y}_{1:t-1} &\sim \mathcal{N}(\boldsymbol{f}_t, Q_t),\\ \boldsymbol{\varphi}_t|\boldsymbol{y}_{1:t} &\sim \mathcal{N}(\boldsymbol{m}_t, C_t)\end{aligned}$$

where

$$\boldsymbol{a}_t = \boldsymbol{\Lambda}\boldsymbol{m}_{t-1} + \boldsymbol{\Theta}, \quad R_t = \boldsymbol{\Lambda}C_{t-1}\boldsymbol{\Lambda}^\top + \mathrm{Y}, \tag{12.16a}$$

$$\boldsymbol{f}_t = \boldsymbol{\alpha} + \mathbf{B}\boldsymbol{a}_t, \quad Q_t = \mathbf{B}R_t\mathbf{B}^\top + \sigma_\varepsilon^2\mathbb{I}_p, \tag{12.16b}$$

$$\boldsymbol{m}_t = \boldsymbol{a}_t + R_t\mathbf{B}^\top Q_t^{-1}(\boldsymbol{y}_t - \boldsymbol{f}_t), \quad C_t = R_t - R_t\mathbf{B}^\top Q_t^{-1}\mathbf{B}R_t. \tag{12.16c}$$

for $t = 1, \ldots, T$.The backward sampling follows $\boldsymbol{\varphi}_t|_{\boldsymbol{\varphi}_{t+1},\boldsymbol{\psi},\boldsymbol{y}_{1:T}} \sim \mathcal{N}(\boldsymbol{h}_t, H_t)$ where

$$\boldsymbol{h}_t = \boldsymbol{m}_t + C_t\boldsymbol{\Lambda}^\top R_{t+1}^{-1}(\boldsymbol{\varphi}_{t+1} - \boldsymbol{a}_{t+1}), \tag{12.17a}$$

$$H_t = C_t - C_t\boldsymbol{\Lambda}^\top R_{t+1}^{-1}\boldsymbol{\Lambda}C_t, \tag{12.17b}$$

For the extended models, we simply need to replace the vectors and matrices of the LC cohort model with objects stressed by tildes from Section 12.3.

12.6.3. Posteriors for Static Parameters in the Cohort Model

To sample the posterior distribution of the static parameters, we assume the following independent conjugate priors:

$$\begin{aligned}&\alpha_x \sim \mathcal{N}(\tilde{\mu}_\alpha, \tilde{\sigma}_\alpha^2), \quad \beta_x \sim \mathcal{N}(\tilde{\mu}_\beta, \tilde{\sigma}_\beta^2), \quad \theta \sim \mathcal{N}(\tilde{\mu}_\theta, \tilde{\sigma}_\theta^2), \quad \eta \sim \mathcal{N}(\tilde{\mu}_\eta, \tilde{\sigma}_\eta^2)\\ &\lambda \sim \mathcal{N}_{[-1,1]}(\tilde{\mu}_\lambda, \tilde{\sigma}_\lambda^2), \quad \sigma_\varepsilon^2 \sim \mathrm{IG}(\tilde{a}_\varepsilon, \tilde{b}_\varepsilon), \quad \sigma_\kappa^2 \sim \mathrm{IG}(\tilde{a}_\kappa, \tilde{b}_\kappa), \quad \sigma_\gamma^2 \sim \mathrm{IG}(\tilde{a}_\gamma, \tilde{b}_\gamma)\end{aligned} \tag{12.18a}$$

where $\mathrm{N}_{[-1,1]}$ denotes a truncated Gaussian with support $[-1, 1]$ and $\mathrm{IG}(\tilde{a}, \tilde{b})$ denotes an inverse-gamma distribution with mean $\tilde{b}/(\tilde{a}-1)$ and variance $\tilde{b}^2/((\tilde{a}-1)^2(\tilde{a}-2))$ for $\tilde{a} > 2$. The posteriors of the static parameters are then obtained as follows:[1]

$$\begin{aligned}\alpha_x|\boldsymbol{y},\boldsymbol{\varphi},\boldsymbol{\psi}_{-\alpha_x} &\sim \mathcal{N}\left(\frac{\tilde{\sigma}_\alpha^2\sum_{t=1}^T(y_{x,t} - \beta_x\kappa_t - \gamma_t^x) + \tilde{\mu}_\alpha\sigma_\varepsilon^2}{\tilde{\sigma}_\alpha^2 T + \sigma_\varepsilon^2}, \frac{\tilde{\sigma}_\alpha^2\sigma_\varepsilon^2}{\tilde{\sigma}_\alpha^2 T + \sigma_\varepsilon^2}\right),\\ \beta_x|\boldsymbol{y},\boldsymbol{\varphi},\boldsymbol{\psi}_{-\beta_x} &\sim \mathcal{N}\left(\frac{\tilde{\sigma}_\beta^2\sum_{t=1}^T(y_{x,t} - (\alpha_x + \gamma_t^x))\kappa_t + \tilde{\mu}_\beta\sigma_\varepsilon^2}{\tilde{\sigma}_\beta^2\sum_{t=1}^T\kappa_t^2 + \sigma_\varepsilon^2}, \frac{\tilde{\sigma}_\beta^2\sigma_\varepsilon^2}{\tilde{\sigma}_\beta^2\sum_{t=1}^T\kappa_t^2 + \sigma_\varepsilon^2}\right),\\ \theta|\boldsymbol{y},\boldsymbol{\varphi},\boldsymbol{\psi}_{-\theta} &\sim \mathcal{N}\left(\frac{\tilde{\sigma}_\theta^2\sum_{t=1}^T(\kappa_t - \kappa_{t-1}) + \tilde{\mu}_\theta\sigma_\kappa^2}{\tilde{\sigma}_\theta^2 T + \sigma_\kappa^2}, \frac{\tilde{\sigma}_\theta^2\sigma_\kappa^2}{\tilde{\sigma}_\theta^2 T + \sigma_\kappa^2}\right),\\ \eta|\boldsymbol{y},\boldsymbol{\varphi},\boldsymbol{\psi}_{-\theta} &\sim \mathcal{N}\left(\frac{\tilde{\sigma}_\eta^2\sum_{t=1}^T(\gamma_t - \lambda\gamma_{t-1}) + \tilde{\mu}_\eta\sigma_\gamma^2}{\tilde{\sigma}_\eta^2 T + \sigma_\gamma^2}, \frac{\tilde{\sigma}_\eta^2\sigma_\gamma^2}{\tilde{\sigma}_\eta^2 T + \sigma_\gamma^2}\right),\end{aligned}$$

[1] For simplicity, we denote $\boldsymbol{y} = \boldsymbol{y}_{1:T}$, $\boldsymbol{\varphi} = \boldsymbol{\varphi}_{0:T}$ and $\boldsymbol{\psi}_{-h} = (\psi_1, \ldots, \psi_{h-1}, \psi_{h+1}, \ldots, \psi_{2p-2})$.

$$\lambda|\boldsymbol{y},\boldsymbol{\varphi},\boldsymbol{\psi}_{-\lambda} \sim \mathcal{N}_{[-1,1]}\left(\frac{\tilde{\sigma}_\lambda^2\sum_{t=1}^{T}((\gamma_t^{x_1}-\eta)\gamma_{t-1}^{x_1})+\tilde{\mu}_\lambda\sigma_\gamma^2}{\tilde{\sigma}_\lambda^2\sum_{t=1}^{T}(\gamma_{t-1}^{x_1})^2+\sigma_\gamma^2}, \frac{\tilde{\sigma}_\lambda^2\sigma_\gamma^2}{\tilde{\sigma}_\lambda^2\sum_{t=1}^{T}(\gamma_{t-1}^{x_1})^2+\sigma_\gamma^2}\right),$$

$$\sigma_\varepsilon^2|\boldsymbol{y},\boldsymbol{\varphi},\boldsymbol{\psi}_{-\sigma_\varepsilon^2} \sim \text{IG}\left(\tilde{a}_\varepsilon+\frac{pT}{2}, \tilde{b}_\varepsilon+\frac{1}{2}\sum_{x=x_1}^{x_p}\sum_{t=1}^{T}(y_{x,t}-(\alpha_x+\beta_x\kappa_t+\gamma_t^x))^2\right),$$

$$\sigma_\kappa^2|\boldsymbol{y},\boldsymbol{\varphi},\boldsymbol{\psi}_{-\sigma_\kappa^2} \sim \text{IG}\left(\tilde{a}_\kappa+\frac{T}{2}, \tilde{b}_\kappa+\frac{1}{2}\sum_{t=1}^{T}(\kappa_t-(\kappa_{t-1}+\theta))^2\right),$$

$$\sigma_\gamma^2|\boldsymbol{y},\boldsymbol{\varphi},\boldsymbol{\psi}_{-\sigma_\gamma^2} \sim \text{IG}\left(\tilde{a}_\gamma+\frac{T}{2}, \tilde{b}_\gamma+\frac{1}{2}\sum_{t=1}^{T}(\gamma_t^{x_1}-\lambda\gamma_{t-1}^{x_1})^2\right).$$

12.6.4. Posteriors for Static Parameters in the Extended Models

In order to develope sampling algorithm for models incorporating European countries population information, we add the conjugate priori assumptions to Equation (12.18) related to static parameters of Equation (12.7)

$$[\boldsymbol{\Omega}]_{i,j} \sim \mathcal{N}(\tilde{\mu}_\Omega, \tilde{\sigma}_\Omega^2), \quad \Psi_j \sim \mathcal{N}(\tilde{\mu}_\Psi, \tilde{\sigma}_\Psi^2), \quad \sigma_\varrho^2 \sim \text{IG}(\tilde{a}_\varrho, \tilde{b}_\varrho) \tag{12.19}$$

The posteriors of static parameters from Equation (12.7) are essential for Gibbs backward sampling regardless of the considered case for the extended model. We refer by small letters i, j to ages x_i, $x_j \in \{x_1,\ldots,x_p\}$, for $i,\ j \in \{1,\ldots,p\}$, and letters $m,\ l \in \{1,\ldots,k\}$ to the components of the matrix $\mathbf{F}_t$. An element labelled as im corresponds to $[\mathbf{F}_t]_{i,m}$ or a latent variable ϱ_t^{im}. Then

$$\Psi_{im}|\boldsymbol{y},\tilde{\boldsymbol{\varphi}},\boldsymbol{\psi}_{-\Psi_{im}} \sim \text{N}\left(\frac{\tilde{\sigma}_\Psi^2\sum_{t=1}^{T}(\varrho_t^{im}-\boldsymbol{\varrho}_{t-1}^T[\boldsymbol{\Omega}]_{im,\cdot})+\tilde{\mu}_\psi\sigma_\varrho^2}{\tilde{\sigma}_\Psi^2 T+\sigma_\varrho^2}, \frac{\tilde{\sigma}_\Psi^2\sigma_\varrho^2}{\tilde{\sigma}_\Psi^2 T+\sigma_\varrho^2}\right),$$

$$[\boldsymbol{\Omega}]_{im,jl}|\boldsymbol{y},\tilde{\boldsymbol{\varphi}},\boldsymbol{\psi}_{-[\boldsymbol{\Omega}]_{im,jl}} \sim$$
$$\mathcal{N}\left(\frac{\tilde{\sigma}_\Omega^2\sum_{t=1}^{T}\left[\left(\varrho_t^{im}-\Psi_{im}-\sum_{xh\neq jm}\varrho_{t-1}^{xh}[\boldsymbol{\Omega}]_{im,xh}\right)\varrho_{t-1}^{jl}\right]+\tilde{\mu}_\Omega\sigma_\varrho^2}{\tilde{\sigma}_\Omega^2\sum_{t=1}^{T}(\varrho_{t-1}^{jl})^2+\sigma_\varrho^2}, \frac{\tilde{\sigma}_\Omega^2\sigma_\varrho^2}{\tilde{\sigma}_\Omega^2\sum_{t=1}^{T}(\varrho_{t-1}^{jl})^2+\sigma_\varrho^2}\right),$$

$$\sigma_\varrho^2|\boldsymbol{y},\tilde{\boldsymbol{\varphi}},\boldsymbol{\psi}_{-\sigma_\varrho^2} \sim \text{IG}\left(\tilde{a}_\varrho+\frac{pkT}{2}, \tilde{b}_\varrho+\frac{1}{2}\sum_{t=1}^{T}\sum_{i=1}^{p}\sum_{m=1}^{k}\left(\varrho_t^{im}-\Psi_{im}-\boldsymbol{\varrho}_{t-1}^T[\boldsymbol{\Omega}]_{im,\cdot}\right)^2\right),$$

where $\tilde{\boldsymbol{\varphi}}$ is a vector of latent variables from Equation (12.8b) and $\boldsymbol{\psi}$ is a vector of static parameters updated to Equation (12.7). Depends on the cases f extended model, we have the following replacement of the posteriori distributions from Section 12.6.3

Case 1 Global factors $\mathbf{F}_t$ in the observation equation

$$\alpha_x|\boldsymbol{y},\tilde{\boldsymbol{\varphi}},\boldsymbol{\psi}_{-\alpha_x} \sim \mathcal{N}\left(\frac{\tilde{\sigma}_\alpha^2\sum_{t=1}^{T}\left(y_{x,t}-[\tilde{\mathbf{B}}_t]_{x,\cdot}^T\tilde{\boldsymbol{\varphi}}_t\right)+\tilde{\mu}_\alpha\sigma_\varepsilon^2}{\tilde{\sigma}_\alpha^2 T+\sigma_\varepsilon^2}, \frac{\tilde{\sigma}_\alpha^2\sigma_\varepsilon^2}{\tilde{\sigma}_\alpha^2 T+\sigma_\varepsilon^2}\right),$$

$$\beta_x|\boldsymbol{y},\tilde{\boldsymbol{\varphi}},\boldsymbol{\psi}_{-\beta_x} \sim \mathcal{N}\left(\frac{\tilde{\sigma}_\beta^2\sum_{t=1}^{T}\left[y_{x,t}-\left(\alpha_x+\boldsymbol{\gamma}_t^x+[\tilde{\mathbf{F}}_t]_{x,\cdot}^T\boldsymbol{\varrho}_t\right)\right]\kappa_t+\tilde{\mu}_\beta\sigma_\varepsilon^2}{\tilde{\sigma}_\beta^2\sum_{t=1}^{T}\kappa_t^2+\sigma_\varepsilon^2},\frac{\tilde{\sigma}_\beta^2\sigma_\varepsilon^2}{\tilde{\sigma}_\beta^2\sum_{t=1}^{T}\kappa_t^2+\sigma_\varepsilon^2}\right)$$

$$\sigma_\varepsilon^2|\boldsymbol{y},\tilde{\boldsymbol{\varphi}},\boldsymbol{\psi}_{-\sigma_\varepsilon^2} \sim \text{IG}\left(\tilde{a}_\varepsilon+\frac{pT}{2},\tilde{b}_\varepsilon+\frac{1}{2}\sum_{x=x_1}^{x_p}\sum_{t=1}^{T}\left[y_{x,t}-\left(\alpha_x+[\tilde{\mathbf{B}}_t]_{x,\cdot}T\tilde{\boldsymbol{\varphi}}_t\right)\right)^2\right]$$

Case 2 Global factors $\mathbf{F}_t$ in the state equation of κ_t

$$\theta|\boldsymbol{y},\tilde{\boldsymbol{\varphi}},\boldsymbol{\psi}_{-\theta} \sim \mathcal{N}\left(\frac{\tilde{\sigma}_\theta^2\sum_{t=1}^{T}(\kappa_t-\kappa_{t-1}-\tilde{\mathbf{f}}_t^T\boldsymbol{\varrho}_t)+\tilde{\mu}_\theta\sigma_\kappa^2}{\tilde{\sigma}_\theta^2T+\sigma_\kappa^2},\frac{\tilde{\sigma}_\theta^2\sigma_\kappa^2}{\tilde{\sigma}_\theta^2T+\sigma_\kappa^2}\right),$$

$$\sigma_\kappa^2|\boldsymbol{y},\boldsymbol{\varphi},\boldsymbol{\psi}_{-\sigma_\kappa^2} \sim \text{IG}\left(\tilde{a}_\kappa+\frac{T}{2},\tilde{b}_\kappa+\frac{1}{2}\sum_{t=1}^{T}\left[\kappa_t-\left(\kappa_{t-1}+\theta+\tilde{\mathbf{f}}_t^T\boldsymbol{\varrho}_t\right)\right]^2\right)$$

Case 3 Global factors $\mathbf{F}_t$ in the state equation of $\boldsymbol{\gamma}_t$

$$\eta|\boldsymbol{y},\tilde{\boldsymbol{\varphi}},\boldsymbol{\psi}_{-\theta} \sim \mathcal{N}\left(\frac{\tilde{\sigma}_\eta^2\sum_{t=1}^{T}\left(\gamma_t^{x_1}-\lambda\gamma_{t-1}^{x_1}-\boldsymbol{\varrho}_t^T[\tilde{\mathbf{F}}_t]_{x_1,\cdot}\right)+\tilde{\mu}_\eta\sigma_\gamma^2}{\tilde{\sigma}_\eta^2T+\sigma_\gamma^2},\frac{\tilde{\sigma}_\eta^2\sigma_\gamma^2}{\tilde{\sigma}_\eta^2T+\sigma_\gamma^2}\right),$$

$$\lambda|\boldsymbol{y},\boldsymbol{\varphi},\boldsymbol{\psi}_{-\lambda} \sim \mathcal{N}_{[-1,1]}\left(\frac{\tilde{\sigma}_\lambda^2\sum_{t=1}^{T}((\gamma_t^{x_1}-\eta-\boldsymbol{\varrho}_t^T[\tilde{\mathbf{F}}_t]_{x_1,\cdot})\gamma_{t-1}^{x_1})+\tilde{\mu}_\lambda\sigma_\gamma^2}{\tilde{\sigma}_\lambda^2\sum_{t=1}^{T}(\gamma_{t-1}^{x_1})^2+\sigma_\gamma^2},\frac{\tilde{\sigma}_\lambda^2\sigma_\gamma^2}{\tilde{\sigma}_\lambda^2\sum_{t=1}^{T}(\gamma_{t-1}^{x_1})^2+\sigma_\gamma^2}\right),$$

$$\sigma_\gamma^2|\boldsymbol{y},\boldsymbol{\varphi},\boldsymbol{\psi}_{-\sigma_\gamma^2} \sim \text{IG}\left(\tilde{a}_\gamma+\frac{T}{2},\tilde{b}_\gamma+\frac{1}{2}\sum_{t=1}^{T}\left(\gamma_t^{x_1}-\lambda\gamma_{t-1}^{x_1}\right)^2\right).$$

12.6.5. Application of the Constraints

The constraints are applied for every iteration of the sampler. Let the current iteration be denoted by i, then the following procedures are performed:

1. The constraints of the vector $\boldsymbol{\beta}^{(i)}$ are applied after sampling the arbitrary vector of static parameters $\boldsymbol{\beta}^{(i)}$, the vector is mapped into a vector of transformed parameters, $\tilde{\boldsymbol{\beta}}^{(i)}$ by the following rescaling $\tilde{\boldsymbol{\beta}}^{(i)}=\frac{\boldsymbol{\beta}^{(i)}}{\sum_x\beta_x^{(i)}}$. Then we replace $\boldsymbol{\beta}^{(i)}$ with $\tilde{\boldsymbol{\beta}}^{(i)}$ and proceed to the next steps of the sampler.
2. The constraints for the latent processes $\boldsymbol{\kappa}_t^{(i)}$ and $\boldsymbol{\gamma}_{t-x}$ are applied after the finalisation of Forward Backward. The arbitrary filtered estimates of the processes are transformed to $\tilde{\kappa}_t^{(i)}=\kappa_t^{(i)}-\bar{\kappa}^{(i)}$ and $\tilde{\gamma}_{t-x}^{(i)}=\gamma_{t-x}^{(i)}-\bar{\gamma}^{(i)}$ for $\bar{\kappa}^{(i)}=\frac{1}{N}\sum_{k=1}^{N}\kappa_{t_k}^{(i)}$ and $\bar{\gamma}^{(i)}=\frac{1}{N+p-1)}\sum_{c=t_1-x_p}^{t_N-x_1}\gamma_c^{(i)}$. Then we replace $\kappa_t^{(i)}$ with $\tilde{\kappa}_t^{(i)}$ and $\boldsymbol{\gamma}_t^{(i)}$ with $\tilde{\boldsymbol{\gamma}}_t^{(i)}$ and and proceed to the next steps of the sampler.

If one models the full cohort model the constraints of the vector $\boldsymbol{\beta}^\gamma$ are applied applied in the same fashion as for the vector $\boldsymbol{\beta}$. For a simplified cohort model, the vector of parameters is set to the vector of ones and is not sampled.

12.7. Stochastic Mortality Models for UK utilizing Factor Extraction from European Demographic Data

In this section we demonstrate the results of incorporating features extracted from European demographic data in the stochastic mortality model for British female mortality data over a study period of 1922 to 2014 with 10 years ahead forecasting validation.

The key findings are that the utilisation of demographic features improves the in-sample and out-of-sample predictive posterior mean Bayesian point estimation and forecasts for the log death rates. Additionally, the employed robustification methodology reduces the variances of the error terms in both observation and state equations and produces a better out-of-sample fit than its non-robust alternative. It indicates that the features extracted in the robust manner are more consistent over time and capture better the information about the true distribution of the demographic data.

The model which has the smallest mean square error of estimation and prediction adds the age-specific components to the latent process. It is later referred as DFM-PC-B. However, the other examined models, which also improve the predictability of the log death rates, are useful in terms of the interpretation as they reveal the individual country specific impacts of each of the European countries data on British female log death rates.

12.7.1. Description of the Models

Before presenting the real data example we note that all the Bayesian models developed and Markov chain samplers constructed were first tested on synthetic case studies in which the true parameters and state variables are known. The performance was found to be very good and this provides confidence in the accuracy and performance behaviour of the methods and models developed. The synthetic study results are provided in technical appendix and are not discussed in this paper. The use of synthetic data enables us to validate the estimation by Forward-Backward Kalman Filter with Gibbs Sampler. The models that we considered in our simulations and empirical studies are labelled by

LCC: Lee-Carter model with the stochastic cohort effect given by Equations (12.2) and (12.3);

DFM-PC: Demographic factor model which incorporates ϱ_t into LCC given by Equation (12.8). Please refer to Section C.2 in Appendix C for illustration how the models from this class are created;

- **DFM-PC-B:** The mean of first principal component of Birth counts as a static parameter, age specific element of ϱ_t;
- **DFM-PC-D-r:** The first principal component of Death counts (which is age and country specific) as an exogenous factor, one element of ϱ_t corresponds to a country specific subvector of the component, robust standardisation;
- **DFM-PC-D-s:** The first principal component of Death counts (which is age and country specific) as an exogenous factor, one element of ϱ_t corresponds to a country specific subvector of the component, non-robust standardisation;
- **DFM-PC-Mx-r:** The first principal component of Death Rates (which is age and country specific) as an exogenous factor, one element of ϱ_t corresponds to a country specific subvector of the component, robust standardisation;
- **DFM-PC-Mx-s:** The first principal component of Death Rates (which is age and country specific) as an exogenous factor, one element of ϱ_t corresponds to a country specific subvector of the component, non-robust standardisation;

The models of the class DFM-PC address **Case 1** from Section 12.3, where factors are incorporated into the observation Equation (12.2). The factors are obtained by performing PPCA jointly on the set of data for all countries listed in Table 12.1 excluding the following specific countries: United Kingdom (as it is our response variable), Greece and Slovakia (due to short time series).

DFM-PC-B incorporates the mean of the first principal component of the Birth counts which is a country specific vector. The matrix $\tilde{\mathbf{F}}$ is a 21×21 diagonal matrix with the mean on the diagonal. Hence, $\boldsymbol{\varrho}_t$ which correspond to the model DFM-PC-B, is a 21 dimensional, age specific state process and attempts to capture an age-specific dynamic in addition to the cohort-period effects.

The models DFM-PC-D-r, DFM-PC-D-s and DFM-PC-Mx-r and DFM-PC-Mx-s incorporate the first component of Death counts and Death Rates, respectively. Recall that the components for these data sets can be presented as age specific and country specific matrices as shown in Section 12.5. Due to the high dimensionality of the problem, we want the one element of $\boldsymbol{\varrho}_t$ to correspond to the subvector of the first component which is specific only for one country. Such a subvector has 21 dimensions which correspond to the age groups. Hence, $\boldsymbol{\varrho}_t$ is a 28 dimensional country specific state process. The country specific subvectors of the first components are placed in the columns of the 21×28 matrix $\tilde{\mathbf{F}}$. The last letter of the name of the models DFM-PC-D and DFM-PC-Mx denotes the type of the standardisation which is applied to the data before performing PPCA: robust (by M - estimator) or non-robust (sample estimator).

In the following part we analyse the population mortality from United Kingdom based on the models listed above and Bayesian methodology studied in this paper. We then examine the models in terms of the forecasting properties of death rates.

12.7.2. Setup

For the Bayesian estimation of models, we assume the priors given in the Appendix 12.6.3 and 12.6.4 to be

$$\begin{aligned}
&\kappa_0 \sim \mathcal{N}(0, 10^2), \quad , \gamma_0 \sim \mathcal{N}(0, 10^2), \quad \alpha_x \sim \mathcal{N}(0, 10^2), \quad \beta_x \sim \mathcal{N}(0, 10^2), \\
&\sigma_\varepsilon^2 \sim \mathrm{IG}(2.01, 0.01), \quad \theta \sim \mathcal{N}(0, 10^2), \eta \sim \mathcal{N}(0, 10^2), \quad \lambda \sim \mathcal{N}_{[-1,1]}(0, 10^2), \\
&\sigma_\kappa^2 \sim \mathrm{IG}(2.01, 0.01), \quad \sigma_\gamma^2 \sim \mathrm{IG}(2.01, 0.01), \quad \varrho_0^i \sim \mathcal{N}(0, 10^2,) \\
&[\boldsymbol{\Omega}]_{i,j} \sim \mathcal{N}(0, 10^2), \quad \Psi_j \sim \mathcal{N}(0, 10^2), \quad \sigma_\varrho^2 \sim \mathrm{IG}(2.01, 0.01).
\end{aligned}$$

The number of iterations of the Markov chain is 50,000 for LCC model and 200,000 for other models with 90% burn-in. The chain is initialised at $\boldsymbol{\alpha} = \bar{\mathbf{y}}_{1:T}$, $\beta_x = \frac{1}{21}$, $\sigma_\epsilon^2 = 0.0005$, $\theta = -0.005$, $\eta = -0.02$, $\sigma_\kappa^2 = 0.01$, $\sigma_\gamma^2 = 0.0005$, $\sigma_\varrho^2 = 1.0$, $[\boldsymbol{\Omega}]_{i,i} = \frac{1}{m}$ for m being either number of countries or number of age groups (depending on the model). The convergence of the sampler has been tested on synthetic data studies. The synthetic data study revealed that the estimation of the drift parameters corresponding to the factor state process model $\boldsymbol{\varrho}_t$ converges very slowly for shorter time series such as those found in mortality data. Thus, we decided to set these parameters to zero and do not sample them in this study.

12.7.3. Estimation of Static Parameters

Estimated values of the static parameters (except $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\Omega}$) for the British female mortality data (1922–2003) are shown in Table 12.2. The rest of the estimated static parameters is displayed in Figures 12.23, 12.24 and 12.25. The results are shown under different models listed in the first columns of the table or indicated by the colour of lines on the plots.

The static parameters of the factor process under DFM-PC-B model are age specific. In addition to the cohort and period effect, they provide supplementary information related to the corresponding age groups. Figure 12.23 shows the estimated diagonal elements of the transition matrix $\boldsymbol{\Omega}$ under the model. The parameters with values close to unity indicate that the factor state process corresponding to these parameters have a slowly decreasing dynamic. The elements of the state processes which correspond to the values of parameters closer to zero, are characterized by higher decrease. The parameters which are negative and close to zero indicate that the corresponding latent state processes fluctuate around zero.

With regards to DFM-PC-B model which incorporates age specific latent processes (supplementary to cohort effect), the elements of $\boldsymbol{\Omega}$ which are positive and close to zero, describe the decreasing dynamic of the corresponding age specific processes. These latent processes are shown to have more significant impact on death rate modelling when the sample starts, however, this impact decreases over time and the cohort effect becomes sufficient to model log death rates in these age groups. For instance, recall the age groups between 70–80 in Figure 12.23. Such process can be interpreted as a period effects which are specific for particular age groups. The elements of $\boldsymbol{\Omega}$ which are close to unity describe the latent process which have consistent impact or its lack over the time. If they have an effect on log death rates (i.e., their domain is not close to zero), they covey age-specific information which is supplementary to cohort and period effect and consistently demanded by the model over the time.

With regards to $\boldsymbol{\Omega}$ estimated under models DFM-PC-D-r, DFM-PC-D-s and DFM-PC-Mx-r and DFM-PC-Mx-s, it refers to the country specific features. Here the parameters are related to the influences of the specific countries on British female log death rates. The plots with the estimates and their confidence intervals are displayed in Figure 12.25. For instance, estimates of $\boldsymbol{\Omega}$ under all models agrees on lack of effects of Austrian or Bulgarian demographic data on the log death rates over whole sample span. On the other hand, the estimate of the parameter corresponding to Belarusian eigenvector is close to unity under all models and therefore highlights the informative effect of the feature on British log mortality rates which is consistent over the times.

It is worth to point out that the values of estimated variances of the observation and state equations error terms are higher for DFM-PC-D-s and DFM-PC-Mx-s where the data has been non-robustly standardized. These models are examined to have a greater mean square error of in-sample and out-of-sample fit than their robust alternatives as shown in Table 12.4. Hence, the robustification procedure employed in this study improved the overall goodness of fit of the considered models. The features which have been extracted from European demographic data by means of robust estimators of mean and covariance are shown to provide the information which

is consistent over the times and conveys the better knowledge about the true distribution of the demographic data sets.

Table 12.2. Bayesian posterior mean estimators with 95% posterior credible intervals for the estimation of the static parameters $\left(\lambda,\ \theta,\ \eta, \sigma_\epsilon^2,\ \sigma_\gamma^2,\ \sigma_\kappa^2, \sigma_\varrho^2\right)$ of $\log m_{x,t}$.

Model	λ	θ	η	σ_ϵ^2
LCC	0.998 (0.994; 1)	−0.154 (−0.331; 0.026)	−0.024 (−0.034; −0.014)	6.4×10^{-3} $(6\times10^{-3}; 6.9\times10^{-3})$
DFM-PC-B	0.991 (0.968; 1)	−0.332 (−0.53; −0.137)	−0.005 (−0.01; 0.002)	8×10^{-4} $(6\times10^{-4}; 9\times10^{-4})$
DFM-PC-D-r	0.949 (0.913; 0.992)	−0.246 (−0.415; −0.101)	0.011 (−0.002; 0.021)	1×10^{-3} $(9\times10^{-4}; 1.1\times10^{-3})$
DFM-PC-D-s	0.998 (0.993; 1)	−0.093 (−0.221; 0.029)	−0.019 (−0.025; −0.013)	1.3×10^{-4} $(1.1\times10^{-4}; 1.4\times10^{-4})$
DFM-PC-Mx-r	0.985 (0.959; 0.999)	−0.042 (−0.115; 0)	−0.013 (−0.02; −0.007)	8×10^{-4} $(7\times10^{-4}; 1\times10^{-3})$
DFM-PC-Mx-s	0.999 (0.995; 1)	−0.024 (−0.111; 0.044)	−0.02 (−0.03; −0.01)	1.2×10^{-4} $(8\times10^{-4}; 2.1\times10^{-4})$
textbfModel	σ_γ^2	σ_κ^2	σ_ϱ^2	
LCC	2×10^{-3} $(6\times10^{-3}; 6.9\times10^{-3})$	0.663 (0.449; 0.96) $(1.4\times10^{-3}; 2.8\times10^{-3})$		
DFM-PC-B	5×10^{-4} $(4\times10^{-4}; 7\times10^{-4})$	0.753 (0.537; 1.055)	0.049 (0.042; 0.057)	
DFM-PC-D-r	5×10^{-4} $(4\times10^{-4}; 8\times10^{-4})$	0.39 (0.227; 0.739)	0.092 (0.074; 0.113)	
DFM-PC-D-s	8×10^{-4} $(5\times10^{-4}; 1.1\times10^{-3})$	0.324 (0.152; 0.616)	0.144 (0.114; 0.178)	
DFM-PC-Mx-r	6×10^{-4} $(4\times10^{-4}; 8\times10^{-4})$	0.044 (0.002; 0.116)	0.08 (0.066; 0.094)	
DFM-PC-Mx-s	1.7×10^{-3} $(1.2\times10^{-3}; 2.3\times10^{-3})$	0.036 (0.001; 0.137)	0.834 (0.594; 0.994)	

We did not choose to calculate the MLE estimates of the parameter's in our models as it has been documented that even for the standard period-cohort type Lee-Carter stochastic mortality models, the classical MLE estimation frameworks can produce convergence and estimation challenges due to gradient based and method of scoring recursive optimization methods getting stuck in local optima of the marginal likelihood surface. We refer the interested reader to the paper (Fung et al. 2018) where we discuss such issues in more depth. Therefore, instead of resolving the known problems that may arise with classical MLE estimations of such models, which may be further compounded in the extended models we developed in the frequentist setting, we have chosen to stick with the Bayesian modelling paradigm and to report an analogue result to the MLE that may be obtained from Bayesian inference, in the case of uninformative priors. That is we have relative uninformative priors and so we can report the Maximum a-postiori (MAP) posterior mode estimator for the parameters as the Bayesian analogue of the MLE, defined as

$$\boldsymbol{\psi}^{MAP} = \underset{\boldsymbol{\psi}}{\operatorname{argmax}}\, \pi_{\boldsymbol{\varrho}_{0:T},\boldsymbol{\Psi}|\mathbf{Y}_{1:T}}(\boldsymbol{\varrho}_{0:T}.\boldsymbol{\psi}) \tag{12.20}$$

The MAP estimates and MLE should be similar in the case of uninformative priors, with the advantage that the MAP estimation is obtained via an MCMC sampler output, which is less prone to the types of estimation challenges experienced in gradient descent methods working directly with the Instead, as we used fairly uninformative priors, note that the MAP estimate of the posterior is a case of uninformative priors will correspond to MLE estimates. Please refer to Table 12.3 for the analogous of point estimates to the outcomes of Table 12.2.

Table 12.3. The MAP estimates of the static parameters $\left(\lambda,\ \theta,\ \eta, \sigma_\epsilon^2,\ \sigma_\gamma^2,\ \sigma_\kappa^2, \sigma_\varrho^2\right)$ of $\log m_{x,t}$.

Model	λ	θ	η	σ_ϵ^2	σ_γ^2	σ_κ^2	σ_ϱ^2
LCC	0.999	−0.155	−0.024	0.0064	0.0019	0.6172	–
DFM-PC-B	0.998	−0.331	−0.005	8.00×10^{-4}	5.00×10^{-4}	0.7279	0.0487
DFM-PC-D-r	0.948	−0.253	0.012	0.001	5.00×10^{-4}	0.3234	0.0895
DFM-PC-D-s	0.999	−0.095	−0.02	0.0013	7.00×10^{-4}	0.3005	0.1416
DFM-PC-Mx-r	0.995	−0.023	−0.012	7.00×10^{-4}	6.00×10^{-4}	0.029	0.0818
DFM-PC-Mx-s	1	−0.023	−0.021	9.00×10^{-4}	0.0016	0.0088	0.8583

12.7.4. Filtering of Latent Variables

The Bayesian posterior mean estimates of the latent stochastic mortality factors in the models for κ_t in the top panel and for γ_t^0 in the bottom panel of Figure 12.26. The colours of lines denote the filtered processes under different models. As expected, adding new state variables related to the factors significantly changes the dynamics of the period and cohort effect state processes. The blue line correspond to the cohort=period only LCC model. The increase of κ_t and decrease of γ_t^0 at the end of the sample is greater for this model in contrast to the the other examined models.

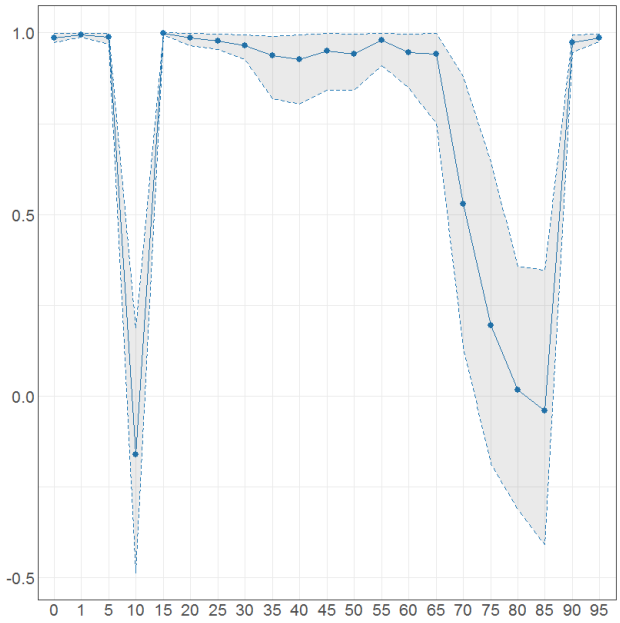


Figure 12.23. Bayesian posterior mean estimators with 95% posterior credible intervals for the estimation of the age-specific diagonal elements of the transition matrix $\mathbf{\Omega}$ (x axis) under DFM-PC-B.

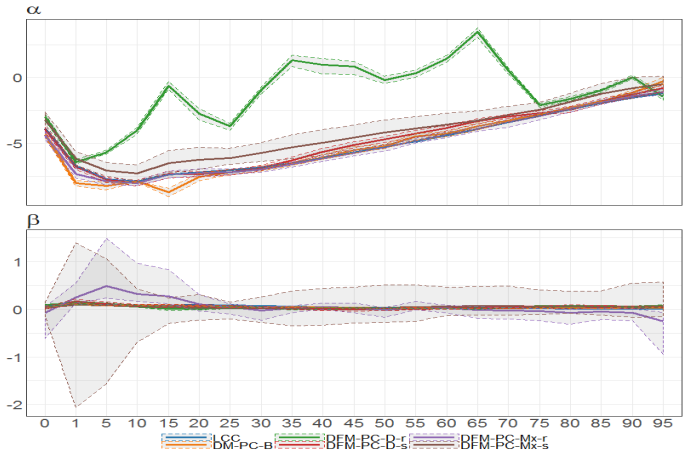


Figure 12.24. Bayesian posterior estimators with 95% posterior credible intervals for the estimation of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ under different models (colours of lines) for British female mortality data (1922–2002).

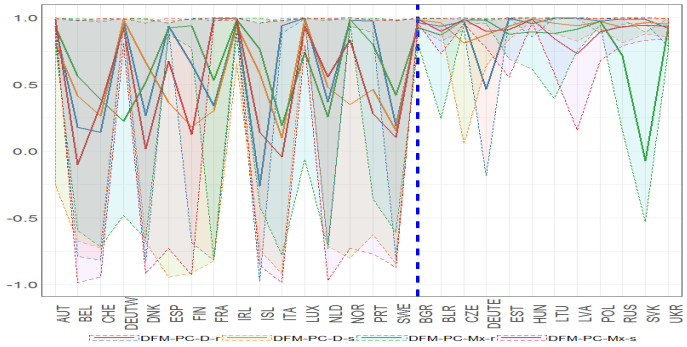


Figure 12.25. Bayesian posterior mean estimators with 95% posterior credible intervals for the estimation of the diagonal elements of the transition matrix $\boldsymbol{\Omega}$ (x axis) under DFM-PC-D-r, DFM-PC-D-s, DFM-PC-Mx-r and DFM-PC-Mx-s models (colours of lines). The dashed blue line divides the set of countries into developed (on the left side) and developing (on the right hand side) European countries, respectively.

The dynamic of the cohort effect latent process vector γ_t over time, which is age group specific, is shown in Figure 12.28. The panels correspond to the Bayesian posterior mean estimates of the process under different models. The age group specific features, which has been utilised in DFM-PC-B model, clearly provide LCC model with supplementary information to the cohort effect state processes. The corresponding $\boldsymbol{\varrho}_t$ reduces the variability of the cohort effect process with comparison to the cohort effect estimated under LCC model (the colours of the surface on DFM-PC-B panel are plain and variance of the error term is smaller). The state processes corresponding to the factors under DFM-PC-B model are shown in Figure 12.27. Recall that in the contrast to the cohort and period effects processes, the latent process vector $\boldsymbol{\varrho}_t$ under the model DFM-PC-B has age group specific stochastic components (κ_t is calendar year specific and γ_t is '0' age group and calendar year specific). As κ_t latent process estimated under DFM-PC-B is not distant from the corresponding estimated under LCC model for majority of the sample, we can conclude that the additional state processes given by this model provide supplementary, age-specific information to the cohort effect process γ_t^0.

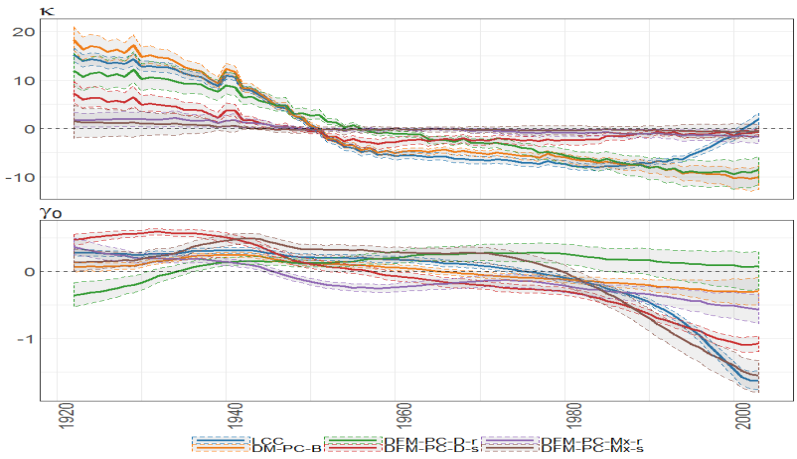


Figure 12.26. The Bayesian posterior mean estimates with 95% posterior credible intervals for κ_t (upper panel) and cohort effect state process γ_t^0 (lower panel) under different models (colours of lines) for British female log death rates during 1922–2002.

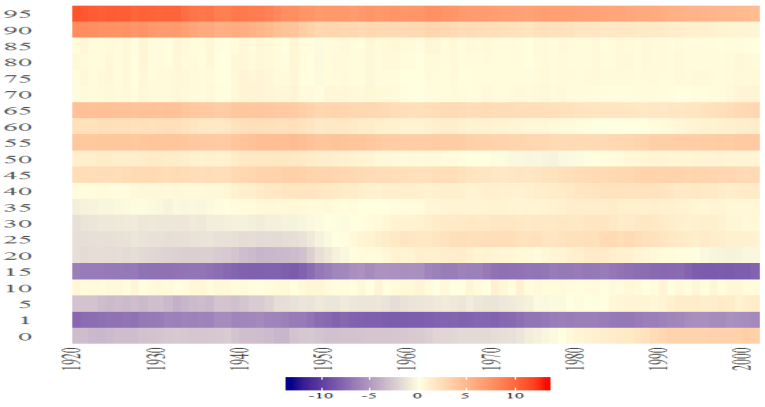


Figure 12.27. The Bayesian posterior mean estimates for ϱ_t across age groups (y axis) over time (x axis) under DFM-PC-B model for British female log death rates during 1922–2002.

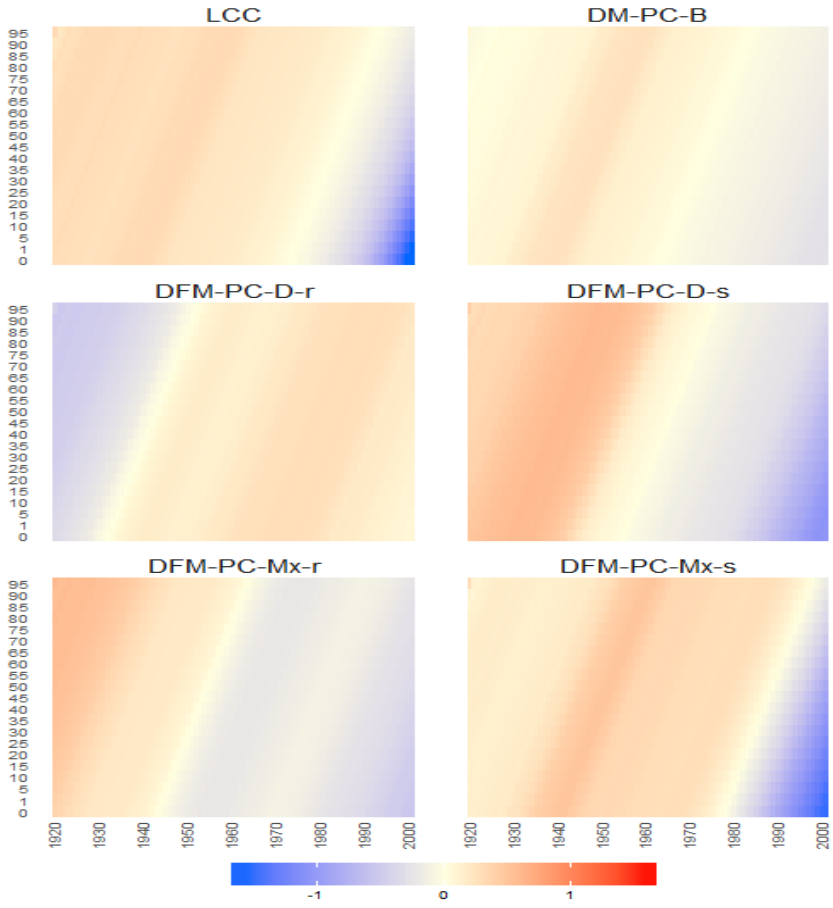


Figure 12.28. The Bayesian posterior mean estimates for the cohort effect latent processes vector γ_t across age groups (y axis) over time (x axis) under different models for British female log death rates during 1922–2002.

The models DFM-PC-D-r, DFM-PC-D-s, DFM-PC-Mx-r and DFM-PC-Mx-s are characterized by the processes which correspond to the country specific vectors with elements related to age groups. The models attempt to address the question whether the structure of demographic data from European countries can be an efficient explanatory variable for British mortality. Figures 12.26 and 12.28 show that dynamics of period and cohort effect state processes are sensitive to different sets of features extracted from demographic data as well as a standardisation methodology. Figure 12.29 and 12.27 provide insight into how these variables change the information extracted from the models. The information highlight the impact of the first components of the considered countries on British mortality rates. As the components are orthonormal, $\varrho_t^{country}$ indicates the magnitude of this effect. To be consistent with the country set specific notation from Section 12.5, the blue vertical line divides the results into two categories: results for the developed countries (below the blue line) and developing countries (above the blue line). The processes which moves closely to zero give information that the data of country they correspond to has small influence on the mortality rates from United Kingdom.

The developed models when incorporating the demographic features from the European countries, indicate the significance of the factor loading from a given country on the mortality of UK. Hence, we can specify the countries which has a positive effect on the mortality of United

Kingdom (factor state processes are negative), neutral (factor state processes are fluctuate around zero) or negative (factor state processes are positive and enlarge log death rates).

For a given model it is not true that all factors which correspond to the age specific vectors of features from European countries, have an influence in the causal fashion on the UK morality data in the same way. As such, what we are showing in the plots Figure 12.25 (or Figure 12.23 for DFM-PC-B models) that some countries have very wide posterior credible intervals (the flat posterior) for $\mathbf{\Omega}$ which is in an alignment with the findings in Figure 12.30 (Figure 12.27 for DFM-PC-B model) where we see that indeed for those countries the $\boldsymbol{\varrho}_t$ upon a model indicate insignificant effect on UK log mortality rate. Let us consider the example of DFM-PC-Mx-s (the red colours of lines on both of the plots) for Belgium (BEL) and Austria (AUT). We see in Figure 12.30, that the effect of Austrian factor to the British mortality data, expressed by the dynamic of ϱ_{AUT}, is non-zero over the time. On the other has, the element of $\boldsymbol{\varrho}_t$ corresponding to Belgium, labelled by ϱ_{BEL} does not load significantly in any way on the UK mortality experience. As a consequence we see that the posterior for this country on $\mathbf{\Omega}$ is also very flat when the credit intervals of Ω_{AUT} on Austria are significantly narrower. This simply means that the factor loading of Belgium does not influence the UK mortality experience.

Hence, to conclude this discussion, we note that when we look at the results in Figure 12.30, they show the effect of each individual countries influence on the UK mortality experience. In fact, what we learn is that some countries have a mean of 0 with large uncertainty, these countries maybe interpreted as not having an influence on the mortality experience of the UK.

The four models are more consistent about the set of countries which does not have any effect on the log death rates of United Kingdom Females. The models corresponding to the non-robust standardisation indicate bigger impact of western Europe countries whereas their robust alternatives indicate the significance of the patterns from Easter and Central Europe countries as Lithua, Poland or Russia.

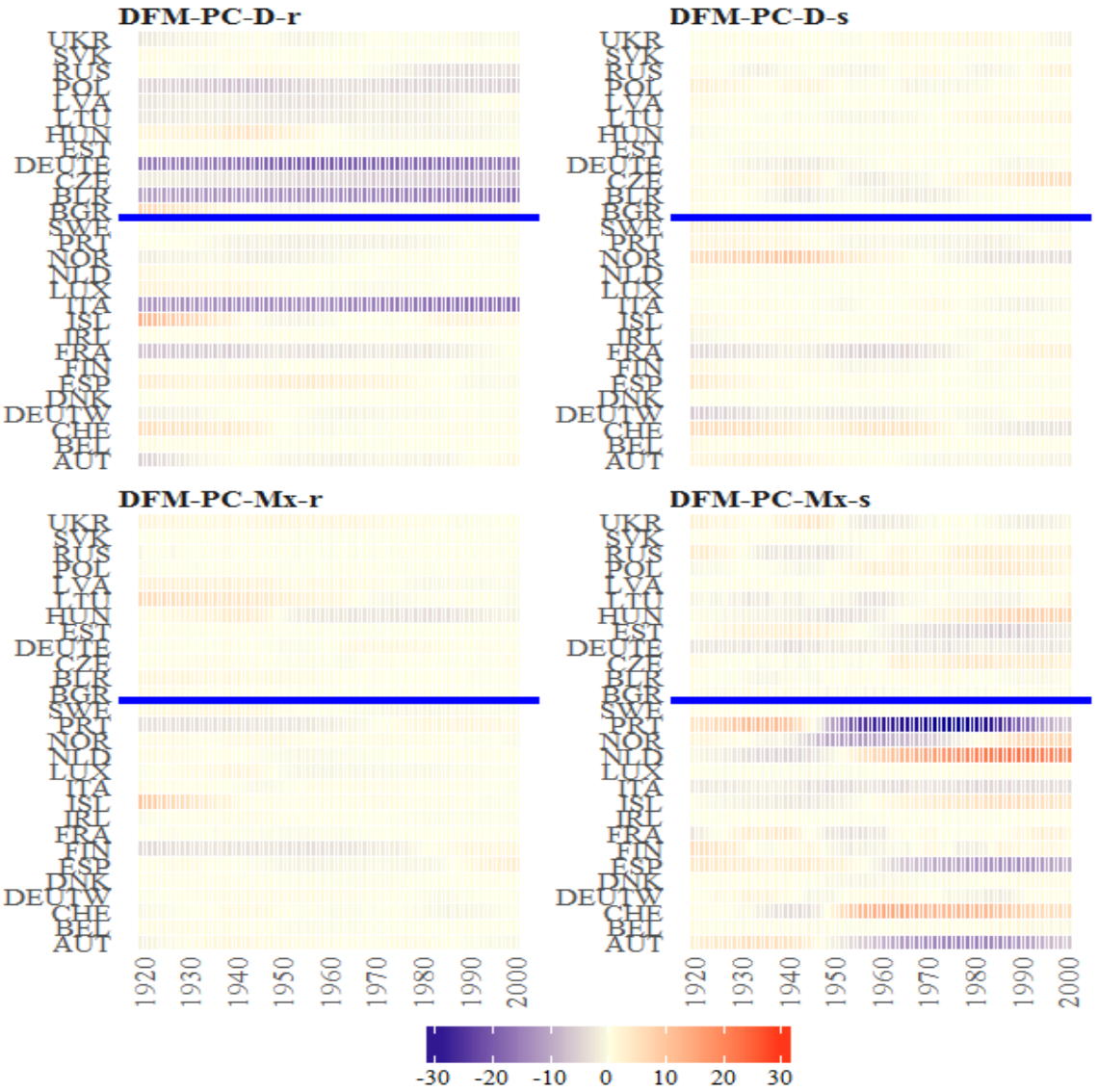


Figure 12.29. The Bayesian posterior mean estimates for ϱ_t across countries (y axis) over time (x axis) under the models from the classes DFM-PC-D and DFM-PC-Mx for British female log death rates during 1922–2002. The vertical blue line divides sets into developed (on the left side) and developing (on the right sie) European countries.

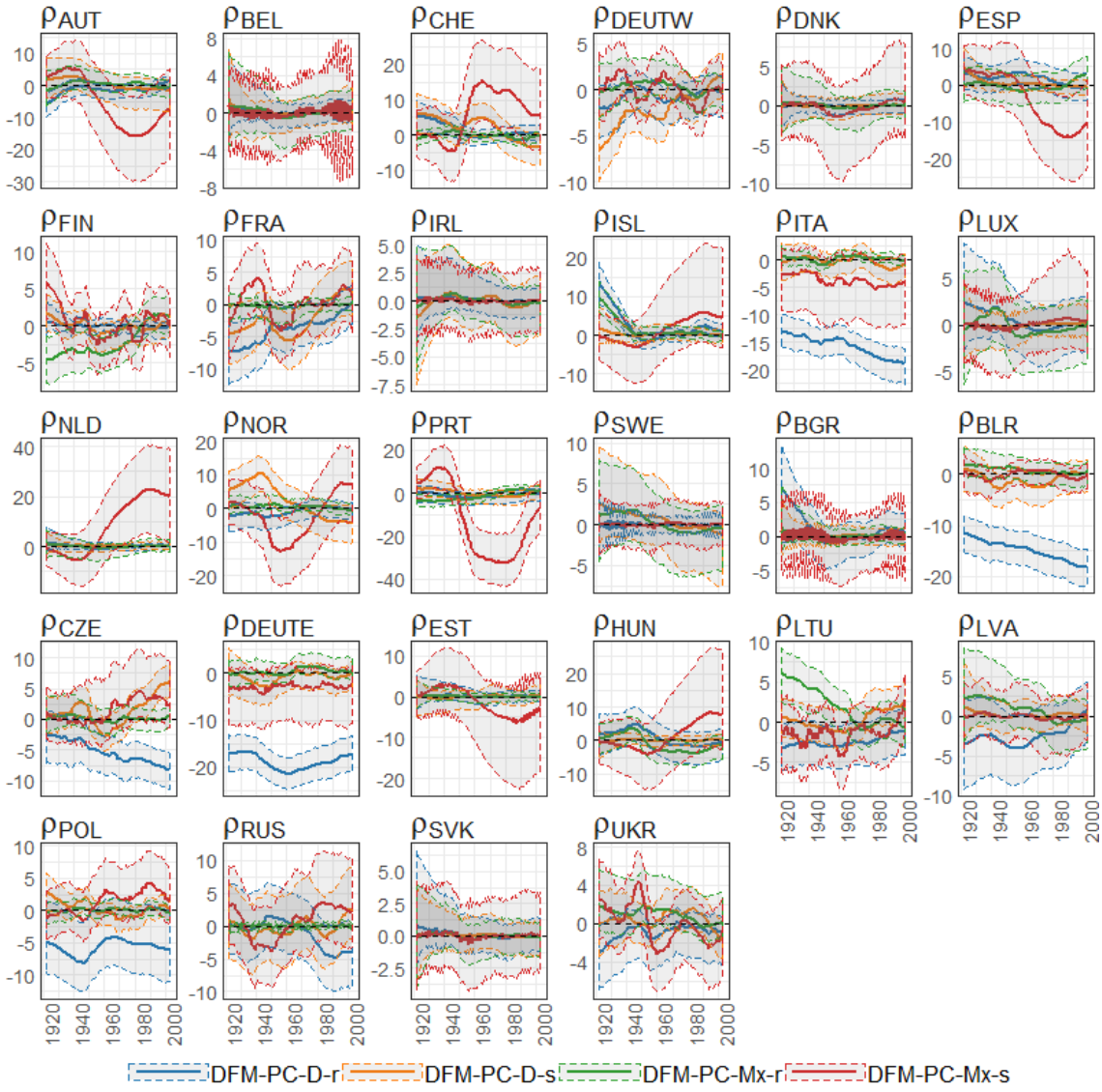


Figure 12.30. The Bayesian posterior mean estimates with 95% posterior credible intervals for ϱ_t across countries (different panels) overtime (x axis) under the models from the classes DFM-PC-D and DFM-PC-Mx (colours of lines) for British female log death rates during 1922–2002.

12.7.5. In-Sample and Out-of-Sample Performance

In this section, we investigate the in-sample and out-of-sample performance of the mortality models summarized in the introduction to this section. The model selection is based on two in-sample performance measures, MSE (Mean Square Error) and DIC (Deviance Information Criterion) as defined in Chapter 9, whereas the forecasting performance examined based on MSEP (Mean Square Error of Prediction) using two forecasting distributions of log death rates, the one obtained by Gibbs sampler and the one provided by Kalman Filter defined in Subsection 8.2.2.

The forecasting distribution given by FFBS algorithm, obtained by sampling recursively, when (i) denotes a sample,

$$\begin{aligned} \boldsymbol{\varphi}_{T+k}^{(i)} &\sim \mathcal{N}\left(\tilde{\boldsymbol{\Lambda}}^{(i)}\boldsymbol{\varphi}_{T+k-1}^{(i)} + \tilde{\boldsymbol{\Theta}}^{(i)}, \tilde{\boldsymbol{\Psi}}^{(i)}\right) \\ \mathbf{y}_{T+k}^{(i)} &\sim \mathcal{N}\left(\boldsymbol{\alpha}^{(i)} + \tilde{\mathbf{B}}_t^{(i)}\boldsymbol{\varphi}_{T+k-1}^{(i)} + \tilde{\boldsymbol{\Theta}}^{(i)}, \sigma_\epsilon^{2(i)}\mathbb{I}_d\right). \end{aligned} \tag{12.21}$$

The forecasting distribution given by the Kalman Filter is given by

$$\begin{aligned} \boldsymbol{\varphi}_{T+k} &\sim \mathcal{N}\left(\tilde{\boldsymbol{\Lambda}}\boldsymbol{\varphi}_{T+k-1} + \tilde{\boldsymbol{\Theta}}, \tilde{\boldsymbol{\Psi}}\right) \\ \mathbf{y}_{T+k} &\sim \mathcal{N}\left(\boldsymbol{\alpha} + \tilde{\mathbf{B}}_t\boldsymbol{\varphi}_{T+k-1} + \tilde{\boldsymbol{\Theta}}, \sigma_\epsilon^2\mathbb{I}_d\right). \end{aligned} \tag{12.22}$$

for the static parameters which has been estimated by averages within sampled realisation provided by the Gibbs sampler.

12.7.5.1. Comparison of the Models

We choose for out-of-sample study last 10 years of the available sample for British Female death rates. The calibration period is 1922–2002. The Table 12.4 summarizes the calculated mean squared errors of the estimated observations using Kalman Filter (MSE), deviance information criterion (DIC) and mean square errors of predictions using the MCMC distribution (MSEP_{MCMC}) and the Kalman Filter distribution (MSEP_{Kalman}). The results confirm that adding the features, which has been extracted from demographic data, as an additional explanatory variable to the LCC model improves both in-sample fit out-of-sample fit and therefore the predictability of log death rates. The plots with age group specific prediction results can be found in Figure 12.35.

Table 12.4. Mean square error of the fit of the models to the data (MSE), deviance information criterion (DIC) and mean square errors of predictions using forecasting distributions given by MCMC samples (MSEP_{MCMC}) and Kalman Filter (MSEP_{Kalman}).

Model	MSE	DIC	MSEP_{MCMC}	MSEP_{Kalman}
LCC	0.0097	−3627	0.1778	0.1774
DFM-PC-B	0.0072	−6500	0.0057	0.0062
DFM-PC-D-r	0.0182	−6380	0.0177	0.0251
DFM-PC-D-s	0.0065	−5996	0.0185	0.0156
DFM-PC-Mx-r	0.0081	−8225	0.0111	0.0129
DFM-PC-Mx-s	0.0174	−3951	0.0692	0.0285

For the in-sample performance, the MSE and DIC agree to the group of two best performing models, DFM-PC-B and DFM-PC-Mx-r, however are conflicted with regards to the group of two worse performing models. Due to assessing the performance of the model considering its complexity, DIC more successfully captures the models which result in poorest out-of-sample performance, LCC and DFM-PC-Ms respectively. Especially it is worth to notice significant over-fitting of the LCC model which is further investigated in Section 12.8. Recall, that using MSE as model selection criterion would not be sufficient to choose the model with good performance. In terms of MSE, the in-sample performance of LCC model is comparable to DFM-PC-B model, while DIC labels the model as one with the worse explanatory power.

12.8. Additional Remarks on Modelling and Forecasting Results

While conducting the study we encountered two issues which are worth separate discussion, the influence of the stratification on the class of Lee-Carter models and the intuition behind the vector $\boldsymbol{\alpha}$ in the model Equation (12.8) when we incorporate the demographic features.

12.8.1. The Affect of the Stratification and Identification Constraints on Estimation of Stochastic Lee-Carter Type Models

We draw the readers attention at this point to the substantially lower predictive accuracy of the Lee-Carter cohort (LCC) model in comparison to the models which employ demographic factors, as shown in Table 12.4. This is especially important since such LCC models, without factors have been previously documented to have better out-of-sample performance for UK data when no age stratification is applied.

We explain the steps we have taken to explore this feature that the reviewer has pointed out. Firstly, we clarify this is not a problem with the sampler or the prior specification, rather it is related to the particular suitability of different model structures under particular assumptions in the model. We explore this, specifically with respect to age group stratification and its influence on the model fit and performance.

Please note that the stratification was adopted, where we looked at 20 sets of age groups in 5 year buckets to reduce the dimensionality of the model, of course this can influence the model fit and the assumptions made regarding model simplification and cohort interpretation. Therefore, we decided to include additional studies to investigate these effects more carefully and in the process we believe we may also address the question raised by the reviewer on this point. Our attention has been drawn to two points: rapid increase of κ_t in 2000 and the substantially lower predictive accuracy of Lee-Carter cohort (LCC) model in comparison to the models which employ demographic factors as shown in Table 12.4. Since the LCC model has been documented to has better out-of-sample explanatory power for the UK data when there is no stratification, we decided to undertake an additional investigation. In particular, in this new class of studies, we explore in more detail the effect of age stratification and the appropriate choice of adjustment of model assumptions and identification constraints to be performed in order to compare models in a meaningful manner. The details are described below and in the manuscript in Section 10. The notation of the models which we decided to analyse are the following

LC: Lee-Carter model

$$\log m_{x,t} = \alpha_x + \beta_x \kappa_t,$$

with the constraints

$$\sum_x \beta_x = 1, \sum_t \kappa_t = 0,$$

LCC: Simplified Lee-Carter cohort model

$$\log m_{x,t} = \alpha_x + \beta_x \kappa_t + \gamma_{t-x},$$

with the constraints

$$\sum_x \beta_x = 1\,, \sum_t \kappa_t = 0, \sum_{c=t_1-x_p}^{t_N-x_1} \gamma_c = 0.$$

LCCF: Lee-Carter full cohort model

$$\log m_{x,t} = \alpha_x + \beta_x \kappa_t + \beta_x^{\gamma} \gamma_{t-x},$$

with the constraints

$$\sum_x \beta_x = 1, \sum_x \beta_x^{\gamma} = 1, \sum_t \kappa_t = 0, \sum_{c=t_1-x_p}^{t_N-x_1} \gamma_c = 0.$$

In this study we consider two age group stratifications, the 1×1 study which has 100 age groups per year and the more parsimonious class of models given by the 5×1 age group stratification with 21 age groups. Since the number of ages groups differs among stratified and non-stratified mortality data, that is, we have 21 age groups for the data in "5×1" format, we expect that the parameters and latent variables, which are estimated using above constraints, may differ in magnitude. Therefore, in order to ensure comparability of the results when we examine the stratification effect on the family of Lee-Carter models, one must therefore to standardize the magnitude of parameters and variables of models in order to compare between the stratified and non-stratified case. We demonstrate in the new studies that such a problem may be resolved via a simple scaling adjustment to the identification constraints in order to resolve this issue. Hence, we introduce the scaling parameter $a > 0$. The models with imposed adjustment are denoted by lower index adj as follows

LC$_{adj}$: Lee-Carter model

$$\log m_{x,t} = \alpha_x + \beta_x \kappa_t,$$

with the constraints

$$\sum_x \beta_x = \frac{1}{a}, \sum_t \kappa_t = 0,$$

LCC$_{adj}$: Simplified Lee-Carter cohort model

$$\log m_{x,t} = \alpha_x + \beta_x \kappa_t + \gamma_{t-x},$$

with the constraints

$$\sum_x \beta_x = \frac{1}{a}, \sum_t \kappa_t = 0, \sum_{c=t_1-x_p}^{t_N-x_1} \gamma_c = 0.$$

LCCF$_{adj}$: Lee-Carter full cohort model

$$\log m_{x,t} = \alpha_x + \beta_x \kappa_t + \beta_x^{\gamma} \gamma_{t-x},$$

with the constraints

$$\sum_x \beta_x = \frac{1}{a}, \sum_x \beta_x^{\gamma} = \frac{1}{a}, \sum_t \kappa_t = 0, \sum_{c=t_1-x_p}^{t_N-x_1} \gamma_c = 0.$$

Lastly, in order to distinguish between the results of the models for stratified and non-stratified data, we denote the models for stratified data with the lower index "5×1" and for non-stratified data with "1×1", for instance, the results for the Lee-Carter model LC for stratified data are denoted by $\text{LC}_{5\times1}$ and the results for the same model for a data without any stratification are denoted by $\text{LC}_{1\times1}$.

12.8.1.1. The Estimates of the Static Parameters and Filtered Latent Variables

The list of models, which have been examined and are discussed in this subsection, is given in the first column of Table 12.5. In addition to LC, LCC and LCCF models, we include into the comparison also the models with adjusted constraints for the mortality data in format "5×1" with the adjustment parameter $a = \frac{\text{\# age groups 1x1}}{\text{\# age groups 5x1}} = \frac{100}{21} \approx 4.762$ being a proportion between the number of age groups in in the 1×1 stratification and the 5×5 stratification of the age groups in the mortality data.

The Bayesian posterior estimators of the static parameters α_x and β across age groups are shown in Figure 12.31 when the Bayesian posterior mean estimates of period effect κ_t and cohort effect state process γ_t^0 are shown in Figure 12.32.

The first straightforward remark on the investigation is to note that there is an inconsequential influence of both the stratification and the adjustment to the identification constraint for stratification, when investigating the estimation of the level vector of the model, as denoted by parameter vector α.

Also, the basic Lee-Carter models is not affected by the stratification as both its in-sample and out-of-sample quality of fit is comparable among data with format "5×1" and "1×1". However, the estimates of $\boldsymbol{\beta}$ for $\text{LC}_{5\times1}$ are greater in the magnitude in comparison to $\text{LC}_{1\times1}$ model which appears to be an offset by a smaller slope of the filtered κ_t. Importantly, we note that when the adjustment to the number of age groups is imposed, the $\boldsymbol{\beta}$ and κ_t for $\text{LC}_{5\times1,adj}$ are in line to those of $\text{LC}_{1\times1}$ and the model still keeps comparable explanatory power. This observation gives us the intuition that the stratification influences mainly the cohort effect. This is something that we intuitively can understand due to the interplay between age stratification and cohort effect.

Table 12.5. Mean square error of the fit of the models to the data (MSE) and mean square errors of predictions using forecasting distributions given by MCMC samples (MSEP_{MCMC}) and Kalman Filter (MSEP_{Kalman}). The models highlated by bold font exhibit significant levels of over-fitting.

Model	MSE	MSEP_{MCMC}	MSEP_{Kalman}
$\text{LC}_{1\times1}$	0.0128	0.0577	0.0568
$\text{LC}_{5\times1}$	0.0113	0.0457	0.0457
$\text{LC}_{5\times1,adj}$	0.0116	0.0512	0.0516
$\text{LCC}_{1\times1}$	0.0079	0.0249	0.0243
$\text{LCC}_{5\times1}$	**0.0097**	**0.1778**	**0.1774**
$\text{LCC}_{5\times1,adj}$	0.0099	0.1664	0.1625
$\text{LCCF}_{1\times1}$	**0.2588**	**0.4735**	**0.6150**
$\text{LCCF}_{5\times1}$	0.0107	0.0458	0.0464
$\text{LCCF}_{5\times1,adj}$	0.0131	0.0481	0.0508

The results in Table 12.5 shows that the out-of-sample quality of the fit for the $\text{LCC}_{5\times1}$ model is significantly lower than the corresponding result for $\text{LCC}_{1\times1}$. As the adjustment, that is $\text{LCC}_{5\times1,adj}$ model, produces similar outcomes, the discrepancy in the out-of-sample quality of fit between LCC model applied to "5×1" and "1×1" data is not caused by the smaller number of age groups

and therefore different magnitude of estimates of $\boldsymbol{\beta}$ and κ_t (recall Figures 12.31 and 12.32). Since the discrepancy of the in-sample explanatory power is smaller between the models, we observe that the LCC model tends to over-fit when applied to stratified data.

We begin discussion of these results by noting the following finding. There appears to be an interplay present between the model parsimony and the bias and variance in the results for both in-sample fits and out-of-sample forecasts, as reflected by the Mean Squared Error (MSE) results, which is more largely affected by the model structure rather than stratification effects.

For instance, we see that the more parsimonious model choices, corresponding to say the three LC sub-family of models always had a larger MSE than the less parsimonious class of simplified LCC model. That is the in-sample MSE improved by around an order of magnitude when we incorporated extra structure corresponding to the cohort feature. This was not influenced by the age stratification reformed. We conjecture that although the LC models will have potentially lower variance, due to less model parameters to be estimated, the in-sample MSE is still worse generally due to increased bias that may arise from not capturing sufficiently the stochastic structure of the data.

Furthermore, we also see a pronounced effect of stratification on the out-of-sample forecast performance of the simplified class of LCC models in which no adjustment was made for the stratification effect. This indicates that the adjustment we propose to use when undertaking age-group stratification can substantially reduce the bias in the resulting model estimates when we compare between the simplified $\text{LCC}_{5\times1}$ model and the adjusted form.

Thirdly, we observe that the most flexible class of LCC model, the non-simplified LCCF class of models was significantly affected by removing the age stratification of 5×1 compared to the 1×1 case. To understand this, we have significantly increased the dimension of the model parameters to be estimated in the $\text{LCCF}_{1\times1}$ compared to the $\text{LCCF}_{5\times1}$. This we believe produces a poor in sample and out-of-sample MSE and MSEP due to the resulting over-fitting and increased variance in the model estimates, compared to the simplified LCC model equivalents. However, importantly the stratification effect is significant here, the dimension reduction in model parameters in the $\text{LCCF}_{5\times1}$ compared to the $\text{LCCF}_{1\times1}$ reduces the variance in the estimates of the mortality in sample and out-of-sample as well as providing additional degrees of freedom to also reduce the bias that arises from the constrained version of the $\text{LCC}_{5\times1}$ model, resulting in the optimal MSE and MSEP performance.

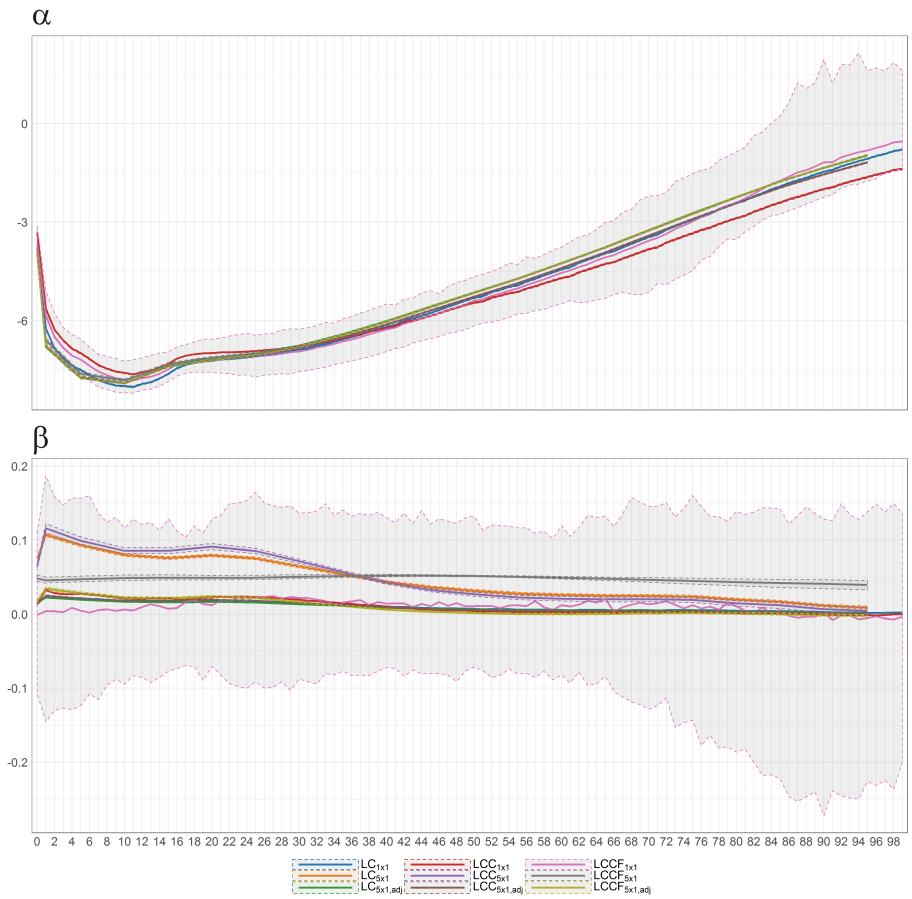


Figure 12.31. Bayesian posterior estimators with 95% posterior credible intervals for the estimation of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ under different models (colours of lines) for British female mortality data (1922–2002).

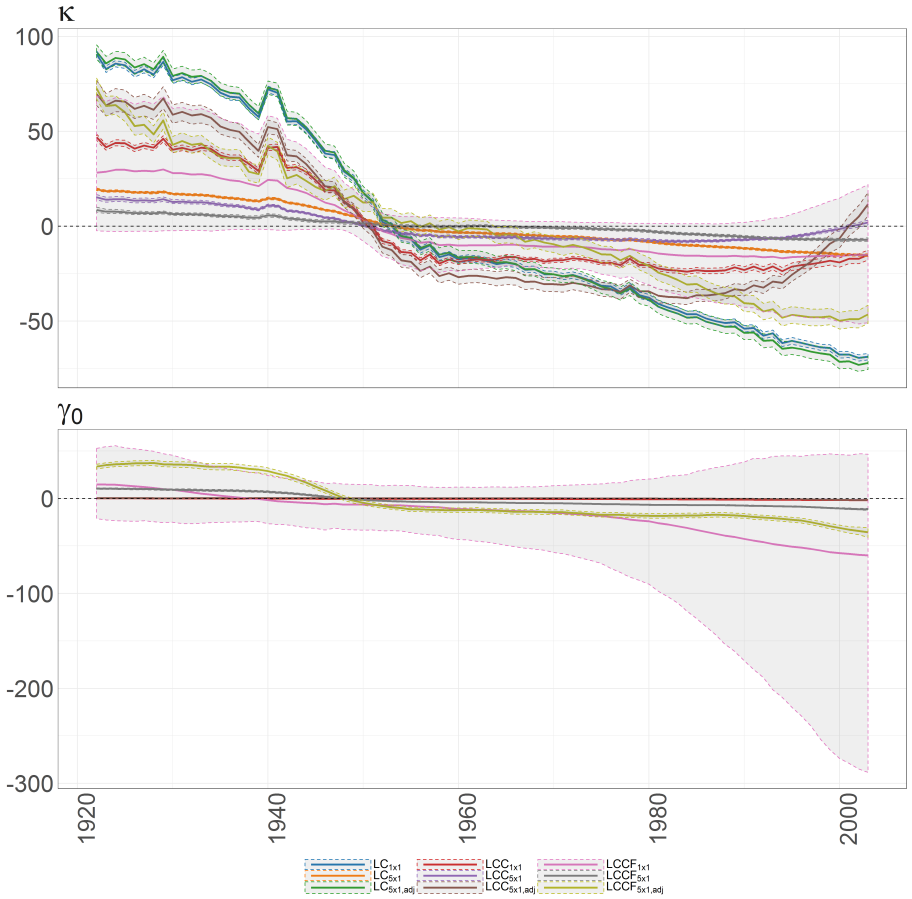


Figure 12.32. The Bayesian posterior mean estimates with 95% posterior credible intervals for κ_t (**upper panel**) and cohort effect state process γ_t^0 (**lower panel**) under different models (colours of lines) for British female log death rates during 1922–2002.

12.8.2. The Estimates of the Intercept for the Factor Model DFM-PC-D-r

The following study addresses the interpretation of the estimates of $\boldsymbol{\alpha}$ under the the new class of stochastic mortality factor model in comparison to the standard Lee-Carter models without the matrix with demographic factors. Please refer to Section 12.8.2. The argument we proposed to interpret $\boldsymbol{\alpha}$ is adjusted to the fact that we have the exogenous factors incorporated into our model compared to standard Lee-Carter model. As such, we argue that the the interpretation of the intercept should now incorporate both $\boldsymbol{\alpha}$, the classical intercept, and the term $\tilde{\mathbf{F}}_t \boldsymbol{\varrho}_t$ corresponding to the intercept which arises from the exogenous factors. In the time series context this is considered as a stochastic intercept. Hence the interpretation of $\boldsymbol{\alpha}$ typically adopted in the classical stochastic Lee-Carter type period-cohort models, does not hold under the new model, since we have now incorporated the additional structure corresponding to the regression term from the demographic factors. The expression $\tilde{\mathbf{F}}_t \boldsymbol{\varrho}_t$ which is added to the observation equation provides with time-varying supplementary information to the static level given by $\boldsymbol{\alpha}$. To validate this claim we have undertaken the additional studies which demonstrate when we combine the $\boldsymbol{\alpha}$ with this component of the model, the posterior mean of this quantity behaves in analogues fashion to what you would expect on the posterior mean of $\boldsymbol{\alpha}$ in the standard, non-factor class of Lee-Carter models. This is

interesting as it shows the factor influence and additional interpretation to the level contributed by the long-term demographic exogenous factors. The plot in Figures 12.33 shows the Bayesian posterior mean estimates with 95% creditable intervals of $\boldsymbol{\alpha} + \tilde{\mathbf{F}}_t \boldsymbol{\varrho}_t^T$ averaged over time, whereas Figure 12.34 illustrates the posteriori mean over time. The level of the expression on either of plots is below zero and behaves in a fashion we would have expected from $\boldsymbol{\alpha}$ in standard Lee-Carter models. It confirms our interpretation as well as answers the question asked by the reviewer.

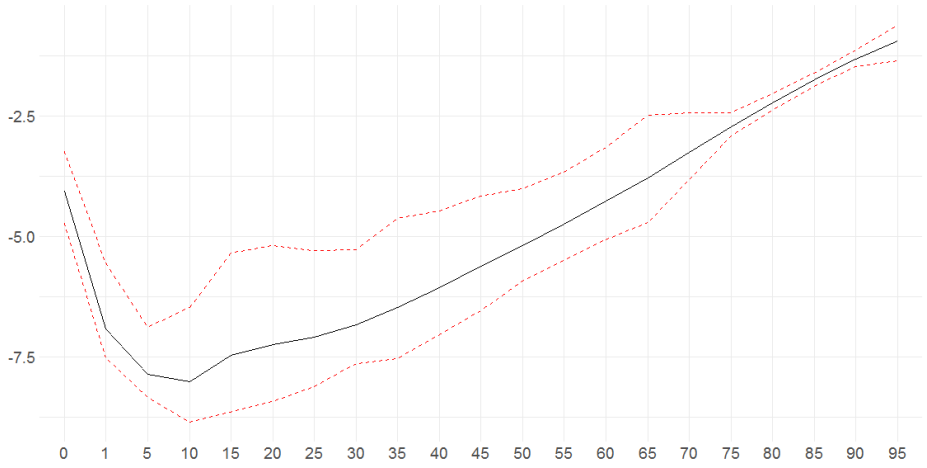


Figure 12.33. The Bayesian posterior mean estimates with 95% posterior credible intervals average over time for $\boldsymbol{\alpha} + \tilde{\mathbf{F}}_t \boldsymbol{\varrho}_t$ for DFM-PC-D-r.

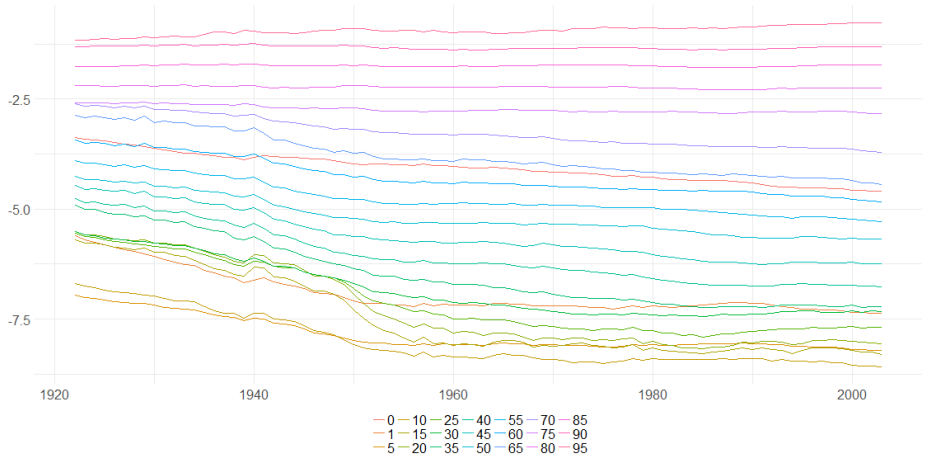


Figure 12.34. The Bayesian posterior mean estimates of $\boldsymbol{\alpha} + \tilde{\mathbf{F}}_t \boldsymbol{\varrho}_t^T$ over time for DFM-PC-D-r. Colours of lines are related to the age groups (the elements of the vector $\boldsymbol{\alpha}$).

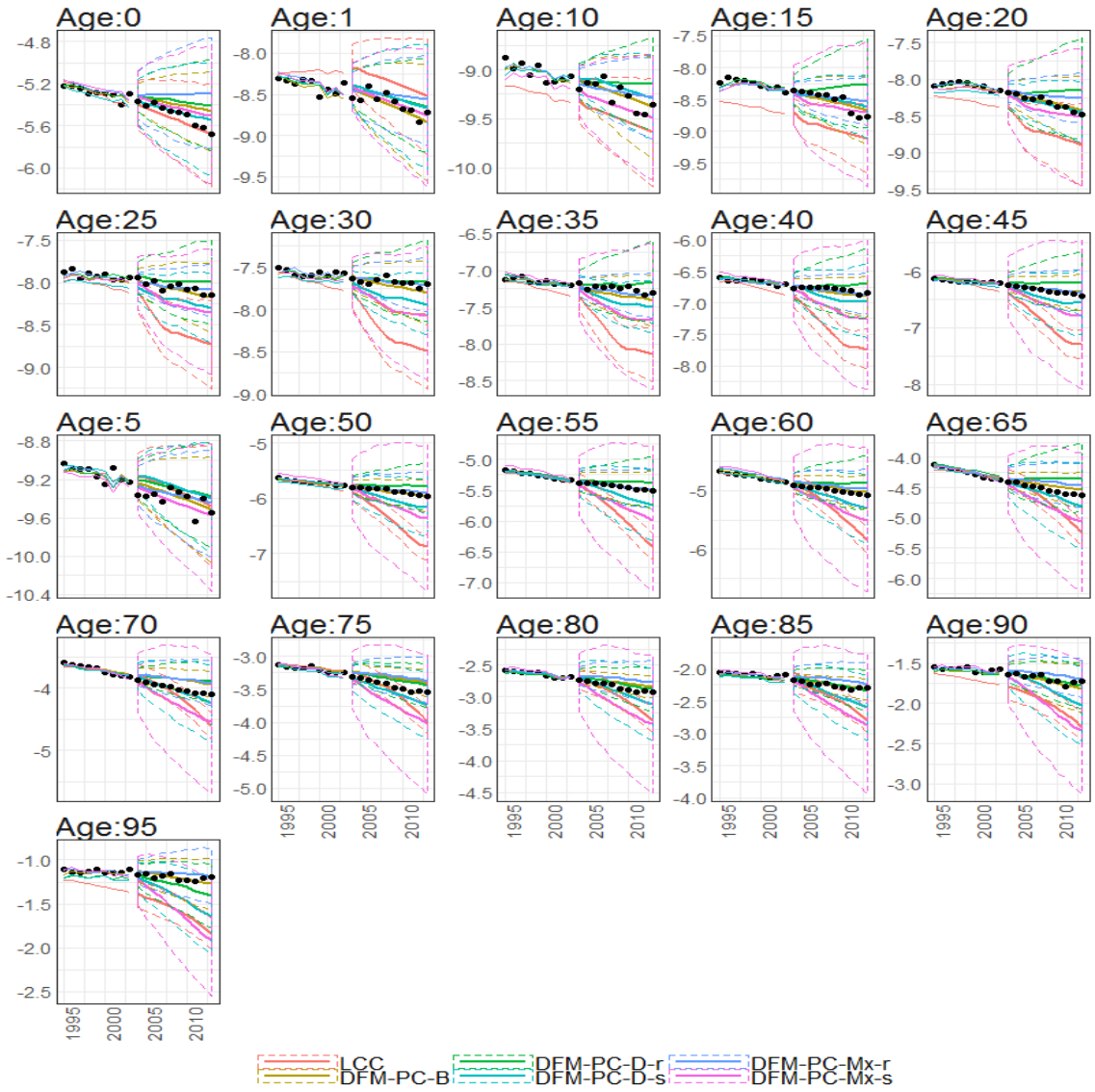


Figure 12.35. 10-year out-of-sample forecasted log death (y axis) rates of different age groups (different subplots) under different models (colours of lines) with corresponding prediction intervals. Calibration period: 1922–2002

Part IV

Conclusion

13. Conclusion

The research work presented in this thesis constitutes important and novel contributions towards probabilistic feature extraction methods and their application to statistical modelling. The primary objective of this thesis was to highlight the advantages of the investigated models from a practical perspective. We focused on developing a dimensionality reduction methodology, which addresses a difficult but not uncommon situation when the underlying observation data is not fully observed; that is, it both contains missing information and is corrupted by noise. We propose various methods of dimensionality reduction which adapt Probabilistic Principal Component Analysis (PPCA) and extend the standard assumptions on the distribution of the observation data to heavy-tailed and skewed distributions.

13.1. Probabilistic Principal Component Analysis and Its Estimation Framework Conclusions

Three complementary feature extraction approaches are developed in this thesis, which addresses the challenge of dimensionality reduction in the presence of multivariate heavy-tailed and asymmetric distributions.

First, we demonstrate how to improve the robustness of the standard Probabilistic Principal Component Analysis, denoted by the Gaussian PPCA, by adapting the concept of robust mean and covariance estimation within the standard framework. The application of the robust mean and covariance matrix is further is explored in both initialisation of the EM algorithm, and its iterative steps.

The remaining two families of feature extraction methods extend the Gaussian PPCA by introducing distribution-based robustifications and replacing the assumed Gaussian distribution.

The second proposed framework revisits the t-Student PPCA, which accounts for a heavy-tailed distribution of the observation data. The components of the probabilistic model of the PPCA are assumed to share the same tail-dependency structure, which is also identical across elements of the multivariate latent vectors. We investigate two t-Student frameworks; one in which the observations are assumed to be independent over time, which leads to different probability density functions over time, and the second in which the observation vectors are assumed to be identically distributed and conditionally independent over time. These two assumptions have significant consequences. The independence over time of the random vectors limits the latent vectors to time-specific tail dependence and, therefore, time-specific probability functions. On the other hand, when we restrict the random vectors to be identically distributed over time, the realisations of the variables are only conditionally independent. We propose an estimation framework based on the EM algorithm, which enjoys many useful convergence properties for investigated models.

We investigate the links between the three PPCA frameworks, the Gaussian PPCA, the robust Gaussian PPCA with robust estimators of mean and covariance matrix and two variants of the t-Student PPCA, from the perspective of covariance and mean estimation. We show that the t-Student PPCA, with observations assumed to be distributed independently over time, can be analysed through the robust estimation theory. It allows us to introduce the t-Student PPCA as a method closely related to the Gaussian PPCA, with a robust estimation of covariance matrix by use of the Tyler estimators. Also, we prove equivalence between the singular vectors of the projection matrix in this PPCA model to the eigenvectors of the covariance matrix of the observation process.

The data case studies undertaken aim to identify which methodology is the most robust in the presence of row-wise or element-wise perturbations under different sample sizes, the dimensionality of a single observation, the proportion of missing values or the proportion of the corrupted sample. We begin with analysis of the effect of robust initialisation on the performance of the PPCA algorithms. Robust initialisation improves the accuracy of the estimation in low sample-size cases, especially of the vector with interceptions of the model and when the perturbation pattern is observation-specific, not element-wise. The study shows that, when the sample size increases, the accurately of the robust initialisation step is outstripped by the standard, sample initialisation. The decline in performance is more significant when the proportion of missing data increases.

The studies show that the robust Gaussian PPCA methodology is very sensitive to different initialisations of the robust estimation algorithms. What is more, when the proportion of missing values and dimensionality are low, the accuracy of the estimation provided by the robust Gaussian PPCA is very similar to the Gaussian PPCA. In other scenarios, the standard PPCA framework provides better results.

In the presence of element-wise perturbation, t-Student IND PPCA estimates the mean vector of the PPCA model the most accurately. This result holds regardless of the input parameters. For the majority of cases, the three first eigenvectors of the estimated covariance matrix are the most accurately estimated, either by t-Student IND or Gaussian PPCA. The t-Student IID PPCA outperforms the other methodologies in the presence of high-dimensional data and high proportion of corruption, especially if the perturbation is row-wise. The presented investigations on the synthetic studies provide indications as to which of the proposed methods should be applied given various characteristics of an observation set.

Encouraged by the described outcomes, we propose a generalisation of the t-Student PPCA to skewed distributions and various heavy-tails-dependence structures across elements of the PPCA model. We call this framework a Generalized Skew t-Student PPCA. The method addresses a common situation, in which the elements of an observation vector have non-trivial dependency structures. It is especially relevant to large data problems when subsets of the observation vector represent different, complex information, often combined from various data sets. Therefore, we introduce a method which assumes many sources of corruption which are no longer identical across the dimensionality of the observation vector. Crucially, the model improves the ability to capture tail dependence patterns in the multivariate data analysis. The adaptation of the skew-t

distribution in the PPCA framework adds flexibility to account for asymmetric distributions if they are relevant to an observation set.

The steps of an estimation framework for the Generalized Skew t-Student PPCA using the EM algorithm are no longer explicitly formulated. In this thesis, we address two aspects which pose a difficulty for the estimation process. First, we introduce an approach one may follow to obtain approximated solutions of the E-step in the algorithm. We show that a complex multidimensional integration problem can be simplified to the integration of a two-dimensional unit cube. Next, we discuss how to obtain efficient approximations of solutions to the system of equations which define the M-step of the EM algorithm. We propose a gradient descent step to approximate the solution for the projection matrix, which results in explicit formulas for all remaining static parameters of the PPCA model.

13.2. Application-related Conclusions

After having discussed robust probabilistic feature extraction methods, we applied them to modelling problems which address the use of large data sets in the statistical analysis of important dynamics present in financial and insurance markets

13.2.1. Modelling Yield Curves Using International Exogenous Marco & Financial Factors

In this study, we have performed an extensive analysis of real data in what we believe is the most complete analysis of yield curve modelling to date in the literature, where we incorporated country-specific financial and macroeconomic factors and comprehensively examined variants of the structure of the models.

In order to conduct a parsimonious dimensionality reduction of big financial datasets, we applied the proposed PPCA frameworks to handle missing data in the analysed data sets and to more accurately capture the information contained in the examined datasets. The frameworks were examined on real data, conducting a detailed analysis of results. We highlight the differences between the features provided by different PPCA frameworks and focus on their interpretation.

We provide a coherent approach which uses global macroeconomic and financial market data sets in the analysis of one of the core reference interest rates, the London Interbank Offered Rate (Libor). The rate has served as a key Eurozone market indicator and, until recently, was considered a risk-free interest rate. The Libor rate remains widely used as a reference for pricing and hedging for the majority of financial instruments. We focused on investigating features which measure Euro-Zone market conditions and their decomposition into influences from the domestic and international markets. The Euro Libor yield curve represents the Euro-zone market, whereas exogenous influences are given by dynamics from leading domestic and global economies, such as sovereign rates, inflation rates, productivity, liquidity and credit proxies, and foreign exchange rates. We show that the dimensionality reduction methodologies employed provide information which varies the most before and during periods of increased volatility, such as a financial crisis.

It highlights the importance of the choice of feature extraction frameworks to account for these changes.

In the next step, we combined the extracted features with a Nelson-Siegel regression model for the Euro Libor yield curve, in order to create a hybrid multi-factor regression state-space model. The relevance of market- and economy-related information contained within the exogenous factors is assessed based on the performance of corresponding hybrid models. Our objective is to verify whether the yearly market features provide additional explanatory power in the presence of the standard components of the Nelson-Siegel state-space model.

Following an extensive model selection exercise, we obtain promising outcomes. The hybrid multi-factor model outperforms the standard Nelson-Siegel model in both forecast and in-sample analysis of Euro Libor yield. The utilisation of the market features provides valuable information about the impacts of international data on the Euro Libor curve and, by proxy, on the Euro Zone economy.

The hybrid multi-factor models decompose the yield curve into variables which reflect the changes of the standard components of the yield over time and economic factors which are defined by the features extracted from informational financial and macroeconomic data sets. The model structure gives a straightforward interpretation of forces which influence the Libor curve. It is also especially applicable to the design of stress-testing scenarios in a multi-curve setting, when one might analyse an effect of dynamic cross-country dependence in financial instruments, such as sovereign yields, and their interactions with the Euro-Zone economy.

13.2.2. Incorporating Demographic Information into Mortality Model

We developed and presented a comprehensive study which focuses on the analysis and the incorporation of demographic data into a state-space framework for stochastic mortality modelling. We have extended the well-known Lee-Carter model with stochastic cohort effect by introducing new state processes which correspond to the age-specific dynamic of European female log death rates.

We investigated various ideas for extracting meaningful information from the demographic data of European countries and applying them as explanatory variables to mortality estimation and forecasting. In the presence of short time-series and different types of missing data, the suggested methodology aims to be as parsimonious as possible. By analysing the extracted features, we found more evidence about region-specific mortality structures. Also, the features exhibit significant sensitivity to the methodology of estimation of moments. The robust alternatives to the sample estimators provide more consistent results, regardless of the number of missing entries, especially if the data has not been pre-processed or smoothed.

The results of applying the studied models to the British female log mortality data showed that incorporating the features extracted from European demographic data provides valuable information about the mortality forces which affect the British population. Also, the models with dynamic factors exhibit better in-sample and out-of-sample fit than the Lee-Carter model with the cohort effect.

As an additional outcome of the study, we analyse the effect of the stratification of the data on the family of the Lee-Carter Model. We argue that the stratification mainly influences the cohort effect process and requires more flexibility of modelling than provided by the simplified Lee-Carter cohort model. The investigation showed that the standard Lee-Carter model has similar performance for stratified and non-stratified data, whereas the simplified Lee-Carter cohort model is prone to significant overfitting when fitted to stratified data. Also, the investigation shows that stratification helps to resolve the issues of overfitting related to the Lee-Carter full cohort, due to the smaller dimensionality of the observation vector.

There are a few ways in which this study can be further extended. First of all, with appropriate data, it would be straightforward to conduct similar analysis for extended data sets, as for gender- and region-specific disaggregation of mortality and demographic data, with the inclusion of various population-related factors such as cause-of-death (Murray and Lopez 1997); (Girosi and King 2008) or, more recently, (Arnold (-Gaille) and Sherris 2015); midlife conditions as in (Gavrilov and Gavrilova 2015) or migration. Secondly, considering different distribution prior assumptions and introducing the methodology to handle more advanced patterns of missingness would give the benefit of better explanatory power and even more consistent interpretability. We can also improve the framework of feature extraction to account for fat-tail distributions by means of different robustification methodologies or extensions to Principal Component Analysis, such as Independent Component Analysis, and their functional alternatives, as in (Shang and Hyndman 2017).

13.3. Future Extensions

The primary objective of this research was to extend feature extraction methods in the probabilistic paradigm. The methods explored are particularly relevant when one wishes to apply feature extraction methods in contexts in which both interpretation and quantification of probabilistic confidence of the explanatory power of the results are required. The investigated frameworks facilitate prior assumptions of the data distribution which can be feasibly tailored to our knowledge and ensure a robust analysis of large, incomplete datasets. The proposed methodologies enjoy many encouraging properties which enable one to determine, interpret and respond to different interactions present in large data sets and appear as a consequence of the data' complexity.

The studies which we conducted on financial and insurance datasets demonstrate that the employment of well-designed probabilistic feature extraction methodologies increases the explanatory power of the extracted features used to model key market dynamics, such as interest rates, given global financial and macroeconomic sets, or mortality rates, given the patterns present in international demographic data sets.

These topics that we explored in this research have exposed us to the fields of statistics related to the estimation of probabilistic models with many variables in a high dimensional setting. The introduced EM algorithm-based estimation frameworks can be easily extended to Bayesian approaches, such as Markov Chain Monte Carol (MCMC) algorithms or the emerging field of fast

approximate inference methods, which provide an efficient and powerful framework to solve the problems we studied.

The problem of incomplete data which we have been exploring, often originates from irregular availability, frequency and length of time series one wants to model. A natural set of statistical methodologies to work with when one assumes no fixed intervals between observations of modelled time series, especially in high-dimensional settings, is a functional regression. When, additionally, the available observations vary in number of dimensions, the Variable-Domain functional regression might be a better choice. Recently, Bayesian inference has been argued to be an excellent estimation technique for application and development in functional data as innovative new approaches are developed to work with infinite dimensionality posterior structures. One can certainly now define posteriors on function spaces, as well as run MCMC samplers on such spaces. The usage of the fast approximation inference approach is a faster technique to estimate functional model parameters and construct their credible intervals than standard approaches such as MCMC.

In addition, a natural follow up of the feature extraction methodologies we have studied is the consideration of greater constraints on obtained features seeking their independence. Since independence in probability theory requires identifying the factorisation of the joint probability function of an observation vector, the estimation framework should explore various probability density functions and verification of interaction between them. Independent Component Analysis (ICA), being a methodology based on a linear projection of observed data into a set of independent components, is a newly introduced field of statistics addressing this problem. Its estimation framework is mostly based on optimisation of the projection pursuit index, in a majority of works expressed by mutual information. However, the direct specification of mutual information is not a trivial task, and feature extraction by ICA typically exploits approximate measures of mutual information under the assumption that the data does not contain any error, which is not the case in real applications. Another, in our opinion, more practical approach, is to design parametric models that link ICA to the probabilistic set of methodologies which we have been studying. The flexibility and efficiency of the estimation provided by Bayesian inference and, hereafter, fast approximate inference, is also applicable in this research objective.

Therefore, a natural extension of the research described in this thesis is the further exploration of aspects of functional data analysis in the context of feature extraction methodologies, such as extensions to robust probabilistic Principal Component Analysis or Independent Component Analysis, and their estimation by fast approximate inference in a research capacity. We consider the efficient estimation of the investigated models as a crucial element of providing a powerful framework for flexible and robust dependency modelling, highly applicable to problems involving large datasets.

References

Abbritti, M., S. D. Erba, A. Moreno, and S. Sola. 2013. Global Factors in the Term Structure of Interest Rates.

Abramowitz, M. and I. A. Stegun. 1974. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. New York: Dover Publications.

Akaike, H.. 1973. Information Theory and an Extension of the Maximum Likelihood Principle. In *2nd International Symposium on Information Theory*, Budapest, pp. 267–281. Akademia Kiado.

Akaike, H.. 1974. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control 19*(6), 716–723.

Akaike, H.. 1978. A New Look at the Bayes Procedure. *Biometrika 65*(1), 53–59.

Akaike, H.. 1983. Information Measure and Model Selection. *Bulletin of the International Statistical Institute 44*, 277–291.

Alspach, D. L.. 1974. Gaussian Sum Approximation in Nonlinear Filtering and Control. *Information Science 7*, 271–290.

Anderson, B. D. O. and J. B. Moore. 1979. *Optimal Filtering*. Englewood Cliffs: Prentice-Hall.

Andrieu, C. and A. Doucet. 2010. Particle Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society, Series B 72*(3), 269–342.

Ang, A. and M. Piazzesi. 2003. A No-arbitrage Vector Autoregression of Term Structure Dynamics with Macroeconomic and Latent Variables. *Journal of Monetary Economics 50*(4), 745–787.

Archambeau, C.. 2005. *Probabilistic Models in Noisy Environments And their Application to a Visual Prosthesis for the Blind*. Phd, Universite Catholique de Louvain.

Archambeau, C., N. Delannay, and M. Verleysen. 2006. Robust Probabilistic Projections. In W. Cohen and A. Moore (Eds.), *Proceedings of the 23 rd International Conference on Machine Learning*, Pittsburgh, pp. 33–40. Omni Press.

Armano, G., M. Marchesi, and A. Murru. 2005. A Hybrid Genetic-Nneural Architecture for Stock Indexes Forecasting. *Information Sciences 170*(1), 3–33.

Arnold (-Gaille), S. and M. Sherris. 2015. Causes-of-Death Mortality: What Do We Know on Their Dependence? *North American Actuarial Journal 19*(2), 116–128.

Arulampalam, S., N. Gordon, and B. Ristic. 2004. *Beyond the Kalman Filter: Particle Filters for Tracking Applications Partical Filters for Tracking Applications*. Artech House, Boston | London.

Athans, M., R. Wishner, and A. Bertolini. 1968. Suboptimal State Estimation for Continuous-Time Nonlinear Systems From Discrete Noisy Measurements. *IEEE Transactions on Automatic Control 13*(5), 504–514.

Austin, U. T. and J. Ghosh. 2015. Sparse Submodular Probabilistic PCA. In G. Lebanon and S. V. N. Vishwanathan (Eds.), *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, Volume 38, San Diego, California, pp. 453–461. PMLR.

Azzalini, A. and A. Capitanio. 2003. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t distribution. *Journal of the Royal Statistical Society. Series B (Statistical Methodology) 65*(3), 367–389.

Bach, F. R. and C. Archambeau. 2009. Sparse probabilistic projections. In *Advances in Neural Information Processing Systems 21*, pp. 73–80.

Bain, A. and D. Crisan. 2009. *Fundamentals of Stochastic Filtering*. New York: Springer-Verlag New York.

Bar-Shalom, Y., X. R. Li, and T. Kirubarajan. 2001. *Estimation with Application to Tracking and Navigation*. New York: John Wiley & Sons, Inc.

Barrett, W.. 1961. Convergence Properties of Gaussian Quadrature Formulae. *The Computer Journal 3*(4), 272–277.

Basilevsky, A.. 1994. *Statistical Factor Analysis and Related Methods*. New York: John Wiley & Sons, Inc.

Berger, J. O.. 1985. *Statistical Decision Theory and Bayesian Analysis*. New York: Springer Verlag New York.

Bianchi, F., H. Mumtaz, and P. Surico. 2009. Dynamics of the Term Structure of UK Interest Rates.

Blackwell, D.. 1947. Conditional Expectation and Unbiased Sequential Estimation. *Annals of Mathematical Statistics 18*(1), 105–110.

Borovykh, A., S. Bohte, and C. W. Oosterlee. 2017. Conditional Time Series Forecasting with Convolutional Neural Networks. In *International Conference on Artificial Neural Networks*, pp. 22.

Bouwmans, T., A. Sobral, S. Javed, S. Ki, and E.-h. Zahzah. 2017. Decomposition into Low-Rank plus Additive Matrices for Background / Foreground Separation : A Review for a Comparative Evaluation with a Large-Scale Dataset. *Computer Science Review 23*, 1–71.

Boyles, R. A.. 1983. On the Convergence of the EM Algorithm. *Journal of the Royal Statistical Society B 45*(1), 47–50.

Briers, M., A. Doucet, and S. Maskell. 2010. Smoothing Algorithms for State-Space Models. *Annals of the Institute of Statistical Mathematics 62*(61), 61–89.

Brooks, S., A. Gelman, G. Jones, and X.-L. Meng. 2011. *Handbook of Markov Chain Monte Carlo*. Boston: Chapman & Hall/CRC.

Bucy, R. S.. 1969. Bayes' Theorem and Digital Realization for Nonlinear Filters. *The Journal of the Astronautical Sciences 17*, 80–94.

Bucy, R. S. and P. D. Joseph. 1968. *Filtering for Stochastic Processes with Applications to Guidance*. New York: AMS Chelsea Publishing.

Bucy, R. S. and K. D. Senne. 1971. Digital Synthesis of Nonlinear Filters. *Automatica 7*(3), 287–298.

Burzuini, C., N. G. Best, W. R. Gilks, and C. Larizza. 1997. Dynamic Conditional Independence Models and Markov Chain Monte Carlo Methods. *Journal of the American Statistical Association 92*(440), 1403–1412.

Butler, R. W., P. L. Davies, and M. Jhun. 1993. Asymptotics for the Minimum Covariance Determinant Estimator. *The Annals of Statistics 21*(3), 1385–1400.

C. Croux, P. Filzmoser, and H. Fritz. 2013. Robust Sparse Principal Component Analysis. *Technometrics 55*(2), 29.

Cairns, A., D. Blake, K. Dowd, G. Coughlan, D. Epstein, A. Ong, and I. Balevich. 2009. A Quantitative Comparison of Stochastic Mortality Models using Data from England and Wales and the United States. *North American Actuarial Journal 13*(1), 1–35.

Cairns, A. J. and D. J. Pritchard. 2001. Stability of Descriptive Models for the Term Structure of Interest Rates with Applications to German Market Data. *British Actuarial Journal 7*(3), 467–507.

Candes, E., X. Li, Y. Ma, J. Wright, S.-Y. Huang, Y.-R. Yeh, and S. Eguchi. 2009. Robust principal component analysis? *Neural computation 21*(11), 3179–3213.

Cao, L. J. and F. E. H. Tay. 2003. Support Vector Machine with Adaptive Parameters in Financial Time Series Forecasting. *IEEE Transactions on Neural Networks 14*(6), 1506–1518.

Cappe, O., E. Moulines, and T. Ryden. 2005. *Inference in Hidden Markov Models*. New York: Springer Verlag New York.

Carlin, B. P.. 1992. State Space Modeling of Non-standard Actuarial Time Series. *Insurance: Mathematics and Economics 11*(3), 209–222.

Carlin, B. P., N. G. Polson, and D. S. Stoffer. 1992. A Monte Carlo Approach to Nonnormal and Nonlinear State-Space Modeling. *Journal of the American Statistical Association 87*(418), 493–500.

Carter, C. K. and R. Kohn. 1994a. On Gibbs Sampling for State Space Models. *Biometrika 81*(3), 541–553.

Carter, C. K. and R. Kohn. 1994b. On Gibbs Sampling for State-Space Models. *Biometrika 81*(3), 541–553.

Carvalho, C. M., M. S. Johannes, H. F. Lopes, and N. G. Polson. 2010. Particle Learninig and Smoothing. *Statistical Science 25*(1), 88–106.

Casella, G. and R. L. Berger. 2001. *Statistical Inference*. Pacific Grove: Duxbury Resource Center.

Casella, G. and E. I. George. 1992. Explaining the Gibbs sampler. *The American Statistician 46*(3), 167–174.

Casella, G. and C. P. Robert. 1996. Rao-Blackwellization of sampling schemes. *Biometrika 83*(1), 81–94.

Celeux, C., F. Forbes, C. P. Robert, and D. M. Titterington. 2006. Deviance Unformation Criteria for Missing Data Models. *Bayesian Analysis 1*(4), 651–674.

Chawla, M. M.. 1968. Asymptotic Estimates for the Error of the Gauss-Legendre Quadrature Formula. *The Computer Journal 11*(3), 339–340.

Chawla, M. M. and M. K. Jain. 1968. Asymptotic Error Estimates for the Gauss Quadrature Formula. *Mathematics of Computation 22*(101), 91–97.

Chen, M.-H., Q.-M. Shao, and J. G. Ibrahim. 2000. *Monte Carlo Methods in Bayesian Computation*. New York: Springer-Verlag New York.

Chen, T., E. Martin, and G. Montague. 2009. Robust Probabilistic PCA with Missing Data and Contribution Analysis for Outlier Detection. *Computational Statistics and Data Analysis 53*(10), 3706–3716.

Chen, T., J. Morris, and E. Martin. 2006. Probability Density Estimation via Infinite Gaussian Mixture Model : Application to Statistical Process Monitoring. *Journal of the Royal Statistical Society Series C 55*(5), 699–715.

Chib, S. and B. Ergashev. 2009. Analysis of Multi-Factor Affine Yield Curve. *Journal of American Statistical Association 104*(488), 1324–1337.

Chib, S. and E. Greenberg. 1995. Understanding the Metropolis-Hastings Algorithms. *The American Statistician 49*(4), 327–335.

Chopin, N.. 2002. A Sequential Particle Filter Method for Static Models. *Biometrika 89*(3), 539–551.

Chopin, N.. 2004. Central Limit Theorem for Sequential Monte Carlo and Its Application to Bayesian Inference. *Annals of Statistics 32*(6), 2385–2411.

Coroneo, L., D. Giannone, and M. Modugno. 2016. Unspanned Macroeconomic Factors in the Yield Curve. *Journal of Business & Economic Statistics 34*(3), 472–485.

Crisan, D. and B. Rozovskii. 2011. *The Oxford Handbook of Nonlinear Filtering*. Oxford: Oxford University Press.

Croux, C. and P. Filzmoser. 1998. Robust Factorization of a Data Matrix. In *Compstat*, pp. 245–250.

Croux, C., P. Filzmoser, and M. R. Oliveira. 2007. Algorithms for Projection–Pursuit Robust Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems 87*(2), 218–225.

Croux, C. and G. Haesbroeck. 1999. Influence Function and Efficiency of the Minimum Covariance Determinant Scatter Matrix Estimator. *Journal of Multivariate Analysis 71*(2), 161–190.

Croux, C. and G. Haesbroeck. 2000. Principal Component Analysis based on Robust Estimators of the Covariance or Correlation Matrix: Influence Functions and Efficiencies. *Biometrika 87*(3), 603–618.

Croux, C. and G. Haesbroeck. 2001. Maxbias curves of robust scale estimators based on subranges. *Metrika 53*(2), 101–122.

Croux, C. and A. Ruiz-Gazen. 2005. High Breakdown Estimators for Principal Components: The Projection-Pursuit Approach Revisited. *Journal of Multivariate Analysis 95*(1), 206–226.

Cuchiero, C., C. Fontana, and A. Gnoatto. 2016a. A General HJM Framework for Multiple Yield Curve Modeling. *Finance and Stochastics 20*(2), 267–320.

Cuchiero, C., C. Fontana, and A. Gnoatto. 2016b. Affine Multiple Yield Curve Models. *Available at arXiv:1603.00527v1*, 1–30.

Daul, S., F. Lindskog, and A. McNeil. 2003. The Grouped t-copula with an Application to Credit Risk. *Available at SSRN: https://ssrn.com/abstract=1358956 16*.

Davies, L.. 1992, dec. The Asymptotics of Rousseeuw's Minimum Volume Ellipsoid Estimator. *The Annals of Statistics 20*(4), 1828–1843.

Davies, P. L.. 1987. Asymptotic Behaviour of S-Estimates of Multivariate Location Parameters and Dispersion Matrices. *The Annals of Statistics 15*(3), 1269–1292.

Davis, P. J. and P. Rabinowitz. 1984. *Methods of Numerical Integration*. San Diego: Academic Press.

de Jong, P.. 1989. Smoothing and Interpolation with the State-Space Model. *Journal of the American Statistical Association 84*(408), 1085–1088.

de Jong, P. and N. Shephard. 1995. The Simulation Smoother for Time Series Models. *Biometrika 82*(2), 339–350.

De la Torre, F. and M. J. Black. 2002. Robust Principal Component Analysis for Computer Vision. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, Vancouver, pp. 362–369.

de Ridder, D. and V. Franc. 2003. Robust Subspace Mixture Models Using t -distributions. In *in 14th British Machine Vision Conference*, pp. 319–328.

Del Moral, P., A. Doucet, and A. Jasra. 2006. Sequential Monte Carlo samplers. *Journal of Royal Statistical Society. Series B (Methodological) 68*(3), 11–436.

Dellaportas, P. and A. F. M. Smith. 1993. Bayesian Inference for Generalized Linear and Proportional Hazards Models via Gibbs Sampling. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 42*(3), 443–459.

Dempster, A. P., M. N. Laird, D. B. Rubin, S. Url, R. S. Society, and R. S. Society. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of Royal Statistical Society. Series B (Methodological) 39*(1), 1–38.

Devijver, P. A. and J. Kittler. 1987. *Pattern Recognition Theory and Applications*. Berlin: Springer-Verlag Berlin Heidelberg.

Devroye, L.. 1986. *Non-Uniform Random Variate Generation*. New York: Springer-Verlag New York.

Diebold, F. X., C. Li, G. D. Rudebusch, and C. Li. 2006. Forecasting the Term Structure of Government Bond Yields. *Journal of Econometrics 130*(2), 337–364.

Ding, C., D. Zhou, X. He, and H. Zha. 2006. R1-PCA : Rotational Invariant L1 -norm Principal Component Analysis for Robust Subspace Factorization. In *Proceedings of the 23rd International Conference on Machine Learning, ICML 2006*, pp. 281–288.

Djuric, P. and J. Miguez. 2002. Sequential Particle Filtering in the Presence of Additive Gaussian Noise with Unknown Parameters. In *IEEE Iinternational Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1621–1624.

Doucet, A., M. Briers, and S. Senecal. 2006. Efficient Block Sampling Strategies for Sequential Monte Carlo. *Journal of Computational and Graphical Statistics 15*(3), 693–711.

Doucet, A., N. de Freitas, and N. Gordon. 2001. An Introduction to Sequential Monte Carlo Methods. In A. Doucet, N. de Freitas, and N. Gordon (Eds.), *Sequential; Monte Carlo Methods in Practice*, pp. 3–14. New York: Springer, New York.

Doucet, A., S. Godsill, and C. Andrieu. 2000. On Sequential Monte Carlo Sampling Methods for Bayesian Filtering. *Statistics and Computing 10*(3), 197–208.

Doucet, A. and A. M. Johansen. 2009. A Tutorial on Particle Filtering and Smoothing: Fifteen Years Later. In D. Crisan and B. Rozovskii (Eds.), *The Oxford Handbook of Nonlinear Filtering*, pp. 656–704. Oxford: Oxford University Press.

Duncan, D. and S. Horn. 1972. Linear Dynamic Recursive Estimation from the Viewpoint of Regression Analysis. *Journal of the American Statistical Association 67*(340), 815–821.

Durbin, J. and S. J. Koopman. 2012. *Time Series Analysis by State Space Methods*. Oxford: Oxford University Press.

Efron, B.. 1979. Bootstrap Methods: Another Look at the Jackknife. *Annals of Statistics 7*(1), 1–26.

Eidukas, A.. 2016. Coverage differences between the Harmonised Index of Consumer Prices and national Consumer Prices Indices. Technical report, Office for National Statistics.

Emanuel Moench. 2008. Forecasting the Yield Curve in a Data-Rich Environment: A No-Arbitrage Factor-Augmented VAR Approach. *Journal of Econometrics 146*(1), 26–43.

Erbas, B., M. Akram, D. M. Gertig, D. English, J. L. Hopper, A. Kavanagh, and R. Hyndman. 2010. Using Functional Data Analysis Models to Estimate Future Time Trends in Age-Specific Breast Cancer Mortality for the United States and England-Wales. *Journal of Epidemiology 20*(2), 159–65.

Eriksson, A. and A. V. D. Hengel. 2012. Efficient Computation of Robust Weighted Low-Rank Matrix Approximations Using the L1 Norm. *IEEE Transactions on Pattern Analysis and Machine Intelligence 34*(9), 1681–1690.

Evans, M. and T. Swartz. 2000. *Approximating Integrals via Monte Carlo and Deterministic Method*. New York: Springer-Verlag New York.

Faa di Bruno, F.. 1857. On a New Formula of Differential Calculus. *The Quarterly Journal of Pure and Applied Mathematics 1*.

Fang, Y. and M. K. Jeong. 2008. Robust Probabilistic Multivariate Calibration Model. *Technometrics 50*(3), 305–316.

Fearnhead, P.. 1998. *Sequential Monte Carlo in Filter Theory*. Ph. D. thesis, University of Oxford.

Fearnhead, P.. 2002. MCMC, Sufficient Statistics and Particle Filters. *Journal of Computational and Graphical Statistics 11*(4), 848–862.

Filipovic, D. and A. B. Trolle. 2013. The Term Structure of Interbank Risk. *Journal of Financial Economics 109*(4), 707–733.

Fishman, G.. 1996. *Monte Carlo: Concepts, Algorithms and Applications*. New York: Springer-Verlag New York.

Frahm, G.. 2004. *Generalized Elliptical Distributions: Theory and Applications*. Ph. D. thesis, University of Cologne.

Frahm, G. and U. Jaekel. 2010. A Generalization of Tyler's M-estimators to the Case of Incomplete Data. *Computational Statistics & Data Analysis 54*(2), 374–393.

Frank, K. and S. Heinrich. 1996. Computing discrepancies of Smolyak quadrature rules. *Journal of Complexity 12*(4), 287–314.

Fraser, D. C.. 1967. A New Technique for the Optimal Smoothing of Data. Technical report, M.I. T, Instrumentation Lab.

Friedman, J. H. H. and J. W. W. Tukey. 1974. A Projection Pursuit Algorithm for Exploratory Data Analysis. *IEEE Transactions on Computers C-23*(9), 881–890.

Frühwirth-Schnatter, S.. 1993. Applied State Space Modelling of Non-Gaussian Time Series using Integration-based Kalman-filtering. *Statistics and Computing 4*(4), 259–269.

Fruhwirth-Schnatter, S.. 1994. Data Augmentation and Dynamic Linear Models. *Journal of Time Series Analysis 15*(2), 183–202.

Fruhwirth-Schnatter, S.. 2006. *Finite Mixture and Markov Switching Models in Matlab* (Springer S ed.). Springer.

Fung, M. C., G. W. Peters, and P. V. Shevchenko. 2017. A Unified Approach to Mortality Modelling Using State-Space Framework: Characterisation, Identification, Estimation and Forecasting. *Annals of Actuarial Science 11*(2), 343–389.

Fung, M. C., G. W. Peters, and P. V. Shevchenko. 2018. Cohort Effects in Mortality Modelling: A Bayesian State-Space Analysis. *Annals of Actuarial Science*, 1–36.

Gamerman, D. and H. F. Lopes. 2006. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, Second Edition*. Boston: Chapman & Hall/CRC.

Gao, J.. 2008. Robust L1 Principal Component Analysis and Its. *Neural Computation 20*(2), 555–572.

Gavrilov, L. A. and N. S. Gavrilova. 2015. Predictors of Exceptional Longevity: Effects of Early-Life and Midlife Conditions, and Familial Longevity. *North America Actuarial Journal 19*(3), 174–186.

Gelfand, A. E. and A. F. M. Smith. 1990. Sampling-based Approaches to Calculating Marginal Densities. *Journal of the American statistical association 85*(410), 398–409.

Geman, S. and D. German. 1984. Stochastic Relaxation,Gibbs Distributions, and the Bayesian Restoration of Image. *IEEE Transactions on Pattern Analysis and Machine Intelligence 6*(6), 721–741.

George, E. I., U. E. Makov, and A. F. M. Smith. 1993. Conjugate Likelihood Distributions. *Wiley on behalf of Board of the Foundation of the Scandinavian Journal of Statistics 20*(2), 147–156.

Geva, T. and J. Zahavi. 2014. Empirical Evaluation of an Automated Intraday Stock Recommendation System Incorporating Both Market Data and Textual News. *Decision Support Systems 57*, 212–223.

Geweke, J.. 1988. Antithetic Acceleration of Monte Carlo Integration in Bayesian Inference. *Journal of Econometrics 38*(1-2), 73–89.

Geweke, J.. 1989. Bayesian Inference in Econometric Models Using Monte Carlo Integration. *Econometrica 57*(6), 1317–1339.

Geweke, J. and H. Tanizaki. 1999. On Markov Chain Monte Carlo Methods for Nonlinear and Non-Gaussian State-Space Models. *Communications in Statistics - Simulation and Computation 28*(4), 867–894.

Geweke, J. and H. Tanizaki. 2001. Bayesian Estimation of State-Space Models using the Metropolis-Hastings Algorithm within Gibbs Sampler. *Computational Statistics & Data Analysis 37*(2), 151–170.

Gilks, W. R. and C. Berzuini. 2001. Following a moving target - Monte Carlo Inference for Dynamic Bayesian Models. *Journal of Royal Statistical Society. Series B (Methodological) 63*(1), 127–146.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter. 1995. *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall/CRC.

Gilks, W. R. and P. Wild. 1992. Adaptive Rejection Sampling for Gibbs Sampling. *Applied Statistics 41*(2), 337–348.

Girosi, F. and G. King. 2008. *Demographic Forecasting*. Princeton: Princeton University Press.

Gordon, N., D. Salmond, and C. Ewing. 1995. Bayesian State Estimation for Tracking and Guidance Using the Bootstrap Filter. *Journal of Guidance, Control and Dynamics 18*(6), 1434–1443.

Gordon, S., D. J. Salmond, A. F. M. Smith, N. J. Gordon, D. J. Salmond, and A. F. M. Smith. 1993. Novel Approach to Nonlinear/Non-Gaussian Bayesian State Estimation. *IEE Proceedings F: Radar and Signal Processing 140*(2), 107 – 113.

Grbac, Z. and W. J. Runggaldier. 2015. *Interest Rate Modeling: Post-Crisis Challenges and Approaches*. Springer International Publishing.

Greenacre, M.. 1984. *Theory and Applications of Correspondence Analysis*. London: Academic Press.

Guan, Y. and J. G. Dy. 2009. Sparse Probabilistic Principal Component Analysis. In D. A. van Dyk and M. Welling (Eds.), *12th International Conference on Artificial Intelligence and Statistics (AISTATS 2009)*, Hilton Clearwater Beach Resort, pp. 185–192. JMLR.org.

Gupta, A. K.. 2003. Multivariate Skew t-Distribution. *Statistics: A Journal of Theoretical and Applied Statistics 37*(4), 359–363.

Gupta, A. K. and D. K. Nagar. 1999. *Matrix Variate Distributions*. New York: Chapman and Hall/CRC.

Haberman, S. and A. E. Renshaw. 2011. A Comparative Study of Parametric Mortality Projection Models. *Insurance: Mathematics and Economics 48*(1), 35–55.

Hamilton, J. D.. 1994. Time Series Analysis.

Hammersley, J. M. and D. C. Handscomb. 1964. *Monte Carlo Methods*. London: Chapman and Hall.

Hammersley, J. M. and K. W. Morton. 1954. Poor Man's Monte Carlo. *Journal of the Royal Statistical Society. Series B (Methodological) 16*(1), 23–38.

Handschin, J. E. and Q. Mayne. 1969. Monte Carlo Techniques to Estimate Conditional Expectation in Multi-State Non-Linear Filtering. *International Journal of Control 9*(5), 547–559.

Hanewald, K.. 2011. Explaining Mortality Dynamics: The Role of Macroeconomic Fluctuations and Cause of Death Trends. *North American Actuarial Journal, Series B 15*(2), 290–314.

Harrison, P. J. and C. F. Stevens. 1976. Bayesian Forecasting. *Journal of the Royal Statistical Society. Series B (Methodological) 38*(3), 205–247.

Harvey, A. C.. 1990. *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: Cambridge University Press.

Harvey, A. C. and G. D. A. Philips. 1979. Maximum Likelihood Estimation of Regression Models with Autoregressive- Moving Average Disturbance. *Biometrika 66*(1), 49–58.

Hastings, W. K.. 1970. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika 57*(1), 97–109.

Higuchi, T.. 1997. Monte Carlo filter Using the Genetic Algorithm Operators. *Journal of Statistical Computation and Simulation 59*(1), 1–23.

Hotelling, H.. 1933. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology 24*(6), 417–441.

Huber, P. J.. 1964. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics 35*(1), 73–101.

Huber, P. J. and E. M. Ronchetti. 2009. *Robust Statistics*. New York: Wiley New York.

Hubert, M., P. J. Rousseeuw, and K. Vanden Branden. 2005. ROBPCA: A New Approach to Robust Principal Component Analysis. *Technometrics 47*(1), 64–79.

Hunt, A. and A. M. Villegas. 2015. Robustness and Convergence in the Lee-Carter Model with Cohort Effects. *Insurance: Mathematics and Economics 64*, 186–202.

Hurzeler, M. and H. Kunsch. 1995. Monte Carlo Approximations for General State Space Models. *Journal of Computational and Graphical Statistics 7*(2), 175–192.

Hyndman, R. J. and F. Yasmeen. 2012. Common Functional Principal Component Models for Mortality Forecasting. In E. G. Bongiorno, A. Goia, E. Salinelli, and P. Vieu (Eds.), *Contributions in infinite-dimensional statistics and related topics*, pp. 161–166. Bologna: Societa Editrice Esculapio.

Ilin, A. and T. Aiko. 2010. Practical Approaches to Principal Component Analysis in the Presence of Missing Values. *Journal ofMachine Learning Research 11*(July), 1957–2000.

Isaacson, E. and H. B. Keller. 1966. *Analysis of Numerical Methods*. New York: Dover Publications.

Ito, K. and K. Xiong. 2000. Gaussian Filters for Nonlinear Filtering Problems. *IEEE Transactions on Automatic Control2 45*(5), 910 – 927.

Jazwinski, A. H.. 1970. *Stochastic Processes and Filtering Theory*. New York: Academic Press.

Jolliffe, I., N. Trendafilov, and M. Uddin. 2003. A Modified Principal Component Technique Based On the LASSO. *Journal of Computational and Graphical Statistics 12*(3), 531–547.

Jolliffe, I. T.. 2002. *Principal Component Analysis*. New York: Springer-Verlag New York.

Jones, R. H.. 1981. Fitting a Continuous Time Autoregression to Discrete Data. In D. F. ed Time (Ed.), *Applied Time Series Analysis*, New York, pp. Findley. Series Analysi.

Joslin, S., M. Priebsch, and K. J. Singleton. 2014. Risk Premiums in Dynamic Term Structure Models with Unspanned Macro Risks. *The Journal of Finance 69*(3), 1197–1233.

Joslin, S. S. W. S., K. J. Singleton, and H. Zhu. 2010. A New Perspective on Gaussian Dynamic Term Structure Models. *Review of Financial Studies 24*(3), 926–970.

Julier, S. J. and J. K. Uhlmann. 1997. A New Extension of the Kalman Filter to Nonlinear Systems. In *AeroSense: The 11th International Symposium on Aerospace/Defense Sensing, Simulation and Controls*, pp. 182–193.

Kallianpur, G.. 1980. *Stochastic Filtering Theory*. New York: Springer-Verlag New York.

Kalman, R. E.. 1960. A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME - Journal of basic Engineering 82*(Series D), 35–45.

Kalman, R. E.. 1962. Mathematical Description of Linear Dynamical System. *Journal of the Society for Industrial and Applied Mathematics Series A Control 1*(2), 152–192.

Kalman, R. E.. 1963. New Methods in Wiener Filtering Theory. In J. L. Bogdanoff and F. Kozin (Eds.), *First Symposium on Engineering Applications of Random Function Theory and Probability*, new York. John Wiley \& Sons, Inc.

Kalman, R. E. and R. S. Bucy. 1961. New Results in Linear Filtering and Prediction Theory. *Journal of Basic Engineering 83*(1), 95–108.

Karimalis, E., I. Kosmidis, and G. W. Peters. 2017. Multi Yield Curve Stress-Testing Framework Incorporating Temporal and Cross Tenor Structural Dependencies.

Ke, Q. and T. Kanade. 2005. Robust L1 Norm Factorization in the Presence of Outliers and Missing Data by Alternative Convex Programming. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, pp. 739–746.

Kercheval, A. N. and Y. Zhang. 2015. Modeling High-Frequency Limit Order Book Dynamics with Support Vector Machines. *Quantitative Finance 15*(8), 1315–1329.

Khan, Z. and F. Dellaert. 2003. Robust Generative Subspace Modeling : The Subspace t Distribution. In *British Machine Vision Conference 2003*, pp. 1–17.

Kim, C.-J. and C. R. Nelson. 1999. *State-Space Models with Regime Switching: Classical and Gibbs-Sampling Approaches with Applications*. Cambridge: MIT Press.

Kitagawa, G.. 1983. Non-Gaussian State Space Modeling of Nonstationary Time Series. *Journal of American Statistical Association 82*(400), 1032–1041.

Kitagawa, G.. 1994. The Two-filter Formula for Smoothing and an Implementation of the Gaussian-sum Smoother. *Annals of the Institute of Statistical Mathematics 46*(4), 605–623.

Kitagawa, G.. 1996. Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State Space Models. *Journal of Computational and Graphical Statistics 5*(1), 1–25.

Kitagawa, G.. 1998. A Self-Organizing State-Space Model. *Journal of the American Statistical Association 93*(443), 1203–1215.

Kitagawa, G. and W. Gersh. 1996. *Smoothness Priors Analysis of Time Series*. New York: Springer-Verlag New York.

Kloek, T. and K. van Dijk. 1978. Bayesian Estimates of Equation Systems Parameters: An Application of Integration by Monte Carlo. *Econometrica 46*(1), 1–19.

Kogure, A. and Y. Kurachi. 2010. A Bayesian Approach to Pricing Longevity Risk based on Risk-neutral Predictive Distributions. *Insurance: Mathematics and Economics 46*(1), 162–172.

Kohn, R. and C. F. Ansley. 1989. Filtering and Smoothing Algorithms for State Space Models. *Computers & Mathematics with Applications 18*(6-7), 515–528.

Kolmogorov, A. N.. 1941. Stationary Sequences in Hilbert spaces. *Bull. Math Univ. Moscow 2*(6), 1–40.

Kolmogorov, A. N.. 1950. Unbiased Estimates. *Izvestiya Akad. Nauk SSSR 14*(4), 303–326.

Kong, A.. 1992. A Note on Importance Sampling Using Standardized Weights. Technical report, Department of Statistics, University of Chicago, Chicago.

Kong, A., J. S. Liu, and W. H. Wong. 1994. Sequential Imputations and Bayesian Missing Data Problems. *Journal of the American Statistical Association 89*(425), 278–288.

Kotz, S. and S. Nadarajah. 2004. *Multivariate t Distributions and Their Applications*. Cambridge: Cambridge University Press.

Kramer, S. C. and H. W. Sorenson. 1988. Recursive Bayesian Estimation Using Pointwise Constant Approximation. *Automatica 24*(6), 789–801.

Kushner, H. J.. 1967. Dynamical Equations for Optimal Nonlinear Filtering. *Journal of Differential Equations 3*(2), 179–190.

Kwon, Y.-k. and B.-r. Moon. 2007. A Hybrid Neurogenetic Approach for Stock Forecasting. *IEEE Transactions on Neural Networks 18*(3), 851–864.

Lange, K. L.. 1995. A Gradient Algorithm Locally Equivalent to the EM Algorithm. *Journal of Royal Statistical Society. Series B (Methodological) 57*(2), 425–437.

Lange, K. L., R. J. A. Little, J. M. G. Taylor, K. L. Lange, R. J. A. Little, and J. M. G. Taylor. 1989. Robust Statistical Modeling Using the t Distribution. *Journal of the American Statistical Association 84*(408), 881–896.

Lauritzen, S. L.. 1981. Time Series Analysis in 1880: a Discussion of Contributions Made by T. N. Thiele. *International Statistical Review / Revue Internationale de Statistique 49*(3), 319–33.

Lee, R. D. and L. R. Carter. 1992. Modeling and Forecasting US Mortality. *Journal of the American Statistical Association 87*(419), 659–671.

Lehmann, E. L. and G. Casella. 1998. *Theory of Point Estimation*. New York: Springer-Verlag New York.

Lether, F.. 1970. An Error Representation for Product Cubature Rules. *SIAM Journal of Numerical Analysis 7*(3), 363–365.

Lether, F.. 1971a. Error Bounds for Fully Symmetric Cubature Rules. *SIAM Journal of Numerical Analysis 8*(1), 49–60.

Lether, F. G.. 1971b. Cubature Error Bounds for Gauss-Legendre Product Rules. *SIAM Journal of Numerical Analysis 8*(1), 36–42.

Lether, F. G.. 1980. Error Estimates for Gaussian Quadrature. *Applied Mathematics and Computation 7*(3), 237–246.

Little, R. J. A. and D. B. Rubin. 2002. *Statistical Analysis with Missing Data*. New York: John Wiley & Sons, Inc.

Liu, C. and D. B. Rubin. 1995. ML Eestimation of the t Distribution Using EM and Its Extensions , ECM and ECME. *Statistica Sinica 5*, 19–39.

Liu, J. and R. Chen. 1998. Sequential Monte Carlo Methods for Dynamic Systems. *Journal of American Statistical Association 93*(443), 1032–1044.

Liu, J. and M. West. 2001. Combined Parameter and State Estimation in Simulation-Based Filtering. In A. Doucet, N. de Freitas, and N. Gordon (Eds.), *Sequential Monte Carlo Methods in Practice*, pp. 197–223. New York: Springer Verlag New York.

Liu, J. S.. 2004. *Monte Carlo Strategies in Scientific Computing* (Springer S ed.). New Yor: Springer-Verlag New York.

Liu, J. S. and R. Chen. 1995. Blind Deconvolution via Sequential Imputations. *Journal of the American Statistical Association 90*(430), 567–576.

Liu, J. S., W. H. Wong, and K. A.. 1995. Covariance Structure and Convergence Rate of the Gibbs Sampler with Various Scans. *Journal of the Royal Statistical Society. Series B (Methodological)1 57*(1), 157–169.

Liu, J. S., W. H. Wong, and A. Kong. 1994. Covariance Structure of the Gibbs Sampler with Applications to the Comparisons of Estimators and Augmentation Schemes. *Biometrika 81*(1), 27–40.

Ljung, L. and T. Soderstrom. 1983. *Theory and Practice of Recursive Identification*. Cambridge: MIT Press.

Lopes, H. F. and R. S. Tsay. 2011. Particle Filters and Bayesian Inference in Financial Econometrics. *Journal of Forecasting 30*(1), 168–209.

Lopuhaa, H. P.. 1989. On the Relation between S-Estimators and M-Estimators of Multivariate Location and Covariance. *The Annals of Statistics 17*(4), 1662–1683.

Ludvigson, S. and S. Ng. 2010. A Factor Analysis of Bond Risk Premia. In A. Uhla and D. E. A. Gile (Eds.), *Handbook of Empirical Economics and Finance*, pp. 313–372. Boca Raton: Chapman and Hall.

Luo, X. and P. V. Shevchenko. 2012. Bayesian Model Choice of Grouped t-copula. *Methodology and Computing in Applied Probability 14*(4), 1097–1119.

MacEachern, S. N., M. A. Clyde, and J. S. Liu. 1999. Sequential Importance Sampling for Nonparametric Bayes Models: The Next Generation. *The Canadian Journal of Statistics 27*(2), 251–267.

Mackay, D. J. C.. 1998. Choice of Basis for Laplace Approximation. *Machine Learninig 33*(1), 77–86.

Maronna, R. A.. 1976. Robust M-Estimators of Multivariate Location and Scatter. *The Annals of Statistics 4*(1), 51–67.

Marsaglia, G.. 1977. The Squeeze Method for Generating Gamma Variates. *Computers & Mathematics with Application 3*(4), 321–325.

Marshall, A.. 1956. The Use of Multi-Stage Sampling Schemes in Monte Carlo Computations. In H. A. Meyer (Ed.), *Symposium on Monte Carlo Methods*, pp. 20. New York: John Wiley & Sons.

McLachlan, G. and T. Krishnan. 1997. *The EM Algorithm and Extensions*. New York: Wiley.

Meinhold, R. J. and N. D. Singpurwalla. 1983. Understanding the Kalman Filter. *American Statistician 37*(2), 123–127.

Mengersen, K. L. and R. L. Tweedie. 1996. Rates of Convergence of the Hastings and Metropilos Algorithms. *Annals of Statistics 24*(1), 101–121.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics 21*(6), 1087–1092.

Metropolis, N. and S. Ulam. 1949. The Monte Carlo Method. *Journal of American Statistical Association 44*(247), 335–341.

Muller, P.. 1991. A Generic Approach to Posterior Integration and Gibbs Sampling. Technical report, Institute of Statistics and Decision Sciences, Duke University.

Murray, C. J. and A. D. Lopez. 1997. Alternative Projections of Mortality and Disability by Cause 1990-2020: Global Burden of Disease Study. *The Lancet 349*(9064), 1498–1504.

Musso, C., N. Oudjane, and F. Le Gland. 2001. Improving Regularized Particle Filters. In A. Doucet, N. de Freitas, and N. Gordon (Eds.), *Sequential Monte Carlo Methods in Practice*, pp. 247–271. New York: Springer-Verlag New York.

Naveau, P., M. G. Genton, and X. Shen. 2005. A Skewed Kalman filter. *Journal of Multivariate Analysis2 94*(2), 382–400.

Neal, R. M.. 1993. Probabilistic Inference Using Markov Chain Monte Carlo Methods. Technical report, Department of Computer Science, University of Toronto.

Nelson, C. R. and A. F. Siegel. 1987. Parsimoniuos Modeling of Yield Curves. *The Journal of Business 60*(4), 473–489.

Niu, G. and B. Melenberg. 2014. Trends in Mortality Decrease and Economic Growth. *Demography 51*(5), 1755–1773.

Novak, E. and F. Schiller. 1999. Simple Cubature Formulas with High Polynomial Exactness. *Constructive Approximation 15*(4), 499–522.

O'Hagan, A., C. E. Buck, A. Daneshkhah, P. H. Eiser, J. Richard Garthwaite, D. J. Jenkinson, J. E. Oakley, and T. Rakow. 2006. *Uncertain Judgements: Eliciting Experts' Probabilities*. Chichester: John Wiley & Sons.

Okatani, T. and K. Deguchi. 2007. On the Wiberg Algorithm for Matrix Factorization in the Presence of Missing Components. *International Journal of Computer Vision 72*(3), 329–337.

Park, J., Z. Zhang, H. Zha, and R. Kasturi. 2004. Local Smoothing for Manifold Learning. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004*, Washington,D.C., pp. 452–459. IEEE Press.

Pedroza, C.. 2006. A Bayesian Forecasting Model: Predicting US Male Mortality. *Biostatistics 7*(4), 530–550.

Peel, D. and J. G. McLachlan. 2000. Robust Mixture Modelling Using the t distribution. *Statistics and Computing 10*(4), 339–348.

Peters, G. W.. 2005. *Topics in Sequencial Monte Carlo Samplers*. Ph. D. thesis, University of Cambridge.

Pitt, M. and N. Shephard. 1999. Filtering via Simulation: Auxillary Particle Filter. *Journal of American Statistical Association 94*(446), 590–599.

Pitt, M. and N. Shephard. 2001. Auxiliary Variable Based Particle Filters. In A. Doucet, N. de Freitas, and N. Gordon (Eds.), *Sequential Monte Carlo Methods in Practice*. new York: Springer Verlag New York.

Pollock, M.. 2010. Introduction to Particle Filtering.

Polson, N. G., J. Stroud, and P. Muller. 2008. Practical Filtering With Sequential Parameter Learning. *Journal of the Royal Statistical Society Series B 70*(2), 413–428.

Poyiadjis, G., A. Doucet, and S. S. Singh. 2005. Particle Methods For Optimal Filter Derivative: Application to Parameter Estimation. In *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing*, Philadelphia, pp. 925–928. IEEE Press.

Poyiadjis, G., A. Doucet, and S. S. Singh. 2011. Particle Approximations of the Score and Observed Information Matrix in State Space Models with Application to Parameter Estimation. *Biometrika 98*(1), 65–80.

Raiffa, H. and R. Schlaifer. 1961. *Applied Statistical Decision Theory*. Boston: Division of Research, Graduate School of Business Administration, Harvard University.

Rao, C. R.. 1945. Information and Accuracy Attainable in the Estimation of Statistical Parameters. *Bulletin of the Calcutta Mathematical Society. 37*(3), 81–91.

Rauch, H. E., C. T. Striebel, and F. Tung. 1965, aug. Maximum likelihood estimates of linear dynamic systems. *AIAA Journal 3*(8), 1445–1450.

Redner, R. A. and H. F. Walker. 1984. Mixture Densities, Maximum Likelihood and the EM Algorithm. *SIAM Review 26*(2), 195–239.

Renshaw, A. E. and S. Haberman. 2006. A Cohort-based Extension to the Lee-Carter model for Mortality Reduction Factors. *Insurance: Mathematics and Economics 38*(3), 556–570.

Rober, C. P. and G. Casella. 2004. *Monte Carlo Statistical Methods* (Springer T ed.). New York: Springer Verlag, New York.

Robert, C. P.. 1998. *Discretization and MCMC Convergence Assessment* (Lecture No ed.). New York: Springer-Verlag New York.

Robert, C. P.. 2007. *The Bayesian Choice: From Decision-Theoretic Foundations to Computentional Implementation*. new York: Springer.

Robert, P. and Y. Escoufier. 1976. A Unifying Tool for Linear Multivariate Statistical Methods: The RV-Coefficient. *Applied Statistics 25*(3), 257–265.

Roberts, G. O. and N. G. Polson. 1994. On the Geometric Convergence of the Gibbs Sampler. *Journal of the Royal Statistical Society. Series B (Methodological) 56*(2), 377–384.

Roberts, G. O. and A. F. M. Smith. 1994. Simple Conditions for the Convergence of the Gibbs Sampler and Metropolis-Hastings Algorithms. *Stochastic Processes and their Applications 49*(2), 207–2016.

Roberts, R. O. and R. L. Tweedie. 1996. Geometric Convergence and Central Limit Theorems for Multidimensional Hastings and Metropolis algorithms. *Biometrika 83*(1), 95–110.

Rosenbluth, M. N. and A. W. Rosenbluth. 1955. Monte Carlo Calculation of Average Extention of Molecular Chains. *The Journal of Chemical Physics 23*(2), 356–359.

Rousseeuw, P.. 1985. Multivariate Estimation With High Breakdown Point. In W. Grossmann, G. Pflug, I. Vincze, and W. Wertz (Eds.), *Mathematical Statistic and Applications*, pp. 283–297. Dordrecht: Reidel Publishing Company.

Rousseeuw, P. and V. Yohai. 1984. Robust Regression by Means of S-Estimators. In J. Franke, W. Hardle, and D. Martin (Eds.), *Robust and Nonlinear Time Series Analysis*, pp. 256–272. Heidelberg: Springer-Verlag New York.

Roweis, S.. 1998. EM Algorithms for PCA and SPCA. In M. Jordan, M. Kearns, and S. Solla (Eds.), *Advances in Neural Information Processing Systems*, Denver, pp. 626–632. MIT Press.

Rubin, D. B.. 1987. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, Inc.

Rubin, D. B.. 1988. Using the SIR Algorithm to Simulate Posterior Distributions. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith (Eds.), *Bayesian Statistics 3: Proceedings of the Third Valencia International Meeting*, Oxford, pp. 395–402. Oxford University Press.

Rubin, D. B. and D. T. Thayer. 1982. EM Algorithms for ML Factor Analysis. *Psychometrika 47*(1), 69–76.

Saha, S., E. Ozkan, and F. Gustafsson. 2010. Marginalized Particle Filters for Bayesian Estimation of Gaussian Noise Parameters. In *13th Conference on Information Fusion*, Edinburgh, pp. 1–8.

Särkkä, S.. 2013. *Bayesian Filtering and Smoothing*. Cambridge: Cambridge University Press.

Sayed, A. H. and T. Kailath. 1994. A State-Space Approach to Adaptive RSL Filtering. *IEEE Signal Processing Magazine 11*(3), 18–60.

Schervish, M. J.. 1995. *Theory of Statistics* (Springer S ed.). New York: Springer Verlag, New York.

Schervish, M. J. and B. P. Carlin. 1992. On the Convergence of Successive Substitution Sampling. *Journal of Computational and Graphical Statistics 1*(2), 111–127.

Schnatter, S.. 1992. Integration-based Kalman-Filtering for a Dynamic Generalized Linear Trend Model. *Computational Statistics and Data Analysis 13*(4), 447–459.

Schulz, A. and J. Stapf. 2014. Price Discovery on Traded Inflation Expectations: Does the Financial Crisis Matter? *The European Journal of Finance 20*(11), 1037–1063.

Shang, H. L. L. and R. J. Hyndman. 2017. Grouped Functional Time Series Forecasting: an Application to Age-Specific Mortality Rates. *Journal of Computational and Graphical Statistics 26*(2), 330–343.

Shao, J.. 1989. Monte Carlo Approximations in Bayesian Decision Theory. *Journal of the American Statistical Association 42*(407), 727–732.

Shiryaev, A. N.. 1996. *Probability*. New York: Springer Verlag New York.

Sirignano, J.. 2016. Deep Learning for Limit Order Books. *Available at SSRN: https://ssrn.com/abstract=2710331*, 53.

Smith, A. F. M. and A. E. Gelfand. 1992. Bayesian Statistics without Tears: A Sampling-Resampling Perspective. *The American Statistician 46*(2), 84–88.

Smolyak, S. A.. 1963. Quadrature and Interpolation Formulas for Tensor Products of Certain Classes of Functions. *Dokl. Akad. Nauk SSSR 148*(5), 1042–1045.

Sorenson, H. W.. 1985. *Kalman Filtering: Theory and Application*. New York: IEEE Press.

Sorenson, H. W. and D. L. Alspach. 1971. Recursive Bayesian Estimation Using Gaussian Sums. *Automatica 7*(4), 465–479.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. van der Linde. 2002. Bayesian Measures of Model Complexity and Fit. *Journal of the Royal Statistical Society, Series B 64*(4), 583–639.

Stratonovich, R. L.. 1968. *Conditional Markov Processes and Their Application to the Theory of Optimal Control*. New York: Elsevier Science & Technology.

Stroud, A. H. and D. Secrest. 1966. *Gaussian Quadrature Formulas*. Englewood Cliffs: Prentice-Hall.

Strovik, G.. 2002. Particle Filters for State-Space Models With the Presence of Uknown Static Parameters. *IEEE Transactions on Signal Processing 50*(2), 281–289.

Sun, Y., P. Babu, and D. P. Palomar. 2016. Robust Estimation of Structured Covariance Matrix for Heavy-Tailed Elliptical Distributions. *IEEE Transactions on Signal Processing 64*(14), 3576–3590.

Tanizaki, H.. 1993. *Nonlinear Filters Estimation and Applications*. Berlin: Springer-Verlag Berlin Heidelberg.

Tanizaki, H. and R. S. Mariano. 1996. Nonlinear Filters Based on Taylor Series Expansions. *Communications in Statistics - Theory and Methods 25*(6), 1261–1282.

Tanizaki, H. and R. S. Mariano. 1998. Nonlinear and non-Gaussian State Space Modelling with Monte Carlo Simulations. *Journal of Econometrics 83*(1-2), 263–290.

Teichmann, J. and M. V. Wuthrich. 2016. Consistent Yield Curve Prediction. *ASTIN Bulletin: The Journal of IAA 46*(2), 191–224.

Thornton, C. L.. 1976. *Triangular Covariance Factorizations for Kalman Filtering*. Ph. D. thesis.

Thornton, C. L. and G. J. Bierman. 1980. $UDUˆT$ Covariance Factorization for Kalman Filtering. *Control and Dynamic Systems 16*, 177–248.

Tibshirani, R.. 1996. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological) 58*(1), 267–288.

Tierney, L.. 1994. Markov Chain For Exploring Posterior Distributions. *The Annals of Statistics 22*(4), 1701–1728.

Tierney, L., R. Kass, and J. Kadane. 1989. Fully Exponential Laplace Approximations to Expectations and Variances of Nonpositive Functions. *Journal of the American Statistical Association 84*(407), 710–716.

Tipping, M. E. and C. M. Bishop. 1999a. Mixtures of Probabilistic Principal Component Analysers. *Neural Computation 11*(2), 443–482.

Tipping, M. E. and C. M. Bishop. 1999b. Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society. Series B (Statistical Methodology) 61*(3), 622–661.

Toczydlowska, D. and G. W. Peters. 2018. Financial Big Data Solutions for State Space Panel Regression in Interest Rate Dynamics. *Econometrics 6*(3), 1–45.

Toczydlowska, D., G. W. Peters, M. C. Fung, and P. V. Shevchenko. 2017. Stochastic Period and Cohort Effect State-Space Mortality Models Incorporating Demographic Factors via Probabilistic Robust Principal Components. *Risks 5*(3).

Tuan, P. D.. 1971. On the Remainder in Quadrature Rules. *Mathematics of Computation 25*(11), 819–826.

Tyler, D. E.. 1987a. A Distribution-Free M-Estimator of Multivariate Scatter. *The Annals of Statistics 15*(1), 234–251.

Tyler, D. E.. 1987b. Statistical Analysis for the Angular Central Gaussian Distribution on the Sphere. *Biometrika 74*(3), 579–589.

Ukkelberg, A. and O. Borgen. 1993. Outlier Detection by Robust Alternating Regression. *Analytica Chimica Acta 277*(2), 489–494.

van der Merwe, R.. 2004. *Sigma-Point Kalman Filters for Probabilistic Inference in Dynamic State-Space Models*. Ph. D. thesis, Oregon Health & Science University.

van der Merwe, R., A. Doucet, N. de Freitas, and E. Wan. 2000. The Unscented Particle Filter. In T. K. Leen, T. G. Dietterich, and V. Tresp (Eds.), *13th International Conference on Neural Information Processing Systems*, Cambridge, pp. 563–569. MIT Press Cambridge.

von Neumann, J.. 1951. Various Techniques Used in Connection with Random Digits. In A. S. Householder, G. E. Forsyth, and H. H. Germond (Eds.), *Monte Carlo Method*, Volume 12, pp. 36–38. National Bureau of Standards Applied Mathematics Series.

Wan, E. A. and R. van der Merwe. 2000. The Unscented Kalman Filter for Nonlinear Estimation. In *Adaptive Systems for Signal Processing, Communications, and Control Symposium 2000. AS-SPCC. The IEEE 2000*, Lake Louise, Alberta, pp. 153–158.

Wasilkowski, G. W. and H. Wozniakowski. 1995. Explicit Cost Bounds of Algorithms for Multivariate Tensor Products Problems. *Journal of Complexity 11*(1), 1–56.

West, M. and J. Harrison. 1997. *Bayesian Forecasting and Dynamic Models*. New York: Springer Verlag, New York.

Wiberg, T.. 1976. Computation of Principal Components When Data Are Missing. In *Proceedings of the Second Symposium of Computational Statistics*, Berlin, pp. 229–236.

Wiener, N. and E. Hopf. 1931. On a Class of Singular Integral Equations. In *Proc. Prussian Acad. Math*, pp. 696.

Wild, P. and W. R. Gilks. 1993. Algorithm AS 287: Adaptive Rejection Sampling from Log-Concave Density Functions. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 42*(4), 701–709.

Willets, R. C.. 2004. The Cohort Effect: Insights and Explanations. *British Actuarial Journal 10*(4), 833–877.

Wilmoth, J. R., K. Andreev, and D. Jdanov. 2007. Methods Protocol for the Human Mortality Database. Technical Report June 2000.

Wonham, W. M.. 1965. Some Applications of Stochastic Differential Equations to Optimal Nonlinear Filtering. *Journal of the Society for Industrial and Applied Mathematics Series A Control 2*(3), 347–369.

Wright, J., A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. 2009. Robust Face Recognition via Sparse Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence 31*(2), 210–227.

Wright, J. H.. 2011. Term Premia and Inflation Uncertainty: Empirical Evidence from an International Panel Dataset. *American Economic Review 101*(4), 1514–1534.

Wu, C. F. J.. 1983. On the Convergence properties of the EM Algorithm. *The Annals of Statistics 11*(1), 95–103.

Xie, P. and E. Xing. 2015. Cauchy Principal Component Analysis. *Available at arXiv preprint arXiv:1412.6506*, 1–14.

Xu, H., C. Caramanis, and S. Sanghavi. 2012. Robust PCA via Outlier Pursuit. *IEEE Transactions on Information Theory 58*(5), 3047–3064.

Xu, L. and A. L. Yuille. 1995. Robust Principal Component Analysis by Self-Organizing Rules Based on Statistical Physics Approach. *IEEE Transactions on Neural Networks 6*(1), 131–143.

Young, T. . and T. W. Calvert. 1974. *Classification, Estimation and Pattern Recognition*. Oxford: Elsevier Science and Technology.

Zadeh, L. A.. 1953. Optimum Nonlinear Filters. *Journal of Applied Physics 24*(4), 396–404.

Zadeh, L. A. and C. A. Desoer. 1963. *Linear System Theory: The State Space Approach*. McGraw-Hill Book Company.

Zakai, M.. 1969. On the Optimal Filtering of Diffusion Processes. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 11*(3), 230–243.

Zaritskii, V., V. Svetnik, and L. Shimelevich. 1976. Monte Carlo Technique in Problems of Optimal Information Processing. *Automation and Remote Control 36*(12), 2015–2022.

Zhao, D., Z. Lin, and X. Tang. 2007. Laplacian PCA and Its Applications. In *IEEE 11th International Conference on Computer Vision*, Rio de Janeiro, pp. 8.

Zhou, Z., X. Li, J. Wright, E. Candes, and Y. Ma. 2010. Stable Principal Component Pursuit. In *IEEE International Symposium on Information Theory*, Austin, pp. 5.

Appendix

A. Proofs of Steps of EM Algorithms

The following equalities are used in all proofs presented in the following chapter

$$\frac{\partial \mathbf{A}\mathbf{W}^T\mathbf{B}}{\partial \mathbf{W}} = \mathbf{B}\mathbf{A} \qquad \frac{\partial \mathbf{A}\mathbf{W}\mathbf{B}}{\partial \mathbf{W}} = \mathbf{A}^T\mathbf{B}^T$$
$$\frac{\partial \operatorname{Tr}\left\{\mathbf{A}\mathbf{W}^T\right\}}{\partial \mathbf{W}} = \mathbf{A} \qquad \frac{\partial \operatorname{Tr}\left\{\mathbf{A}\mathbf{W}^T\mathbf{W}\right\}}{\partial \mathbf{W}} = \mathbf{W}\mathbf{A} + \mathbf{W}\mathbf{A}^T.$$

A.1. Supporting Lemmas

Lemma A1. *Let* $\mathbf{X}$ *be a d-dimensional random vectors such that* $\mathbf{X} = \left[\mathbf{X}_1, \mathbf{X}_2\right]$ *for* d_1*-dimensional subvector* $\mathbf{X}_1$ *and* d_2*-dimensional subvector* $\mathbf{X}_2$ *,* $d_1 + d_2 = d$*. If* $\mathbf{X}$ *follows multivariate Gaussian distribution, that is*

$$\mathbf{X} = \left[\mathbf{X}_1, \mathbf{X}_2\right] \sim \mathcal{N}\left(\left[\boldsymbol{\mu}_1, \boldsymbol{\mu}_2\right], \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right)$$

then for $i, j = 1, 2$ *and* $i \neq j$*, we have the following*

1. $\mathbf{X}_i | \mathbf{X}_j \sim \mathcal{N}\left(\boldsymbol{\mu}_i + \left[\mathbf{X}_j - \boldsymbol{\mu}_j\right]\boldsymbol{\Sigma}_{jj}^{-1}\boldsymbol{\Sigma}_{ji}, \boldsymbol{\Sigma}_{ii} - \boldsymbol{\Sigma}_{ij}\boldsymbol{\Sigma}_{jj}^{-1}\boldsymbol{\Sigma}_{ji}\right)$
2. $\mathbf{X}_i \sim \mathcal{N}\left(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_{ii}\right)$

where $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ *are* d_1*- and* d_2*-dimensional real valued vectors, respectively,* $\boldsymbol{\Sigma}_{11}$ *and* $\boldsymbol{\Sigma}_{11}$ *are* $d_1 \times d_1$ *and* $d_2 \times d_2$ *symmetric, positive-definite real valued matrices, and* $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}^T$ *is a* $d_1 \times d_2$ *real valued matrix.*

Proof. Please refer to Theorem 2.3.12 in (Gupta 2003). □

A.2. Gaussian Probabilistic Principal Component Analysis

A.2.1. Proof of Theorem 4

Proof. The E-step (2.14) can be written in terms of integrals as

$$\begin{aligned}
Q(\Psi, \Psi^*) &= \mathbb{E}_{\mathbf{X}_{1:N}|\mathbf{Y}_{1:N},\Psi}\left[\log \pi_{\mathbf{Y}_{1:N},\mathbf{X}_{1:N}|\Psi^*}(\mathbf{y}_{1:N}, \mathbf{x}_{1:N})\right] = \int_{\mathbb{R}^{N\times k}} \log \pi_{\mathbf{Y}_{1:N},\mathbf{X}_{1:N}|\Psi^*}(\mathbf{y}_{1:N}, \mathbf{x}_{1:N})\, \pi_{\mathbf{X}_{1:N}|\mathbf{Y}_{1:N},\Psi}(\mathbf{x}_{1:N})\, d\mathbf{x}_{1:N} \\
&= \int_{\mathbb{R}^{N\times k}} \log\left(\prod_{t=1}^{N} \pi_{\mathbf{Y}_t,\mathbf{X}_t|\Psi^*}(\mathbf{y}_t, \mathbf{x}_t)\right) \prod_{s=1}^{N} \pi_{\mathbf{X}_s|\mathbf{Y}_s,\Psi}(\mathbf{x}_s)\, d\mathbf{x}_{1:N} = \sum_{t=1}^{N}\left\{\int_{\mathbb{R}^{N\times k}} \log\left(\pi_{\mathbf{Y}_t,\mathbf{X}_t|\Psi^*}(\mathbf{y}_t, \mathbf{x}_t)\right) \prod_{s=1}^{N} \pi_{\mathbf{X}_s|\mathbf{Y}_s,\Psi}(\mathbf{x}_s)\, d\mathbf{x}_{1:N}\right\} \\
&= \sum_{t=1}^{N}\left\{\int_{\mathbb{R}^{k}} \log\left(\pi_{\mathbf{Y}_t,\mathbf{X}_t|\Psi^*}(\mathbf{y}_t, \mathbf{x}_t)\right) \pi_{\mathbf{X}_t|\mathbf{Y}_t,\Psi}(\mathbf{x}_t)\, d\mathbf{x}_t \underbrace{\int_{\mathbb{R}^{(N-1)\times k}} \left(\prod_{s=1,\, s\neq t}^{N} \pi_{\mathbf{X}_s|\mathbf{Y}_s,\Psi}(\mathbf{x}_s)\right) d\mathbf{x}_{1:N}}_{N-1 \text{ without } t \text{ element}}\right\} \\
&= \sum_{t=1}^{N}\left(\int_{\mathbb{R}^{k}} \log\left(\pi_{\mathbf{Y}_t,\mathbf{X}_t|\Psi^*}(\mathbf{y}_t, \mathbf{x}_t)\right) \pi_{\mathbf{X}_t|\mathbf{Y}_t,\Psi}(\mathbf{x}_t)\, d\mathbf{x}_t \underbrace{\prod_{s=1,\, s\neq t}^{N} \int_{\mathbb{R}^{k}} \pi_{\mathbf{X}_s|\mathbf{Y}_s,\Psi}(\mathbf{x}_s)\, d\mathbf{x}_s}_{=1 \text{ since a density}}\right) \\
&= \sum_{t=1}^{N} \int_{\mathbb{R}^{k}} \log\left(\pi_{\mathbf{Y}_t,\mathbf{X}_t|\Psi^*}(\mathbf{y}_t, \mathbf{x}_t)\right) \pi_{\mathbf{X}_t|\mathbf{Y}_t,\Psi}(\mathbf{x}_t)\, d\mathbf{x}_t
\end{aligned}$$

Let us recall the decomposition of the joint probability density function from Equation (2.5) and the distribution of $\mathbf{Y}_t$ conditioned on $\mathbf{X}_t$. Finding the expectation (2.14) requires specifying the conditional distribution of $\mathbf{X}_t$ given observations $\mathbf{Y}_t$ and parameters. It is given via Bayes' rule as

$$\pi(\mathbf{X}_t|\mathbf{Y}_t,\Psi) = \frac{\pi(\mathbf{Y}_t,\mathbf{X}_t|\Psi)}{\pi(\mathbf{Y}_t|\Psi)} = \frac{\pi(\mathbf{Y}_t|\mathbf{X}_t,\Psi)\pi(\mathbf{X}_t|\Psi)}{\pi(\mathbf{Y}_t|\Psi)}$$

and results in $\mathbf{X}_t|\mathbf{Y}_t,\Psi \sim \mathcal{N}\left((\mathbf{Y}_t-\boldsymbol{\mu})\,\mathbf{W}\mathbf{M}^{-1},\sigma^2\mathbf{M}^{-1}\right)$ for $\mathbf{M} = \mathbf{W}^T\mathbf{W}+\sigma^2\mathbb{I}_k$ as shown in (Rubin and Thayer 1982). Therefore, the E-step from Equation (2.14) becomes

$$\begin{aligned}
Q(\Psi,\Psi^*) &= -\frac{1}{2}\sum_{t=1}^{N}\Bigg\{(d+k)\log 2\pi + d\log\sigma^{*2} + \frac{1}{\sigma^{*2}}\left(\mathbf{y}_t-\boldsymbol{\mu}^*\right)\left(\mathbf{y}_t-\boldsymbol{\mu}^*\right)^T \\
&\quad - \frac{2}{\sigma^{*2}}\mathbb{E}_{\mathbf{X}_t|\mathbf{Y}_t,\Psi}\left[\mathbf{X}_t\right]\mathbf{W}^{*T}\left(\mathbf{y}_t-\boldsymbol{\mu}^*\right)^T + \operatorname{Tr}\left\{\left(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{W}^*+\mathbb{I}_k\right)\mathbb{E}_{\mathbf{X}_t|\mathbf{Y}_t,\Psi}\left[\mathbf{X}_t^T\mathbf{X}_t\right]\right\}\Bigg\} \\
&= -\frac{1}{2}\sum_{t=1}^{N}\Bigg\{(d+k)\log 2\pi + d\log\sigma^{*2} + \frac{1}{\sigma^{*2}}\Bigg(\operatorname{Tr}\left\{\mathbf{y}_t^T\mathbf{y}_t\right\} - 2\mathbf{y}_t\boldsymbol{\mu}^{*T} + \boldsymbol{\mu}^*\boldsymbol{\mu}^{*T} \\
&\quad - 2\operatorname{Tr}\left\{\mathbf{W}^*\mathbb{E}_{\mathbf{X}_t|\mathbf{Y}_t,\Psi}\left[\mathbf{X}_t\right]^T\mathbf{y}_t\right\} + 2\mathbb{E}_{\mathbf{X}_t|\mathbf{Y}_t,\Psi}\left[\mathbf{X}_t\right]\mathbf{W}^{*T}\boldsymbol{\mu}^{*T}\Bigg) + \operatorname{Tr}\left\{\left(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{W}^*+\mathbb{I}_k\right)\mathbb{E}_{\mathbf{X}_t|\mathbf{Y}_t,\Psi}\left[\mathbf{X}_t^T\mathbf{X}_t\right]\right\}\Bigg\}
\end{aligned}$$

for the corresponding moments of the conditional distribution $\mathbf{X}_t|\mathbf{Y}_t,\Psi$

$$\begin{aligned}
\mathbb{E}_{\mathbf{X}_t|\mathbf{Y}_t,\Psi}\left[\mathbf{X}_t\right]_{1\times k} &= \left(\mathbf{y}_t-\boldsymbol{\mu}\right)\mathbf{W}\mathbf{M}^{-1}, \\
\mathbb{E}_{\mathbf{X}_t|\mathbf{Y}_t,\Psi}\left[\mathbf{X}_t^T\mathbf{X}_t\right]_{k\times k} &= \sigma^2\mathbf{M}^{-1} + \mathbf{M}^{-1}\mathbf{W}^T\left(\mathbf{y}_t-\boldsymbol{\mu}\right)^T\left(\mathbf{y}_t-\boldsymbol{\mu}\right)\mathbf{W}\mathbf{M}^{-1}.
\end{aligned}$$

□

A.2.2. Proof of Theorem 5

Proof. The maximizers of the function Q defined in Equation (2.14) and specified in Theorem 4 are obtained as the solutions to the system of equations specified by $\nabla_{\Psi^*}Q(\Psi,\Psi^*) = \mathbf{0}$ and are the following

$$\begin{aligned}
&\frac{\partial Q(\Psi,\Psi^*)}{\partial\boldsymbol{\mu}^*} = \frac{1}{2\sigma^*}\sum_{t=1}^{N}\left\{2\left(\mathbf{y}_t-\boldsymbol{\mu}^*\right) - 2\mathbb{E}_{\mathbf{X}_t|\mathbf{Y}_t,\Psi}\left[\mathbf{X}_t\right]\mathbf{W}^{*T}\right\} = \mathbf{0} \\
&\quad\iff \sum_{t=1}^{N}\left\{\mathbf{y}_t-\boldsymbol{\mu}^* - \left(\mathbf{y}_t-\boldsymbol{\mu}\right)\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\right\} = \mathbf{0} \\
&\quad\iff \left(\sum_{t=1}^{N}\mathbf{y}_t\right)\left(\mathbb{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\right) - N\boldsymbol{\mu}^* + N\boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T} = \mathbf{0} \\
&\quad\iff \boldsymbol{\mu}^* = \bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N};\Psi)\left(\mathbb{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\right) + \boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}
\end{aligned}$$

$$\begin{aligned}
&\frac{\partial Q\left(\Psi,\Psi^*\right)}{\partial\mathbf{W}^*} = \frac{-1}{2\sigma^{*2}}\sum_{t=1}^{N}\left\{-2\left(\mathbf{y}_t-\boldsymbol{\mu}^*\right)^T\mathbb{E}_{\mathbf{X}_t|\mathbf{Y}_t,\Psi}\left[\mathbf{X}_t\right] + 2\mathbf{W}^*\mathbb{E}_{\mathbf{X}_t|\mathbf{Y}_t,\Psi}\left[\mathbf{X}_t^T\mathbf{X}_t\right]\right\} = \mathbf{0} \\
&\quad\iff \sum_{t=1}^{N}\left\{\left(\mathbf{y}_t-\boldsymbol{\mu}^*\right)^T\left(\mathbf{y}_t-\boldsymbol{\mu}\right)\mathbf{W}\mathbf{M}^{-1} - \mathbf{W}^*\left(\sigma^2\mathbf{M}^{-1} + \mathbf{M}^{-1}\mathbf{W}^T\left(\mathbf{y}_t-\boldsymbol{\mu}\right)^T\left(\mathbf{y}_t-\boldsymbol{\mu}\right)\mathbf{W}\mathbf{M}^{-1}\right)\right\} = \mathbf{0} \\
&\quad\iff \mathbf{W}^* = \bar{\mathbf{C}}_{\mu,\mu^*}(\mathbf{y}_{1:N};\Psi,\Psi^*)\mathbf{W}\mathbf{M}^{-1}\left(\sigma^2\mathbf{M}^{-1} + \mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\mu}(\mathbf{y}_{1:N};\Psi)\mathbf{W}\mathbf{M}^{-1}\right)^{-1}
\end{aligned}$$

$$\begin{aligned}
&\frac{\partial Q\left(\Psi,\Psi^*\right)}{\partial\sigma^{*2}} = -\frac{1}{2}\sum_{t=1}^{N}\Bigg\{\frac{d}{\sigma^{*2}} - \frac{1}{(\sigma^{*2})^2}\Bigg(\operatorname{Tr}\left\{\left(\mathbf{y}_t-\boldsymbol{\mu}^*\right)^T\left(\mathbf{y}_t-\boldsymbol{\mu}^*\right)\right\} \\
&\qquad - 2\operatorname{Tr}\left\{\mathbf{W}^{*T}\mathbb{E}_{\mathbf{X}_t|\mathbf{Y}_t,\Psi}\left[\mathbf{X}_t\right)^T\left(\mathbf{y}_t-\boldsymbol{\mu}^*\right]\right\} + \operatorname{Tr}\left\{\mathbf{W}^{*T}\mathbf{W}^*\mathbb{E}_{\mathbf{X}_t|\mathbf{Y}_t,\Psi}\left[\mathbf{X}_t^T\mathbf{X}_t\right]\right\}\Bigg)\Bigg\} = 0 \\
&\quad\iff \sum_{t=1}^{N}\Bigg\{\frac{d}{\sigma^{*2}} - \frac{1}{(\sigma^{*2})^2}\Bigg(\operatorname{Tr}\left\{\left(\mathbf{y}_t-\boldsymbol{\mu}^*\right)^T\left(\mathbf{y}_t-\boldsymbol{\mu}^*\right)\right\} - 2\operatorname{Tr}\left\{\mathbf{W}^{*T}\mathbf{M}^{-1}\mathbf{W}^T\left(\mathbf{y}_t-\boldsymbol{\mu}\right)^T\left(\mathbf{y}_t-\boldsymbol{\mu}^*\right)\right\} \\
&\qquad + \operatorname{Tr}\left\{\mathbf{W}^{*T}\mathbf{W}^*\left(\sigma^2\mathbf{M}^{-1} + \mathbf{M}^{-1}\mathbf{W}^T\left(\mathbf{y}_t-\boldsymbol{\mu}\right)^T\left(\mathbf{y}_t-\boldsymbol{\mu}\right)\mathbf{W}\mathbf{M}^{-1}\right)\right\}\Bigg)\Bigg\} = 0 \\
&\quad\iff \sigma^{*2} = \frac{1}{d}\operatorname{Tr}\left\{\bar{\mathbf{C}}_{\mu^*}(\mathbf{y}_{1:N};\Psi^*) - 2\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\mu,\mu^*}(\mathbf{y}_{1:N};\Psi,\Psi^*) + \mathbf{W}^{*T}\mathbf{W}^*\left(\sigma^2\mathbf{M}^{-1} + \mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\mu}(\mathbf{y}_{1:N};\Psi)\mathbf{W}\mathbf{M}^{-1}\right)\right\}
\end{aligned}$$

where

$$\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N};\Psi)_{1\times d} = \frac{1}{N}\sum_{t=1}^{N}\mathbf{y}_t,$$
$$\bar{\mathbf{S}}(\mathbf{y}_{1:N};\Psi)_{d\times d} = \frac{1}{N}\sum_{t=1}^{N}\mathbf{y}_t^T\mathbf{y}_t,$$
$$\bar{\mathbf{C}}_{\boldsymbol{\mu}}(\mathbf{y}_{1:N};\Psi)_{d\times d} = \bar{\mathbf{S}}(\mathbf{y}_{1:N};\Psi) - \boldsymbol{\mu}^T\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N};\Psi) - \bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N};\Psi)^T\boldsymbol{\mu} + \boldsymbol{\mu}^T\boldsymbol{\mu},$$
$$\bar{\mathbf{C}}_{\boldsymbol{\mu}^*}(\mathbf{y}_{1:N};\Psi^*)_{d\times d} = \bar{\mathbf{S}}(\mathbf{y}_{1:N};\Psi) - \boldsymbol{\mu}^{*T}\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N};\Psi) - \bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N};\Psi)^T\boldsymbol{\mu}^* + \boldsymbol{\mu}^{*T}\boldsymbol{\mu}^*,$$
$$\bar{\mathbf{C}}_{\boldsymbol{\mu},\boldsymbol{\mu}^*}(\mathbf{y}_{1:N};\Psi,\Psi^*)_{d\times d} = \bar{\mathbf{S}}(\mathbf{y}_{1:N};\Psi) - \boldsymbol{\mu}^{*T}\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N};\Psi) - \bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N};\Psi)^T\boldsymbol{\mu} + \boldsymbol{\mu}^{*T}\boldsymbol{\mu}.$$

□

A.2.3. Proof of Theorem 19

Proof. The steps of proof of Theorem 19 follows the reasoning of the proof of Theorem 4 but in addition we consider the distribution of the vector with missing observation $\mathbf{Y}_t^m$, given the partition of the vector , $\mathbf{Y}_t = [\mathbf{Y}_t^o, \mathbf{Y}_t^m]$, that is

$$\begin{aligned}
\tilde{Q}(\Psi,\Psi^*) =& \ \mathbb{E}_{\mathbf{Y}_{1:N}^m,\mathbf{X}_{1:N}|\mathbf{Y}_{1:N}^o,\Psi}\left[\log \pi_{\mathbf{Y}_{1:N},\mathbf{X}_{1:N}|\Psi^*}(\mathbf{Y}_{1:N},\mathbf{X}_{1:N})\right] \\
=& \int_{\mathbb{R}^{N\times d_m}}\int_{\mathbb{R}^{N\times k}} \log\left(\pi_{\mathbf{Y}_{1:N},\mathbf{X}_{1:N}|\Psi^*}(\mathbf{y}_{1:N},\mathbf{x}_{1:N})\right)\pi_{\mathbf{Y}_{1:N}^m,\mathbf{X}_{1:N}|\mathbf{Y}_{1:N}^o,\Psi}(\mathbf{y}_{1:N}^m,\mathbf{x}_{1:N})\,d\mathbf{x}_{1:N}\,d\mathbf{y}_{1:N}^m \\
=& \sum_{t=1}^{N}\left\{\int_{\mathbb{R}^{d_m}}\int_{\mathbb{R}^k}\log\left(\pi_{\mathbf{Y}_t,\mathbf{X}_t|\Psi^*}(\mathbf{y}_t,\mathbf{x}_t)\right)\pi_{\mathbf{Y}_t^m,\mathbf{X}_t|\mathbf{Y}_t^o,\Psi}(\mathbf{x}_t,\mathbf{y}_t^m)\,d\mathbf{x}_t\,d\mathbf{y}_t^m\right\} \\
=& -\frac{1}{2}\sum_{t=1}^{N}\left\{(d+k)\log 2\pi + d\log\sigma^{*2} + \frac{1}{\sigma^{*2}}\mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t|\mathbf{Y}_t^o,\Psi}\left[(\mathbf{Y}_t-\boldsymbol{\mu}^*)(\mathbf{Y}_t-\boldsymbol{\mu}^*)^T\right]\right. \\
&\left. - \frac{2}{\sigma^{*2}}\mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t|\mathbf{Y}_t^o,\Psi}\left[\mathbf{X}_t\mathbf{W}^{*T}(\mathbf{Y}_t-\boldsymbol{\mu}^*)^T\right] + \mathrm{Tr}\left\{\left(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{W}^* + \mathbb{I}_k\right)\mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t|\mathbf{Y}_t^o,\Psi}\left[\mathbf{X}_t^T\mathbf{X}_t\right]\right\}\right\} \\
=& -\frac{1}{2}\sum_{t=1}^{N}\left\{(d+k)\log 2\pi + d\log\sigma^{*2} + \frac{1}{\sigma^{*2}}\,\mathrm{Tr}\left\{\mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t|\mathbf{Y}_t^o,\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right]\right\} - \frac{2}{\sigma^{*2}}\mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t|\mathbf{Y}_t^o,\Psi}\left[\mathbf{Y}_t\right]\boldsymbol{\mu}^{*T}\right. \\
&\left. + \frac{1}{\sigma^{*2}}\boldsymbol{\mu}^*\boldsymbol{\mu}^{*T} - \frac{2}{\sigma^{*2}}\,\mathrm{Tr}\left\{\mathbf{W}^*\mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t|\mathbf{Y}_t^o,\Psi}\left[\mathbf{X}_t^T\mathbf{Y}_t\right]\right\} + \frac{2}{\sigma^{*2}}\mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t|\mathbf{Y}_t^o,\Psi}\left[\mathbf{X}_t\right]\mathbf{W}^{*T}\boldsymbol{\mu}^{*T} + \mathrm{Tr}\left\{\left(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{W}^* + \mathbb{I}_k\right)\mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t|\mathbf{Y}_t^o,\Psi}\left[\mathbf{X}_t^T\mathbf{X}_t\right]\right\}\right\}.
\end{aligned}$$

In order to specify the moments of the conditional distribution $\mathbf{Y}_t^m, \mathbf{X}_t|\mathbf{Y}_t^o, \Psi$, we apply the chain rule of probabilities and recall that

$$\pi_{\mathbf{Y}_t^m,\mathbf{X}_t|\mathbf{Y}_t^o,\Psi}(\mathbf{y}_t^m,\mathbf{x}_t) = \pi_{\mathbf{X}_t|\mathbf{Y}_t,\Psi}(\mathbf{x}_t)\ \pi_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\Psi}(\mathbf{y}_t^m)$$

The moments of the distribution $\mathbf{Y}_t^m|\mathbf{Y}_t^o, \Psi$ in the Gaussian PPCA setting can be obtained straightforwardly as $\mathbf{Y}_t$ is assumed to be Gaussian as shown in Lemma A1. The conditional distribution of $\mathbf{Y}_t^m|\mathbf{Y}_t^o, \Psi$ is the following

$$\mathbf{Y}_t^m|\mathbf{Y}_t^o,\Psi \sim \mathcal{N}\left(\boldsymbol{\mu}_m + (\mathbf{Y}_t^o-\boldsymbol{\mu}_o)\left(\mathbf{W}_o\mathbf{W}_o^T + \sigma^2\mathbb{I}_{d_o}\right)^{-1}\mathbf{W}_o\mathbf{W}_m^T,\ \mathbf{C}_{mm} - \mathbf{W}_m\mathbf{W}_o^T\mathbf{C}_{oo}^{-1}\mathbf{W}_o\mathbf{W}_m^T\right)$$

and given the definition of the vector $\mathbf{Y}_t = [\mathbf{Y}_t^o, \mathbf{Y}_t^m]$ we obtain the following moments

$$\mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t|\mathbf{Y}_t^o,\Psi}\left[\mathbf{Y}_t\right]_{1\times d} = \mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\Psi}\left[\mathbf{Y}_t\right] = \begin{bmatrix}\mathbf{Y}_t^o \\ \boldsymbol{\mu}_m + (\mathbf{Y}_t^o-\boldsymbol{\mu}_o)\left(\mathbf{W}_o\mathbf{W}_o^T + \sigma^2\mathbb{I}_{d_o}\right)^{-1}\mathbf{W}_o\mathbf{W}_m^T\end{bmatrix}$$
$$\mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t|\mathbf{Y}_t^o,\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right]_{d\times d} = \mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right] = \begin{bmatrix}\mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{mm} - \mathbf{W}_m\mathbf{W}_o^T\mathbf{C}_{oo}^{-1}\mathbf{W}_o\mathbf{W}_m^T\end{bmatrix} + \mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t|\mathbf{Y}_t^o,\Psi}\left[\mathbf{Y}_t\right]^T\mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t|\mathbf{Y}_t^o,\Psi}\left[\mathbf{Y}_t\right]$$

what implies that

$$\mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t|\mathbf{Y}_t^o,\Psi}\left[\mathbf{X}_t\right]_{1\times k} = \mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\Psi}\left[\mathbb{E}_{\mathbf{X}_t|\mathbf{Y}_t,\Psi}\left[\mathbf{X}_t\right]\right] = \left(\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\Psi}\left[\mathbf{Y}_t\right] - \boldsymbol{\mu}\right)\mathbf{W}\mathbf{M}^{-1},$$

$$\begin{aligned}\mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t|\mathbf{Y}_t^o,\Psi}\left[\mathbf{X}_t^T\mathbf{X}_t\right]_{k\times k} &= \mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\Psi}\left[\mathbb{E}_{\mathbf{X}_t|\mathbf{Y}_t,\Psi}\left[\mathbf{X}_t^T\mathbf{X}_t\right]\right] = \sigma^2\mathbf{M}^{-1} + \mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\Psi}\left[\mathbb{E}_{\mathbf{X}_t|\mathbf{Y}_t,\Psi}\left[\mathbf{X}_t\right]^T\mathbb{E}_{\mathbf{X}_t|\mathbf{Y}_t,\Psi}\left[\mathbf{X}_t\right]\right] \\ &= \sigma^2\mathbf{M}^{-1} + \mathbf{M}^{-1}\mathbf{W}^T\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\Psi}\left[\left(\mathbf{Y}_t-\boldsymbol{\mu}\right)^T\left(\mathbf{Y}_t-\boldsymbol{\mu}\right)\right]\mathbf{W}\mathbf{M}^{-1},\end{aligned}$$

$$\begin{aligned}\mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t|\mathbf{Y}_t^o,\Psi}\left[\mathbf{X}_t^T\mathbf{Y}_t\right]_{k\times d} &= \mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\Psi}\left[\mathbb{E}_{\mathbf{X}_t|\mathbf{Y}_t,\Psi}\left[\mathbf{X}_t\right]\mathbf{Y}_t\right] = \left(\mathbf{W}^T\mathbf{W}+\sigma^2\mathbb{I}_k\right)^{-1}\mathbf{W}^T\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\Psi}\left[\left(\mathbf{Y}_t-\boldsymbol{\mu}\right)^T\mathbf{Y}_t\right] \\ &= \left(\mathbf{W}^T\mathbf{W}+\sigma^2\mathbb{I}_k\right)^{-1}\mathbf{W}^T\left(\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right] - \boldsymbol{\mu}^T\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\Psi}\left[\mathbf{Y}_t\right]\right).\end{aligned}$$

□

A.2.4. Proof of Theorem 20

Proof. The steps of proof of Theorem 20 follows the reasoning of the proof of Theorem 5 but the introduction of missingness to the PPCA setting requires applying the moments of $\mathbf{Y}_t^m, \mathbf{X}_t|\mathbf{Y}_t^o, \Psi$ derived in Theorem 19. The maximizers of the function Q^m defined in Theorem 19 are obtained as the solutions to the system of equations specified by $\nabla_{\Psi^*}\tilde{Q}(\Psi,\Psi^*) = \mathbf{0}$ and are the following

$$\begin{aligned}\frac{\partial\tilde{Q}(\Psi,\Psi^*)}{\partial\boldsymbol{\mu}^*} &= \frac{1}{2\sigma^*}\sum_{t=1}^{N}\left\{2\left(\mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t|\mathbf{Y}_t^o,\Psi}\left[\mathbf{Y}_t\right]-\boldsymbol{\mu}^*\right) - 2\mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t|\mathbf{Y}_t^o,\Psi}\left[\mathbf{X}_t\right]\mathbf{W}^{*T}\right\} = \mathbf{0} \\ &\iff \sum_{t=1}^{N}\left\{\mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t|\mathbf{Y}_t^o,\Psi}\left[\mathbf{Y}_t\right]-\boldsymbol{\mu}^* - \left(\mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t|\mathbf{Y}_t^o,\Psi}\left[\mathbf{Y}_t\right]-\boldsymbol{\mu}\right)\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\right\} = \mathbf{0} \\ &\iff \boldsymbol{\mu}^* = \bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o;\Psi)\left(\mathbb{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\right) + \boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\end{aligned}$$

$$\begin{aligned}\frac{\partial\tilde{Q}(\Psi,\Psi^*)}{\partial\mathbf{W}^*} &= -\frac{1}{2\sigma^{*2}}\sum_{t=1}^{N}\left\{-2\mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t|\mathbf{Y}_t^o,\Psi}\left[\mathbf{Y}_t^T\mathbf{X}_t\right]\right\} + 2\boldsymbol{\mu}^{*T}\mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t|\mathbf{Y}_t^o,\Psi}\left[\mathbf{X}_t\right] + 2\mathbf{W}^*\mathbb{E}_{\mathbf{X}_t|\mathbf{Y}_t,\Psi}\left[\mathbf{X}_t^T\mathbf{X}_t\right]\Big\} = \mathbf{0} \\ &\iff \mathbf{W}^* = \bar{\mathbf{C}}_{\boldsymbol{\mu},\boldsymbol{\mu}^*}^m(\mathbf{y}_{1:N}^o;\Psi,\Psi^*)\mathbf{W}\mathbf{M}^{-1}\left(\sigma^2\mathbf{M}^{-1} + \mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}\mathbf{M}^{-1}\right)^{-1}\end{aligned}$$

$$\begin{aligned}\frac{\partial\tilde{Q}(\Psi,\Psi^*)}{\partial\sigma^{*2}} &= -\frac{1}{2}\sum_{t=1}^{N}\left\{\frac{d}{\sigma^{*2}} - \frac{1}{(\sigma^{*2})^2}\left(\operatorname{Tr}\left\{\mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t|\mathbf{Y}_t^o,\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right]\right\} - 2\mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t|\mathbf{Y}_t^o,\Psi}\left[\mathbf{Y}_t\right]\boldsymbol{\mu}^{*T} + \boldsymbol{\mu}^*\boldsymbol{\mu}^{*T}\right.\right. \\ &\left.\left. - 2\operatorname{Tr}\left\{\mathbf{W}^*\mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t|\mathbf{Y}_t^o,\Psi}\left[\mathbf{X}_t^T\mathbf{Y}_t\right]\right\} + 2\mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t|\mathbf{Y}_t^o,\Psi}\left[\mathbf{X}_t\right]\mathbf{W}^{*T}\boldsymbol{\mu}^{*T} + \operatorname{Tr}\left\{\mathbf{W}^{*T}\mathbf{W}^*\mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t|\mathbf{Y}_t^o,\Psi}\left[\mathbf{X}_t^T\mathbf{X}_t\right]\right\}\right)\right\} = 0 \\ &\iff \sigma^{*2} = \frac{1}{d}\operatorname{Tr}\left\{\bar{\mathbf{C}}_{\boldsymbol{\mu}^*}^m(\mathbf{y}_{1:N}^o;\Psi^*) - 2\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\boldsymbol{\mu},\boldsymbol{\mu}^*}^m(\mathbf{y}_{1:N}^o;\Psi,\Psi^*) + \mathbf{W}^*\left(\sigma^2\mathbf{M}^{-1} + \mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}\mathbf{M}^{-1}\right)\mathbf{W}^{*T}\right\}\end{aligned}$$

where

$$\begin{aligned}\boldsymbol{\mu}^* &= \bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o;\Psi)\left(\mathbb{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\right) + \boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T} \\ \mathbf{W}^* &= \bar{\mathbf{C}}_{\boldsymbol{\mu},\boldsymbol{\mu}^*}^m(\mathbf{y}_{1:N}^o;\Psi,\Psi^*)\mathbf{W}\mathbf{M}^{-1}\left(\sigma^2\mathbf{M}^{-1} + \mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}\mathbf{M}^{-1}\right)^{-1} \\ \sigma^{*2} &= \frac{1}{d}\operatorname{Tr}\left\{\bar{\mathbf{C}}_{\boldsymbol{\mu}^*}^m(\mathbf{y}_{1:N}^o;\Psi^*) - 2\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\boldsymbol{\mu},\boldsymbol{\mu}^*}^m(\mathbf{y}_{1:N}^o;\Psi,\Psi^*) + \mathbf{W}^*\left(\sigma^2\mathbf{M}^{-1} + \mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}\mathbf{M}^{-1}\right)\mathbf{W}^{*T}\right\}\end{aligned}$$

where

$$\begin{aligned}\bar{\mathbf{Q}} &= \begin{bmatrix}\mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_m\mathbf{W}_m^T + \sigma^2\mathbb{I}_{d_m} - \mathbf{W}_m\mathbf{W}_o^T\left(\mathbf{W}_o\mathbf{W}_o^T + \sigma^2\mathbb{I}_{d_o}\right)^{-1}\mathbf{W}_o\mathbf{W}_m^T\end{bmatrix}, \\ \bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o;\Psi) &= \frac{1}{N}\sum_{t=1}^{N}\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\Psi}\left[\mathbf{Y}_t\right], \\ \bar{\mathbf{S}}^m(\mathbf{y}_{1:N}^o;\Psi) &= \frac{1}{N}\sum_{t=1}^{N}\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\Psi}\left[\mathbf{Y}_t\right]^T\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\Psi}\left[\mathbf{Y}_t\right],\end{aligned}$$

$$\bar{\mathbf{C}}^m_{\boldsymbol{\mu}}(\mathbf{y}^o_{1:N};\Psi) = \bar{\mathbf{S}}^m(\mathbf{y}^o_{1:N};\Psi) + \bar{\mathbf{Q}} - \boldsymbol{\mu}^T\bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N};\Psi) - \bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N};\Psi)^T\boldsymbol{\mu} + \boldsymbol{\mu}^T\boldsymbol{\mu},$$
$$\bar{\mathbf{C}}^m_{\boldsymbol{\mu}^*}(\mathbf{y}^o_{1:N};\Psi^*) = \bar{\mathbf{S}}^m(\mathbf{y}^o_{1:N};\Psi) + \bar{\mathbf{Q}} - \boldsymbol{\mu}^{*T}\bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N};\Psi) - \bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N};\Psi)^T\boldsymbol{\mu}^* + \boldsymbol{\mu}^{*T}\boldsymbol{\mu}^*,$$
$$\bar{\mathbf{C}}^m_{\boldsymbol{\mu},\boldsymbol{\mu}^*}(\mathbf{y}^o_{1:N};\Psi,\Psi^*) = \bar{\mathbf{S}}^m(\mathbf{y}^o_{1:N};\Psi) + \bar{\mathbf{Q}} - \boldsymbol{\mu}^{*T}\bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N};\Psi) - \bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N};\Psi)^T\boldsymbol{\mu} + \boldsymbol{\mu}^{*T}\boldsymbol{\mu}.$$

□

A.3. t-Student Probabilistic Principal Component Analysis

A.3.1. Proof of Theorem 9

Proof. Given the assumption of the time independence of $\mathbf{X}_t$ and $\mathbf{U}_t$ we may express the conditional distribution of $\mathbf{X}_{1:N}, \mathbf{U}_{1:N}|\mathbf{Y}_{1:N}$ as

$$\pi_{\mathbf{X}_{1:N},U_{1:N}|\mathbf{Y}_{1:N},\Psi}(\mathbf{x}_{1:N},u_{1:N}) = \prod_{t=1}^{N}\pi_{\mathbf{X}_t,U_t|\mathbf{Y}_t,\Psi}(\mathbf{x}_t,u_t)$$

Hence, the E-step from Theorem 9 can be obtained as

$$\begin{aligned}
&Q_{TS,ind}(\Psi,\Psi^*) = \mathbb{E}_{\mathbf{X}_{1:N},U_{1:N}|\mathbf{Y}_{1:N},\Psi}\left[\log\pi_{\mathbf{Y}_{1:N},\mathbf{X}_{1:N},U_{1:N}|\Psi^*}(\mathbf{y}_{1:N},\mathbf{X}_{1:N},U_{1:N})\right]\\
&= \int_{\mathbb{R}_+^N}\int_{\mathbb{R}^{N\times k}}\log\left(\prod_{t=1}^{N}\pi_{\mathbf{Y}_t,\mathbf{X}_t,U_t|\Psi^*}(\mathbf{y}_t,\mathbf{x}_t,u_t)\right)\pi_{\mathbf{X}_{1:N},U_{1:N}|\mathbf{Y}_{1:N},\Psi}(\mathbf{x}_{1:N},u_{1:N})\,d\mathbf{x}_{1:N}\,du_{1:N}\\
&= \int_{\mathbb{R}_+^N}\int_{\mathbb{R}^{N\times k}}\log\left(\prod_{t=1}^{N}\pi_{\mathbf{Y}_t,\mathbf{X}_t,U_t|\Psi^*}(\mathbf{y}_t,\mathbf{x}_t,u_t)\right)\prod_{s=1}^{N}\pi_{\mathbf{X}_s,U_s|\mathbf{Y}_s,\Psi}(\mathbf{x}_s,u_s)\,d\mathbf{x}_{1:N}\,du_{1:N}\\
&= \sum_{t=1}^{N}\int_{\mathbb{R}_+^N}\int_{\mathbb{R}^{N\times k}}\log\left(\pi_{\mathbf{Y}_t,\mathbf{X}_t,U_t|\Psi^*}(\mathbf{y}_t,\mathbf{x}_t,u_t)\right)\prod_{s=1}^{N}\pi_{\mathbf{X}_s,U_s|\mathbf{Y}_s,\Psi}(\mathbf{x}_s,u_s)\,d\mathbf{x}_{1:N}\,du_{1:N}\\
&= \sum_{t=1}^{N}\left\{\int_{\mathbb{R}_+}\int_{\mathbb{R}^k}\log\left(\pi_{\mathbf{Y}_t,\mathbf{X}_t,U_t|\Psi^*}(\mathbf{y}_t,\mathbf{x}_t,u_t)\right)\cdot\pi_{\mathbf{X}_t,U_t|\mathbf{Y}_t,\Psi}(\mathbf{x}_t,u_t)d\mathbf{x}_t\,du_t\underbrace{\int_{\mathbb{R}_+^{N-1}}\int_{\mathbb{R}^{(N-1)\times k}}\left(\prod_{s=1,\,s\neq t}^{N}\pi_{\mathbf{X}_s,U_s|\mathbf{Y}_s,\Psi}(\mathbf{x}_s,u_s)\right)d\mathbf{x}_{1:N}\,du_{1:N}}_{N-1\text{ without }t\text{ element}}\right\}\\
&= \sum_{t=1}^{N}\left\{\int_{\mathbb{R}_+}\int_{\mathbb{R}^k}\log\left(\pi_{\mathbf{Y}_t,\mathbf{X}_t,U_t|\Psi^*}(\mathbf{y}_t,\mathbf{x}_t,u_t)\right)\pi_{\mathbf{X}_t,U_t|\mathbf{Y}_t,\Psi}(\mathbf{x}_t,u_t)d\mathbf{x}_t\,du_t\left(\prod_{s=1,\,s\neq t}^{N}\underbrace{\int_{\mathbb{R}_+}\int_{\mathbb{R}^k}\pi_{\mathbf{X}_s,U_s|\mathbf{Y}_s,\Psi}(\mathbf{x}_s,u_s)d\mathbf{x}_s\,du_s}_{=1\text{ since a density function}}\right)\right\}\\
&= \sum_{t=1}^{N}\int_{\mathbb{R}_+}\int_{\mathbb{R}^k}\log\left(\pi_{\mathbf{Y}_t,\mathbf{X}_t,U_t|\Psi^*}(\mathbf{y}_t,\mathbf{x}_t,u_t)\right)\pi_{\mathbf{X}_t,U_t|\mathbf{Y}_t,\Psi}(\mathbf{x}_t,u_t)d\mathbf{x}_t\,du_t\\
&= \sum_{t=1}^{N}\mathbb{E}_{\mathbf{X}_t,U_t|\mathbf{Y}_t,\Psi}\left[\log\left(\pi_{\mathbf{Y}_t,\mathbf{X}_t,U_t|\Psi^*}(\mathbf{y}_t,\mathbf{x}_t,u_t)\right)\right]\\
&= -\frac{1}{2}\sum_{t=1}^{N}\left\{(d+k-\frac{\nu^*}{2})\log 2\pi + d\log\sigma^{*2} - (d+k+\nu^*-2)\mathbb{E}_{U_t|\mathbf{Y}_t,\Psi}\left[\log U_t\right]\right.\\
&\quad + \frac{1}{\sigma^{*2}}\mathbb{E}_{U_t|\mathbf{Y}_t,\Psi}\left[U_t\right](\mathbf{y}_t-\boldsymbol{\mu}^*)(\mathbf{y}_t-\boldsymbol{\mu}^*)^T - \frac{2}{\sigma^{*2}}\mathbb{E}_{\mathbf{X}_t,U_t|\mathbf{Y}_t,\Psi}\left[U_t\mathbf{X}_t\right]\mathbf{W}^{*T}(\mathbf{y}_t-\boldsymbol{\mu}^*)^T\\
&\quad \left.+\operatorname{Tr}\left\{\left(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{W}^*+\mathbb{I}_k\right)\mathbb{E}_{\mathbf{X}_t,U_t|\mathbf{Y}_t,\Psi}\left[U_t\mathbf{X}_t^T\mathbf{X}_t\right]\right\} + \nu^*\mathbb{E}_{U_t|\mathbf{Y}_t,\Psi}\left[U_t\right] - 2\log\left(\frac{\left(\frac{\nu^*}{2}\right)^{\frac{\nu^*}{2}}}{\Gamma\left(\frac{\nu^*}{2}\right)}\right)\right\}
\end{aligned}$$

Recall, that the marginal distribution of $\mathbf{Y}_t$, as a convolution of two central t-Student distributions with the same degrees of freedom ν results in a t-Student distribution with ν degrees of freedom given by

$$\mathbf{Y}_t = \boldsymbol{\mu} + \frac{1}{\sqrt{U_t}}\mathbf{Z}_{x,t}\mathbf{W}^T + \sqrt{\frac{\sigma^2}{U_t}}\mathbf{Z}_{\epsilon,t} = \boldsymbol{\mu} + \frac{1}{\sqrt{U_t}}\left(\mathbf{Z}_{x,t}\mathbf{W}^T + \sqrt{\sigma^2}\mathbf{Z}_{\epsilon,t}\right) = \boldsymbol{\mu} + \frac{1}{\sqrt{U_t}}\mathbf{Z}_{y,t}$$

for $\mathbf{Z}_{y,t} \sim \mathcal{N}\left(0, \mathbf{W}\mathbf{W}^T + \sigma^2\mathbb{I}_d\right)$ being a sum of two independent multivariate Gaussian random variables. Notice that we used the stochastic representation of $\mathbf{X}_t$ and $\boldsymbol{\epsilon}_t$ specified in Equation (3.21) as well as the general model of PPCA from (2.7). Therefore, the marginal distribution of $\mathbf{Y}_t$ is t-Student with ν degrees of freedom, the mean $\boldsymbol{\mu}$ and the covariance matrix $\mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbb{I}_d$. The random vector $\mathbf{Y}_t$ is Gaussian when conditioned on the Gamma random variable U_t, that is $\mathbf{Y}_t|U_t,\Psi \sim \mathcal{N}\left(\boldsymbol{\mu}, \frac{1}{U_t}\mathbf{C}\right)$.

Using the above properties and Bayes' rule, we may specify the conditional distribution of U_t, that is

$$\pi_{U_t|\mathbf{Y}_t,\Psi}(u_t) = \frac{\pi_{\mathbf{Y}_t,U_t|\Psi}(\mathbf{y}_t,u_t)}{\pi_{\mathbf{Y}_t|\Psi}(\mathbf{y}_t)} = \frac{\pi_{\mathbf{Y}_t|U_T,\Psi}(\mathbf{y}_t)\cdot\pi_{U_t|\Psi}(u_t)}{\pi_{\mathbf{Y}_t|\Psi}(\mathbf{y}_t)}$$

The conditional distribution is proven in (Archambeau et al. 2006), (Fang and Jeong 2008), (Chen et al. 2009) to also be a univariate Gamma random variable due to conjugacy, however the degrees of freedom are adjusted as follows

$$U_t|\mathbf{Y}_t,\Psi \sim \Gamma\left(\frac{\nu+d}{2},\frac{\nu+D_t}{2}\right)$$

where $D_t = (\mathbf{Y}_t-\boldsymbol{\mu})\,\mathbf{C}^{-1}\,(\mathbf{Y}_t-\boldsymbol{\mu})^T$ which results in the first moment of the random variable $U_t|\mathbf{Y}_t,\Psi$ to be given by

$$\mathbb{E}_{\mathbf{X}_t,U_t|\mathbf{Y}_t,\Psi}[U_t] = \mathbb{E}_{U_t|\mathbf{Y}_t,\Psi}[U_t] = \frac{\nu+d}{\nu+D_t}$$

Consequently the conditional expectation of $\log U_t$ is given by

$$\mathbb{E}_{\mathbf{X}_t,U_t|\mathbf{Y}_t,\Psi}[\log U_t] = \mathbb{E}_{U_t|\mathbf{Y}_t,\Psi}[\log U_t] = \psi\left(\frac{\nu+d}{2}\right) - \log\left(\frac{\nu+D_t}{2}\right).$$

Further application of Bayes' rule gives the conditional distribution of $\mathbf{X}_t|\mathbf{Y}_t,U_t,\Psi$ as

$$\pi_{\mathbf{X}_t|\mathbf{Y}_t,U_t,\Psi}(\mathbf{x}_t) = \frac{\pi_{\mathbf{X}_t,\mathbf{Y}_t|U_t,\Psi}(\mathbf{x}_t,\mathbf{y}_t)}{\pi_{\mathbf{Y}_t|U_t,\Psi}(\mathbf{y}_t)} = \frac{\pi_{\mathbf{Y}_t|\mathbf{X}_t,U_t,\Psi}(\mathbf{y}_t)\cdot\pi_{\mathbf{X}_t|U_t,\Psi}(\mathbf{x}_t)}{\pi_{\mathbf{Y}_t|U_t,\Psi}(\mathbf{y}_t)}.$$

It can be shown that $\mathbf{X}_t|\mathbf{Y}_t,U_t,\Psi$ has a multivariate Gaussian distribution

$$\mathbf{X}_t|\mathbf{Y}_t,U_t,\Psi \sim \mathcal{N}\left((\mathbf{Y}_t-\boldsymbol{\mu})\,\mathbf{W}\mathbf{M}^{-1},\ \frac{\sigma^2}{U_t}\mathbf{M}^{-1}\right)$$

where $\mathbf{M}_{k\times k} = \mathbf{W}^T\mathbf{W}+\sigma^2\mathbb{I}_k$. Therefore, the joint moments of latent variable $\mathbf{X}_t$ and U_t conditioned on the vector of observations $\mathbf{Y}_t$ are obtained using the Law of Total Expectation as follows

$$\begin{aligned}
&\mathbb{E}_{\mathbf{X}_t,U_t|\mathbf{Y}_t,\Psi}[U_t\mathbf{X}_t]_{1\times k} = \mathbb{E}_{U_t|\mathbf{Y}_t,\Psi}\Big[U_t\mathbb{E}_{\mathbf{X}_t|\mathbf{Y}_t,U_t,\Psi}[\mathbf{X}_t]\Big] = \mathbb{E}_{U_t|\mathbf{Y}_t,\Psi}[U_t\,(\mathbf{Y}_t-\boldsymbol{\mu})\,\mathbf{W}\mathbf{M}^{-1}] = \frac{\nu+d}{\nu+D_t}\,(\mathbf{Y}_t-\boldsymbol{\mu})\,\mathbf{W}\mathbf{M}^{-1}\\
&\mathbb{E}_{\mathbf{X}_t,U_t|\mathbf{Y}_t,\Psi}[U_t\mathbf{X}_t^T\mathbf{X}_t]_{k\times k} = \mathbb{E}_{U_t|\mathbf{Y}_t,\Psi}\Big[U_t\mathbb{E}_{\mathbf{X}_t|\mathbf{Y}_t,U_t,\Psi}[\mathbf{X}_t^T\mathbf{X}_t]\Big]\\
&\quad = \mathbb{E}_{U_t|\mathbf{Y}_t,\Psi}\Big[\sigma^2\mathbf{M}^{-1}+U_t\mathbb{E}_{\mathbf{X}_t|\mathbf{Y}_t,U_t,\Psi}[\mathbf{X}_t]^T\mathbb{E}_{\mathbf{X}_t|\mathbf{Y}_t,U_t,\Psi}[\mathbf{X}_t]\Big] = \sigma^2\mathbf{M}^{-1}+\frac{\nu+d}{\nu+D_t}\mathbf{M}^{-1}\mathbf{W}^T\,(\mathbf{Y}_t-\boldsymbol{\mu})^T\,(\mathbf{Y}_t-\boldsymbol{\mu})\,\mathbf{W}\mathbf{M}^{-1}.
\end{aligned}$$

The final step it to replace $\mathbf{Y}_t$ with its realisation in time t, $\mathbf{y}_t$. □

A.3.2. Proof of Theorem 10

Proof. Using the notations introduced in the proof of the Theorem 9, the maximizers of the function

$$\begin{aligned}
Q_{TS,ind}\,(\Psi,\Psi^*) = -\frac{1}{2}\sum_{t=1}^{N}\Bigg\{&(d+k-\frac{\nu^*}{2})\log 2\pi + d\log\sigma^{*2} - (d+k+\nu^*-2)\mathbb{E}_{U_t|\mathbf{Y}_t,\Psi}[\log U_t]\\
&+\frac{1}{\sigma^{*2}}\mathbb{E}_{U_t|\mathbf{Y}_t,\Psi}[U_t]\,(\mathbf{y}_t-\boldsymbol{\mu}^*)\,(\mathbf{y}_t-\boldsymbol{\mu}^*)^T - \frac{2}{\sigma^{*2}}\mathbb{E}_{\mathbf{X}_t,U_t|\mathbf{Y}_t,\Psi}[U_t\mathbf{X}_t]\mathbf{W}^{*T}\,(\mathbf{y}_t-\boldsymbol{\mu}^*)^T\\
&+\mathrm{Tr}\left\{\left(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{W}^*+\mathbb{I}_k\right)\mathbb{E}_{\mathbf{X}_t,U_t|\mathbf{Y}_t,\Psi}[U_t\mathbf{X}_t^T\mathbf{X}_t]\right\} + \nu^*\mathbb{E}_{U_t|\mathbf{Y}_t,\Psi}[U_t] - 2\log\left(\frac{\left(\frac{\nu^*}{2}\right)^{\frac{\nu^*}{2}}}{\Gamma\left(\frac{\nu^*}{2}\right)}\right)\Bigg\}
\end{aligned}$$

with respect to the vector of the static parameters $\Psi^* = (\mathbf{W}^*,\boldsymbol{\mu}^*,\sigma^{*2})$, which are the solutions to the set of the equations $\nabla_{\Psi^*}Q = \mathbf{0}$, are given by

$$\begin{aligned}
&\frac{\partial Q_{TS,ind}(\Psi,\Psi^*)}{\partial\boldsymbol{\mu}^*} = -\frac{1}{2\sigma^*}\sum_{t=1}^{N}-2\mathbb{E}_{\mathbf{X}_t,U_t|\mathbf{Y}_t,\Psi}[U_t]\,(\mathbf{y}_t-\boldsymbol{\mu}^*)+2\mathbb{E}_{\mathbf{X}_t,U_t|\mathbf{Y}_t,\Psi}[U_t\mathbf{X}_t]\mathbf{W}^{*T} = \mathbf{0}\\
&\iff \sum_{t=1}^{N}\frac{\nu+d}{\nu+d_n}\,(\mathbf{y}_t-\boldsymbol{\mu}^*) - \frac{\nu+d}{\nu+D_t}\,(\mathbf{y}_t-\boldsymbol{\mu})\,\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T} = \mathbf{0}\\
&\iff \sum_{t=1}^{N}\left(\mathbf{y}_t\frac{\nu+d}{\nu+D_t}\right)\left(\mathbb{I}_d-\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\right)+\sum_{t=1}^{N}\left(\frac{\nu+d}{\nu+D_t}\right)\boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}-\boldsymbol{\mu}^*\sum_{t=1}^{N}\frac{\nu+d}{\nu+D_t} = \mathbf{0}
\end{aligned}$$

$$
\begin{aligned}
&\iff \underbrace{\frac{1}{N}\sum_{t=1}^{N}\left(\mathbf{y}_t\frac{\nu+d}{\nu+D_t}\right)}_{\bar{\boldsymbol{\mu}}_{ts}}\left(\mathbb{I}_d-\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\right)+\boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\underbrace{\frac{1}{N}\sum_{t=1}^{N}\frac{\nu+d}{\nu+D_t}}_{\bar{u}}-\boldsymbol{\mu}^*\frac{1}{N}\sum_{t=1}^{N}\frac{\nu+d}{\nu+D_t}=\mathbf{0}\\
&\iff \boldsymbol{\mu}^*=\bar{u}^{-1}\left(\bar{\boldsymbol{\mu}}_{ts}\left(\mathbb{I}_d-\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\right)+\bar{u}\boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\right)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial Q_{TS,ind}\left(\Psi,\Psi^*\right)}{\partial\mathbf{W}^*}&=-\frac{1}{2\sigma^{*2}}\sum_{t=1}^{N}\left\{-2\left(\mathbf{y}_t-\boldsymbol{\mu}^*\right)^T\mathbb{E}_{\mathbf{X}_t,U_t|\mathbf{Y}_t,\Psi}\left[U_t\mathbf{X}_t\right]+2\mathbf{W}^*\mathbb{E}_{\mathbf{X}_t,U_t|\mathbf{Y}_t,\Psi}\left[U_t\mathbf{X}_t^T\mathbf{X}_t\right]\right\}=\mathbf{0}\\
&\iff \sum_{t=1}^{N}\left\{\frac{\nu+d}{\nu+D_t}\left(\mathbf{y}_t-\boldsymbol{\mu}^*\right)^T\left(\mathbf{y}_t-\boldsymbol{\mu}\right)\mathbf{W}\mathbf{M}^{-1}\right.\\
&\left.-\mathbf{W}^*\left(\sigma^2\mathbf{M}^{-1}+\frac{\nu+d}{\nu+D_t}\mathbf{M}^{-1}\mathbf{W}^T\left(\mathbf{y}_t-\boldsymbol{\mu}\right)^T\left(\mathbf{y}_t-\boldsymbol{\mu}\right)\mathbf{W}\mathbf{M}^{-1}\right)\right\}=\mathbf{0}\\
&\iff \underbrace{\left(\frac{1}{N}\sum_{t=1}^{N}\frac{\nu+d}{\nu+D_t}\left(\mathbf{y}_t-\boldsymbol{\mu}^*\right)^T\left(\mathbf{y}_t-\boldsymbol{\mu}\right)\right)}_{\bar{\mathbf{C}}_{\mu,\mu^*}}\mathbf{W}\mathbf{M}^{-1}\\
&-\mathbf{W}^*\left(\sigma^2\mathbf{M}^{-1}+\mathbf{M}^{-1}\mathbf{W}^T\underbrace{\left(\frac{1}{N}\sum_{t=1}^{N}\frac{\nu+d}{\nu+D_t}\left(\mathbf{y}_t-\boldsymbol{\mu}\right)^T\left(\mathbf{y}_t-\boldsymbol{\mu}\right)\right)}_{\bar{\mathbf{C}}_{\mu}}\mathbf{W}\mathbf{M}^{-1}\right)=\mathbf{0}\\
&\iff \mathbf{W}^*=\bar{\mathbf{C}}_{\mu,\mu^*}\mathbf{W}\mathbf{M}^{-1}\left(\sigma^2\mathbf{M}^{-1}+\mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\mu}\mathbf{W}\mathbf{M}^{-1}\right)^{-1}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial Q_{TS,ind}\left(\Psi,\Psi^*\right)}{\partial\sigma^{*2}}&=-\frac{1}{2}\sum_{t=1}^{N}\left\{\frac{d}{\sigma^{*2}}-\frac{1}{(\sigma^{*2})^2}\left(\mathbb{E}_{\mathbf{X}_t,U_t|\mathbf{Y}_t,\Psi}\left[U_t\right]\operatorname{Tr}\left\{\left(\mathbf{y}_t-\boldsymbol{\mu}^*\right)^T\left(\mathbf{y}_t-\boldsymbol{\mu}^*\right)\right\}\right.\right.\\
&\left.\left.-2\operatorname{Tr}\left\{\mathbf{W}^{*T}\mathbb{E}_{\mathbf{X}_t,U_t|\mathbf{Y}_t,\Psi}\left[U_t\mathbf{X}_t\right)^T\left(\mathbf{y}_t-\boldsymbol{\mu}^*\right]\right\}+\operatorname{Tr}\left\{\mathbf{W}^{*T}\mathbf{W}^*\mathbb{E}_{\mathbf{X}_t,U_t|\mathbf{Y}_t,\Psi}\left[U_t\mathbf{X}_t^T\mathbf{X}_t\right]\right\}\right)\right\}=0\\
&\iff \sum_{t=1}^{N}\left\{\frac{d}{\sigma^{*2}}-\frac{1}{(\sigma^{*2})^2}\left(\frac{\nu+d}{\nu+D_t}\operatorname{Tr}\left\{\left(\mathbf{y}_t-\boldsymbol{\mu}^*\right)^T\left(\mathbf{y}_t-\boldsymbol{\mu}^*\right)\right\}\right.\right.\\
&-2\frac{\nu+d}{\nu+D_t}\operatorname{Tr}\left\{\mathbf{W}^{*T}\mathbf{M}^{-1}\mathbf{W}^T\left(\mathbf{y}_t-\boldsymbol{\mu}\right)^T\left(\mathbf{y}_t-\boldsymbol{\mu}^*\right)\right\}\\
&\left.\left.+\operatorname{Tr}\left\{\mathbf{W}^{*T}\mathbf{W}^*\left(\sigma^2\mathbf{M}^{-1}+\frac{\nu+d}{\nu+D_t}\mathbf{M}^{-1}\mathbf{W}^T\left(\mathbf{y}_t-\boldsymbol{\mu}\right)^T\left(\mathbf{y}_t-\boldsymbol{\mu}\right)\mathbf{W}\mathbf{M}^{-1}\right)\right\}\right)\right\}=0\\
&\iff \frac{d}{\sigma^{*2}}-\frac{1}{(\sigma^{*2})^2}\left(\operatorname{Tr}\left\{\underbrace{\frac{1}{N}\sum_{t=1}^{N}\frac{\nu+d}{\nu+D_t}\left(\mathbf{y}_t-\boldsymbol{\mu}^*\right)^T\left(\mathbf{y}_t-\boldsymbol{\mu}^*\right)}_{\bar{\mathbf{C}}_{\mu^*}}\right\}\right.\\
&-2\operatorname{Tr}\left\{\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\underbrace{\left(\frac{1}{N}\sum_{t=1}^{N}\frac{\nu+d}{\nu+D_t}\left(\mathbf{y}_t-\boldsymbol{\mu}\right)^T\left(\mathbf{y}_t-\boldsymbol{\mu}^*\right)\right)}_{\bar{\mathbf{C}}_{\mu,\mu^*}}\right\}\\
&\left.+\operatorname{Tr}\left\{\mathbf{W}^*\left(\sigma^2\mathbf{M}^{-1}+\mathbf{M}^{-1}\mathbf{W}^T\underbrace{\left(\frac{1}{N}\sum_{t=1}^{N}\frac{\nu+d}{\nu+D_t}\left(\mathbf{y}_t-\boldsymbol{\mu}\right)^T\left(\mathbf{y}_t-\boldsymbol{\mu}\right)\right)}_{\bar{\mathbf{C}}_{\mu}}\mathbf{W}\mathbf{M}^{-1}\right)\mathbf{W}^{*T}\right\}\right)=0\\
&\iff \sigma^{*2}=\frac{1}{d}\operatorname{Tr}\left\{\bar{\mathbf{C}}_{\mu^*}-2\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\mu,\mu^*}+\mathbf{W}^*\left(\sigma^2\mathbf{M}^{-1}+\mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\mu}\mathbf{W}\mathbf{M}^{-1}\right)\mathbf{W}^{*T}\right\}
\end{aligned}
$$

where

$$
\begin{aligned}
\bar{u}&=\frac{1}{N}\sum_{t=1}^{N}\frac{\nu+d}{\nu+D_t} & \bar{\boldsymbol{\mu}}_{ts}&=\frac{1}{N}\sum_{t=1}^{N}\mathbf{y}_t\frac{\nu+d}{\nu+D_t},\\
\bar{\mathbf{C}}_{\mu}&=\bar{\mathbf{S}}_{ts}-\boldsymbol{\mu}^T\bar{\boldsymbol{\mu}}_{ts}-\bar{\boldsymbol{\mu}}_{ts}^T\boldsymbol{\mu}+\bar{u}\boldsymbol{\mu}^T\boldsymbol{\mu}, & \bar{\mathbf{S}}_{ts}&=\frac{1}{N}\sum_{t=1}^{N}\frac{\nu+d}{\nu+D_t}\mathbf{y}_t^T\mathbf{y}_t,\\
\bar{\mathbf{C}}_{\mu^*}&=\bar{\mathbf{S}}_{ts}-\boldsymbol{\mu}^{*T}\bar{\boldsymbol{\mu}}_{ts}-\bar{\boldsymbol{\mu}}_{ts}^T\boldsymbol{\mu}^*+\bar{u}\boldsymbol{\mu}^{*T}\boldsymbol{\mu}^*,\\
\bar{\mathbf{C}}_{\mu,\mu^*}&=\bar{\mathbf{S}}_{ts}-\boldsymbol{\mu}^{*T}\bar{\boldsymbol{\mu}}_{ts}-\bar{\boldsymbol{\mu}}_{ts}^T\boldsymbol{\mu}+\bar{u}\boldsymbol{\mu}^{*T}\boldsymbol{\mu}.
\end{aligned}
$$

Hence, the maximizers of Q with respect to the vector of static parameters Ψ^* are the following

$$\begin{aligned}
\boldsymbol{\mu}^* &= \bar{u}^{-1}\left(\bar{\boldsymbol{\mu}}_{ts}\left(\mathbb{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\right) + \bar{u}\boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\right)\\
\mathbf{W}^* &= \bar{\mathbf{C}}_{\mu,\mu^*}\mathbf{W}\mathbf{M}^{-1}\left(\sigma^2\mathbf{M}^{-1} + \mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\mu}\mathbf{W}\mathbf{M}^{-1}\right)^{-1}\\
\sigma^{*2} &= \frac{1}{d}\operatorname{Tr}\left\{\bar{\mathbf{C}}_{\mu^*} - 2\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\mu,\mu^*} + \mathbf{W}^*\left(\sigma^2\mathbf{M}^{-1} + \mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\mu}\mathbf{W}\mathbf{M}^{-1}\right)\mathbf{W}^{*T}\right\}.
\end{aligned}$$

The reader may notice that the vector $\boldsymbol{\mu}^*$ is a linear function $\mathbf{W}^*$ and vice versa. Hence, we can solve the following linear system of equations

$$\begin{aligned}
\boldsymbol{\mu}^* &= \bar{u}^{-1}\left(\bar{\boldsymbol{\mu}}_{ts}\left(\mathbb{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\right) + \bar{u}\boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\right)\\
\mathbf{W}^* &= \bar{\mathbf{C}}_{\mu,\mu^*}\mathbf{W}\mathbf{M}^{-1}\left(\sigma^2\mathbf{M}^{-1} + \mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\mu}\mathbf{W}\mathbf{M}^{-1}\right)^{-1}
\end{aligned}$$

by substituting the values of $\mathbf{W}^*$ in the top equation by its values in the bottom equation, that is

$$\begin{aligned}
\bar{u}\boldsymbol{\mu}^* &= \bar{\boldsymbol{\mu}}_{ts} + (\bar{u}\boldsymbol{\mu} - \bar{\boldsymbol{\mu}}_{ts})\mathbf{W}\mathbf{M}^{-1}\Big(\underbrace{\bar{\mathbf{C}}_{\mu,\mu^*}\mathbf{W}\mathbf{M}^{-1}\left(\sigma^2\mathbf{M}^{-1} + \mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\mu}\mathbf{W}\mathbf{M}^{-1}\right)^{-1}}_{\mathbf{W}^*}\Big)^T\\
\bar{u}\boldsymbol{\mu}^* &= \bar{\boldsymbol{\mu}}_{ts} + (\bar{u}\boldsymbol{\mu} - \bar{\boldsymbol{\mu}}_{ts})\mathbf{W}\mathbf{M}^{-1}\left(\sigma^2\mathbf{M}^{-1} + \mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\mu}\mathbf{W}\mathbf{M}^{-1}\right)^{-1}\mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\mu,\mu^*}.
\end{aligned}$$

Next we substitute $\bar{\mathbf{C}}_{\mu,\mu^*} = \bar{\mathbf{S}}_{ts} - \boldsymbol{\mu}^{*T}\bar{\boldsymbol{\mu}}_{ts} - \bar{\boldsymbol{\mu}}_{ts}^T\boldsymbol{\mu} + \bar{u}\boldsymbol{\mu}^{*T}\boldsymbol{\mu}$ which contains the vector $\boldsymbol{\mu}^*$. We obtain that

$$\begin{aligned}
\bar{u}\boldsymbol{\mu}^* &= \bar{\boldsymbol{\mu}}_{ts} + (\bar{u}\boldsymbol{\mu} - \bar{\boldsymbol{\mu}}_{ts})\mathbf{W}\mathbf{M}^{-1}\left(\sigma^2\mathbf{M}^{-1} + \mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\mu}\mathbf{W}\mathbf{M}^{-1}\right)^{-1}\mathbf{M}^{-1}\mathbf{W}^T\left(\bar{\mathbf{S}}_{ts} - \bar{\boldsymbol{\mu}}_{ts}^T\boldsymbol{\mu}\right)\\
&\quad + (\bar{u}\boldsymbol{\mu} - \bar{\boldsymbol{\mu}}_{ts})\mathbf{W}\mathbf{M}^{-1}\left(\sigma^2\mathbf{M}^{-1} + \mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\mu}\mathbf{W}\mathbf{M}^{-1}\right)^{-1}\mathbf{M}^{-1}\mathbf{W}^T\left(\bar{u}\boldsymbol{\mu}^{*T}\boldsymbol{\mu} - \boldsymbol{\mu}^{*T}\bar{\boldsymbol{\mu}}_{ts}\right)\\
&\iff\\
&\left(\bar{u} - (\bar{u}\boldsymbol{\mu} - \bar{\boldsymbol{\mu}}_{ts})\mathbf{W}\mathbf{M}^{-1}\left(\sigma^2\mathbf{M}^{-1} + \mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\mu}\mathbf{W}\mathbf{M}^{-1}\right)^{-1}\mathbf{M}^{-1}\mathbf{W}^T\left(\bar{u}\boldsymbol{\mu} - \bar{\boldsymbol{\mu}}_{ts}\right)^T\right)\boldsymbol{\mu}^*\\
&\quad = \bar{\boldsymbol{\mu}}_{ts} + (\bar{u}\boldsymbol{\mu} - \bar{\boldsymbol{\mu}}_{ts})\mathbf{W}\mathbf{M}^{-1}\left(\sigma^2\mathbf{M}^{-1} + \mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\mu}\mathbf{W}\mathbf{M}^{-1}\right)^{-1}\mathbf{M}^{-1}\mathbf{W}^T\left(\bar{\mathbf{S}}_{ts} - \bar{\boldsymbol{\mu}}_{ts}^T\boldsymbol{\mu}\right)\\
&\iff\\
\boldsymbol{\mu}^* &= \left(\bar{u} - (\bar{u}\boldsymbol{\mu} - \bar{\boldsymbol{\mu}}_{ts})\mathbf{W}\mathbf{M}^{-1}\left(\sigma^2\mathbf{M}^{-1} + \mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\mu}\mathbf{W}\mathbf{M}^{-1}\right)^{-1}\mathbf{M}^{-1}\mathbf{W}^T\left(\bar{u}\boldsymbol{\mu} - \bar{\boldsymbol{\mu}}_{ts}\right)^T\right)^{-1}\\
&\quad \times\left(\bar{\boldsymbol{\mu}}_{ts} + (\bar{u}\boldsymbol{\mu} - \bar{\boldsymbol{\mu}}_{ts})\mathbf{W}\mathbf{M}^{-1}\left(\sigma^2\mathbf{M}^{-1} + \mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\mu}\mathbf{W}\mathbf{M}^{-1}\right)^{-1}\mathbf{M}^{-1}\mathbf{W}^T\left(\bar{\mathbf{S}}_{ts} - \bar{\boldsymbol{\mu}}_{ts}^T\boldsymbol{\mu}\right)\right)
\end{aligned}$$

for

$$\begin{aligned}
\mathbf{W}^* &= \bar{\mathbf{C}}_{\mu,\mu^*}\mathbf{W}\mathbf{M}^{-1}\left(\sigma^2\mathbf{M}^{-1} + \mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\mu}\mathbf{W}\mathbf{M}^{-1}\right)^{-1}\\
\sigma^{*2} &= \frac{1}{d}\operatorname{Tr}\left\{\bar{\mathbf{C}}_{\mu^*} - 2\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\mu,\mu^*} + \mathbf{W}^*\left(\sigma^2\mathbf{M}^{-1} + \mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\mu}\mathbf{W}\mathbf{M}^{-1}\right)\mathbf{W}^{*T}\right\}.
\end{aligned}$$

□

A.3.3. Proof of Theorem 11

Proof. The E-step of Identical and Conditionally Independent tStudent PPCA case is derived as follows

$$\begin{aligned}
&Q_{TS,iid}(\Psi,\Psi^*) = \mathbb{E}_{\mathbf{X}_{1:N},U|\mathbf{Y}_{1:N},\Psi}\left[\log\pi_{\mathbf{Y}_{1:N},\mathbf{X}_{1:N},U|\Psi^*}\left(\mathbf{y}_{1:N},\mathbf{X}_{1:N},U\right)\right]\\
&= \int_{\mathbb{R}_+}\int_{\mathbb{R}^{N\times k}}\log\left(\pi_{U|\Psi^*}(u)\prod_{t=1}^{N}\pi_{\mathbf{Y}_t,\mathbf{X}_t|U,\Psi^*}(\mathbf{y}_t,\mathbf{x}_t,u)\right)\pi_{\mathbf{X}_{1:N},U|\mathbf{Y}_{1:N},\Psi}(\mathbf{x}_{1:N},u)\,d\mathbf{x}_{1:N}\,du\\
&= \int_{\mathbb{R}_+}\int_{\mathbb{R}^{N\times k}}\log\left(\pi_{U|\Psi^*}(u)\prod_{t=1}^{N}\pi_{\mathbf{Y}_t,\mathbf{X}_t|U,\Psi^*}(\mathbf{y}_t,\mathbf{x}_t,u)\right)\frac{\pi_{\mathbf{Y}_{1:N},\mathbf{X}_{1:N},U|\Psi}(\mathbf{y}_{1:N},\mathbf{x}_{1:N},u)}{\pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N})}\,d\mathbf{x}_{1:N}\,du
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N})} \int_{\mathbb{R}_+} \int_{\mathbb{R}^{N\times k}} \left\{ \log\left(\pi_{U|\Psi^*}(u) \prod_{t=1}^{N} \pi_{\mathbf{Y}_t,\mathbf{X}_t|U,\Psi^*}(\mathbf{y}_t,\mathbf{x}_t,u) \right) \right. \\
&\quad \left. \times\, \pi_{\mathbf{Y}_{1:N}|\mathbf{X}_{1:N},U,\Psi}(\mathbf{y}_{1:N}) \cdot \pi_{\mathbf{X}_{1:N}|U,\Psi}(\mathbf{x}_{1:N}) \cdot \pi_{U|\Psi}(u) \right\} d\mathbf{x}_{1:N}\, du \\
&= \frac{1}{\pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N})} \int_{\mathbb{R}_+} \pi_{U|\Psi}(u) \int_{\mathbb{R}^{N\times k}} \left\{ \log\left(\pi_{U|\Psi^*}(u) \prod_{t=1}^{N} \pi_{\mathbf{Y}_t,\mathbf{X}_t|U,\Psi^*}(\mathbf{y}_t,\mathbf{x}_t,u) \right) \right. \\
&\quad \left. \times \prod_{s=1}^{N} \pi_{\mathbf{Y}_s|\mathbf{X}_s,U,\Psi}(\mathbf{y}_s) \cdot \pi_{\mathbf{X}_s|U,\Psi}(\mathbf{x}_s) \right\} d\mathbf{x}_{1:N}\, du \\
&= \underbrace{\frac{1}{\pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N})} \int_{\mathbb{R}_+} \pi_{U|\Psi}(u) \int_{\mathbb{R}^{N\times k}} \log \pi_{U|\Psi^*}(u) \prod_{s=1}^{N} \pi_{\mathbf{Y}_s,\mathbf{X}_s|U,\Psi}(\mathbf{y}_s,\mathbf{x}_s) \right\} d\mathbf{x}_{1:N}\, du}_{f(\nu^*)} \\
&\quad + \frac{1}{\pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N})} \sum_{t=1}^{N} \int_{\mathbb{R}_+} \pi_{U|\Psi}(u) \int_{\mathbb{R}^{N\times k}} \left\{ \log\left(\pi_{\mathbf{Y}_t,\mathbf{X}_t|U,\Psi^*}(\mathbf{y}_t,\mathbf{x}_t) \right) \prod_{s=1}^{N} \pi_{\mathbf{Y}_s|\mathbf{X}_s,U,\Psi}(\mathbf{y}_s) \pi_{\mathbf{X}_s|U,\Psi}(\mathbf{x}_s) \right\} d\mathbf{x}_{1:N}\, du \\
&= \frac{1}{\pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N})} \sum_{t=1}^{N} \int_{\mathbb{R}_+} \left\{ \pi_{U|\Psi}(u) \int_{\mathbb{R}^k} \log\left(\pi_{\mathbf{Y}_t,\mathbf{X}_t|U,\Psi^*}(\mathbf{y}_t,\mathbf{x}_t) \right) \pi_{\mathbf{Y}_t,\mathbf{X}_t|U,\Psi}(\mathbf{y}_t,\mathbf{x}_t) d\mathbf{x}_t \right. \\
&\quad \left. \times \underbrace{\left(\int_{\mathbb{R}^{(N-1)\times k}} \prod_{s=1,s\neq t}^{N} \pi_{\mathbf{Y}_s|\mathbf{X}_s,U,\Psi}(\mathbf{y}_s) \cdot \pi_{\mathbf{X}_s|U,\Psi}(\mathbf{x}_s)\, d\mathbf{x}_{1:N} \right)}_{\text{without } t \text{ element}} \right\} du + f(\nu^*).
\end{aligned}
$$

Based on the assumption that the variables $\mathbf{X}_t$ and $\boldsymbol{\epsilon}_t$ conditioned on U are independent Gaussian random vectors, Lemma A2 shows that we may find a function $h : \mathbb{R}^d \times \mathbb{R}_+ \to \mathbb{R}$ which satisfies

$$\pi_{\mathbf{Y}_t|\mathbf{X}_t,U,\Psi}(\mathbf{y}_t) \cdot \pi_{\mathbf{X}_t|U,\Psi}(\mathbf{x}_t) = h(\mathbf{y}_t,u;\Psi) \cdot \varphi(\mathbf{x}_t;\tilde{\boldsymbol{\mu}}_t,\tilde{\boldsymbol{\Sigma}})$$

for $\tilde{\boldsymbol{\mu}}_t \in \mathbb{R}^k$, a positive-definite covariance matrix $\tilde{\boldsymbol{\Sigma}} \in \mathbb{R}^{k\times k}$ and $\varphi(\mathbf{x};\tilde{\boldsymbol{\mu}},\tilde{\boldsymbol{\Sigma}})$ being a density function of the Gaussian random variable $\mathbf{X} \in \mathbb{R}^k$ with a vector of means $\tilde{\boldsymbol{\mu}}$ and a covariance matrix $\tilde{\boldsymbol{\Sigma}}$. Hence

$$
\begin{aligned}
Q_{TS,iid}(\Psi,\Psi^*) &= \frac{1}{\pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N})} \sum_{t=1}^{N} \int_{\mathbb{R}_+} \left\{ \pi_{U|\Psi}(u) \int_{\mathbb{R}^k} \log\left(\pi_{\mathbf{Y}_t,\mathbf{X}_t|U,\Psi^*}(\mathbf{y}_t,\mathbf{x}_t) \right) h(\mathbf{y}_t,u;\Psi) \varphi(\mathbf{x}_t;\tilde{\boldsymbol{\mu}}_t,\tilde{\boldsymbol{\Sigma}}) d\mathbf{x}_t \right. \\
&\quad \left. \times \underbrace{\left(\int_{\mathbb{R}^{(N-1)\times k}} \prod_{s=1,s\neq t}^{N} h(\mathbf{y}_s,u;\Psi) \varphi(\mathbf{x}_s;\tilde{\boldsymbol{\mu}}_t,\tilde{\boldsymbol{\Sigma}})\, d\mathbf{x}_{1:N} \right)}_{\text{without } t \text{ element}} \right\} du + f(\nu^*) \\
&= \frac{1}{\pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N})} \sum_{t=1}^{N} \int_{\mathbb{R}_+} \left\{ \pi_{U|\Psi}(u) \int_{\mathbb{R}^k} \log\left(\pi_{\mathbf{Y}_t,\mathbf{X}_t|U,\Psi^*}(\mathbf{y}_t,\mathbf{x}_t) \right) h(\mathbf{y}_t,u;\Psi) \varphi(\mathbf{x}_t;\tilde{\boldsymbol{\mu}}_t,\tilde{\boldsymbol{\Sigma}}) d\mathbf{x}_t \right. \\
&\quad \left. \times \underbrace{\left(\prod_{s=1,s\neq t}^{N} \int_{\mathbb{R}^k} h(\mathbf{y}_s,u;\Psi) \varphi(\mathbf{x}_s;\tilde{\boldsymbol{\mu}}_t,\tilde{\boldsymbol{\Sigma}})\, d\mathbf{x}_s \right)}_{\text{without } t \text{ element}} \right\} du + f(\nu^*) \\
&= \frac{1}{\pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N})} \sum_{t=1}^{N} \int_{\mathbb{R}_+} \left\{ \pi_{U|\Psi}(u) \int_{\mathbb{R}^k} \log\left(\pi_{\mathbf{Y}_t,\mathbf{X}_t|U,\Psi^*}(\mathbf{y}_t,\mathbf{x}_t) \right) h(\mathbf{y}_t,u;\Psi) \varphi(\mathbf{x}_t;\tilde{\boldsymbol{\mu}}_t,\tilde{\boldsymbol{\Sigma}}) d\mathbf{x}_t \right. \\
&\quad \left. \times \left(\prod_{s=1,s\neq t}^{N} h(\mathbf{y}_s,u;\Psi) \underbrace{\int_{\mathbb{R}^k} \varphi(\mathbf{x}_s;\tilde{\boldsymbol{\mu}}_t,\tilde{\boldsymbol{\Sigma}})\, d\mathbf{x}_s}_{=1 \text{ since a density function}} \right) \right\} du + f(\nu^*) \\
&= \frac{1}{\pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N})} \sum_{t=1}^{N} \int_{\mathbb{R}_+} \pi_{U|\Psi}(u) \cdot w(\mathbf{y}_t,u;\Psi,\Psi^*) \cdot \left(\prod_{s=1}^{N} h(\mathbf{y}_s,u;\Psi) \right) du + f(\nu^*) \\
&= \frac{1}{\pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N})} \int_{\mathbb{R}_+} \pi_{U|\Psi}(u) \cdot \left(\prod_{s=1}^{N} h(\mathbf{y}_s,u;\Psi) \right) \cdot \sum_{t=1}^{N} w(\mathbf{y}_t,u;\Psi,\Psi^*)\, du + f(\nu^*).
\end{aligned}
$$

where function $w : \mathbb{R}^d \times \mathbb{R}_+ \to \mathbb{R}$ is specified in Lemma A3.

The next step is to use Lemma A4 which shows that given the function $H(\mathbf{y}_{1:N},u;\Psi) := \prod_{t=1}^{N} h(\mathbf{y}_t,u;\Psi)$, the expression $\pi_{U|\Psi}(u) \cdot H(\mathbf{y}_{1:N},u;\Psi)$ is proportional to the univariate Gamma distribution, that is

$$
\begin{aligned}
\pi_{U|\Psi}(u) \cdot H(\mathbf{y}_{1:N},u;\Psi) &= C_H(\mathbf{y}_{1:N};\Psi) \frac{\beta(\mathbf{y}_{1:N};\Psi)^{\alpha}}{\Gamma(\alpha)} u^{\alpha-1} e^{-u\beta(\mathbf{y}_{1:N};\Psi)} \\
&= C_H(\mathbf{y}_{1:N};\Psi) \cdot \tilde{\pi}_{U|\Psi}(u).
\end{aligned}
$$

for $\alpha = \frac{\nu}{2} + \frac{dN}{2}$ and $\beta, C_H : \mathbb{R}^{N\times d} \to \mathbb{R}_+$ being defined as

$$\beta(\mathbf{y}_{1:N}; \Psi) = \frac{\nu}{2} + \frac{1}{2\sigma^2}\sum_{t=1}^{N}(\mathbf{y}_t - \boldsymbol{\mu})\left(\mathbb{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^T\right)(\mathbf{y}_t - \boldsymbol{\mu})^T,$$

$$C_H(\mathbf{y}_{1:N}; \Psi) = (2\pi)^{-\frac{dN}{2}}(\sigma^2)^{-N\frac{d-k}{2}}\left|\mathbf{M}\right|^{-\frac{N}{2}}\frac{\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}\Gamma(\alpha)}{\Gamma\left(\frac{\nu}{2}\right)}\beta(\mathbf{y}_{1:N}; \Psi)^{-\alpha}$$

The function $\tilde{\pi}_{U|\Psi}(u)$ is a density of $U \sim \Gamma(\alpha, \beta)$. Hence the E-step of EM algorithm for Identical and Conditionally Independent tStudent PPCA is given by

$$Q_{TS,iid}(\Psi, \Psi^*) = \frac{C_H(\mathbf{y}_{1:N}; \Psi)}{\pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N})}\int_{\mathbb{R}_+}\tilde{\pi}_{U|\Psi}(u)\Big(\log\pi_{U|\Psi^*}(u) + \sum_{t=1}^{N} w(\mathbf{y}_t, u; \Psi, \Psi^*)\Big)\,du$$

The reader may notice that the previously defined expression $f(\nu^*)$ becomes

$$f(\nu^*) = \frac{C_H(\mathbf{y}_{1:N}; \Psi)}{\pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N})}\int_{\mathbb{R}_+}\tilde{\pi}_{U|\Psi}(u)\log\pi_{U|\Psi^*}(u)\,du.$$

□

A.3.4. Proof of Theorem 12

Proof. The maximizers of the E-step from Theorem 11, the function $Q_{TS,iid}$, with respect to the vectors of parameters $\Psi^* = [\mathbf{W}^*, \boldsymbol{\mu}^*, \sigma^{*2}]$ are the solution to the set of the equations $\nabla_{\Psi^*}\tilde{Q} = \mathbf{0}$ and calculated as follows

$$\begin{aligned}\frac{Q_{TS,iid}(\Psi, \Psi^*)}{\partial\Psi^*} &= \frac{\partial}{\partial\Psi^*}\left(\frac{C_H(\mathbf{y}_{1:N}; \Psi)}{\pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N})}\int_{\mathbb{R}_+}\tilde{\pi}_{U|\Psi}(u)\Big(\log\pi_{U|\Psi^*}(u) + \sum_{t=1}^{N} w(\mathbf{y}_t, u; \Psi, \Psi^*)\Big)\,du\right)\\ &= \frac{C_H(\mathbf{y}_{1:N}; \Psi)}{\pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N})}\int_{\mathbb{R}_+}\tilde{\pi}_{U|\Psi}(u)\Big(\frac{\partial}{\partial\Psi^*}\log\pi_{U|\Psi^*}(u) + \sum_{t=1}^{N}\frac{\partial}{\partial\Psi^*} w(\mathbf{y}_t, u; \Psi, \Psi^*)\Big)\,du\end{aligned}$$

We applied Leibniz integral rule to swap the order or differentiating and integrating as the function

$$f(\Psi^*, u) = \tilde{\pi}_{U|\Psi}(u)\Big(\log\pi_{U|\Psi^*}(u) + \sum_{t=1}^{N} w(\mathbf{y}_t, u; \Psi, \Psi^*)\Big)$$

is integrable with respect to u and differentiable at Ψ^* and its gradient $\nabla_{\Psi^*} f$ if bounded by a function of u.

Notice that the component $\tilde{\pi}_{U|\Psi}(u)\log\pi_{U|\Psi^*}(u)$, is only dependent on the parameter ν^*. Since we do not maximize $Q_{TS,iid}$ with respect to the parameter ν^*, the derivative $\frac{\partial\log\pi_{U|\Psi^*}(u)}{\partial\Psi^*}$ is equal to zero. Therefore we can solve a dual problem

$$\int_{\mathbb{R}_+}\tilde{\pi}_{U|\Psi}(u)\frac{\partial}{\partial\Psi^*}\sum_{t=1}^{N} w(\mathbf{y}_t, u; \Psi, \Psi^*)\,du = \int_{\mathbb{R}_+}\tilde{\pi}_{U|\Psi}(u)\sum_{t=1}^{N}\frac{\partial}{\partial\Psi^*} w(\mathbf{y}_t, u; \Psi, \Psi^*)\,du = \mathbf{0}$$

Recalling the function w derived in Lemma A3,

$$\begin{aligned}w(\mathbf{y}_t, u; \Psi, \Psi^*) = &-\frac{d+k}{2}\log(2\pi) - \frac{d}{2}\log\left(\sigma^{*2}\right) + \frac{d+k}{2}\log u\\ &- \frac{u}{2\sigma^{*2}}\operatorname{Tr}\left\{(\mathbf{y}_t - \boldsymbol{\mu}^*)^T(\mathbf{y}_t - \boldsymbol{\mu}^*) - 2(\mathbf{y}_t - \boldsymbol{\mu})^T(\mathbf{y}_t - \boldsymbol{\mu}^*)\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\right\}\\ &- \frac{1}{2\sigma^{*2}}\operatorname{Tr}\left\{\left(u\mathbf{M}^{-1}\mathbf{W}^T(\mathbf{y}_t - \boldsymbol{\mu})^T(\mathbf{y}_t - \boldsymbol{\mu})\mathbf{W}\mathbf{M}^{-1} + \sigma^2\mathbf{M}^{-1}\right)\left(\mathbf{W}^{*T}\mathbf{W}^* + \sigma^{*2}\mathbb{I}_k\right)\right\}\end{aligned}$$

the partial derivatives of the sum of function w over N realisations of $\mathbf{Y}_t$ with respect to the parameters $\boldsymbol{\mu}^*$, $\mathbf{W}^*$, σ^{*2} are given by

$$\sum_{t=1}^{N}\frac{\partial w(\mathbf{y}_t, u; \Psi, \Psi^*)}{\partial\boldsymbol{\mu}^*} = -\frac{u}{2\sigma^{*2}}\sum_{t=1}^{N}\left\{-2(\mathbf{y}_t - \boldsymbol{\mu}^*) + 2(\mathbf{y}_t - \boldsymbol{\mu})\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\right\} = -N\frac{u}{\sigma^{*2}}\left[\boldsymbol{\mu}^* - \bar{\boldsymbol{\mu}}(\mathbb{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}) - \boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\right]$$

$$\begin{aligned}\sum_{t=1}^{N}\frac{\partial w(\mathbf{y}_t, u; \Psi, \Psi^*)}{\partial\boldsymbol{W}^*} &= -\frac{1}{2\sigma^{*2}}\sum_{t=1}^{N}\left\{u\Big(-2(\mathbf{y}_t - \boldsymbol{\mu}^*)^T(\mathbf{y}_t - \boldsymbol{\mu})\mathbf{W}\mathbf{M}^{-1} + 2\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T(\mathbf{y}_t - \boldsymbol{\mu})^T(\mathbf{y}_t - \boldsymbol{\mu})\mathbf{W}\mathbf{M}^{-1}\Big) + 2\sigma^2\mathbf{W}^*\mathbf{M}^{-1}\right\}\\ &= \frac{N}{\sigma^{*2}}\left[u\bar{\mathbf{C}}_{\mu,\mu^*}\mathbf{W}\mathbf{M}^{-1} - \mathbf{W}^*\left(u\mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\mu}\mathbf{W}\mathbf{M}^{-1} + \sigma^2\mathbf{M}^{-1}\right)\right]\end{aligned}$$

$$\sum_{t=1}^{N} \frac{\partial w(\mathbf{y}_t, u; \Psi, \Psi^*)}{\partial \sigma^{*2}} = \sum_{t=1}^{N} \Bigg\{ -\frac{d}{2\sigma^{*2}} + \frac{1}{2(\sigma^{*2})^2} \Bigg(u \operatorname{Tr}\Big\{ (\mathbf{y}_t - \boldsymbol{\mu}^*)^T (\mathbf{y}_t - \boldsymbol{\mu}^*) \Big\} - 2u \operatorname{Tr}\Big\{ (\mathbf{y}_t - \boldsymbol{\mu})^T (\mathbf{y}_t - \boldsymbol{\mu}^*) \mathbf{W}^* \mathbf{M}^{-1} \mathbf{W}^T \Big\}$$
$$+ \operatorname{Tr}\Big\{ u\mathbf{M}^{-1}\mathbf{W}^T (\mathbf{y}_t - \boldsymbol{\mu})^T (\mathbf{y}_t - \boldsymbol{\mu}) \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{W}^* + \sigma^2 \mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{W}^* \Big\} \Bigg) \Bigg\}$$
$$= -\frac{dN}{2\sigma^{*2}} + \frac{N}{2(\sigma^{*2})^2} \Bigg(\operatorname{Tr}\Big\{ u\bar{\mathbf{C}}_{\mu^*} - 2u\bar{\mathbf{C}}_{\mu,\mu^*}\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T \Big\} + \operatorname{Tr}\Big\{ u\mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_\mu \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{W}^* + \sigma^2\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{W}^* \Big\} \Bigg)$$

for

$$\bar{\boldsymbol{\mu}} = \frac{1}{N}\sum_{t=1}^{N} \mathbf{y}_t, \qquad \bar{\mathbf{C}}_\mu = \bar{\mathbf{S}} - \boldsymbol{\mu}^T\bar{\boldsymbol{\mu}} - \bar{\boldsymbol{\mu}}^T\boldsymbol{\mu} + \boldsymbol{\mu}^T\boldsymbol{\mu},$$
$$\bar{\mathbf{S}} = \frac{1}{N}\sum_{t=1}^{N} \mathbf{y}_t^T\mathbf{y}_t, \qquad \bar{\mathbf{C}}_{\mu^*} = \bar{\mathbf{S}} - \boldsymbol{\mu}^{*T}\bar{\boldsymbol{\mu}} - \bar{\boldsymbol{\mu}}^T\boldsymbol{\mu}^* + \boldsymbol{\mu}^{*T}\boldsymbol{\mu}^*,$$
$$\mathbf{M} = \mathbf{W}^T\mathbf{W} + \sigma^2\mathbb{I}_k \qquad \bar{\mathbf{C}}_{\mu,\mu^*} = \bar{\mathbf{S}} - \boldsymbol{\mu}^{*T}\bar{\boldsymbol{\mu}} - \bar{\boldsymbol{\mu}}^T\boldsymbol{\mu} + \boldsymbol{\mu}^{*T}\boldsymbol{\mu}.$$

Hence the maximizers of $\tilde{Q}$ are

$$\frac{Q_{TS,iid}(\Psi, \Psi^*)}{\partial \boldsymbol{\mu}^*} = \mathbf{0}$$
$$\iff \underbrace{\int_{\mathbb{R}_+} u \cdot \tilde{\pi}_{U|\Psi}(u)\, du}_{A(\mathbf{y}_{1:N};\Psi)} \Big(\boldsymbol{\mu}^* - \bar{\boldsymbol{\mu}}(\mathbb{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}) - \boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T} \Big) = \mathbf{0}$$
$$\iff A(\mathbf{y}_{1:N};\Psi) \Big(\boldsymbol{\mu}^* - \bar{\boldsymbol{\mu}}(\mathbb{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}) - \boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T} \Big) = \mathbf{0}$$
$$\iff \boldsymbol{\mu}^* = \bar{\boldsymbol{\mu}}(\mathbb{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}) + \boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}$$

$$\frac{Q_{TS,iid}(\Psi, \Psi^*)}{\partial \boldsymbol{W}^*} = \mathbf{0}$$
$$\iff \int_{\mathbb{R}_+} \tilde{\pi}_{U|\Psi}(u) \frac{N}{\sigma^{*2}} \Big(u\bar{\mathbf{C}}_{\mu,\mu^*}\mathbf{W}\mathbf{M}^{-1} - \mathbf{W}^* \Big(u\mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_\mu\mathbf{W}\mathbf{M}^{-1} + \sigma^2\mathbf{M}^{-1} \Big) \Big)\, du = \mathbf{0}$$
$$\iff \underbrace{\int_{\mathbb{R}_+} u\, \tilde{\pi}_{U|\Psi}(u)\, du}_{A(\mathbf{y}_{1:N};\Psi)} \bar{\mathbf{C}}_{\mu,\mu^*}\mathbf{W}\mathbf{M}^{-1} - \mathbf{W}^* \Big(\underbrace{\int_{\mathbb{R}_+} u\, \tilde{\pi}_{U|\Psi}(u)\, du}_{A(\mathbf{y}_{1:N};\Psi)} \mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_\mu\mathbf{W}\mathbf{M}^{-1} + \sigma^2\mathbf{M}^{-1} \Big) = \mathbf{0}$$
$$\iff \mathbf{W}^* = A(\mathbf{y}_{1:N};\Psi)\bar{\mathbf{C}}_{\mu,\mu^*}\mathbf{W}\mathbf{M}^{-1} \Big(A(\mathbf{y}_{1:N};\Psi)\mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_\mu\mathbf{W}\mathbf{M}^{-1} + \sigma^2\mathbf{M}^{-1} \Big)^{-1}$$

$$\frac{Q_{TS,iid}(\Psi, \Psi^*)}{\partial \sigma^{*2}} = 0$$
$$\iff \int_{\mathbb{R}_+} \tilde{\pi}_{U|\Psi}(u) \Bigg(-\frac{dN}{2\sigma^{*2}} + \frac{N}{2(\sigma^{*2})^2} \Bigg(\operatorname{Tr}\Big\{ u\bar{\mathbf{C}}_{\mu^*} - 2u\bar{\mathbf{C}}_{\mu,\mu^*}\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T \Big\}$$
$$+ \operatorname{Tr}\Big\{ u\mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_\mu\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{W}^* + \sigma^2\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{W}^* \Big\} \Bigg) \Bigg) du = 0$$
$$\iff \Bigg(-\frac{d}{\sigma^{*2}} + \frac{1}{(\sigma^{*2})^2} \Bigg(\operatorname{Tr}\Big\{ \underbrace{\int_{\mathbb{R}_+} u\, \tilde{\pi}_{U|\Psi}(u)\, du}_{A(\mathbf{y}_{1:N};\Psi)} \bar{\mathbf{C}}_{\mu^*} - 2 \underbrace{\int_{\mathbb{R}_+} u\, \tilde{\pi}_{U|\Psi}(u)\, du}_{A(\mathbf{y}_{1:N};\Psi)} \bar{\mathbf{C}}_{\mu,\mu^*}\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T \Big\}$$
$$+ \operatorname{Tr}\Big\{ \underbrace{\int_{\mathbb{R}_+} u\, \tilde{\pi}_{U|\Psi}(u)\, du}_{A(\mathbf{y}_{1:N};\Psi)} \mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_\mu\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{W}^* + \sigma^2\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{W}^* \Big\} \Bigg) \Bigg) du = 0$$
$$\iff \sigma^{*2} = \frac{1}{d}\operatorname{Tr}\Big\{ A(\mathbf{y}_{1:N};\Psi)\bar{\mathbf{C}}_{\mu^*} - 2A(\mathbf{y}_{1:N};\Psi)\bar{\mathbf{C}}_{\mu,\mu^*}\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T \Big\}$$
$$+ \frac{1}{d}\operatorname{Tr}\Big\{ \Big(A(\mathbf{y}_{1:N};\Psi)\mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_\mu\mathbf{W}\mathbf{M}^{-1} + \sigma^2\mathbf{M}^{-1} \Big)\mathbf{W}^{*T}\mathbf{W}^* \Big\}$$

where used the fact that $\int_{\mathbb{R}_+} \tilde{\pi}_{U|\Psi}(u)\, du = 1$ as $\tilde{\pi}_{U|\Psi}(u)$ is a density function as stated in Lemma A4. Given this fact we can easily derive the explicit solution to the integration problem A since the problem represents the expectation of the univariate Gamma variable, that is

$$A(\mathbf{y}_{1:N}; \Psi) = \int_{\mathbb{R}_+} u\, \tilde{\pi}_{U|\Psi}(u)\, du = \frac{\alpha}{\beta(\mathbf{y}_{1:N}; \Psi)}$$

where $\alpha = \frac{\nu}{2} + \frac{dN}{2}$ and $\beta : \mathbb{R}^{N \times d} \to \mathbb{R}_+$ being a function of $\mathbf{y}_{1:N}$, that is

$$\beta(\mathbf{y}_{1:N}; \Psi) = \frac{\nu}{2} + \frac{1}{2\sigma^2} \sum_{t=1}^{N} (\mathbf{y}_t - \boldsymbol{\mu}) \left(\mathbb{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^T \right) (\mathbf{y}_t - \boldsymbol{\mu})^T.$$

Therefore the solutions to the set of equation given by $\nabla_{\Psi^*} \tilde{Q} = \mathbf{0}$ where $\Psi^* = (\boldsymbol{\mu}^*, \mathbf{W}^*, \sigma^{*2})$

$$\begin{aligned}
\boldsymbol{\mu}^* &= \bar{\boldsymbol{\mu}}(\mathbb{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}) + \boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T} \\
\mathbf{W}^* &= \bar{\mathbf{C}}_{\mu,\mu^*}\mathbf{W}\mathbf{M}^{-1} \left(\mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\mu}\mathbf{W}\mathbf{M}^{-1} + \frac{\beta(\mathbf{y}_{1:N}; \Psi)}{\alpha}\sigma^2\mathbf{M}^{-1} \right)^{-1} \\
\sigma^{*2} &= \frac{1}{d} \operatorname{Tr}\left\{ \frac{\alpha}{\beta(\mathbf{y}_{1:N}; \Psi)} \left(\bar{\mathbf{C}}_{\mu^*} - 2\bar{\mathbf{C}}_{\mu,\mu^*}\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T \right) \right\} + \frac{1}{d} \operatorname{Tr}\left\{ \left(\frac{\alpha}{\beta(\mathbf{y}_{1:N}; \Psi)}\mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\mu}\mathbf{W}\mathbf{M}^{-1} + \sigma^2\mathbf{M}^{-1} \right)\mathbf{W}^{*T}\mathbf{W}^* \right\}.
\end{aligned}$$

The reader may notice that the vector $\boldsymbol{\mu}^*$ is a linear function $\mathbf{W}^*$ and vice versa. Hence, we can solve the following linear system of equations

$$\begin{aligned}
\boldsymbol{\mu}^* &= \bar{\boldsymbol{\mu}}(\mathbb{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}) + \boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T} \\
\mathbf{W}^* &= \bar{\mathbf{C}}_{\mu,\mu^*}\mathbf{W}\mathbf{M}^{-1} \left(\mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\mu}\mathbf{W}\mathbf{M}^{-1} + \frac{\beta(\mathbf{y}_{1:N}; \Psi)}{\alpha}\sigma^2\mathbf{M}^{-1} \right)^{-1}
\end{aligned}$$

by substituting the values of $\mathbf{W}^*$ in the top equation by its values in the bottom equation, that is

$$\begin{aligned}
&\boldsymbol{\mu}^* = \bar{\boldsymbol{\mu}} + (\boldsymbol{\mu} - \bar{\boldsymbol{\mu}})\mathbf{W}\mathbf{M}^{-1} \Bigg(\underbrace{\bar{\mathbf{C}}_{\mu,\mu^*}\mathbf{W}\mathbf{M}^{-1} \left(\mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\mu}\mathbf{W}\mathbf{M}^{-1} + \frac{\beta(\mathbf{y}_{1:N}; \Psi)}{\alpha}\sigma^2\mathbf{M}^{-1} \right)^{-1}}_{\mathbf{W}^*} \Bigg)^T \\
&\iff \\
&\boldsymbol{\mu}^* = \bar{\boldsymbol{\mu}} + (\boldsymbol{\mu} - \bar{\boldsymbol{\mu}})\mathbf{W}\mathbf{M}^{-1} \left(\mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\mu}\mathbf{W}\mathbf{M}^{-1} + \frac{\beta(\mathbf{y}_{1:N}; \Psi)}{\alpha}\sigma^2\mathbf{M}^{-1} \right)^{-1} \mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\mu,\mu^*}
\end{aligned}$$

Next we substitute $\bar{\mathbf{C}}_{\mu,\mu^*} = \bar{\mathbf{S}} - \boldsymbol{\mu}^{*T}\bar{\boldsymbol{\mu}} - \bar{\boldsymbol{\mu}}^T\boldsymbol{\mu} + \boldsymbol{\mu}^{*T}\boldsymbol{\mu}$ which contains the vector $\boldsymbol{\mu}^*$. We obtain that

$$\begin{aligned}
\boldsymbol{\mu}^* &= \bar{\boldsymbol{\mu}} + (\boldsymbol{\mu} - \bar{\boldsymbol{\mu}})\mathbf{W}\mathbf{M}^{-1} \left(\mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\mu}\mathbf{W}\mathbf{M}^{-1} + \frac{\beta(\mathbf{y}_{1:N}; \Psi)}{\alpha}\sigma^2\mathbf{M}^{-1} \right)^{-1} \mathbf{M}^{-1}\mathbf{W}^T \left(\bar{\mathbf{S}} - \bar{\boldsymbol{\mu}}^T\boldsymbol{\mu} \right) \\
&\quad + (\boldsymbol{\mu} - \bar{\boldsymbol{\mu}})\mathbf{W}\mathbf{M}^{-1} \left(\mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\mu}\mathbf{W}\mathbf{M}^{-1} + \frac{\beta(\mathbf{y}_{1:N}; \Psi)}{\alpha}\sigma^2\mathbf{M}^{-1} \right)^{-1} \mathbf{M}^{-1}\mathbf{W}^T \left(\boldsymbol{\mu}^T\boldsymbol{\mu}^* - \bar{\boldsymbol{\mu}}^T\boldsymbol{\mu}^* \right) \\
&\iff \\
&\left(1 - (\boldsymbol{\mu} - \bar{\boldsymbol{\mu}})\mathbf{W}\mathbf{M}^{-1} \left(\mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\mu}\mathbf{W}\mathbf{M}^{-1} + \frac{\beta(\mathbf{y}_{1:N}; \Psi)}{\alpha}\sigma^2\mathbf{M}^{-1} \right)^{-1} \mathbf{M}^{-1}\mathbf{W}^T \left(\boldsymbol{\mu} - \bar{\boldsymbol{\mu}} \right)^T \right) \boldsymbol{\mu}^* \\
&\quad = \bar{\boldsymbol{\mu}} + (\boldsymbol{\mu} - \bar{\boldsymbol{\mu}})\mathbf{W}\mathbf{M}^{-1} \left(\mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\mu}\mathbf{W}\mathbf{M}^{-1} + \frac{\beta(\mathbf{y}_{1:N}; \Psi)}{\alpha}\sigma^2\mathbf{M}^{-1} \right)^{-1} \mathbf{M}^{-1}\mathbf{W}^T \left(\bar{\mathbf{S}} - \bar{\boldsymbol{\mu}}^T\boldsymbol{\mu} \right)
\end{aligned}$$

for

$$\begin{aligned}
\mathbf{W}^* &= \bar{\mathbf{C}}_{\mu,\mu^*}\mathbf{W}\mathbf{M}^{-1} \left(\mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\mu}\mathbf{W}\mathbf{M}^{-1} + \frac{\beta(\mathbf{y}_{1:N}; \Psi)}{\alpha}\sigma^2\mathbf{M}^{-1} \right)^{-1} \\
\sigma^{*2} &= \frac{1}{d} \operatorname{Tr}\left\{ \frac{\alpha}{\beta(\mathbf{y}_{1:N}; \Psi)} \left(\bar{\mathbf{C}}_{\mu^*} - 2\bar{\mathbf{C}}_{\mu,\mu^*}\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T \right) \right\} + \frac{1}{d} \operatorname{Tr}\left\{ \left(\frac{\alpha}{\beta(\mathbf{y}_{1:N}; \Psi)}\mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}_{\mu}\mathbf{W}\mathbf{M}^{-1} + \sigma^2\mathbf{M}^{-1} \right)\mathbf{W}^{*T}\mathbf{W}^* \right\}.
\end{aligned}$$

□

A.3.5. Proof of Theorem 21

Proof. The steps of proof of Theorem 21 follows the reasoning of the proof of Theorem 9. Given the N realisation of the vector with observed values, $\mathbf{y}_{1:N}^o = \mathbf{y}_1^o, \ldots, \mathbf{y}_N^o$ and recalling that the observation vector is partitioned into two subvectors, $\mathbf{Y}_t = [\mathbf{Y}_t^o, \mathbf{Y}_t^m]$ with observed and unobserved entries of $\mathbf{Y}_t$, the E-step from Theorem 21 is obtained by

$$\begin{aligned}
&\tilde{Q}_{TS,ind}(\Psi, \Psi^*) = \mathbb{E}_{\mathbf{Y}_{1:N}^m, \mathbf{X}_{1:N}, U_{1:N}|\mathbf{Y}_{1:N}^o, \Psi}\Big[\log \pi_{\mathbf{Y}_{1:N}, \mathbf{X}_{1:N}, U_{1:N}|\Psi^*}(\mathbf{Y}_{1:N}, \mathbf{X}_{1:N}, U_{1:N})\Big] \\
&= \int_{\mathbb{R}_+^N}\int_{\mathbb{R}^{N\times k}}\int_{\mathbb{R}^{N\times d_m}} \left\{\log\left(\prod_{t=1}^N \pi_{\mathbf{Y}_t, \mathbf{X}_t, U_t|\Psi^*}(\mathbf{y}_t, \mathbf{x}_t, u_t)\right)\pi_{\mathbf{Y}_{1:N}^m, \mathbf{X}_{1:N}, U_{1:N}|\mathbf{Y}_{1:N}^o, \Psi}(\mathbf{y}_{1:N}^m, \mathbf{x}_{1:N}, u_{1:N})\right\} d\mathbf{y}_{1:N}^m\, d\mathbf{x}_{1:N}\, du_{1:N} \\
&= \sum_{t=1}^N \int_{\mathbb{R}_+}\int_{\mathbb{R}^k}\int_{\mathbb{R}^{d_m}} \log\left(\pi_{\mathbf{Y}_t, \mathbf{X}_t, U_t|\Psi^*}(\mathbf{y}_t, \mathbf{x}_t, u_t)\right)\pi_{\mathbf{Y}_t^m, \mathbf{X}_t, U_t|\mathbf{Y}_t^o, \Psi}(\mathbf{y}_t^m, \mathbf{x}_t, u_t)\, d\mathbf{y}_t^m\, d\mathbf{x}_t\, du_t
\end{aligned}$$

Hence we have that

$$\begin{aligned}
&\tilde{Q}_{TS,ind}(\Psi, \Psi^*) = \sum_{t=1}^N \mathbb{E}_{\mathbf{Y}_t^m, \mathbf{X}_t, U_t|\mathbf{Y}_t^o, \Psi}\Big[\log \pi_{\mathbf{Y}_t, \mathbf{X}_t, U_t|\Psi^*}(\mathbf{y}_t, \mathbf{x}_t, u_t)\Big] \\
&= -\frac{1}{2}\sum_{t=1}^N \Big\{\Big(d + k - \frac{v^*}{2}\Big)\log 2\pi + d\log\sigma^{*2} - (d + k + v^* - 2)\mathbb{E}_{U_t|\mathbf{Y}_t^o, \Psi}\big[\log U_t\big] \\
&+ \frac{1}{\sigma^{*2}}\mathbb{E}_{\mathbf{Y}_t^m, U_t|\mathbf{Y}_t^o, \Psi}\big[U_t\,(\mathbf{Y}_t - \boldsymbol{\mu}^*)\,(\mathbf{Y}_t - \boldsymbol{\mu}^*)^T\big] - \frac{2}{\sigma^{*2}} tr\Big\{\mathbf{W}^*\mathbb{E}_{\mathbf{Y}_t^m, \mathbf{X}_t, U_t|\mathbf{Y}_t^o, \Psi}\big[U_t\mathbf{X}_t^T\,(\mathbf{Y}_t - \boldsymbol{\mu}^*)\big]\Big\} \\
&+ tr\Big\{\Big(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{W}^* + \mathbb{I}_k\Big)\mathbb{E}_{\mathbf{X}_t, U_t|\mathbf{Y}_t^o, \Psi}\big[U_t\mathbf{X}_t^T\mathbf{X}_t\big]\Big\} + v^*\mathbb{E}_{U_t|\mathbf{Y}_t^o, \Psi}\big[U_t\big] - 2\log\left(\frac{\left(\frac{v^*}{2}\right)^{\frac{v^*}{2}}}{\Gamma\left(\frac{v^*}{2}\right)}\right)\Big\} \\
&= -\frac{1}{2}\sum_{t=1}^N \Big\{\Big(d + k - \frac{v^*}{2}\Big)\log 2\pi + d\log\sigma^{*2} - (d + k + v^* - 2)\mathbb{E}_{U_t|\mathbf{Y}_t^o, \Psi}\big[\log U_t\big] \\
&+ \frac{1}{\sigma^{*2}}\,\mathrm{Tr}\Big\{\mathbb{E}_{\mathbf{Y}_t^m, U_t|\mathbf{Y}_t^o, \Psi}\big[U_t\,(\mathbf{Y}_t - \boldsymbol{\mu}^*)^T\,(\mathbf{Y}_t - \boldsymbol{\mu}^*)\big]\Big\} - \frac{2}{\sigma^{*2}} tr\Big\{\mathbf{W}^*\Big(\mathbb{E}_{\mathbf{Y}_t^m, \mathbf{X}_t, U_t|\mathbf{Y}_t^o, \Psi}\big[U_t\mathbf{X}_t^T\mathbf{Y}_t\big] - \mathbb{E}_{\mathbf{X}_t, U_t|\mathbf{Y}_t^o, \Psi}\big[U_t\mathbf{X}_t\big]^T\boldsymbol{\mu}^*\Big)\Big\} \\
&+ tr\Big\{\Big(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{W}^* + \mathbb{I}_k\Big)\mathbb{E}_{\mathbf{X}_t, U_t|\mathbf{Y}_t^o, \Psi}\big[U_t\mathbf{X}_t^T\mathbf{X}_t\big]\Big\} + v^*\mathbb{E}_{U_t|\mathbf{Y}_t^o, \Psi}\big[U_t\big] - 2\log\left(\frac{\left(\frac{v^*}{2}\right)^{\frac{v^*}{2}}}{\Gamma\left(\frac{v^*}{2}\right)}\right)\Big\}.
\end{aligned}$$

Since the distribution $\mathbf{Y}_t|U_t, \Psi$ is Gaussian, it is shown in Lemma A1 that the distribution of the vector $\mathbf{Y}_t$ conditioned on the subvector of the observed values and the variable U_t, is specified as follows

$$\mathbf{Y}_t|\mathbf{Y}_t^o, U_t \sim \mathcal{N}\left(\begin{bmatrix}\mathbf{Y}_t^o \\ \boldsymbol{\mu}_m + (\mathbf{Y}_t^o - \boldsymbol{\mu}_o)\,\mathbf{C}_{oo}^{-1}\mathbf{C}_{om}\end{bmatrix}, \frac{1}{U_t}\begin{bmatrix}\mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{mm} - \mathbf{C}_{mo}\mathbf{C}_{oo}^{-1}\mathbf{C}_{om}\end{bmatrix}\right)$$

where $\mathbf{C}_{oo} = \mathbf{W}_o\mathbf{W}_o^T + \sigma^2\mathbb{I}_{d_o}$ and $\mathbf{C}_{mm} = \mathbf{W}_m\mathbf{W}_m^T + \sigma^2\mathbb{I}_{d_m}$ are square, symmetric submatrices of $\mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbb{I}_d$ with rows (left side index) and columns (right side index) corresponding to observed entries and missing entries, respectively. The matrices $\mathbf{C}_{om} = \mathbf{W}_o\mathbf{W}_m^T$ is a submatrix of the matrix $\mathbf{C}$ which rows correspond to observed entries and columns to missing entries of the vector $\mathbf{Y}_t$, and $\mathbf{C}_{mo} = \mathbf{C}_{om}^T$. Let us denote

$$\mathbf{Q} = \begin{bmatrix}\mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{mm} - \mathbf{C}_{mo}\mathbf{C}_{oo}^{-1}\mathbf{C}_{om}\end{bmatrix}$$

We can specify the following moments which are useful to derive the conditional moments of the distribution $\mathbf{Y}_t^m, \mathbf{X}_t, U_t|\mathbf{Y}_t^o$

$$\begin{aligned}
\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o, U_t, \Psi}\big[\mathbf{Y}_t\big]_{1\times d} &= \begin{bmatrix}\mathbf{Y}_t^o \\ \boldsymbol{\mu}_m + (\mathbf{Y}_t^o - \boldsymbol{\mu}_o)\,\mathbf{C}_{oo}^{-1}\mathbf{C}_{om}\end{bmatrix}, \\
\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o, U_t, \Psi}\big[\mathbf{Y}_t^T\mathbf{Y}_t\big]_{d\times d} &= \frac{1}{U_t}\mathbf{Q} + \mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o, \Psi}\big[\mathbf{Y}_t\big]^T\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o, \Psi}\big[\mathbf{Y}_t\big].
\end{aligned}$$

Let us define $D_t^o = (\mathbf{Y}_t^o - \boldsymbol{\mu}_o)\,\mathbf{C}_{oo}^{-1}\,(\mathbf{Y}_t^o - \boldsymbol{\mu}_o)^T$ being a squared Mahalanobias distance of the observed data subvector $\mathbf{Y}_t^o$ with marginal distribution $\mathbf{Y}_t^o \sim \mathcal{N}(\boldsymbol{\mu}_o, \mathbf{C}_{oo})$. The moments of the variables $U_t|\mathbf{Y}_t^o, \Psi$ and $\log U_t|\mathbf{Y}_t^o, \Psi$ are calculated similarly as their equivalents in complete data setting, using the marginal distribution of $\mathbf{Y}_t^o$, and are given by

$$\begin{aligned}
\mathbb{E}_{U_t|\mathbf{Y}_t^o, \Psi}\big[U_t\big] &= \frac{v + d_o}{v + D_t^o}, \\
\mathbb{E}_{U_t|\mathbf{Y}_t^o, \Psi}\big[\log U_t\big] &= \psi\left(\frac{v + d_o}{2}\right) - \log\left(\frac{v + D_t^o}{2}\right).
\end{aligned}$$

The moments of the joint distribution of the latent variables $\mathbf{X}_t, \mathbf{Y}_t^m$ and U_t conditioned on the subvector of observed entries $\mathbf{Y}_t^o$ are obtained using the Law of Total Expectation. Applying the chain rule of probability to the join distribution

$$\pi_{\mathbf{Y}_t^m,\mathbf{X}_t,U_t|\mathbf{Y}_t^o,\Psi} = \pi_{\mathbf{X}_t|\mathbf{Y}_t,U_t,\Psi} \cdot \pi_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U_t,\Psi} \cdot \pi_{U_t|\mathbf{Y}_t^o,\Psi}$$

and the fact that conditional means of $\mathbf{Y}_t^m|\mathbf{Y}_t^o, U_t$ and $\mathbf{X}_t|\mathbf{Y}_t^o, U_t$ are not dependent on $\mathbf{U}_t$, we obtain

$$\begin{aligned}
\mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t,U_t|\mathbf{Y}_t^o,\Psi}\left[U_t\mathbf{Y}_t\right]_{1\times d} &= \mathbb{E}_{\mathbf{Y}_t^m,U_t|\mathbf{Y}_t^o,\Psi}\left[U_t\mathbf{Y}_t\right] = \mathbb{E}_{U_t|\mathbf{Y}_t^o,\Psi}\left[U_t\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U_t,\Psi}\left[\mathbf{Y}_t\right]\right] = \mathbb{E}_{U_t|\mathbf{Y}_t^o,\Psi}\left[U_t\right]\begin{bmatrix}\mathbf{Y}_t^o\\ \boldsymbol{\mu}_m + \left(\mathbf{Y}_t^o - \boldsymbol{\mu}_o\right)\mathbf{C}_{oo}^{-1}\mathbf{C}_{om}\end{bmatrix}\\
&= \frac{v+d_o}{v+D_t^o}\begin{bmatrix}\mathbf{Y}_t^o\\ \boldsymbol{\mu}_m + \left(\mathbf{Y}_t^o - \boldsymbol{\mu}_o\right)\mathbf{C}_{oo}^{-1}\mathbf{C}_{om}\end{bmatrix},
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t,U_t|\mathbf{Y}_t^o,\Psi}\left[U_t\mathbf{Y}_t^T\mathbf{Y}_t\right]_{d\times d} &= \mathbb{E}_{\mathbf{Y}_t^m,U_t|\mathbf{Y}_t^o,\Psi}\left[U_t\mathbf{Y}_t^T\mathbf{Y}_t\right] = \mathbb{E}_{U_t|\mathbf{Y}_t^o,\Psi}\left[U_t\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U_t,\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right]\right]\\
&= \mathbf{Q} + \mathbb{E}_{U_t|\mathbf{Y}_t^o,\Psi}\left[U_t\right]\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U_t,\Psi}\left[\mathbf{Y}_t\right]^T\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U_t,\Psi}\left[\mathbf{Y}_t\right]\\
&= \mathbf{Q} + \frac{v+d_o}{v+D_t^o}\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U_t,\Psi}\left[\mathbf{Y}_t\right]^T\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U_t,\Psi}\left[\mathbf{Y}_t\right],
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t,U_t|\mathbf{Y}_t^o,\Psi}\left[U_t\mathbf{X}_t\right]_{1\times k} &= \mathbb{E}_{U_t|\mathbf{Y}_t^o,\Psi}\left[U_t\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U_t,\Psi}\left[\mathbb{E}_{\mathbf{X}_t|\mathbf{Y}_t,U_t,\Psi}\left(\mathbf{X}_t\right)\right]\right]\\
&= \mathbb{E}_{U_t|\mathbf{Y}_t^o,\Psi}\left[U_t\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U_t\Psi}\left[\mathbf{Y}_t\right]\right]\mathbf{W}\mathbf{M}^{-1} = \mathbb{E}_{U_t|\mathbf{Y}_t^o,\Psi}\left[U_t\right]\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U_t\Psi}\left[\mathbf{Y}_t\right]\mathbf{W}\mathbf{M}^{-1}\\
&= \frac{v+d_o}{v+D_t^o}\begin{bmatrix}\mathbf{Y}_t^o\\ \boldsymbol{\mu}_m + \left(\mathbf{Y}_t^o - \boldsymbol{\mu}_o\right)\mathbf{C}_{oo}^{-1}\mathbf{C}_{om}\end{bmatrix}\mathbf{W}\mathbf{M}^{-1},
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t,U_t|\mathbf{Y}_t^o,\Psi}\left[U_t\mathbf{X}_t^T\mathbf{X}_t\right]_{k\times k} &= \mathbb{E}_{U_t|\mathbf{Y}_t^o,\Psi}\left[U_t\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U_t,\Psi}\left[\mathbb{E}_{\mathbf{X}_t|\mathbf{Y}_t,U_t,\Psi}\left[\mathbf{X}_t^T\mathbf{X}_t\right]\right]\right]\\
&= \sigma^2\mathbf{M}^{-1} + \mathbf{M}^{-1}\mathbf{W}^T\mathbb{E}_{\mathbf{Y}_t^m,U_t|\mathbf{Y}_t^o,\Psi}\left[U_t\left(\mathbf{Y}_t - \boldsymbol{\mu}\right)^T\left(\mathbf{Y}_t - \boldsymbol{\mu}\right)\right]\mathbf{W}\mathbf{M}^{-1},
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t,U_t|\mathbf{Y}_t^o,\Psi}\left[U_t\left(\mathbf{Y}_t - \boldsymbol{\mu}\right)^T\left(\mathbf{Y}_t - \boldsymbol{\mu}\right)\right]_{d\times d} &= \mathbb{E}_{\mathbf{Y}_t^m,U_t|\mathbf{Y}_t^o,\Psi}\left[U_t\left(\mathbf{Y}_t - \boldsymbol{\mu}\right)^T\left(\mathbf{Y}_t - \boldsymbol{\mu}\right)\right]\\
&= \mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t,U_t|\mathbf{Y}_t^o,\Psi}\left[U_t\mathbf{Y}_t^T\mathbf{Y}_t\right] - 2\boldsymbol{\mu}^T\mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t,U_t|\mathbf{Y}_t^o,\Psi}\left[U_t\mathbf{Y}_t\right] + \mathbb{E}_{U_t|\mathbf{Y}_t^o,\Psi}\left[U_t\right]\boldsymbol{\mu}^T\boldsymbol{\mu},
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t,U_t|\mathbf{Y}_t^o,\Psi}\left[U_t\mathbf{Y}_t^T\mathbf{X}_t\right]_{d\times k} &= \mathbb{E}_{U_t|\mathbf{Y}_t^o,\Psi}\left[U_t\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U_t,\Psi}\left[\mathbf{Y}_t^T\mathbb{E}_{\mathbf{X}_t|\mathbf{Y}_t,U_t,\Psi}\left[\mathbf{X}_t\right]\right]\right] = \mathbb{E}_{U_t|\mathbf{Y}_t^o,\Psi}\left[\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U_t,\Psi}\left[U_t\mathbf{Y}_t^T\left(\mathbf{Y}_t - \boldsymbol{\mu}\right)\right]\right]\mathbf{W}\mathbf{M}^{-1}\\
&= \left(\mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t,U_t|\mathbf{Y}_t^o,\Psi}\left[U_t\mathbf{Y}_t^T\mathbf{Y}_t\right] - \mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t,U_t|\mathbf{Y}_t^o,\Psi}\left[U_t\mathbf{Y}_t\right]^T\boldsymbol{\mu}\right)\mathbf{W}\mathbf{M}^{-1}.
\end{aligned}$$

□

A.3.6. Proof of Theorem 22

Proof. The steps of proof of Theorem 22 follows the reasoning of the proof of Theorem 10 but the introduction of missingness to the PPCA setting requires applying the moments of $\mathbf{Y}_t^m, \mathbf{X}_t|\mathbf{Y}_t^o, \Psi$ derived in Theorem 21. The maximizers of the function $\tilde{Q}_{TS,ind}$ defined in Theorem 21 with respecto $\Psi^* = \left[\mathbf{W}^*, \boldsymbol{\mu}^*, \sigma^{*2}\right]$ are obtained as the solutions to the system of equations specified by $\nabla_{\Psi^*}\tilde{Q}_{TS,ind}(\Psi, \Psi^*) = \mathbf{0}$ and are the following

$$\begin{aligned}
&\frac{\partial\tilde{Q}_{TS,ind}(\Psi,\Psi^*)}{\partial\boldsymbol{\mu}^*} = -\frac{1}{2\sigma^*}\sum_{t=1}^{N}\left\{-2\mathbb{E}_{\mathbf{Y}_t^m,U_t|\mathbf{Y}_t^o,\Psi}\left[U_t\left(\mathbf{Y}_t - \boldsymbol{\mu}^*\right)\right] + 2\mathbb{E}_{\mathbf{X}_t,\mathbf{Y}_t^m,U_t|\mathbf{Y}_t^o,\Psi}\left[U_t\mathbf{X}_t\right]\mathbf{W}^{*T}\right\} = \mathbf{0}\\
&\iff \sum_{t=1}^{N}\left\{\frac{v+d_o}{v+D_t^o}\left(\mathbb{E}_{\mathbf{Y}_t^m,U_t|\mathbf{Y}_t^o,\Psi}\left[\mathbf{Y}_t\right] - \boldsymbol{\mu}^*\right) - \frac{v+d_o}{v+D_t^o}\left(\mathbb{E}_{\mathbf{Y}_t^m,U_t|\mathbf{Y}_t^o,\Psi}\left[\mathbf{Y}_t\right] - \boldsymbol{\mu}\right)\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\right\} = \mathbf{0}\\
&\iff \underbrace{\frac{1}{N}\sum_{t=1}^{N}\left(\mathbb{E}_{\mathbf{Y}_t^m,U_t|\mathbf{Y}_t^o,\Psi}\left[\mathbf{Y}_t\right]\frac{v+d_o}{v+D_t^o}\right)}_{\bar{\boldsymbol{\mu}}_{ts}}\left(\mathbb{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\right) + \underbrace{\frac{1}{N}\sum_{t=1}^{N}\left(\frac{v+d}{v+D_t^o}\right)}_{\bar{u}}\boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T} - \boldsymbol{\mu}^*\frac{1}{N}\sum_{t=1}^{N}\frac{v+d}{v+D_t^o} = \mathbf{0}\\
&\iff \boldsymbol{\mu}^* = \frac{1}{\bar{u}(\mathbf{y}_{1:N}^o;\Psi)}\left(\bar{\boldsymbol{\mu}}_{ts}(\mathbf{y}_{1:N}^o;\Psi)\left(\mathbb{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\right) + \bar{u}(\mathbf{y}_{1:N}^o;\Psi)\boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\right)
\end{aligned}$$

$$\frac{\partial \tilde{Q}_{TS,ind}(\Psi,\Psi^*)}{\partial \mathbf{W}^*} = -\frac{1}{2\sigma^{*2}}\sum_{t=1}^{N}\Big\{ -2\mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t,U_t|\mathbf{Y}_t^o,\Psi}\left[U_t\left(\mathbf{Y}_t-\boldsymbol{\mu}^*\right)^T\mathbf{X}_t\right] + 2\mathbf{W}^*\mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t,U|\mathbf{Y}_t^o,\Psi}\left[U_t\mathbf{X}_t^T\mathbf{X}_t\right]\Big\} = \mathbf{0}$$

$$\iff \sum_{t=1}^{N}\Big\{ \mathbb{E}_{\mathbf{Y}_t^m,U_t|\mathbf{Y}_t^o,\Psi}\left[U_t\left(\mathbf{Y}_t-\boldsymbol{\mu}^*\right)^T\left(\mathbf{Y}_t-\boldsymbol{\mu}\right)\right]\mathbf{W}\mathbf{M}^{-1} - \mathbf{W}^*\left(\sigma^2\mathbf{M}^{-1}+\mathbf{M}^{-1}\mathbf{W}^T\mathbb{E}_{\mathbf{Y}_t^m,U_t|\mathbf{Y}_t^o,\Psi}\left[U_t\left(\mathbf{Y}_t-\boldsymbol{\mu}\right)^T\left(\mathbf{Y}_t-\boldsymbol{\mu}\right)\right]\mathbf{W}\mathbf{M}^{-1}\right)\Big\} = \mathbf{0}$$

$$\iff \underbrace{\frac{1}{N}\sum_{t=1}^{N}\left(\mathbb{E}_{\mathbf{Y}_t^m,U_t|\mathbf{Y}_t^o,\Psi}\left[U_t\left(\mathbf{Y}_t-\boldsymbol{\mu}^*\right)^T\left(\mathbf{Y}_t-\boldsymbol{\mu}\right)\right]\right)}_{\bar{\mathbf{C}}^{ts}_{\mu,\mu^*}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)}\mathbf{W}\mathbf{M}^{-1}$$

$$-\mathbf{W}^*\Big(\sigma^2\mathbf{M}^{-1}+\mathbf{M}^{-1}\mathbf{W}^T\underbrace{\frac{1}{N}\sum_{t=1}^{N}\left(\mathbb{E}_{\mathbf{Y}_t^m,U_t|\mathbf{Y}_t^o,\Psi}\left[U_t\left(\mathbf{Y}_t-\boldsymbol{\mu}\right)^T\left(\mathbf{Y}_t-\boldsymbol{\mu}\right)\right]\right)}_{\bar{\mathbf{C}}^{ts}_{\mu}(\mathbf{y}^o_{1:N};\Psi)}\mathbf{W}\mathbf{M}^{-1}\Big) = \mathbf{0}$$

$$\iff \mathbf{W}^* = \bar{\mathbf{C}}^{ts}_{\mu,\mu^*}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\mathbf{W}\mathbf{M}^{-1}\left(\sigma^2\mathbf{M}^{-1}+\mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}^{ts}_{\mu}(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}\mathbf{M}^{-1}\right)^{-1}$$

$$\frac{\partial \tilde{Q}_{TS,ind}(\Psi,\Psi^*)}{\partial \sigma^{*2}} = -\frac{1}{2}\sum_{t=1}^{N}\Big\{\frac{d}{\sigma^{*2}} - \frac{1}{(\sigma^{*2})^2}\Big(\mathrm{Tr}\left\{\mathbb{E}_{\mathbf{Y}_t^m,U_t|\mathbf{Y}_t^o,\Psi}\left[U_t\left(\mathbf{Y}_t-\boldsymbol{\mu}^*\right)^T\left(\mathbf{Y}_t-\boldsymbol{\mu}^*\right)\right]\right\}$$

$$-2\,\mathrm{Tr}\left\{\mathbf{W}^*\mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t,U_t|\mathbf{Y}_t^o,\Psi}\left[U_t\mathbf{X}_t^T\left(\mathbf{Y}_t-\boldsymbol{\mu}^*\right)\right]\right\} + \mathrm{Tr}\left\{\mathbf{W}^{*T}\mathbf{W}^*\mathbb{E}_{\mathbf{Y}_t^m,\mathbf{X}_t,U_t|\mathbf{Y}_t^o,\Psi}\left[U_t\mathbf{X}_t^T\mathbf{X}_t\right]\right\}\Big)\Big\} = 0$$

$$\iff \sum_{t=1}^{N}\Big\{\frac{d}{\sigma^{*2}} - \frac{1}{(\sigma^{*2})^2}\Big(\mathrm{Tr}\left\{\mathbb{E}_{\mathbf{Y}_t^m,U_t|\mathbf{Y}_t^o,\Psi}\left[U_t\left(\mathbf{Y}_t-\boldsymbol{\mu}^*\right)^T\left(\mathbf{Y}_t-\boldsymbol{\mu}^*\right)\right]\right\} - 2\,\mathrm{Tr}\left\{\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\mathbb{E}_{\mathbf{Y}_t^m,U_t|\mathbf{Y}_t^o,\Psi}\left[U_t\left(\mathbf{Y}_t-\boldsymbol{\mu}\right)^T\left(\mathbf{Y}_t-\boldsymbol{\mu}^*\right)\right]\right\}$$

$$+\mathrm{Tr}\left\{\mathbf{W}^{*T}\mathbf{W}^*\left(\sigma^2\mathbf{M}^{-1}+\mathbf{M}^{-1}\mathbf{W}^T\mathbb{E}_{\mathbf{Y}_t^m,U_t|\mathbf{Y}_t^o,\Psi}\left[U_t\left(\mathbf{Y}_t-\boldsymbol{\mu}\right)^T\left(\mathbf{Y}_t-\boldsymbol{\mu}\right)\right]\mathbf{W}\mathbf{M}^{-1}\right)\right\}\Big)\Big\} = 0$$

$$\iff \frac{d}{\sigma^{*2}} - \frac{1}{(\sigma^{*2})^2}\,\mathrm{Tr}\Big\{\underbrace{\frac{1}{N}\sum_{t=1}^{N}\left(\mathbb{E}_{\mathbf{Y}_t^m,U_t|\mathbf{Y}_t^o,\Psi}\left[U_t\left(\mathbf{Y}_t-\boldsymbol{\mu}^*\right)^T\left(\mathbf{Y}_t-\boldsymbol{\mu}^*\right)\right]\right)}_{\bar{\mathbf{C}}^{ts}_{\mu^*}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)}$$

$$-2\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\underbrace{\frac{1}{N}\sum_{t=1}^{N}\left(\mathbb{E}_{\mathbf{Y}_t^m,U_t|\mathbf{Y}_t^o,\Psi}\left[U_t\left(\mathbf{Y}_t-\boldsymbol{\mu}\right)^T\left(\mathbf{Y}_t-\boldsymbol{\mu}^*\right)\right]\right)}_{\bar{\mathbf{C}}^{ts}_{\mu,\mu^*}(\mathbf{y}^o_{1:N};\Psi,\Psi)}$$

$$+\mathbf{W}^*\Big(\sigma^2\mathbf{M}^{-1}+\mathbf{M}^{-1}\mathbf{W}^T\underbrace{\frac{1}{N}\sum_{t=1}^{N}\left(\mathbb{E}_{\mathbf{Y}_t^m,U_t|\mathbf{Y}_t^o,\Psi}\left[U_t\left(\mathbf{Y}_t-\boldsymbol{\mu}\right)^T\left(\mathbf{Y}_t-\boldsymbol{\mu}\right)\right]\right)}_{\bar{\mathbf{C}}^{ts}_{\mu}(\mathbf{y}^o_{1:N};\Psi)}\mathbf{W}\mathbf{M}^{-1}\Big)\mathbf{W}^{*T}\Big\} = 0$$

$$\iff \sigma^{*2} = \frac{1}{d}\,\mathrm{Tr}\left\{\bar{\mathbf{C}}^{ts}_{\mu^*}(\mathbf{y}^o_{1:N};\Psi,\Psi^*) - 2\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}^{ts}_{\mu,\mu^*}(\mathbf{y}^o_{1:N};\Psi,\Psi)\right\} + \mathrm{Tr}\left\{\mathbf{W}^*\left(\sigma^2\mathbf{M}^{-1}+\mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{C}}^{ts}_{\mu}(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}\mathbf{M}^{-1}\right)\mathbf{W}^{*T}\right\}$$

where

$$\bar{u}(\mathbf{y}^o_{1:N};\Psi) = \frac{1}{N}\sum_{t=1}^{N}\frac{v+d_o}{v+D^o_t},$$

$$\bar{\boldsymbol{\mu}}_{ts}(\mathbf{y}^o_{1:N};\Psi) = \frac{1}{N}\sum_{t=1}^{N}\frac{v+d_o}{v+D^o_t}\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U_t,\Psi}\left(\mathbf{Y}_t\right),$$

$$\bar{\mathbf{S}}_{ts}(\mathbf{y}^o_{1:N};\Psi) = \frac{1}{N}\sum_{t=1}^{N}\frac{v+d_o}{v+D^o_t(\mathbf{y}^o_{1:N};\Psi)}\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U_t,\Psi}\left(\mathbf{Y}_t\right)^T\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U_t,\Psi}\left(\mathbf{Y}_t\right),$$

$$\bar{\mathbf{Q}} = \begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{W}_m\mathbf{W}_m^T+\sigma^2\mathbb{I}_{d_m}-\mathbf{W}_m\mathbf{W}_o^T\left(\mathbf{W}_o\mathbf{W}_o^T+\sigma^2\mathbb{I}_{d_o}\right)^{-1}\mathbf{W}_o\mathbf{W}_m^T\end{bmatrix},$$

$$\bar{\mathbf{C}}^{ts}_{\mu}(\mathbf{y}^o_{1:N};\Psi) = \bar{\mathbf{S}}_{ts}(\mathbf{y}^o_{1:N};\Psi) + \bar{\mathbf{Q}} - \boldsymbol{\mu}^T\bar{\boldsymbol{\mu}}_{ts}(\mathbf{y}^o_{1:N};\Psi) - \bar{\boldsymbol{\mu}}_{ts}(\mathbf{y}^o_{1:N};\Psi)^T\boldsymbol{\mu} + \bar{u}(\mathbf{y}^o_{1:N};\Psi)\boldsymbol{\mu}^T\boldsymbol{\mu},$$

$$\bar{\mathbf{C}}^{ts}_{\mu^*}(\mathbf{y}^o_{1:N};\Psi,\Psi^*) = \bar{\mathbf{S}}_{ts}(\mathbf{y}^o_{1:N};\Psi) + \bar{\mathbf{Q}} - \boldsymbol{\mu}^{*T}\bar{\boldsymbol{\mu}}_{ts}(\mathbf{y}^o_{1:N};\Psi) - \bar{\boldsymbol{\mu}}_{ts}(\mathbf{y}^o_{1:N};\Psi)^T\boldsymbol{\mu}^* + \bar{u}(\mathbf{y}^o_{1:N};\Psi)\boldsymbol{\mu}^{*T}\boldsymbol{\mu}^*,$$

$$\bar{\mathbf{C}}^{ts}_{\mu,\mu^*}(\mathbf{y}^o_{1:N};\Psi,\Psi^*) = \bar{\mathbf{S}}_{ts}(\mathbf{y}^o_{1:N};\Psi) + \bar{\mathbf{Q}} - \boldsymbol{\mu}^{*T}\bar{\boldsymbol{\mu}}_{ts}(\mathbf{y}^o_{1:N};\Psi) - \bar{\boldsymbol{\mu}}_{ts}(\mathbf{y}^o_{1:N};\Psi)^T\boldsymbol{\mu} + \bar{u}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\boldsymbol{\mu}^{*T}\boldsymbol{\mu},$$

for $\mathbf{M}_{k\times k} = \mathbf{W}^T\mathbf{W}+\sigma^2\mathbb{I}_k$, $D^o_t(\mathbf{y}^o_{1:N};\Psi) = \left(\mathbf{y}^o_t-\boldsymbol{\mu}_o\right)\mathbf{C}_{oo}^{-1}\left(\mathbf{y}^o_t-\boldsymbol{\mu}_o\right)^T$ where $\mathbf{C}_{oo} = \mathbf{W}_o\mathbf{W}_o^T+\sigma^2\mathbb{I}_{d_o}$. □

A.3.7. Proof of Theorem 23

Proof. The steps of proof of Theorem 23 follows the reasoning of the proof of Theorem 11. Given the N realisation of the vector with observed values, $\mathbf{y}^o_{1:N} = \mathbf{y}^o_1, \ldots, \mathbf{y}^o_N$ and recalling that the observation vector is partitioned into two subvectors, $\mathbf{Y}_t = [\mathbf{Y}^o_t, \mathbf{Y}^m_t]$ with observed and unobserved entries of $\mathbf{Y}_t$, the E-step from Theorem 23 is obtained by

$$
\begin{aligned}
\tilde{Q}_{TS,id}(\Psi, \Psi^*) &= \mathbb{E}_{\mathbf{Y}^m_{1:N}, \mathbf{X}_{1:N}, U|\mathbf{Y}^o_{1:N}, \Psi}\Big[\log \pi_{\mathbf{Y}_{1:N}, \mathbf{X}_{1:N}, U|\Psi^*}(\mathbf{Y}_{1:N}, \mathbf{X}_{1:N}, U)\Big] \\
&= \int_{\mathbb{R}_+}\int_{\mathbb{R}^{N\times d_m}}\int_{\mathbb{R}^{N\times k}} \left\{ \log\Big(\pi_{\mathbf{Y}_{1:N}, \mathbf{X}_{1:N}, U|\Psi^*}(\mathbf{y}_{1:N}, \mathbf{x}_{1:N}, u)\Big) \times \frac{\pi_{\mathbf{Y}_{1:N}, \mathbf{X}_{1:N}, U|\Psi}(\mathbf{y}_{1:N}, \mathbf{x}_{1:N}, u)}{\pi_{\mathbf{Y}^o_{1:N}|\Psi}(\mathbf{y}^o_{1:N})} \right\} d\mathbf{x}_{1:N}\, d\mathbf{y}^m_{1:N}\, du \\
&= \frac{1}{\pi_{\mathbf{Y}^o_{1:N}|\Psi}(\mathbf{y}^o_{1:N})}\int_{\mathbb{R}_+} \pi_{U|\Psi}(u) \int_{\mathbb{R}^{N\times d_m}}\int_{\mathbb{R}^{N\times k}} \left\{ \log\Big(\pi_{\mathbf{Y}_{1:N}, \mathbf{X}_{1:N}, U|\Psi^*}(\mathbf{y}_{1:N}, \mathbf{x}_{1:N}, u)\Big) \right. \\
&\quad \left. \times\, \pi_{\mathbf{Y}_{1:N}|\mathbf{X}_{1:N}, U, \Psi}(\mathbf{y}_{1:N})\pi_{\mathbf{X}_{1:N}|U,\Psi}(\mathbf{x}_{1:N}) \right\} d\mathbf{x}_{1:N}\, \mathbf{y}^m_{1:N}\, du \xrightarrow{\text{applying Lemma A2 and A3}} \\
&= \frac{1}{\pi_{\mathbf{Y}^o_{1:N}|\Psi}(\mathbf{y}^o_{1:N})}\int_{\mathbb{R}_+} \pi_{U|\Psi}(u) \int_{\mathbb{R}^{N\times d_m}} \prod_{s=1}^{N} h(\mathbf{y}_s, u; \Psi)\Big(\log \pi_{U|\Psi^*}(u) + \sum_{t=1}^{N} w(\mathbf{y}_t, u; \Psi, \Psi^*)\Big)\, d\mathbf{y}^m_{1:N}\, du \\
&= \frac{1}{\pi_{\mathbf{Y}^o_{1:N}|\Psi}(\mathbf{y}^o_{1:N})}\int_{\mathbb{R}_+} \pi_{U|\Psi}(u) \left\{ \int_{\mathbb{R}^{N\times d_m}} C_H(\Psi)\cdot\Big(\prod_{s=1}^{N}\tilde{\pi}_{\mathbf{Y}_s|U,\Psi}(\mathbf{y}_s)\Big)\log \pi_{U|\Psi^*}(u)\mathbf{y}^m_{1:N} \right. \\
&\quad \left. + \int_{\mathbb{R}^{N\times d_m}} C_H(\Psi)\cdot\Big(\prod_{s=1}^{N}\tilde{\pi}_{\mathbf{Y}_s|U,\Psi}(\mathbf{y}_s)\Big)\Big(\sum_{t=1}^{N} w(\mathbf{y}_t, u; \Psi, \Psi^*)\Big)\, d\mathbf{y}^m_{1:N} \right\} du
\end{aligned}
$$

where $C_H(\Psi)$ and $\tilde{\pi}_{\mathbf{Y}_t|U,\Psi}(\mathbf{y}_t)$ are specified in Lemma A5. Given the partition of the observation vector $\mathbf{Y}_t$ into the vector of observed and missing entries in time t, $\mathbf{Y}^o_t$ and $\mathbf{Y}^m_t$, respectively, that is $\mathbf{Y}_t = [\mathbf{Y}^o_t, \mathbf{Y}^m_t]$ and applying Chain rule of probability, we obtain the following equality

$$
\tilde{\pi}_{\mathbf{Y}_t|U,\Psi}(\mathbf{y}_t) = \tilde{\pi}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,U,\Psi}(\mathbf{y}^m_t)\cdot\tilde{\pi}_{\mathbf{Y}^o_t|U,\Psi}(\mathbf{y}^o_t)
$$

we have that

$$
\begin{aligned}
\tilde{Q}_{TS,id}(\Psi, \Psi^*) &= \frac{C_H(\Psi)}{\pi_{\mathbf{Y}^o_{1:N}|\Psi}(\mathbf{y}^o_{1:N})}\int_{\mathbb{R}_+} \pi_{U|\Psi}(u)\int_{\mathbb{R}^{N\times d_m}} \left\{ \Big(\log \pi_{U|\Psi^*}(u) + \sum_{t=1}^{N} w(\mathbf{y}_t, u; \Psi, \Psi^*)\Big)\prod_{t=1}^{N}\tilde{\pi}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,U,\Psi}(\mathbf{y}^m_t)\tilde{\pi}_{\mathbf{Y}^o_t|U,\Psi}(\mathbf{y}^o_t) \right\} d\mathbf{y}^m_{1:N}\, du \\
&= \frac{C_H(\Psi)}{\pi_{\mathbf{Y}^o_{1:N}|\Psi}(\mathbf{y}^o_{1:N})}\left\{ \int_{\mathbb{R}_+} \pi_{U|\Psi}(u)\Big(\prod_{t=1}^{N}\tilde{\pi}_{\mathbf{Y}^o_t|U,\Psi}(\mathbf{y}^o_t)\Big)\log \pi_{U|\Psi^*}(u) \underbrace{\int_{\mathbb{R}^{N\times d_m}}\prod_{t=1}^{N}\tilde{\pi}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,U,\Psi}(\mathbf{y}^m_t)\, d\mathbf{y}^m_{1:N}}_{=1} du \right. \\
&\quad \left. + \int_{\mathbb{R}_+} \pi_{U|\Psi}(u)\Big(\prod_{t=1}^{N}\tilde{\pi}_{\mathbf{Y}^o_t|U,\Psi}(\mathbf{y}^o_t)\Big)\int_{\mathbb{R}^{N\times d_m}}\Big(\sum_{t=1}^{N} w(\mathbf{y}_t, u; \Psi, \Psi^*)\Big)\prod_{t=1}^{N}\tilde{\pi}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,U,\Psi}(\mathbf{y}^m_t)\, d\mathbf{y}^m_{1:N}\, du \right\} \\
&= \frac{C_H(\Psi)}{\pi_{\mathbf{Y}^o_{1:N}|\Psi}(\mathbf{y}^o_{1:N})}\left\{ \int_{\mathbb{R}_+} \pi_{U|\Psi}(u)\Big(\prod_{t=1}^{N}\tilde{\pi}_{\mathbf{Y}^o_t|U,\Psi}(\mathbf{y}^o_t)\Big)\log \pi_{U|\Psi^*}(u)\, du \right. \\
&\quad \left. + \int_{\mathbb{R}_+} \pi_{U|\Psi}(u)\Big(\prod_{t=1}^{N}\tilde{\pi}_{\mathbf{Y}^o_t|U,\Psi}(\mathbf{y}^o_t)\Big)\sum_{t=1}^{N}\left\{ \int_{\mathbb{R}^{d_m}} w(\mathbf{y}_t, u; \Psi, \Psi^*)\tilde{\pi}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,U,\Psi}(\mathbf{y}^m_t)\, d\mathbf{y}^m_t \right\} du \right\}
\end{aligned}
$$

Applying Lemma A6 and A7, we obtain the following simplified form of the function $\tilde{Q}_{TS,id}$

$$
\tilde{Q}_{TS,id}(\Psi, \Psi^*) = \frac{C_H(\Psi)C_\beta(\mathbf{y}^o_{1:N}, \Psi)}{\pi_{\mathbf{Y}^o_{1:N}|\Psi}(\mathbf{y}^o_{1:N})}\int_{\mathbb{R}_+} \tilde{\pi}_{U|\Psi}(u)\Big(\log \pi_{U|\Psi^*}(u) + \sum_{t=1}^{N}\tilde{w}(\mathbf{y}^o_t; \Psi, \Psi^*)\Big)du.
$$

where $\tilde{w}(\mathbf{y}^o_t; \Psi, \Psi^*)$ is derived in Lemma A6 and the density function $\tilde{\pi}_{U|\Psi}(u)$ is given in Lemma A7. □

A.3.8. Proof of Theorem 24

Proof. The maximizers of $\tilde{Q}_{TS,id}$ defined in Theorem 23 with respect to the vector of static parameters $\Psi^* = [\mathbf{W}^*, \boldsymbol{\mu}^*, \sigma^{*2}]$ are calculates as

$$
\nabla_{\Psi^*}\tilde{Q}_{TS,id}(\Psi, \Psi^*) = \frac{\partial}{\partial \Psi^*}\tilde{C}(\mathbf{y}^o_{1:N}, \Psi)\left\{ \int_{\mathbb{R}_+} \tilde{\pi}_{U|\Psi}(u)\log \pi_{U|\Psi^*}(u)du + \int_{\mathbb{R}_+} \tilde{\pi}_{U|\Psi}(u)\sum_{t=1}^{N}\tilde{w}(\mathbf{y}^o_t, u; \Psi, \Psi^*)\, du \right\}
$$

$$= \tilde{C}(\mathbf{y}^o_{1:N}, \Psi)\left\{ \int_{\mathbb{R}_+} \tilde{\pi}_{U|\Psi}(u) \underbrace{\frac{\partial}{\partial \Psi^*} \log \pi_{U|\Psi^*}(u)}_{=0 \text{ as includes only } v^*} du + \int_{\mathbb{R}_+} \tilde{\pi}_{U|\Psi}(u) \frac{\partial}{\partial \Psi^*} \sum_{t=1}^{N} \tilde{w}(\mathbf{y}^o_t, u; \Psi, \Psi^*)\, du \right\}$$

$$= \tilde{C}(\mathbf{y}^o_{1:N}, \Psi) \int_{\mathbb{R}_+} \tilde{\pi}_{U|\Psi}(u) \sum_{t=1}^{N} \frac{\partial}{\partial \Psi^*} \tilde{w}(\mathbf{y}^o_t, u; \Psi, \Psi^*)\, du$$

for $\tilde{C}(\mathbf{y}^o_{1:N}, \Psi) = \frac{C_H(\Psi) C_\beta(\mathbf{y}^o_{1:N}, \Psi)}{\pi_{\mathbf{Y}^o_{1:N}|\Psi}(\mathbf{y}^o_{1:N})}$. We applied Leibniz integral rule to swap the order of differentiating and integrating as the function $f(\Psi^*, u) = \tilde{\pi}_{U|\Psi}(u)\Big(\log \pi_{U|\Psi^*}(u) + \sum_{t=1}^{N} w(\mathbf{y}_t, u; \Psi, \Psi^*)\Big)$ is a linear function of u. It is integrable with respect to u and differentiable at Ψ^* and its gradient $\nabla_{\Psi^*} f$ is bounded by a function of u. Next we calculate the gradient of $\sum_{t=1}^{N} \tilde{w}(\mathbf{y}^o_t, u; \Psi, \Psi^*)$ with respect to $[\mathbf{W}^*, \boldsymbol{\mu}^*, \sigma^{*2}]$, that is

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\mu}^*} \sum_{t=1}^{N} \tilde{w}(\mathbf{y}^o_t, u; \Psi, \Psi^*) &= -\frac{u}{2\sigma^{*2}} \sum_{t=1}^{N} -2\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t, U, \Psi}\left[\mathbf{Y}_t\right] + 2\boldsymbol{\mu}^* + 2\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t, U, \Psi}\left[\mathbf{Y}_t\right]\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T} - 2\boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T} \\
&= \frac{u}{\sigma^{*2}} \sum_{t=1}^{N} \mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t, U, \Psi}\left[\mathbf{Y}_t\right]\left(\mathbf{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\right) - \boldsymbol{\mu}^* + \boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T} \\
&= u\frac{N}{\sigma^{*2}}\left(\underbrace{\frac{1}{N}\sum_{t=1}^{N} \mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t, U, \Psi}\left[\mathbf{Y}_t\right]}_{\bar{\boldsymbol{\mu}}(\mathbf{y}^o_{1:N}; \Psi)} \left(\mathbf{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\right) - \boldsymbol{\mu}^* + \boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\right) \\
&= u\frac{N}{\sigma^{*2}}\left(\bar{\boldsymbol{\mu}}(\mathbf{y}^o_{1:N}; \Psi,)\left(\mathbf{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\right) - \boldsymbol{\mu}^* + \boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\right)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{W}^*} \sum_{t=1}^{N} \tilde{w}(\mathbf{y}^o_t, u; \Psi, \Psi^*) &= \sum_{t=1}^{N} \left\{ \frac{u}{\sigma^{*2}}\left(\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t, U, \Psi}\left[\mathbf{Y}_t\mathbf{Y}_t^T\right] - \mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t, U, \Psi}\left[\mathbf{Y}_t\right]\left(\boldsymbol{\mu} + \boldsymbol{\mu}^*\right)^T + \boldsymbol{\mu}^{*\,T}\boldsymbol{\mu}\right)\mathbf{W}\mathbf{M}^{-1} \right. \\
&\quad \left. - \frac{u}{\sigma^{*2}}\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\left(\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t, U, \Psi}\left[\mathbf{Y}_t\mathbf{Y}_t^T\right] - 2\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t, U, \Psi}\left[\mathbf{Y}_t\right]\boldsymbol{\mu}^T + \boldsymbol{\mu}\boldsymbol{\mu}^T\right)\mathbf{W}\mathbf{M}^{-1} - \frac{\sigma^2}{\sigma^{*2}}\mathbf{W}^*\mathbf{M}^{-1} \right\} \\
&= \frac{N}{\sigma^{*2}}\left(\frac{u}{N}\sum_{t=1}^{N}\left\{ \frac{\sigma^2}{u}\mathbf{Q} + \mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t, U, \Psi}\left[\mathbf{Y}_t\right]^T\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t, U, \Psi}\left[\mathbf{Y}_t\right] - \mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t, U, \Psi}\left[\mathbf{Y}_t\right]\left(\boldsymbol{\mu} + \boldsymbol{\mu}^*\right)^T + \boldsymbol{\mu}^{*\,T}\boldsymbol{\mu}\right\}\mathbf{W}\mathbf{M}^{-1} \right. \\
&\quad \left. - \mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\frac{u}{N}\sum_{t=1}^{N}\left\{ \frac{\sigma^2}{u}\mathbf{Q} + \mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t, U, \Psi}\left[\mathbf{Y}_t\right]^T\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t, U, \Psi}\left[\mathbf{Y}_t\right] - 2\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t, U, \Psi}\left[\mathbf{Y}_t\right]\boldsymbol{\mu}^T + \boldsymbol{\mu}\boldsymbol{\mu}^T\right\}\mathbf{W}\mathbf{M}^{-1} - \sigma^2\mathbf{W}^*\mathbf{M}^{-1}\right) \\
&= \frac{N}{\sigma^{*2}}\left(\sigma^2 \underbrace{\frac{1}{N}\sum_{t=1}^{N}\mathbf{Q}}_{\bar{\mathbf{Q}}}\mathbf{M}^{-1}\mathbf{W}^T - \sigma^2\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\frac{1}{N}\sum_{t=1}^{N}\mathbf{Q}\,\mathbf{W}\mathbf{M}^{-1} - \sigma^2\mathbf{W}^*\mathbf{M}^{-1} \right. \\
&\quad + u\underbrace{\frac{1}{N}\sum_{t=1}^{N}\left\{ \mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t, U, \Psi}\left[\mathbf{Y}_t\right]^T\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t, U, \Psi}\left[\mathbf{Y}_t\right] - \mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t, U, \Psi}\left[\mathbf{Y}_t\right]^T\left(\boldsymbol{\mu} + \boldsymbol{\mu}^*\right) + \boldsymbol{\mu}^{*T}\boldsymbol{\mu}\right\}}_{\bar{\mathbf{C}}_{\mu,\mu^*}(\mathbf{y}^o_{1:N}; \Psi, \Psi^*)}\mathbf{W}\mathbf{M}^{-1} \\
&\quad \left. - u\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\underbrace{\frac{1}{N}\sum_{t=1}^{N}\left\{ \mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t, U, \Psi}\left[\mathbf{Y}_t\right]^T\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t, U, \Psi}\left[\mathbf{Y}_t\right] - 2\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t, U, \Psi}\left[\mathbf{Y}_t\right]^T\boldsymbol{\mu} + \boldsymbol{\mu}^T\boldsymbol{\mu}\right\}}_{\bar{\mathbf{C}}_{\mu}(\mathbf{y}^o_{1:N}; \Psi, \Psi^*)}\mathbf{W}\mathbf{M}^{-1}\right) \\
&= \frac{N}{\sigma^{*2}}\left(\left(\sigma^2\bar{\mathbf{Q}} + u\bar{\mathbf{C}}_{\mu,\mu^*}(\mathbf{y}^o_{1:N}; \Psi, \Psi^*)\right)\mathbf{W}\mathbf{M}^{-1} - \mathbf{W}^*\left(\mathbf{M}^{-1}\mathbf{W}^T\left(\sigma^2\bar{\mathbf{Q}} + u\bar{\mathbf{C}}_{\mu}(\mathbf{y}^o_{1:N}; \Psi, \Psi^*)\right)\mathbf{W}\mathbf{M}^{-1} + \sigma^2\mathbf{M}^{-1}\right)\right)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial \sigma^{*2}} \sum_{t=1}^{N} \tilde{w}(\mathbf{y}^o_t, u; \Psi, \Psi^*) &= -\frac{dN}{2\sigma^{*2}} + \frac{1}{2(\sigma^{*2})^2}\sum_{t=1}^{N}\left(\sigma^2\,\mathrm{Tr}\left\{\mathbf{M}^{-1}\mathbf{W}^{*\,T}\mathbf{W}^*\right\} + u\left(\mathrm{Tr}\left\{ \mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t, U, \Psi}\left[\mathbf{Y}_t\mathbf{Y}_t^T\right] - 2\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t, U, \Psi}\left[\mathbf{Y}_t\right]\boldsymbol{\mu}^{*\,T} + \boldsymbol{\mu}^*\boldsymbol{\mu}^{*\,T}\right\} \right. \right. \\
&\quad - 2\,\mathrm{Tr}\left\{ \left(\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t, U, \Psi}\left[\mathbf{Y}_t\mathbf{Y}_t^T\right] - \mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t, U, \Psi}\left[\mathbf{Y}_t\right]\left(\boldsymbol{\mu} + \boldsymbol{\mu}^*\right)^T + \boldsymbol{\mu}^{*T}\boldsymbol{\mu}\right)\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T \right\} \\
&\quad \left. \left. + \mathrm{Tr}\left\{ \mathbf{M}^{-1}\mathbf{W}^T\left(\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t, U, \Psi}\left[\mathbf{Y}_t\mathbf{Y}_t^T\right] - 2\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t, U, \Psi}\left[\mathbf{Y}_t\right]\boldsymbol{\mu}^T + \boldsymbol{\mu}\boldsymbol{\mu}^T\right)\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{W}^* \right\} \right) \right) \\
&= -\frac{dN}{2\sigma^{*2}} + \frac{N}{2(\sigma^{*2})^2}\left(\sigma^2\,\mathrm{Tr}\left\{\mathbf{M}^{-1}\mathbf{W}^{*\,T}\mathbf{W}^*\right\} + \mathrm{Tr}\left\{ \sigma^2\bar{\mathbf{Q}} + u\bar{\mathbf{C}}_{\mu^*}(\mathbf{y}^o_{1:N}; \Psi, \Psi^*) \right\} \right. \\
&\quad \left. - 2\,\mathrm{Tr}\left\{ \left(\sigma^2\bar{\mathbf{Q}} + u\bar{\mathbf{C}}_{\mu,\mu^*}(\mathbf{y}^o_{1:N}; \Psi, \Psi^*)\right)\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T \right\} \right.
\end{aligned}$$

$$+ \operatorname{Tr}\left\{\mathbf{M}^{-1}\mathbf{W}^T\Big(\sigma^2\bar{\mathbf{Q}} + u\bar{\mathbf{C}}_\mu(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\Big)\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{W}^*\right\}\Bigg)$$

for the following notation

$$\begin{aligned}
\bar{\boldsymbol{\mu}}(\mathbf{y}^o_{1:N};\Psi) &= \frac{1}{N}\sum_{t=1}^N \mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,U,\Psi}\big[\mathbf{Y}_t\big],\\
\bar{\mathbf{S}}(\mathbf{y}^o_{1:N};\Psi) &= \frac{1}{N}\sum_{t=1}^N \mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,U,\Psi}\big[\mathbf{Y}_t\big]^T\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,U,\Psi}\big[\mathbf{Y}_t\big]\\
\bar{\mathbf{Q}} &= \frac{1}{N}\sum_{t=1}^N\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{V}^{-1}_{mm} - \mathbf{V}^{-1}_{mo}\mathbf{V}_{oo}\mathbf{V}^{-1}_{om}\end{bmatrix},\\
\bar{\mathbf{C}}_{\mu,\mu^*}(\mathbf{y}^o_{1:N};\Psi,\Psi^*) &= \bar{\mathbf{S}}(\mathbf{y}^o_{1:N};\Psi) - \boldsymbol{\mu}^{*T}\bar{\boldsymbol{\mu}}(\mathbf{y}^o_{1:N};\Psi) - \bar{\boldsymbol{\mu}}(\mathbf{y}^o_{1:N};\Psi)^T\boldsymbol{\mu}^T + \boldsymbol{\mu}^{*T}\boldsymbol{\mu},\\
\bar{\mathbf{C}}_{\mu}(\mathbf{y}^o_{1:N};\Psi,\Psi^*) &= \bar{\mathbf{S}}(\mathbf{y}^o_{1:N};\Psi) - \bar{\boldsymbol{\mu}}(\mathbf{y}^o_{1:N};\Psi)^T\boldsymbol{\mu} - \boldsymbol{\mu}^T\bar{\boldsymbol{\mu}}(\mathbf{y}^o_{1:N};\Psi) + \boldsymbol{\mu}^T\boldsymbol{\mu},\\
\bar{\mathbf{C}}_{\mu^*}(\mathbf{y}^o_{1:N};\Psi,\Psi^*) &= \bar{\mathbf{S}}(\mathbf{y}^o_{1:N};\Psi) - \bar{\boldsymbol{\mu}}(\mathbf{y}^o_{1:N};\Psi)^T\boldsymbol{\mu}^* - \boldsymbol{\mu}^{*T}\bar{\boldsymbol{\mu}}(\mathbf{y}^o_{1:N};\Psi) + \boldsymbol{\mu}^{*T}\boldsymbol{\mu}^*.
\end{aligned}$$

Hence, the maximizers of $\tilde{Q}_{TS,id}$ with respect to the vector of static parameters $\Psi^* = [\mathbf{W}^*, \boldsymbol{\mu}^*, \sigma^{*2}]$ are calculated by solving the set of equation $\nabla_{\Psi^*}\tilde{Q}_{TS,id} = \mathbf{0}$ what results in the following

$$\begin{aligned}
&\frac{\tilde{Q}_{TS,id}(\Psi,\Psi^*)}{\partial\boldsymbol{\mu}^*} = \tilde{C}(\mathbf{y}^o_{1:N},\Psi)\int_{\mathbb{R}_+}\tilde{\pi}_{U|\Psi}(u)\sum_{t=1}^N\frac{\partial}{\partial\boldsymbol{\mu}^*}\tilde{w}(\mathbf{y}^o_t,u;\Psi,\Psi^*)\,du = \mathbf{0}\\
&\iff \frac{N}{\sigma^{*2}}\int_{\mathbb{R}_+} u\,\tilde{\pi}_{U|\Psi}(u)\left(\bar{\boldsymbol{\mu}}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\Big(\mathbf{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\Big) - \boldsymbol{\mu}^* + \boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\right)du = \mathbf{0}\\
&\iff \left(\bar{\boldsymbol{\mu}}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\Big(\mathbf{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\Big) - \boldsymbol{\mu}^* + \boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\right)\int_{\mathbb{R}_+} u\,\tilde{\pi}_{U|\Psi}(u)\,du = \mathbf{0}\\
&\iff \bar{\boldsymbol{\mu}}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\Big(\mathbf{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\Big) - \boldsymbol{\mu}^* + \boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T} = \mathbf{0}\\
&\iff \boldsymbol{\mu}^* = \bar{\boldsymbol{\mu}}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\Big(\mathbf{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\Big) + \boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}
\end{aligned}$$

$$\begin{aligned}
&\frac{\tilde{Q}_{TS,id}(\Psi,\Psi^*)}{\partial\mathbf{W}^*} = \tilde{C}(\mathbf{y}^o_{1:N},\Psi)\int_{\mathbb{R}_+}\tilde{\pi}_{U|\Psi}(u)\sum_{t=1}^N\frac{\partial}{\partial\mathbf{W}^*}\tilde{w}(\mathbf{y}^o_t,u;\Psi,\Psi^*)\,du = \mathbf{0}\\
&\iff \int_{\mathbb{R}_+}\tilde{\pi}_{U|\Psi}(u)\Big(\sigma^2\bar{\mathbf{Q}} + u\bar{\mathbf{C}}_{\mu,\mu^*}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\Big)\mathbf{W}\mathbf{M}^{-1}du\\
&\quad - \int_{\mathbb{R}_+}\tilde{\pi}_{U|\Psi}(u)\mathbf{W}^*\left(\mathbf{M}^{-1}\mathbf{W}^T\Big(\sigma^2\bar{\mathbf{Q}} + u\bar{\mathbf{C}}_{\mu}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\Big)\mathbf{W}\mathbf{M}^{-1} + \sigma^2\mathbf{M}^{-1}\right)du = \mathbf{0}\\
&\iff \Big(\sigma^2\bar{\mathbf{Q}} + \underbrace{\int_{\mathbb{R}_+} u\,\tilde{\pi}_{U|\Psi}(u)\,du}_{A(\mathbf{y}^o_{1:N},\Psi,\Psi^*)}\bar{\mathbf{C}}_{\mu,\mu^*}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\Big)\mathbf{W}\mathbf{M}^{-1}\\
&\quad - \mathbf{W}^*\left(\mathbf{M}^{-1}\mathbf{W}^T\Big(\sigma^2\bar{\mathbf{Q}} + \underbrace{\int_{\mathbb{R}_+} u\,\tilde{\pi}_{U|\Psi}(u)\,du}_{A(\mathbf{y}^o_{1:N},\Psi,\Psi^*)}\bar{\mathbf{C}}_{\mu}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\Big)\mathbf{W}\mathbf{M}^{-1} + \sigma^2\mathbf{M}^{-1}\right) = \mathbf{0}\\
&\iff \Big(\sigma^2\bar{\mathbf{Q}} + A(\mathbf{y}^o_{1:N},\Psi,\Psi^*)\bar{\mathbf{C}}_{\mu,\mu^*}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\Big)\mathbf{W}\mathbf{M}^{-1}\\
&\quad - \mathbf{W}^*\left(\mathbf{M}^{-1}\mathbf{W}^T\Big(\sigma^2\bar{\mathbf{Q}} + A(\mathbf{y}^o_{1:N},\Psi,\Psi^*)\bar{\mathbf{C}}_{\mu}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\Big)\mathbf{W}\mathbf{M}^{-1} + \sigma^2\mathbf{M}^{-1}\right) = \mathbf{0}\\
&\iff \mathbf{W}^* = \Big(\sigma^2\bar{\mathbf{Q}} + A(\mathbf{y}^o_{1:N},\Psi,\Psi^*)\bar{\mathbf{C}}_{\mu,\mu^*}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\Big)\mathbf{W}\mathbf{M}^{-1}\\
&\quad \times\left(\mathbf{M}^{-1}\mathbf{W}^T\Big(\sigma^2\bar{\mathbf{Q}} + A(\mathbf{y}^o_{1:N},\Psi,\Psi^*)\bar{\mathbf{C}}_{\mu}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\Big)\mathbf{W}\mathbf{M}^{-1} + \sigma^2\mathbf{M}^{-1}\right)^{-1}
\end{aligned}$$

$$\begin{aligned}
&\frac{\tilde{Q}_{TS,id}(\Psi,\Psi^*)}{\partial\sigma^{*2}} = \tilde{C}(\mathbf{y}^o_{1:N},\Psi)\int_{\mathbb{R}_+}\tilde{\pi}_{U|\Psi}(u)\sum_{t=1}^N\frac{\partial}{\partial\sigma^{*2}}\tilde{w}(\mathbf{y}^o_t,u;\Psi,\Psi^*)\,du = \mathbf{0}\\
&\iff \int_{\mathbb{R}_+}\tilde{\pi}_{U|\Psi}(u)\Bigg\{-\frac{dN}{2\sigma^{*2}} + \frac{N}{2(\sigma^{*2})^2}\left(\sigma^2\operatorname{Tr}\Big\{\mathbf{M}^{-1}\mathbf{W}^{*\,T}\mathbf{W}^*\Big\} + \operatorname{Tr}\Big\{\sigma^2\bar{\mathbf{Q}} + u\bar{\mathbf{C}}_{\mu^*}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\Big\}\right.
\end{aligned}$$

$$
\begin{aligned}
&- 2\,\mathrm{Tr}\left\{\left(\sigma^2\bar{\mathbf{Q}} + u\bar{\mathbf{C}}_{\mu,\mu^*}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\right)\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\right\}\\
&+ \mathrm{Tr}\left\{\mathbf{M}^{-1}\mathbf{W}^T\left(\sigma^2\bar{\mathbf{Q}} + u\bar{\mathbf{C}}_{\mu}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\right)\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{W}^*\right\}\Bigg)\Bigg\}\,du = 0\\
\iff &-\frac{d}{\sigma^{*2}} + \frac{1}{(\sigma^{*2})^2}\Bigg(\sigma^2\,\mathrm{Tr}\left\{\mathbf{M}^{-1}\mathbf{W}^{*\,T}\mathbf{W}^*\right\} + \mathrm{Tr}\left\{\sigma^2\bar{\mathbf{Q}} + \underbrace{\int_{\mathbb{R}_+} u\;\tilde{\pi}_{U|\Psi}(u)\,du}_{A(\mathbf{y}^o_{1:N},\Psi,\Psi^*)}\bar{\mathbf{C}}_{\mu^*}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\right\}\\
&- 2\,\mathrm{Tr}\left\{\left(\sigma^2\bar{\mathbf{Q}} + \underbrace{\int_{\mathbb{R}_+} u\;\tilde{\pi}_{U|\Psi}(u)\,du}_{A(\mathbf{y}^o_{1:N},\Psi,\Psi^*)}\bar{\mathbf{C}}_{\mu,\mu^*}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\right)\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\right\}\\
&+ \mathrm{Tr}\left\{\mathbf{M}^{-1}\mathbf{W}^T\left(\sigma^2\bar{\mathbf{Q}} + \underbrace{\int_{\mathbb{R}_+} u\;\tilde{\pi}_{U|\Psi}(u)\,du}_{A(\mathbf{y}^o_{1:N},\Psi,\Psi^*)}\bar{\mathbf{C}}_{\mu}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\right)\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{W}^*\right\}\Bigg) = 0\\
\iff &\sigma^{*2} = \frac{1}{d}\Bigg(\sigma^2\,\mathrm{Tr}\left\{\mathbf{M}^{-1}\mathbf{W}^{*\,T}\mathbf{W}^*\right\} + \mathrm{Tr}\left\{\sigma^2\bar{\mathbf{Q}} + A(\mathbf{y}^o_{1:N},\Psi,\Psi^*)\bar{\mathbf{C}}_{\mu^*}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\right\}\\
&- 2\,\mathrm{Tr}\left\{\left(\sigma^2\bar{\mathbf{Q}} + A(\mathbf{y}^o_{1:N},\Psi,\Psi^*)\bar{\mathbf{C}}_{\mu,\mu^*}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\right)\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\right\}\\
&+ \mathrm{Tr}\left\{\mathbf{M}^{-1}\mathbf{W}^T\left(\sigma^2\bar{\mathbf{Q}} + A(\mathbf{y}^o_{1:N},\Psi,\Psi^*)\bar{\mathbf{C}}_{\mu}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\right)\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{W}^*\right\}.
\end{aligned}
$$

Given Lemma A7, $\tilde{\pi}_{U|\Psi}(u)$ is the density function of univariate Gamma random variable, $U \sim \Gamma\left(\alpha, \beta(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\right)$ and therefore we used the fact $\int_{\mathbb{R}_+}\tilde{\pi}_{U|\Psi}(u)\,du = 1$. Its expectation can be easily derived, that is

$$
A(\mathbf{y}^o_{1:N},\Psi,\Psi^*) = \int_{\mathbb{R}_+} u\;\tilde{\pi}_{U|\Psi}(u)\,du = \frac{\alpha}{\beta(\mathbf{y}^o_{1:N};\Psi,\Psi^*)}
$$

Therefore, the solution to the set of equations given by $\nabla_{\Psi^*}\tilde{Q} = \mathbf{0}$ are the following

$$
\begin{aligned}
\boldsymbol{\mu}^* &= \bar{\boldsymbol{\mu}}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\left(\mathbf{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\right) + \boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\\
\mathbf{W}^* &= \left(\sigma^2\bar{\mathbf{Q}} + \frac{\alpha}{\beta(\mathbf{y}^o_{1:N};\Psi,\Psi^*)}\bar{\mathbf{C}}_{\mu,\mu^*}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\right)\mathbf{W}\mathbf{M}^{-1}\left(\mathbf{M}^{-1}\mathbf{W}^T\left(\sigma^2\bar{\mathbf{Q}} + \frac{\alpha}{\beta(\mathbf{y}^o_{1:N};\Psi,\Psi^*)}\bar{\mathbf{C}}_{\mu}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\right)\mathbf{W}\mathbf{M}^{-1} + \sigma^2\mathbf{M}^{-1}\right)^{-1}\\
\sigma^{*2} &= \frac{1}{d}\Bigg(\sigma^2\,\mathrm{Tr}\left\{\mathbf{M}^{-1}\mathbf{W}^{*\,T}\mathbf{W}^*\right\} + \mathrm{Tr}\left\{\sigma^2\bar{\mathbf{Q}} + \frac{\alpha}{\beta(\mathbf{y}^o_{1:N};\Psi,\Psi^*)}\bar{\mathbf{C}}^m_{\mu^*}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\right\}\\
&\quad - 2\,\mathrm{Tr}\left\{\left(\sigma^2\bar{\mathbf{Q}} + \frac{\alpha}{\beta(\mathbf{y}^o_{1:N};\Psi,\Psi^*)}\bar{\mathbf{C}}_{\mu,\mu^*}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\right)\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\right\}\\
&\quad + \mathrm{Tr}\left\{\mathbf{M}^{-1}\mathbf{W}^T\left(\sigma^2\bar{\mathbf{Q}} + \frac{\alpha}{\beta(\mathbf{y}^o_{1:N};\Psi,\Psi^*)}\bar{\mathbf{C}}^m_{\mu}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\right)\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{W}^*\right\}
\end{aligned}
$$

□

A.3.9. Supporting Lemmas

Lemma A2. *Let a d dimensional observation vector $\mathbf{Y}_t$ be modelled as in Equation (2.7) with the assumptions given in Equation (3.21). The function $h : \mathbb{R}^d \times \mathbb{R}_+ \to \mathbb{R}$ of single realisations of $\mathbf{Y}_t$ and a Gamma random variable U under the assumption of the Identical and Conditionally Independent t-Student case from Subsection 3.2.2.2 is defined as*

$$
\begin{aligned}
h(\mathbf{y}_t, u;\Psi) &= \pi_{\mathbf{Y}_t|\mathbf{X}_t,U\Psi}(\mathbf{y}_t)\cdot\pi_{\mathbf{X}_t|U,\Psi}(\mathbf{x}_t)\cdot\varphi\left(\mathbf{x};\tilde{\boldsymbol{\mu}}_t,\tilde{\boldsymbol{\Sigma}}\right)^{-1}\\
&= (2\pi)^{-\frac{d}{2}}(\sigma^2)^{-\frac{d-k}{2}}u^{\frac{d}{2}}\left|\mathbf{M}\right|^{-\frac{1}{2}}\exp\left\{-\frac{u}{2\sigma^2}(\mathbf{y}_t-\boldsymbol{\mu})\left(\mathbb{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^T\right)(\mathbf{y}_t-\boldsymbol{\mu})^T\right\}
\end{aligned}
$$

for $\tilde{\boldsymbol{\mu}}_{t,\,1\times k} = (\mathbf{y}_t-\boldsymbol{\mu})\mathbf{W}\mathbf{M}^{-1}$, $\tilde{\boldsymbol{\Sigma}}_{k\times k} = \frac{\sigma^2}{u}\mathbf{M}^{-1}$, $\mathbf{M}_{k\times k} = \mathbf{W}^T\mathbf{W} + \sigma^2\mathbb{I}_k$ and $\varphi\left(\mathbf{x};\tilde{\boldsymbol{\mu}}_t,\tilde{\boldsymbol{\Sigma}}\right)$ being a density function of k dimensional Gaussian random vector $\mathbf{X}$ with a mean vector $\tilde{\boldsymbol{\mu}}_t$ and a covariance matrix $\tilde{\boldsymbol{\Sigma}}$.

Proof. Given the assumptions of the identically and conditionally independently distributed $\mathbf{X}_t$ and $\boldsymbol{\epsilon}$ and their stochastic representations from Subsection 3.2.2.2, $\pi_{\mathbf{Y}_t|\mathbf{X}_t,U\Psi} \cdot \pi_{\mathbf{X}_t|U,\Psi}$ is a convolution of two Gaussian densities, that is

$$\begin{aligned}
&\pi_{\mathbf{Y}_t|\mathbf{X}_t,U\Psi}(\mathbf{y}_t) \cdot \pi_{\mathbf{X}_t|U,\Psi}(\mathbf{x}_t) = (2\pi)^{-\frac{k+d}{2}} (\sigma^2)^{-\frac{d}{2}} u^{\frac{d+k}{2}} \exp\left\{ -\frac{u}{2\sigma^2}\left(\left(\mathbf{y}_t - \boldsymbol{\mu} - \mathbf{x}_t\mathbf{W}^T\right)\left(\mathbf{y}_t - \boldsymbol{\mu} - \mathbf{x}_t\mathbf{W}^T\right)^T + \sigma^2\mathbf{x}_t\mathbf{x}_t^T\right)\right\} \\
&= (2\pi)^{-\frac{k+d}{2}} (\sigma^2)^{-\frac{d}{2}} u^{\frac{d+k}{2}} \exp\left\{ -\frac{u}{2\sigma^2}\left(\mathbf{x}_t\left(\mathbf{W}^T\mathbf{W} + \sigma^2\mathbb{I}_k\right)\mathbf{x}_t^T - 2\mathbf{x}_t\mathbf{W}^T\left(\mathbf{y}_t - \boldsymbol{\mu}\right)^T + \left(\mathbf{y}_t - \boldsymbol{\mu}\right)\left(\mathbf{y}_t - \boldsymbol{\mu}\right)^T\right)\right\} \\
&= (2\pi)^{-\frac{k+d}{2}} (\sigma^2)^{-\frac{d}{2}} u^{\frac{d+k}{2}} \exp\left\{ -\frac{1}{2}\left(\mathbf{x}_t - \tilde{\boldsymbol{\mu}}_t\right)\tilde{\boldsymbol{\Sigma}}^{-1}\left(\mathbf{x}_t - \tilde{\boldsymbol{\mu}}_t\right)^T\right\} \exp\left\{ -\frac{1}{2}\left(\frac{u}{\sigma^2}\left(\mathbf{y}_t - \boldsymbol{\mu}\right)\left(\mathbf{y}_t - \boldsymbol{\mu}\right)^T - \tilde{\boldsymbol{\mu}}_t\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\boldsymbol{\mu}}_t^T\right)\right\}
\end{aligned}$$

where $\tilde{\boldsymbol{\mu}}_{t,\,1\times k} = \left(\mathbf{y}_t - \boldsymbol{\mu}\right)\mathbf{W}\mathbf{M}^{-1}$, $\tilde{\boldsymbol{\Sigma}}_{k\times k} = \frac{\sigma^2}{u}\mathbf{M}^{-1}$ and $\mathbf{M}_{k\times k} = \mathbf{W}^T\mathbf{W} + \sigma^2\mathbb{I}_k$. Let us define $\varphi(\mathbf{x};\tilde{\boldsymbol{\mu}}_t,\tilde{\boldsymbol{\Sigma}})$ as a density function of a Gaussian random variable $\mathbf{X} \in \mathbb{R}^k$ with a vector of means $\tilde{\boldsymbol{\mu}}_t$ and a covariance matrix $\tilde{\boldsymbol{\Sigma}}$. We have that

$$\begin{aligned}
&\pi_{\mathbf{Y}_t|\mathbf{X}_t,U\Psi}(\mathbf{y}_t) \cdot \pi_{\mathbf{X}_t|U,\Psi}(\mathbf{x}_t) = (2\pi)^{-\frac{d}{2}} (\sigma^2)^{-\frac{d}{2}} u^{\frac{d+k}{2}} \left|\tilde{\boldsymbol{\Sigma}}\right|^{\frac{1}{2}} \varphi\left(\mathbf{x}_t;\tilde{\boldsymbol{\mu}}_t,\tilde{\boldsymbol{\Sigma}}\right) \times \exp\left\{ -\frac{1}{2}\left(\mathbf{y}_t - \boldsymbol{\mu}\right)\left(\frac{u}{\sigma^2}\mathbb{I}_d - \mathbf{W}\tilde{\boldsymbol{\Sigma}}\mathbf{W}^T\right)\left(\mathbf{y}_t - \boldsymbol{\mu}\right)^T\right\} \\
&= (2\pi)^{-\frac{d}{2}} (\sigma^2)^{-\frac{d-k}{2}} u^{\frac{d}{2}} \left|\mathbf{M}\right|^{-\frac{1}{2}} \varphi\left(\mathbf{x}_t;\tilde{\boldsymbol{\mu}}_t,\tilde{\boldsymbol{\Sigma}}\right) \exp\left\{ -\frac{1}{2}\left(\mathbf{y}_t - \boldsymbol{\mu}\right)\left(\frac{u}{\sigma^2}\mathbb{I}_d - \mathbf{W}\mathbf{M}^{-1}\tilde{\boldsymbol{\Sigma}}^{-1}\mathbf{M}^{-1}\mathbf{W}^T\right)\left(\mathbf{y}_t - \boldsymbol{\mu}\right)^T\right\} \\
&= (2\pi)^{-\frac{d}{2}} (\sigma^2)^{-\frac{d-k}{2}} u^{\frac{d}{2}} \left|\mathbf{M}\right|^{-\frac{1}{2}} \varphi\left(\mathbf{x}_t;\tilde{\boldsymbol{\mu}}_t,\tilde{\boldsymbol{\Sigma}}\right) \exp\left\{ -\frac{u}{2\sigma^2}\left(\mathbf{y}_t - \boldsymbol{\mu}\right)\left(\mathbb{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^T\right)\left(\mathbf{y}_t - \boldsymbol{\mu}\right)^T\right\} \\
&= h(\mathbf{y}_t, u; \Psi) \cdot \varphi\left(\mathbf{x}_t;\tilde{\boldsymbol{\mu}}_t,\tilde{\boldsymbol{\Sigma}}\right)
\end{aligned}$$

where function $h : \mathbb{R}^d \times \mathbb{R}_+ \to \mathbb{R}$ is defined as

$$h(\mathbf{y}_t, u; \Psi) = (2\pi)^{-\frac{d}{2}} (\sigma^2)^{-\frac{d-k}{2}} u^{\frac{d}{2}} \left|\mathbf{M}\right|^{-\frac{1}{2}} \exp\left\{ -\frac{u}{2\sigma^2}\left(\mathbf{y}_t - \boldsymbol{\mu}\right)\left(\mathbb{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^T\right)\left(\mathbf{y}_t - \boldsymbol{\mu}\right)^T\right\}.$$

□

Lemma A3. *Let a d dimensional observation vector $\mathbf{Y}_t$ be modelled as in Equation (2.7) with the assumptions given in Equation (3.21). The function $w : \mathbb{R}^d \times \mathbb{R}_+ \to \mathbb{R}$ of single realisations of $\mathbf{Y}_t$ and a Gamma random variable U under the assumption of the Identical and Conditionally Independent t-Student case from Subsection 3.2.2.2 is defined as*

$$\begin{aligned}
&w(\mathbf{y}_t, u; \Psi, \Psi^*) = \int_{\mathbb{R}^k} \log\left(\pi_{\mathbf{Y}_t,\mathbf{X}_t|U,\Psi^*}(\mathbf{y}_t,\mathbf{x}_t)\right)\varphi\left(\mathbf{x}_t;\tilde{\boldsymbol{\mu}}_t,\tilde{\boldsymbol{\Sigma}}\right)d\mathbf{x}_t = -\frac{d+k}{2}\log(2\pi) - \frac{d}{2}\log\left(\sigma^{*2}\right) + \frac{d+k}{2}\log u \\
&- \frac{u}{2\sigma^{*2}}\,Tr\left\{\left(\mathbf{y}_t - \boldsymbol{\mu}^*\right)^T\left(\mathbf{y}_t - \boldsymbol{\mu}^*\right) - 2\left(\mathbf{y}_t - \boldsymbol{\mu}\right)^T\left(\mathbf{y}_t - \boldsymbol{\mu}^*\right)\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\right\} \\
&- \frac{1}{2\sigma^{*2}}\,Tr\left\{\left(u\mathbf{M}^{-1}\mathbf{W}^T\left(\mathbf{y}_t - \boldsymbol{\mu}\right)^T\left(\mathbf{y}_t - \boldsymbol{\mu}\right)\mathbf{W}\mathbf{M}^{-1} + \sigma^2\mathbf{M}^{-1}\right)\left(\mathbf{W}^{*T}\mathbf{W}^* + \sigma^{*2}\mathbb{I}_k\right)\right\}
\end{aligned}$$

for $\tilde{\boldsymbol{\mu}}_{t,\,1\times k} = \left(\mathbf{y}_t - \boldsymbol{\mu}\right)\mathbf{W}\mathbf{M}^{-1}$, $\tilde{\boldsymbol{\Sigma}}_{k\times k} = \frac{\sigma^2}{u}\mathbf{M}^{-1}$, $\mathbf{M}_{k\times k} = \mathbf{W}^T\mathbf{W} + \sigma^2\mathbb{I}_k$ and $\varphi(\mathbf{x};\tilde{\boldsymbol{\mu}}_t,\tilde{\boldsymbol{\Sigma}})$ being a distribution function of k dimensional Gaussian random vector $\mathbf{X}$ with mean $\tilde{\boldsymbol{\mu}}_t$ and covariance matrix $\tilde{\boldsymbol{\Sigma}}$.

Proof. Given the definition of $\mathbf{X}_t$ and $\boldsymbol{\epsilon}_t$ from Equation (3.21) and the complete data likelihood from Equation (3.22) of the Identical and Conditionally Independent t-Student case, let us define function $w : \mathbb{R}^d \times \mathbb{R}_+ \leftarrow \mathbb{R}$ as a solution to the following integration problem

$$w(\mathbf{y}_t, u; \Psi, \Psi^*) = \int_{\mathbb{R}^k} \log\left(\pi_{\mathbf{Y}_t,\mathbf{X}_t,U|\Psi^*}(\mathbf{y}_t,\mathbf{x}_t,u)\right)\varphi\left(\mathbf{x}_t;\tilde{\boldsymbol{\mu}}_t,\tilde{\boldsymbol{\Sigma}}\right)d\mathbf{x}_t$$

where $\tilde{\boldsymbol{\mu}}_{t,\,1\times k} = \left(\mathbf{y}_t - \boldsymbol{\mu}\right)\mathbf{W}\mathbf{M}^{-1}$, $\tilde{\boldsymbol{\Sigma}}_{k\times k} = \frac{\sigma^2}{u}\mathbf{M}^{-1}$ and $\mathbf{M}_{k\times k} = \mathbf{W}^T\mathbf{W} + \sigma^2\mathbb{I}_k$ are specified as in Lemma A2. The function $\varphi(\mathbf{x};\tilde{\boldsymbol{\mu}}_t,\tilde{\boldsymbol{\Sigma}})$ is a density function of the k-dimensional Gaussian random variable $\mathbf{X} \in \mathbb{R}^k$ with a vector of means $\tilde{\boldsymbol{\mu}}_t$ and a

covariance matrix $\tilde{\mathbf{\Sigma}}$. Therefore, using the Chain Rule of Probability and the fact that $\mathbf{Y}_t$ and $\mathbf{X}_t$ are conditionally Gaussian, we calculate the function w as follows

$$
\begin{aligned}
w(\mathbf{y}_t, u; \Psi, \Psi^*) &= \int_{\mathbb{R}^k} \log\Big(\pi_{\mathbf{Y}_t|\mathbf{X}_t,U,\Psi^*}(\mathbf{y}_t)\cdot\pi_{\mathbf{X}_t|U,\Psi^*}(\mathbf{x}_t)\Big)\varphi(\mathbf{x}_t;\tilde{\boldsymbol{\mu}}_t,\tilde{\mathbf{\Sigma}})d\mathbf{x}_t \\
&= \int_{\mathbb{R}^k} \log\Big((2\pi)^{-\frac{d+k}{2}}\left(\sigma^{*2}\right)^{-\frac{d}{2}}u^{\frac{d+k}{2}}\Big)\varphi(\mathbf{x}_t;\tilde{\boldsymbol{\mu}}_t,\tilde{\mathbf{\Sigma}})d\mathbf{x}_t \\
&\quad + \int_{\mathbb{R}^k} \log\Bigg(\exp\Bigg\{-\frac{u}{2\sigma^{*2}}\left(\mathbf{y}_t-\boldsymbol{\mu}^*-\mathbf{x}_t\mathbf{W}^{*T}\right)\left(\mathbf{y}_t-\boldsymbol{\mu}^*-\mathbf{x}_t\mathbf{W}^{*T}\right)^T-\frac{u}{2}\mathbf{x}_t\mathbf{x}_t^T\Bigg\}\Bigg)\varphi(\mathbf{x}_t;\tilde{\boldsymbol{\mu}}_t,\tilde{\mathbf{\Sigma}})d\mathbf{x}_t \\
&= -\frac{d+k}{2}\log(2\pi)-\frac{d}{2}\log\left(\sigma^{*2}\right)+\frac{d+k}{2}\log u \\
&\quad -\frac{u}{2}\int_{\mathbb{R}^k}\Bigg(\frac{1}{\sigma^{*2}}\left(\mathbf{y}_t-\boldsymbol{\mu}^*-\mathbf{x}_t\mathbf{W}^{*T}\right)\left(\mathbf{y}_t-\boldsymbol{\mu}^*-\mathbf{x}_t\mathbf{W}^{*T}\right)^T+\mathbf{x}_t\mathbf{x}_t^T\Bigg)\varphi(\mathbf{x}_t;\tilde{\boldsymbol{\mu}}_t,\tilde{\mathbf{\Sigma}})d\mathbf{x}_t \\
&= -\frac{d+k}{2}\log(2\pi)-\frac{d}{2}\log\left(\sigma^{*2}\right)+\frac{d+k}{2}\log u+\log\pi_{U|\Psi^*}(u) \\
&\quad -\frac{u}{2}\int_{\mathbb{R}^k}\Bigg(\frac{1}{\sigma^{*2}}\left(\mathbf{y}_t-\boldsymbol{\mu}^*\right)\left(\mathbf{y}_t-\boldsymbol{\mu}^*\right)^T-\frac{2}{\sigma^{*2}}\left(\mathbf{y}_t-\boldsymbol{\mu}^*\right)\mathbf{W}^*\mathbf{x}_t^T+\mathbf{x}_t\Big(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{W}^*+\mathbb{I}_k\Big)\mathbf{x}_t^T\Bigg)\varphi(\mathbf{x}_t;\tilde{\boldsymbol{\mu}}_t,\tilde{\mathbf{\Sigma}})d\mathbf{x}_t \\
&= -\frac{d+k}{2}\log(2\pi)-\frac{d}{2}\log\left(\sigma^{*2}\right)+\frac{d+k}{2}\log u-\frac{u}{2\sigma^{*2}}\operatorname{Tr}\left\{\left(\mathbf{y}_t-\boldsymbol{\mu}^*\right)^T\left(\mathbf{y}_t-\boldsymbol{\mu}^*\right)\right\} \\
&\quad -\frac{u}{2}\Bigg(\frac{1}{\sigma^{*2}}\int_{\mathbb{R}^k}\operatorname{Tr}\Big\{\mathbf{x}_t^T\mathbf{x}_t\Big(\mathbf{W}^{*T}\mathbf{W}^*+\sigma^{*2}\mathbb{I}_k\Big)\Big\}\varphi(\mathbf{x}_t;\tilde{\boldsymbol{\mu}}_t,\tilde{\mathbf{\Sigma}})d\mathbf{x}_t-\frac{2}{\sigma^{*2}}\left(\mathbf{y}_t-\boldsymbol{\mu}^*\right)\mathbf{W}^*\int_{\mathbb{R}^k}\mathbf{x}_t^T\varphi(\mathbf{x}_t;\tilde{\boldsymbol{\mu}}_t,\tilde{\mathbf{\Sigma}})d\mathbf{x}_t\Bigg) \\
&= -\frac{d+k}{2}\log(2\pi)-\frac{d}{2}\log\left(\sigma^{*2}\right)+\frac{d+k}{2}\log u+\log\pi_{U|\Psi^*}(u) \\
&\quad -\frac{u}{2}\Bigg(\frac{1}{\sigma^{*2}}\operatorname{Tr}\left\{\left(\mathbf{y}_t-\boldsymbol{\mu}^*\right)^T\left(\mathbf{y}_t-\boldsymbol{\mu}^*\right)\right\}-\frac{2}{\sigma^{*2}}\left(\mathbf{y}_t-\boldsymbol{\mu}^*\right)\mathbf{W}^*\tilde{\boldsymbol{\mu}}_t^T+\frac{1}{\sigma^{*2}}\operatorname{Tr}\Big\{\Big(\tilde{\mathbf{\Sigma}}+\tilde{\boldsymbol{\mu}}_t^T\tilde{\boldsymbol{\mu}}_t\Big)\mathbf{M}^*\Big\}\Bigg) \\
&= -\frac{d+k}{2}\log(2\pi)-\frac{d}{2}\log\left(\sigma^{*2}\right)+\frac{d+k}{2}\log u-\frac{u}{2\sigma^{*2}}\operatorname{Tr}\left\{\left(\mathbf{y}_t-\boldsymbol{\mu}^*\right)^T\left(\mathbf{y}_t-\boldsymbol{\mu}^*\right)\right\} \\
&\quad -\frac{u}{2}\Bigg(\frac{1}{\sigma^{*2}}\operatorname{Tr}\Big\{\frac{\sigma^2}{u}\mathbf{M}^{-1}\mathbf{M}^*+\mathbf{M}^{-1}\mathbf{W}^T\left(\mathbf{y}_t-\boldsymbol{\mu}\right)^T\left(\mathbf{y}_t-\boldsymbol{\mu}\right)\mathbf{W}\mathbf{M}^{-1}\mathbf{M}^*\Big\}-\frac{2}{\sigma^{*2}}\left(\mathbf{y}_t-\boldsymbol{\mu}^*\right)\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\left(\mathbf{y}_t-\boldsymbol{\mu}\right)^T\Bigg) \\
&= -\frac{d+k}{2}\log(2\pi)-\frac{d}{2}\log\left(\sigma^{*2}\right)+\frac{d+k}{2}\log u-\frac{u}{2\sigma^{*2}}\operatorname{Tr}\left\{\left(\mathbf{y}_t-\boldsymbol{\mu}^*\right)^T\left(\mathbf{y}_t-\boldsymbol{\mu}^*\right)\right\} \\
&\quad -\frac{u}{2\sigma^{*2}}\operatorname{Tr}\Big\{\mathbf{M}^{-1}\mathbf{W}^T\left(\mathbf{y}_t-\boldsymbol{\mu}\right)^T\left(\mathbf{y}_t-\boldsymbol{\mu}\right)\mathbf{W}\mathbf{M}^{-1}\mathbf{M}^*--2\left(\mathbf{y}_t-\boldsymbol{\mu}\right)^T\left(\mathbf{y}_t-\boldsymbol{\mu}^*\right)\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\Big\}-\frac{\sigma^2}{2\sigma^{*2}}\operatorname{Tr}\left\{\mathbf{M}^{-1}\mathbf{M}^*\right\}
\end{aligned}
$$

where $\mathbf{M}^* = \mathbf{W}^{*T}\mathbf{W}^* + \sigma^{*2}\mathbb{I}_k$. □

Lemma A4. *Let the d dimensional observation vector $\mathbf{Y}_t$ be modelled as in Equation (2.7) with the assumptions given in Equation (3.21) for scalar random variable Gamma $U \sim \Gamma(\frac{v}{2}, \frac{v}{2})$. We define function $H : \mathbb{R}^{N\times d}\times\mathbb{R}_+ \to \mathbb{R}$, $H(\mathbf{y}_{1:N}, u; \Psi) := \prod_{t=1}^N h(\mathbf{y}_t, u; \Psi)$ being a product of functions $h : \mathbf{R}^d \times \mathbb{R}_+ \to \mathbb{R}$ defined in Lemma A2 over N realisation of vector $\mathbf{Y}_t$, $\mathbf{y}_{1:N}$, in time $t = 1, \ldots, N$. Let u be a single realisation of U. It is true that*

$$\pi_{U|\Psi}(u)\cdot H(\mathbf{y}_{1:N}, u; \Psi) = C_H(\mathbf{y}_{1:N}; \Psi)\cdot\tilde{\pi}_{U|\Psi}(u)$$

where $\tilde{\pi}_{U|\Psi}(u)$ is a density function of a Gamma random variable $U \sim \Gamma\Big(\alpha, \beta(\mathbf{y}_{1:N;\Psi})\Big)$ for $\alpha = \frac{v}{2} + \frac{dN}{2}$ and $\beta, C_H : \mathbb{R}^{N\times d} \to \mathbb{R}$ being defined as

$$\beta(\mathbf{y}_{1:N}; \Psi) := \frac{v}{2} + \frac{1}{2\sigma^2}\sum_{t=1}^N\left(\mathbf{y}_t-\boldsymbol{\mu}\right)\Big(\mathbb{I}_d-\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^T\Big)\left(\mathbf{y}_t-\boldsymbol{\mu}\right)^T$$

$$C_H(\mathbf{y}_{1:N}; \Psi) := (2\pi)^{-\frac{dN}{2}}(\sigma^2)^{-N\frac{d-k}{2}}\left|\mathbf{M}\right|^{-\frac{N}{2}}\frac{\left(\frac{v}{2}\right)^{\frac{v}{2}}\Gamma(\alpha)}{\Gamma\left(\frac{v}{2}\right)}\beta(\mathbf{y}_{1:N}; \Psi)^{-\alpha}.$$

Proof. Let us recall the notation and the definition of the function h from Lemma A2, that is

$$h(\mathbf{y}_t, u; \Psi) = (2\pi)^{-\frac{d}{2}}(\sigma^2)^{-\frac{d-k}{2}}u^{\frac{d}{2}}\left|\mathbf{M}\right|^{-\frac{1}{2}}\exp\Bigg\{-\frac{u}{2\sigma^2}\left(\mathbf{y}_t-\boldsymbol{\mu}\right)\Big(\mathbb{I}_d-\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^T\Big)\left(\mathbf{y}_t-\boldsymbol{\mu}\right)^T\Bigg\}$$

The function $H : \mathbb{R}^{N\times d} \times \mathbb{R}_+ \to \mathbb{R}$, $H(\mathbf{y}_{1:N}, u; \Psi) := \prod_{t=1}^{N} h(\mathbf{y}_t, u; \Psi)$, is calculated as follows

$$\begin{aligned} H(\mathbf{y}_{1:N}, u; \Psi) &= \prod_{t=1}^{N} \left((2\pi)^{-\frac{d}{2}} (\sigma^2)^{-\frac{d-k}{2}} u^{\frac{d}{2}} \left|\mathbf{M}\right|^{-\frac{1}{2}} \exp\left\{ -\frac{u}{2\sigma^2} (\mathbf{y}_t - \boldsymbol{\mu}) \left(\mathbb{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^T\right) (\mathbf{y}_t - \boldsymbol{\mu})^T \right\} \right) \\ &= (2\pi)^{-\frac{dN}{2}} (\sigma^2)^{-N\frac{d-k}{2}} u^{\frac{dN}{2}} \left|\mathbf{M}\right|^{-\frac{N}{2}} \exp\left\{ -\frac{u}{2\sigma^2} \sum_{t=1}^{N} (\mathbf{y}_t - \boldsymbol{\mu}) \left(\mathbb{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^T\right) (\mathbf{y}_t - \boldsymbol{\mu})^T \right\} \end{aligned}$$

We show in the next part that given $U \sim \Gamma\left(\frac{v}{2}, \frac{v}{2}\right)$, the expression $\pi_{U|\Psi}(u) \cdot H(\mathbf{y}_{1:N}, u; \Psi)$ is proportional to a gamma distribution, that is

$$\begin{aligned} \pi_{U|\Psi}(u) \cdot H(\mathbf{y}_{1:N}, u; \Psi) &= (2\pi)^{-\frac{dN}{2}} (\sigma^2)^{-N\frac{d-k}{2}} \left|\mathbf{M}\right|^{-\frac{N}{2}} \frac{\left(\frac{v}{2}\right)^{\frac{v}{2}}}{\Gamma\left(\frac{v}{2}\right)} \\ &\quad \times u^{\frac{v}{2}-1} e^{-\frac{v}{2}u} u^{\frac{dN}{2}} \exp\left\{ -\frac{u}{2\sigma^2} \sum_{t=1}^{N} (\mathbf{y}_t - \boldsymbol{\mu}) \left(\mathbb{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^T\right) (\mathbf{y}_t - \boldsymbol{\mu})^T \right\} \\ &\propto u^{\frac{v}{2}+\frac{dN}{2}-1} \exp\left\{ -u\left(\frac{v}{2} + \frac{1}{2\sigma^2} \sum_{t=1}^{N} (\mathbf{y}_t - \boldsymbol{\mu}) \left(\mathbb{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^T\right) (\mathbf{y}_t - \boldsymbol{\mu})^T \right) \right\} \\ &\propto \frac{\beta(\mathbf{y}_{1:N}; \Psi)^{\alpha}}{\Gamma(\alpha)} u^{\alpha-1} e^{-u\beta(\mathbf{y}_{1:N}; \Psi)} =: \tilde{\pi}_{U|\Psi}(u) \end{aligned}$$

for $\alpha = \frac{v}{2} + \frac{dN}{2}$ and $\beta : \mathbb{R}^{N\times d} \to \mathbb{R}_+$ being a function defined as

$$\beta(\mathbf{y}_{1:N}; \Psi) = \frac{v}{2} + \frac{1}{2\sigma^2} \sum_{t=1}^{N} (\mathbf{y}_t - \boldsymbol{\mu}) \left(\mathbb{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^T\right) (\mathbf{y}_t - \boldsymbol{\mu})^T.$$

Recall that the relation of proportionality means that we can find C_H which is not a function of u, such that

$$\pi_{U|\Psi}(u) \cdot H(\mathbf{y}_{1:N}, u; \Psi) = C_H \cdot \tilde{\pi}_{U|\Psi}(u).$$

In our case, C_H is a function of $\mathbf{y}_{1:N}$ as

$$C_H(\mathbf{y}_{1:N}; \Psi) = \frac{\pi_{U|\Psi}(u) H(\mathbf{y}_{1:N}, u; \Psi)}{\tilde{\pi}_{U|\Psi}(u)} = (2\pi)^{-\frac{dN}{2}} (\sigma^2)^{-N\frac{d-k}{2}} \left|\mathbf{M}\right|^{-\frac{N}{2}} \frac{\left(\frac{v}{2}\right)^{\frac{v}{2}} \Gamma(\alpha)}{\Gamma\left(\frac{v}{2}\right)} \beta(\mathbf{y}_{1:N}; \Psi)^{-\alpha}$$

□

Lemma A5. *Let a d dimensional observation vector $\mathbf{Y}_t$ be modelled as in Equation (2.7) with assumptions given in Equation (3.21) for a scalar random variable Gamma $U \sim \Gamma(\frac{v}{2}, \frac{v}{2})$. We define function $H : \mathbb{R}^{N\times d} \times \mathbb{R}_+ \to \mathbb{R}$, $H(\mathbf{y}_{1:N}, u; \Psi) := \prod_{s=1}^{N} h(\mathbf{y}_s, u; \Psi)$ being a product of functions $h : \mathbf{R}^d \times \mathbb{R}_+ \to \mathbb{R}$ defined in Lemma A2 over N realisation of vector $\mathbf{Y}_t$, $\mathbf{y}_{1:N}$, in time $t = 1, \ldots, N$. Let u be a single realisation of U. It is true that there exists a scalar $C_H(\Psi) \in \mathbb{R}$ equal to $C_H(\Psi) = (\sigma^2)^{\frac{kN}{2}} \left|\mathbf{M}\right|^{-\frac{N}{2}} \left|\left(\mathbb{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^T\right)\right|^{-\frac{N}{2}}$ such that*

$$H(\mathbf{y}_{1:N}, u; \Psi) = C_H(\Psi) \cdot \prod_{t=1}^{N} \tilde{\pi}_{\mathbf{Y}_t|U,\Psi}(\mathbf{y}_t)$$

where $\tilde{\pi}_{\mathbf{Y}|U,\Psi}(\mathbf{y}_t) := \phi\left(\mathbf{y}_t, \boldsymbol{\mu}, \tilde{\boldsymbol{\Sigma}}\right)$ for $\tilde{\boldsymbol{\Sigma}} = \frac{\sigma^2}{u}\left(\mathbb{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^T\right)^{-1}$ and $\phi\left(\mathbf{y}, \tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}\right)$ being probability density function of d-dimensional Gaussian random vector with mean $\tilde{\boldsymbol{\mu}}$ and a covariance matrix $\tilde{\boldsymbol{\Sigma}}$.

Proof. Given the definition of function h from Lemma A2

$$h(\mathbf{y}_t, u; \Psi) = (2\pi)^{-\frac{d}{2}} (\sigma^2)^{-\frac{d-k}{2}} u^{\frac{d}{2}} \left|\mathbf{M}\right|^{-\frac{1}{2}} \exp\left\{ -\frac{u}{2\sigma^2} (\mathbf{y}_t - \boldsymbol{\mu}) \left(\mathbb{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^T\right) (\mathbf{y}_t - \boldsymbol{\mu})^T \right\},$$

the function $H : \mathbb{R}^{N\times d} \times \mathbb{R}_+ \to \mathbb{R}$, which equals to $H(\mathbf{y}_{1:N}, u; \Psi) := \prod_{s=1}^{N} h(\mathbf{y}_s, u; \Psi)$, is calculated as follows

$$H(\mathbf{y}_{1:N}, u; \Psi) = \prod_{s=1}^{N} \left((2\pi)^{-\frac{d}{2}} (\sigma^2)^{-\frac{d-k}{2}} u^{\frac{d}{2}} \left|\mathbf{M}\right|^{-\frac{1}{2}} \exp\left\{ -\frac{u}{2\sigma^2} (\mathbf{y}_t - \boldsymbol{\mu}) \left(\mathbb{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^T\right) (\mathbf{y}_t - \boldsymbol{\mu})^T \right\} \right)$$

Let us denote $\tilde{\mathbf{\Sigma}} = \frac{\sigma^2}{u}\left(\mathbb{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^T\right)^{-1}$. Then

$$\begin{aligned}
H(\mathbf{y}_{1:N}, u; \Psi) &= \prod_{s=1}^{N}\left((2\pi)^{-\frac{d}{2}}(\sigma^2)^{-\frac{d-k}{2}}u^{\frac{d}{2}}\left|\mathbf{M}\right|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\mathbf{y}_t-\boldsymbol{\mu})\tilde{\mathbf{\Sigma}}^{-1}(\mathbf{y}_t-\boldsymbol{\mu})^T\right\}\right)\\
&= \prod_{s=1}^{N}\left((2\pi)^{-\frac{d}{2}}(\sigma^2)^{-\frac{d-k}{2}}u^{\frac{d}{2}}\left|\mathbf{M}\right|^{-\frac{1}{2}}|\tilde{\mathbf{\Sigma}}|^{\frac{1}{2}}|\tilde{\mathbf{\Sigma}}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\mathbf{y}_t-\boldsymbol{\mu})\tilde{\mathbf{\Sigma}}^{-1}(\mathbf{y}_t-\boldsymbol{\mu})^T\right\}\right)\\
&= \prod_{s=1}^{N}\Bigg((\sigma^2)^{-\frac{d-k}{2}}u^{\frac{d}{2}}\left|\mathbf{M}\right|^{-\frac{1}{2}}\left|\tilde{\mathbf{\Sigma}}\right|^{\frac{1}{2}}\underbrace{(2\pi)^{-\frac{d}{2}}\left|\tilde{\mathbf{\Sigma}}\right|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\mathbf{y}_t-\boldsymbol{\mu})\tilde{\mathbf{\Sigma}}^{-1}(\mathbf{y}_t-\boldsymbol{\mu})^T\right\}}_{\phi(\mathbf{y}_t;\boldsymbol{\mu},\tilde{\mathbf{\Sigma}})}\Bigg)\\
&= \prod_{s=1}^{N}\left((\sigma^2)^{-\frac{d-k}{2}}u^{\frac{d}{2}}\left|\mathbf{M}\right|^{-\frac{1}{2}}\left|\frac{\sigma^2}{u}\left(\mathbb{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^T\right)^{-1}\right|^{\frac{1}{2}}\phi(\mathbf{y}_t;\boldsymbol{\mu},\tilde{\mathbf{\Sigma}})\right)\\
&= \prod_{t=1}^{N}\left((\sigma^2)^{\frac{k}{2}}\left|\mathbf{M}\right|^{-\frac{1}{2}}\left|\left(\mathbb{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^T\right)\right|^{-\frac{1}{2}}\phi(\mathbf{y}_t;\boldsymbol{\mu},\tilde{\mathbf{\Sigma}})\right)
\end{aligned}$$

By denoting $C_H(\Psi) = (\sigma^2)^{\frac{kN}{2}}\left|\mathbf{M}\right|^{-\frac{N}{2}}\left|\left(\mathbb{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^T\right)\right|^{-\frac{N}{2}}$ and $\tilde{\pi}_{\mathbf{Y}_t|U,\Psi}(\mathbf{y}_t) := \phi(\mathbf{y}_t;\boldsymbol{\mu},\tilde{\mathbf{\Sigma}})$ we have that the function H can be decomposed into a scalar and product of Gaussian, multivariate densities with the same mean and covariance

$$H(\mathbf{y}_{1:N}, u; \Psi) = C_H(\Psi)\cdot\prod_{t=1}^{N}\tilde{\pi}_{\mathbf{Y}_t|U,\Psi}(\mathbf{y}_t)$$

□

Lemma A6. *Let a d dimensional observation vector $\mathbf{Y}_t$ has a probability density function $\tilde{\pi}_{\mathbf{Y}_t|U,\Psi}(\mathbf{y}_t)$ specified in Lemma A5 and let U be specified as in Subsection 3.2.2.2. Given the partition of the vector , $\mathbf{Y}_t = [\mathbf{Y}_t^o, \mathbf{Y}_t^m]$, the conditional distribution of $\mathbf{Y}_t|\mathbf{Y}_t^m, U, \Psi$ is the following*

$$\mathbf{Y}_t|\mathbf{Y}_t^o, U, \Psi \sim \mathcal{N}\left(\begin{bmatrix}\mathbf{Y}_t^o\\ \boldsymbol{\mu}_m + (\mathbf{Y}_t^o - \boldsymbol{\mu}_o)\,\mathbf{V}_{oo}\mathbf{V}_{om}^{-1}\end{bmatrix}, \frac{\sigma^2}{U}\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{V}_{mm}^{-1} - \mathbf{V}_{mo}^{-1}\mathbf{V}_{oo}\mathbf{V}_{om}^{-1}\end{bmatrix}\right)$$

for $\mathbf{V} = \mathbb{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^T$, where $\mathbf{V}_{oo}$ and $\mathbf{V}_{mm}$ or $\mathbf{V}_{oo}^{-1}$ and $\mathbf{V}_{mm}^{-1}$ are $d_o \times d_o$ and $d_m \times d_m$ square, symmetric submatrices of $\mathbf{V}$ or $\mathbf{V}^{-1}$ with rows (the left side lower index) and columns (the right side lower index) corresponding to observed entries and missing entries of the vector $\mathbf{Y}_t$, respectively. The matrices $\mathbf{V}_{om} = \mathbf{V}_{mo}^T$ or $\mathbf{V}_{om}^{-1} = \mathbf{V}_{mo}^{-1\,T}$ is a submatrix of the matrix $\mathbf{V}$ or $\mathbf{V}^{-1}$, respectively, which rows correspond to observed entries and columns to missing entries of the vector $\mathbf{Y}_t$.

Then we have the following solution to the integration problem

$$\begin{aligned}
\tilde{w}(\mathbf{y}_t^o, u; \Psi, \Psi^*) &= \int_{\mathbb{R}^{d_m}} w(\mathbf{y}_t, u; \Psi, \Psi^*)\tilde{\pi}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U,\Psi}(\mathbf{y}_t^m)\,d\mathbf{y}_t^m\\
&= -\frac{d+k}{2}\log(2\pi) - \frac{d}{2}\log\left(\sigma^{*2}\right) + \frac{d+k}{2}\log u - \frac{\sigma^2}{2\sigma^{*2}}\,Tr\left\{\mathbf{M}^{-1}\mathbf{M}^*\right\}\\
&\quad - \frac{u}{2\sigma^{*2}}\Bigg(Tr\left\{\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U,\Psi}\left[\mathbf{Y}_t\mathbf{Y}_t^T\right] - 2\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U,\Psi}\left[\mathbf{Y}_t\right]\boldsymbol{\mu}^{*\,T} + \boldsymbol{\mu}^*\boldsymbol{\mu}^{*\,T}\right\}\\
&\quad - 2\,Tr\left\{\left(\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U,\Psi}\left[\mathbf{Y}_t\mathbf{Y}_t^T\right] - \mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U,\Psi}\left[\mathbf{Y}_t\right]\left(\boldsymbol{\mu}+\boldsymbol{\mu}^*\right)^T + \boldsymbol{\mu}^{*\,T}\boldsymbol{\mu}\right)\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\right\}\\
&\quad + Tr\left\{\mathbf{M}^{-1}\mathbf{W}^T\left(\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U,\Psi}\left[\mathbf{Y}_t\mathbf{Y}_t^T\right] - 2\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U,\Psi}\left[\mathbf{Y}_t\right]\boldsymbol{\mu}^T + \boldsymbol{\mu}\boldsymbol{\mu}^T\right)\mathbf{W}\mathbf{M}^{-1}\mathbf{M}^*\right\}
\end{aligned}$$

for the corresponding moments of the conditional distribution

$$\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U,\Psi}\left[\mathbf{Y}_t\right] = \begin{bmatrix} \mathbf{Y}_t^o \\ \boldsymbol{\mu}_m + \left(\mathbf{Y}_t^o - \boldsymbol{\mu}_o\right)\mathbf{V}_{oo}\mathbf{V}_{om}^{-1} \end{bmatrix},$$

$$\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U,\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right] = \frac{\sigma^2}{U}\begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_{mm}^{-1} - \mathbf{V}_{mo}^{-1}\mathbf{V}_{oo}\mathbf{V}_{om}^{-1} \end{bmatrix} + \mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U,\Psi}\left[\mathbf{Y}_t\right]^T\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U,\Psi}\left[\mathbf{Y}_t\right].$$

Proof. Let us denote $\tilde{w}(\mathbf{y}_t^o, u; \Psi, \Psi^*) := \int_{\mathbb{R}^{d_m}} w(\mathbf{y}_t, u; \Psi, \Psi^*)\tilde{\pi}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U,\Psi}(\mathbf{y}_t^m)\, d\mathbf{y}_t^m$. Using the definition of the function w from Lemma A3, we have that

$$\begin{aligned}
\tilde{w}(\mathbf{y}_t^o, u; \Psi, \Psi^*) &= \int_{\mathbb{R}^{d_m}} w(\mathbf{y}_t, u; \Psi, \Psi^*)\tilde{\pi}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U,\Psi}(\mathbf{y}_t^m)\, d\mathbf{y}_t^m \\
&= \int_{\mathbb{R}^{d_m}} \left(-\frac{d+k}{2}\log(2\pi) - \frac{d}{2}\log\left(\sigma^{*2}\right) + \frac{d+k}{2}\log u \right)\tilde{\pi}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U,\Psi}(\mathbf{y}_t^m)\, d\mathbf{y}_t^m \\
&\quad - \frac{u}{2\sigma^{*2}}\int_{\mathbb{R}^{d_m}} \operatorname{Tr}\left\{ \left(\mathbf{y}_t - \boldsymbol{\mu}^*\right)^T\left(\mathbf{y}_t - \boldsymbol{\mu}^*\right) - 2\left(\mathbf{y}_t - \boldsymbol{\mu}\right)^T\left(\mathbf{y}_t - \boldsymbol{\mu}^*\right)\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T \right\}\tilde{\pi}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U,\Psi}(\mathbf{y}_t^m)\, d\mathbf{y}_t^m \\
&\quad - \frac{u}{2\sigma^{*2}}\int_{\mathbb{R}^{d_m}} \operatorname{Tr}\left\{ \mathbf{M}^{-1}\mathbf{W}^T\left(\mathbf{y}_t - \boldsymbol{\mu}\right)^T\left(\mathbf{y}_t - \boldsymbol{\mu}\right)\mathbf{W}\mathbf{M}^{-1}\mathbf{M}^* \right\}\tilde{\pi}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U,\Psi}(\mathbf{y}_t^m)\, d\mathbf{y}_t^m \\
&\quad - \frac{\sigma^2}{2\sigma^{*2}}\int_{\mathbb{R}^{d_m}} \operatorname{Tr}\left\{ \mathbf{M}^{-1}\mathbf{M}^* \right\}\tilde{\pi}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U,\Psi}(\mathbf{y}_t^m)\, d\mathbf{y}_t^m \\
&= -\frac{d+k}{2}\log(2\pi) - \frac{d}{2}\log\left(\sigma^{*2}\right) + \frac{d+k}{2}\log u - \frac{\sigma^2}{2\sigma^{*2}}\operatorname{Tr}\left\{ \mathbf{M}^{-1}\mathbf{M}^* \right\} \\
&\quad - \frac{u}{2\sigma^{*2}}\left(\operatorname{Tr}\left\{ \mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U,\Psi}\left[\mathbf{Y}_t\mathbf{Y}_t^T\right] - 2\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U,\Psi}\left[\mathbf{Y}_t\right]\boldsymbol{\mu}^{*\,T} + \boldsymbol{\mu}^*\boldsymbol{\mu}^{*\,T} \right\} \right. \\
&\quad \left. - 2\operatorname{Tr}\left\{ \left(\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U,\Psi}\left[\mathbf{Y}_t\mathbf{Y}_t^T\right] - \mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U,\Psi}\left[\mathbf{Y}_t\right]\left(\boldsymbol{\mu} + \boldsymbol{\mu}^*\right)^T + \boldsymbol{\mu}^{*\,T}\boldsymbol{\mu}\right)\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T \right\} \right. \\
&\quad \left. + \operatorname{Tr}\left\{ \mathbf{M}^{-1}\mathbf{W}^T\left(\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U,\Psi}\left[\mathbf{Y}_t\mathbf{Y}_t^T\right] - 2\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U,\Psi}\left[\mathbf{Y}_t\right]\boldsymbol{\mu}^T + \boldsymbol{\mu}\boldsymbol{\mu}^T\right)\mathbf{W}\mathbf{M}^{-1}\mathbf{M}^* \right\} \right)
\end{aligned}$$

Let us recall the fact $\tilde{\pi}_{\mathbf{Y}_t|U,\Psi}(\mathbf{y}_t)$ is a probability density function of multivariate Gaussian random variable, that is, $\mathbf{Y}_t = [\mathbf{Y}_t^o, \mathbf{Y}_t^m]$ conditioned on U is a multivariate Gaussian random vector with the mean $\boldsymbol{\mu}$ and the covariance matrix $\tilde{\boldsymbol{\Sigma}}$.

We denote $\mathbf{V} := \mathbb{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^T$ such that $\tilde{\boldsymbol{\Sigma}} = \frac{\sigma^2}{u}\mathbf{V}^{-1}$. The distribution of a subvector of Gaussian random variable conditioned of the other subvector of the same random variable in given in Lemma A1. Hence, the distribution of $\mathbf{Y}_t^m|\mathbf{Y}_t^o, U, \Psi$ is multivariate Gaussian, that is

$$\mathbf{Y}_t|\mathbf{Y}_t^o, U, \Psi \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{Y}_t^o \\ \boldsymbol{\mu}_m + \left(\mathbf{Y}_t^o - \boldsymbol{\mu}_o\right)\mathbf{V}_{oo}\mathbf{V}_{om}^{-1} \end{bmatrix}, \frac{\sigma^2}{U}\begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_{mm}^{-1} - \mathbf{V}_{mo}^{-1}\mathbf{V}_{oo}\mathbf{V}_{om}^{-1} \end{bmatrix} \right)$$

where $\mathbf{V}_{oo}$ and $\mathbf{V}_{mm}$ or $\mathbf{V}_{oo}^{-1}$ and $\mathbf{V}_{mm}^{-1}$ are $d_o \times d_o$ and $d_m \times d_m$ square, symmetric submatrices of $\mathbf{V}$ or $\mathbf{V}^{-1}$ with rows (the left side lower index) and columns (the right side lower index) corresponding to observed entries and missing entries of the vector $\mathbf{Y}_t$, respectively. The matrices $\mathbf{V}_{om} = \mathbf{V}_{mo}^T$ or $\mathbf{V}_{om}^{-1} = \mathbf{V}_{mo}^{-1\,T}$ is a submatrix of the matrix $\mathbf{V}$ or $\mathbf{V}^{-1}$, respectively, which rows correspond to observed entries and columns to missing entries of the vector $\mathbf{Y}_t$. Let us denote

$$\mathbf{Q} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_{mm}^{-1} - \mathbf{V}_{mo}^{-1}\mathbf{V}_{oo}\mathbf{V}_{om}^{-1} \end{bmatrix}.$$

The required moments can be easily derived given the conditional distribution of $\mathbf{Y}_t|\mathbf{Y}_t^o, U, \Psi$, that is

$$\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U,\Psi}\left[\mathbf{Y}_t\right] = \begin{bmatrix} \mathbf{Y}_t^o \\ \boldsymbol{\mu}_m + \left(\mathbf{Y}_t^o - \boldsymbol{\mu}_o\right)\mathbf{V}_{oo}^{-1}\mathbf{V}_{om} \end{bmatrix},$$

$$\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U,\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right] = \frac{\sigma^2}{U}\mathbf{Q} + \mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U,\Psi}\left[\mathbf{Y}_t\right]^T\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U,\Psi}\left[\mathbf{Y}_t\right].$$

□

Lemma A7. *Let a d dimensional observation vector* $\mathbf{Y}_t$ *has a probability density function* $\tilde{\pi}_{\mathbf{Y}_t|U,\Psi}(\mathbf{y}_t)$ *specified in Lemma A5 and let* U *be specified as in Subsection 3.2.2.2 . Given the partition of the vector ,* $\mathbf{Y}_t = [\mathbf{Y}_t^o, \mathbf{Y}_t^m]$*, the marginal distribution of* $\mathbf{Y}_t^o|U,\Psi$ *is the following*

$$\mathbf{Y}_t^o|U,\Psi \sim \mathcal{N}\Big(\boldsymbol{\mu}_o, \frac{\sigma^2}{U}\mathbf{V}_{oo}^{-1}\Big)$$

where $\boldsymbol{\mu}_o$ *is a subvector of* $\boldsymbol{\mu}$ *which elements correspond to the vector with the observed entries* $\mathbf{Y}_o$ *and* $\mathbf{V}_{oo}^{-1}$ *is a square, symmetric submatrix of* $\mathbf{V}^{-1} = \Big(\mathbb{I}_d - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^T\Big)^{-1}$ *which correspond to the vector with the observed entries by rows and columns.*

Let $\mathbf{y}_{1:N}^o$ *be N realisations of vector* $\mathbf{Y}_t^o$*. We can specify the function* $C_\beta : \mathbb{R}^{d_o \times N} \to \mathbb{R}$

$$C_\beta(\mathbf{y}_{1:N}^o, \Psi) = \frac{(2\pi)^{-\frac{d_o N}{2}}(\sigma^2)^{-\frac{d_o N}{2}}\Big|\mathbf{V}_{oo}\Big|^{\frac{N}{2}}\left(\frac{v}{2}\right)^{\frac{v}{2}}\Gamma(\alpha)}{\Gamma\left(\frac{v}{2}\right)}\beta(\mathbf{y}_{1:N}^o;\Psi)^{-\alpha},$$

which is constant with respect to U*, such that*

$$\pi_{U|\Psi}(u)\cdot\left(\prod_{t=1}^N \tilde{\pi}_{\mathbf{Y}_t^o|U,\Psi}(\mathbf{y}_t^o)\right) = C_\beta(\mathbf{y}_{1:N}^o,\Psi)\cdot\tilde{\pi}_{U|\Psi}(u)$$

where $\tilde{\pi}_{U|\Psi}(u)$ *is a density of univariate Gamma random variable* $U \sim \Gamma(\alpha, \beta(\mathbf{y}_{1:N}^o;\Psi))$ *for the scalar* $\alpha = \frac{v+d_o N}{2}$ *and a function* $\beta : \mathbb{R}^{d_o\times N} \to \mathbb{R}$

$$\beta(\mathbf{y}_{1:N}^o;\Psi) = \frac{v}{2} + \frac{1}{2\sigma^2}\sum_{t=1}^N (\mathbf{y}_t^o - \boldsymbol{\mu}_o)\mathbf{V}_{oo}(\mathbf{y}_t^o - \boldsymbol{\mu}_o)^T.$$

Proof. We show that $\pi_{U|\Psi}(u)\Big(\prod_{t=1}^N \tilde{\pi}_{\mathbf{Y}_t^o|U,\Psi}(\mathbf{y}_t^o)\Big) \propto \tilde{\pi}_{U|\Psi}(u)$, where $\tilde{\pi}_{U|\Psi}(u)$ is a density of univariate Gamma random variable. The density $\tilde{\pi}_{\mathbf{Y}_t^o|U,\Psi}(\mathbf{y}_t^o)$ is a probability function of $\mathbf{Y}_t^o$ which is multivariate Gaussian with the mean $\boldsymbol{\mu}_o$ and the covariance matrix $\tilde{\boldsymbol{\Sigma}}_{oo} = \frac{\sigma^2}{u}\mathbf{V}_{oo}^{-1}$. The derivation of a marginal distribution of a subvector of multivariate Gaussian random variable are given in Lemma A1. Using this fact, we have

$$\begin{aligned}
\pi_{U|\Psi}(u)\cdot\left(\prod_{t=1}^N \tilde{\pi}_{\mathbf{Y}_t^o|U,\Psi}(\mathbf{y}_t^o)\right) &= \frac{\left(\frac{v}{2}\right)^{\frac{v}{2}}}{\Gamma\left(\frac{v}{2}\right)}u^{\frac{v}{2}-1}e^{-\frac{v}{2}u}\prod_{t=1}^N(2\pi)^{-\frac{d}{2}}\Big|\frac{\sigma^2}{u}\mathbf{V}_{oo}^{-1}\Big|^{-\frac{1}{2}}\exp\left\{-\frac{u}{2\sigma^2}(\mathbf{y}_t^o-\boldsymbol{\mu}_o)\mathbf{V}_{oo}(\mathbf{y}_t^o-\boldsymbol{\mu}_o)^T\right\}\\
&= \frac{\left(\frac{v}{2}\right)^{\frac{v}{2}}}{\Gamma\left(\frac{v}{2}\right)}u^{\frac{v}{2}-1}e^{-\frac{v}{2}u}(2\pi)^{-\frac{d_oN}{2}}(\sigma^2)^{-\frac{d_oN}{2}}u^{\frac{d_oN}{2}}\Big|\mathbf{V}_{oo}\Big|^{\frac{N}{2}}\exp\left\{-\frac{u}{2\sigma^2}\sum_{t=1}^N(\mathbf{y}_t^o-\boldsymbol{\mu}_o)\mathbf{V}_{oo}(\mathbf{y}_t^o-\boldsymbol{\mu}_o)^T\right\}\\
&= (2\pi)^{-\frac{d_oN}{2}}(\sigma^2)^{-\frac{d_oN}{2}}\Big|\mathbf{V}_{oo}\Big|^{\frac{N}{2}}\frac{\left(\frac{v}{2}\right)^{\frac{v}{2}}}{\Gamma\left(\frac{v}{2}\right)}u^{\frac{v+d_oN}{2}-1}\exp\left\{-u\left(\frac{v}{2}+\frac{1}{2\sigma^2}\sum_{t=1}^N(\mathbf{y}_t^o-\boldsymbol{\mu}_o)\mathbf{V}_{oo}(\mathbf{y}_t^o-\boldsymbol{\mu}_o)^T\right)\right\}.
\end{aligned}$$

Let us denote the scalar $\alpha = \frac{v+d_oN}{2}$ and define a function $\beta : \mathbb{R}^{d_o\times N} \to \mathbb{R}$ such that

$$\beta(\mathbf{y}_{1:N}^o;\Psi) = \frac{v}{2} + \frac{1}{2\sigma^2}\sum_{t=1}^N(\mathbf{y}_t^o-\boldsymbol{\mu}_o)\mathbf{V}_{oo}(\mathbf{y}_t^o-\boldsymbol{\mu}_o)^T.$$

We obtain the following proportionality

$$\pi_{U|\Psi}(u)\cdot\left(\prod_{t=1}^N\tilde{\pi}_{\mathbf{Y}_t^o|U,\Psi}(\mathbf{y}_t^o)\right) \propto u^{\alpha-1}\exp\Big\{-u\beta(\mathbf{y}_{1:N}^o;\Psi)\Big\} \propto \tilde{\pi}_{U|\Psi}(u),$$

where $\tilde{\pi}_{U|\Psi}(u)$ is a probability density function of the a univariate random variable $U \sim \Gamma(\alpha, \beta(\mathbf{y}^o_{1:N}; \Psi))$. We can specify the function $C_\beta : \mathbb{R}^{d_o \times N} \to \mathbb{R}$, which is a constant with respect to U, such that

$$\pi_{U|\Psi}(u) \cdot \left(\prod_{t=1}^{N} \tilde{\pi}_{\mathbf{Y}^o_t|U,\Psi}(\mathbf{y}^o_t) \right) = C_\beta(\mathbf{y}^o_{1:N}, \Psi) \cdot \tilde{\pi}_{U|\Psi}(u)$$

After some derivation, we obtain that

$$C_\beta(\mathbf{y}^o_{1:N}, \Psi) = \frac{\pi_{U|\Psi}(u) \cdot \left(\prod_{t=1}^{N} \tilde{\pi}_{\mathbf{Y}^o_t|U,\Psi}(\mathbf{y}^o_t) \right)}{\tilde{\pi}_{U|\Psi}(u)} = \frac{(2\pi)^{-\frac{d_o N}{2}} (\sigma^2)^{-\frac{d_o N}{2}} \left| \mathbf{V}_{oo} \right|^{\frac{N}{2}} \frac{\left(\frac{v}{2}\right)^{\frac{v}{2}}}{\Gamma\left(\frac{v}{2}\right)}}{\frac{\beta(\mathbf{y}^o_{1:N}, \Psi)^\alpha}{\Gamma(\alpha)}}$$

$$= \frac{(2\pi)^{-\frac{d_o N}{2}} (\sigma^2)^{-\frac{d_o N}{2}} \left| \mathbf{V}_{oo} \right|^{\frac{N}{2}} \left(\frac{v}{2}\right)^{\frac{v}{2}} \Gamma(\alpha)}{\Gamma\left(\frac{v}{2}\right)} \beta(\mathbf{y}^o_{1:N}; \Psi)^{-\alpha}.$$

□

A.4. Generalized Skew t-Student Probabilistic Principal Component Analysis

A.4.1. IND: Complete Data Setting

A.4.1.1. Proof of Theorem 13

Proof. The joint probability function of the variables $\mathbf{Y}_{1:N}$, $\mathbf{X}_{1:N}, \mathbf{U}_{1:N}$ and $\mathbf{V}_{1:N}$ for Independent Generalized Skew t-Student PPCA model is specified in Equation (3.29). Given the time independence of the random vectors $\mathbf{Y}_t$, $\mathbf{X}_t, \mathbf{U}_t$ and $\mathbf{V}_t$, the E-step of the corresponding EM algorithm can be simplified to

$$Q_{GStS,ind}(\Psi, \Psi^*) = \mathbb{E}_{\mathbf{X}_{1:N},\mathbf{U}_{1:N},\mathbf{V}_{1:N}|\mathbf{Y}_{1:N},\Psi} \left[\log \pi_{\mathbf{Y}_{1:N},\mathbf{X}_{1:N},\mathbf{U}_{1:N},\mathbf{V}_{1:N}|\Psi^*} \left(\mathbf{Y}_{1:N}, \mathbf{X}_{1:N}, \mathbf{U}_{1:N}, \mathbf{V}_{1:N} \right) \right]$$

$$= \frac{1}{\pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N})} \int_{\mathbb{R}_+^{N\times d}} \int_{\mathbb{R}_+^{N\times k}} \sum_{t=1}^{N} \left\{ \int_{\mathbb{R}^{N\times k}} \left[\log \left(\pi_{\mathbf{Y}_t,\mathbf{X}_t,\mathbf{U}_t,\mathbf{V}_t|\Psi^*}(\mathbf{y}_t, \mathbf{x}_t, \mathbf{u}_t, \mathbf{v}_t) \right) \prod_{s=1}^{N} \pi_{\mathbf{Y}_s,\mathbf{X}_s,\mathbf{U}_s,\mathbf{V}_s|\Psi}(\mathbf{y}_s, \mathbf{x}_s, \mathbf{u}_s, \mathbf{v}_s) \right] d\mathbf{x}_{1:N} \right\} d\mathbf{u}_{1:N}\, d\mathbf{v}_{1:N}.$$

Using Lemma A8, we can decompose the probability function $\pi_{\mathbf{Y}_t,\mathbf{X}_t,\mathbf{U}_t,\mathbf{V}_t|\Psi}(\mathbf{y}_t, \mathbf{x}_t, \mathbf{u}_t, \mathbf{v}_t)$ into the product of three functions

$$\pi_{\mathbf{Y}_t,\mathbf{X}_t,\mathbf{U}_t,\mathbf{V}_t|\Psi}(\mathbf{y}_t, \mathbf{x}_t, \mathbf{u}_t, \mathbf{v}_t) = \pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{x}_t) \cdot \pi_{\mathbf{Y}_t|\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{y}_t) \cdot C(\mathbf{u}_t, \mathbf{v}_t; \Psi)$$

Please refer to the result derived in Lemma A8 to find the definitions of the probability functions and the function $C(\mathbf{u}_t, \mathbf{v}_t; \Psi) : \mathbb{R}^d_+ \times \mathbb{R}^k_+ \longrightarrow \mathbb{R}$. Therefore, we can further simplify the formulation of the Q function

$$Q_{GStS,ind}(\Psi, \Psi^*) = \frac{1}{\pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N})} \int_{\mathbb{R}_+^{N\times d}} \int_{\mathbb{R}_+^{N\times k}} \sum_{t=1}^{N} \left\{ \int_{\mathbb{R}^k} \log \left(\pi_{\mathbf{Y}_t,\mathbf{X}_t,\mathbf{U}_t,\mathbf{V}_t|\Psi^*}(\mathbf{y}_t, \mathbf{x}_t, \mathbf{u}_t, \mathbf{v}_t) \right) \pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{x}_t)\, d\mathbf{x}_t \right.$$

$$\left. \times \underbrace{\int_{\mathbb{R}^{(N-1)\times k}} \prod_{s=1, s\neq t}^{N} \pi_{\mathbf{X}_s|\mathbf{Y}_s,\mathbf{U}_s,\mathbf{V}_s,\Psi}(\mathbf{x}_s)\, d\mathbf{x}_{1:N}}_{\text{without t element}} \right\} \left(\prod_{s=1}^{N} \pi_{\mathbf{Y}_s|\mathbf{U}_s,\mathbf{V}_s,\Psi}(\mathbf{y}_s) \right) \left(\prod_{s=1}^{N} C(\mathbf{u}_s, \mathbf{v}_s; \Psi) \right) d\mathbf{u}_{1:N}\, d\mathbf{v}_{1:N}$$

$$= \frac{1}{\pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N})} \int_{\mathbb{R}_+^{N\times d}} \int_{\mathbb{R}_+^{N\times k}} \left(\sum_{t=1}^{N} w(\mathbf{y}_t, \mathbf{u}_t, \mathbf{v}_t; \Psi, \Psi^*) \right) \left(\prod_{s=1}^{N} \pi_{\mathbf{Y}_s|\mathbf{U}_s,\mathbf{V}_s,\Psi}(\mathbf{y}_s) \right) \left(\prod_{s=1}^{N} C(\mathbf{u}_t, \mathbf{v}_t; \Psi) \right) d\mathbf{u}_{1:N}\, d\mathbf{v}_{1:N}$$

for the function $w : \mathbb{R}^d \times \mathbb{R}^d_+ \times \mathbb{R}^k_+ \longrightarrow \mathbb{R}$ defined in Lemma A9. Let us denote for the convenience of the notation

$$h(\mathbf{y}_t, \mathbf{u}_t, \mathbf{v}_t; \Psi) := \pi_{\mathbf{Y}_t|\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{y}_t) \cdot C(\mathbf{u}_t, \mathbf{v}_t; \Psi) = \pi_{\mathbf{Y}_t|\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{y}_t) \cdot \pi_{\mathbf{U}_t|\Psi}(\mathbf{u}_t) \cdot \pi_{\mathbf{V}_t|\Psi}(\mathbf{v}_t)$$

$$\times (\sigma^2)^{\frac{k}{2}} \left| \mathbf{D}_{\epsilon,t} \right|^{\frac{1}{2}} \left| \mathbf{D}_{x,t} \right|^{\frac{1}{2}} \left| \mathbf{M}_t \right|^{-\frac{1}{2}} \left| \mathbf{N}_t \right|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} \delta_x \left(\mathbf{D}_{x,t}^{-1} - \sigma^2 \mathbf{M}_t^{-1} \mathbf{W}^T \mathbf{D}_{\epsilon,t} \mathbf{N}_t^{-1} \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1} - \sigma^2 \mathbf{M}_t^{-1} \right) \delta_x^T \right\}$$

the E-step of the Generalized Skew t-Student IND PPCA is the following

$$Q_{GStS,ind}(\Psi,\Psi^*) = \frac{1}{\pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N})}\sum_{t=1}^{N}\left\{\int_{\mathbb{R}_+^d}\int_{\mathbb{R}_+^k} w(\mathbf{y}_t,\mathbf{u}_t,\mathbf{v}_t;\Psi,\Psi^*)\,h(\mathbf{y}_t,\mathbf{u}_t,\mathbf{v}_t;\Psi)\,d\mathbf{u}_t\,d\mathbf{v}_t\prod_{s=1,s\neq t}^{N}\int_{\mathbb{R}_+^d}\int_{\mathbb{R}_+^k} h(\mathbf{y}_s,\mathbf{u}_s,\mathbf{v}_s;\Psi)\,d\mathbf{u}_s\,d\mathbf{v}_s\right\}. \tag{A1}$$

Let us recall the definition of the mutually independent multivariate Gamma random vectors $\mathbf{U}_t$ and $\mathbf{V}_t$ from Equation (3.27) in Subsubsection 3.4.1.1 as transformations of the mutually independent univariate uniform random variables $S_{\epsilon,t},\ S_{x,t} \in \mathcal{U}(0,1)$ by the quantile functions $T_\epsilon : [0,1] \to \mathbb{R}_+^d$ and $T_x : [0,1] \to \mathbb{R}_+^k$ such that

$$T_\epsilon(s) := \left(\frac{\chi^{-1}_{\nu_\epsilon^1}(s)}{\nu_\epsilon^1},\ldots,\frac{\chi^{-1}_{\nu_\epsilon^d}(s)}{\nu_\epsilon^d}\right)_{1\times d} \text{ and } T_x(s) := \left(\frac{\chi^{-1}_{\nu_x^1}(s)}{\nu_x^1},\ldots,\frac{\chi^{-1}_{\nu_x^k}(s)}{\nu_x^k}\right)_{1\times k} \tag{A2}$$

where χ_ν^{-1} is the quantile distribution function of a univariate chi-square random variable with ν degrees of freedom. By the change of variables $S_{\epsilon,t} = T_\epsilon^{-1}(\mathbf{U}_t)$ and $S_{x,t} = T_x^{-1}(\mathbf{V}_t)$, we can reduce the $d \times k$ dimensional integration problems from Equation (A1) to the 2-dimensional integrations on the unit hypercube $[0,1]^2$.

Firstly, we remark that the corresponding density functions $\pi_{\mathbf{U}_t|\Psi}(\mathbf{u}_t)$ and $\pi_{\mathbf{V}_t|\Psi}(\mathbf{v}_t)$ can be expressed by the density function of the random variables $S_{\epsilon,t}$ and $S_{x,t}$ defined in Subsubsection 3.4.1.1, that is

$$\begin{aligned}
\pi_{\mathbf{U}_t|\Psi}(\mathbf{u}_t) &= \pi_{S_{\epsilon,t}|\Psi}(s_{\epsilon,t})\left|\frac{\partial T_\epsilon(s_{\epsilon,t})}{\partial s_{\epsilon,t}}\right|^{-1} = \mathbf{1}_{[0,1]}(s_{\epsilon,t})\left|\frac{\partial T_\epsilon(s_{\epsilon,t})}{\partial s_{\epsilon,t}}\right|^{-1}\\
\pi_{\mathbf{V}_t|\Psi}(\mathbf{v}_t) &= \pi_{S_{x,t}|\Psi}(s_{x,t})\left|\frac{\partial T_x(s_{x,t})}{\partial s_{x,t}}\right|^{-1} = \mathbf{1}_{[0,1]}(s_{x,t})\left|\frac{\partial T_x(s_{x,t})}{\partial s_{x,t}}\right|^{-1}.
\end{aligned} \tag{A3}$$

The distribution of $\mathbf{Y}_t^o$ conditioned on $\mathbf{U}_t$ and $\mathbf{V}_t$ can also by expressed by $S_{\epsilon,t}$ and $S_{x,t}$ as

$$\pi_{\mathbf{Y}_t|T_\epsilon(S_{\epsilon,t}),T_x(S_{x,t}),\Psi}(\mathbf{y}_t) = \pi_{\mathbf{Y}_t|S_{\epsilon,t},S_{x,t},\Psi}(\mathbf{y}_t).$$

Recall that the definitions of the matrices $\mathbf{D}_{\epsilon,t}$ and $\mathbf{D}_{x,t}$ under the transformation $\mathbf{U}_t = T_\epsilon(S_{\epsilon,t})$ and $\mathbf{V}_t = T_x(S_{x,t})$ become

$$\mathbf{D}_{\epsilon,t} = \begin{pmatrix} \frac{\chi^{-1}_{\nu_\epsilon^1}(S_{\epsilon,t})}{\nu_\epsilon^1} & 0 & 0\\ 0 & \ddots & 0\\ 0 & 0 & \frac{\chi^{-1}_{\nu_\epsilon^d}(S_\epsilon)}{\nu_\epsilon^d}\end{pmatrix}_{d\times d},\ \mathbf{D}_{x,t} = \begin{pmatrix} \frac{\chi^{-1}_{\nu_x^1}(S_{x,t})}{\nu_x^1} & 0 & 0\\ 0 & \ddots & 0\\ 0 & 0 & \frac{\chi^{-1}_{\nu_x^k}(S_{x,t})}{\nu_x^k}\end{pmatrix}_{k\times k}.$$

Hence, the E-step of the EM algorithm for the Independent Generalized Skew t-Student PPCA from Equation (A1) can be expressed by the sum of 2-dimensional integration problems

$$Q_{GStS,ind}(\Psi,\Psi^*) = \frac{1}{\pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N})}\sum_{t=1}^{N}\left\{I_1(\mathbf{y}_t;\Psi,\Psi^*)\prod_{s=1,s\neq t}^{N} I_2(\mathbf{y}_s;\Psi)\right\}$$

for the functions $I_1,\ I_2 : \mathbb{R}^d \to \mathbb{R}$

$$\begin{aligned}
I_1(\mathbf{y}_t;\Psi,\Psi^*) &:= \int_0^1\int_0^1 w(\mathbf{y}_t,T_\epsilon(s_{\epsilon,t}),T_x(s_{x,t});\Psi,\Psi^*)\,m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\\
I_2(\mathbf{y}_t;\Psi) &:= \int_0^1\int_0^1 m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}
\end{aligned}$$

and the function $m : \mathbb{R}^d \times [0,1]^2 \longrightarrow \mathbb{R}$

$$m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi) = \pi_{\mathbf{Y}_t|S_{\epsilon,t},S_{x,t},\Psi}(\mathbf{y}_t)(\sigma^2)^{\frac{k}{2}}\left|\mathbf{D}_{\epsilon,t}\right|^{\frac{1}{2}}\left|\mathbf{D}_{x,t}\right|^{\frac{1}{2}}\left|\mathbf{M}_t\right|^{-\frac{1}{2}}\left|\mathbf{N}_t\right|^{-\frac{1}{2}}e^{-\frac{1}{2}\delta_x\left(\mathbf{D}_{x,t}^{-1}-\sigma^2\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbf{N}_t^{-1}\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}-\sigma^2\mathbf{M}_t^{-1}\right)\delta_x^T}.$$

□

A.4.1.2. Proof of Corollary 2

Proof. Let us remark that in under the assumptions of Special Case 2 from Subsection 3.4.1.3, the random vectors $\mathbf{U}_t$ and $\mathbf{V}_t$ reduce to the scalar random variables U_t and V_t. Following the steps from the proof of Theorem 13 for Generalized Skew t-Student IND PPCA, the E-step of the EM algorithm for Special Case 2: Skew t-Student PPCA is the following

$$Q_{GStS,ind,2}(\Psi,\Psi^*) = \frac{1}{\pi_{\mathbf{Y}^o_{1:N}|\Psi}(\mathbf{y}^o_{1:N})}\int_{\mathbb{R}^N_+}\int_{\mathbb{R}^N_+}\int_{\mathbb{R}^{N\times d_m}}\sum_{t=1}^{N}\Bigg\{\int_{\mathbb{R}^{N\times k}}\Bigg[\log\Big(\pi_{\mathbf{Y}_t,\mathbf{X}_t,U_t,V_t|\Psi^*}(\mathbf{y}_t,\mathbf{x}_t,u_t,v_t)\Big)$$
$$\times\prod_{s=1}^{N}\pi_{\mathbf{Y}_s,\mathbf{X}_s,U_s,V_s|\Psi}(\mathbf{y}_s,\mathbf{x}_s,u_s,v_s)\Bigg]\,d\mathbf{x}_{1:N}\Bigg\}du_{1:N}\,dv_{1:N}.$$

Let us remark, that the diagonal matrices $\mathbf{D}_{\epsilon,t}$ and $\mathbf{D}_{x,t}$ from Subsection 3.4.2.1 reduce to the product of scalars and identity matrices, that is

$$\mathbf{D}_{\epsilon,t} = U_t\mathbb{I}_d \text{ and } \mathbf{D}_{x,t} = V_t\mathbb{I}_k \tag{A4}$$

Analogously to the Generalized Skew t-Student IND PPCA case, the probability function $\pi_{\mathbf{Y}_t,\mathbf{X}_t,U_t,V_t|\Psi}(\mathbf{y}_t,\mathbf{x}_t,u_t,v_t)$ can be factorised using derivations from Lemma A8 and the new definitions of the matrices $\mathbf{D}_{\epsilon,t}$ and $\mathbf{D}_{x,t}$, as following

$$\pi_{\mathbf{Y}_t,\mathbf{X}_t,U_t,V_t|\Psi}(\mathbf{y}_t,\mathbf{x}_t,u_t,v_t) = \pi_{\mathbf{X}_t|\mathbf{Y}_t,U_t,V_t,\Psi}(\mathbf{x}_t)\cdot\pi_{\mathbf{Y}_t|U_t,V_t,\Psi}(\mathbf{y}_t)\cdot C(u_t,v_t;\Psi)$$

where $\pi_{\mathbf{X}_t|\mathbf{Y}_t,U_t,V_t,\Psi}(\mathbf{x}_t)$ represents a probability density function of the random vector $\mathbf{X}_t|\mathbf{Y}_t,U_t,V_t,\Psi \sim \mathcal{N}\Big(\boldsymbol{\mu}_{x,t},\boldsymbol{\Sigma}_{x,t}\Big)$ and $\pi_{\mathbf{Y}_t|U_t,V_t,\Psi}(\mathbf{x}_t)$ is a probability density function of the random vector $\mathbf{Y}_t|U_t,V_t,\Psi \sim \mathcal{N}\Big(\boldsymbol{\mu}_{y,t},\boldsymbol{\Sigma}_{y,t}\Big)$ for the corresponding moments

$$\begin{aligned}\boldsymbol{\mu}_{x,t} &= \Big(U_t(\mathbf{y}_t-\boldsymbol{\mu}-U_t^{-1}\boldsymbol{\delta}_\epsilon)\mathbf{W}+\sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}_t^{-1} \text{ and } \boldsymbol{\Sigma}_{x,t} = \sigma^2\mathbf{M}_t^{-1},\\ \boldsymbol{\mu}_{y,t} &= \boldsymbol{\mu}+U_t^{-1}\boldsymbol{\delta}_\epsilon+\sigma^2\boldsymbol{\delta}_x U_t\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{N}_t^{-1} \text{ and } \boldsymbol{\Sigma}_{y,t} = \sigma^2\mathbf{N}_t^{-1}\end{aligned} \tag{A5}$$

for $\mathbf{M}_t = \sigma^2V_t\mathbb{I}_k + U_t\mathbf{W}^T\mathbf{W}$ and $\mathbf{N}_t = U_t\big(\mathbb{I}_d - U_t\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^T\big)$. The function $C:\mathbb{R}^2_+ \longrightarrow \mathbb{R}$ is given by

$$C(u_t,v_t;\Psi) = \pi_{U_t|\Psi}(u_t)\cdot\pi_{V_t|\Psi}(v_t)\big(\sigma^2\big)^{\frac{k}{2}}u_T^{\frac{d}{2}}u_t^{\frac{k}{2}}\Big|\mathbf{M}_t\Big|^{-\frac{1}{2}}\Big|\mathbf{N}_t\Big|^{-\frac{1}{2}}e^{-\frac{1}{2}\boldsymbol{\delta}_x\left(v_t^{-1}\mathbb{I}_k-\sigma^2u_t^2\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{N}_t^{-1}\mathbf{W}\mathbf{M}_t^{-1}-\sigma^2\mathbf{M}_t^{-1}\right)\boldsymbol{\delta}_x^T}.$$

We obtain the following simplification of the function $Q_{GStS,ind,2}$

$$Q_{GStS,ind,2}(\Psi,\Psi^*) = \frac{1}{\pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N})}\sum_{t=1}^{N}\Bigg\{\prod_{s=1,s\neq t}^{N}\int_{\mathbb{R}_+}\int_{\mathbb{R}_+}\pi_{\mathbf{Y}_s|U_s,V_s,\Psi}(\mathbf{y}_s)\,C(u_s,v_s;\Psi)\,du_s\,dv_s$$
$$\times\int_{\mathbb{R}_+}\int_{\mathbb{R}_+}\int_{\mathbb{R}^{d_m}}w(\mathbf{y}_t,u_t,v_t;\Psi,\Psi^*)\,\pi_{\mathbf{Y}_t|U_t,V_t,\Psi}(\mathbf{y}_t)\,C(u_t,v_t;\Psi)\,du_t\,dv_t\Bigg\}.$$

for the special case of the function $w(\mathbf{y}_t,u_t,v_t;\Psi,\Psi^*)$ derived Lemma A9 given new definitions of $\mathbf{D}_{\epsilon,t}$ and $\mathbf{D}_{x,t}$ from Equation A11, that is

$$\begin{aligned}w(\mathbf{y}_t,u_t,v_t;\Psi,\Psi^*) &= \log\pi_{U_t|\Psi^*}(u_t)+\log\pi_{\mathbf{V}_t|\Psi^*}(\mathbf{v}_t)-\frac{k+d}{2}\log 2\pi-\frac{d}{2}\log\sigma^{*2}+\frac{d}{2}\log u_t+\frac{k}{2}\log v_t-\frac{v_t^{-1}}{2}\boldsymbol{\delta}_x^*\boldsymbol{\delta}_x^{*T}\\
&-\frac{u_t}{2\sigma^{*2}}\operatorname{Tr}\Big\{\mathbf{y}_t^T\mathbf{y}_t\Big\}+\frac{1}{\sigma^{*2}}\mathbf{y}_t\Big(u_t\boldsymbol{\mu}^*+\boldsymbol{\delta}_\epsilon^*\Big)^T-\frac{u_t}{2\sigma^{*2}}\boldsymbol{\mu}^*\boldsymbol{\mu}^{*T}-\frac{1}{\sigma^{*2}}\boldsymbol{\mu}^*\boldsymbol{\delta}_\epsilon^{*T}-\frac{u_t^{-1}}{2\sigma^{*2}}\boldsymbol{\delta}_\epsilon^*\boldsymbol{\delta}_\epsilon^{*T}\\
&+\frac{u_t^2}{\sigma^{*2}}\operatorname{Tr}\Big\{\mathbf{y}_t^T\mathbf{y}_t\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}\Big\}+u_t\mathbf{y}_t\mathbf{W}\mathbf{M}_t^{-1}\Big(-\frac{u_t}{\sigma^{*2}}\mathbf{W}^{*T}\boldsymbol{\mu}^{*T}-\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\boldsymbol{\delta}_\epsilon^{*T}+\boldsymbol{\delta}_x^{*T}\Big)\\
&+\Big(\frac{u_t}{\sigma^{*2}}\mathbf{y}_t\mathbf{W}^*-\frac{u_t}{\sigma^{*2}}\boldsymbol{\mu}^*\mathbf{W}^*-\frac{1}{\sigma^{*2}}\boldsymbol{\delta}_\epsilon^*\mathbf{W}^*+\boldsymbol{\delta}_x^*\Big)\mathbf{M}_t^{-1}\Big(-u_t\boldsymbol{\mu}\mathbf{W}-\boldsymbol{\delta}_\epsilon\mathbf{W}+\sigma^2\boldsymbol{\delta}_x\Big)^T\\
&-\frac{u_t^2}{2}\operatorname{Tr}\Big\{\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{y}_t^T\mathbf{y}_t\mathbf{W}\mathbf{M}_t^{-1}\Big(\frac{u_t}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{W}^*+v_t\mathbb{I}_k\Big)\Big\}\\
&-\frac{u_t}{2}\operatorname{Tr}\Big\{\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{y}_t^T\Big(-u_t\boldsymbol{\mu}\mathbf{W}-\boldsymbol{\delta}_\epsilon\mathbf{W}+\sigma^2\boldsymbol{\delta}_x\Big)\Big)\mathbf{M}_t^{-1}\Big(\frac{u_t}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{W}^*+v_t\mathbb{I}_k\Big)\Big\}\\
&-\frac{u_t}{2}\operatorname{Tr}\Big\{\mathbf{M}_t^{-1}\Big(-u_t\boldsymbol{\mu}\mathbf{W}-\boldsymbol{\delta}_\epsilon\mathbf{W}+\sigma^2\boldsymbol{\delta}_x\Big)\Big)^T\mathbf{y}_t\mathbf{W}\mathbf{M}_t^{-1}\Big(\frac{u_t}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{W}^*+v_t\mathbb{I}_k\Big)\Big\}\\
&-\frac{1}{2}\operatorname{Tr}\Big\{\mathbf{M}_t^{-1}\Big(-u_t\boldsymbol{\mu}\mathbf{W}-\boldsymbol{\delta}_\epsilon\mathbf{W}+\sigma^2\boldsymbol{\delta}_x\Big)^T\Big(-u_t\boldsymbol{\mu}\mathbf{W}-\boldsymbol{\delta}_\epsilon\mathbf{W}+\sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}_t^{-1}\Big(\frac{u_t}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{W}^*+v_t\mathbb{I}_k\Big)\Big\}\end{aligned}$$

$$-\frac{\sigma^2}{2}\operatorname{Tr}\left\{\mathbf{M}_t^{-1}\left(\frac{u_t}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{W}^*+v_t\mathbb{I}_k\right)\right\}.$$

In order to obtain an efficient quadrature rule to approximate the solution of $Q_{GStS,ind,2}(\Psi,\Psi^*)$ we recall the quantile transformation from Equation (3.27). Under the assumption of Special Case 2: Skew t-Student IND PPCA, the random variable U_t and V_t are obtained by quantile transformations of the mutually independent univariate uniform random variables $S_{\epsilon,t},\ S_{x,t}\in\mathcal{U}(0,1)$ by functions $T_\epsilon:[0,1]\to\mathbb{R}_+$ and $T_x:[0,1]\to\mathbb{R}_+$ such that

$$T_\epsilon(s):=\frac{\chi^{-1}_{\nu_\epsilon}(s)}{\nu_\epsilon}\text{ and }T_x(s):=\frac{\chi^{-1}_{\nu_x}(s)}{\nu_x}$$

where χ^{-1}_ν is a quantile function of univariate chi-square distribution function with ν degrees of freedom. By the change of variables $S_{\epsilon,t}=T_\epsilon^{-1}(U_t)$ and $S_{x,t}=T_x^{-1}(V_t)$, we can simplify the integration problems from Equation (A13) to the 2-dimensional integrations on $[0,1]^2$ given by

$$Q_{GStS,ind,2}(\Psi,\Psi^*)=\frac{1}{\pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N})}\sum_{t=1}^N\left\{\int_0^1\int_0^1\tilde{w}(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi,\Psi^*)\,m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\prod_{s=1,s\neq t}^N\int_0^1\int_0^1 m(\mathbf{y}_s,s_{\epsilon,s},s_{x,s};\Psi)\,ds_{\epsilon,s}\,ds_{x,s}\right\}$$

for

$$m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)=\pi_{\mathbf{Y}_t^o|S_{\epsilon,t},S_{x,t},\Psi}(\mathbf{y}_t^o)(\sigma^2)^{\frac{k}{2}}T_\epsilon(s_{\epsilon,t})^{\frac{d}{2}}T_x(s_{x,t})^{\frac{k}{2}}\left|\mathbf{M}_t\right|^{-\frac{1}{2}}\left|\mathbf{N}_t\right|^{-\frac{1}{2}}e^{-\frac{1}{2}\delta_x\left(T_x(s_{x,t})^{-1}\mathbb{I}_k-\sigma^2T_\epsilon(s_{\epsilon,t})^2\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{N}_t^{-1}\mathbf{W}\mathbf{M}_t^{-1}-\sigma^2\mathbf{M}_t^{-1}\right)\delta_x^T},$$

the probability function $\pi_{\mathbf{Y}_t|S_{\epsilon,t},S_{x,t},\Psi}(\mathbf{y}_t):=\pi_{\mathbf{Y}_t|T_\epsilon(S_{\epsilon,t}),T_x(S_{x,t}),\Psi}(\mathbf{y}_t)$ and $\tilde{w}(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi,\Psi^*):=w(\mathbf{y}_t,T_\epsilon(s_{\epsilon,t}),T_x(s_{x,t});\Psi,\Psi^*)$. □

A.4.1.3. Proof of Theorem 14

Proof. The M-step of the EM algorithm for Identical and Conditionally Independent Generalized Skew t-Student PPCA evaluates the maximizers of the function $Q_{GStS,ind}(\Psi,\Psi^*)$ defined in Theorem 13 with respect to the vector of static parameters $\Psi^*=\left[\boldsymbol{\mu}^*,\mathbf{W}^*,\sigma^{*2},\delta_\epsilon^*,\delta_x\right]$. Recall that the only element of the function Q which contains components of the vector Ψ^* is the function w specified in Lemma A9 being a component of the integral $I_1(\mathbf{y}_t;\Psi,\Psi^*)$. Hence, the maximizers of the function $Q_{GStS,ind}$ with respect to the vector of static parameters Ψ^*, which are the solution to the system of equations $\nabla_{\Psi^*}Q_{GStS,ind}(\Psi,\Psi^*)=\mathbf{0}$, are calculated as follows

$$\frac{\partial}{\partial\Psi^*}Q_{GStS,ind}(\Psi,\Psi^*)=\frac{\partial}{\partial\Psi^*}\frac{1}{\pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N})}\sum_{t=1}^N\left\{H(\mathbf{y}_{1:N/t};\Psi)I_1(\mathbf{y}_t;\Psi,\Psi^*)\right\}=\mathbf{0}\iff\sum_{t=1}^N\left\{H(\mathbf{y}_{1:N/t};\Psi)\frac{\partial}{\partial\Psi^*}I_1(\mathbf{y}_t;\Psi,\Psi^*)\right\}=\mathbf{0}$$
$$\iff\sum_{t=1}^N\left\{H(\mathbf{y}_{1:N/t};\Psi)\int_0^1\int_0^1\frac{\partial}{\partial\Psi^*}\tilde{w}(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi,\Psi^*)\,m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\right\}=\mathbf{0}.\tag{A6}$$

for $H(\mathbf{y}_{1:N/t};\Psi):=\prod_{s=1,s\neq t}^N I_2(\mathbf{y}_s;\Psi)$ defined as in Equation (3.40). We applied Leibniz integral rule to swap the order of differentiating and integration. The gradient $\frac{\partial}{\partial\Psi^*}\tilde{w}(\mathbf{y}_ts_{\epsilon,t},s_{x,t};\Psi,\Psi^*)$ is determined as following

$$\begin{aligned}\frac{\partial}{\partial\boldsymbol{\mu}^*}\tilde{w}(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi,\Psi^*)&=\frac{1}{\sigma^{*2}}\mathbf{y}_t\mathbf{D}_{\epsilon,t}-\frac{1}{\sigma^{*2}}\boldsymbol{\mu}^*\mathbf{D}_{\epsilon,t}-\frac{1}{\sigma^{*2}}\delta_\epsilon^*-\frac{1}{\sigma^{*2}}\mathbf{y}_t\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\\&+\frac{1}{\sigma^{*2}}\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}+\frac{1}{\sigma^{*2}}\delta_\epsilon\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}-\frac{\sigma^2}{\sigma^{*2}}\delta_x\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\end{aligned}$$

$$\begin{aligned}\frac{\partial}{\partial\mathbf{W}^*}\tilde{w}(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi,\Psi^*)&=\frac{1}{\sigma^{*2}}\mathbf{D}_{\epsilon,t}\mathbf{y}_t^T\mathbf{y}_t\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}-\frac{1}{\sigma^{*2}}\left(\mathbf{D}_{\epsilon,t}\boldsymbol{\mu}^{*T}+\delta_\epsilon^{*T}\right)\mathbf{y}_t\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\\&+\frac{1}{\sigma^{*2}}\left(\mathbf{y}_t\mathbf{D}_{\epsilon,t}-\boldsymbol{\mu}^*\mathbf{D}_{\epsilon,t}-\delta_\epsilon^*\right)^T\left(-\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}-\delta_\epsilon\mathbf{W}+\sigma^2\delta_x\right)\mathbf{M}_t^{-1}-\frac{1}{\sigma^{*2}}\mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbf{y}_t^T\mathbf{y}_t\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\\&+\frac{1}{\sigma^{*2}}\mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}_t^{-1}\left(\mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbf{y}_t^T\left(\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}+\delta_\epsilon\mathbf{W}-\sigma^2\delta_x\right)+\left(\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}+\delta_\epsilon\mathbf{W}-\sigma^2\delta_x\right)^T\mathbf{y}_t\mathbf{D}_{\epsilon,t}\mathbf{W}\right)\mathbf{M}_t^{-1}\\&-\frac{1}{\sigma^{*2}}\mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}_t^{-1}\left(\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}+\delta_\epsilon\mathbf{W}-\sigma^2\delta_x\right)^T\left(\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}+\delta_\epsilon\mathbf{W}-\sigma^2\delta_x\right)\mathbf{M}_t^{-1}-\frac{\sigma^2}{\sigma^{*2}}\mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}_t^{-1}.\end{aligned}$$

The gradient of the function $\tilde{w}$ with respect to the scalar σ^{*2} is given by

$$\frac{\partial}{\partial\sigma^{*2}}\tilde{w}(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi,\Psi^*)=-\frac{d}{2\sigma^{*2}}-\frac{1}{2(\sigma^{*2})^2}\left[-\operatorname{Tr}\left\{\mathbf{y}_t^T\mathbf{y}_t\mathbf{D}_{\epsilon,t}\right\}+2\mathbf{y}_t\left(\mathbf{D}_{\epsilon,t}\boldsymbol{\mu}^{*T}+\delta_\epsilon^{*T}\right)\right.$$

$$
\begin{aligned}
&- \boldsymbol{\mu}^* \mathbf{D}_{\epsilon,t} \boldsymbol{\mu}^{*T} - 2\boldsymbol{\mu}^* \boldsymbol{\delta}_\epsilon^{*T} - \boldsymbol{\delta}_\epsilon^* \mathbf{D}_{\epsilon,t}^{-1} \boldsymbol{\delta}_\epsilon^{*T} + 2\,\mathrm{Tr}\left\{ \mathbf{y}_t^T \mathbf{y}_t \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1} \mathbf{W}^{*T} \mathbf{D}_{\epsilon,t} \right\} \\
&- 2\big(\mathbf{y}_t \mathbf{D}_{\epsilon,t} \mathbf{W} - \boldsymbol{\mu} \mathbf{D}_{\epsilon,t} \mathbf{W} - \boldsymbol{\delta}_\epsilon \mathbf{W} + \sigma^2 \boldsymbol{\delta}_x\big) \mathbf{M}_t^{-1} \mathbf{W}^{*T} \big(\boldsymbol{\mu}^* \mathbf{D}_{\epsilon,t} + \boldsymbol{\delta}_\epsilon^*\big)^T \\
&+ 2\mathbf{y}_t \mathbf{D}_{\epsilon,t} \mathbf{W}^* \mathbf{M}_t^{-1} \big(- \boldsymbol{\mu} \mathbf{D}_{\epsilon,t} \mathbf{W} - \boldsymbol{\delta}_\epsilon \mathbf{W} + \sigma^2 \boldsymbol{\delta}_x\big)^T - \mathrm{Tr}\left\{ \mathbf{M}_t^{-1} \mathbf{W}^T \mathbf{D}_{\epsilon,t} \mathbf{y}_t^T \mathbf{y}_t \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1} \mathbf{W}^{*T} \mathbf{D}_{\epsilon,t} \mathbf{W}^* \right\} \\
&+ \mathrm{Tr}\left\{ \mathbf{M}_t^{-1} \Big(\mathbf{W}^T \mathbf{D}_{\epsilon,t} \mathbf{y}_t^T \big(\boldsymbol{\mu} \mathbf{D}_{\epsilon,t} \mathbf{W} + \boldsymbol{\delta}_\epsilon \mathbf{W} - \sigma^2 \boldsymbol{\delta}_x\big) + \big(\boldsymbol{\mu} \mathbf{D}_{\epsilon,t} \mathbf{W} + \boldsymbol{\delta}_\epsilon \mathbf{W} - \sigma^2 \boldsymbol{\delta}_x\big)^T \mathbf{y}_t \mathbf{D}_{\epsilon,t} \mathbf{W} \Big) \mathbf{M}_t^{-1} \mathbf{W}^{*T} \mathbf{D}_{\epsilon,t} \mathbf{W}^* \right\} \\
&- \mathrm{Tr}\left\{ \mathbf{M}_t^{-1} \big(\boldsymbol{\mu} \mathbf{D}_{\epsilon,t} \mathbf{W} + \boldsymbol{\delta}_\epsilon \mathbf{W} - \sigma^2 \boldsymbol{\delta}_x\big)^T \big(\boldsymbol{\mu} \mathbf{D}_{\epsilon,t} \mathbf{W} + \boldsymbol{\delta}_\epsilon \mathbf{W} \sigma^2 \boldsymbol{\delta}_x\big) \mathbf{M}_t^{-1} \mathbf{W}^{*T} \mathbf{D}_{\epsilon,t} \mathbf{W}^* \right\} - \sigma^2\,\mathrm{Tr}\left\{ \mathbf{M}_t^{-1} \mathbf{W}^{*T} \mathbf{D}_{\epsilon,t} \mathbf{W}^* \right\} \Bigg],
\end{aligned}
$$

whereas the gradients with respect to the vectors $\boldsymbol{\delta}_\epsilon^*$ and $\boldsymbol{\delta}_x^*$ are given by

$$
\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\delta}_\epsilon^*} \tilde{w}\big(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi, \Psi^*\big) &= \frac{1}{\sigma^{*2}} \mathbf{y}_t - \frac{1}{\sigma^{*2}} \boldsymbol{\mu}^* - \frac{1}{\sigma^{*2}} \boldsymbol{\delta}_\epsilon^* \mathbf{D}_{\epsilon,t}^{-1} - \frac{1}{\sigma^{*2}} \mathbf{y}_t \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1} \mathbf{W}^{*T} \\
&+ \frac{1}{\sigma^{*2}} \boldsymbol{\mu} \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1} \mathbf{W}^{*T} + \frac{1}{\sigma^{*2}} \boldsymbol{\delta}_\epsilon \mathbf{W} \mathbf{M}_t^{-1} \mathbf{W}^{*T} - \frac{\sigma^2}{\sigma^{*2}} \boldsymbol{\delta}_x \mathbf{M}_t^{-1} \mathbf{W}^{*T}
\end{aligned}
$$

$$
\frac{\partial}{\partial \boldsymbol{\delta}_x^*} \tilde{w}\big(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi, \Psi^*\big) = -\boldsymbol{\delta}_x^* \mathbf{D}_{x,t}^{-1} + \mathbf{y}_t \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1} - \boldsymbol{\mu} \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1} - \boldsymbol{\delta}_\epsilon \mathbf{W} \mathbf{M}_t^{-1} + \sigma^2 \boldsymbol{\delta}_x \mathbf{M}_t^{-1}.
$$

□

A.4.1.4. Proof of Corollary 4

Proof. Given the formulation of the gradient $\nabla_{\Psi^*} Q_{GStS,ind}$ from Equation (A6) and the system of equations specified in Theorem 14, the maximizers of the function $Q_{GStS,ind}$ with respect to $\boldsymbol{\mu}^*, \sigma^{*2}, \boldsymbol{\delta}_\epsilon^*$ and $\boldsymbol{\delta}_x^*$ have closed form solutions defined by sequences of specific 2-dimensional integration problem determined as following

$$
\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\mu}^*} Q_{GStS,ind}(\Psi, \Psi^*) = \mathbf{0} \Longleftrightarrow \sum_{t=1}^{N} \left\{ H\Big(\mathbf{y}_{1:N/t}^o; \Psi\Big) \int_0^1 \int_0^1 \frac{\partial}{\partial \boldsymbol{\mu}^*} \tilde{w}\big(\mathbf{y}_t^o, s_{\epsilon,t}, s_{x,t}; \Psi, \Psi^*\big)\, m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \right\} = \mathbf{0} \\
&\Longleftrightarrow \sum_{t=1}^{N} \left\{ H\Big(\mathbf{y}_{1:N/t}; \Psi\Big)\, \mathbf{y}_t \int_0^1 \int_0^1 \mathbf{D}_{\epsilon,t}\, m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \right\} \\
&\quad - \boldsymbol{\mu}^* \sum_{t=1}^{N} \left\{ H\Big(\mathbf{y}_{1:N/t}; \Psi\Big) \int_0^1 \int_0^1 \mathbf{D}_{\epsilon,t}\, m(\mathbf{y}_t^o, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \right\} \\
&\quad - \boldsymbol{\delta}_\epsilon^* \sum_{t=1}^{N} \left\{ H\Big(\mathbf{y}_{1:N/t}; \Psi\Big) \int_0^1 \int_0^1 m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \right\} \\
&\quad - \sum_{t=1}^{N} \left\{ H\Big(\mathbf{y}_{1:N/t}; \Psi\Big)\, \mathbf{y}_t \int_0^1 \int_0^1 \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1} \mathbf{W}^{*T} \mathbf{D}_{\epsilon,t}\, m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \right\} \\
&\quad + \boldsymbol{\mu} \sum_{t=1}^{N} \left\{ H\Big(\mathbf{y}_{1:N/t}; \Psi\Big) \int_0^1 \int_0^1 \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1} \mathbf{W}^{*T} \mathbf{D}_{\epsilon,t}\, m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \right\} \\
&\quad + \Big(\boldsymbol{\delta}_\epsilon \mathbf{W} - \sigma^2 \boldsymbol{\delta}_x\Big) \sum_{t=1}^{N} \left\{ H\Big(\mathbf{y}_{1:N/t}; \Psi\Big) \int_0^1 \int_0^1 \mathbf{M}_t^{-1} \mathbf{W}^{*T} \mathbf{D}_{\epsilon,t}\, m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \right\} = \mathbf{0} \\
&\Longleftrightarrow \boldsymbol{\mu}^* = \Big[\mathbf{A}_6(\mathbf{y}_{1:N}^o; \Psi) - \mathbf{A}_{10}(\mathbf{y}_{1:N}; \Psi, \Psi^*) - \boldsymbol{\delta}_\epsilon^* A_0(\mathbf{y}_{1:N}; \Psi) + \boldsymbol{\mu} \mathbf{A}_{13}(\mathbf{y}_{1:N}; \Psi, \Psi^*) \\
&\quad + \Big(\boldsymbol{\delta}_\epsilon \mathbf{W} - \sigma^2 \boldsymbol{\delta}_x\Big) \mathbf{A}_{12}(\mathbf{y}_{1:N}; \Psi, \Psi^*)^T \Big] \mathbf{A}_1(\mathbf{y}_{1:N}; \Psi)^{-1}
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\delta}_\epsilon^*} Q_{GStS,ind}(\Psi, \Psi^*) = \mathbf{0} \Longleftrightarrow \sum_{t=1}^{N} \left\{ H\Big(\mathbf{y}_{1:N/t}; \Psi\Big) \int_0^1 \int_0^1 \frac{\partial}{\partial \boldsymbol{\delta}_\epsilon^*} \tilde{w}\big(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi, \Psi^*\big)\, m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \right\} = \mathbf{0} \\
&\Longleftrightarrow \sum_{t=1}^{N} \left\{ H\Big(\mathbf{y}_{1:N/t}; \Psi\Big)\, \mathbf{y}_t \int_0^1 \int_0^1 m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \right\} \\
&\quad - \boldsymbol{\mu}^* \sum_{t=1}^{N} \left\{ H(\mathbf{y}_{1:N/t}; \Psi) \int_0^1 \int_0^1 m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \right\} \\
&\quad - \boldsymbol{\delta}_\epsilon^* \sum_{t=1}^{N} \left\{ H(\mathbf{y}_{1:N/t}; \Psi) \int_0^1 \int_0^1 \mathbf{D}_{\epsilon,t}^{-1}\, m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \right\} \\
&\quad - \sum_{t=1}^{N} \left\{ H\Big(\mathbf{y}_{1:N/t}; \Psi\Big)\, \mathbf{y}_t \int_0^1 \int_0^1 \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1}\, m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \right\} \mathbf{W}^{*T}
\end{aligned}
$$

$$
\begin{aligned}
&+ \boldsymbol{\mu} \sum_{t=1}^{N} \left\{ H\Big(\mathbf{y}_{1:N/t}; \Psi\Big) \int_0^1 \int_0^1 \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1} \, m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi) \, ds_{\epsilon,t} \, ds_{x,t} \right\} \mathbf{W}^{*T} \\
&+ (\delta_\epsilon \mathbf{W} - \sigma^2 \delta_x) \sum_{t=1}^{N} \left\{ H\Big(\mathbf{y}_{1:N/t}; \Psi\Big) \int_0^1 \int_0^1 \mathbf{M}_t^{-1} \, m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi) \, ds_{\epsilon,t} \, ds_{x,t} \right\} \mathbf{W}^{*T} = \mathbf{0} \\
\Longleftrightarrow \delta_\epsilon^* = &\Big[\mathbf{A}_5(\mathbf{y}_{1:N}; \Psi) - \boldsymbol{\mu}^* A_0(\mathbf{y}_{1:N}; \Psi) - \mathbf{A}_8(\mathbf{y}_{1:N}; \Psi) \mathbf{W}^{*T} + \boldsymbol{\mu} \mathbf{A}_{11}(\mathbf{y}_{1:N}; \Psi) \mathbf{W}^{*T} \\
&+ (\delta_\epsilon \mathbf{W} - \sigma^2 \delta_x) \mathbf{A}_4(\mathbf{y}_{1:N}; \Psi) \mathbf{W}^{*T} \Big] \mathbf{A}_2(\mathbf{y}_{1:N}; \Psi)^{-1}
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\partial}{\partial \delta_x^*} Q_{GStS,ind}(\Psi, \Psi^*) = \mathbf{0} \Longleftrightarrow \sum_{t=1}^{N} \left\{ H\Big(\mathbf{y}_{1:N/t}^o; \Psi\Big) \int_0^1 \int_0^1 \frac{\partial}{\partial \delta_x^*} \tilde{w}(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi, \Psi^*) \, m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi) \, ds_{\epsilon,t} \, ds_{x,t} = \mathbf{0} \right. \\
\Longleftrightarrow &- \delta_x^* \sum_{t=1}^{N} \left\{ H\Big(\mathbf{y}_{1:N/t}; \Psi\Big) \int_0^1 \int_0^1 \mathbf{D}_{x,t}^{-1} \, m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi) \, ds_{\epsilon,t} \, ds_{x,t} \right\} \\
&+ \sum_{t=1}^{N} \left\{ H\Big(\mathbf{y}_{1:N/t}; \Psi\Big) \, \mathbf{y}_t \int_0^1 \int_0^1 \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1} \, m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi) \, ds_{\epsilon,t} \, ds_{x,t} \right\} \\
&- \boldsymbol{\mu} \sum_{t=1}^{N} \left\{ H(\mathbf{y}_{1:N/t}; \Psi) \int_0^1 \int_0^1 \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1} \, m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi) \, ds_{\epsilon,t} \, ds_{x,t} \right\} \\
&- (\delta_\epsilon \mathbf{W} - \sigma^2 \delta_x) \sum_{t=1}^{N} \left\{ H(\mathbf{y}_{1:N/t}; \Psi) \int_0^1 \int_0^1 \mathbf{M}_t^{-1} \, m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi) \, ds_{\epsilon,t} \, ds_{x,t} \right\} = \mathbf{0} \\
\Longleftrightarrow \delta_x^* = &\Big[\mathbf{A}_8(\mathbf{y}_{1:N}; \Psi) - \boldsymbol{\mu} \mathbf{A}_{11}(\mathbf{y}_{1:N}; \Psi) - (\delta_\epsilon \mathbf{W} - \sigma^2 \delta_x) \mathbf{A}_4(\mathbf{y}_{1:N}; \Psi) \Big] \mathbf{A}_3(\mathbf{y}_{1:N}, \Psi)^{-1}
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\partial}{\partial \sigma^{*2}} Q_{GStS,ind}(\Psi, \Psi^*) = \mathbf{0} \Longleftrightarrow \sum_{t=1}^{N} \left\{ H\Big(\mathbf{y}_{1:N/t}^o; \Psi\Big) \int_0^1 \int_0^1 \frac{\partial}{\partial \sigma^{*2}} \tilde{w}(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi, \Psi^*) \, m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi) \, ds_{\epsilon,t} \, ds_{x,t} = \mathbf{0} \right. \\
\Longleftrightarrow &- \frac{d}{2\sigma^{*2}} \sum_{t=1}^{N} \left\{ H(\mathbf{y}_{1:N/t}; \Psi) \int_0^1 \int_0^1 m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi) \, ds_{\epsilon,t} \, ds_{x,t} \right\} \\
&- \frac{1}{2(\sigma^{*2})^2} \Bigg[- \operatorname{Tr} \left\{ \sum_{t=1}^{N} \left\{ H(\mathbf{y}_{1:N/t}; \Psi) \, \mathbf{y}_t^T \mathbf{y}_t \int_0^1 \int_0^1 \mathbf{D}_{\epsilon,t} \, m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi) \, ds_{\epsilon,t} \, ds_{x,t} \right\} \right\} \\
&+ 2 \sum_{t=1}^{N} \left\{ H\Big(\mathbf{y}_{1:N/t}; \Psi\Big) \, \mathbf{y}_t \int_0^1 \int_0^1 \mathbf{D}_{\epsilon,t} \, m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi) \, ds_{\epsilon,t} \, ds_{x,t} \right\} \boldsymbol{\mu}^{*T} \\
&+ 2 \sum_{t=1}^{N} \left\{ H\Big(\mathbf{y}_{1:N/t}; \Psi\Big) \, \mathbf{y}_t \int_0^1 \int_0^1 m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi) \, ds_{\epsilon,t} \, ds_{x,t} \right\} \delta_\epsilon^{*T} \\
&- \boldsymbol{\mu}^* \sum_{t=1}^{N} \left\{ H(\mathbf{y}_{1:N/t}; \Psi) \int_0^1 \int_0^1 \mathbf{D}_{\epsilon,t} \, m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi) \, ds_{\epsilon,t} \, ds_{x,t} \right\} \boldsymbol{\mu}^{*T} \\
&- 2 \boldsymbol{\mu}^* \delta_\epsilon^{*T} \sum_{t=1}^{N} \left\{ H\Big(\mathbf{y}_{1:N/t}; \Psi\Big) \int_0^1 \int_0^1 m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi) \, ds_{\epsilon,t} \, ds_{x,t} \right\} \\
&- \delta_\epsilon^* \sum_{t=1}^{N} \left\{ H\Big(\mathbf{y}_{1:N/t}; \Psi\Big) \int_0^1 \int_0^1 \mathbf{D}_{\epsilon,t}^{-1} \, m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi) \, ds_{\epsilon,t} \, ds_{x,t} \right\} \delta_\epsilon^{*T} \\
&+ 2 \operatorname{Tr} \left\{ \sum_{t=1}^{N} \left\{ H\Big(\mathbf{y}_{1:N/t}; \Psi\Big) \, \mathbf{y}_t^T \mathbf{y}_t \int_0^1 \int_0^1 \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1} \mathbf{W}^{*T} \mathbf{D}_{\epsilon,t} \, m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi) \, ds_{\epsilon,t} \, ds_{x,t} \right\} \right\} \\
&- 2 \sum_{t=1}^{N} \left\{ H\Big(\mathbf{y}_{1:N/t}; \Psi\Big) \, \mathbf{y}_t \int_0^1 \int_0^1 \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1} \mathbf{W}^{*T} \mathbf{D}_{\epsilon,t} \, m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi) \, ds_{\epsilon,t} \, ds_{x,t} \right\} \boldsymbol{\mu}^{*T} \\
&+ 2 \boldsymbol{\mu} \sum_{t=1}^{N} \left\{ H\Big(\mathbf{y}_{1:N/t}^o; \Psi\Big) \int_0^1 \int_0^1 \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1} \mathbf{W}^{*T} \mathbf{D}_{\epsilon,t} \, m(\mathbf{y}_t^o, s_{\epsilon,t}, s_{x,t}; \Psi) \, ds_{\epsilon,t} \, ds_{x,t} \right\} \boldsymbol{\mu}^{*T} \\
&+ 2 \Big(\delta_\epsilon \mathbf{W} - \sigma^2 \delta_x \Big) \sum_{t=1}^{N} \left\{ H(\mathbf{y}_{1:N/t}; \Psi) \int_0^1 \int_0^1 \mathbf{M}_t^{-1} \mathbf{W}^{*T} \mathbf{D}_{\epsilon,t} \, m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi) \, ds_{\epsilon,t} \, ds_{x,t} \right\} \boldsymbol{\mu}^{*T} \\
&- 2 \sum_{t=1}^{N} \left\{ H(\mathbf{y}_{1:N/t}; \Psi) \, \mathbf{y}_t \int_0^1 \int_0^1 \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1} \, m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi) \, ds_{\epsilon,t} \, ds_{x,t} \right\} \mathbf{W}^{*T} \delta_\epsilon^{*T} \\
&+ 2 \boldsymbol{\mu} \sum_{t=1}^{N} \left\{ H(\mathbf{y}_{1:N/t}; \Psi) \int_0^1 \int_0^1 \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1} \, m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi) \, ds_{\epsilon,t} \, ds_{x,t} \right\} \mathbf{W}^{*T} \delta_\epsilon^{*T} \\
&+ 2 \Big(\delta_\epsilon \mathbf{W} - \sigma^2 \delta_x \Big) \sum_{t=1}^{N} \left\{ H(\mathbf{y}_{1:N/t}; \Psi) \int_0^1 \int_0^1 \mathbf{M}_t^{-1} \, m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi) \, ds_{\epsilon,t} \, ds_{x,t} \right\} \mathbf{W}^{*T} \delta_\epsilon^{*T}
\end{aligned}
$$

$$
\begin{aligned}
&- 2\sum_{t=1}^{N}\Big\{H(\mathbf{y}_{1:N/t};\Psi)\ \mathbf{y}_t \int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\ m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\Big\}\boldsymbol{\mu}^T\\
&- 2\sum_{t=1}^{N}\Big\{H(\mathbf{y}_{1:N/t};\Psi)\ \mathbf{y}_t\ \int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}_t^{-1}\ m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\Big\}\left(\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\right)^T\\
&- \sigma^2\operatorname{Tr}\Bigg\{\sum_{t=1}^{N}\Big\{H(\mathbf{y}_{1:N/t};\Psi)\ \int_0^1\int_0^1 \mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\Big\}\mathbf{W}^*\Bigg\}\\
&- \operatorname{Tr}\Bigg\{\sum_{t=1}^{N}\Big\{H(\mathbf{y}_{1:N/t};\Psi)\ \int_0^1\int_0^1 \Big(\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbf{y}_t^T\mathbf{y}_t\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\\
&\quad - \mathbf{M}_t^{-1}\Big(\mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbf{y}_t^T\left(\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}+\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\right)+\left(\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}+\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\right)^T\mathbf{y}_t\mathbf{D}_{\epsilon,t}\mathbf{W}\Big)\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\\
&\quad + \mathbf{M}_t^{-1}\left(\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}+\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\right)^T\left(\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}+\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\right)\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\Big)\ m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\mathbf{W}^*\Big\}\Bigg\} = 0
\end{aligned}
$$

$$
\begin{aligned}
\iff \sigma^{*2} = \frac{1}{dA_0(\mathbf{y}_{1:N};\Psi)}\Big[&A_{20}(\mathbf{y}_{1:N};\Psi)+\boldsymbol{\mu}^*\mathbf{A}_1(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^{*T}+\boldsymbol{\delta}_\epsilon^*\mathbf{A}_2(\mathbf{y}_{1:N};\Psi)\boldsymbol{\delta}_\epsilon^{*T}+2\mathbf{A}_9(\mathbf{y}_{1:N};\Psi,\Psi^*)\boldsymbol{\mu}^T\\
&+2\Big(\mathbf{A}_{10}(\mathbf{y}_{1:N};\Psi,\Psi^*)-\mathbf{A}_6(\mathbf{y}_{1:N};\Psi)-\boldsymbol{\mu}\mathbf{A}_{13}(\mathbf{y}_{1:N};\Psi,\Psi^*)\Big)\boldsymbol{\mu}^{*T}-2A_{21}(\mathbf{y}_{1:N}^o;\Psi,\Psi^*)+\sigma^2\operatorname{Tr}\Big\{\mathbf{A}_{12}(\mathbf{y}_{1:N};\Psi)^T\mathbf{W}^*\Big\}\\
&+\operatorname{Tr}\Big\{\mathbf{A}_{17}(\mathbf{y}_{1:N};\Psi)^T\mathbf{W}^*\Big\}-\operatorname{Tr}\Big\{\mathbf{A}_{18.1}(\mathbf{y}_{1:N};\Psi)^T\mathbf{W}^*\Big\}-\operatorname{Tr}\Big\{\mathbf{A}_{18.2}(\mathbf{y}_{1:N};\Psi)^T\mathbf{W}^*\Big\}+\operatorname{Tr}\Big\{\mathbf{A}_{19}(\mathbf{y}_{1:N};\Psi)^T\mathbf{W}^*\Big\}\\
&+2\Big(\boldsymbol{\mu}^*A_0(\mathbf{y}_{1:N};\Psi)-\mathbf{A}_5(\mathbf{y}_{1:N};\Psi)+\mathbf{A}_8(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}-\boldsymbol{\mu}\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\Big)\boldsymbol{\delta}_\epsilon^{*T}\\
&+2\Big(\mathbf{A}_7(\mathbf{y}_{1:N};\Psi,\Psi^*)-\boldsymbol{\mu}^*\mathbf{A}_{12}(\mathbf{y}_{1:N};\Psi,\Psi^*)-\boldsymbol{\delta}_\epsilon^*\mathbf{W}^*\mathbf{A}_4(\mathbf{y}_{1:N};\Psi)\Big)\left(\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\right)^T\Big]
\end{aligned}
$$

where

$$
\begin{aligned}
A_0(\mathbf{y}_{1:N};\Psi)_{1\times 1} &:= \sum_{t=1}^{N}\Big\{H(\mathbf{y}_{1:N/t};\Psi)\ \int_0^1\int_0^1\ m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\Big\}\\
\mathbf{A}_1(\mathbf{y}_{1:N};\Psi)_{d\times d} &:= \sum_{t=1}^{N}\Big\{H(\mathbf{y}_{1:N/t};\Psi)\ \int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}\ m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\Big\}\\
\mathbf{A}_2(\mathbf{y}_{1:N};\Psi)_{d\times d} &:= \sum_{t=1}^{N}\Big\{H(\mathbf{y}_{1:N/t};\Psi)\ \int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}^{-1}\ m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\Big\}\\
\mathbf{A}_3(\mathbf{y}_{1:N};\Psi)_{k\times k} &:= \sum_{t=1}^{N}\Big\{H(\mathbf{y}_{1:N/t}^o;\Psi)\ \int_0^1\int_0^1 \mathbf{D}_{x,t}^{-1}\ m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\Big\}\\
\mathbf{A}_4(\mathbf{y}_{1:N};\Psi)_{k\times k} &:= \sum_{t=1}^{N}\Big\{H(\mathbf{y}_{1:N/t};\Psi)\ \int_0^1\int_0^1 \mathbf{M}_t^{-1}\ m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\Big\}\\
\mathbf{A}_5(\mathbf{y}_{1:N};\Psi)_{1\times d} &:= \sum_{t=1}^{N}\Big\{H(\mathbf{y}_{1:N/t};\Psi)\ \mathbf{y}_t\ \int_0^1\int_0^1 m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\Big\}\\
\mathbf{A}_6(\mathbf{y}_{1:N};\Psi)_{1\times d} &:= \sum_{t=1}^{N}\Big\{H(\mathbf{y}_{1:N/t};\Psi)\ \mathbf{y}_t\ \int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}\ m(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\Big\}\\
\mathbf{A}_7(\mathbf{y}_{1:N};\Psi,\Psi^*)_{1\times k} &:= \sum_{t=1}^{N}\Big\{H(\mathbf{y}_{1:N/t};\Psi)\ \mathbf{y}_t\ \int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}_t^{-1}\ m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\Big\}\\
\mathbf{A}_8(\mathbf{y}_{1:N}^o;\Psi)_{1\times k} &:= \sum_{t=1}^{N}\Big\{H(\mathbf{y}_{1:N/t};\Psi)\ \mathbf{y}_t\ \int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\ m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\Big\}\\
\mathbf{A}_9(\mathbf{y}_{1:N};\Psi,\Psi^*)_{1\times d} &:= \sum_{t=1}^{N}\Big\{H(\mathbf{y}_{1:N/t};\Psi)\ \mathbf{y}_t\ \int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\ m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\Big\}\\
\mathbf{A}_{10}(\mathbf{y}_{1:N};\Psi,\Psi^*)_{1\times d} &:= \sum_{t=1}^{N}\Big\{H(\mathbf{y}_{1:N/t};\Psi)\ \mathbf{y}_t\ \int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\ m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\Big\}\\
\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi)_{d\times k} &:= \sum_{t=1}^{N}\Big\{H(\mathbf{y}_{1:N/t};\Psi)\ \int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\ m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\Big\}\\
\mathbf{A}_{12}(\mathbf{y}_{1:N};\Psi,\Psi^*)_{d\times k} &:= \sum_{t=1}^{N}\Big\{H(\mathbf{y}_{1:N/t};\Psi)\ \int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}_t^{-1}\ m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\Big\}\\
\mathbf{A}_{13}(\mathbf{y}_{1:N};\Psi,\Psi^*)_{d\times d} &:= \sum_{t=1}^{N}\Big\{H(\mathbf{y}_{1:N/t};\Psi)\ \int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\ m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\Big\}\\
\mathbf{A}_{14}(\mathbf{y}_{1:N};\Psi)_{d\times k} &:= \sum_{t=1}^{N}\Big\{H(\mathbf{y}_{1:N/t};\Psi)\ \int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}\mathbf{y}_t^T\mathbf{y}_t\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\ m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\Big\}\\
\mathbf{A}_{15}(\mathbf{y}_{1:N};\Psi,\Psi^*)_{d\times k} &:= \sum_{t=1}^{N}\Big\{H(\mathbf{y}_{1:N/t};\Psi)\ \int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}\boldsymbol{\mu}^{*T}\mathbf{y}_t\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\ m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\Big\}
\end{aligned}
$$

$$\mathbf{A}_{16}(\mathbf{y}_{1:N};\Psi,\Psi^*)_{d\times k} := \sum_{t=1}^{N}\left\{H(\mathbf{y}_{1:N/t};\Psi)\ \int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}(\mathbf{y}_t-\boldsymbol{\mu}^*)^T(\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}+\delta_\epsilon\mathbf{W}-\sigma^2\delta_x)\mathbf{M}_t^{-1}\ m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\right\}$$

$$\mathbf{A}_{17}(\mathbf{y}_{1:N};\Psi,\Psi^*)_{k\times d} := \sum_{t=1}^{N}\left\{H(\mathbf{y}_{1:N/t};\Psi)\ \int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbf{y}_t^T\mathbf{y}_t\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\ m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\right\}$$

$$\mathbf{A}_{18.1}(\mathbf{y}_{1:N};\Psi,\Psi^*)_{k\times d} := \sum_{t=1}^{N}\left\{H(\mathbf{y}_{1:N/t};\Psi)\ \int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbf{y}_t^T\left(\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}+\delta_\epsilon\mathbf{W}-\sigma^2\delta_x\right)\mathbf{M}_t^{-1}\ m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\right\}$$

$$\mathbf{A}_{18.2}(\mathbf{y}_{1:N};\Psi,\Psi^*)_{k\times d} := \sum_{t=1}^{N}\left\{H(\mathbf{y}_{1:N/t};\Psi)\ \int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}_t^{-1}\left(\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}+\delta_\epsilon\mathbf{W}-\sigma^2\delta_x\right)^T\mathbf{y}_t\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\ m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\right\}$$

$$\begin{aligned}\mathbf{A}_{19}(\mathbf{y}_{1:N};\Psi,\Psi^*)_{k\times d} := \sum_{t=1}^{N}\Big\{&H(\mathbf{y}_{1:N/t};\Psi)\ \int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}_t^{-1}\left(\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}+\delta_\epsilon\mathbf{W}-\sigma^2\delta_x\right)^T\left(\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}+\delta_\epsilon\mathbf{W}-\sigma^2\delta_x\right)\\ &\times\mathbf{M}_t^{-1}\ m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\Big\}\end{aligned}$$

$$A_{20}(\mathbf{y}_{1:N};\Psi)_{1\times 1} := \mathrm{Tr}\left\{\sum_{t=1}^{N}\left\{H(\mathbf{y}_{1:N/t};\Psi)\ \mathbf{y}_t^T\mathbf{y}_t\ \int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}\ m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\right\}\right\}$$

$$A_{21}(\mathbf{y}_{1:N};\Psi,\Psi^*)_{1\times 1} := \mathrm{Tr}\left\{\sum_{t=1}^{N}\left\{H(\mathbf{y}_{1:N/t};\Psi)\ \mathbf{y}_t^T\mathbf{y}_t\ \int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\ m(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\right\}\right\}.$$

The iterative updates of the parameters $\boldsymbol{\mu}$, δ_ϵ, δ_x and σ^2, being the maximizers of the function $Q_{GStS,ind}$ with respect to $\boldsymbol{\mu}^*$, δ_ϵ^*, δ_x^* and σ^{*2}, are obtained in the closed forms determined by the sequences of various integration problems specified on 2-dimensional hypercube $[0,1]^2$. Recall that the equations defining these maximizers are dependent on $\mathbf{W}^*$. The maximizer which correspond to the parameter $\mathbf{W}^*$ is more difficult to obtain since the partial derivative of the function $Q_{GStS,ind}$ with respect to $\mathbf{W}^*$ requires integrating the matrix products which contain $\mathbf{W}^*$ and cannot be further simplified, that is

$$\begin{aligned}&\frac{\partial}{\partial\mathbf{W}^*}Q_{GStS,ind}(\Psi,\Psi^*)=\mathbf{0}\Longleftrightarrow\sum_{t=1}^{N}\left\{H(\mathbf{y}_{1:N/t};\Psi)\ \int_0^1\int_0^1\frac{\partial}{\partial\mathbf{W}^*}\tilde{w}(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi,\Psi^*)\ m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}=\mathbf{0}\right.\\
&\Longleftrightarrow\sum_{t=1}^{N}\left\{H(\mathbf{y}^o_{1:N/t};\Psi)\ \int_0^1\int_0^1\mathbf{D}_{\epsilon,t}\mathbf{y}_t^T\mathbf{y}_t\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\ m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\right\}\\
&-\sum_{t=1}^{N}\left\{H(\mathbf{y}_{1:N/t};\Psi)\right\}\int_0^1\int_0^1\mathbf{D}_{\epsilon,t}\boldsymbol{\mu}^{*T}\mathbf{y}_t\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\ m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\Big\}\\
&-\delta_\epsilon^{*T}\sum_{t=1}^{N}\left\{H(\mathbf{y}_{1:N/t};\Psi)\ \mathbf{y}_t\ \int_0^1\int_0^1\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\ m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\right\}\\
&-\sum_{t=1}^{N}\left\{H(\mathbf{y}_{1:N/t};\Psi)\ \int_0^1\int_0^1\mathbf{D}_{\epsilon,t}(\mathbf{y}_t-\boldsymbol{\mu}^*)^T(\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}+\delta_\epsilon\mathbf{W}-\sigma^2\delta_x)\mathbf{M}_t^{-1}\ m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\right\}\\
&+\delta_\epsilon^{*T}\boldsymbol{\mu}\sum_{t=1}^{N}\left\{H(\mathbf{y}_{1:N/t};\Psi)\ \int_0^1\int_0^1\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\ m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\right\}\\
&+\delta_\epsilon^{*T}(\delta_\epsilon\mathbf{W}-\sigma^2\delta_x)\sum_{t=1}^{N}\left\{H(\mathbf{y}_{1:N/t};\Psi)\ \int_0^1\int_0^1\mathbf{M}_t^{-1}\ m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\right\}\\
&-\sum_{t=1}^{N}\left\{H(\mathbf{y}_{1:N/t};\Psi)\ \int_0^1\int_0^1\mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbf{y}_t^T\mathbf{y}_t\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\ m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\right\}\\
&+\sum_{t=1}^{N}\left\{H(\mathbf{y}_{1:N/t};\Psi)\ \int_0^1\int_0^1\mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbf{y}_t^T\left(\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}+\delta_\epsilon\mathbf{W}-\sigma^2\delta_x\right)\mathbf{M}_t^{-1}\ m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\right\}\\
&+\sum_{t=1}^{N}\left\{H(\mathbf{y}_{1:N/t};\Psi)\ \int_0^1\int_0^1\mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}_t^{-1}\left(\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}+\delta_\epsilon\mathbf{W}-\sigma^2\delta_x\right)^T\mathbf{y}_t\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\ m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\right\}\\
&-\sum_{t=1}^{N}\left\{H(\mathbf{y}_{1:N/t};\Psi)\ \int_0^1\int_0^1\mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}_t^{-1}\left(\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}+\delta_\epsilon\mathbf{W}-\sigma^2\delta_x\right)^T\left(\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}+\delta_\epsilon\mathbf{W}-\sigma^2\delta_x\right)\mathbf{M}_t^{-1}\ m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\right\}\\
&-\sigma^2\sum_{t=1}^{N}\left\{H(\mathbf{y}^o_{1:N/t};\Psi)\ \int_0^1\int_0^1\mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}_t^{-1}\ m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\right\}=\mathbf{0}\\
&\Longleftrightarrow\mathbf{A}_{14}(\mathbf{y}_{1:N};\Psi)-\mathbf{A}_{15}(\mathbf{y}_{1:N};\Psi,\Psi^*)-\delta_\epsilon^{*T}\mathbf{A}_8(\mathbf{y}_{1:N};\Psi)-\mathbf{A}_{16}(\mathbf{y}_{1:N};\Psi,\Psi^*)+\delta_\epsilon^{*T}\boldsymbol{\mu}\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi)\\
&+\delta_\epsilon^{*T}(\delta\mathbf{W}-\sigma^2\mathbf{W})\mathbf{A}_4(\mathbf{y}_{1:N};\Psi)-\mathbf{A}_{17}(\mathbf{y}_{1:N};\Psi,\Psi^*)+\mathbf{A}_{18.1}(\mathbf{y}_{1:N};\Psi,\Psi^*)+\mathbf{A}_{18.2}(\mathbf{y}_{1:N};\Psi,\Psi^*)\\
&-\mathbf{A}_{19}(\mathbf{y}_{1:N};\Psi,\Psi^*)-\sigma^2\mathbf{A}_{12}(\mathbf{y}_{1:N};\Psi,\Psi^*)^T=\mathbf{0}\end{aligned}$$

The maximizer with respect to $\mathbf{W}^*$ requires solving a root finding problem $\frac{\partial Q_{GStS,ind}(\Psi,\Psi^*)}{\partial \mathbf{W}^*} = \mathbf{0}$. Recall that $\frac{\partial Q(\Psi,\Psi^*)}{\partial \mathbf{W}^*}$ contains elements of $\boldsymbol{\mu}^*$ and δ_ϵ^*. By denoting $f : \mathbb{R}^{d\times k} \to \mathbb{R}^{d\times k}$ such that $f(\mathbf{W}^*;\Psi,\Psi^*) := \frac{\partial Q(\Psi,\Psi^*)}{\partial \mathbf{W}^*}$, the system of equations $\nabla_{\Psi^*} Q(\Psi,\Psi^*) = \mathbf{0}$ reduces to

$$
\begin{cases}
\boldsymbol{\mu}^* = \Big[\mathbf{A}_6(\mathbf{y}^o_{1:N};\Psi) - \mathbf{A}_{10}(\mathbf{y}_{1:N};\Psi,\Psi^*) - \delta_\epsilon^* A_0(\mathbf{y}_{1:N};\Psi) + \boldsymbol{\mu}\mathbf{A}_{13}(\mathbf{y}_{1:N};\Psi,\Psi^*) + \Big(\delta_\epsilon\mathbf{W} - \sigma^2\delta_x\Big)\mathbf{A}_{12}(\mathbf{y}_{1:N};\Psi,\Psi^*)^T\Big]\mathbf{A}_1(\mathbf{y}_{1:N};\Psi)^{-1} \\
\delta_\epsilon^* = \Big[\mathbf{A}_5(\mathbf{y}_{1:N};\Psi) - \boldsymbol{\mu}^* A_0(\mathbf{y}_{1:N};\Psi) - \mathbf{A}_8(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T} + \boldsymbol{\mu}\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T} + (\delta_\epsilon\mathbf{W} - \sigma^2\delta_x)\mathbf{A}_4(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\Big]\mathbf{A}_2(\mathbf{y}_{1:N};\Psi)^{-1} \\
\delta_x^* = \Big[\mathbf{A}_8(\mathbf{y}^o_{1:N}) - \boldsymbol{\mu}\mathbf{A}_{11}(\mathbf{y}^o_{1:N}) - (\delta_\epsilon\mathbf{W} - \sigma^2\delta_x)\mathbf{A}_4(\mathbf{y}^o_{1:N})\Big]\mathbf{A}_3(\mathbf{y}_{1:N},\Psi)^{-1} \\
\sigma^{*2} = \frac{1}{dA_0(\mathbf{y}^o_{1:N};\Psi)}\Big[A_{20}(\mathbf{y}^o_{1:N};\Psi) + \boldsymbol{\mu}^*\mathbf{A}_1(\mathbf{y}^o_{1:N};\Psi)\boldsymbol{\mu}^{*T} + \delta_\epsilon^*\mathbf{A}_2(\mathbf{y}^o_{1:N};\Psi)\delta_\epsilon^{*T} + 2\mathbf{A}_9(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\boldsymbol{\mu}^T \\
\qquad +2\Big(\mathbf{A}_{10}(\mathbf{y}^o_{1:N};\Psi,\Psi^*) - \mathbf{A}_6(\mathbf{y}^o_{1:N};\Psi) - \boldsymbol{\mu}\mathbf{A}_{13}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\Big)\boldsymbol{\mu}^{*T} - 2A_{21}(\mathbf{y}^o_{1:N};\Psi,\Psi^*) + \sigma^2\,\mathrm{Tr}\left\{\mathbf{A}_{12}(\mathbf{y}^o_{1:N};\Psi)^T\mathbf{W}^*\right\} \\
\qquad +\mathrm{Tr}\left\{\mathbf{A}_{17}(\mathbf{y}^o_{1:N};\Psi)^T\mathbf{W}^*\right\} - \mathrm{Tr}\left\{\mathbf{A}_{18.1}(\mathbf{y}^o_{1:N};\Psi)^T\mathbf{W}^*\right\} - \mathrm{Tr}\left\{\mathbf{A}_{18.2}(\mathbf{y}^o_{1:N};\Psi)^T\mathbf{W}^*\right\} + \mathrm{Tr}\left\{\mathbf{A}_{19}(\mathbf{y}^o_{1:N};\Psi)^T\mathbf{W}^*\right\} \\
\qquad +2\Big(\boldsymbol{\mu}^* A_0(\mathbf{y}^o_{1:N};\Psi) - \mathbf{A}_5(\mathbf{y}^o_{1:N};\Psi) + \mathbf{A}_8(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}^{*T} - \boldsymbol{\mu}\mathbf{A}_{11}(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}^{*T}\Big)\delta_\epsilon^{*T} \\
\qquad +2\Big(\mathbf{A}_7(\mathbf{y}^o_{1:N};\Psi,\Psi^*) - \boldsymbol{\mu}^*\mathbf{A}_{12}(\mathbf{y}^o_{1:N};\Psi,\Psi^*) - \delta_\epsilon^*\mathbf{W}^*\mathbf{A}_4(\mathbf{y}^o_{1:N};\Psi)\Big)(\delta_\epsilon\mathbf{W} - \sigma^2\delta_x)^T\Big] \\
f(\mathbf{W}^*;\Psi,\Psi^*) = \mathbf{0}.
\end{cases}
$$

□

A.4.1.5. Proof of Corollary 6

Proof. Given the formulation of the gradient $\nabla_{\Psi^*} Q_{GStS,ind}$ from Equation (A6), system of equations specified in Theorem 14 and the assumptions of Special Case 2 from Subsection 3.4.1.3, the M-step for Generalized Skew t-Student IND PPCA is specified finding the the maximizers of the function $Q_{GStS,ind,2}$ from Corollary 2 with respect to $\boldsymbol{\mu}^*$, σ^{*2}, δ_ϵ^* and δ_x^* as following

$$
\begin{aligned}
&\frac{\partial \tilde{w}(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t};\Psi,\Psi^*)}{\partial \boldsymbol{\mu}^*} = \frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})\mathbf{y}_t - \frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}^* - \frac{1}{\sigma^{*2}}\delta_\epsilon^* - \frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})^2\mathbf{y}_t\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T} \\
&\quad - \frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})\Big(-T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W} - \delta_\epsilon\mathbf{W} + \sigma^2\delta_x\Big)\mathbf{M}_t^{-1}\mathbf{W}^{*T} \\
&\implies \frac{\partial Q_{GStS,ind,2}(\Psi,\Psi^*)}{\partial \boldsymbol{\mu}^*} = \mathbf{0} \\
&\iff \sum_{t=1}^{N}\Bigg\{H(\mathbf{y}_{1:N\ t};\Psi)\int_0^1\int_0^1\Bigg[\frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})\mathbf{y}_t - \frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}^* - \frac{1}{\sigma^{*2}}\delta_\epsilon^* - \frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})^2\mathbf{y}_t\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T} \\
&\quad - \frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})\Big(-T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W} - \delta_\epsilon\mathbf{W} + \sigma^2\delta_x\Big)\mathbf{M}_t^{-1}\mathbf{W}^{*T}\Bigg]m(\mathbf{y}_t^o, s_{\epsilon,t}, s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\Bigg\} = \mathbf{0} \\
&\implies \mathbf{A}_6(\mathbf{y}_{1:N};\Psi) - A_1(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^* - A_0(\mathbf{y}_{1:N};\Psi)\delta_\epsilon^* - \mathbf{A}_{10}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T} + \boldsymbol{\mu}\mathbf{W}\mathbf{A}_{13}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T} + (\delta_\epsilon\mathbf{W} + \sigma^2\delta_x)\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T} = \mathbf{0} \\
&\implies \boldsymbol{\mu}^* = A_1(\mathbf{y}_{1:N};\Psi)^{-1}\Big[\mathbf{A}_6(\mathbf{y}_{1:N};\Psi) - A_0(\mathbf{y}_{1:N};\Psi)\delta_\epsilon^* - \mathbf{A}_{10}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T} + \boldsymbol{\mu}\mathbf{W}\mathbf{A}_{13}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T} \\
&\qquad + (\delta_\epsilon\mathbf{W} + \sigma^2\delta_x)\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\Big]
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\partial \tilde{w}(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t};\Psi,\Psi^*)}{\partial \delta_\epsilon^*} = \frac{1}{\sigma^{*2}}\mathbf{Y}_t - \frac{1}{\sigma^{*2}}\boldsymbol{\mu}^* - \frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})^{-1}\delta_\epsilon^* - \frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})\mathbf{y}_t\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T} + \frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T} \\
&\qquad + \frac{1}{\sigma^{*2}}\Big(\delta_\epsilon\mathbf{W} - \sigma^2\delta_x\Big)\mathbf{M}_t^{-1}\mathbf{W}^{*T} \\
&\implies \frac{\partial Q_{GStS,ind,2}(\Psi,\Psi^*)}{\partial \delta_\epsilon^*} = \mathbf{0} \\
&\iff \sum_{t=1}^{N}\Bigg\{H(\mathbf{y}_{1:N\ t};\Psi)\int_0^1\int_0^1\Bigg[\mathbf{y}_t - \boldsymbol{\mu}^* - T_\epsilon(s_{\epsilon,t})^{-1}\delta_\epsilon^* - T_\epsilon(s_{\epsilon,t})\mathbf{Y}_t\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T} + T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T} \\
&\qquad + \Big(\delta_\epsilon\mathbf{W} - \sigma^2\delta_x\Big)\mathbf{M}_t^{-1}\mathbf{W}^{*T}\Bigg]m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\Bigg\} = \mathbf{0} \\
&\implies \mathbf{A}_5(\mathbf{y}_{1:N};\Psi) - A_0(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^* - A_2(\mathbf{y}_{1:N};\Psi)\delta_\epsilon^* - \mathbf{A}_8(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T} + \boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T} + \Big(\delta_\epsilon\mathbf{W} - \sigma^2\delta_x\Big)\mathbf{A}_4(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T} = \mathbf{0} \\
&\implies \delta_\epsilon^* = A_2(\mathbf{y}_{1:N};\Psi)^{-1}\Big(\mathbf{A}_5(\mathbf{y}_{1:N};\Psi) - A_0(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^* - \mathbf{A}_8(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T} + \boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T} + \Big(\delta_\epsilon\mathbf{W} - \sigma^2\delta_x\Big)\mathbf{A}_4(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\Big)
\end{aligned}
$$

$$\frac{\partial \tilde{w}(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi, \Psi^*)}{\partial \boldsymbol{\delta}_x^*} = -T_x(s_{x,t})^{-1}\boldsymbol{\delta}_x^* + T_\epsilon(s_{\epsilon,t})\mathbf{y}_t\mathbf{W}\mathbf{M}_t^{-1} - T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W}\mathbf{M}_t^{-1} - \big(\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\big)\mathbf{M}_t^{-1}$$
$$\implies \frac{\partial Q_{GStS,ind,2}(\Psi, \Psi^*)}{\partial \boldsymbol{\delta}_x^*} = \mathbf{0}$$
$$\iff \sum_{t=1}^{N}\Bigg\{H\big(\mathbf{y}_{1:N\,t};\Psi\big)\int_0^1\int_0^1\Big[-T_x(s_{x,t})^{-1}\boldsymbol{\delta}_x^* + T_\epsilon(s_{\epsilon,t})\mathbf{y}_t\mathbf{W}\mathbf{M}_t^{-1} - T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W}\mathbf{M}_t^{-1} - \big(\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\big)\mathbf{M}_t^{-1}\Big]$$
$$m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t}\Bigg\} = \mathbf{0}$$
$$\implies -A_3(\mathbf{y}_{1:N};\Psi)\boldsymbol{\delta}_x^* + \mathbf{A}_8(\mathbf{y}_{1:N};\Psi) - \boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi) - \big(\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\big)\mathbf{A}_4(\mathbf{y}_{1:N};\Psi) = \mathbf{0}$$
$$\implies \boldsymbol{\delta}_x^* = A_3(\mathbf{y}_{1:N};\Psi)^{-1}\Big(\mathbf{A}_8(\mathbf{y}_{1:N};\Psi) - \boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi) - \big(\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\big)\mathbf{A}_4(\mathbf{y}_{1:N};\Psi)\Big)$$

$$\frac{\partial \tilde{w}(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi, \Psi^*)}{\partial \sigma^{*2}} = -\frac{d}{2\sigma^{*2}} + \frac{1}{2\big(\sigma^{*2}\big)^2}\Bigg[T_\epsilon(s_{\epsilon,t})\operatorname{Tr}\Big\{\mathbf{y}_t^T\mathbf{y}_t\Big\} - 2\mathbf{y}_t\Big(T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}^{*T} + \boldsymbol{\delta}_\epsilon^{*T}\Big) + T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}^*\boldsymbol{\mu}^{*T} + 2\boldsymbol{\mu}^*\boldsymbol{\delta}_\epsilon^{*T} + T_\epsilon(s_{\epsilon,t})^{-1}\boldsymbol{\delta}_\epsilon^*\boldsymbol{\delta}_\epsilon^{*T}$$
$$- 2T_\epsilon(s_{\epsilon,t})^2\operatorname{Tr}\Big\{\mathbf{y}_t^T\mathbf{y}_t\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}\Big\} + 2T_\epsilon(s_{\epsilon,t})\mathbf{y}_t\Big]\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}\Big(T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}^{*T} + \boldsymbol{\delta}_\epsilon^{*T}\Big)$$
$$- 2T_\epsilon(s_{\epsilon,t})\mathbf{y}_t\mathbf{W}^*\mathbf{M}_t^{-1}\Big(-\boldsymbol{\mu}T_\epsilon(s_{\epsilon,t})\mathbf{W} - \boldsymbol{\delta}_\epsilon\mathbf{W} + \sigma^2\boldsymbol{\delta}_x\Big)^T$$
$$+ 2\Big(-T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W} - \boldsymbol{\delta}_\epsilon\mathbf{W} + \sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}_t^{-1}\mathbf{W}^{*T}\Big(T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}^{*T} + \boldsymbol{\delta}_\epsilon^{*T}\Big) + T_\epsilon(s_{\epsilon,t})^3\operatorname{Tr}\Big\{\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{y}_t^T\mathbf{y}_t\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{W}^*\Big\}$$
$$+ T_\epsilon(s_{\epsilon,t})^2\operatorname{Tr}\Big\{\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{y}_t^T\Big(-T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W} - \boldsymbol{\delta}_\epsilon\mathbf{W} + \sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{W}^*\Big\}$$
$$+ T_\epsilon(s_{\epsilon,t})^2\operatorname{Tr}\Big\{\mathbf{M}_t^{-1}\Big(-T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W} - \boldsymbol{\delta}_\epsilon\mathbf{W} + \sigma^2\boldsymbol{\delta}_x\Big)^T\mathbf{y}_t\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{W}^*\Big\}$$
$$+ T_\epsilon(s_{\epsilon,t})\operatorname{Tr}\Big\{\mathbf{M}_t^{-1}\Big(-T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W} - \boldsymbol{\delta}_\epsilon\mathbf{W} + \sigma^2\boldsymbol{\delta}_x\Big)^T\Big(-T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W} - \boldsymbol{\delta}_\epsilon\mathbf{W} + \sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{W}^*\Big\} + \sigma^2T_\epsilon(s_{\epsilon,t})\operatorname{Tr}\Big\{\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{W}^*\Big\}\Bigg]$$
$$\implies \frac{\partial Q_{GStS,ind,2}(\Psi, \Psi^*)}{\partial \sigma^{*2}} = \mathbf{0}$$
$$\iff -\frac{d}{2\sigma^{*2}}A_0(\mathbf{y}_{1:N};\Psi) + \frac{1}{2\big(\sigma^{*2}\big)^2}\Bigg[A_{20}(\mathbf{y}_{1:N};\Psi) - 2\mathbf{A}_6(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^{*T} - 2\mathbf{A}_5(\mathbf{y}_{1:N};\Psi)\boldsymbol{\delta}_\epsilon^{*T}$$
$$+ A_1(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^*\boldsymbol{\mu}^{*T} + 2A_0(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^*\boldsymbol{\delta}_\epsilon^{*T} + A_2(\mathbf{y}_{1:N};\Psi)\boldsymbol{\delta}_\epsilon^*\boldsymbol{\delta}_\epsilon^{*T} - 2\operatorname{Tr}\Big\{\mathbf{A}_{14}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\Big\} + 2\mathbf{A}_{10}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\boldsymbol{\mu}^{*T}$$
$$+ 2\mathbf{A}_8(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\boldsymbol{\delta}_\epsilon^{*T} + 2\mathbf{A}_9(\mathbf{y}_{1:N};\Psi,\Psi^*)\mathbf{W}^T\boldsymbol{\mu}^T + 2\mathbf{A}_7(\mathbf{y}_{1:N};\Psi,\Psi)\mathbf{W}^T\boldsymbol{\delta}_\epsilon^T$$
$$- 2\sigma^2\mathbf{A}_7(\mathbf{y}_{1:N};\Psi,\Psi)\boldsymbol{\delta}_x^T - 2\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{13}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\boldsymbol{\mu}^{*T} - 2\big(\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\big)\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\boldsymbol{\mu}^{*T}$$
$$- 2\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\boldsymbol{\delta}_\epsilon^{*T} - 2\big(\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\big)\mathbf{A}_4(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}\boldsymbol{\delta}_\epsilon^{*T} + \operatorname{Tr}\Big\{\mathbf{A}_{17}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\mathbf{W}^*\Big\}$$
$$- \operatorname{Tr}\Big\{\Big(\mathbf{A}_{18}(\mathbf{y}_{1:N};\Psi) + \mathbf{A}_{18}(\mathbf{y}_{1:N};\Psi)^T\Big)\mathbf{W}^{*T}\mathbf{W}^*\Big\} + \operatorname{Tr}\Big\{\mathbf{A}_{19}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\mathbf{W}^*\Big\} + \sigma^2\operatorname{Tr}\Big\{\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\mathbf{W}^*\Big\}\Bigg] = 0$$
$$\iff \sigma^{*2} = \frac{1}{dA_0(\mathbf{y}_{1:N};\Psi)}\Bigg[A_{20}(\mathbf{y}_{1:N};\Psi) - 2\mathbf{A}_6(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^{*T} - 2\mathbf{A}_5(\mathbf{y}_{1:N};\Psi)\boldsymbol{\delta}_\epsilon^{*T} + A_1(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^*\boldsymbol{\mu}^{*T} + 2A_0(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^*\boldsymbol{\delta}_\epsilon^{*T}$$
$$+ A_2(\mathbf{y}_{1:N};\Psi)\boldsymbol{\delta}_\epsilon^*\boldsymbol{\delta}_\epsilon^{*T} - 2\operatorname{Tr}\Big\{\mathbf{A}_{14}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\Big\} + 2\mathbf{A}_9(\mathbf{y}_{1:N};\Psi,\Psi^*)\mathbf{W}^T\boldsymbol{\mu}^T + 2\mathbf{A}_7(\mathbf{y}_{1:N};\Psi,\Psi)\mathbf{W}^T\boldsymbol{\delta}_\epsilon^T$$
$$- 2\sigma^2\mathbf{A}_7(\mathbf{y}_{1:N};\Psi,\Psi)\boldsymbol{\delta}_x^T + 2\Big(\mathbf{A}_{10}(\mathbf{y}_{1:N};\Psi) - \boldsymbol{\mu}\mathbf{W}\mathbf{A}_{13}(\mathbf{y}_{1:N};\Psi) - \big(\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\big)\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi)\Big)\mathbf{W}^{*T}\boldsymbol{\mu}^{*T}$$
$$+ 2\Big(\mathbf{A}_8(\mathbf{y}_{1:N};\Psi) - \boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}_{1:N}^o;\Psi) - \big(\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\big)\mathbf{A}_4(\mathbf{y}_{1:N};\Psi)\Big)\mathbf{W}^{*T}\boldsymbol{\delta}_\epsilon^{*T} + \operatorname{Tr}\Big\{\mathbf{A}_{17}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\mathbf{W}^*\Big\}$$
$$- \operatorname{Tr}\Big\{\Big(\mathbf{A}_{18}(\mathbf{y}_{1:N};\Psi) + \mathbf{A}_{18}(\mathbf{y}_{1:N};\Psi)^T\Big)\mathbf{W}^{*T}\mathbf{W}^*\Big\} + \operatorname{Tr}\Big\{\mathbf{A}_{19}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\mathbf{W}^*\Big\} + \sigma^2\operatorname{Tr}\Big\{\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\mathbf{W}^*\Big\}\Bigg]$$

$$\frac{\partial \tilde{w}(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi, \Psi^*)}{\partial \mathbf{W}^*} = \frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})^2\mathbf{y}_t^T\mathbf{y}_t\mathbf{W}\mathbf{M}_t^{-1} - \frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})^2\boldsymbol{\mu}^{*T}\mathbf{y}_t\mathbf{W}\mathbf{M}_t^{-1} - \frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})\boldsymbol{\delta}_\epsilon^{*T}\mathbf{y}_t\mathbf{W}\mathbf{M}_t^{-1} - \frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})^2\mathbf{y}_t^T\boldsymbol{\mu}\mathbf{W}\mathbf{M}_t^{-1}$$
$$- \frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})\mathbf{y}_t^T\Big(\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}_t^{-1} + \frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})^2\boldsymbol{\mu}^{*T}\boldsymbol{\mu}\mathbf{W}\mathbf{M}_t^{-1} + \frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}^{*T}\Big(\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}_t^{-1} + \frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})\boldsymbol{\delta}_\epsilon^{*T}\boldsymbol{\mu}\mathbf{W}\mathbf{M}_t^{-1}$$
$$+ \frac{1}{\sigma^{*2}}\boldsymbol{\delta}_\epsilon^{*T}\Big(\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}_t^{-1} - \frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})^3\mathbf{W}^*\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{y}_t^T\mathbf{y}_t\mathbf{W}\mathbf{M}_t^{-1} + \frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})^2\mathbf{W}^*\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{y}_t^T\Big(T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W} + \boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}_t^{-1}$$

$$
\begin{aligned}
&+ \frac{1}{\sigma^{*2}} T_\epsilon(s_{\epsilon,t})^2 \mathbf{W}^* \mathbf{M}_t^{-1} \Big(T_\epsilon(s_{\epsilon,t}) \boldsymbol{\mu}\mathbf{W} + \boldsymbol{\delta}_\epsilon \mathbf{W} - \sigma^2 \boldsymbol{\delta}_x \Big)^T \mathbf{y}_t \mathbf{W}^T \mathbf{M}_t^{-1} \\
&- \frac{1}{\sigma^{*2}} T_\epsilon(s_{\epsilon,t}) \mathbf{W}^* \mathbf{M}_t^{-1} \Big(T_\epsilon(s_{\epsilon,t}) \boldsymbol{\mu}\mathbf{W} + \boldsymbol{\delta}_\epsilon \mathbf{W} - \sigma^2 \boldsymbol{\delta}_x \Big)^T \Big(T_\epsilon(s_{\epsilon,t}) \boldsymbol{\mu}\mathbf{W} + \boldsymbol{\delta}_\epsilon \mathbf{W} - \sigma^2 \boldsymbol{\delta}_x \Big) \mathbf{M}_t^{-1} - \frac{\sigma^2}{\sigma^{*2}} T_\epsilon(s_{\epsilon,t}) \mathbf{W}^* \mathbf{M}_t^{-1} \\
\Longrightarrow\ & \frac{\partial Q_{GStS,ind,2}(\Psi, \Psi^*)}{\partial \mathbf{W}^*} = \mathbf{0} \\
\Longleftrightarrow\ & \sum_{t=1}^{N} \Bigg\{ H(\mathbf{y}_{1:N\ t}; \Psi) \int_0^1 \int_0^1 \Big[T_\epsilon(s_{\epsilon,t})^2 \mathbf{y}_t^T \mathbf{y}_t \mathbf{W} \mathbf{M}_t^{-1} - T_\epsilon(s_{\epsilon,t})^2 \boldsymbol{\mu}^{*T} \mathbf{y}_t \mathbf{W} \mathbf{M}_t^{-1} - T_\epsilon(s_{\epsilon,t}) \boldsymbol{\delta}_\epsilon^{*T} \mathbf{y}_t \mathbf{W} \mathbf{M}_t^{-1} - T_\epsilon(s_{\epsilon,t})^2 \mathbf{y}_t^T \boldsymbol{\mu} \mathbf{W} \mathbf{M}_t^{-1} \\
&- T_\epsilon(s_{\epsilon,t}) \mathbf{y}_t^T \Big(\boldsymbol{\delta}_\epsilon \mathbf{W} - \sigma^2 \boldsymbol{\delta}_x \Big) \mathbf{M}_t^{-1} + T_\epsilon(s_{\epsilon,t})^2 \boldsymbol{\mu}^{*T} \boldsymbol{\mu} \mathbf{W} \mathbf{M}_t^{-1} + T_\epsilon(s_{\epsilon,t}) \boldsymbol{\mu}^{*T} \Big(\boldsymbol{\delta}_\epsilon \mathbf{W} - \sigma^2 \boldsymbol{\delta}_x \Big) \mathbf{M}_t^{-1} + T_\epsilon(s_{\epsilon,t}) \boldsymbol{\delta}_\epsilon^{*T} \boldsymbol{\mu} \mathbf{W} \mathbf{M}_t^{-1} \\
&+ \boldsymbol{\delta}_\epsilon^{*T} \Big(\boldsymbol{\delta}_\epsilon \mathbf{W} - \sigma^2 \boldsymbol{\delta}_x \Big) \mathbf{M}_t^{-1} - T_\epsilon(s_{\epsilon,t})^3 \mathbf{W}^* \mathbf{M}_t^{-1} \mathbf{W}^T \mathbf{y}_t^T \mathbf{y}_t \mathbf{W} \mathbf{M}_t^{-1} + T_\epsilon(s_{\epsilon,t})^2 \mathbf{W}^* \mathbf{M}_t^{-1} \mathbf{W}^T \mathbf{y}_t^T \Big(T_\epsilon(s_{\epsilon,t}) \boldsymbol{\mu}\mathbf{W} + \boldsymbol{\delta}_\epsilon \mathbf{W} - \sigma^2 \boldsymbol{\delta}_x \Big) \mathbf{M}_t^{-1} \\
&+ T_\epsilon(s_{\epsilon,t})^2 \mathbf{W}^* \mathbf{M}_t^{-1} \Big(T_\epsilon(s_{\epsilon,t}) \boldsymbol{\mu}\mathbf{W} + \boldsymbol{\delta}_\epsilon \mathbf{W} - \sigma^2 \boldsymbol{\delta}_x \Big)^T \mathbf{y}_t \mathbf{W} \mathbf{M}_t^{-1} - T_\epsilon(s_{\epsilon,t}) \mathbf{W}^* \mathbf{M}_t^{-1} \Big(T_\epsilon(s_{\epsilon,t}) \boldsymbol{\mu}\mathbf{W} + \boldsymbol{\delta}_\epsilon \mathbf{W} \\
&- \sigma^2 \boldsymbol{\delta}_x \Big)^T \Big(T_\epsilon(s_{\epsilon,t}) \boldsymbol{\mu}\mathbf{W} + \boldsymbol{\delta}_\epsilon \mathbf{W} - \sigma^2 \boldsymbol{\delta}_x \Big) \mathbf{M}_t^{-1} - \sigma^2 T_\epsilon(s_{\epsilon,t}) \mathbf{W}^* \mathbf{M}_t^{-1} \Big] m(\mathbf{y}_t^o, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \Bigg\} = \mathbf{0} \\
\Longleftrightarrow\ & \mathbf{A}_{14}(\mathbf{y}_{1:N}; \Psi) - \boldsymbol{\mu}^{*T} \mathbf{A}_{10}(\mathbf{y}_{1:N}; \Psi) - \boldsymbol{\delta}_\epsilon^* \mathbf{A}_8(\mathbf{y}_{1:N}; \Psi) - \mathbf{A}_{16}(\mathbf{y}_{1:N}; \Psi) + \boldsymbol{\mu}^{*T} \boldsymbol{\mu} \mathbf{W} \mathbf{A}_{13}(\mathbf{y}_{1:N}; \Psi) + \boldsymbol{\mu}^{*T} \big(\boldsymbol{\delta}_\epsilon - \sigma^2 \boldsymbol{\delta}_x \big) \mathbf{A}_{11}(\mathbf{y}_{1:N}; \Psi) \\
&+ \boldsymbol{\delta}_\epsilon^{*T} \boldsymbol{\mu} \mathbf{W} \mathbf{A}_{11}(\mathbf{y}_{1:N}; \Psi) + \boldsymbol{\delta}_\epsilon^{*T} \big(\boldsymbol{\delta}_\epsilon \mathbf{W} - \sigma^2 \boldsymbol{\delta}_x \big) \mathbf{A}_4(\mathbf{y}_{1:N}; \Psi) \\
&- \mathbf{W}^* \Big(\mathbf{A}_{17}(\mathbf{y}_{1:N}; \Psi) - \mathbf{A}_{18}(\mathbf{y}_{1:N}; \Psi) - \mathbf{A}_{18}(\mathbf{y}_{1:N}; \Psi)^T + \mathbf{A}_{19}(\mathbf{y}_{1:N}; \Psi) + \sigma^2 \mathbf{A}_{11}(\mathbf{y}_{1:N}; \Psi) \Big) = \mathbf{0} \\
\Longleftrightarrow\ & \mathbf{W}^* = \Big[\mathbf{A}_{14}(\mathbf{y}_{1:N}; \Psi) - \boldsymbol{\mu}^{*T} \mathbf{A}_{10}(\mathbf{y}_{1:N}; \Psi) - \boldsymbol{\delta}_\epsilon^* \mathbf{A}_8(\mathbf{y}_{1:N}; \Psi) - \mathbf{A}_{16}(\mathbf{y}_{1:N}; \Psi) + \boldsymbol{\mu}^{*T} \boldsymbol{\mu} \mathbf{W} \mathbf{A}_{13}(\mathbf{y}_{1:N}; \Psi) \\
&+ \boldsymbol{\mu}^{*T} \big(\boldsymbol{\delta}_\epsilon - \sigma^2 \boldsymbol{\delta}_x \big) \mathbf{A}_{11}(\mathbf{y}_{1:N}; \Psi) + \boldsymbol{\delta}_\epsilon^{*T} \boldsymbol{\mu} \mathbf{W} \mathbf{A}_{11}(\mathbf{y}_{1:N}; \Psi) + \boldsymbol{\delta}_\epsilon^{*T} \big(\boldsymbol{\delta}_\epsilon \mathbf{W} - \sigma^2 \boldsymbol{\delta}_x \big) \mathbf{A}_4(\mathbf{y}_{1:N}; \Psi) \Big] \\
&\times \Big[\mathbf{A}_{17}(\mathbf{y}_{1:N}; \Psi) - \mathbf{A}_{18}(\mathbf{y}_{1:N}; \Psi) - \mathbf{A}_{18}(\mathbf{y}_{1:N}; \Psi)^T + \mathbf{A}_{19}(\mathbf{y}_{1:N}; \Psi) + \sigma^2 \mathbf{A}_{11}(\mathbf{y}_{1:N}; \Psi) \Big]^{-1}
\end{aligned}
$$

for

$$
\begin{aligned}
A_0(\mathbf{y}_{1:N}; \Psi)_{1\times 1} &:= \sum_{t=1}^{N} \Big\{ H(\mathbf{y}_{1:N/t}; \Psi) \int_0^1 \int_0^1 m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \Big\} \\
A_1(\mathbf{y}_{1:N}; \Psi)_{1\times 1} &:= \sum_{t=1}^{N} \Big\{ H(\mathbf{y}_{1:N/t}; \Psi) \int_0^1 \int_0^1 T_\epsilon(s_{\epsilon,t})\, m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \Big\} \\
A_2(\mathbf{y}_{1:N}; \Psi)_{1\times 1} &:= \sum_{t=1}^{N} \Big\{ H(\mathbf{y}_{1:N/t}; \Psi) \int_0^1 \int_0^1 T_\epsilon(s_{\epsilon,t})^{-1}\, m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \Big\} \\
A_3(\mathbf{y}_{1:N}; \Psi)_{1\times 1} &:= \sum_{t=1}^{N} \Big\{ H(\mathbf{y}_{1:N/t}; \Psi) \int_0^1 \int_0^1 T_x(s_{x,t})^{-1}\, m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \Big\} \\
\mathbf{A}_4(\mathbf{y}_{1:N}; \Psi)_{k\times k} &:= \sum_{t=1}^{N} \Big\{ H(\mathbf{y}_{1:N/t}; \Psi) \int_0^1 \int_0^1 \mathbf{M}_t^{-1}\, m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \Big\} \\
\mathbf{A}_5(\mathbf{y}_{1:N}; \Psi)_{1\times d} &:= \sum_{t=1}^{N} \Big\{ H(\mathbf{y}_{1:N/t}; \Psi)\, \mathbf{y}_t \int_0^1 \int_0^1 m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \Big\} \\
\mathbf{A}_6(\mathbf{y}_{1:N}; \Psi)_{1\times d} &:= \sum_{t=1}^{N} \Big\{ H(\mathbf{y}_{1:N/t}; \Psi)\, \mathbf{y}_t \int_0^1 \int_0^1 T_\epsilon(s_{\epsilon,t})\, m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \Big\} \\
\mathbf{A}_7(\mathbf{y}_{1:N}; \Psi, \Psi^*)_{1\times k} &:= \sum_{t=1}^{N} \Big\{ H(\mathbf{y}_{1:N/t}; \Psi)\, \mathbf{y}_t \int_0^1 \int_0^1 T_\epsilon(s_{\epsilon,t}) \mathbf{W}^* \mathbf{M}_t^{-1}\, m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \Big\} \\
\mathbf{A}_8(\mathbf{y}_{1:N}; \Psi)_{1\times k} &:= \sum_{t=1}^{N} \Big\{ H(\mathbf{y}_{1:N/t}; \Psi)\, \mathbf{y}_t \mathbf{W} \int_0^1 \int_0^1 T_\epsilon(s_{\epsilon,t}) \mathbf{M}_t^{-1}\, m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \Big\} \\
\mathbf{A}_9(\mathbf{y}_{1:N}; \Psi, \Psi^*)_{1\times d} &:= \sum_{t=1}^{N} \Big\{ H(\mathbf{y}_{1:N/t}; \Psi)\, \mathbf{y}_t \mathbf{W}^* \int_0^1 \int_0^1 T_\epsilon(s_{\epsilon,t})^2 \mathbf{M}_t^{-1}\, m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \Big\} \\
\mathbf{A}_{10}(\mathbf{y}_{1:N}; \Psi)_{1\times d} &:= \sum_{t=1}^{N} \Big\{ H(\mathbf{y}_{1:N/t}; \Psi)\, \mathbf{y}_t \mathbf{W} \int_0^1 \int_0^1 T_\epsilon(s_{\epsilon,t})^2 \mathbf{M}_t^{-1}\, m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \Big\} \\
\mathbf{A}_{11}(\mathbf{y}_{1:N}; \Psi)_{k\times k} &:= \sum_{t=1}^{N} \Big\{ H(\mathbf{y}^o_{1:N/t}; \Psi) \int_0^1 \int_0^1 T_\epsilon(s_{\epsilon,t}) \mathbf{M}_t^{-1}\, m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \Big\} \\
\mathbf{A}_{13}(\mathbf{y}^o_{1:N}; \Psi)_{k\times k} &:= \sum_{t=1}^{N} \Big\{ H(\mathbf{y}_{1:N/t}; \Psi) \int_0^1 \int_0^1 T_\epsilon(s_{\epsilon,t})^2 \mathbf{M}_t^{-1}\, m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \Big\} \\
\mathbf{A}_{14}(\mathbf{y}_{1:N}; \Psi)_{d\times k} &:= \sum_{t=1}^{N} \Big\{ H(\mathbf{y}_{1:N/t}; \Psi)\, \mathbf{y}_t^T \mathbf{y}_t \mathbf{W} \int_0^1 \int_0^1 T_\epsilon(s_{\epsilon,t})^2 \mathbf{M}_t^{-1}\, m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \Big\}
\end{aligned}
$$

$$\mathbf{A}_{16}(\mathbf{y}_{1:N};\Psi,\Psi^*)_{d\times k} := \sum_{t=1}^{N}\left\{H(\mathbf{y}_{1:N/t};\Psi)\ \mathbf{y}_t^T\int_0^1\int_0^1 T_\epsilon(s_{\epsilon,t})\Big(T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W}+\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}_t^{-1}\ m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\right\}$$

$$\mathbf{A}_{17}(\mathbf{y}_{1:N};\Psi)_{k\times k} := \sum_{t=1}^{N}\left\{H(\mathbf{y}_{1:N/t};\Psi)\ \int_0^1\int_0^1 T_\epsilon(s_{\epsilon,t})^3\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{y}_t^T\mathbf{y}_t\mathbf{W}\mathbf{M}_t^{-1}\ m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\right\}$$

$$\mathbf{A}_{18}(\mathbf{y}^o_{1:N};\Psi)_{k\times k} := \sum_{t=1}^{N}\left\{H(\mathbf{y}_{1:N/t};\Psi)\ \int_0^1\int_0^1 T_\epsilon(s_{\epsilon,t})^2\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{y}_t^T\Big(T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W}+\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}_t^{-1}m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\right\}$$

$$\mathbf{A}_{19}(\mathbf{y}_{1:N};\Psi)_{k\times k} := \sum_{t=1}^{N}\left\{H(\mathbf{y}_{1:N/t};\Psi)\ \int_0^1\int_0^1 T_\epsilon(s_{\epsilon,t})\mathbf{M}_t^{-1}\Big(T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W}+\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\Big)^T\Big(T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W}+\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}_t^{-1}\right.$$
$$\left.\times\ m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\right\}$$

$$A_{20}(\mathbf{y}_{1:N};\Psi)_{1\times 1} := \sum_{t=1}^{N}\left\{H(\mathbf{y}_{1:N/t};\Psi)\ \mathrm{Tr}\left\{\mathbf{y}_t^T\mathbf{y}_t\right\}\ \int_0^1\int_0^1 T_\epsilon(s_{\epsilon,t})\ m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\right\}$$

$$A_{21}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)_{1\times 1} := \sum_{t=1}^{N}\left\{H(\mathbf{y}_{1:N/t};\Psi)\ \int_0^1\int_0^1 T_\epsilon(s_{\epsilon,t})^2\,\mathrm{Tr}\left\{\mathbf{y}_t^T\mathbf{y}_t\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}\right\}\ m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\right\}.$$

Hence, the system of equations $\nabla_{\Psi^*}Q_{GStS,ind,2}(\Psi,\Psi^*)=\mathbf{0}$ is the following

$$\begin{cases}
\boldsymbol{\mu}^* = A_1(\mathbf{y}_{1:N};\Psi)^{-1}\Big[\mathbf{A}_6(\mathbf{y}_{1:N};\Psi)-A_0(\mathbf{y}_{1:N};\Psi)\boldsymbol{\delta}_\epsilon^*-\mathbf{A}_{10}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}+\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{13}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}+\big(\boldsymbol{\delta}_\epsilon\mathbf{W}+\sigma^2\boldsymbol{\delta}_x\big)\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\Big]\\
\boldsymbol{\delta}_\epsilon^* = A_2(\mathbf{y}_{1:N};\Psi)^{-1}\Big(\mathbf{A}_5(\mathbf{y}_{1:N};\Psi)-A_0(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^*-\mathbf{A}_8(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}+\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}+\Big(\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\Big)\mathbf{A}_4(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\Big)\\
\boldsymbol{\delta}_x^* = A_3(\mathbf{y}_{1:N};\Psi)^{-1}\Big(\mathbf{A}_8(\mathbf{y}_{1:N};\Psi)-\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi)-\big(\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\big)\mathbf{A}_4(\mathbf{y}_{1:N};\Psi)\Big)\\
\mathbf{W}^* = \Big[\mathbf{A}_{14}(\mathbf{y}_{1:N};\Psi)-\boldsymbol{\mu}^{*T}\mathbf{A}_{10}(\mathbf{y}_{1:N};\Psi)-\boldsymbol{\delta}_\epsilon^*\mathbf{A}_8(\mathbf{y}_{1:N};\Psi)-\mathbf{A}_{16}(\mathbf{y}_{1:N};\Psi)+\boldsymbol{\mu}^{*T}\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{13}(\mathbf{y}_{1:N};\Psi)+\boldsymbol{\mu}^{*T}\big(\boldsymbol{\delta}_\epsilon-\sigma^2\boldsymbol{\delta}_x\big)\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi)\\
\quad+\boldsymbol{\delta}_\epsilon^{*T}\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi)+\boldsymbol{\delta}_\epsilon^{*T}\big(\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\big)\mathbf{A}_4(\mathbf{y}_{1:N};\Psi)\Big]\\
\qquad\Big[\mathbf{A}_{17}(\mathbf{y}_{1:N};\Psi)-\mathbf{A}_{18}(\mathbf{y}_{1:N};\Psi)-\mathbf{A}_{18}(\mathbf{y}_{1:N};\Psi)^T+\mathbf{A}_{19}(\mathbf{y}_{1:N};\Psi)+\sigma^2\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi)\Big]^{-1}\\
\sigma^{*2} = \frac{1}{dA_0(\mathbf{y}_{1:N};\Psi)}\Big[A_{20}(\mathbf{y}_{1:N};\Psi)-2\mathbf{A}_6(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^{*T}-2\mathbf{A}_5(\mathbf{y}_{1:N};\Psi)\boldsymbol{\delta}_\epsilon^{*T}+A_1(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^*\boldsymbol{\mu}^{*T}+2A_0(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^*\boldsymbol{\delta}_\epsilon^{*T}\\
\quad+A_2(\mathbf{y}_{1:N};\Psi)\boldsymbol{\delta}_\epsilon^*\boldsymbol{\delta}_\epsilon^{*T}-2\,\mathrm{Tr}\left\{\mathbf{A}_{14}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\right\}+2\mathbf{A}_9(\mathbf{y}_{1:N};\Psi,\Psi^*)\mathbf{W}^T\boldsymbol{\mu}^T+2\mathbf{A}_7(\mathbf{y}_{1:N};\Psi,\Psi)\mathbf{W}^T\boldsymbol{\delta}_\epsilon^T\\
\quad-2\sigma^2\mathbf{A}_7(\mathbf{y}_{1:N};\Psi,\Psi)\boldsymbol{\delta}_x^T+2\Big(\mathbf{A}_{10}(\mathbf{y}_{1:N};\Psi)-\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{13}(\mathbf{y}_{1:N};\Psi)-\big(\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\big)\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi)\Big)\mathbf{W}^{*T}\boldsymbol{\mu}^{*T}\\
\quad+2\Big(\mathbf{A}_8(\mathbf{y}_{1:N};\Psi)-\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}^o_{1:N};\Psi)-\big(\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\big)\mathbf{A}_4(\mathbf{y}_{1:N};\Psi)\Big)\mathbf{W}^{*T}\boldsymbol{\delta}_\epsilon^{*T}+\mathrm{Tr}\left\{\mathbf{A}_{17}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\mathbf{W}^*\right\}\\
\quad-\mathrm{Tr}\left\{\Big(\mathbf{A}_{18}(\mathbf{y}_{1:N};\Psi)+\mathbf{A}_{18}(\mathbf{y}_{1:N};\Psi)^T\Big)\mathbf{W}^{*T}\mathbf{W}^*\right\}+\mathrm{Tr}\left\{\mathbf{A}_{19}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\mathbf{W}^*\right\}+\sigma^2\,\mathrm{Tr}\left\{\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\mathbf{W}^*\right\}\Big].
\end{cases}$$

□

A.4.2. IND: Incomplete Data Setting

A.4.2.1. Proof of Theorem 25

Proof. Let the observation vector $\mathbf{Y}_t$ is partitioned into two subvectors, $\mathbf{Y}_t=[\mathbf{Y}_t^o,\mathbf{Y}_t^m]$ with observed and unobserved entries of $\mathbf{Y}_t$, respectively. The joint probability function of the variables $\mathbf{Y}_t$, $\mathbf{X}_t,\mathbf{U}_t$ and $\mathbf{V}_t$ for Generalized Skew t-Student IND PPCA model is specified in Equation (3.29). Given the time independence of the random vectors $\mathbf{Y}_t$, $\mathbf{X}_t,\mathbf{U}_t$ and $\mathbf{V}_t$, the E-step of the corresponding EM algorithm can be written as

$$\begin{aligned}
\tilde{Q}_{GStS,ind}(\Psi,\Psi^*) &= \mathbb{E}_{\mathbf{Y}^m_{1:N},\mathbf{X}_{1:N},\mathbf{U}_{1:N},\mathbf{V}_{1:N}|\mathbf{Y}^o_{1:N},\Psi}\Big[\log\pi_{\mathbf{Y}_{1:N},\mathbf{X}_{1:N},\mathbf{U}_{1:N},\mathbf{V}_{1:N}|\Psi^*}(\mathbf{Y}_{1:N},\mathbf{X}_{1:N},\mathbf{U}_{1:N},\mathbf{V}_{1:N})\Big]\\
&= \frac{1}{\pi_{\mathbf{Y}^o_{1:N}|\Psi}(\mathbf{y}^o_{1:N})}\int_{\mathbb{R}_+^{N\times d}}\int_{\mathbb{R}_+^{N\times k}}\int_{\mathbb{R}^{N\times d_m}}\sum_{t=1}^{N}\Bigg\{\int_{\mathbb{R}^{N\times k}}\Big[\log\Big(\pi_{\mathbf{Y}_t,\mathbf{X}_t,\mathbf{U}_t,\mathbf{V}_t|\Psi^*}(\mathbf{y}_t,\mathbf{x}_t,\mathbf{u}_t,\mathbf{v}_t)\Big)\\
&\prod_{s=1}^{N}\pi_{\mathbf{Y}_s,\mathbf{X}_s,\mathbf{U}_s,\mathbf{V}_s|\Psi}(\mathbf{y}_s,\mathbf{x}_s,\mathbf{u}_s,\mathbf{v}_s)\Big]\,d\mathbf{x}_{1:N}\Bigg\}d\mathbf{y}^m_{1:N}\,d\mathbf{u}_{1:N}\,d\mathbf{v}_{1:N}.
\end{aligned}\tag{A7}$$

Using Lemma A8, we can decompose the probability function $\pi_{\mathbf{Y}_t,\mathbf{X}_t,\mathbf{U}_t,\mathbf{V}_t|\Psi}(\mathbf{y}_t,\mathbf{x}_t,\mathbf{u}_t,\mathbf{v}_t)$ into the product of three functions

$$\pi_{\mathbf{Y}_t,\mathbf{X}_t,\mathbf{U}_t,\mathbf{V}_t|\Psi}(\mathbf{y}_t,\mathbf{x}_t,\mathbf{u}_t,\mathbf{v}_t)=\pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{x}_t)\cdot\pi_{\mathbf{Y}_t|\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{y}_t)\cdot C(\mathbf{u}_t,\mathbf{v}_t;\Psi)$$

Please refer to the result derived in Lemma A8 to find the definitions of the probability functions and the function $C(\mathbf{u}_t, \mathbf{v}_t; \Psi) : \mathbb{R}_+^d \times \mathbb{R}_+^k \longrightarrow \mathbb{R}$. Therefore, we can further simplify the formulation of the Q function

$$\tilde{Q}_{GStS,ind}(\Psi, \Psi^*) = \frac{1}{\pi_{\mathbf{Y}^o_{1:N}|\Psi}(\mathbf{y}^o_{1:N})} \int_{\mathbb{R}_+^{N\times d}} \int_{\mathbb{R}_+^{N\times k}} \int_{\mathbb{R}^{N\times d_m}} \sum_{t=1}^{N} \left\{ \int_{\mathbb{R}^k} \log\left(\pi_{\mathbf{Y}_t,\mathbf{X}_t,\mathbf{U}_t,\mathbf{V}_t|\Psi^*}(\mathbf{y}_t, \mathbf{x}_t, \mathbf{u}_t, \mathbf{v}_t)\right) \pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{x}_t)\, d\mathbf{x}_t \right.$$
$$\left. \times \underbrace{\int_{\mathbb{R}^{(N-1)\times k}} \prod_{s=1, s\neq t}^{N} \pi_{\mathbf{X}_s|\mathbf{Y}_s,\mathbf{U}_s,\mathbf{V}_s,\Psi}(\mathbf{x}_s)\, d\mathbf{x}_{1:N}}_{\text{without t element}} \right\} \left(\prod_{s=1}^{N} \pi_{\mathbf{Y}_s|\mathbf{U}_s,\mathbf{V}_s,\Psi}(\mathbf{y}_s)\right) d\mathbf{y}^m_{1:N} \left(\prod_{s=1}^{N} C(\mathbf{u}_s, \mathbf{v}_s; \Psi)\right) d\mathbf{u}_{1:N}\, d\mathbf{v}_{1:N}$$
$$= \frac{1}{\pi_{\mathbf{Y}^o_{1:N}|\Psi}(\mathbf{y}^o_{1:N})} \int_{\mathbb{R}_+^{N\times d}} \int_{\mathbb{R}_+^{N\times k}} \int_{\mathbb{R}^{N\times d_m}} \left(\sum_{t=1}^{N} w(\mathbf{y}_t, \mathbf{u}_t, \mathbf{v}_t; \Psi, \Psi^*)\right) \left(\prod_{s=1}^{N} \pi_{\mathbf{Y}_s|\mathbf{U}_s,\mathbf{V}_s,\Psi}(\mathbf{y}_s)\right) d\mathbf{y}^m_{1:N} \left(\prod_{s=1}^{N} C(\mathbf{u}_t, \mathbf{v}_t; \Psi)\right) d\mathbf{u}_{1:N}\, d\mathbf{v}_{1:N}$$

for the function $w : \mathbb{R}^d \times \mathbb{R}_+^d \times \mathbb{R}_+^k \longrightarrow \mathbb{R}$ defined in Lemma A9. We can follow the similar steps to simplify the integration over the vector $\mathbf{y}^m_{1:N}$ using the fact that $\mathbf{y}_t$ are independent in time and obtain that

$$\tilde{Q}_{GStS,ind}(\Psi, \Psi^*) = \frac{1}{\pi_{\mathbf{Y}^o_{1:N}|\Psi}(\mathbf{y}^o_{1:N})} \int_{\mathbb{R}_+^{N\times d}} \int_{\mathbb{R}_+^{N\times k}} \left(\prod_{s=1}^{N} C(\mathbf{u}_t, \mathbf{v}_t; \Psi)\right) \sum_{t=1}^{N} \left\{ \int_{\mathbb{R}^{d_m}} w(\mathbf{y}_t, \mathbf{u}_t, \mathbf{v}_t; \Psi, \Psi^*) \pi_{\mathbf{Y}_t|\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{y}_t)\, d\mathbf{y}^m_t \right.$$
$$\left. \times \left(\prod_{s=1, s\neq t}^{N} \int_{\mathbb{R}^{d_m}} \pi_{\mathbf{Y}_s|\mathbf{U}_s,\mathbf{V}_s,\Psi}(\mathbf{y}_s)\, d\mathbf{y}^m_s\right) \right\} d\mathbf{u}_{1:N}\, d\mathbf{v}_{1:N}.$$

The next step is to specify the solutions to the integration problems

$$\int_{\mathbb{R}^{d_m}} \pi_{\mathbf{Y}_t|\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{y}_t)\, d\mathbf{y}^m_t \text{ and } \int_{\mathbb{R}^{d_m}} w(\mathbf{y}_t, \mathbf{u}_t, \mathbf{v}_t; \Psi, \Psi^*) \pi_{\mathbf{Y}_t|\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{y}_t)\, d\mathbf{y}^m_t$$

which rely on calculation of the conditional distribution $\pi_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{y}^m_t)$. Fortunately, the probability function $\pi_{\mathbf{Y}_t|\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{y}_t)$ is Gaussian and the conditional distributions of the Gaussian random vectors are broadly known. We derive the probability functions in Lemma A10 such that

$$\pi_{\mathbf{Y}_t|\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{y}_t) = \pi_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{y}^m_t) \cdot \pi_{\mathbf{Y}^o_t|\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{y}^o_t)$$

and use them to obtain the formulation of the Q function as following

$$\tilde{Q}_{GStS,ind}(\Psi, \Psi^*) = \frac{1}{\pi_{\mathbf{Y}^o_{1:N}|\Psi}(\mathbf{y}^o_{1:N})} \int_{\mathbb{R}_+^{N\times d}} \int_{\mathbb{R}_+^{N\times k}} \underbrace{\sum_{t=1}^{N} \left\{ \int_{\mathbb{R}^{d_m}} w(\mathbf{y}_t, \mathbf{u}_t, \mathbf{v}_t; \Psi, \Psi^*) \cdot \pi_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{y}^m_t)\, d\mathbf{y}^m_t \right\}}_{v(\mathbf{y}^o_{1:N}, \mathbf{u}_{1:N}, \mathbf{v}_{1:N}; \Psi, \Psi^*)}$$
$$\underbrace{\left(\prod_{s=1}^{N} \pi_{\mathbf{Y}^o_t|\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{y}^o_t) C(\mathbf{u}_t, \mathbf{v}_t; \Psi)\right)}_{h(\mathbf{y}^o_{1:N}, \mathbf{u}_{1:N}, \mathbf{v}_{1:N}; \Psi)} d\mathbf{u}_{1:N}\, d\mathbf{v}_{1:N}.$$

Given the solution to the integration problem $v(\mathbf{y}^o_{1:N}, \mathbf{u}_{1:N}, \mathbf{v}_{1:N}; \Psi, \Psi^*)$ from Lemma A11 and denoting for the convenience of the notation

$$h(\mathbf{y}^o_t, \mathbf{u}_t, \mathbf{v}_t; \Psi) = \pi_{\mathbf{Y}^o_t|\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{y}^o_t) \cdot C(\mathbf{u}_t, \mathbf{v}_t; \Psi) = \pi_{\mathbf{Y}^o_t|\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{y}^o_t) \cdot \pi_{\mathbf{U}_t|\Psi}(\mathbf{u}_t) \cdot \pi_{\mathbf{V}_t|\Psi}(\mathbf{v}_t)$$
$$\times (\sigma^2)^{\frac{k}{2}} \left|\mathbf{D}^{-1}_{x,t}\right|^{-\frac{1}{2}} \left|\mathbf{M}_t\right|^{-\frac{1}{2}} \left|\mathbf{N}_t^{-1}\right|^{\frac{1}{2}} \exp\left\{ -\frac{1}{2}\delta_x \left(\mathbf{D}^{-1}_{x,t} - \sigma^2 \mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{N}_t^{-1}\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1} - \sigma^2\mathbf{M}_t^{-1}\right)\delta_x^T \right\}$$

the E-step of the Generalized Skew t-Student IND PPCA is the following

$$\tilde{Q}_{GStS,ind}(\Psi, \Psi^*) = \frac{1}{\pi_{\mathbf{Y}^o_{1:N}|\Psi}(\mathbf{y}^o_{1:N})} \sum_{t=1}^{N} \left\{ \int_{\mathbb{R}_+^d} \int_{\mathbb{R}_+^k} v(\mathbf{y}^o_t, \mathbf{u}_t, \mathbf{v}_t; \Psi, \Psi^*) h(\mathbf{y}^o_t, \mathbf{u}_t, \mathbf{v}_t; \Psi)\, d\mathbf{u}_t\, d\mathbf{v}_t \prod_{s=1, s\neq t}^{N} \int_{\mathbb{R}_+^d} \int_{\mathbb{R}_+^k} h(\mathbf{y}^o_s, \mathbf{u}_s, \mathbf{v}_s; \Psi)\, d\mathbf{u}_s\, d\mathbf{v}_s \right\}. \tag{A8}$$

Let us recall the definition of the mutually independent multivariate Gamma random vectors $\mathbf{U}_t$ and $\mathbf{V}_t$ from Equation (3.27) in Subsection 3.4.2.1 as transformations of the mutually independent univariate uniform random variables $S_{\epsilon,t},\ S_{x,t} \in \mathcal{U}(0,1)$ by the quantile functions $T_\epsilon : [0,1] \to \mathbb{R}_+^d$ and $T_x : [0,1] \to \mathbb{R}_+^k$ such that

$$T_\epsilon(s) := \left(\frac{\chi^{-1}_{\nu^1_\epsilon}(s)}{\nu^1_\epsilon}, \ldots, \frac{\chi^{-1}_{\nu^d_\epsilon}(s)}{\nu^d_\epsilon}\right)_{1\times d} \text{ and } T_x(s) := \left(\frac{\chi^{-1}_{\nu^1_x}(s)}{\nu^1_x}, \ldots, \frac{\chi^{-1}_{\nu^k_x}(s)}{\nu^k_x}\right)_{1\times k} \tag{A9}$$

where χ_ν^{-1} is the quantile distribution function of a univariate chi-square random variable with ν degrees of freedom. By the change of variables $S_{\epsilon,t} = T_\epsilon^{-1}(\mathbf{U}_t)$ and $S_{x,t} = T_x^{-1}(\mathbf{V}_t)$, we can reduce the $d \times k$ dimensional integration problems from Equation (A1) to the 2-dimensional integrations on the unit hypercube $[0,1]^2$.

Firstly, we remark that the corresponding density functions $\pi_{\mathbf{U}_t|\Psi}(\mathbf{u}_t)$ and $\pi_{\mathbf{V}_t|\Psi}(\mathbf{v}_t)$ can be expressed by the density function of the random variables $S_{\epsilon,t}$ and $S_{x,t}$ defined in Subsection 3.4.2.1, that is

$$\begin{aligned}\pi_{\mathbf{U}_t|\Psi}(\mathbf{u}_t) &= \pi_{S_{\epsilon,t}|\Psi}(s_{\epsilon,t})\left|\frac{\partial T_\epsilon(s_{\epsilon,t})}{\partial s_{\epsilon,t}}\right|^{-1} = \mathbf{1}_{[0,1]}(s_{\epsilon,t})\left|\frac{\partial T_\epsilon(s_{\epsilon,t})}{\partial s_{\epsilon,t}}\right|^{-1}\\ \pi_{\mathbf{V}_t|\Psi}(\mathbf{v}_t) &= \pi_{S_{x,t}|\Psi}(s_{x,t})\left|\frac{\partial T_x(s_{x,t})}{\partial s_{x,t}}\right|^{-1} = \mathbf{1}_{[0,1]}(s_{x,t})\left|\frac{\partial T_x(s_{x,t})}{\partial s_{x,t}}\right|^{-1}\end{aligned} \tag{A10}$$

The distribution of $\mathbf{Y}_t^o$ conditioned on $\mathbf{U}_t$ and $\mathbf{V}_t$ can also by expressed by $S_{\epsilon,t}$ and $S_{x,t}$ as

$$\pi_{\mathbf{Y}_t^o|T_\epsilon(S_{\epsilon,t}),T_x(S_{x,t}),\Psi}(\mathbf{y}_t^o) = \pi_{\mathbf{Y}_t^o|S_{\epsilon,t},S_{x,t},\Psi}(\mathbf{y}_t^o).$$

Recall that the definitions of the matrices $\mathbf{D}_{\epsilon,t}$ and $\mathbf{D}_{x,t}$ under the transformation $\mathbf{U}_t = T_\epsilon(S_{\epsilon,t})$ and $\mathbf{V}_t = T_x(S_{x,t})$ become

$$\mathbf{D}_{\epsilon,t} = \begin{pmatrix} \frac{\chi^{-1}_{\nu_\epsilon^1}(S_{\epsilon,t})}{\nu_\epsilon^1} & 0 & 0\\ 0 & \ddots & 0\\ 0 & 0 & \frac{\chi^{-1}_{\nu_\epsilon^d}(S_\epsilon)}{\nu_\epsilon^d}\end{pmatrix}_{d\times d}, \mathbf{D}_{x,t} = \begin{pmatrix} \frac{\chi^{-1}_{\nu_x^1}(S_{x,t})}{\nu_x^1} & 0 & 0\\ 0 & \ddots & 0\\ 0 & 0 & \frac{\chi^{-1}_{\nu_x^k}(S_{x,t})}{\nu_x^k}\end{pmatrix}_{k\times k}.$$

Hence, the E-step of the EM algorithm for the Independent Generalized Skew t-Student PPCA from Equation (A8) can be expressed by the sum of 2-dimensional integration problems

$$\tilde{Q}_{GStS,ind}(\Psi,\Psi^*) = \frac{1}{\pi_{\mathbf{Y}_{1:N}^o|\Psi}(\mathbf{y}_{1:N}^o)}\sum_{t=1}^N\left\{I_1(\mathbf{y}_t^o;\Psi,\Psi^*)\prod_{s=1,s\neq t}^N I_2(\mathbf{y}_s^o;\Psi)\right\}$$

for the functions $I_1,\ I_2 : \mathbb{R}^{d_o} \to \mathbb{R}$

$$\begin{aligned}I_1(\mathbf{y}_t^o;\Psi,\Psi^*) &:= \int_0^1\int_0^1 v(\mathbf{y}_t^o, T_\epsilon(s_{\epsilon,t}), T_x(s_{x,t});\Psi,\Psi^*)\, m(\mathbf{y}_t^o, s_{\epsilon,t}, s_{x,t};\Psi)\, ds_{\epsilon,t}\, ds_{x,t}\\ I_2(\mathbf{y}_t^o;\Psi) &:= \int_0^1\int_0^1 m(\mathbf{y}_t^o, s_{\epsilon,t}, s_{x,t};\Psi)\, ds_{\epsilon,t}\, ds_{x,t}\end{aligned}$$

and the function $m : \mathbb{R}^{d_o}\times[0,1]^2 \longrightarrow \mathbb{R}$

$$m(\mathbf{y}_t^o, s_{\epsilon,t}, s_{x,t};\Psi) = \pi_{\mathbf{Y}_t^o|S_{\epsilon,t},S_{x,t},\Psi}(\mathbf{y}_t^o)(\sigma^2)^{\frac{k}{d}}\left|\mathbf{D}_{x,t}^{-1}\right|^{-\frac{1}{2}}\left|\mathbf{M}_t\right|^{-\frac{1}{2}}\left|\mathbf{N}_t\right|^{-\frac{1}{2}} e^{-\frac{1}{2}\delta_x\left(\mathbf{D}_{x,t}^{-1}-\sigma^2\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{N}_t^{-1}\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}-\sigma^2\mathbf{M}_t^{-1}\right)\delta_x^T}.$$

□

A.4.2.2. Proof of Corollary 18

Proof. Let us remark that in under the assumptions of Special Case 2 from Subsection 3.4.1.3, the random vectors $\mathbf{U}_t$ and $\mathbf{V}_t$ reduce to the scalar random variables U_t and V_t. Following the steps from the proof of Theorem 13, the E-step of the EM algorithm for Special Case 2 Skew t-Student IND PPCA is the following

$$\begin{aligned}\tilde{Q}_{GStS,ind,2}(\Psi,\Psi^*) = \frac{1}{\pi_{\mathbf{Y}_{1:N}^o|\Psi}(\mathbf{y}_{1:N}^o)}\int_{\mathbb{R}_+^N}\int_{\mathbb{R}_+^N}\int_{\mathbb{R}^{N\times d_m}}\sum_{t=1}^N\Bigg\{\int_{\mathbb{R}^{N\times k}}\Bigg[\log\left(\pi_{\mathbf{Y}_t,\mathbf{X}_t,U_t,V_t|\Psi^*}(\mathbf{y}_t,\mathbf{x}_t,u_t,v_t)\right)\\ \times\prod_{s=1}^N\pi_{\mathbf{Y}_s,\mathbf{X}_s,U_s,V_s|\Psi}(\mathbf{y}_s,\mathbf{x}_s,u_s,v_s)\Bigg]d\mathbf{x}_{1:N}\Bigg\}d\mathbf{y}_{1:N}^m\, du_{1:N}\, dv_{1:N}.\end{aligned}$$

Let us remark, that the diagonal matrices $\mathbf{D}_{\epsilon,t}$ and $\mathbf{D}_{x,t}$ from Subsection 3.4.2.1 reduce to the product of scalars and identity matrices, that is

$$\mathbf{D}_{\epsilon,t} = U_t\mathbb{I}_d \text{ and } \mathbf{D}_{x,t} = V_t\mathbb{I}_k \tag{A11}$$

Analogously to the Independent Generalized Skew t-Student PPCA case, we decompose the probability function $\pi_{\mathbf{Y}_t,\mathbf{X}_t,U_t,V_t|\Psi}(\mathbf{y}_t,\mathbf{x}_t,u_t,v_t)$ by subsection the new formulations of the matrices $\mathbf{D}_{\epsilon,t}$ and $\mathbf{D}_{x,t}$ into Lemma A8 and straightforwardly obtain that

$$\pi_{\mathbf{Y}_t,\mathbf{X}_t,U_t,V_t|\Psi}(\mathbf{y}_t,\mathbf{x}_t,u_t,v_t) = \pi_{\mathbf{X}_t|\mathbf{Y}_t,U_t,V_t,\Psi}(\mathbf{x}_t)\cdot\pi_{\mathbf{Y}_t|U_t,V_t,\Psi}(\mathbf{y}_t)\cdot C(u_t,v_t;\Psi)$$

where $\pi_{\mathbf{X}_t|\mathbf{Y}_t,U_t,V_t,\Psi}(\mathbf{x}_t)$ is a probability density function of the random vector $\mathbf{X}_t|\mathbf{Y}_t,U_t,V_t,\Psi \sim \mathcal{N}\left(\boldsymbol{\mu}_{x,t},\boldsymbol{\Sigma}_{x,t}\right)$ and $\pi_{\mathbf{Y}_t|U_t,V_t,\Psi}(\mathbf{x}_t)$ is a probability density function of the random vector $\mathbf{Y}_t|U_t,V_t,\Psi \sim \mathcal{N}\left(\boldsymbol{\mu}_{y,t},\boldsymbol{\Sigma}_{y,t}\right)$ for the corresponding moments specified as follows

$$\begin{aligned}\boldsymbol{\mu}_{x,t} &= \left(U_t(\mathbf{y}_t-\boldsymbol{\mu}-U_t^{-1}\delta_\epsilon)\mathbf{W}+\sigma^2\delta_x\right)\mathbf{M}_t^{-1} \text{ and } \boldsymbol{\Sigma}_{x,t}=\sigma^2\mathbf{M}_t^{-1},\\ \boldsymbol{\mu}_{y,t} &= \boldsymbol{\mu}+U_t^{-1}\delta_\epsilon+\sigma^2\delta_x\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{N}_t^{-1} \text{ and } \boldsymbol{\Sigma}_{y,t}=\sigma^2U_t^{-1}\mathbf{N}_t^{-1}\end{aligned} \tag{A12}$$

for $\mathbf{M}_t=\sigma^2V_t\mathbb{I}_k+U_t\mathbf{W}^T\mathbf{W}$ and $\mathbf{N}_t=\mathbb{I}_d-U_t\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^T$. The function $C:\mathbb{R}_+^2\longrightarrow\mathbb{R}$ is given by

$$C(u_t,v_t;\Psi)=\pi_{U_t|\Psi}(u_t)\cdot\pi_{V_t|\Psi}(v_t)\left(\sigma^2\right)^{\frac{k}{2}}V_t^{\frac{k}{2}}\left|\mathbf{M}_t\right|^{-\frac{1}{2}}\left|\mathbf{N}_t\right|^{-\frac{1}{2}}e^{-\frac{1}{2}\delta_x\left(V_t^{-1}\mathbb{I}_k-\sigma^2U_t\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{N}_t^{-1}\mathbf{W}\mathbf{M}_t^{-1}-\sigma^2\mathbf{M}_t^{-1}\right)\delta_x^T}.$$

The further simplification of the function Q_2 lead to its following formulation

$$\begin{aligned}\tilde{Q}_{GStS,ind,2}(\Psi,\Psi^*) = \frac{1}{\pi_{\mathbf{Y}_{1:N}^o|\Psi}(\mathbf{y}_{1:N}^o)}\sum_{t=1}^{N}\Bigg\{&\prod_{s=1,s\neq t}^{N}\int_{\mathbb{R}_+}\int_{\mathbb{R}_+}\underbrace{\int_{\mathbb{R}^{d_m}}\pi_{\mathbf{Y}_s|U_s,V_s,\Psi}(\mathbf{y}_s)\,d\mathbf{y}_s^m}_{\pi_{\mathbf{Y}_s^o|U_s,V_s,\Psi}(\mathbf{y}_s^o)}\,C(u_t,v_t;\Psi)\,du_s\,dv_s\\ \times\int_{\mathbb{R}_+}\int_{\mathbb{R}_+}\int_{\mathbb{R}^{d_m}}&\underbrace{\int_{\mathbb{R}^k}\log\left(\pi_{\mathbf{Y}_t,\mathbf{X}_t,U_t,V_t|\Psi^*}(\mathbf{y}_t,\mathbf{x}_t,u_t,v_t)\right)\pi_{\mathbf{X}_t|\mathbf{Y}_t,U_t,V_t,\Psi}(\mathbf{x}_t)\,d\mathbf{x}_t}_{w(\mathbf{y}_t,u_t,v_t;\Psi,\Psi^*)}\,\pi_{\mathbf{Y}_t|U_t,V_t,\Psi}(\mathbf{y}_t)\,d\mathbf{y}_t^m\,C(u_t,v_t;\Psi)\,du_t\,dv_t\Bigg\}.\end{aligned}$$

Analogously to the results summarized in Lemma A10, $\pi_{\mathbf{Y}_t^o|U_t,V_t,\Psi}(\mathbf{y}_t^o)$ is a probability density function of the d_o-dimensional random vector $\mathbf{Y}_t^o|U_t,V_t,\Psi\sim\mathcal{N}(\boldsymbol{\mu}_{y,t}^o,\boldsymbol{\Sigma}_{y,t}^{oo})$, where $\boldsymbol{\mu}_{y,t}^o$ contains elements of the vector $\boldsymbol{\mu}_{y,t}$ which correspond to the observed entries of the vector $\mathbf{Y}_t$ and $\boldsymbol{\Sigma}_{y,t}^{oo}$ is a submatrix of $\boldsymbol{\Sigma}_{y,t}$ which correspond by rows and columns to observed elements of $\mathbf{Y}_t$. The function $w(\mathbf{y}_t^o,u_t,v_t;\Psi,\Psi^*)$ is derived by using Lemma A9 and substituting the new formulations of the diagonal matrices $\mathbf{D}_{\epsilon,t}$ and $\mathbf{D}_{x,t}$ from Equation A11, that is

$$\begin{aligned}w(\mathbf{y}_t,u_t,v_t;\Psi,\Psi^*) &= \log\pi_{U_t|\Psi^*}(u_t)+\log\pi_{V_t|\Psi^*}(v_t)-\frac{k+d}{2}\log 2\pi-\frac{d}{2}\log\sigma^{*2}+\frac{d}{2}\log u_t+\frac{k}{2}\log v_t\\ &-\frac{v_t^{-1}}{2}\delta_x^*\delta_x^{*T}-\frac{1}{2\sigma^{*2}}u_t\mathbf{y}_t\mathbf{y}_t^T+\frac{1}{\sigma^{*2}}u_t\mathbf{y}_t\boldsymbol{\mu}^{*T}+\frac{1}{\sigma^{*2}}\mathbf{y}_t\delta_\epsilon^{*T}-\frac{1}{2\sigma^{*2}}u_t\boldsymbol{\mu}^*\boldsymbol{\mu}^{*T}-\frac{1}{\sigma^{*2}}\boldsymbol{\mu}^*\delta_\epsilon^{*T}-\frac{1}{2\sigma^{*2}}u_t^{-1}\delta_\epsilon^*\delta_\epsilon^{*T}\\ &+\left(u_t\mathbf{y}_t\mathbf{W}-u_t\boldsymbol{\mu}\mathbf{W}-\delta_\epsilon\mathbf{W}+\sigma^2\delta_x\right)\mathbf{M}_t^{-1}\left(\frac{1}{\sigma^{*2}}u_t\mathbf{W}^{*T}\mathbf{y}_t^T-\frac{1}{\sigma^{*2}}u_t\mathbf{W}^{*T}\boldsymbol{\mu}^{*T}-\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\delta_\epsilon^{*T}+\delta_x^{*T}\right)\\ &-\frac{1}{2}\operatorname{Tr}\left\{\sigma^2\mathbf{M}_t^{-1}\left(\frac{1}{\sigma^{*2}}u_t\mathbf{W}^{*T}\mathbf{W}^*+v_t\mathbb{I}_k\right)+\boldsymbol{\mu}_{x,t}^T\boldsymbol{\mu}_{x,t}\left(\frac{1}{\sigma^{*2}}u_t\mathbf{W}^{*T}\mathbf{W}^*+v_t\mathbb{I}_k\right)\right\}\end{aligned}$$

for $\mathbf{M}_t$, $\mathbf{N}_t$ and $\boldsymbol{\mu}_{x,t}$ specified as in Equation (A12). The next step is to specify a function $v:\mathbb{R}^{d_o}\times[0,1]^2\longrightarrow\mathbb{R}$ such that

$$\int_{\mathbb{R}^{d_m}}w(\mathbf{y}_t,u_t,v_t;\Psi,\Psi^*)\,\pi_{\mathbf{Y}_t|U_t,V_t,\Psi}(\mathbf{y}_t)\,d\mathbf{y}_t^m=\pi_{\mathbf{Y}_t^o|U_t,V_t,\Psi}(\mathbf{y}_t^o)\,v(\mathbf{y}_t^o,u_t,v_t;\Psi,\Psi^*).$$

Again, we can use the modified Lemma A11 from the proof of Theorem 13 in order to specify the function. Applying the assumptions of the Skew t-Student PPCA we obtain that

$$\begin{aligned}v(\mathbf{y}_t^o,u_t,v_t;\Psi,\Psi^*) &= \log\pi_{U_t|\Psi^*}(u_t)+\log\pi_{\mathbf{V}_t|\Psi^*}(\mathbf{v}_t)-\frac{k+d}{2}\log 2\pi-\frac{d}{2}\log\sigma^{*2}+\frac{d}{2}\log u_t+\frac{k}{2}\log v_t-\frac{v_t^{-1}}{2}\delta_x^*\delta_x^{*T}\\ &-\frac{u_t}{2\sigma^{*2}}\operatorname{Tr}\left\{\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U_t,V_t,\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right]\right\}+\frac{1}{\sigma^{*2}}\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U_t,V_t,\Psi}\left[\mathbf{Y}_t\right]\left(u_t\boldsymbol{\mu}^*+\delta_\epsilon^*\right)^T-\frac{u_t}{2\sigma^{*2}}\boldsymbol{\mu}^*\boldsymbol{\mu}^{*T}-\frac{1}{\sigma^{*2}}\boldsymbol{\mu}^*\delta_\epsilon^{*T}-\frac{u_t^{-1}}{2\sigma^{*2}}\delta_\epsilon^*\delta_\epsilon^{*T}\\ &+\frac{u_t^2}{\sigma^{*2}}\operatorname{Tr}\left\{\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U_t,V_t,\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right]\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}\right\}+u_t\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U_t,V_t,\Psi}\left[\mathbf{Y}_t\right]\mathbf{W}\mathbf{M}_t^{-1}\left(-\frac{u_t}{\sigma^{*2}}\mathbf{W}^{*T}\boldsymbol{\mu}^{*T}-\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\delta_\epsilon^{*T}+\delta_x^{*T}\right)\\ &+\left(\frac{u_t}{\sigma^{*2}}\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U_t,V_t,\Psi}\left[\mathbf{Y}_t\right]\mathbf{W}^*-\frac{u_t}{\sigma^{*2}}\boldsymbol{\mu}^*\mathbf{W}^*-\frac{1}{\sigma^{*2}}\delta_\epsilon^*\mathbf{W}^*+\delta_x^*\right)\mathbf{M}_t^{-1}\left(-u_t\boldsymbol{\mu}\mathbf{W}-\delta_\epsilon\mathbf{W}+\sigma^2\delta_x\right)^T\\ &-\frac{1}{2}\operatorname{Tr}\left\{\mathbf{M}_t^{-1}\left(u_t^2\mathbf{W}^T\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U_t,V_t,\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right]\mathbf{W}\mathbf{M}_t^{-1}\left(\frac{u_t}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{W}^*+v_t\mathbb{I}_k\right)\right)\right\}\\ &+u_t\operatorname{Tr}\left\{\mathbf{M}_t^{-1}\mathbf{W}^T\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U_t,V_t,\Psi}\left[\mathbf{Y}_t\right]\left(-u_t\boldsymbol{\mu}\mathbf{W}-\delta_\epsilon\mathbf{W}+\sigma^2\delta_x\right)\mathbf{M}_t^{-1}\left(\frac{u_t}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{W}^*+v_t\mathbb{I}_k\right)\right\}\\ &-\frac{1}{2}\operatorname{Tr}\left\{\mathbf{M}_t^{-1}\left(-u_t\boldsymbol{\mu}\mathbf{W}-\delta_\epsilon\mathbf{W}+\sigma^2\delta_x\right)^T\left(-u_t\boldsymbol{\mu}\mathbf{W}-\delta_\epsilon\mathbf{W}+\sigma^2\delta_x\right)\mathbf{M}_t^{-1}\left(\frac{u_t}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{W}^*+v_t\mathbb{I}_k\right)\right\}\\ &-\frac{1}{2}\operatorname{Tr}\left\{\sigma^2\mathbf{M}_t^{-1}\left(\frac{u_t}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{W}^*+v_t\mathbb{I}_k\right)\right\}.\end{aligned}$$

The non-central moments

$$\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U_t,V_t,\Psi}\left[\mathbf{Y}_t\right] = \left[\mathbf{y}_t^o,\ \tilde{\boldsymbol{\mu}}_{Y^m,t}\right],$$
$$\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U_t,V_t,\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right] = \begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \tilde{\boldsymbol{\Sigma}}_{Y^m,t}\end{bmatrix} + \mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U_t,V_t,\Psi}\left[\mathbf{Y}_t\right]^T\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,U_t,\mathbf{V}_t,\Psi}\left[\mathbf{Y}_t\right],$$

are determined using the reasoning from Lemma A10 and the modification from Equation (A12) which result in the distribution of the random vector $\mathbf{Y}_t^m|\mathbf{Y}_t^o,U_t,V_t,\Psi \sim \mathcal{N}\left(\tilde{\boldsymbol{\mu}}_{Y^m,t},\tilde{\boldsymbol{\Sigma}}_{Y^m,t}\right)$ specified by

$$\tilde{\boldsymbol{\mu}}_{Y^m,t} = \boldsymbol{\mu}_{y,t}^m + (\mathbf{Y}_t^o - \boldsymbol{\mu}_{y,t}^m)\boldsymbol{\Sigma}_{y,t}^{oo\ -1}\boldsymbol{\Sigma}_{y,t}^{om}$$
$$\tilde{\boldsymbol{\Sigma}}_{Y^m,t} = \boldsymbol{\Sigma}_{y,t}^{mm} - \boldsymbol{\Sigma}_{y,t}^{mo}\boldsymbol{\Sigma}_{y,t}^{oo\ -1}\boldsymbol{\Sigma}_{y,t}^{om}$$

where $\boldsymbol{\mu}_{y,t}^m$ contains elements of the vector $\boldsymbol{\mu}_{y,t}$ which correspond to the non-observed entries of the vector $\mathbf{Y}_t$ and $\boldsymbol{\Sigma}_{y,t}^{mm}$ is a square submatrix of $\boldsymbol{\Sigma}_{y,t}$ which correspond by rows and columns to missing elements of $\mathbf{Y}_t$. The elements of the $d_m \times d_o$ submatrix $\boldsymbol{\Sigma}_{y,t}^{mo} = \boldsymbol{\Sigma}_{y,t}^{om\ T}$ correspond by rows to missing entries of the vector $\mathbf{Y}_t$ and by columns to its observed elements. Therefore, the E-step of the EM algorithm for the Independent Skew t-Student PPCA described in Subsection 3.4.1.3 can be represented by the sum of two dimensional integration problems over $\mathbb{R}_+^2$, that is

$$\tilde{Q}_{GStS,ind,2}(\Psi,\Psi^*) = \frac{1}{\pi_{\mathbf{Y}_{1:N}^o|\Psi}(\mathbf{y}_{1:N}^o)}\sum_{t=1}^{N}\left\{\prod_{s=1,s\neq t}^{N}\int_{\mathbb{R}_+}\int_{\mathbb{R}_+}\pi_{\mathbf{Y}_s^o|U_s,V_s,\Psi}(\mathbf{y}_s^o)\ C(u_t,v_t;\Psi)\ du_s\ dv_s\right.$$
$$\left.\times\int_{\mathbb{R}_+}\int_{\mathbb{R}_+} v(\mathbf{y}_t^o,u_t,v_t;\Psi,\Psi^*)\ \pi_{\mathbf{Y}_t^o|U_t,V_t,\Psi}(\mathbf{y}_t^o)\ C(u_t,v_t;\Psi)\ du_t\ dv_t\right\}. \tag{A13}$$

In order to obtain an efficient quadrature rule to approximate the solution of $Q_2(\Psi,\Psi^*)$ we recall the quantile transformation from Equation (3.27). Under the assumption of Independent Skew t-Student PPCA, the random variable U_t and V_t are as quantile transformations of the mutually independent univariate uniform random variables $S_{\epsilon,t},\ S_{x,t} \in \mathcal{U}(0,1)$ by functions $T_\epsilon : [0,1] \to \mathbb{R}_+$ and $T_x : [0,1] \to \mathbb{R}_+$ such that

$$T_\epsilon(s) := \frac{\chi_{\nu_\epsilon}^{-1}(s)}{\nu_\epsilon} \text{ and } T_x(s) := \frac{\chi_{\nu_x}^{-1}(s)}{\nu_x}$$

where χ_ν^{-1} is a quantile function of univariate chi-square distribution function with ν degrees of freedom. By the change of variables $S_{\epsilon,t} = T_\epsilon^{-1}(U_t)$ and $S_{x,t} = T_x^{-1}(V_t)$, we can simplify the integration problems from Equation (A13) to the 2-dimensional integrations on $[0,1]^2$ given by

$$\tilde{Q}_{GStS,ind,2}(\Psi,\Psi^*) = \frac{1}{\pi_{\mathbf{Y}_{1:N}^o|\Psi}(\mathbf{y}_{1:N}^o)}\sum_{t=1}^{N}\left\{\int_0^1\int_0^1 v(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi,\Psi^*)\ m(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\right.$$
$$\left.\times\prod_{s=1,s\neq t}^{N}\int_0^1\int_0^1 m(\mathbf{y}_s^o,s_{\epsilon,s},s_{x,s};\Psi)\ ds_{\epsilon,s}\ ds_{x,s}\right\}$$

for

$$m(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi) = \pi_{\mathbf{Y}_t^o|S_{\epsilon,t},S_{x,t},\Psi}(\mathbf{y}_t^o)(\sigma^2)^{\frac{k}{2}}v_t^{\frac{k}{2}}\left|\mathbf{M}_t\right|^{-\frac{1}{2}}\left|\mathbf{N}_t\right|^{-\frac{1}{2}}e^{-\frac{1}{2}\delta_x\left(v_t^{-1}\mathbb{I}_k-\sigma^2u_t\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{N}_t^{-1}\mathbf{W}\mathbf{M}_t^{-1}-\sigma^2\mathbf{M}_t^{-1}\right)\delta_x^T}.$$

and the probability function $\pi_{\mathbf{Y}_t^o|S_{\epsilon,t},S_{x,t},\Psi}(\mathbf{y}_t^o) := \pi_{\mathbf{Y}_t^o|T_\epsilon(S_{\epsilon,t}),T_x(S_{x,t}),\Psi}(\mathbf{y}_t^o)$ and redefine function v such that $v(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi,\Psi^*) := v(\mathbf{y}_t^o,T_\epsilon(s_{\epsilon,t}),T_x(s_{x,t});\Psi,\Psi^*)$. □

A.4.2.3. Proof of Theorem 26

Proof. The M-step of the EM algorithm evaluates the maximizers of the function $\tilde{Q}_{GStS,ind}(\Psi,\Psi^*)$ defined in Theorem 25 with respect to the vector of static parameters $\Psi^* = \left[\boldsymbol{\mu}^*,\mathbf{W}^*,\sigma^{*2},\delta_\epsilon^*,\boldsymbol{\delta}_x\right]$. Recall that the only element of the function $\tilde{Q}_{GStS,ind}$ which contains components of the vector Ψ^* is the function $\tilde{v}$ being a component of the integral $I_1(\mathbf{y}_t^o;\Psi,\Psi^*)$. Hence, the maximizers of the function $\tilde{Q}_{GStS,ind}$ with respect to the vector of static parameters Ψ^*, which are the solution to the set of equations $\nabla_{\Psi^*}\tilde{Q}_{GStS,ind}(\Psi,\Psi^*) = \mathbf{0}$, are calculated as follows

$$\frac{\partial}{\partial\Psi^*}\tilde{Q}_{GStS,ind}(\Psi,\Psi^*) = \frac{\partial}{\partial\Psi^*}\frac{1}{\pi_{\mathbf{Y}_{1:N}^o|\Psi}(\mathbf{y}_{1:N}^o)}\sum_{t=1}^{N}\left\{H(\mathbf{y}_{1:N/t}^o;\Psi)I_1(\mathbf{y}_t^o;\Psi,\Psi^*)\right\} = \mathbf{0} \iff \sum_{t=1}^{N}\left\{H(\mathbf{y}_{1:N/t}^o;\Psi)\frac{\partial}{\partial\Psi^*}I_1(\mathbf{y}_t^o;\Psi,\Psi^*)\right\} = \mathbf{0}$$

$$\iff \sum_{t=1}^{N} \left\{ H(\mathbf{y}^o_{1:N/t}; \Psi) \int_0^1 \int_0^1 \frac{\partial}{\partial \Psi^*} \tilde{v}(\mathbf{y}^o_t, s_{\epsilon,t}, s_{x,t}; \Psi, \Psi^*)\; m(\mathbf{y}^o_t, s_{\epsilon,t}, s_{x,t}; \Psi)\; ds_{\epsilon,t}\; ds_{x,t} \right\} = \mathbf{0}. \tag{A14}$$

We applied Leibniz integral rule to swap the order of differentiating and integration. The gradient $\frac{\partial}{\partial \Psi^*} \tilde{v}(\mathbf{y}^o_t, s_{\epsilon,t}, s_{x,t}; \Psi, \Psi^*)$ is derived as following

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\mu}^*} \tilde{v}(\mathbf{y}^o_t, s_{\epsilon,t}, s_{x,t}; \Psi, \Psi^*) &= \frac{1}{\sigma^{*2}} \mathbb{E}_{\mathbf{Y}^m_t | \mathbf{Y}^o_t, S_{\epsilon,t}, S_{x,t}, \Psi}[\mathbf{Y}_t] \mathbf{D}_{\epsilon,t} - \frac{1}{\sigma^{*2}} \boldsymbol{\mu}^* \mathbf{D}_{\epsilon,t} - \frac{1}{\sigma^{*2}} \boldsymbol{\delta}^*_\epsilon - \frac{1}{\sigma^{*2}} \mathbb{E}_{\mathbf{Y}^m_t | \mathbf{Y}^o_t, S_{\epsilon,t}, S_{x,t}, \Psi}[\mathbf{Y}_t] \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1} \mathbf{W}^{*T} \mathbf{D}_{\epsilon,t} \\
&+ \frac{1}{\sigma^{*2}} \boldsymbol{\mu} \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1} \mathbf{W}^{*T} \mathbf{D}_{\epsilon,t} + \frac{1}{\sigma^{*2}} \boldsymbol{\delta}_\epsilon \mathbf{W} \mathbf{M}_t^{-1} \mathbf{W}^{*T} \mathbf{D}_{\epsilon,t} - \frac{\sigma^2}{\sigma^{*2}} \boldsymbol{\delta}_x \mathbf{M}_t^{-1} \mathbf{W}^{*T} \mathbf{D}_{\epsilon,t}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{W}^*} \tilde{v}(\mathbf{y}^o_t, s_{\epsilon,t}, s_{x,t}; \Psi, \Psi^*) &= \frac{1}{\sigma^{*2}} \mathbf{D}_{\epsilon,t} \mathbb{E}_{\mathbf{Y}^m_t | \mathbf{Y}^o_t, S_{\epsilon,t}, S_{x,t}, \Psi}[\mathbf{Y}_t^T \mathbf{Y}_t] \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1} - \frac{1}{\sigma^{*2}} \left(\mathbf{D}_{\epsilon,t} \boldsymbol{\mu}^{*T} + \boldsymbol{\delta}^{*T}_\epsilon \right) \mathbb{E}_{\mathbf{Y}^m_t | \mathbf{Y}^o_t, S_{\epsilon,t}, S_{x,t}, \Psi}[\mathbf{Y}_t] \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1} \\
&+ \frac{1}{\sigma^{*2}} \left(\mathbb{E}_{\mathbf{Y}^m_t | \mathbf{Y}^o_t, S_{\epsilon,t}, S_{x,t}, \Psi}[\mathbf{Y}_t] \mathbf{D}_{\epsilon,t} - \boldsymbol{\mu}^* \mathbf{D}_{\epsilon,t} - \boldsymbol{\delta}^*_\epsilon \right)^T \left(- \boldsymbol{\mu} \mathbf{D}_{\epsilon,t} \mathbf{W} - \boldsymbol{\delta}_\epsilon \mathbf{W} + \sigma^2 \boldsymbol{\delta}_x \right) \mathbf{M}_t^{-1} \\
&- \frac{1}{\sigma^{*2}} \mathbf{D}_{\epsilon,t} \mathbf{W}^* \mathbf{M}_t^{-1} \mathbf{W}^T \mathbf{D}_{\epsilon,t} \mathbb{E}_{\mathbf{Y}^m_t | \mathbf{Y}^o_t, S_{\epsilon,t}, S_{x,t}, \Psi}[\mathbf{Y}_t^T \mathbf{Y}_t] \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1} \\
&- \frac{1}{\sigma^{*2}} \mathbf{D}_{\epsilon,t} \mathbf{W}^* \mathbf{M}_t^{-1} \mathbf{W}^T \mathbf{D}_{\epsilon,t} \mathbb{E}_{\mathbf{Y}^m_t | \mathbf{Y}^o_t, S_{\epsilon,t}, S_{x,t}, \Psi}[\mathbf{Y}_t]^T \left(- \boldsymbol{\mu} \mathbf{D}_{\epsilon,t} \mathbf{W} - \boldsymbol{\delta}_\epsilon \mathbf{W} + \sigma^2 \boldsymbol{\delta}_x \right) \mathbf{M}_t^{-1} \\
&- \frac{1}{\sigma^{*2}} \mathbf{D}_{\epsilon,t} \mathbf{W}^* \mathbf{M}_t^{-1} \left(- \boldsymbol{\mu} \mathbf{D}_{\epsilon,t} \mathbf{W} - \boldsymbol{\delta}_\epsilon \mathbf{W} + \sigma^2 \boldsymbol{\delta}_x \right)^T \mathbb{E}_{\mathbf{Y}^m_t | \mathbf{Y}^o_t, S_{\epsilon,t}, S_{x,t}, \Psi}[\mathbf{Y}_t] \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1} \\
&- \frac{1}{\sigma^{*2}} \mathbf{D}_{\epsilon,t} \mathbf{W}^* \mathbf{M}_t^{-1} \left(- \boldsymbol{\mu} \mathbf{D}_{\epsilon,t} \mathbf{W} - \boldsymbol{\delta}_\epsilon \mathbf{W} + \sigma^2 \boldsymbol{\delta}_x \right)^T \left(- \boldsymbol{\mu} \mathbf{D}_{\epsilon,t} \mathbf{W} - \boldsymbol{\delta}_\epsilon \mathbf{W} + \sigma^2 \boldsymbol{\delta}_x \right) \mathbf{M}_t^{-1} - \frac{\sigma^2}{\sigma^{*2}} \mathbf{D}_{\epsilon,t} \mathbf{W}^* \mathbf{M}_t^{-1}
\end{aligned}$$

The gradient of the function $\tilde{v}$ with respect to the scalar σ^{*2} is given by

$$\begin{aligned}
\frac{\partial}{\partial \sigma^{*2}} \tilde{v}(\mathbf{y}^o_t, s_{\epsilon,t}, s_{x,t}; \Psi, \Psi^*) &= -\frac{d}{2\sigma^{*2}} - \frac{1}{2(\sigma^{*2})^2} \Bigg[- \operatorname{Tr} \left\{ \mathbb{E}_{\mathbf{Y}^m_t | \mathbf{Y}^o_t, S_{\epsilon,t}, S_{x,t}, \Psi}[\mathbf{Y}_t^T \mathbf{Y}_t] \mathbf{D}_{\epsilon,t} \right\} + 2 \mathbb{E}_{\mathbf{Y}^m_t | \mathbf{Y}^o_t, S_{\epsilon,t}, S_{x,t}, \Psi}[\mathbf{Y}_t] \left(\mathbf{D}_{\epsilon,t} \boldsymbol{\mu}^{*T} + \boldsymbol{\delta}^{*T}_\epsilon \right) \\
&- \boldsymbol{\mu}^* \mathbf{D}_{\epsilon,t} \boldsymbol{\mu}^{*T} - 2 \boldsymbol{\mu}^* \boldsymbol{\delta}^{*T}_\epsilon - \boldsymbol{\delta}^*_\epsilon \mathbf{D}^{-1}_{\epsilon,t} \boldsymbol{\delta}^{*T}_\epsilon + 2 \operatorname{Tr} \left\{ \mathbb{E}_{\mathbf{Y}^m_t | \mathbf{Y}^o_t, S_{\epsilon,t}, S_{x,t}, \Psi}[\mathbf{Y}_t^T \mathbf{Y}_t] \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1} \mathbf{W}^{*T} \mathbf{D}_{\epsilon,t} \right\} \\
&- 2 \left(\mathbb{E}_{\mathbf{Y}^m_t | \mathbf{Y}^o_t, S_{\epsilon,t}, S_{x,t}, \Psi}[\mathbf{Y}_t] \mathbf{D}_{\epsilon,t} \mathbf{W} - \boldsymbol{\mu} \mathbf{D}_{\epsilon,t} \mathbf{W} - \boldsymbol{\delta}_\epsilon \mathbf{W} + \sigma^2 \boldsymbol{\delta}_x \right) \mathbf{M}_t^{-1} \mathbf{W}^{*T} \left(\boldsymbol{\mu}^* \mathbf{D}_{\epsilon,t} + \boldsymbol{\delta}^*_\epsilon \right)^T \\
&+ 2 \mathbb{E}_{\mathbf{Y}^m_t | \mathbf{Y}^o_t, S_{\epsilon,t}, S_{x,t}, \Psi}[\mathbf{Y}_t] \mathbf{D}_{\epsilon,t} \mathbf{W}^* \mathbf{M}_t^{-1} \left(- \boldsymbol{\mu} \mathbf{D}_{\epsilon,t} \mathbf{W} - \boldsymbol{\delta}_\epsilon \mathbf{W} + \sigma^2 \boldsymbol{\delta}_x \right)^T - \sigma^2 \operatorname{Tr} \left\{ \mathbf{M}_t^{-1} \mathbf{W}^{*T} \mathbf{D}_{\epsilon,t} \mathbf{W}^* \right\} \\
&- \operatorname{Tr} \left\{ \mathbf{M}_t^{-1} \left(\mathbf{W}^T \mathbf{D}_{\epsilon,t} \mathbb{E}_{\mathbf{Y}^m_t | \mathbf{Y}^o_t, S_{\epsilon,t}, S_{x,t}, \Psi}[\mathbf{Y}_t^T \mathbf{Y}_t] \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1} \mathbf{W}^{*T} \mathbf{D}_{\epsilon,t} \mathbf{W}^* \right\} \right. \\
&- \operatorname{Tr} \left\{ \mathbf{M}_t^{-1} \mathbf{W}^T \mathbf{D}_{\epsilon,t} \mathbb{E}_{\mathbf{Y}^m_t | \mathbf{Y}^o_t, S_{\epsilon,t}, S_{x,t}, \Psi}[\mathbf{Y}_t]^T \left(- \boldsymbol{\mu} \mathbf{D}_{\epsilon,t} \mathbf{W} - \boldsymbol{\delta}_\epsilon \mathbf{W} + \sigma^2 \boldsymbol{\delta}_x \right) \mathbf{M}_t^{-1} \mathbf{W}^{*T} \mathbf{D}_{\epsilon,t} \mathbf{W}^* \right\} \\
&- \operatorname{Tr} \left\{ \mathbf{M}_t^{-1} \left(- \boldsymbol{\mu} \mathbf{D}_{\epsilon,t} \mathbf{W} - \boldsymbol{\delta}_\epsilon \mathbf{W} + \sigma^2 \boldsymbol{\delta}_x \right)^T \mathbb{E}_{\mathbf{Y}^m_t | \mathbf{Y}^o_t, S_{\epsilon,t}, S_{x,t}, \Psi}[\mathbf{Y}_t] \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1} \mathbf{W}^{*T} \mathbf{D}_{\epsilon,t} \mathbf{W}^* \right\} \\
&- \left. \operatorname{Tr} \left\{ \mathbf{M}_t^{-1} \left(- \boldsymbol{\mu} \mathbf{D}_{\epsilon,t} \mathbf{W} - \boldsymbol{\delta}_\epsilon \mathbf{W} + \sigma^2 \boldsymbol{\delta}_x \right)^T \left(- \boldsymbol{\mu} \mathbf{D}_{\epsilon,t} \mathbf{W} - \boldsymbol{\delta}_\epsilon \mathbf{W} + \sigma^2 \boldsymbol{\delta}_x \right) \mathbf{M}_t^{-1} \mathbf{W}^{*T} \mathbf{D}_{\epsilon,t} \mathbf{W}^* \right\} \right]
\end{aligned}$$

whereas the gradients with respect to the vectors $\boldsymbol{\delta}^*_\epsilon$ and $\boldsymbol{\delta}^*_x$ are given by

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\delta}^*_\epsilon} \tilde{v}\left(\mathbf{y}^o_t, s_{\epsilon,t}, s_{x,t}; \Psi, \Psi^*\right) &= \frac{1}{\sigma^{*2}} \mathbb{E}_{\mathbf{Y}^m_t | \mathbf{Y}^o_t, S_{\epsilon,t}, S_{x,t}, \Psi}[\mathbf{Y}_t] - \frac{1}{\sigma^{*2}} \boldsymbol{\mu}^* - \frac{1}{\sigma^{*2}} \boldsymbol{\delta}^*_\epsilon \mathbf{D}^{-1}_{\epsilon,t} - \frac{1}{\sigma^{*2}} \mathbb{E}_{\mathbf{Y}^m_t | \mathbf{Y}^o_t, S_{\epsilon,t}, S_{x,t}, \Psi}[\mathbf{Y}_t] \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1} \mathbf{W}^{*T} \\
&+ \frac{1}{\sigma^{*2}} \boldsymbol{\mu} \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1} \mathbf{W}^{*T} + \frac{1}{\sigma^{*2}} \boldsymbol{\delta}_\epsilon \mathbf{W} \mathbf{M}_t^{-1} \mathbf{W}^{*T} - \frac{\sigma^2}{\sigma^{*2}} \boldsymbol{\delta}_x \mathbf{M}_t^{-1} \mathbf{W}^{*T}
\end{aligned}$$

$$\frac{\partial}{\partial \boldsymbol{\delta}^*_x} \tilde{v}(\mathbf{y}^o_t, s_{\epsilon,t}, s_{x,t}; \Psi, \Psi^*) = -\boldsymbol{\delta}^*_x \mathbf{D}^{-1}_{x,t} + \mathbb{E}_{\mathbf{Y}^m_t | \mathbf{Y}^o_t, S_{\epsilon,t}, S_{x,t}, \Psi}[\mathbf{Y}_t] \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1} - \boldsymbol{\mu} \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1} - \boldsymbol{\delta}_\epsilon \mathbf{W} \mathbf{M}_t^{-1} + \sigma^2 \boldsymbol{\delta}_x \mathbf{M}_t^{-1}.$$

□

A.4.2.4. Proof of Corollary 15

Proof. Given the formulation of the gradient $\nabla_{\Psi^*} \tilde{Q}_{GStS,ind}$ from Equation (A14) and the system of equations which determines the maximizers of the function $\tilde{Q}_{GStS,ind}$ with respect to the vector of parameters $\Psi^* = [\boldsymbol{\mu}^*, \mathbf{W}^*, \sigma^{*2}, \boldsymbol{\delta}^*_\epsilon, \boldsymbol{\delta}^*_x]$

specified in Theorem 26, the maximizers of the function $\tilde{Q}_{GStS,ind}$ with respect to $\boldsymbol{\mu}^*$,σ^{*2}, δ_ϵ^* and δ_x^* have closed form solutions defined by sequences of various 2-dimensional integration problem as follows

$$
\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\mu}^*}\tilde{Q}_{GStS,ind}(\Psi,\Psi^*)=\mathbf{0}\Longleftrightarrow\sum_{t=1}^{N}\left\{H\Big(\mathbf{y}_{1:N/t}^{o};\Psi\Big)\int_0^1\int_0^1\frac{\partial}{\partial \boldsymbol{\mu}^*}\tilde{v}(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi,\Psi^*)\,m(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}=\mathbf{0}\right.\\
&\Longleftrightarrow\sum_{t=1}^{N}\left\{H\Big(\mathbf{y}_{1:N/t}^{o};\Psi\Big)\int_0^1\int_0^1\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,s_{\epsilon,t},s_{x,t},\Psi}\big[\mathbf{Y}_t\big]\mathbf{D}_{\epsilon,t}\,m(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\right\}\\
&\quad-\boldsymbol{\mu}^*\sum_{t=1}^{N}\left\{H\Big(\mathbf{y}_{1:N/t}^{o};\Psi\Big)\int_0^1\int_0^1\mathbf{D}_{\epsilon,t}\,m(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\right\}\\
&\quad-\delta_\epsilon^*\sum_{t=1}^{N}\left\{H\Big(\mathbf{y}_{1:N/t}^{o};\Psi\Big)\int_0^1\int_0^1m(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\right\}\\
&\quad-\sum_{t=1}^{N}\left\{H\Big(\mathbf{y}_{1:N/t}^{o};\Psi\Big)\int_0^1\int_0^1\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,s_{\epsilon,t},s_{x,t},\Psi}\big[\mathbf{Y}_t\big]\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\,m(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\right\}\\
&\quad+\boldsymbol{\mu}\sum_{t=1}^{N}\left\{H\Big(\mathbf{y}_{1:N/t}^{o};\Psi\Big)\int_0^1\int_0^1\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\,m(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\right\}\\
&\quad+\left(\delta_\epsilon\mathbf{W}-\sigma^2\delta_x\right)\sum_{t=1}^{N}\left\{H\Big(\mathbf{y}_{1:N/t}^{o};\Psi\Big)\int_0^1\int_0^1\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\,m(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\right\}=\mathbf{0}\\
&\Longleftrightarrow\boldsymbol{\mu}^*=\Big[\mathbf{A}_6(\mathbf{y}_{1:N}^{o};\Psi)-\mathbf{A}_{10}(\mathbf{y}_{1:N}^{o};\Psi,\Psi^*)-\delta_\epsilon^*A_0(\mathbf{y}_{1:N}^{o};\Psi)+\boldsymbol{\mu}\mathbf{A}_{13}(\mathbf{y}_{1:N}^{o};\Psi,\Psi^*)+\left(\delta_\epsilon\mathbf{W}-\sigma^2\delta_x\right)\mathbf{A}_{12}(\mathbf{y}_{1:N}^{o};\Psi,\Psi^*)^T\Big]\mathbf{A}_1(\mathbf{y}_{1:N}^{o};\Psi)^{-1}
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\partial}{\partial \delta_\epsilon^*}\tilde{Q}_{GStS,ind}(\Psi,\Psi^*)=\mathbf{0}\Longleftrightarrow\sum_{t=1}^{N}\left\{H\Big(\mathbf{y}_{1:N/t}^{o};\Psi\Big)\int_0^1\int_0^1\frac{\partial}{\partial \delta_\epsilon^*}\tilde{v}(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi,\Psi^*)\,m(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}=\mathbf{0}\right.\\
&\Longleftrightarrow\sum_{t=1}^{N}\left\{H\Big(\mathbf{y}_{1:N/t}^{o};\Psi\Big)\int_0^1\int_0^1\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,s_{\epsilon,t},s_{x,t},\Psi}\big[\mathbf{Y}_t\big]\,m(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\right\}\\
&\quad-\boldsymbol{\mu}^*\sum_{t=1}^{N}\left\{H\Big(\mathbf{y}_{1:N/t}^{o};\Psi\Big)\int_0^1\int_0^1m(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\right\}\\
&\quad-\delta_\epsilon^*\sum_{t=1}^{N}\left\{H\Big(\mathbf{y}_{1:N/t}^{o};\Psi\Big)\int_0^1\int_0^1\mathbf{D}_{\epsilon,t}^{-1}\,m(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\right\}\\
&\quad-\sum_{t=1}^{N}\left\{H\Big(\mathbf{y}_{1:N/t}^{o};\Psi\Big)\int_0^1\int_0^1\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,s_{\epsilon,t},s_{x,t},\Psi}\big[\mathbf{Y}_t\big]\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\,m(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\right\}\mathbf{W}^{*T}\\
&\quad+\boldsymbol{\mu}\sum_{t=1}^{N}\left\{H\Big(\mathbf{y}_{1:N/t}^{o};\Psi\Big)\int_0^1\int_0^1\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\,m(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\right\}\mathbf{W}^{*T}\\
&\quad+\left(\delta_\epsilon\mathbf{W}-\sigma^2\delta_x\right)\sum_{t=1}^{N}\left\{H\Big(\mathbf{y}_{1:N/t}^{o};\Psi\Big)\int_0^1\int_0^1\mathbf{M}_t^{-1}\,m(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\right\}\mathbf{W}^{*T}=\mathbf{0}\\
&\Longleftrightarrow\delta_\epsilon^*=\Big[\mathbf{A}_5(\mathbf{y}_{1:N}^{o};\Psi)-\boldsymbol{\mu}^*A_0(\mathbf{y}_{1:N}^{o};\Psi)-\mathbf{A}_8(\mathbf{y}_{1:N}^{o};\Psi)\mathbf{W}^{*T}+\boldsymbol{\mu}\mathbf{A}_{11}(\mathbf{y}_{1:N}^{o};\Psi)\mathbf{W}^{*T}+\left(\delta_\epsilon\mathbf{W}-\sigma^2\delta_x\right)\mathbf{A}_4(\mathbf{y}_{1:N}^{o};\Psi)\mathbf{W}^{*T}\Big]\mathbf{A}_2(\mathbf{y}_{1:N}^{o};\Psi)^{-1}
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\partial}{\partial \delta_x^*}\tilde{Q}_{GStS,ind}(\Psi,\Psi^*)=\mathbf{0}\Longleftrightarrow\sum_{t=1}^{N}\left\{H\Big(\mathbf{y}_{1:N/t}^{o};\Psi\Big)\int_0^1\int_0^1\frac{\partial}{\partial \delta_x^*}\tilde{v}(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi,\Psi^*)\,m(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}=\mathbf{0}\right.\\
&\Longleftrightarrow-\delta_x^*\sum_{t=1}^{N}\left\{H\Big(\mathbf{y}_{1:N/t}^{o};\Psi\Big)\int_0^1\int_0^1\mathbf{D}_{x,t}^{-1}\,m(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\right\}\\
&\quad+\sum_{t=1}^{N}\left\{H\Big(\mathbf{y}_{1:N/t}^{o};\Psi\Big)\int_0^1\int_0^1\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,s_{\epsilon,t},s_{x,t},\Psi}\big[\mathbf{Y}_t\big]\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\,m(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\right\}\\
&\quad-\boldsymbol{\mu}\sum_{t=1}^{N}\left\{H\Big(\mathbf{y}_{1:N/t}^{o};\Psi\Big)\int_0^1\int_0^1\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\,m(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\right\}\\
&\quad-\left(\delta_\epsilon\mathbf{W}-\sigma^2\delta_x\right)\sum_{t=1}^{N}\left\{H\Big(\mathbf{y}_{1:N/t}^{o};\Psi\Big)\int_0^1\int_0^1\mathbf{M}_t^{-1}\,m(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\right\}=\mathbf{0}\\
&\Longleftrightarrow\delta_x^*=\Big[\mathbf{A}_8(\mathbf{y}_{1:N}^{o};\Psi)-\boldsymbol{\mu}\mathbf{A}_{11}(\mathbf{y}_{1:N}^{o};\Psi)-\left(\delta_\epsilon\mathbf{W}-\sigma^2\delta_x\right)\mathbf{A}_4(\mathbf{y}_{1:N}^{o};\Psi)\Big]\mathbf{A}_3(\mathbf{y}_{1:N},\Psi)^{-1}
\end{aligned}
$$

$$
\frac{\partial}{\partial \sigma^{*2}}\tilde{Q}_{GStS,ind}(\Psi,\Psi^*)=\mathbf{0}\Longleftrightarrow\sum_{t=1}^{N}\left\{H\Big(\mathbf{y}_{1:N/t}^{o};\Psi\Big)\int_0^1\int_0^1\frac{\partial}{\partial \sigma^{*2}}\tilde{v}(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi,\Psi^*)\,m(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}=\mathbf{0}\right.
$$

$$
\begin{aligned}
\iff & -\frac{d}{2\sigma^{*2}} \sum_{t=1}^{N} \left\{ H\left(\mathbf{y}^o_{1:N/t}; \Psi\right) \int_0^1 \int_0^1 m(\mathbf{y}^o_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \right\} \\
& - \frac{1}{2(\sigma^{*2})^2} \Bigg[- \sum_{t=1}^{N} \left\{ H\left(\mathbf{y}^o_{1:N/t}; \Psi\right) \int_0^1 \int_0^1 \mathrm{Tr}\left\{ \mathbb{E}_{\mathbf{Y}^m_t | \mathbf{Y}^o_t, s_{\epsilon,t}, s_{x,t}, \Psi}\left[\mathbf{Y}_t^T \mathbf{Y}_t\right] \mathbf{D}_{\epsilon,t} \right\} m(\mathbf{y}^o_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \right\} \\
& + 2 \sum_{t=1}^{N} \left\{ H\left(\mathbf{y}^o_{1:N/t}; \Psi\right) \int_0^1 \int_0^1 \mathbb{E}_{\mathbf{Y}^m_t | \mathbf{Y}^o_t, s_{\epsilon,t}, s_{x,t}, \Psi}\left[\mathbf{Y}_t\right] \mathbf{D}_{\epsilon,t}\, m(\mathbf{y}^o_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \right\} \boldsymbol{\mu}^{*T} \\
& + 2 \sum_{t=1}^{N} \left\{ H\left(\mathbf{y}^o_{1:N/t}; \Psi\right) \int_0^1 \int_0^1 \mathbb{E}_{\mathbf{Y}^m_t | \mathbf{Y}^o_t, s_{\epsilon,t}, s_{x,t}, \Psi}\left[\mathbf{Y}_t\right] m(\mathbf{y}^o_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \right\} \boldsymbol{\delta}_\epsilon^{*T} \\
& - \boldsymbol{\mu}^* \sum_{t=1}^{N} \left\{ H\left(\mathbf{y}^o_{1:N/t}; \Psi\right) \int_0^1 \int_0^1 \mathbf{D}_{\epsilon,t}\, m(\mathbf{y}^o_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \right\} \boldsymbol{\mu}^{*T} \\
& - 2\boldsymbol{\mu}^* \boldsymbol{\delta}_\epsilon^{*T} \sum_{t=1}^{N} \left\{ H\left(\mathbf{y}^o_{1:N/t}; \Psi\right) \int_0^1 \int_0^1 m(\mathbf{y}^o_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \right\} \\
& - \boldsymbol{\delta}_\epsilon^* \sum_{t=1}^{N} \left\{ H\left(\mathbf{y}^o_{1:N/t}; \Psi\right) \int_0^1 \int_0^1 \mathbf{D}_{\epsilon,t}^{-1}\, m(\mathbf{y}^o_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \right\} \boldsymbol{\delta}_\epsilon^{*T} \\
& + 2 \sum_{t=1}^{N} \left\{ H\left(\mathbf{y}^o_{1:N/t}; \Psi\right) \int_0^1 \int_0^1 \mathrm{Tr}\left\{ \mathbb{E}_{\mathbf{Y}^m_t | \mathbf{Y}^o_t, s_{\epsilon,t}, s_{x,t}, \Psi}\left[\mathbf{Y}_t^T \mathbf{Y}_t\right] \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1} \mathbf{W}^{*T} \mathbf{D}_{\epsilon,t} \right\} m(\mathbf{y}^o_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \right\} \\
& - 2 \sum_{t=1}^{N} \left\{ H\left(\mathbf{y}^o_{1:N/t}; \Psi\right) \int_0^1 \int_0^1 \mathbb{E}_{\mathbf{Y}^m_t | \mathbf{Y}^o_t, s_{\epsilon,t}, s_{x,t}, \Psi}\left[\mathbf{Y}_t\right] \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1} \mathbf{W}^{*T} \mathbf{D}_{\epsilon,t}\, m(\mathbf{y}^o_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \right\} \boldsymbol{\mu}^{*T} \\
& + 2\boldsymbol{\mu} \sum_{t=1}^{N} \left\{ H\left(\mathbf{y}^o_{1:N/t}; \Psi\right) \int_0^1 \int_0^1 \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1} \mathbf{W}^{*T} \mathbf{D}_{\epsilon,t}\, m(\mathbf{y}^o_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \right\} \boldsymbol{\mu}^{*T} \\
& + 2\left(\boldsymbol{\delta}_\epsilon \mathbf{W} - \sigma^2 \boldsymbol{\delta}_x\right) \sum_{t=1}^{N} \left\{ H\left(\mathbf{y}^o_{1:N/t}; \Psi\right) \int_0^1 \int_0^1 \mathbf{M}_t^{-1} \mathbf{W}^{*T} \mathbf{D}_{\epsilon,t}\, m(\mathbf{y}^o_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \right\} \boldsymbol{\mu}^{*T} \\
& - 2 \sum_{t=1}^{N} \left\{ H\left(\mathbf{y}^o_{1:N/t}; \Psi\right) \int_0^1 \int_0^1 \mathbb{E}_{\mathbf{Y}^m_t | \mathbf{Y}^o_t, s_{\epsilon,t}, s_{x,t}, \Psi}\left[\mathbf{Y}_t\right] \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1}\, m(\mathbf{y}^o_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \right\} \mathbf{W}^{*T} \boldsymbol{\delta}_\epsilon^{*T} \\
& + 2\boldsymbol{\mu} \sum_{t=1}^{N} \left\{ H\left(\mathbf{y}^o_{1:N/t}; \Psi\right) \int_0^1 \int_0^1 \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1}\, m(\mathbf{y}^o_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \right\} \mathbf{W}^{*T} \boldsymbol{\delta}_\epsilon^{*T} \\
& + 2\left(\boldsymbol{\delta}_\epsilon \mathbf{W} - \sigma^2 \boldsymbol{\delta}_x\right) \sum_{t=1}^{N} \left\{ H\left(\mathbf{y}^o_{1:N/t}; \Psi\right) \int_0^1 \int_0^1 \mathbf{M}_t^{-1}\, m(\mathbf{y}^o_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \right\} \mathbf{W}^{*T} \boldsymbol{\delta}_\epsilon^{*T} \\
& - 2 \sum_{t=1}^{N} \left\{ H\left(\mathbf{y}^o_{1:N/t}; \Psi\right) \int_0^1 \int_0^1 \mathbb{E}_{\mathbf{Y}^m_t | \mathbf{Y}^o_t, s_{\epsilon,t}, s_{x,t}, \Psi}\left[\mathbf{Y}_t\right] \mathbf{D}_{\epsilon,t} \mathbf{W}^* \mathbf{M}_t^{-1} \mathbf{W}^T \mathbf{D}_{\epsilon,t}\, m(\mathbf{y}^o_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \right\} \boldsymbol{\mu}^T \\
& - 2 \sum_{t=1}^{N} \left\{ H\left(\mathbf{y}^o_{1:N/t}; \Psi\right) \int_0^1 \int_0^1 \mathbb{E}_{\mathbf{Y}^m_t | \mathbf{Y}^o_t, s_{\epsilon,t}, s_{x,t}, \Psi}\left[\mathbf{Y}_t\right] \mathbf{D}_{\epsilon,t} \mathbf{W}^* \mathbf{M}_t^{-1}\, m(\mathbf{y}^o_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \right\} \left(\boldsymbol{\delta}_\epsilon \mathbf{W} - \sigma^2 \boldsymbol{\delta}_x\right)^T \\
& - \sigma^2 \sum_{t=1}^{N} \left\{ H\left(\mathbf{y}^o_{1:N/t}; \Psi\right) \int_0^1 \int_0^1 \mathrm{Tr}\left\{ \mathbf{M}_t^{-1} \mathbf{W}^{*T} \mathbf{D}_{\epsilon,t} \mathbf{W}^* \right\} m(\mathbf{y}^o_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \right\} \\
& - \sum_{t=1}^{N} \left\{ H\left(\mathbf{y}^o_{1:N/t}; \Psi\right) \int_0^1 \int_0^1 \mathrm{Tr}\left\{ \mathbf{M}_t^{-1} \mathbf{W}^T \mathbf{D}_{\epsilon,t} \mathbb{E}_{\mathbf{Y}^m_t | \mathbf{Y}^o_t, s_{\epsilon,t}, s_{x,t}, \Psi}\left[\mathbf{Y}_t^T \mathbf{Y}_t\right] \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1} \mathbf{W}^{*T} \mathbf{D}_{\epsilon,t} \mathbf{W}^* \right\} \right. \\
& \quad - 2\,\mathrm{Tr}\left\{ \mathbf{M}_t^{-1} \mathbf{W}^T \mathbf{D}_{\epsilon,t} \mathbb{E}_{\mathbf{Y}^m_t | \mathbf{Y}^o_t, s_{\epsilon,t}, s_{x,t}, \Psi}\left[\mathbf{Y}_t\right]^T \left(- \boldsymbol{\mu} \mathbf{D}_{\epsilon,t} \mathbf{W} - \boldsymbol{\delta}_\epsilon \mathbf{W} + \sigma^2 \boldsymbol{\delta}_x \right) \mathbf{M}_t^{-1} \mathbf{W}^{*T} \mathbf{D}_{\epsilon,t} \mathbf{W}^* \right\} \\
& \quad \left. - \mathrm{Tr}\left\{ \mathbf{M}_t^{-1} \left(- \boldsymbol{\mu} \mathbf{D}_{\epsilon,t} \mathbf{W} - \boldsymbol{\delta}_\epsilon \mathbf{W} + \sigma^2 \boldsymbol{\delta}_x \right)^T \left(- \boldsymbol{\mu} \mathbf{D}_{\epsilon,t} \mathbf{W} - \boldsymbol{\delta}_\epsilon \mathbf{W} + \sigma^2 \boldsymbol{\delta}_x \right) \mathbf{M}_t^{-1} \mathbf{W}^{*T} \mathbf{D}_{\epsilon,t} \mathbf{W}^* \right\} m(\mathbf{y}^o_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \right\} \Bigg] = 0 \\
\iff & \sigma^{*2} = \frac{1}{d A_0(\mathbf{y}^o_{1:N}; \Psi)} \Big[A_{20}(\mathbf{y}^o_{1:N}; \Psi) + \boldsymbol{\mu}^* \mathbf{A}_1(\mathbf{y}^o_{1:N}; \Psi) \boldsymbol{\mu}^{*T} + \boldsymbol{\delta}_\epsilon^* \mathbf{A}_2(\mathbf{y}^o_{1:N}; \Psi) \boldsymbol{\delta}_\epsilon^{*T} + 2\mathbf{A}_9(\mathbf{y}^o_{1:N}; \Psi, \Psi^*) \boldsymbol{\mu}^T \\
& \quad + 2\left(\mathbf{A}_{10}(\mathbf{y}^o_{1:N}; \Psi, \Psi^*) - \mathbf{A}_6(\mathbf{y}^o_{1:N}; \Psi) - \boldsymbol{\mu} \mathbf{A}_{13}(\mathbf{y}^o_{1:N}; \Psi, \Psi^*) \right) \boldsymbol{\mu}^{*T} - 2A_{21}(\mathbf{y}^o_{1:N}; \Psi, \Psi^*) + \sigma^2\, \mathrm{Tr}\left\{ \mathbf{A}_{12}(\mathbf{y}^o_{1:N}; \Psi)^T \mathbf{W}^* \right\} \\
& \quad + \mathrm{Tr}\left\{ \mathbf{A}_{17}(\mathbf{y}^o_{1:N}; \Psi)^T \mathbf{W}^* \right\} - \mathrm{Tr}\left\{ \mathbf{A}_{18.1}(\mathbf{y}^o_{1:N}; \Psi)^T \mathbf{W}^* \right\} - \mathrm{Tr}\left\{ \mathbf{A}_{18.2}(\mathbf{y}^o_{1:N}; \Psi)^T \mathbf{W}^* \right\} + \mathrm{Tr}\left\{ \mathbf{A}_{19}(\mathbf{y}^o_{1:N}; \Psi)^T \mathbf{W}^* \right\} \\
& \quad + 2\left(\boldsymbol{\mu}^* A_0(\mathbf{y}^o_{1:N}; \Psi) - \mathbf{A}_5(\mathbf{y}^o_{1:N}; \Psi) + \mathbf{A}_8(\mathbf{y}^o_{1:N}; \Psi) \mathbf{W}^{*T} - \boldsymbol{\mu} \mathbf{A}_{11}(\mathbf{y}^o_{1:N}; \Psi) \mathbf{W}^{*T} \right) \boldsymbol{\delta}_\epsilon^{*T} \\
& \quad + 2\left(\mathbf{A}_7(\mathbf{y}^o_{1:N}; \Psi, \Psi^*) - \boldsymbol{\mu}^* \mathbf{A}_{12}(\mathbf{y}^o_{1:N}; \Psi, \Psi^*) - \boldsymbol{\delta}_\epsilon^* \mathbf{W}^* \mathbf{A}_4(\mathbf{y}^o_{1:N}; \Psi) \right) \left(\boldsymbol{\delta}_\epsilon \mathbf{W} - \sigma^2 \boldsymbol{\delta}_x\right)^T \Big]
\end{aligned}
$$

where

$$
\begin{aligned}
A_0(\mathbf{y}^o_{1:N}; \Psi)_{1\times 1} &:= \sum_{t=1}^{N} \left\{ H\left(\mathbf{y}^o_{1:N/t}; \Psi\right) \int_0^1 \int_0^1 m(\mathbf{y}^o_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \right\} \\
\mathbf{A}_1(\mathbf{y}^o_{1:N}; \Psi)_{d\times d} &:= \sum_{t=1}^{N} \left\{ H\left(\mathbf{y}^o_{1:N/t}; \Psi\right) \int_0^1 \int_0^1 \mathbf{D}_{\epsilon,t}\, m(\mathbf{y}^o_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} \right\}
\end{aligned}
$$

$$\mathbf{A}_2(\mathbf{y}^o_{1:N};\Psi)_{d\times d} := \sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\ \int_0^1\int_0^1 \mathbf{D}^{-1}_{\epsilon,t}\ m(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\right\}$$

$$\mathbf{A}_3(\mathbf{y}^o_{1:N};\Psi)_{k\times k} := \sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\ \int_0^1\int_0^1 \mathbf{D}^{-1}_{x,t}\ m(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\right\}$$

$$\mathbf{A}_4(\mathbf{y}^o_{1:N};\Psi)_{k\times k} := \sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\ \int_0^1\int_0^1 \mathbf{M}^{-1}_{t}\ m(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\right\}$$

$$\mathbf{A}_5(\mathbf{y}^o_{1:N};\Psi)_{1\times d} := \sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\ \int_0^1\int_0^1 \mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]\ m(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\right\}$$

$$\mathbf{A}_6(\mathbf{y}^o_{1:N};\Psi)_{1\times d} := \sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\ \int_0^1\int_0^1 \mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]\mathbf{D}_{\epsilon,t}\ m(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\right\}$$

$$\mathbf{A}_7(\mathbf{y}^o_{1:N};\Psi,\Psi^*)_{1\times k} := \sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\ \int_0^1\int_0^1 \mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]\mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}^{-1}_{t}\ m(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\right\}$$

$$\mathbf{A}_8(\mathbf{y}^o_{1:N};\Psi)_{1\times k} := \sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\ \int_0^1\int_0^1 \mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}^{-1}_{t}\ m(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\right\}$$

$$\mathbf{A}_9(\mathbf{y}^o_{1:N};\Psi,\Psi^*)_{1\times d} := \sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\ \int_0^1\int_0^1 \mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]\mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}^{-1}_{t}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\ m(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\right\}$$

$$\mathbf{A}_{10}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)_{1\times d} := \sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\ \int_0^1\int_0^1 \mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}^{-1}_{t}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\ m(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\right\}$$

$$\mathbf{A}_{11}(\mathbf{y}^o_{1:N};\Psi)_{d\times k} := \sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\ \int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}^{-1}_{t}\ m(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\right\}$$

$$\mathbf{A}_{12}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)_{d\times k} := \sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\ \int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}^{-1}_{t}\ m(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\right\}$$

$$\mathbf{A}_{13}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)_{d\times d} := \sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\ \int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}^{-1}_{t}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\ m(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\right\}$$

$$\mathbf{A}_{14}(\mathbf{y}^o_{1:N};\Psi)_{d\times k} := \sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\ \int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}^T_t\mathbf{Y}_t\right]\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}^{-1}_{t}\ m(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\right\}$$

$$\mathbf{A}_{15}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)_{d\times k} := \sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\ \int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}\boldsymbol{\mu}^{*T}\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}^{-1}_{t}\ m(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\right\}$$

$$\mathbf{A}_{16}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)_{d\times k} := \sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\ \int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}\left(\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]-\boldsymbol{\mu}^*\right)^T\left(\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}+\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\right)\mathbf{M}^{-1}_{t}\right.$$
$$\left.\times\ m(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\right\}$$

$$\mathbf{A}_{17}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)_{k\times d} := \sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\ \int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}^{-1}_{t}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}^T_t\mathbf{Y}_t\right]\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}^{-1}_{t}\right.$$
$$\left.\times\ m(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\right\}$$

$$\mathbf{A}_{18.1}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)_{k\times d} := \sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\ \int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}^{-1}_{t}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]^T\left(\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}+\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\right)\mathbf{M}^{-1}_{t}\right.$$
$$\left.\times\ m(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\right\}$$

$$\mathbf{A}_{18.2}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)_{k\times d} := \sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\ \int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}^{-1}_{t}\left(\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}+\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\right)^T\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}^{-1}_{t}\right.$$
$$\left.\times\ m(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\right\}$$

$$\mathbf{A}_{19}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)_{k\times d} := \sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\ \int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}^{-1}_{t}\left(\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}+\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\right)^T\left(\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}+\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\right)\mathbf{M}^{-1}_{t}\right.$$
$$\left.\times\ m(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\right\}$$

$$A_{20}(\mathbf{y}^o_{1:N};\Psi)_{1\times 1} := \sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\ \int_0^1\int_0^1 \mathrm{Tr}\left\{\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}^T_t\mathbf{Y}_t\right]\mathbf{D}_{\epsilon,t}\right\}\ m(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\right\}$$

$$A_{21}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)_{1\times 1} := \sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\ \int_0^1\int_0^1 \mathrm{Tr}\left\{\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}^T_t\mathbf{Y}_t\right]\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}^{-1}_{t}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\right\}\ m(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi)\ ds_{\epsilon,t}\ ds_{x,t}\right\}$$

The maximizer which correspond to the parameter $\mathbf{W}^*$ is more difficult to obtain since the partial derivative of the function $\tilde{Q}_{GStS,ind}$ with respect to $\mathbf{W}^*$ requires integrating the matrix products which contain $\mathbf{W}^*$ and cannot be further simplified, that is

$$
\begin{aligned}
&\frac{\partial}{\partial \mathbf{W}^*}\tilde{Q}_{GStS,ind}(\Psi,\Psi^*)=\mathbf{0} \Longleftrightarrow \sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\int_0^1\int_0^1 \frac{\partial}{\partial \mathbf{W}^*}\tilde{v}\left(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi,\Psi^*\right)\, m\left(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi\right)\, ds_{\epsilon,t}\, ds_{x,t}\right\}=\mathbf{0}\\
&\Longleftrightarrow \sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right]\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\, m\left(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi\right)\, ds_{\epsilon,t}\, ds_{x,t}\right\}\\
&\quad-\sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}\boldsymbol{\mu}^{*T}\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\, m\left(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi\right)\, ds_{\epsilon,t}\, ds_{x,t}\right\}\\
&\quad-\boldsymbol{\delta}_\epsilon^{*T}\sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\int_0^1\int_0^1 \mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\, m\left(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi\right)\, ds_{\epsilon,t}\, ds_{x,t}\right\}\\
&\quad-\sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}\left(\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]-\boldsymbol{\mu}^*\right)^T\left(\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}+\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\right)\mathbf{M}_t^{-1}\, m\left(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi\right)\, ds_{\epsilon,t}\, ds_{x,t}\right\}\\
&\quad+\boldsymbol{\delta}_\epsilon^{*T}\boldsymbol{\mu}\sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\, m\left(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi\right)\, ds_{\epsilon,t}\, ds_{x,t}\right\}\\
&\quad+\boldsymbol{\delta}_\epsilon^{*T}\left(\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\right)\sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\int_0^1\int_0^1 \mathbf{M}_t^{-1}\, m\left(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi\right)\, ds_{\epsilon,t}\, ds_{x,t}\right\}\\
&\quad-\sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right]\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\, m\left(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi\right)\, ds_{\epsilon,t}\, ds_{x,t}\right\}\\
&\quad+\sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]^T\left(\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}+\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\right)\mathbf{M}_t^{-1}\, m\left(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi\right)\, ds_{\epsilon,t}\, ds_{x,t}\right\}\\
&\quad+\sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}_t^{-1}\left(\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}+\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\right)^T\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\, m\left(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi\right)\, ds_{\epsilon,t}\, ds_{x,t}\right\}\\
&\quad-\sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}_t^{-1}\left(\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}+\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\right)^T\left(\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}+\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\right)\mathbf{M}_t^{-1}\, m\left(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi\right)\, ds_{\epsilon,t}\, ds_{x,t}\right\}\\
&\quad-\sigma^2\sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\int_0^1\int_0^1 \mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}_t^{-1}\, m\left(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi\right)\, ds_{\epsilon,t}\, ds_{x,t}\right\}=\mathbf{0}\\
&\Longleftrightarrow \mathbf{A}_{14}(\mathbf{y}^o_{1:N};\Psi)-\mathbf{A}_{15}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)-\boldsymbol{\delta}_\epsilon^{*T}\mathbf{A}_8(\mathbf{y}^o_{1:N};\Psi)-\mathbf{A}_{16}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)+\boldsymbol{\delta}_\epsilon^{*T}\boldsymbol{\mu}\mathbf{A}_{11}(\mathbf{y}^o_{1:N};\Psi)\\
&\qquad+\boldsymbol{\delta}_\epsilon^{*T}(\boldsymbol{\delta}\mathbf{W}-\sigma^2\mathbf{W})\mathbf{A}_4(\mathbf{y}^o_{1:N};\Psi)-\mathbf{A}_{17}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)+\mathbf{A}_{18.1}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)+\mathbf{A}_{18.2}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\\
&\qquad-\mathbf{A}_{19}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)-\sigma^2\mathbf{A}_{12}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)=\mathbf{0}
\end{aligned}
$$

The maximizer with respect to $\mathbf{W}^*$ requires solving a root finding problem $\frac{\partial\tilde{Q}_{GStS,ind}(\Psi,\Psi^*)}{\partial\mathbf{W}^*}=\mathbf{0}$. Recall that $\frac{\partial\tilde{Q}_{GStS,ind}(\Psi,\Psi^*)}{\partial\mathbf{W}^*}$ contains elements of $\boldsymbol{\mu}^*$ and $\boldsymbol{\delta}_\epsilon^*$. By denoting $f:\mathbb{R}^{d\times k}\rightarrow\mathbb{R}^{d\times k}$ such that $f(\mathbf{W}^*;\Psi,\Psi^*):=\frac{\partial\tilde{Q}_{GStS,ind}(\Psi,\Psi^*)}{\partial\mathbf{W}^*}$ is given by

$$
\begin{aligned}
f(\mathbf{W}^*;\Psi,\Psi^*)&=\mathbf{A}_{14}(\mathbf{y}^o_{1:N};\Psi)-\mathbf{A}_{15}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)-\boldsymbol{\delta}_\epsilon^{*T}\mathbf{A}_8(\mathbf{y}^o_{1:N};\Psi)-\mathbf{A}_{16}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)+\boldsymbol{\delta}_\epsilon^{*T}\boldsymbol{\mu}\mathbf{A}_{11}(\mathbf{y}^o_{1:N};\Psi)\\
&\quad+\boldsymbol{\delta}_\epsilon^{*T}(\boldsymbol{\delta}\mathbf{W}-\sigma^2\mathbf{W})\mathbf{A}_4(\mathbf{y}^o_{1:N};\Psi)-\mathbf{A}_{17}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)+\mathbf{A}_{18.1}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)+\mathbf{A}_{18.2}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\\
&\quad-\mathbf{A}_{19}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)-\sigma^2\mathbf{A}_{12}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)^T
\end{aligned}
$$

the system of equations $\nabla_{\Psi^*}\tilde{Q}_{GStS,ind}(\Psi,\Psi^*)=\mathbf{0}$ reduces to

$$
\begin{cases}
\boldsymbol{\mu}^* = A_1(\mathbf{y}^o_{1:N};\Psi)^{-1}\Big[\mathbf{A}_6(\mathbf{y}^o_{1:N};\Psi) - A_0(\mathbf{y}^o_{1:N};\Psi)\boldsymbol{\delta}^*_\epsilon - \mathbf{A}_{10}(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}^{*T} + \boldsymbol{\mu}\mathbf{W}\mathbf{A}_{13}(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}^{*T} + \big(\delta_\epsilon\mathbf{W}+\sigma^2\delta_x\big)\mathbf{A}_{11}(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}^{*T}\Big]\\
\boldsymbol{\delta}^*_\epsilon = A_2(\mathbf{y}^o_{1:N};\Psi)^{-1}\Big[\mathbf{A}_5(\mathbf{y}_{1:N};\Psi) - A_0(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^* - \mathbf{A}_8(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}^{*T} + \boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}^{*T} + \Big(\delta_\epsilon\mathbf{W}-\sigma^2\delta_x\Big)\mathbf{A}_4(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}^{*T}\Big]\\
\delta^*_x = A_3(\mathbf{y}^o_{1:N};\Psi)^{-1}\Big(\mathbf{A}_8(\mathbf{y}_{1:N};\Psi) - \boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}^o_{1:N};\Psi) - \big(\delta_\epsilon\mathbf{W}-\sigma^2\delta_x\big)\mathbf{A}_4(\mathbf{y}^o_{1:N};\Psi)\Big)\\
\mathbf{W}^* = \Big[\mathbf{A}_{14}(\mathbf{y}^o_{1:N};\Psi) - \boldsymbol{\mu}^{*T}\mathbf{A}_{10}(\mathbf{y}^o_{1:N};\Psi) - \boldsymbol{\delta}^*_\epsilon\mathbf{A}_8(\mathbf{y}^o_{1:N};\Psi) - \mathbf{A}_{16}(\mathbf{y}^o_{1:N};\Psi) + \boldsymbol{\mu}^{*T}\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{13}(\mathbf{y}^o_{1:N};\Psi) + \boldsymbol{\mu}^{*T}\big(\delta_\epsilon-\sigma^2\delta_x\big)\mathbf{A}_{11}(\mathbf{y}^o_{1:N};\Psi)\\
\quad+\boldsymbol{\delta}^{*T}_\epsilon\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}^o_{1:N};\Psi) + \boldsymbol{\delta}^{*T}_\epsilon\big(\delta_\epsilon\mathbf{W}-\sigma^2\delta_x\big)\mathbf{A}_4(\mathbf{y}^o_{1:N};\Psi)\Big]\\
\quad\times\Big[\mathbf{A}_{17}(\mathbf{y}^o_{1:N};\Psi) - \mathbf{A}_{18}(\mathbf{y}^o_{1:N};\Psi) - \mathbf{A}_{18}(\mathbf{y}^o_{1:N};\Psi)^T + \mathbf{A}_{19}(\mathbf{y}^o_{1:N};\Psi) + \sigma^2\mathbf{A}_{11}(\mathbf{y}^o_{1:N};\Psi)\Big]^{-1}\\
\sigma^{*2} = \frac{1}{dA_0(\mathbf{y}^o_{1:N};\Psi)}\Big[A_{20}(\mathbf{y}^o_{1:N};\Psi) - 2\mathbf{A}_6(\mathbf{y}^o_{1:N};\Psi)\boldsymbol{\mu}^{*T} - 2\mathbf{A}_5(\mathbf{y}^o_{1:N};\Psi)\boldsymbol{\delta}^{*T}_\epsilon + A_1(\mathbf{y}^o_{1:N};\Psi)\boldsymbol{\mu}^*\boldsymbol{\mu}^{*T} + 2A_0(\mathbf{y}^o_{1:N};\Psi)\boldsymbol{\mu}^*\boldsymbol{\delta}^{*T}_\epsilon\\
\quad+A_2(\mathbf{y}^o_{1:N};\Psi)\boldsymbol{\delta}^*_\epsilon\boldsymbol{\delta}^{*T}_\epsilon - 2\operatorname{Tr}\Big\{\mathbf{A}_{14}(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}^{*T}\Big\} + 2\mathbf{A}_9(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\mathbf{W}^T\boldsymbol{\mu}^T + 2\mathbf{A}_7(\mathbf{y}^o_{1:N};\Psi,\Psi)\mathbf{W}^T\boldsymbol{\delta}^T_\epsilon\\
\quad-2\sigma^2\mathbf{A}_7(\mathbf{y}^o_{1:N};\Psi,\Psi)\delta^T_x + 2\Big(\mathbf{A}_{10}(\mathbf{y}^o_{1:N};\Psi) - \boldsymbol{\mu}\mathbf{W}\mathbf{A}_{13}(\mathbf{y}^o_{1:N};\Psi) - \big(\delta_\epsilon\mathbf{W}-\sigma^2\delta_x\big)\mathbf{A}_{11}(\mathbf{y}^o_{1:N};\Psi)\Big)\mathbf{W}^{*T}\boldsymbol{\mu}^{*T}\\
\quad+2\Big(\mathbf{A}_8(\mathbf{y}^o_{1:N};\Psi) - \boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}^o_{1:N};\Psi) - \big(\delta_\epsilon\mathbf{W}-\sigma^2\delta_x\big)\mathbf{A}_4(\mathbf{y}^o_{1:N};\Psi)\Big)\mathbf{W}^{*T}\boldsymbol{\delta}^{*T}_\epsilon + \operatorname{Tr}\Big\{\mathbf{A}_{17}(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}^{*T}\mathbf{W}^*\Big\}\\
\quad-\operatorname{Tr}\Big\{\Big(\mathbf{A}_{18}(\mathbf{y}^o_{1:N};\Psi) + \mathbf{A}_{18}(\mathbf{y}^o_{1:N};\Psi)^T\Big)\mathbf{W}^{*T}\mathbf{W}^*\Big\} + \operatorname{Tr}\Big\{\mathbf{A}_{19}(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}^{*T}\mathbf{W}^*\Big\} + \sigma^2\operatorname{Tr}\Big\{\mathbf{A}_{11}(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}^{*T}\mathbf{W}^*\Big\}\Big]
\end{cases}
$$

□

A.4.2.5. Proof of Corollary 19

Proof. Given the formulation of the gradient $\nabla_{\Psi^*}\tilde{Q}_{GStS,ind,2}$ from Equation (A14), system of equations specified in Theorem 26 and the assumptions of Special Case 2 from Subsection 3.4.1.3, the M-step for Generalized Skew t-Student IND PPCA is specified by finding the the maximizers of the function $\tilde{Q}_{GStS,ind,2}$ from Corollary 18 with respect to $\boldsymbol{\mu}^*$, σ^{*2}, $\boldsymbol{\delta}^*_\epsilon$ and $\boldsymbol{\delta}^*_x$ as following

$$
\begin{aligned}
&\frac{\partial\tilde{v}(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi,\Psi^*)}{\partial\boldsymbol{\mu}^*} = \frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\big[\mathbf{Y}_t\big] - \frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}^* - \frac{1}{\sigma^{*2}}\boldsymbol{\delta}^*_\epsilon - \frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})^2\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\big[\mathbf{Y}_t\big]\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}\\
&\quad-\frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})\Big(-T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W}-\boldsymbol{\delta}_\epsilon\mathbf{W}+\sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}_t^{-1}\mathbf{W}^{*T}\\
&\Longrightarrow \frac{\partial\tilde{Q}_{GStS,ind,2}(\Psi,\Psi^*)}{\partial\boldsymbol{\mu}^*} = \mathbf{0}\\
&\Longleftrightarrow \sum_{t=1}^{N}\Bigg\{H\Big(\mathbf{y}_{1:N\,t};\Psi\Big)\int_0^1\int_0^1\Big[\frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\big[\mathbf{Y}_t\big] - \frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}^* - \frac{1}{\sigma^{*2}}\boldsymbol{\delta}^*_\epsilon\\
&\quad-\frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})^2\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\big[\mathbf{Y}_t\big]\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T} - \frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})\Big(-T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W}-\boldsymbol{\delta}_\epsilon\mathbf{W}+\sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}_t^{-1}\mathbf{W}^{*T}\Big]m(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\Bigg\} = \mathbf{0}\\
&\Longrightarrow \mathbf{A}_6(\mathbf{y}_{1:N};\Psi) - A_1(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^* - A_0(\mathbf{y}_{1:N};\Psi)\boldsymbol{\delta}^*_\epsilon - \mathbf{A}_{10}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T} + \boldsymbol{\mu}\mathbf{W}\mathbf{A}_{13}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T} + \big(\boldsymbol{\delta}_\epsilon\mathbf{W}+\sigma^2\boldsymbol{\delta}_x\big)\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T} = \mathbf{0}\\
&\Longrightarrow \boldsymbol{\mu}^* = A_1(\mathbf{y}_{1:N};\Psi)^{-1}\Big[\mathbf{A}_6(\mathbf{y}_{1:N};\Psi) - A_0(\mathbf{y}_{1:N};\Psi)\boldsymbol{\delta}^*_\epsilon - \mathbf{A}_{10}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T} + \boldsymbol{\mu}\mathbf{W}\mathbf{A}_{13}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T} + \big(\boldsymbol{\delta}_\epsilon\mathbf{W}+\sigma^2\boldsymbol{\delta}_x\big)\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\Big]
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\partial\tilde{v}(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi,\Psi^*)}{\partial\mathbf{W}^*} = \frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})^2\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,\mathbf{U}_t,\mathbf{V}_t,\Psi}\big[\mathbf{Y}^T_t\mathbf{Y}_t\big]\mathbf{W}\mathbf{M}_t^{-1} - \frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})^2\boldsymbol{\mu}^{*T}\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\big[\mathbf{Y}_t\big]\mathbf{W}\mathbf{M}_t^{-1}\\
&\quad-\frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})\boldsymbol{\delta}^{*T}_\epsilon\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\big[\mathbf{Y}_t\big]\mathbf{W}\mathbf{M}_t^{-1} - \frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})^2\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\big[\mathbf{Y}_t\big]^T\boldsymbol{\mu}\mathbf{W}\mathbf{M}_t^{-1}\\
&\quad-\frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\big[\mathbf{Y}_t\big]^T\Big(\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}_t^{-1}\\
&\quad+\frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})^2\boldsymbol{\mu}^{*T}\boldsymbol{\mu}\mathbf{W}\mathbf{M}_t^{-1} + \frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}^{*T}\Big(\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}_t^{-1} + \frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})\boldsymbol{\delta}^{*T}_\epsilon\boldsymbol{\mu}\mathbf{W}\mathbf{M}_t^{-1} + \frac{1}{\sigma^{*2}}\boldsymbol{\delta}^{*T}_\epsilon\Big(\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}_t^{-1}\\
&\quad-\frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})^3\mathbf{W}^*\mathbf{M}_t^{-1}\mathbf{W}^T\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\big[\mathbf{Y}^T_t\mathbf{Y}_t\big]\mathbf{W}\mathbf{M}_t^{-1}\\
&\quad+\frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})^2\mathbf{W}^*\mathbf{M}_t^{-1}\mathbf{W}^T\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\big[\mathbf{Y}_t\big]^T\Big(T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W}+\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}_t^{-1}\\
&\quad+\frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})^2\mathbf{W}^*\mathbf{M}_t^{-1}\Big(T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W}+\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\Big)^T\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\big[\mathbf{Y}_t\big]\mathbf{W}^T\mathbf{M}_t^{-1}\\
&\quad-\frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})\mathbf{W}^*\mathbf{M}_t^{-1}\Big(T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W}+\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\Big)^T\Big(T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W}+\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}_t^{-1} - \frac{\sigma^2}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})\mathbf{W}^*\mathbf{M}_t^{-1}
\end{aligned}
$$

$$
\begin{aligned}
&\implies \frac{\partial \tilde{Q}_{GStS,ind,2}(\Psi,\Psi^*)}{\partial \mathbf{W}^*} = \mathbf{0} \\
&\iff \sum_{t=1}^{N}\Bigg\{ H\Big(\mathbf{y}_{1:N\ t};\Psi\Big)\ \int_0^1\int_0^1 \Big[T_\epsilon(s_{\epsilon,t})^2 \mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U}_t,\mathbf{V}_t,\Psi}\big[\mathbf{Y}_t^T\mathbf{Y}_t\big]\mathbf{W}\mathbf{M}_t^{-1} - T_\epsilon(s_{\epsilon,t})^2\boldsymbol{\mu}^{*T}\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\big[\mathbf{Y}_t\big]\mathbf{W}\mathbf{M}_t^{-1} \\
&\quad - T_\epsilon(s_{\epsilon,t})\boldsymbol{\delta}_\epsilon^{*T}\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\big[\mathbf{Y}_t\big]\mathbf{W}\mathbf{M}_t^{-1} - T_\epsilon(s_{\epsilon,t})^2\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\big[\mathbf{Y}_t\big]^T\boldsymbol{\mu}\mathbf{W}\mathbf{M}_t^{-1} - T_\epsilon(s_{\epsilon,t})\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\big[\mathbf{Y}_t\big]^T\Big(\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}_t^{-1} \\
&\quad + T_\epsilon(s_{\epsilon,t})^2\boldsymbol{\mu}^{*T}\boldsymbol{\mu}\mathbf{W}\mathbf{M}_t^{-1} + T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}^{*T}\Big(\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}_t^{-1} + T_\epsilon(s_{\epsilon,t})\boldsymbol{\delta}_\epsilon^{*T}\boldsymbol{\mu}\mathbf{W}\mathbf{M}_t^{-1} + \boldsymbol{\delta}_\epsilon^{*T}\Big(\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}_t^{-1} \\
&\quad - T_\epsilon(s_{\epsilon,t})^3\mathbf{W}^*\mathbf{M}_t^{-1}\mathbf{W}^T\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\big[\mathbf{Y}_t^T\mathbf{Y}_t\big]\mathbf{W}\mathbf{M}_t^{-1} + T_\epsilon(s_{\epsilon,t})^2\mathbf{W}^*\mathbf{M}_t^{-1}\mathbf{W}^T\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\big[\mathbf{Y}_t\big]^T\Big(T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W}+\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}_t^{-1} \\
&\quad + T_\epsilon(s_{\epsilon,t})^2\mathbf{W}^*\mathbf{M}_t^{-1}\Big(T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W}+\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\Big)^T\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\big[\mathbf{Y}_t\big]\mathbf{W}^T\mathbf{M}_t^{-1} \\
&\quad - T_\epsilon(s_{\epsilon,t})\mathbf{W}^*\mathbf{M}_t^{-1}\Big(T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W}+\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\Big)^T\Big(T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W}+\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}_t^{-1} - \sigma^2 T_\epsilon(s_{\epsilon,t})\mathbf{W}^*\mathbf{M}_t^{-1}\Big] \\
&\quad \times m(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\Bigg\} = \mathbf{0} \\
&\iff \mathbf{A}_{14}(\mathbf{y}_{1:N};\Psi) - \boldsymbol{\mu}^{*T}\mathbf{A}_{10}(\mathbf{y}_{1:N};\Psi) - \boldsymbol{\delta}_\epsilon^*\mathbf{A}_8(\mathbf{y}_{1:N};\Psi) - \mathbf{A}_{16}(\mathbf{y}_{1:N};\Psi) + \boldsymbol{\mu}^{*T}\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{13}(\mathbf{y}_{1:N};\Psi) + \boldsymbol{\mu}^{*T}\big(\boldsymbol{\delta}_\epsilon-\sigma^2\boldsymbol{\delta}_x\big)\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi) \\
&+ \boldsymbol{\delta}_\epsilon^{*T}\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi) + \boldsymbol{\delta}_\epsilon^{*T}\big(\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\big)\mathbf{A}_4(\mathbf{y}_{1:N};\Psi) - \mathbf{W}^*\Big(\mathbf{A}_{17}(\mathbf{y}_{1:N};\Psi) - \mathbf{A}_{18}(\mathbf{y}_{1:N};\Psi) - \mathbf{A}_{18}(\mathbf{y}_{1:N};\Psi)^T \\
&\quad + \mathbf{A}_{19}(\mathbf{y}_{1:N};\Psi) + \sigma^2\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi)\Big) = \mathbf{0} \\
&\iff \mathbf{W}^* = \Big[\mathbf{A}_{14}(\mathbf{y}_{1:N};\Psi) - \boldsymbol{\mu}^{*T}\mathbf{A}_{10}(\mathbf{y}_{1:N};\Psi) - \boldsymbol{\delta}_\epsilon^*\mathbf{A}_8(\mathbf{y}_{1:N};\Psi) - \mathbf{A}_{16}(\mathbf{y}_{1:N};\Psi) + \boldsymbol{\mu}^{*T}\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{13}(\mathbf{y}_{1:N};\Psi) + \boldsymbol{\mu}^{*T}\big(\boldsymbol{\delta}_\epsilon-\sigma^2\boldsymbol{\delta}_x\big)\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi) \\
&+ \boldsymbol{\delta}_\epsilon^{*T}\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi) + \boldsymbol{\delta}_\epsilon^{*T}\big(\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\big)\mathbf{A}_4(\mathbf{y}_{1:N};\Psi)\Big] \\
&\quad \times \Big[\mathbf{A}_{17}(\mathbf{y}_{1:N};\Psi) - \mathbf{A}_{18}(\mathbf{y}_{1:N};\Psi) - \mathbf{A}_{18}(\mathbf{y}_{1:N};\Psi)^T + \mathbf{A}_{19}(\mathbf{y}_{1:N};\Psi) + \sigma^2\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi)\Big]^{-1}
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\partial \tilde{v}(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi,\Psi^*)}{\partial \boldsymbol{\delta}_\epsilon^*} = \frac{1}{\sigma^{*2}}\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\big[\mathbf{Y}_t\big] - \frac{1}{\sigma^{*2}}\boldsymbol{\mu}^* - \frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})^{-1}\boldsymbol{\delta}_\epsilon^* - \frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\big[\mathbf{Y}_t\big]\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T} \\
&\quad + \frac{1}{\sigma^{*2}}T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T} + \frac{1}{\sigma^{*2}}\Big(\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}_t^{-1}\mathbf{W}^{*T} \\
&\implies \frac{\partial \tilde{Q}_{GStS,ind,2}(\Psi,\Psi^*)}{\partial \boldsymbol{\delta}_\epsilon^*} = \mathbf{0} \\
&\iff \sum_{t=1}^{N}\Bigg\{ H\Big(\mathbf{y}_{1:N\ t};\Psi\Big)\ \int_0^1\int_0^1 \Big[\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\big[\mathbf{Y}_t\big] - \boldsymbol{\mu}^* - T_\epsilon(s_{\epsilon,t})^{-1}\boldsymbol{\delta}_\epsilon^* - T_\epsilon(s_{\epsilon,t})\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\big[\mathbf{Y}_t\big]\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T} \\
&\qquad + T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T} + \Big(\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}_t^{-1}\mathbf{W}^{*T}\Big] m(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\Bigg\} = \mathbf{0} \\
&\implies \mathbf{A}_5(\mathbf{y}_{1:N};\Psi) - A_0(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^* - A_2(\mathbf{y}_{1:N};\Psi)\boldsymbol{\delta}_\epsilon^* - \mathbf{A}_8(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T} + \boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T} + \Big(\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\Big)\mathbf{A}_4(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T} = \mathbf{0} \\
&\implies \boldsymbol{\delta}_\epsilon^* = A_2(\mathbf{y}_{1:N};\Psi)^{-1}\Big[\mathbf{A}_5(\mathbf{y}_{1:N};\Psi) - A_0(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^* - \mathbf{A}_8(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T} + \boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T} + \Big(\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\Big)\mathbf{A}_4(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\Big]
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\partial \tilde{v}(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi,\Psi^*)}{\partial \boldsymbol{\delta}_x^*} = -T_x(s_{x,t})^{-1}\boldsymbol{\delta}_x^* + T_\epsilon(s_{\epsilon,t})\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\big[\mathbf{Y}_t\big]\mathbf{W}\mathbf{M}_t^{-1} - T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W}\mathbf{M}_t^{-1} - \big(\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\big)\mathbf{M}_t^{-1} \\
&\implies \frac{\partial \tilde{Q}_{GStS,ind,2}(\Psi,\Psi^*)}{\partial \boldsymbol{\delta}_x^*} = \mathbf{0} \\
&\iff \sum_{t=1}^{N}\Bigg\{ H\big(\mathbf{y}_{1:N\ t};\Psi\big)\ \int_0^1\int_0^1 \Big[- T_x(s_{x,t})^{-1}\boldsymbol{\delta}_x^* + T_\epsilon(s_{\epsilon,t})\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\big[\mathbf{Y}_t\big]\mathbf{W}\mathbf{M}_t^{-1} - T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W}\mathbf{M}_t^{-1} - \big(\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\big)\mathbf{M}_t^{-1}\Big] \\
&\quad \times m(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\Bigg\} = \mathbf{0} \\
&\implies -A_3(\mathbf{y}_{1:N}^o;\Psi)\boldsymbol{\delta}_x^* + \mathbf{A}_8(\mathbf{y}_{1:N}^o;\Psi) - \boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}_{1:N}^o;\Psi) - \big(\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\big)\mathbf{A}_4(\mathbf{y}_{1:N}^o;\Psi) = \mathbf{0} \\
&\implies \boldsymbol{\delta}_x^* = A_3(\mathbf{y}_{1:N}^o;\Psi)^{-1}\Big(\mathbf{A}_8(\mathbf{y}_{1:N};\Psi) - \boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}_{1:N}^o;\Psi) - \big(\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\big)\mathbf{A}_4(\mathbf{y}_{1:N}^o;\Psi)\Big)
\end{aligned}
$$

$$
\frac{\partial \tilde{v}(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi,\Psi^*)}{\partial \sigma^{*2}} = -\frac{d}{2\sigma^{*2}} + \frac{1}{2(\sigma^{*2})^2}\Big[T_\epsilon(s_{\epsilon,t})\,\mathrm{Tr}\Big\{\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\big[\mathbf{Y}_t^T\mathbf{Y}_t\big]\Big\} - 2\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\big[\mathbf{Y}_t\big]\Big(T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}^{*T} + \boldsymbol{\delta}_\epsilon^{*T}\Big)
$$

$$
\begin{aligned}
&+ T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}^*\boldsymbol{\mu}^{*T} + 2\boldsymbol{\mu}^*\boldsymbol{\delta}_\epsilon^{*T} + T_\epsilon(s_{\epsilon,t})^{-1}\boldsymbol{\delta}_\epsilon^*\boldsymbol{\delta}_\epsilon^{*T} - 2T_\epsilon(s_{\epsilon,t})^2 \operatorname{Tr}\left\{\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U}_t,\mathbf{V}_t,\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right]\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}\right\}\\
&+ 2T_\epsilon(s_{\epsilon,t})\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}\Big(T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}^{*T} + \boldsymbol{\delta}_\epsilon^{*T}\Big)\\
&- 2T_\epsilon(s_{\epsilon,t})\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]\mathbf{W}^*\mathbf{M}_t^{-1}\Big(-\boldsymbol{\mu}T_\epsilon(s_{\epsilon,t})\mathbf{W} - \boldsymbol{\delta}_\epsilon\mathbf{W} + \sigma^2\boldsymbol{\delta}_x\Big)^T\\
&+ 2\Big(-T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W} - \boldsymbol{\delta}_\epsilon\mathbf{W} + \sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}_t^{-1}\mathbf{W}^{*T}\Big(T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}^{*T} + \boldsymbol{\delta}_\epsilon^{*T}\Big)\\
&+ T_\epsilon(s_{\epsilon,t})^3 \operatorname{Tr}\left\{\mathbf{M}_t^{-1}\mathbf{W}^T\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right]\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{W}^*\right\}\\
&+ T_\epsilon(s_{\epsilon,t})^2 \operatorname{Tr}\left\{\mathbf{M}_t^{-1}\mathbf{W}^T\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]^T\Big(-T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W} - \boldsymbol{\delta}_\epsilon\mathbf{W} + \sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{W}^*\right\}\\
&+ T_\epsilon(s_{\epsilon,t})^2 \operatorname{Tr}\left\{\mathbf{M}_t^{-1}\Big(-T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W} - \boldsymbol{\delta}_\epsilon\mathbf{W} + \sigma^2\boldsymbol{\delta}_x\Big)^T\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]\mathbf{W}^T\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{W}^*\right\}\\
&+ T_\epsilon(s_{\epsilon,t}) \operatorname{Tr}\left\{\mathbf{M}_t^{-1}\Big(-T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W} - \boldsymbol{\delta}_\epsilon\mathbf{W} + \sigma^2\boldsymbol{\delta}_x\Big)^T\Big(-T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W} - \boldsymbol{\delta}_\epsilon\mathbf{W} + \sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{W}^*\right\}\\
&+ \sigma^2 T_\epsilon(s_{\epsilon,t}) \operatorname{Tr}\left\{\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{W}^*\right\}\Bigg]\\
\implies &\frac{\partial\tilde{Q}_{GStS,ind,2}(\Psi,\Psi^*)}{\partial\sigma^{*2}} = \mathbf{0}\\
\iff &-\frac{d}{2\sigma^{*2}}A_0(\mathbf{y}_{1:N}^o;\Psi) + \frac{1}{2\left(\sigma^{*2}\right)^2}\Bigg[A_{20}(\mathbf{y}_{1:N}^o;\Psi) - 2\mathbf{A}_6(\mathbf{y}_{1:N}^o;\Psi)\boldsymbol{\mu}^{*T} - 2\mathbf{A}_5(\mathbf{y}_{1:N}^o;\Psi)\boldsymbol{\delta}_\epsilon^{*T}\\
&+ A_1(\mathbf{y}_{1:N}^o;\Psi)\boldsymbol{\mu}^*\boldsymbol{\mu}^{*T} + 2A_0(\mathbf{y}_{1:N}^o;\Psi)\boldsymbol{\mu}^*\boldsymbol{\delta}_\epsilon^{*T} + A_2(\mathbf{y}_{1:N}^o;\Psi)\boldsymbol{\delta}_\epsilon^*\boldsymbol{\delta}_\epsilon^{*T} - 2\operatorname{Tr}\left\{\mathbf{A}_{14}(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}\right\} + 2\mathbf{A}_{10}(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}\boldsymbol{\mu}^{*T}\\
&+ 2\mathbf{A}_8(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}\boldsymbol{\delta}_\epsilon^{*T} + 2\mathbf{A}_9(\mathbf{y}_{1:N}^o;\Psi,\Psi^*)\mathbf{W}^T\boldsymbol{\mu}^T + 2\mathbf{A}_7(\mathbf{y}_{1:N}^o;\Psi,\Psi)\mathbf{W}^T\boldsymbol{\delta}_\epsilon^T - 2\sigma^2\mathbf{A}_7(\mathbf{y}_{1:N}^o;\Psi,\Psi)\boldsymbol{\delta}_x^T - 2\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{13}(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}\boldsymbol{\mu}^{*T}\\
&- 2\left(\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\right)\mathbf{A}_{11}(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}\boldsymbol{\mu}^{*T} - 2\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}\boldsymbol{\delta}_\epsilon^{*T} - 2\left(\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\right)\mathbf{A}_4(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}\boldsymbol{\delta}_\epsilon^{*T} + \operatorname{Tr}\left\{\mathbf{A}_{17}(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}\mathbf{W}^*\right\}\\
&- \operatorname{Tr}\left\{\Big(\mathbf{A}_{18}(\mathbf{y}_{1:N}^o;\Psi) + \mathbf{A}_{18}(\mathbf{y}_{1:N}^o;\Psi)^T\Big)\mathbf{W}^{*T}\mathbf{W}^*\right\} + \operatorname{Tr}\left\{\mathbf{A}_{19}(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}\mathbf{W}^*\right\} + \sigma^2\operatorname{Tr}\left\{\mathbf{A}_{11}(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}\mathbf{W}^*\right\}\Bigg] = 0\\
\iff &\sigma^{*2} = \frac{1}{dA_0(\mathbf{y}_{1:N}^o;\Psi)}\Bigg[A_{20}(\mathbf{y}_{1:N}^o;\Psi) - 2\mathbf{A}_6(\mathbf{y}_{1:N}^o;\Psi)\boldsymbol{\mu}^{*T} - 2\mathbf{A}_5(\mathbf{y}_{1:N}^o;\Psi)\boldsymbol{\delta}_\epsilon^{*T} + A_1(\mathbf{y}_{1:N}^o;\Psi)\boldsymbol{\mu}^*\boldsymbol{\mu}^{*T} + 2A_0(\mathbf{y}_{1:N}^o;\Psi)\boldsymbol{\mu}^*\boldsymbol{\delta}_\epsilon^{*T}\\
&+ A_2(\mathbf{y}_{1:N}^o;\Psi)\boldsymbol{\delta}_\epsilon^*\boldsymbol{\delta}_\epsilon^{*T} - 2\operatorname{Tr}\left\{\mathbf{A}_{14}(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}\right\} + 2\mathbf{A}_9(\mathbf{y}_{1:N}^o;\Psi,\Psi^*)\mathbf{W}^T\boldsymbol{\mu}^T + 2\mathbf{A}_7(\mathbf{y}_{1:N}^o;\Psi,\Psi)\mathbf{W}^T\boldsymbol{\delta}_\epsilon^T\\
&- 2\sigma^2\mathbf{A}_7(\mathbf{y}_{1:N}^o;\Psi,\Psi)\boldsymbol{\delta}_x^T + 2\Big(\mathbf{A}_{10}(\mathbf{y}_{1:N}^o;\Psi) - \boldsymbol{\mu}\mathbf{W}\mathbf{A}_{13}(\mathbf{y}_{1:N}^o;\Psi) - \left(\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\right)\mathbf{A}_{11}(\mathbf{y}_{1:N}^o;\Psi)\Big)\mathbf{W}^{*T}\boldsymbol{\mu}^{*T}\\
&+ 2\Big(\mathbf{A}_8(\mathbf{y}_{1:N}^o;\Psi) - \boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}_{1:N}^o;\Psi) - \left(\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\right)\mathbf{A}_4(\mathbf{y}_{1:N}^o;\Psi)\Big)\mathbf{W}^{*T}\boldsymbol{\delta}_\epsilon^{*T} + \operatorname{Tr}\left\{\mathbf{A}_{17}(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}\mathbf{W}^*\right\}\\
&- \operatorname{Tr}\left\{\Big(\mathbf{A}_{18}(\mathbf{y}_{1:N}^o;\Psi) + \mathbf{A}_{18}(\mathbf{y}_{1:N}^o;\Psi)^T\Big)\mathbf{W}^{*T}\mathbf{W}^*\right\} + \operatorname{Tr}\left\{\mathbf{A}_{19}(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}\mathbf{W}^*\right\} + \sigma^2\operatorname{Tr}\left\{\mathbf{A}_{11}(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}\mathbf{W}^*\right\}\Bigg]
\end{aligned}
$$

for

$$
\begin{aligned}
A_0(\mathbf{y}_{1:N}^o;\Psi)_{1\times 1} &:= \sum_{t=1}^N\left\{H\Big(\mathbf{y}_{1:N/t}^o;\Psi\Big)\int_0^1\int_0^1 m(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\right\}\\
A_1(\mathbf{y}_{1:N}^o;\Psi)_{1\times 1} &:= \sum_{t=1}^N\left\{H\Big(\mathbf{y}_{1:N/t}^o;\Psi\Big)\int_0^1\int_0^1 T_\epsilon(s_{\epsilon,t})\,m(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\right\}\\
A_2(\mathbf{y}_{1:N}^o;\Psi)_{1\times 1} &:= \sum_{t=1}^N\left\{H\Big(\mathbf{y}_{1:N/t}^o;\Psi\Big)\int_0^1\int_0^1 T_\epsilon(s_{\epsilon,t})^{-1}\,m(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\right\}\\
A_3(\mathbf{y}_{1:N}^o;\Psi)_{1\times 1} &:= \sum_{t=1}^N\left\{H\Big(\mathbf{y}_{1:N/t}^o;\Psi\Big)\int_0^1\int_0^1 T_x(s_{x,t})^{-1}\,m(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\right\}\\
\mathbf{A}_4(\mathbf{y}_{1:N}^o;\Psi)_{k\times k} &:= \sum_{t=1}^N\left\{H\Big(\mathbf{y}_{1:N/t}^o;\Psi\Big)\int_0^1\int_0^1 \mathbf{M}_t^{-1}\,m(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\right\}\\
\mathbf{A}_5(\mathbf{y}_{1:N}^o;\Psi)_{1\times d} &:= \sum_{t=1}^N\left\{H\Big(\mathbf{y}_{1:N/t}^o;\Psi\Big)\int_0^1\int_0^1 \mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]\,m(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\right\}\\
\mathbf{A}_6(\mathbf{y}_{1:N}^o;\Psi)_{1\times d} &:= \sum_{t=1}^N\left\{H\Big(\mathbf{y}_{1:N/t}^o;\Psi\Big)\int_0^1\int_0^1 T_\epsilon(s_{\epsilon,t})\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]\,m(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\right\}\\
\mathbf{A}_7(\mathbf{y}_{1:N}^o;\Psi,\Psi^*)_{1\times k} &:= \sum_{t=1}^N\left\{H\Big(\mathbf{y}_{1:N/t}^o;\Psi\Big)\int_0^1\int_0^1 T_\epsilon(s_{\epsilon,t})\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]\mathbf{W}^*\mathbf{M}_t^{-1}\,m(\mathbf{y}_t^o,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\right\}
\end{aligned}
$$

$$\mathbf{A}_8(\mathbf{y}^o_{1:N};\Psi)_{1\times k} := \sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\int_0^1\int_0^1 T_\epsilon(s_{\epsilon,t})\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]\mathbf{W}\mathbf{M}_t^{-1}\, m(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\right\}$$

$$\mathbf{A}_9(\mathbf{y}^o_{1:N};\Psi,\Psi^*)_{1\times d} := \sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\int_0^1\int_0^1 T_\epsilon(s_{\epsilon,t})^2\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]\mathbf{W}^*\mathbf{M}_t^{-1}\, m(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\right\}$$

$$\mathbf{A}_{10}(\mathbf{y}^o_{1:N};\Psi)_{1\times d} := \sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\int_0^1\int_0^1 T_\epsilon(s_{\epsilon,t})^2\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]\mathbf{W}\mathbf{M}_t^{-1}\, m(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\right\}$$

$$\mathbf{A}_{11}(\mathbf{y}^o_{1:N};\Psi)_{k\times k} := \sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\int_0^1\int_0^1 T_\epsilon(s_{\epsilon,t})\mathbf{M}_t^{-1}\, m(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\right\}$$

$$\mathbf{A}_{13}(\mathbf{y}^o_{1:N};\Psi)_{k\times k} := \sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\int_0^1\int_0^1 T_\epsilon(s_{\epsilon,t})^2\mathbf{M}_t^{-1}\, m(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\right\}$$

$$\mathbf{A}_{14}(\mathbf{y}^o_{1:N};\Psi)_{d\times k} := \sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\int_0^1\int_0^1 T_\epsilon(s_{\epsilon,t})^2\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right]\mathbf{W}\mathbf{M}_t^{-1}\, m(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\right\}$$

$$\begin{aligned}\mathbf{A}_{16}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)_{d\times k} := \sum_{t=1}^{N}\Bigg\{&H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\int_0^1\int_0^1 T_\epsilon(s_{\epsilon,t})\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]^T\Big(T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W}+\delta_\epsilon\mathbf{W}-\sigma^2\delta_x\Big)\mathbf{M}_t^{-1}\\&\times m(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\Bigg\}\end{aligned}$$

$$\mathbf{A}_{17}(\mathbf{y}^o_{1:N};\Psi)_{k\times k} := \sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\int_0^1\int_0^1 T_\epsilon(s_{\epsilon,t})^3\mathbf{M}_t^{-1}\mathbf{W}^T\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right]\mathbf{W}\mathbf{M}_t^{-1}\, m(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\right\}$$

$$\begin{aligned}\mathbf{A}_{18}(\mathbf{y}^o_{1:N};\Psi)_{k\times k} := \sum_{t=1}^{N}\Bigg\{&H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\int_0^1\int_0^1 T_\epsilon(s_{\epsilon,t})^2\mathbf{M}_t^{-1}\mathbf{W}^T\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t\right]^T\Big(T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W}+\delta_\epsilon\mathbf{W}-\sigma^2\delta_x\Big)\mathbf{M}_t^{-1}\\&\times m(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\Bigg\}\end{aligned}$$

$$\begin{aligned}\mathbf{A}_{19}(\mathbf{y}^o_{1:N};\Psi)_{k\times k} := \sum_{t=1}^{N}\Bigg\{&H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\int_0^1\int_0^1 T_\epsilon(s_{\epsilon,t})\mathbf{M}_t^{-1}\Big(T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W}+\delta_\epsilon\mathbf{W}-\sigma^2\delta_x\Big)^T\Big(T_\epsilon(s_{\epsilon,t})\boldsymbol{\mu}\mathbf{W}+\delta_\epsilon\mathbf{W}-\sigma^2\delta_x\Big)\mathbf{M}_t^{-1}\\&\times m(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\Bigg\}\end{aligned}$$

$$A_{20}(\mathbf{y}^o_{1:N};\Psi)_{1\times 1} := \sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\int_0^1\int_0^1 T_\epsilon(s_{\epsilon,t})\,\mathrm{Tr}\left\{\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right]\right\}\, m(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\right\}$$

$$A_{21}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)_{1\times 1} := \sum_{t=1}^{N}\left\{H\left(\mathbf{y}^o_{1:N/t};\Psi\right)\int_0^1\int_0^1 T_\epsilon(s_{\epsilon,t})^2\,\mathrm{Tr}\left\{\mathbb{E}_{\mathbf{Y}^m_t|\mathbf{Y}^o_t,S_{\epsilon,t},S_{x,t},\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right]\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}\right\}\, m(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi)\,ds_{\epsilon,t}\,ds_{x,t}\right\}.$$

Hence, the system of equations $\nabla_{\Psi^*}\tilde{Q}_{GStS,ind,2}(\Psi,\Psi^*)=\mathbf{0}$ is the following

$$\begin{cases}
\boldsymbol{\mu}^* = A_1(\mathbf{y}_{1:N};\Psi)^{-1}\Big[\mathbf{A}_6(\mathbf{y}_{1:N};\Psi)-A_0(\mathbf{y}_{1:N};\Psi)\delta^*_\epsilon-\mathbf{A}_{10}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}+\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{13}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}+(\delta_\epsilon\mathbf{W}+\sigma^2\delta_x)\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\Big]\\
\delta^*_\epsilon = A_2(\mathbf{y}_{1:N};\Psi)^{-1}\Big[\mathbf{A}_5(\mathbf{y}_{1:N};\Psi)-A_0(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^*-\mathbf{A}_8(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}+\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}+\Big(\delta_\epsilon\mathbf{W}-\sigma^2\delta_x\Big)\mathbf{A}_4(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\Big]\\
\delta^*_x = A_3(\mathbf{y}^o_{1:N};\Psi)^{-1}\Big(\mathbf{A}_8(\mathbf{y}_{1:N};\Psi)-\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}^o_{1:N};\Psi)-(\delta_\epsilon\mathbf{W}-\sigma^2\delta_x)\mathbf{A}_4(\mathbf{y}^o_{1:N};\Psi)\Big)\\
\mathbf{W}^* = \Big[\mathbf{A}_{14}(\mathbf{y}_{1:N};\Psi)-\boldsymbol{\mu}^{*T}\mathbf{A}_{10}(\mathbf{y}_{1:N};\Psi)-\delta^*_\epsilon\mathbf{A}_8(\mathbf{y}_{1:N};\Psi)-\mathbf{A}_{16}(\mathbf{y}_{1:N};\Psi)+\boldsymbol{\mu}^{*T}\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{13}(\mathbf{y}_{1:N};\Psi)+\boldsymbol{\mu}^{*T}(\delta_\epsilon-\sigma^2\delta_x)\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi)\\
\quad+\delta^{*T}_\epsilon\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi)+\delta^{*T}_\epsilon(\delta_\epsilon\mathbf{W}-\sigma^2\delta_x)\mathbf{A}_4(\mathbf{y}_{1:N};\Psi)\Big]\\
\quad\Big[\mathbf{A}_{17}(\mathbf{y}_{1:N};\Psi)-\mathbf{A}_{18}(\mathbf{y}_{1:N};\Psi)-\mathbf{A}_{18}(\mathbf{y}_{1:N};\Psi)^T+\mathbf{A}_{19}(\mathbf{y}_{1:N};\Psi)+\sigma^2\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi)\Big]^{-1}\\
\sigma^{*2} = \frac{1}{dA_0(\mathbf{y}^o_{1:N};\Psi)}\Big[A_{20}(\mathbf{y}^o_{1:N};\Psi)-2\mathbf{A}_6(\mathbf{y}^o_{1:N};\Psi)\boldsymbol{\mu}^{*T}-2\mathbf{A}_5(\mathbf{y}^o_{1:N};\Psi)\delta^{*T}_\epsilon+A_1(\mathbf{y}^o_{1:N};\Psi)\boldsymbol{\mu}^*\boldsymbol{\mu}^{*T}+2A_0(\mathbf{y}^o_{1:N};\Psi)\boldsymbol{\mu}^*\delta^{*T}_\epsilon\\
\quad+A_2(\mathbf{y}^o_{1:N};\Psi)\delta^*_\epsilon\delta^{*T}_\epsilon-2\,\mathrm{Tr}\Big\{\mathbf{A}_{14}(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}^{*T}\Big\}+2\mathbf{A}_9(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\mathbf{W}^T\boldsymbol{\mu}^T+2\mathbf{A}_7(\mathbf{y}^o_{1:N};\Psi,\Psi)\mathbf{W}^T\delta^T_\epsilon\\
\quad-2\sigma^2\mathbf{A}_7(\mathbf{y}^o_{1:N};\Psi,\Psi)\delta^T_x+2\Big(\mathbf{A}_{10}(\mathbf{y}^o_{1:N};\Psi)-\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{13}(\mathbf{y}^o_{1:N};\Psi)-(\delta_\epsilon\mathbf{W}-\sigma^2\delta_x)\mathbf{A}_{11}(\mathbf{y}^o_{1:N};\Psi)\Big)\mathbf{W}^{*T}\boldsymbol{\mu}^{*T}\\
\quad+2\Big(\mathbf{A}_8(\mathbf{y}^o_{1:N};\Psi)-\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}^o_{1:N};\Psi)-(\delta_\epsilon\mathbf{W}-\sigma^2\delta_x)\mathbf{A}_4(\mathbf{y}^o_{1:N};\Psi)\Big)\mathbf{W}^{*T}\delta^{*T}_\epsilon+\mathrm{Tr}\Big\{\mathbf{A}_{17}(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}^{*T}\mathbf{W}^*\Big\}\\
\quad-\mathrm{Tr}\Big\{\Big(\mathbf{A}_{18}(\mathbf{y}^o_{1:N};\Psi)+\mathbf{A}_{18}(\mathbf{y}^o_{1:N};\Psi)^T\Big)\mathbf{W}^{*T}\mathbf{W}^*\Big\}+\mathrm{Tr}\Big\{\mathbf{A}_{19}(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}^{*T}\mathbf{W}^*\Big\}+\sigma^2\,\mathrm{Tr}\Big\{\mathbf{A}_{11}(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}^{*T}\mathbf{W}^*\Big\}\Big]
\end{cases}$$

□

A.4.3. ID: Complete Data Setting

A.4.3.1. Proof of Theorem 15

Proof. The joint probability function of the variables $\mathbf{Y}_{1:N}$, $\mathbf{X}_{1:N}$, $\mathbf{U}$ and $\mathbf{V}$ for Generalized Skew t-Student ID PPCA model is specified in Equation (3.32). Given the conditional time independence of the random vectors $\mathbf{Y}_t$, $\mathbf{X}_t$, $\mathbf{U}$ and $\mathbf{V}$, the E-step of the corresponding EM algorithm can be formalized as follows

$$
\begin{aligned}
Q_{GStS,id}(\Psi,\Psi^*) &= \mathbb{E}_{\mathbf{X}_{1:N},\mathbf{U},\mathbf{V}|\mathbf{Y}_{1:N},\Psi}\Big[\log \pi_{\mathbf{Y}_{1:N},\mathbf{X}_{1:N},\mathbf{U},\mathbf{V}|\Psi^*}(\mathbf{Y}_{1:N},\mathbf{X}_{1:N},\mathbf{U},\mathbf{V})\Big] \\
&= \frac{1}{\pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N})}\int_{\mathbb{R}_+^d}\int_{\mathbb{R}_+^k}\left\{\log\left(\pi_{\mathbf{U},\mathbf{V}|\Psi^*}(\mathbf{u},\mathbf{v})\right)\pi_{\mathbf{U},\mathbf{V}|\Psi}(\mathbf{u},\mathbf{v})\int_{\mathbb{R}^{N\times k}}\prod_{s=1}^{N}\pi_{\mathbf{Y}_s,\mathbf{X}_s|\mathbf{U},\mathbf{V},\Psi}(\mathbf{y}_s,\mathbf{x}_s)d\mathbf{x}_{1:N}\right\}d\mathbf{u}\,d\mathbf{v} \\
&+ \frac{1}{\pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N})}\int_{\mathbb{R}_+^d}\int_{\mathbb{R}_+^k}\sum_{t=1}^{N}\int_{\mathbb{R}^{N\times k}}\left\{\log\left(\pi_{\mathbf{Y}_t,\mathbf{X}_t|\mathbf{U},\mathbf{V},\Psi^*}(\mathbf{y}_t,\mathbf{x}_t,)\right)\pi_{\mathbf{U},\mathbf{V}|\Psi}(\mathbf{u},\mathbf{v})\prod_{s=1}^{N}\pi_{\mathbf{Y}_s,\mathbf{X}_s|\mathbf{U},\mathbf{V},\Psi}(\mathbf{y}_s,\mathbf{x}_s)\right\}d\mathbf{x}_{1:N}\,d\mathbf{u}\,d\mathbf{v}.
\end{aligned}
$$

Using Lemma A12, the probability function $\pi_{\mathbf{Y}_t,\mathbf{X}_t|\mathbf{U},\mathbf{V},\Psi}(\mathbf{y}_t,\mathbf{x}_t)$ can be factorised into the product of three functions

$$
\pi_{\mathbf{Y}_t,\mathbf{X}_t|\mathbf{U},\mathbf{V},\Psi}(\mathbf{y}_t,\mathbf{x}_t) = \pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U},\mathbf{V},\Psi}(\mathbf{x}_t)\cdot\pi_{\mathbf{Y}_t|\mathbf{U},\mathbf{V},\Psi}(\mathbf{y}_t)\cdot C(\mathbf{u},\mathbf{v};\Psi)
$$

Please refer to the result derived in Lemma A12 to find the definitions of the probability functions and the function $C(\mathbf{u},\mathbf{v};\Psi):\mathbb{R}_+^d\times\mathbb{R}_+^k\longrightarrow\mathbb{R}$. Therefore, we can further simplify the formulation of the $Q_{GStS,id}$ function

$$
\begin{aligned}
Q_{GStS,id}(\Psi,\Psi^*) &= \frac{1}{\pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N})}\int_{\mathbb{R}_+^d}\int_{\mathbb{R}_+^k}\left\{\log\left(\pi_{\mathbf{U},\mathbf{V}|\Psi^*}(\mathbf{u},\mathbf{v})\right)\pi_{\mathbf{U},\mathbf{V}|\Psi}(\mathbf{u},\mathbf{v})\,C(\mathbf{u},\mathbf{v};\Psi)^N\right. \\
&\left.\times\prod_{s=1}^{N}\pi_{\mathbf{Y}_s|\mathbf{U},\mathbf{V},\Psi}(\mathbf{y}_s)\underbrace{\prod_{s=1}^{N}\int_{\mathbb{R}^k}\pi_{\mathbf{X}_s|\mathbf{Y}_s,\mathbf{U},\mathbf{V},\Psi}(\mathbf{x}_s)d\mathbf{x}_s}_{=1}\right\}d\mathbf{u}\,d\mathbf{v} \\
&+ \frac{1}{\pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N})}\int_{\mathbb{R}_+^d}\int_{\mathbb{R}_+^k}\sum_{t=1}^{N}\left\{\int_{\mathbb{R}^{k\times N}}\log\left(\pi_{\mathbf{Y}_t,\mathbf{X}_t|\mathbf{U},\mathbf{V},\Psi^*}(\mathbf{y}_t,\mathbf{x}_t)\right)\prod_{s=1}^{N}\pi_{\mathbf{X}_s|\mathbf{Y}_s,\mathbf{U},\mathbf{V},\Psi}(\mathbf{x}_s)d\mathbf{x}_{1:N}\right. \\
&\left.\times\,\pi_{\mathbf{U},\mathbf{V}|\Psi}(\mathbf{u},\mathbf{v})\,C(\mathbf{u},\mathbf{v};\Psi)^N\prod_{s=1}^{N}\pi_{\mathbf{Y}_s|\mathbf{U},\mathbf{V},\Psi}(\mathbf{y}_s)\right\}d\mathbf{u}\,d\mathbf{v} \\
&= \frac{1}{\pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N})}\int_{\mathbb{R}_+^d}\int_{\mathbb{R}_+^k}\left\{\log\left(\pi_{\mathbf{U},\mathbf{V}|\Psi^*}(\mathbf{u},\mathbf{v})\right)\pi_{\mathbf{U},\mathbf{V}|\Psi}(\mathbf{u},\mathbf{v})\,C(\mathbf{u},\mathbf{v};\Psi)^N\prod_{s=1}^{N}\pi_{\mathbf{Y}_s|\mathbf{U},\mathbf{V},\Psi}(\mathbf{y}_s)\right\}d\mathbf{u}\,d\mathbf{v} \\
&+ \frac{1}{\pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N})}\int_{\mathbb{R}_+^d}\int_{\mathbb{R}_+^k}\left\{\pi_{\mathbf{U},\mathbf{V}|\Psi}(\mathbf{u},\mathbf{v})\,C(\mathbf{u},\mathbf{v};\Psi)^N\prod_{s=1}^{N}\pi_{\mathbf{Y}_s|\mathbf{U},\mathbf{V},\Psi}(\mathbf{y}_s)\sum_{t=1}^{N}\underbrace{\int_{\mathbb{R}^k}\log\left(\pi_{\mathbf{Y}_t,\mathbf{X}_t|\mathbf{U},\mathbf{V},\Psi^*}(\mathbf{y}_t,\mathbf{x}_t)\right)\pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U},\mathbf{V},\Psi}(\mathbf{x}_t)\,d\mathbf{x}_t}_{w(\mathbf{y}_t,\mathbf{u},\mathbf{v};\Psi,\Psi^*)}\right\}d\mathbf{u}\,d\mathbf{v}
\end{aligned}
$$

for the function $w:\mathbb{R}^d\times\mathbb{R}_+^d\times\mathbb{R}_+^k\longrightarrow\mathbb{R}$ defined in Lemma A13. Let us denote for the convenience of the notation

$$
\begin{aligned}
h(\mathbf{y}_{1:N},\mathbf{u},\mathbf{v},\Psi) &:= \pi_{\mathbf{U},\mathbf{V}|\Psi}(\mathbf{u},\mathbf{v})\,C(\mathbf{u},\mathbf{v};\Psi)^N\left(\prod_{s=1}^{N}\pi_{\mathbf{Y}_s|\mathbf{U},\mathbf{V},\Psi}(\mathbf{y}_s)\right) = \pi_{\mathbf{U},\mathbf{V}|\Psi}(\mathbf{u},\mathbf{v})\left(\prod_{s=1}^{N}\pi_{\mathbf{Y}_s|\mathbf{U},\mathbf{V},\Psi}(\mathbf{y}_s)\right) \\
&\times(\sigma^2)^{\frac{kN}{2}}\left|\mathbf{D}_\epsilon\right|^{\frac{N}{2}}\left|\mathbf{D}_{x,t}\right|^{\frac{N}{2}}\left|\mathbf{M}\right|^{-\frac{N}{2}}\left|\mathbf{N}\right|^{-\frac{N}{2}}\exp\left\{-\frac{N}{2}\delta_x\left(\mathbf{D}_x^{-1}-\sigma^2\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\mathbf{N}^{-1}\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}-\sigma^2\mathbf{M}^{-1}\right)\delta_x^T\right\}
\end{aligned}
$$

the E-step of the Generalized Skew t-Student IND PPCA is the following

$$
Q_{GStS,id}(\Psi,\Psi^*) = \frac{1}{\pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N})}\int_{\mathbb{R}_+^d}\int_{\mathbb{R}_+^k}\left\{\left(\log\left(\pi_{\mathbf{U},\mathbf{V}|\Psi^*}(\mathbf{u},\mathbf{v})+\sum_{t=1}^{N}w(\mathbf{y}_t,\mathbf{u},\mathbf{v};\Psi,\Psi^*)\right)h(\mathbf{y}_{1:N},\mathbf{u},\mathbf{v};\Psi)\right\}d\mathbf{u}\,d\mathbf{v}. \quad \text{(A15)}
$$

Let us recall the definition the quantile functions $T_\epsilon:[0,1]\to\mathbb{R}_+^d$ and $T_x:[0,1]\to\mathbb{R}_+^k$ from Equation (A9). Following the same steps as in the proof of Theorem 15, we reduce the $d\times k$ dimensional integration problems to the 2-dimensional integrations on the unit hypercube $[0,1]^2$.

Let us denote the variables $S_\epsilon = T_\epsilon^{-1}(\mathbf{U})$ and $S_x = T_x^{-1}(\mathbf{V})$ in Equation (A1) and remark that

$$
\pi_{\mathbf{U}|\Psi}(\mathbf{u}) = \pi_{S_\epsilon|\Psi}(s_\epsilon)\left|\frac{\partial T_\epsilon(s_\epsilon)}{\partial s_\epsilon}\right|^{-1} = \mathbf{1}_{[0,1]}(s_\epsilon)\left|\frac{\partial T_\epsilon(s_\epsilon)}{\partial s_\epsilon}\right|^{-1}
$$

$$\pi_{\mathbf{V}|\Psi}(\mathbf{v}) = \pi_{S_x|\Psi}(s_x)\left|\frac{\partial T_x(s_x)}{\partial s_x}\right|^{-1} = \mathbf{1}_{[0,1]}(s_x)\left|\frac{\partial T_x(s_x)}{\partial s_x}\right|^{-1} \tag{A16}$$

and that $\pi_{\mathbf{Y}_t|T_\epsilon(S_\epsilon),T_x(S_x),\Psi}(\mathbf{y}_t) = \pi_{\mathbf{Y}_t|S_\epsilon,S_x,\Psi}(\mathbf{y}_t)$. We redefine the matrices $\mathbf{D}_{\epsilon,t}$ and $\mathbf{D}_{x,t}$ using the transformations $\mathbf{U} = T_\epsilon(S_\epsilon)$ and $\mathbf{V} = T_x(S_x)$ such that

$$\mathbf{D}_\epsilon = \begin{pmatrix} \frac{\chi^{-1}_{\nu^1_\epsilon}(S_\epsilon)}{\nu^1_\epsilon} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \frac{\chi^{-1}_{\nu^d_\epsilon}(S_\epsilon)}{\nu^d_\epsilon} \end{pmatrix}_{d\times d}, \; \mathbf{D}_{x,t} = \begin{pmatrix} \frac{\chi^{-1}_{\nu^1_x}(S_x)}{\nu^1_x} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \frac{\chi^{-1}_{\nu^k_x}(S_x)}{\nu^k_x} \end{pmatrix}_{k\times k}.$$

Hence, the E-step of the EM algorithm for the Generalized Skew t-Student ID PPCA from Equation (A15) can be expressed by the sum of 2-dimensional integration problems

$$Q_{GStS,id}(\Psi,\Psi^*) = \frac{1}{\pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N})}\int_0^1\int_0^1\left\{\left(-\log\left|\frac{\partial T_\epsilon(s_\epsilon)}{\partial s_\epsilon}\right| - \log\left|\frac{\partial T_x(s_x)}{\partial s_x}\right| + \sum_{t=1}^N \tilde{w}(\mathbf{y}_t,s_\epsilon,s_x;\Psi,\Psi^*)\right) m(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)\right\} ds_\epsilon\, ds_x$$

for the function $m : \mathbb{R}^{dN}\times[0,1]^2 \longrightarrow \mathbb{R}$

$$m(\mathbf{y}_t,s_\epsilon,s_x;\Psi) = \left(\prod_{t=1}^N \pi_{\mathbf{Y}_t|S_\epsilon,S_x,\Psi}(\mathbf{y}_t)\right)(\sigma^2)^{\frac{k}{2}}\left|\mathbf{D}_\epsilon\right|^{\frac{N}{2}}\left|\mathbf{D}_x\right|^{\frac{N}{2}}\left|\mathbf{M}\right|^{-\frac{N}{2}}\left|\mathbf{N}\right|^{-\frac{N}{2}} e^{-\frac{N}{2}\delta_x\left(\mathbf{D}_x^{-1}-\sigma^2\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\mathbf{N}^{-1}\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}_t^{-1}-\sigma^2\mathbf{M}^{-1}\right)\delta_x^T}$$

and $\tilde{w}(\mathbf{y}_t,s_\epsilon,s_x;\Psi,\Psi^*) := w\Big(\mathbf{y}_t,T_\epsilon(s_\epsilon),T_x(s_x);\Psi,\Psi^*\Big)$. □

A.4.3.2. Proof of Theorem 16

Proof. The M-step of the EM algorithm for Identical and Conditionally Independent Generalized Skew t-Student PPCA evaluates the maximizers of the function $Q_{GStS,id}(\Psi,\Psi^*)$ defined in Theorem 15 with respect to the vector of static parameters $\Psi^* = \left[\boldsymbol{\mu}^*,\mathbf{W}^*,\sigma^{*2},\delta^*_\epsilon,\delta_x\right]$. Recall that the only element of the function $Q_{GStS,id}$ which contains components of the vector Ψ^* is the function $\tilde{w}$ specified in Lemma A13 being a component of the integral $I_1(\mathbf{y}_{1:N};\Psi,\Psi^*)$. Hence, the maximizers of the function $Q_{GStS,id}$ with respect to the vector of static parameters Ψ^*, which are the solution to the system of equations $\nabla_{\Psi^*}Q_{GStS,ind}(\Psi,\Psi^*) = \mathbf{0}$, are calculated as follows

$$\begin{aligned} &\frac{\partial}{\partial\Psi^*}Q_{GStS,id}(\Psi,\Psi^*) = \frac{\partial}{\partial\Psi^*}\Big(I_1(\mathbf{y}_{1:N};\Psi,\Psi^*) + I_2(\mathbf{y}_{1:N};\Psi)\Big) = \mathbf{0} \\ &\iff \int_0^1\int_0^1\left(\sum_{t=1}^N\frac{\partial}{\partial\Psi^*}\tilde{w}(\mathbf{y}_t,s_\epsilon,s_x;\Psi,\Psi^*)\right) m(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)\, ds_\epsilon\, ds_x = \mathbf{0}. \end{aligned} \tag{A17}$$

The gradient $\frac{\partial}{\partial\Psi^*}\tilde{w}(\mathbf{y}_t s_{\epsilon,t},s_{x,t};\Psi,\Psi^*)$ similar to the equivalent gradient of the function w defined in Lemma A11 and derived in the proof of Theorem 14 in Appendix A.4.1.3 for Generelized Skew t-Student IND PPCA model. Therefore, its sum over all realisations of the observation vector $\mathbf{Y}_t$ is the following determined as following

$$\begin{aligned} \sum_{t=1}^N\frac{\partial}{\partial\boldsymbol{\mu}^*}\tilde{w}(\mathbf{y}_t,s_\epsilon,s_x;\Psi,\Psi^*) &= \frac{N}{\sigma^{*2}}\Big[\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{D}_\epsilon - \boldsymbol{\mu}^*\mathbf{D}_\epsilon - \delta^*_\epsilon - \bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon + \boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon + \delta_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon \\ &\quad - \sigma^2\delta_x\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon\Big] \end{aligned}$$

$$\begin{aligned} \sum_{t=1}^N\frac{\partial}{\partial\mathbf{W}^*}\tilde{w}(\mathbf{y}_t,s_\epsilon,s_x;\Psi,\Psi^*) &= \frac{N}{\sigma^{*2}}\Big[\mathbf{D}_\epsilon\bar{\mathbb{S}}(\mathbf{y}_{1:N})\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1} - \Big(\mathbf{D}_\epsilon\boldsymbol{\mu}^{*T} + \delta^{*T}_\epsilon\Big)\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1} \\ &\quad + \Big(\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{D}_\epsilon - \boldsymbol{\mu}^*\mathbf{D}_\epsilon - \delta^*_\epsilon\Big)^T\Big(-\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W} - \delta_\epsilon\mathbf{W} + \sigma^2\delta_x\Big)\mathbf{M}^{-1} - \mathbf{D}_\epsilon\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\bar{\mathbf{S}}(\mathbf{y}_{1:N})\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1} \\ &\quad + \mathbf{D}_\epsilon\mathbf{W}^*\mathbf{M}^{-1}\Big(\mathbf{W}^T\mathbf{D}_\epsilon\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})^T\big(\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W} + \delta_\epsilon\mathbf{W} - \sigma^2\delta_x\big) + \big(\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W} + \delta_\epsilon\mathbf{W} - \sigma^2\delta_x\big)^T\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{D}_\epsilon\mathbf{W}\Big)\mathbf{M}^{-1} \\ &\quad - \mathbf{D}_\epsilon\mathbf{W}^*\mathbf{M}^{-1}\Big(\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W} + \delta_\epsilon\mathbf{W} - \sigma^2\delta_x\Big)^T\big(\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W} + \delta_\epsilon\mathbf{W} - \sigma^2\delta_x\big)\mathbf{M}^{-1} - \sigma^2\mathbf{D}_\epsilon\mathbf{W}^*\mathbf{M}^{-1}\Big] \end{aligned}$$

$$\sum_{t=1}^N\frac{\partial}{\partial\sigma^{*2}}\tilde{w}(\mathbf{y}_t,s_\epsilon,s_x;\Psi,\Psi^*) = -\frac{dN}{2\sigma^{*2}} - \frac{N}{2(\sigma^{*2})^2}\Big[-\operatorname{Tr}\Big\{\bar{\mathbf{S}}(\mathbf{y}_{1:N})\mathbf{D}_\epsilon\Big\} + 2\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\Big(\mathbf{D}_\epsilon\boldsymbol{\mu}^{*T} + \delta^{*T}_\epsilon\Big)$$

$$
\begin{aligned}
&- \boldsymbol{\mu}^* \mathbf{D}_\epsilon \boldsymbol{\mu}^{*T} - 2\boldsymbol{\mu}^* \boldsymbol{\delta}_\epsilon^{*T} - \boldsymbol{\delta}_\epsilon^* \mathbf{D}_\epsilon^{-1} \boldsymbol{\delta}_\epsilon^{*T} + 2\,\mathrm{Tr}\left\{\bar{\mathbf{S}}(\mathbf{y}_{1:N}) \mathbf{D}_\epsilon \mathbf{W} \mathbf{M}^{-1} \mathbf{W}^{*T} \mathbf{D}_\epsilon\right\} \\
&- 2\big(\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N}) \mathbf{D}_\epsilon \mathbf{W} - \boldsymbol{\mu} \mathbf{D}_\epsilon \mathbf{W} - \boldsymbol{\delta}_\epsilon \mathbf{W} + \sigma^2 \boldsymbol{\delta}_x\big) \mathbf{M}^{-1} \mathbf{W}^{*T} \big(\boldsymbol{\mu}^* \mathbf{D}_\epsilon + \boldsymbol{\delta}_\epsilon^*\big)^T \\
&+ 2\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N}) \mathbf{D}_\epsilon \mathbf{W}^* \mathbf{M}^{-1} \big(- \boldsymbol{\mu} \mathbf{D}_\epsilon \mathbf{W} - \boldsymbol{\delta}_\epsilon \mathbf{W} + \sigma^2 \boldsymbol{\delta}_x\big)^T - \mathrm{Tr}\left\{\mathbf{M}^{-1} \mathbf{W}^T \mathbf{D}_\epsilon \bar{\mathbf{S}}(\mathbf{y}_{1:N}) \mathbf{D}_\epsilon \mathbf{W} \mathbf{M}^{-1} \mathbf{W}^{*T} \mathbf{D}_\epsilon \mathbf{W}^*\right\} \\
&+ \mathrm{Tr}\left\{\mathbf{M}^{-1} \Big(\mathbf{W}^T \mathbf{D}_\epsilon \bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})^T \big(\boldsymbol{\mu} \mathbf{D}_\epsilon \mathbf{W} + \boldsymbol{\delta}_\epsilon \mathbf{W} - \sigma^2 \boldsymbol{\delta}_x\big) + \big(\boldsymbol{\mu} \mathbf{D}_\epsilon \mathbf{W} + \boldsymbol{\delta}_\epsilon \mathbf{W} - \sigma^2 \boldsymbol{\delta}_x\big)^T \bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N}) \mathbf{D}_\epsilon \mathbf{W}\Big) \mathbf{M}^{-1} \mathbf{W}^{*T} \mathbf{D}_\epsilon \mathbf{W}^*\right\} \\
&- \mathrm{Tr}\left\{\mathbf{M}^{-1} \big(\boldsymbol{\mu} \mathbf{D}_\epsilon \mathbf{W} + \boldsymbol{\delta}_\epsilon \mathbf{W} - \sigma^2 \boldsymbol{\delta}_x\big)^T \big(\boldsymbol{\mu} \mathbf{D}_\epsilon \mathbf{W} + \boldsymbol{\delta}_\epsilon \mathbf{W} \sigma^2 \boldsymbol{\delta}_x\big) \mathbf{M}^{-1} \mathbf{W}^{*T} \mathbf{D}_\epsilon \mathbf{W}^*\right\} - \sigma^2\,\mathrm{Tr}\left\{\mathbf{M}^{-1} \mathbf{W}^{*T} \mathbf{D}_\epsilon \mathbf{W}^*\right\}\Bigg]
\end{aligned}
$$

$$
\sum_{t=1}^{N} \frac{\partial}{\partial \boldsymbol{\delta}_\epsilon^*} \tilde{w}(\mathbf{y}_t, s_\epsilon, s_x; \Psi, \Psi^*) = \frac{N}{\sigma^{*2}} \Big[\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N}) - \boldsymbol{\mu}^* - \boldsymbol{\delta}_\epsilon^* \mathbf{D}_\epsilon^{-1} - \bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N}) \mathbf{D}_\epsilon \mathbf{W} \mathbf{M}^{-1} \mathbf{W}^{*T} + \boldsymbol{\mu} \mathbf{D}_\epsilon \mathbf{W} \mathbf{M}^{-1} \mathbf{W}^{*T} + \boldsymbol{\delta}_\epsilon \mathbf{W} \mathbf{M}^{-1} \mathbf{W}^{*T} - \sigma^2 \boldsymbol{\delta}_x \mathbf{M}^{-1} \mathbf{W}^{*T}\Big]
$$

$$
\sum_{t=1}^{N} \frac{\partial}{\partial \boldsymbol{\delta}_x^*} \tilde{w}(\mathbf{y}_t, s_\epsilon, s_x; \Psi, \Psi^*) = N\Big[- \boldsymbol{\delta}_x^* \mathbf{D}_x^{-1} + \bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N}) \mathbf{D}_\epsilon \mathbf{W} \mathbf{M}^{-1} - \boldsymbol{\mu} \mathbf{D}_\epsilon \mathbf{W} \mathbf{M}^{-1} - \boldsymbol{\delta}_\epsilon \mathbf{W} \mathbf{M}^{-1} + \sigma^2 \boldsymbol{\delta}_x \mathbf{M}^{-1}\Big]
$$

where $\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})_{1\times d} = \frac{1}{N}\sum_{t=1}^{N} \mathbf{y}_t$ and $\bar{\mathbf{S}}(\mathbf{y}_{1:N})_{d\times d} = \frac{1}{N}\sum_{t=1}^{N} \mathbf{y}_t^T \mathbf{y}_t$. □

A.4.3.3. Proof of Corollary 11

Proof. Given the formulation of the gradient $\nabla_{\Psi^*} Q_{GStS,id}$ from Equation (A17) and the system of equations specified in Theorem 16, the maximizers of the function $Q_{GStS,id}$ with respect to $\boldsymbol{\mu}^*$, σ^{*2}, $\boldsymbol{\delta}_\epsilon^*$ and $\boldsymbol{\delta}_x^*$ have closed form solutions defined by sequences of specific 2-dimensional integration problem determined as following

$$
\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\mu}^*} Q_{GStS,id}(\Psi, \Psi^*) = \mathbf{0} \Longleftrightarrow \int_0^1 \int_0^1 \left(\sum_{t=1}^{N} \frac{\partial}{\partial \boldsymbol{\mu}^*} \tilde{w}(\mathbf{y}_t, s_\epsilon, s_x; \Psi, \Psi^*)\right) m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x = \mathbf{0} \\
&\Longleftrightarrow \bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N}) \int_0^1 \int_0^1 \mathbf{D}_\epsilon\, m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x - \boldsymbol{\mu}^* \int_0^1 \int_0^1 \mathbf{D}_\epsilon\, m(\mathbf{y}_{1:N}^o, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x - \boldsymbol{\delta}_\epsilon^* \int_0^1 \int_0^1 m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x \\
&\quad - \bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N}) \int_0^1 \int_0^1 \mathbf{D}_\epsilon \mathbf{W} \mathbf{M}^{-1} \mathbf{W}^{*T} \mathbf{D}_\epsilon\, m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x + \boldsymbol{\mu} \int_0^1 \int_0^1 \mathbf{D}_\epsilon \mathbf{W} \mathbf{M}^{-1} \mathbf{W}^{*T} \mathbf{D}_\epsilon\, m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x \\
&\quad + \Big(\boldsymbol{\delta}_\epsilon \mathbf{W} - \sigma^2 \boldsymbol{\delta}_x\Big) \int_0^1 \int_0^1 \mathbf{M}^{-1} \mathbf{W}^{*T} \mathbf{D}_\epsilon\, m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x = \mathbf{0}
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\delta}_\epsilon^*} Q_{GStS,id}(\Psi, \Psi^*) = \mathbf{0} \Longleftrightarrow \int_0^1 \int_0^1 \left(\sum_{t=1}^{N} \frac{\partial}{\partial \boldsymbol{\delta}_\epsilon^*} \tilde{w}(\mathbf{y}_t, s_\epsilon, s_x; \Psi, \Psi^*)\right) m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x = \mathbf{0} \\
&\Longleftrightarrow \bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N}) \int_0^1 \int_0^1 m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x - \boldsymbol{\mu}^* \int_0^1 \int_0^1 m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x \\
&\quad - \boldsymbol{\delta}_\epsilon^* \int_0^1 \int_0^1 \mathbf{D}_\epsilon^{-1}\, m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x - \bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N}) \int_0^1 \int_0^1 \mathbf{D}_\epsilon \mathbf{W} \mathbf{M}^{-1}\, m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x\, \mathbf{W}^{*T} \\
&\quad + \boldsymbol{\mu} \int_0^1 \int_0^1 \mathbf{D}_\epsilon \mathbf{W} \mathbf{M}^{-1}\, m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x\, \mathbf{W}^{*T} + \big(\boldsymbol{\delta}_\epsilon \mathbf{W} - \sigma^2 \boldsymbol{\delta}_x\big) \int_0^1 \int_0^1 \mathbf{M}^{-1}\, m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x\, \mathbf{W}^{*T} = \mathbf{0}
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\delta}_x^*} Q_{GStS,id}(\Psi, \Psi^*) = \mathbf{0} \Longleftrightarrow \int_0^1 \int_0^1 \left(\sum_{t=1}^{N} \frac{\partial}{\partial \boldsymbol{\delta}_x^*} \tilde{w}(\mathbf{y}_t, s_\epsilon, s_x; \Psi, \Psi^*)\right) m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x = \mathbf{0} \\
&\Longleftrightarrow - \boldsymbol{\delta}_x^* \int_0^1 \int_0^1 \mathbf{D}_x^{-1}\, m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x + \bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N}) \int_0^1 \int_0^1 \mathbf{D}_\epsilon \mathbf{W} \mathbf{M}^{-1}\, m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x \\
&\quad - \boldsymbol{\mu} \int_0^1 \int_0^1 \mathbf{D}_\epsilon \mathbf{W} \mathbf{M}^{-1}\, m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x - \big(\boldsymbol{\delta}_\epsilon \mathbf{W} - \sigma^2 \boldsymbol{\delta}_x\big) \int_0^1 \int_0^1 \mathbf{M}^{-1}\, m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x = \mathbf{0}
\end{aligned}
$$

$$
\begin{aligned}
&\frac{\partial}{\partial \sigma^{*2}} Q_{GStS,id}(\Psi, \Psi^*) = \mathbf{0} \Longleftrightarrow \int_0^1 \int_0^1 \left(\sum_{t=1}^{N} \frac{\partial}{\partial \sigma^{*2}} \tilde{w}(\mathbf{y}_t, s_\epsilon, s_x; \Psi, \Psi^*)\right) m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x = \mathbf{0} \\
&\Longleftrightarrow - \frac{d}{2\sigma^{*2}} \int_0^1 \int_0^1 m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x - \frac{1}{2(\sigma^{*2})^2} \Bigg[- \mathrm{Tr}\left\{\bar{\mathbf{S}}(\mathbf{y}_{1:N}) \int_0^1 \int_0^1 \mathbf{D}_\epsilon\, m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x\right\} \\
&\quad + 2\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N}) \int_0^1 \int_0^1 \mathbf{D}_\epsilon\, m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x\, \boldsymbol{\mu}^{*T} + 2\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N}) \int_0^1 \int_0^1 m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x\, \boldsymbol{\delta}_\epsilon^{*T} \\
&\quad - \boldsymbol{\mu}^* \int_0^1 \int_0^1 \mathbf{D}_\epsilon\, m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x \boldsymbol{\mu}^{*T}
\end{aligned}
$$

$$
\begin{aligned}
&- 2\boldsymbol{\mu}^* \boldsymbol{\delta}_\epsilon^{*T} \int_0^1 \int_0^1 m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x - \boldsymbol{\delta}_\epsilon^* \int_0^1 \int_0^1 \mathbf{D}_\epsilon^{-1}\, m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x\, \boldsymbol{\delta}_\epsilon^{*T} \\
&+ 2\,\mathrm{Tr}\left\{ \bar{\mathbf{S}}(\mathbf{y}_{1:N}) \int_0^1 \int_0^1 \mathbf{D}_{\epsilon,t} \mathbf{W} \mathbf{M}_t^{-1} \mathbf{W}^{*T} \mathbf{D}_\epsilon\, m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x \right\} \\
&- 2\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N}) \int_0^1 \int_0^1 \mathbf{D}_\epsilon \mathbf{W} \mathbf{M}^{-1} \mathbf{W}^{*T} \mathbf{D}_\epsilon\, m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x \boldsymbol{\mu}^{*T} \\
&+ 2\boldsymbol{\mu} \int_0^1 \int_0^1 \mathbf{D}_\epsilon \mathbf{W} \mathbf{M}^{-1} \mathbf{W}^{*T} \mathbf{D}_\epsilon\, m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x \boldsymbol{\mu}^{*T} \\
&+ 2\left(\boldsymbol{\delta}_\epsilon \mathbf{W} - \sigma^2 \boldsymbol{\delta}_x\right) \int_0^1 \int_0^1 \mathbf{M}^{-1} \mathbf{W}^{*T} \mathbf{D}_\epsilon\, m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x\, \boldsymbol{\mu}^{*T} \\
&- 2\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N}) \int_0^1 \int_0^1 \mathbf{D}_\epsilon \mathbf{W} \mathbf{M}^{-1}\, m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x\, \mathbf{W}^{*T} \boldsymbol{\delta}_\epsilon^{*T} + 2\boldsymbol{\mu} \int_0^1 \int_0^1 \mathbf{D}_\epsilon \mathbf{W} \mathbf{M}^{-1}\, m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x \Big\} \mathbf{W}^{*T} \boldsymbol{\delta}_\epsilon^{*T} \\
&+ 2\left(\boldsymbol{\delta}_\epsilon \mathbf{W} - \sigma^2 \boldsymbol{\delta}_x\right) \int_0^1 \int_0^1 \mathbf{M}^{-1}\, m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x \mathbf{W}^{*T} \boldsymbol{\delta}_\epsilon^{*T} \\
&- 2\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N}) \int_0^1 \int_0^1 \mathbf{D}_\epsilon \mathbf{W}^* \mathbf{M}^{-1} \mathbf{W}^T \mathbf{D}_\epsilon\, m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x\, \boldsymbol{\mu}^T \\
&- 2\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N}) \int_0^1 \int_0^1 \mathbf{D}_\epsilon \mathbf{W}^* \mathbf{M}^{-1}\, m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x \left(\boldsymbol{\delta}_\epsilon \mathbf{W} - \sigma^2 \boldsymbol{\delta}_x\right)^T \\
&- \sigma^2\,\mathrm{Tr}\left\{ \int_0^1 \int_0^1 \mathbf{M}^{-1} \mathbf{W}^{*T} \mathbf{D}_\epsilon m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x \mathbf{W}^* \right\} \\
&- \mathrm{Tr}\left\{ \int_0^1 \int_0^1 \mathbf{M}^{-1} \mathbf{W}^T \mathbf{D}_\epsilon \bar{\mathbf{S}}(\mathbf{y}_{1:N}) \mathbf{D}_\epsilon \mathbf{W} \mathbf{M}^{-1} \mathbf{W}^{*T} \mathbf{D}_\epsilon\, m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x \mathbf{W}^* \right\} \\
&+ \mathrm{Tr}\left\{ \int_0^1 \int_0^1 \mathbf{M}^{-1} \Big(\mathbf{W}^T \mathbf{D}_\epsilon \bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})^T \left(\boldsymbol{\mu} \mathbf{D}_\epsilon \mathbf{W} + \boldsymbol{\delta}_\epsilon \mathbf{W} - \sigma^2 \boldsymbol{\delta}_x\right) + \left(\boldsymbol{\mu} \mathbf{D}_\epsilon \mathbf{W} + \boldsymbol{\delta}_\epsilon \mathbf{W} - \sigma^2 \boldsymbol{\delta}_x\right)^T \bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N}) \mathbf{D}_\epsilon \mathbf{W} \Big) \right. \\
&\quad \left. \times \mathbf{M}^{-1} \mathbf{W}^{*T} \mathbf{D}_\epsilon\, m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x \mathbf{W}^* \right\} \\
&- \mathrm{Tr}\left\{ \int_0^1 \int_0^1 \mathbf{M}_t^{-1} \left(\boldsymbol{\mu} \mathbf{D}_{\epsilon,t} \mathbf{W} + \boldsymbol{\delta}_\epsilon \mathbf{W} - \sigma^2 \boldsymbol{\delta}_x\right)^T \left(\boldsymbol{\mu} \mathbf{D}_{\epsilon,t} \mathbf{W} + \boldsymbol{\delta}_\epsilon \mathbf{W} - \sigma^2 \boldsymbol{\delta}_x\right) \mathbf{M}_t^{-1} \mathbf{W}^{*T} \mathbf{D}_{\epsilon,t}\, m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x \mathbf{W}^* \right\} \Big] = 0
\end{aligned}
$$

The iterative updates of the parameters $\boldsymbol{\mu}$, $\boldsymbol{\delta}_\epsilon$, $\boldsymbol{\delta}_x$ and σ^2, being the maximizers of the function $Q_{GStS,id}$ with respect to $\boldsymbol{\mu}^*$, $\boldsymbol{\delta}_\epsilon^*$, $\boldsymbol{\delta}_x^*$ and σ^{*2}, are obtained in the closed forms determined by various integration problems specified on 2-dimensional hypercube $[0,1]^2$. Recall that the equations defining these maximizers are dependent on $\mathbf{W}^*$. The maximizer which correspond to the parameter $\mathbf{W}^*$ is more difficult to obtain since the partial derivative of the function $Q_{GStS,id}$ with respect to $\mathbf{W}^*$ requires integrating the matrix products which contain $\mathbf{W}^*$ and cannot be further simplified, that is

$$
\begin{aligned}
&\frac{\partial}{\partial \mathbf{W}^*} Q_{GStS,ind}(\Psi, \Psi^*) = \mathbf{0} \Longleftrightarrow \int_0^1 \int_0^1 \Big(\sum_{t=1}^N \frac{\partial}{\partial \mathbf{W}^*} \tilde{w}(\mathbf{y}_t, s_\epsilon, s_x; \Psi, \Psi^*) \Big)\, m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x = \mathbf{0} \\
&\Longleftrightarrow \int_0^1 \int_0^1 \mathbf{D}_\epsilon \hat{\mathbf{S}}(\mathbf{y}_{1:N}) \mathbf{D}_\epsilon \mathbf{W} \mathbf{M}^{-1}\, m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x - \int_0^1 \int_0^1 \mathbf{D}_\epsilon \boldsymbol{\mu}^{*T} \hat{\boldsymbol{\mu}}(\mathbf{y}_{1:N}) \mathbf{D}_\epsilon \mathbf{W} \mathbf{M}^{-1}\, m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x \\
&- \boldsymbol{\delta}_\epsilon^{*T} \hat{\boldsymbol{\mu}}(\mathbf{y}_{1:N}) \int_0^1 \int_0^1 \mathbf{D}_\epsilon \mathbf{W} \mathbf{M}^{-1}\, m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x \\
&- \int_0^1 \int_0^1 \mathbf{D}_\epsilon \left(\hat{\boldsymbol{\mu}}(\mathbf{y}_{1:N}) - \boldsymbol{\mu}^*\right)^T \left(\boldsymbol{\mu} \mathbf{D}_\epsilon \mathbf{W} + \boldsymbol{\delta}_\epsilon \mathbf{W} - \sigma^2 \boldsymbol{\delta}_x\right) \mathbf{M}^{-1}\, m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x \\
&+ \boldsymbol{\delta}_\epsilon^{*T} \boldsymbol{\mu} \int_0^1 \int_0^1 \mathbf{D}_\epsilon \mathbf{W} \mathbf{M}^{-1}\, m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x + \boldsymbol{\delta}_\epsilon^{*T} \left(\boldsymbol{\delta}_\epsilon \mathbf{W} - \sigma^2 \boldsymbol{\delta}_x\right) \int_0^1 \int_0^1 \mathbf{M}^{-1}\, m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x \\
&- \int_0^1 \int_0^1 \mathbf{D}_\epsilon \mathbf{W}^* \mathbf{M}^{-1} \mathbf{W}^T \mathbf{D}_\epsilon \hat{\mathbf{S}}(\mathbf{y}_{1:N}) \mathbf{D}_\epsilon \mathbf{W} \mathbf{M}^{-1}\, m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x \\
&+ \int_0^1 \int_0^1 \mathbf{D}_\epsilon \mathbf{W}^* \mathbf{M}^{-1} \mathbf{W}^T \mathbf{D}_\epsilon \hat{\boldsymbol{\mu}}(\mathbf{y}_{1:N})^T \left(\boldsymbol{\mu} \mathbf{D}_\epsilon \mathbf{W} + \boldsymbol{\delta}_\epsilon \mathbf{W} - \sigma^2 \boldsymbol{\delta}_x\right) \mathbf{M}^{-1}\, m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x \\
&+ \int_0^1 \int_0^1 \mathbf{D}_\epsilon \mathbf{W}^* \mathbf{M}^{-1} \left(\boldsymbol{\mu} \mathbf{D}_\epsilon \mathbf{W} + \boldsymbol{\delta}_\epsilon \mathbf{W} - \sigma^2 \boldsymbol{\delta}_x\right)^T \hat{\boldsymbol{\mu}}(\mathbf{y}_{1:N}) \mathbf{D}_\epsilon \mathbf{W} \mathbf{M}^{-1}\, m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x \\
&- \int_0^1 \int_0^1 \mathbf{D}_\epsilon \mathbf{W}^* \mathbf{M}^{-1} \left(\boldsymbol{\mu} \mathbf{D}_\epsilon \mathbf{W} + \boldsymbol{\delta}_\epsilon \mathbf{W} - \sigma^2 \boldsymbol{\delta}_x\right)^T \left(\boldsymbol{\mu} \mathbf{D}_\epsilon \mathbf{W} + \boldsymbol{\delta}_\epsilon \mathbf{W} - \sigma^2 \boldsymbol{\delta}_x\right) \mathbf{M}^{-1}\, m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x \\
&- \sigma^2 \int_0^1 \int_0^1 \mathbf{D}_\epsilon \mathbf{W}^* \mathbf{M}^{-1}\, m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x = \mathbf{0}
\end{aligned}
$$

The maximizer with respect to $\mathbf{W}^*$ requires solving a root finding problem $\frac{\partial Q(\Psi, \Psi^*)}{\partial \mathbf{W}^*} = \mathbf{0}$. Recall that $\frac{\partial Q(\Psi, \Psi^*)}{\partial \mathbf{W}^*}$ contains elements of $\boldsymbol{\mu}^*$ and $\boldsymbol{\delta}_\epsilon^*$. Let us introduce the following notation

$$
A_0(\mathbf{y}_{1:N}; \Psi)_{1\times 1} := \int_0^1 \int_0^1 m(\mathbf{y}_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x
$$

$$\mathbf{A}_1(\mathbf{y}_{1:N};\Psi)_{d\times d} := \int_0^1\int_0^1 \mathbf{D}_\epsilon\, m(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)\, ds_\epsilon\, ds_x$$
$$\mathbf{A}_2(\mathbf{y}_{1:N};\Psi)_{d\times d} := \int_0^1\int_0^1 \mathbf{D}_\epsilon^{-1}\, m(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)\, ds_\epsilon\, ds_x$$
$$\mathbf{A}_3(\mathbf{y}_{1:N};\Psi)_{k\times k} := \int_0^1\int_0^1 \mathbf{D}_x^{-1}\, m(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)\, ds_\epsilon\, ds_x$$
$$\mathbf{A}_4(\mathbf{y}_{1:N};\Psi)_{k\times k} := \int_0^1\int_0^1 \mathbf{M}_t^{-1}\, m(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)\, ds_\epsilon\, ds_x$$
$$\mathbf{A}_5(\mathbf{y}_{1:N};\Psi)_{d\times k} := \int_0^1\int_0^1 \mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\, m(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)\, ds_\epsilon\, ds_x$$
$$\mathbf{A}_6(\mathbf{y}_{1:N};\Psi,\Psi^*)_{d\times k} := \int_0^1\int_0^1 \mathbf{D}_\epsilon\mathbf{W}^*\mathbf{M}^{-1}\, m(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)\, ds_\epsilon\, ds_x$$
$$\mathbf{A}_7(\mathbf{y}_{1:N};\Psi,\Psi^*)_{d\times d} := \int_0^1\int_0^1 \mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon\, m(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)\, ds_\epsilon\, ds_x$$
$$\mathbf{A}_8(\mathbf{y}_{1:N};\Psi)_{d\times k} := \int_0^1\int_0^1 \mathbf{D}_\epsilon\hat{\mathbf{S}}^m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\, m(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)\, ds_\epsilon\, ds_x$$
$$\mathbf{A}_9(\mathbf{y}_{1:N};\Psi,\Psi^*)_{d\times k} := \int_0^1\int_0^1 \mathbf{D}_\epsilon\boldsymbol{\mu}^{*T}\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\, m(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)\, ds_\epsilon\, ds_x$$
$$\mathbf{A}_{10}(\mathbf{y}_{1:N};\Psi,\Psi^*)_{d\times k} := \int_0^1\int_0^1 \mathbf{D}_\epsilon\big(\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)-\boldsymbol{\mu}^*\big)^T\big(\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}+\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\big)\mathbf{M}^{-1}\, m(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)\, ds_\epsilon\, ds_x$$
$$\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi,\Psi^*)_{k\times d} := \int_0^1\int_0^1 \mathbf{D}_\epsilon\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\hat{\mathbf{S}}(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}_t^{-1}\, m(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)\, ds_\epsilon\, ds_x$$
$$\mathbf{A}_{12}(\mathbf{y}_{1:N};\Psi,\Psi^*)_{k\times d} := \int_0^1\int_0^1 \mathbf{D}_\epsilon\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)^T\Big(\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}+\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}^{-1}\, m(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)\, ds_\epsilon\, ds_x$$
$$\mathbf{A}_{13}(\mathbf{y}_{1:N};\Psi,\Psi^*)_{k\times d} := \int_0^1\int_0^1 \mathbf{D}_\epsilon\mathbf{W}^*\mathbf{M}^{-1}\Big(\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}+\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\Big)^T\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\, m(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)\, ds_\epsilon\, ds_x$$
$$\mathbf{A}_{14}(\mathbf{y}_{1:N};\Psi,\Psi^*)_{k\times d} := \int_0^1\int_0^1 \mathbf{D}_\epsilon\mathbf{W}^*\mathbf{M}^{-1}\Big(\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}+\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\Big)^T\Big(\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}+\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}^{-1}\, m(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)\, ds_\epsilon\, ds_x.$$

By denoting $f:\mathbb{R}^{d\times k}\to\mathbb{R}^{d\times k}$ such that $f(\mathbf{W}^*;\Psi,\Psi^*) := \frac{\partial Q(\Psi,\Psi^*)}{\partial\mathbf{W}^*}$ defined as

$$\begin{aligned} f(\mathbf{W}^*;\Psi,\Psi^*) = {}& \mathbf{A}_8(\mathbf{y}_{1:N};\Psi)-\mathbf{A}_9(\mathbf{y}_{1:N};\Psi,\Psi^*)-\boldsymbol{\delta}_\epsilon^{*T}\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{A}_5(\mathbf{y}_{1:N};\Psi)-\mathbf{A}_{10}(\mathbf{y}_{1:N};\Psi,\Psi^*)+\boldsymbol{\delta}_\epsilon^{*T}\boldsymbol{\mu}\mathbf{A}_5(\mathbf{y}_{1:N};\Psi)+\boldsymbol{\delta}_\epsilon^{*T}(\boldsymbol{\delta}\mathbf{W}\\ & -\sigma^2\mathbf{W})\mathbf{A}_4(\mathbf{y}_{1:N};\Psi)-\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi,\Psi^*)+\mathbf{A}_{12}(\mathbf{y}_{1:N};\Psi,\Psi^*)+\mathbf{A}_{13}(\mathbf{y}_{1:N};\Psi,\Psi^*)-\mathbf{A}_{14}(\mathbf{y}_{1:N};\Psi,\Psi^*)-\sigma^2\mathbf{A}_6(\mathbf{y}_{1:N};\Psi,\Psi^*), \end{aligned}$$

the system of equations $\nabla_{\Psi^*}Q_{GStS,id}(\Psi,\Psi^*)=\mathbf{0}$ reduces to

$$\boldsymbol{\mu}^* = \Big[\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\Big(\mathbf{A}_1(\mathbf{y}_{1:N};\Psi)-\mathbf{A}_7(\mathbf{y}_{1:N};\Psi,\Psi^*)\Big)-\boldsymbol{\delta}_\epsilon^*A_0(\mathbf{y}_{1:N};\Psi)+\boldsymbol{\mu}\mathbf{A}_7(\mathbf{y}_{1:N};\Psi,\Psi^*)+\Big(\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\Big)\mathbf{A}_6(\mathbf{y}_{1:N};\Psi,\Psi^*)^T\Big]\mathbf{A}_1(\mathbf{y}_{1:N};\Psi)^{-1}$$
$$\boldsymbol{\delta}_\epsilon^* = \Big[\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\Big(A_0(\mathbf{y}_{1:N};\Psi)-\mathbf{A}_5(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\Big)-\boldsymbol{\mu}^*A_0(\mathbf{y}_{1:N};\Psi)+\boldsymbol{\mu}\mathbf{A}_5(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\Big)+(\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x)\mathbf{A}_4(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\Big]\mathbf{A}_2(\mathbf{y}_{1:N};\Psi)^{-1}$$
$$\boldsymbol{\delta}_x^* = \Big[\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{A}_5(\mathbf{y}_{1:N})-\boldsymbol{\mu}\mathbf{A}_5(\mathbf{y}_{1:N})-(\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x)\mathbf{A}_4(\mathbf{y}_{1:N};\Psi)\Big]\mathbf{A}_3(\mathbf{y}_{1:N},\Psi)^{-1}$$
$$\begin{aligned} \sigma^{*2} = {}& \frac{1}{d\,A_0(\mathbf{y}_{1:N};\Psi)}\Big[\operatorname{Tr}\Big\{\bar{\mathbf{S}}(\mathbf{y}_{1:N})\mathbf{A}_1(\mathbf{y}_{1:N};\Psi)\Big\}-2\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{A}_1(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^*-2\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})A_0(\mathbf{y}_{1:N};\Psi)\boldsymbol{\delta}_\epsilon^*+\boldsymbol{\mu}^*\mathbf{A}_1(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^{*T}\\ & +2\boldsymbol{\mu}^*\boldsymbol{\delta}_\epsilon^{*T}A_0(\mathbf{y}_{1:N};\Psi)+\boldsymbol{\delta}_\epsilon^*\mathbf{A}_2(\mathbf{y}_{1:N};\Psi)\boldsymbol{\delta}_\epsilon^{*T}-2\operatorname{Tr}\Big\{\bar{\mathbf{S}}(\mathbf{y}_{1:N})\mathbf{A}_7(\mathbf{y}_{1:N};\Psi,\Psi^*)\Big\}+2\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{A}_7(\mathbf{y}_{1:N};\Psi,\Psi^*)\boldsymbol{\mu}^{*T}\\ & -2\boldsymbol{\mu}\mathbf{A}_7(\mathbf{y}_{1:N};\Psi,\Psi^*)\boldsymbol{\mu}^{*T}-2(\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x)\mathbf{A}_6(\mathbf{y}_{1:N};\Psi)^T\boldsymbol{\mu}^*\\ & +2\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{A}_5(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\boldsymbol{\delta}_\epsilon^{*T}-2\boldsymbol{\mu}\mathbf{A}_5(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\boldsymbol{\delta}_\epsilon^{*T}-2(\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x)\mathbf{A}_4(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\boldsymbol{\delta}_\epsilon^{*T}+2\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{A}_7(\mathbf{y}_{1:N};\Psi,\Psi^*)^T\boldsymbol{\mu}^T\\ & +2\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{A}_6(\mathbf{y}_{1:N};\Psi,\Psi^*)(\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x)^T+\sigma^2\operatorname{Tr}\Big\{\mathbf{A}_6(\mathbf{y}_{1:N};\Psi,\Psi^*)\mathbf{W}^*\Big\}+\operatorname{Tr}\Big\{\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi,\Psi^*)^T\mathbf{W}^*\Big\}\\ & -\operatorname{Tr}\Big\{\mathbf{A}_{12}(\mathbf{y}_{1:N};\Psi,\Psi^*)^T\mathbf{W}^*\Big\}-\operatorname{Tr}\Big\{\mathbf{A}_{13}(\mathbf{y}_{1:N};\Psi,\Psi^*)^T\mathbf{W}^*\Big\}+\operatorname{Tr}\Big\{\mathbf{A}_{14}(\mathbf{y}_{1:N};\Psi,\Psi^*)^T\mathbf{W}^*\Big\}\Big] \end{aligned}$$

$f(\mathbf{W}^*;\Psi,\Psi^*)=\mathbf{0}$.

□

A.4.3.4. Proof of Corollary 13

Proof. Given the formulation of the gradient $\nabla_{\Psi^*}Q_{GStS,id}$ from Equation (A17), system of equations specified in Theorem 16 and the assumptions of Special Case 2 from Subsection 3.4.1.3, the M-step for Generalized Skew t-Student ID PPCA

is specified finding the the maximizers of the function $Q_{GStS,id,2}$ from Corollary 9 with respect to $\boldsymbol{\mu}^*$, σ^{*2}, δ_ϵ^* and δ_x^* as following

$$
\begin{aligned}
&\sum_{t=1}^{N} \frac{\partial \tilde{w}(\mathbf{y}_t, s_\epsilon, s_x; \Psi, \Psi^*)}{\partial \boldsymbol{\mu}^*} = \frac{N}{\sigma^{*2}} \Big[T_\epsilon(s_\epsilon) \bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N}) - T_\epsilon(s_\epsilon) \boldsymbol{\mu}^* - \delta_\epsilon^* - T_\epsilon(s_\epsilon)^2 \bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N}) \mathbf{W} \mathbf{M}^{-1} \mathbf{W}^{*T} + T_\epsilon(s_\epsilon)^2 \boldsymbol{\mu} \mathbf{W} \mathbf{M}^{-1} \mathbf{W}^{*T} \\
&\qquad + T_\epsilon(s_\epsilon) \big(\delta_\epsilon \mathbf{W} - \sigma^2 \delta_x \big) \mathbf{M}^{-1} \mathbf{W}^{*T} \Big] \\
&\Longrightarrow \frac{\partial Q_{GStS,id,2}(\Psi, \Psi^*)}{\partial \boldsymbol{\mu}^*} = \mathbf{0} \\
&\Longrightarrow \boldsymbol{\mu}^* = A_1(\mathbf{y}_{1:N}; \Psi)^{-1} \Big[\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N}) \Big(A_1(\mathbf{y}_{1:N}; \Psi) - \mathbf{W} \mathbf{A}_6(\mathbf{y}_{1:N}; \Psi) \mathbf{W}^{*T} \Big) - \delta_\epsilon^* A_0(\mathbf{y}_{1:N}; \Psi) + \boldsymbol{\mu} \mathbf{W} \mathbf{A}_6(\mathbf{y}_{1:N}; \Psi) \mathbf{W}^{*T} \\
&\quad + \big(\delta_\epsilon \mathbf{W} - \sigma^2 \delta_x \big) \mathbf{A}_5(\mathbf{y}_{1:N}; \Psi) \mathbf{W}^{*T} \Big]
\end{aligned}
$$

$$
\begin{aligned}
&\sum_{t=1}^{N} \frac{\partial \tilde{w}(\mathbf{y}_t, s_\epsilon, s_x; \Psi, \Psi^*)}{\partial \delta_\epsilon^*} = \frac{N}{\sigma^{*2}} \Big[\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N}) \big(1 - T_\epsilon(s_\epsilon) \mathbf{W} \mathbf{M}^{-1} \mathbf{W}^{*T} \big) - \boldsymbol{\mu}^* - T_\epsilon(s_\epsilon)^{-1} \delta_\epsilon^* + T_\epsilon(s_\epsilon) \boldsymbol{\mu} \mathbf{W} \mathbf{M}^{-1} \mathbf{W}^{*T} + \big(\delta_\epsilon \mathbf{W} - \sigma^2 \delta_x \big) \mathbf{M}^{-1} \mathbf{W}^{*T} \Big] \\
&\Longrightarrow \frac{\partial Q_{GStS,id,2}(\Psi, \Psi^*)}{\partial \delta_\epsilon^*} = \mathbf{0} \\
&\Longleftrightarrow \delta_\epsilon^* = A_2(\mathbf{y}_{1:N}; \Psi)^{-1} \Big[\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N}) \Big(A_0(\mathbf{y}_{1:N}; \Psi) - \mathbf{W} \mathbf{A}_5(\mathbf{y}_{1:N}; \Psi) \mathbf{W}^{*T} \Big) - A_0(\mathbf{y}_{1:N}; \Psi) \boldsymbol{\mu}^* + \boldsymbol{\mu} \mathbf{W} \mathbf{A}_5(\mathbf{y}_{1:N}; \Psi) \mathbf{W}^{*T} \\
&\quad + \big(\delta_\epsilon \mathbf{W} - \sigma^2 \delta_x \big) \mathbf{A}_4(\mathbf{y}_{1:N}; \Psi) \mathbf{W}^{*T} \Big]
\end{aligned}
$$

$$
\begin{aligned}
&\sum_{t=1}^{N} \frac{\partial \tilde{w}(\mathbf{y}_t, s_\epsilon, s_x; \Psi, \Psi^*)}{\partial \delta_x^*} = N \Big[- T_x(s_x)^{-1} \delta_x^* + T_\epsilon(s_\epsilon) \bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N}) \mathbf{W} \mathbf{M}^{-1} - T_\epsilon(s_\epsilon) \boldsymbol{\mu} \mathbf{W} \mathbf{M}^{-1} - \big(\delta_\epsilon \mathbf{W} - \sigma^2 \delta_x \big) \mathbf{M}^{-1} \Big] \\
&\Longrightarrow \frac{\partial Q_{GStS,id,2}(\Psi, \Psi^*)}{\partial \delta_x^*} = \mathbf{0} \\
&\Longrightarrow \delta_x^* = A_3(\mathbf{y}_{1:N}; \Psi)^{-1} \Big[\big(\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N}) - \boldsymbol{\mu} \big) \mathbf{A}_5(\mathbf{y}_{1:N}; \Psi) - \big(\delta_\epsilon \mathbf{W} - \sigma^2 \delta_x \big) \mathbf{A}_4(\mathbf{y}_{1:N}; \Psi) \Big]
\end{aligned}
$$

$$
\begin{aligned}
&\sum_{t=1}^{N} \frac{\partial \tilde{w}(\mathbf{y}_t, s_\epsilon, s_x; \Psi, \Psi^*)}{\partial \sigma^{*2}} = - \frac{dN}{2\sigma^{*2}} + \frac{N}{2(\sigma^{*2})^2} \Big[T_\epsilon(s_\epsilon) \operatorname{Tr} \Big\{ \bar{\mathbf{S}}(\mathbf{y}_{1:N}) \Big\} - 2 \bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N}) \Big(T_\epsilon(s_\epsilon) \boldsymbol{\mu}^{*T} + \delta_\epsilon^{*T} \Big) + T_\epsilon(s_\epsilon) \boldsymbol{\mu}^* \boldsymbol{\mu}^{*T} + 2 \boldsymbol{\mu}^* \delta_\epsilon^{*T} \\
&\quad + T_\epsilon(s_\epsilon)^{-1} \delta_\epsilon^* \delta_\epsilon^{*T} - 2 T_\epsilon(s_\epsilon)^2 \operatorname{Tr} \Big\{ \bar{\mathbf{S}}(\mathbf{y}_{1:N}) \mathbf{W} \mathbf{M}^{-1} \mathbf{W}^{*T} \Big\} + 2 T_\epsilon(s_\epsilon) \bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N}) \mathbf{W} \mathbf{M}^{-1} \mathbf{W}^{*T} \Big(T_\epsilon(s_\epsilon) \boldsymbol{\mu}^{*T} + \delta_\epsilon^{*T} \Big) \\
&\quad + 2 T_\epsilon(s_\epsilon) \bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N}) \mathbf{W}^* \mathbf{M}^{-1} \Big(\boldsymbol{\mu} T_\epsilon(s_\epsilon) \mathbf{W} + \delta_\epsilon \mathbf{W} - \sigma^2 \delta_x \Big)^T - 2 \Big(T_\epsilon(s_\epsilon) \boldsymbol{\mu} \mathbf{W} + \delta_\epsilon \mathbf{W} - \sigma^2 \delta_x \Big) \mathbf{M}^{-1} \mathbf{W}^{*T} \Big(T_\epsilon(s_\epsilon) \boldsymbol{\mu}^{*T} + \delta_\epsilon^{*T} \Big) \\
&\quad + T_\epsilon(s_\epsilon)^3 \operatorname{Tr} \Big\{ \mathbf{M}^{-1} \mathbf{W}^T \bar{\mathbf{S}}(\mathbf{y}_{1:N}) \mathbf{W} \mathbf{M}^{-1} \mathbf{W}^{*T} \mathbf{W}^* \Big\} + T_\epsilon(s_\epsilon)^2 \operatorname{Tr} \Big\{ \mathbf{M}^{-1} \mathbf{W}^T \bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})^T \Big(- T_\epsilon(s_\epsilon) \boldsymbol{\mu} \mathbf{W} - \delta_\epsilon \mathbf{W} + \sigma^2 \delta_x \Big) \mathbf{M}^{-1} \mathbf{W}^{*T} \mathbf{W}^* \Big\} \\
&\quad + T_\epsilon(s_\epsilon)^2 \operatorname{Tr} \Big\{ \mathbf{M}^{-1} \Big(- T_\epsilon(s_\epsilon) \boldsymbol{\mu} \mathbf{W} - \delta_\epsilon \mathbf{W} + \sigma^2 \delta_x \Big)^T \bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N}) \mathbf{W} \mathbf{M}^{-1} \mathbf{W}^{*T} \mathbf{W}^* \Big\} \\
&\quad + T_\epsilon(s_\epsilon) \operatorname{Tr} \Big\{ \mathbf{M}^{-1} \Big(- T_\epsilon(s_\epsilon) \boldsymbol{\mu} \mathbf{W} - \delta_\epsilon \mathbf{W} + \sigma^2 \delta_x \Big)^T \Big(- T_\epsilon(s_\epsilon) \boldsymbol{\mu} \mathbf{W} - \delta_\epsilon \mathbf{W} + \sigma^2 \delta_x \Big) \mathbf{M}^{-1} \mathbf{W}^{*T} \mathbf{W}^* \Big\} + \sigma^2 T_\epsilon(s_\epsilon) \operatorname{Tr} \Big\{ \mathbf{M}^{-1} \mathbf{W}^{*T} \mathbf{W}^* \Big\} \Big] \\
&\Longrightarrow \frac{\partial Q_{GStS,id,2}(\Psi, \Psi^*)}{\partial \sigma^{*2}} = \mathbf{0} \\
&\Longleftrightarrow \sigma^{*2} = \frac{1}{d A_0(\mathbf{y}_{1:N}; \Psi)} \Big[A_1(\mathbf{y}_{1:N}; \Psi) \operatorname{Tr} \Big\{ \bar{\mathbf{S}}(\mathbf{y}_{1:N}) \Big\} - 2 A_1(\mathbf{y}_{1:N}; \Psi) \bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N}) \boldsymbol{\mu}^{*T} - 2 A_0(\mathbf{y}_{1:N}; \Psi) \bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N}) \delta_\epsilon^{*T} + A_1(\mathbf{y}_{1:N}; \Psi) \boldsymbol{\mu}^* \boldsymbol{\mu}^{*T} \\
&\quad + 2 A_0(\mathbf{y}_{1:N}; \Psi) \boldsymbol{\mu}^* \delta_\epsilon^{*T} + A_2(\mathbf{y}_{1:N}; \Psi) \delta_\epsilon^* \delta_\epsilon^{*T} - 2 \operatorname{Tr} \Big\{ \bar{\mathbf{S}}(\mathbf{y}_{1:N}) \mathbf{W} \mathbf{A}_6(\mathbf{y}_{1:N}; \Psi) \mathbf{W}^{*T} \Big\} + 2 \bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N}) \mathbf{W} \mathbf{A}_6(\mathbf{y}_{1:N}) \mathbf{W}^{*T} \boldsymbol{\mu}^{*T} \\
&\quad + 2 \bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N}) \mathbf{W} \mathbf{A}_5(\mathbf{y}_{1:N}) \mathbf{W}^{*T} \delta_\epsilon^{*T} + 2 \bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N}) \mathbf{W}^* \mathbf{A}_6(\mathbf{y}_{1:N}; \Psi) \mathbf{W}^T \boldsymbol{\mu}^T + 2 \bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N}) \mathbf{W}^* \mathbf{A}_5(\mathbf{y}_{1:N}; \Psi) \big(\delta_\epsilon \mathbf{W} - \sigma^2 \delta_x \big)^T \\
&\quad - 2 \boldsymbol{\mu} \mathbf{W} \mathbf{A}_6(\mathbf{y}_{1:N}; \Psi) \mathbf{W}^{*T} \boldsymbol{\mu}^{*T} - 2 \big(\delta_\epsilon \mathbf{W} - \sigma^2 \delta_x \big) \mathbf{A}_5(\mathbf{y}_{1:N}; \Psi) \mathbf{W}^{*T} \boldsymbol{\mu}^{*T} \\
&\quad - 2 \boldsymbol{\mu} \mathbf{W} \mathbf{A}_5(\mathbf{y}_{1:N}; \Psi) \mathbf{W}^{*T} \delta_\epsilon^{*T} - 2 \big(\delta_\epsilon \mathbf{W} - \sigma^2 \delta_x \big) \mathbf{A}_4(\mathbf{y}_{1:N}; \Psi) \mathbf{W}^{*T} \delta_\epsilon^{*T} + \operatorname{Tr} \Big\{ \mathbf{A}_7(\mathbf{y}_{1:N}; \Psi) \mathbf{W}^{*T} \mathbf{W}^* \Big\} + \operatorname{Tr} \Big\{ \mathbf{A}_9(\mathbf{y}_{1:N}; \Psi) \mathbf{W}^{*T} \mathbf{W}^* \Big\} \\
&\quad - \operatorname{Tr} \Big\{ \Big(\mathbf{A}_8(\mathbf{y}_{1:N}; \Psi) + \mathbf{A}_8(\mathbf{y}_{1:N}; \Psi)^T \Big) \mathbf{W}^{*T} \mathbf{W}^* \Big\} + \sigma^2 \operatorname{Tr} \Big\{ \mathbf{A}_5(\mathbf{y}_{1:N}; \Psi) \mathbf{W}^{*T} \mathbf{W}^* \Big\} \Big]
\end{aligned}
$$

$$
\begin{aligned}
&\sum_{t=1}^{N} \frac{\partial \tilde{w}(\mathbf{y}_t, s_\epsilon, s_x; \Psi, \Psi^*)}{\partial \mathbf{W}^*} = \frac{N}{\sigma^{*2}} \Big[T_\epsilon(s_\epsilon)^2 \bar{\mathbf{S}}(\mathbf{y}_{1:N}) \mathbf{W} \mathbf{M}^{-1} - T_\epsilon(s_\epsilon)^2 \boldsymbol{\mu}^{*T} \bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N}) \mathbf{W} \mathbf{M}^{-1} - T_\epsilon(s_\epsilon) \delta_\epsilon^{*T} \bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N}) \mathbf{W} \mathbf{M}^{-1}
\end{aligned}
$$

$$
\begin{aligned}
&- T_\epsilon(s_\epsilon)^2\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})^T\boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1} - T_\epsilon(s_\epsilon)\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})^T\big(\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\big)\mathbf{M}^{-1} + T_\epsilon(s_\epsilon)^2\boldsymbol{\mu}^{*T}\boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1} + T_\epsilon(s_\epsilon)\boldsymbol{\mu}^{*T}\big(\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\big)\mathbf{M}^{-1}\\
&+ T_\epsilon(s_\epsilon)\boldsymbol{\delta}_\epsilon^{*T}\boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1} + \boldsymbol{\delta}_\epsilon^{*T}\big(\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\big)\mathbf{M}^{-1} - T_\epsilon(s_\epsilon)^3\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{S}}(\mathbf{y}_{1:N})\mathbf{W}\mathbf{M}^{-1}\\
&+ T_\epsilon(s_\epsilon)^2\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})^T\Big(T_\epsilon(s_\epsilon)\boldsymbol{\mu}\mathbf{W} + \boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}^{-1} + T_\epsilon(s_\epsilon)^2\mathbf{W}^*\mathbf{M}^{-1}\Big(T_\epsilon(s_\epsilon)\boldsymbol{\mu}\mathbf{W} + \boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\Big)^T\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{W}\mathbf{M}^{-1}\\
&- T_\epsilon(s_\epsilon)\mathbf{W}^*\mathbf{M}^{-1}\Big(T_\epsilon(s_\epsilon)\boldsymbol{\mu}\mathbf{W} + \boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\Big)^T\Big(T_\epsilon(s_\epsilon)\boldsymbol{\mu}\mathbf{W} + \boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}^{-1} - \sigma^2T_\epsilon(s_\epsilon)\mathbf{W}^*\mathbf{M}^{-1}\\
\implies &\frac{\partial Q_{GStS,id,2}(\Psi,\Psi^*)}{\partial\mathbf{W}^*} = \mathbf{0}\\
\iff \mathbf{W}^* = &\Big[\bar{\mathbf{S}}(\mathbf{y}_{1:N})\mathbf{W}\mathbf{A}_6(\mathbf{y}_{1:N};\Psi) - \boldsymbol{\mu}^{*T}\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{W}\mathbf{A}_6(\mathbf{y}_{1:N};\Psi) - \boldsymbol{\delta}_\epsilon^{*T}\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{W}\mathbf{A}_5(\mathbf{y}_{1:N};\Psi) - \bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})^T\boldsymbol{\mu}\mathbf{W}\mathbf{A}_6(\mathbf{y}_{1:N};\Psi)\\
&- \bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})^T\big(\boldsymbol{\delta}_\epsilon - \sigma^2\boldsymbol{\delta}_x\big)\mathbf{A}_5(\mathbf{y}_{1:N};\Psi) + \boldsymbol{\mu}^{*T}\boldsymbol{\mu}\mathbf{W}\mathbf{A}_6(\mathbf{y}_{1:N};\Psi) + \boldsymbol{\mu}^{*T}\big(\boldsymbol{\delta}_\epsilon - \sigma^2\boldsymbol{\delta}_x\big)\mathbf{A}_5(\mathbf{y}_{1:N};\Psi)\\
&+ \boldsymbol{\delta}_\epsilon^{*T}\boldsymbol{\mu}\mathbf{W}\mathbf{A}_5(\mathbf{y}_{1:N};\Psi) + \boldsymbol{\delta}_\epsilon^{*T}\big(\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\big)\mathbf{A}_4(\mathbf{y}_{1:N};\Psi)\Big]\Big[\mathbf{A}_7(\mathbf{y}_{1:N};\Psi) - \mathbf{A}_8(\mathbf{y}_{1:N};\Psi) - \mathbf{A}_8(\mathbf{y}_{1:N};\Psi)^T + \mathbf{A}_9(\mathbf{y}_{1:N};\Psi) + \sigma^2\mathbf{A}_5(\mathbf{y}_{1:N};\Psi)\Big]^{-1}
\end{aligned}
$$

for

$$
\begin{aligned}
A_0(\mathbf{y}_{1:N};\Psi)_{1\times1} &:= \int_0^1\int_0^1 m(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x\\
A_1(\mathbf{y}_{1:N};\Psi)_{1\times1} &:= \int_0^1\int_0^1 T_\epsilon(s_\epsilon)\,m(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x\\
A_2(\mathbf{y}_{1:N};\Psi)_{1\times1} &:= \int_0^1\int_0^1 T_\epsilon(s_\epsilon)^{-1}\,m(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x\\
A_3(\mathbf{y}_{1:N};\Psi)_{1\times1} &:= \int_0^1\int_0^1 T_x(s_x)^{-1}\,m(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x\\
\mathbf{A}_4(\mathbf{y}_{1:N};\Psi)_{k\times k} &:= \int_0^1\int_0^1 \mathbf{M}^{-1}\,m(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x\\
\mathbf{A}_5(\mathbf{y}_{1:N};\Psi)_{k\times k} &:= \int_0^1\int_0^1 T_\epsilon(s_\epsilon)\mathbf{M}^{-1}\,m(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x\\
\mathbf{A}_6(\mathbf{y}^o_{1:N};\Psi)_{k\times k} &:= \int_0^1\int_0^1 T_\epsilon(s_\epsilon)^2\mathbf{M}^{-1}\,m(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_{x,t}\\
\mathbf{A}_7(\mathbf{y}_{1:N};\Psi)_{k\times k} &:= \int_0^1\int_0^1 T_\epsilon(s_\epsilon)^3\mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{S}}(\mathbf{y}_{1:N})\mathbf{W}\mathbf{M}^{-1}\,m(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x\\
\mathbf{A}_8(\mathbf{y}_{1:N};\Psi)_{k\times k} &:= \int_0^1\int_0^1 T_\epsilon(s_\epsilon)^2\mathbf{M}^{-1}\mathbf{W}^T\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\Big(T_\epsilon(s_\epsilon)\boldsymbol{\mu}\mathbf{W} + \boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}^{-1}m(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x\\
\mathbf{A}_9(\mathbf{y}_{1:N};\Psi)_{k\times k} &:= \int_0^1\int_0^1 T_\epsilon(s_\epsilon)\mathbf{M}^{-1}\Big(T_\epsilon(s_\epsilon)\boldsymbol{\mu}\mathbf{W} + \boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\Big)^T\Big(T_\epsilon(s_\epsilon)\boldsymbol{\mu}\mathbf{W} + \boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}^{-1}m(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x.
\end{aligned}
$$

Hence, the system of equations $\nabla_{\Psi^*}Q_{GStS,id,2}(\Psi,\Psi^*) = \mathbf{0}$ is the following

$$
\begin{cases}
\boldsymbol{\mu}^* = A_1(\mathbf{y}_{1:N};\Psi)^{-1}\Big[\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\Big(A_1(\mathbf{y}_{1:N};\Psi) - \mathbf{W}\mathbf{A}_6(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\Big) - \boldsymbol{\delta}_\epsilon^*A_0(\mathbf{y}_{1:N};\Psi) + \boldsymbol{\mu}\mathbf{W}\mathbf{A}_6(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T} + \big(\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\big)\mathbf{A}_5(\mathbf{y}_{1:N};\Psi)\mathbf{W}\\
\boldsymbol{\delta}_\epsilon^* = A_2(\mathbf{y}_{1:N};\Psi)^{-1}\Big[\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\Big(A_0(\mathbf{y}_{1:N};\Psi) - \mathbf{W}\mathbf{A}_5(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\Big) - A_0(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^* + \boldsymbol{\mu}\mathbf{W}\mathbf{A}_5(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\\
\qquad + \big(\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\big)\mathbf{A}_4(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\Big]\\
\boldsymbol{\delta}_x^* = A_3(\mathbf{y}_{1:N};\Psi)^{-1}\Big[\big(\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N}) - \boldsymbol{\mu}\big)\mathbf{A}_5(\mathbf{y}_{1:N};\Psi) - \big(\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\big)\mathbf{A}_4(\mathbf{y}_{1:N};\Psi)\Big]\\
\mathbf{W}^* = \Big[\bar{\mathbf{S}}(\mathbf{y}_{1:N})\mathbf{W}\mathbf{A}_6(\mathbf{y}_{1:N};\Psi) - \boldsymbol{\mu}^{*T}\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{W}\mathbf{A}_6(\mathbf{y}_{1:N};\Psi) - \boldsymbol{\delta}_\epsilon^{*T}\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{W}\mathbf{A}_5(\mathbf{y}_{1:N};\Psi) - \bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})^T\boldsymbol{\mu}\mathbf{W}\mathbf{A}_6(\mathbf{y}_{1:N};\Psi)\\
\qquad -\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})^T\big(\boldsymbol{\delta}_\epsilon - \sigma^2\boldsymbol{\delta}_x\big)\mathbf{A}_5(\mathbf{y}_{1:N};\Psi) + \boldsymbol{\mu}^{*T}\boldsymbol{\mu}\mathbf{W}\mathbf{A}_6(\mathbf{y}_{1:N};\Psi) + \boldsymbol{\mu}^{*T}\big(\boldsymbol{\delta}_\epsilon - \sigma^2\boldsymbol{\delta}_x\big)\mathbf{A}_5(\mathbf{y}_{1:N};\Psi) + \boldsymbol{\delta}_\epsilon^{*T}\boldsymbol{\mu}\mathbf{W}\mathbf{A}_5(\mathbf{y}_{1:N};\Psi)\\
\qquad +\boldsymbol{\delta}_\epsilon^{*T}\big(\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\big)\mathbf{A}_4(\mathbf{y}_{1:N};\Psi)\Big]\Big[\mathbf{A}_7(\mathbf{y}_{1:N};\Psi) - \mathbf{A}_8(\mathbf{y}_{1:N};\Psi) - \mathbf{A}_8(\mathbf{y}_{1:N};\Psi)^T + \mathbf{A}_9(\mathbf{y}_{1:N};\Psi) + \sigma^2\mathbf{A}_5(\mathbf{y}_{1:N};\Psi)\Big]^{-1}\\
\sigma^{*2} = \frac{1}{dA_0(\mathbf{y}_{1:N};\Psi)}\Big[A_1(\mathbf{y}_{1:N};\Psi)\,\mathrm{Tr}\Big\{\bar{\mathbf{S}}(\mathbf{y}_{1:N})\Big\} - 2A_1(\mathbf{y}_{1:N};\Psi)\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\boldsymbol{\mu}^{*T} - 2A_0(\mathbf{y}_{1:N};\Psi)\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\boldsymbol{\delta}_\epsilon^{*T} + A_1(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^*\boldsymbol{\mu}^{*T}\\
\qquad +2A_0(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^*\boldsymbol{\delta}_\epsilon^{*T} + A_2(\mathbf{y}_{1:N};\Psi)\boldsymbol{\delta}_\epsilon^*\boldsymbol{\delta}_\epsilon^{*T} - 2\,\mathrm{Tr}\Big\{\bar{\mathbf{S}}(\mathbf{y}_{1:N})\mathbf{W}\mathbf{A}_6(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\Big\} + 2\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{W}\mathbf{A}_6(\mathbf{y}_{1:N})\mathbf{W}^{*T}\boldsymbol{\mu}^{*T}\\
\qquad +2\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{W}\mathbf{A}_5(\mathbf{y}_{1:N})\mathbf{W}^{*T}\boldsymbol{\delta}_\epsilon^{*T} + 2\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{W}^*\mathbf{A}_6(\mathbf{y}_{1:N};\Psi)\mathbf{W}^T\boldsymbol{\mu}^T + 2\bar{\boldsymbol{\mu}}(\mathbf{y}_{1:N})\mathbf{W}^*\mathbf{A}_5(\mathbf{y}_{1:N};\Psi)\big(\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\big)^T\\
\qquad -2\boldsymbol{\mu}\mathbf{W}\mathbf{A}_6(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\boldsymbol{\mu}^{*T} - 2\big(\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\big)\mathbf{A}_5(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\boldsymbol{\mu}^{*T}\\
\qquad -2\boldsymbol{\mu}\mathbf{W}\mathbf{A}_5(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\boldsymbol{\delta}_\epsilon^{*T} - 2\big(\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\big)\mathbf{A}_4(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\boldsymbol{\delta}_\epsilon^{*T} + \mathrm{Tr}\Big\{\mathbf{A}_7(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\mathbf{W}^*\Big\} + \mathrm{Tr}\Big\{\mathbf{A}_9(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\mathbf{W}^*\Big\}\\
\qquad - \mathrm{Tr}\Big\{\Big(\mathbf{A}_8(\mathbf{y}_{1:N};\Psi) + \mathbf{A}_8(\mathbf{y}_{1:N};\Psi)^T\Big)\mathbf{W}^{*T}\mathbf{W}^*\Big\} + \sigma^2\,\mathrm{Tr}\Big\{\mathbf{A}_5(\mathbf{y}_{1:N};\Psi)\mathbf{W}^{*T}\mathbf{W}^*\Big\}\Big].
\end{cases}
$$

□

A.4.4. ID: Incomplete Data Setting

A.4.4.1. Proof of Theorem 27

Proof. Let the observation vector $\mathbf{Y}_t$ be partitioned into the subvector with d_o observed entries $\mathbf{Y}_t^o$ and the subvector with d_m missing entries $\mathbf{Y}_t^m$ such that $\mathbf{Y} = [\mathbf{Y}_t^o, \mathbf{Y}_t^m]$ and $d = d_o + d_m$. The joint probability function of the variables $\mathbf{Y}_{1:N}$, $\mathbf{X}_{1:N}$, $\mathbf{U}$ and $\mathbf{V}$ for Generalized Skew t-Student ID PPCA model is specified in Equation (3.32). Given the conditional time independence of the random vectors $\mathbf{Y}_t, \mathbf{X}_t|\mathbf{U}$ and $\mathbf{V}$, the E-step of the corresponding EM algorithm in the missing data setting can be formalized as follows

$$
\begin{aligned}
\tilde{Q}_{GStS,id}(\Psi, \Psi^*) &= \mathbb{E}_{\mathbf{Y}_{1:N}^m, \mathbf{X}_{1:N}, \mathbf{U}, \mathbf{V}|\mathbf{Y}_{1:N}, \Psi}\Big[\log \pi_{\mathbf{Y}_{1:N}, \mathbf{X}_{1:N}, \mathbf{U}, \mathbf{V}|\Psi^*}(\mathbf{Y}_{1:N}, \mathbf{X}_{1:N}, \mathbf{U}, \mathbf{V})\Big] \\
&= \frac{1}{\pi_{\mathbf{Y}_{1:N}^o|\Psi}(\mathbf{y}_{1:N}^o)} \int_{\mathbb{R}_+^d}\int_{\mathbb{R}_+^k} \Big\{ \log\left(\pi_{\mathbf{U},\mathbf{V}|\Psi^*}(\mathbf{u},\mathbf{v})\right) \pi_{\mathbf{U},\mathbf{V}|\Psi}(\mathbf{u},\mathbf{v}) \int_{\mathbb{R}^{N\times d_m}}\int_{\mathbb{R}^{N\times k}} \prod_{s=1}^N \pi_{\mathbf{Y}_s,\mathbf{X}_s|\mathbf{U},\mathbf{V},\Psi}(\mathbf{y}_s,\mathbf{x}_s) d\mathbf{x}_{1:N}\, d\mathbf{y}_{1:N}^m \Big\} d\mathbf{u}\, d\mathbf{v} \\
&+ \frac{1}{\pi_{\mathbf{Y}_{1:N}^o|\Psi}(\mathbf{y}_{1:N}^o)} \int_{\mathbb{R}^{d_+}}\int_{\mathbb{R}^{k_+}} \sum_{t=1}^N \int_{\mathbb{R}^{N\times d_m}}\int_{\mathbb{R}^{N\times k}} \Big\{ \log\left(\pi_{\mathbf{Y}_t,\mathbf{X}_t|\mathbf{U},\mathbf{V},\Psi^*}(\mathbf{y}_t,\mathbf{x}_t,)\right) \pi_{\mathbf{U},\mathbf{V}|\Psi}(\mathbf{u},\mathbf{v}) \prod_{s=1}^N \pi_{\mathbf{Y}_s,\mathbf{X}_s|\mathbf{U},\mathbf{V},\Psi}(\mathbf{y}_s,\mathbf{x}_s) \Big\} d\mathbf{x}_{1:N} d\, \mathbf{y}_{1:N}^m\, d\mathbf{u}\, d\mathbf{v}.
\end{aligned}
$$

Using Lemma A12, the probability function $\pi_{\mathbf{Y}_t,\mathbf{X}_t|\mathbf{U},\mathbf{V},\Psi}(\mathbf{y}_t,\mathbf{x}_t)$ can be factorised into the product of three functions

$$
\pi_{\mathbf{Y}_t,\mathbf{X}_t|\mathbf{U},\mathbf{V},\Psi}(\mathbf{y}_t,\mathbf{x}_t) = \pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U},\mathbf{V},\Psi}(\mathbf{x}_t) \cdot \pi_{\mathbf{Y}_t|\mathbf{U},\mathbf{V},\Psi}(\mathbf{y}_t) \cdot C(\mathbf{u},\mathbf{v};\Psi)
$$

Please refer to the result derived in Lemma A12 to find the definitions of the probability functions and the function $C(\mathbf{u},\mathbf{v};\Psi) : \mathbb{R}_+^d \times \mathbb{R}_+^k \longrightarrow \mathbb{R}$. In addition, the random vector $\mathbf{Y}_t|\mathbf{U},\mathbf{V},\Psi$ is Gaussian. Given the partition $\mathbf{Y}_t = [\mathbf{Y}_t^o, \mathbf{Y}_t^m]$ and using the chain rule of probabilities, we can derive the conditional probability density functions such that

$$
\pi_{\mathbf{Y}_t|\mathbf{U},\mathbf{V},\Psi}(\mathbf{y}_t) = \pi_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U},\mathbf{V},\Psi}(\mathbf{y}_t^m) \cdot \pi_{\mathbf{Y}_t^o|\mathbf{U},\mathbf{V},\Psi}(\mathbf{y}_t^o)
$$

where their characteristics are specified in Lemma A14. Therefore, we can further simplify the formulation of the $\tilde{Q}_{GStS,id}$ function

$$
\begin{aligned}
\tilde{Q}_{GStS,id}(\Psi,\Psi^*) &= \frac{1}{\pi_{\mathbf{Y}_{1:N}^o|\Psi}(\mathbf{y}_{1:N}^o)} \int_{\mathbb{R}^{d_+}}\int_{\mathbb{R}^{k_+}} \log\left(\pi_{\mathbf{U},\mathbf{V}|\Psi^*}(\mathbf{u},\mathbf{v})\right) \pi_{\mathbf{U},\mathbf{V}|\Psi}(\mathbf{u},\mathbf{v})\, C(\mathbf{u},\mathbf{v};\Psi)^N \left(\prod_{s=1}^N \int_{\mathbb{R}^{d_m}} \pi_{\mathbf{Y}_s|\mathbf{U},\mathbf{V},\Psi}(\mathbf{y}_s)\, d\mathbf{y}_s^m\right) d\mathbf{u}\, d\mathbf{v} \\
&+ \frac{1}{\pi_{\mathbf{Y}_{1:N}|\Psi}(\mathbf{y}_{1:N})} \int_{\mathbb{R}_+^d}\int_{\mathbb{R}_+^k} \Bigg\{ \pi_{\mathbf{U},\mathbf{V}|\Psi}(\mathbf{u},\mathbf{v})\, C(\mathbf{u},\mathbf{v};\Psi)^N \prod_{s=1}^N \pi_{\mathbf{Y}_s^o|\mathbf{U},\mathbf{V},\Psi}(\mathbf{y}_s^o) \\
&\times \Bigg(\sum_{t=1}^N \int_{\mathbb{R}^{\times d_m}} \pi_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U},\mathbf{V},\Psi}(\mathbf{y}_t^m) \underbrace{\int_{\mathbb{R}^k} \log\left(\pi_{\mathbf{Y}_t,\mathbf{X}_t|\mathbf{U},\mathbf{V},\Psi^*}(\mathbf{y}_t,\mathbf{x}_t)\right) \pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U},\mathbf{V},\Psi}(\mathbf{x}_t)\, d\mathbf{x}_t}_{w(\mathbf{y}_t,\mathbf{u},\mathbf{v};\Psi,\Psi^*) \text{ in Lemma A13}}\, d\mathbf{y}_t^m \Bigg) \Bigg\} d\mathbf{u}\, d\mathbf{v} \\
&= \frac{1}{\pi_{\mathbf{Y}_{1:N}^o|\Psi}(\mathbf{y}_{1:N}^o)} \int_{\mathbb{R}^{d_+}}\int_{\mathbb{R}^{k_+}} \log\left(\pi_{\mathbf{U},\mathbf{V}|\Psi^*}(\mathbf{u},\mathbf{v})\right) \pi_{\mathbf{U},\mathbf{V}|\Psi}(\mathbf{u},\mathbf{v})\, C(\mathbf{u},\mathbf{v};\Psi)^N \left(\prod_{s=1}^N \pi_{\mathbf{Y}_s^o|\mathbf{U},\mathbf{V},\Psi}(\mathbf{y}_s^o)\right) d\mathbf{u}\, d\mathbf{v} \\
&+ \frac{1}{\pi_{\mathbf{Y}_{1:N}^o|\Psi}(\mathbf{y}_{1:N}^o)} \int_{\mathbb{R}_+^d}\int_{\mathbb{R}_+^k} \Bigg\{ \pi_{\mathbf{U},\mathbf{V}|\Psi}(\mathbf{u},\mathbf{v})\, C(\mathbf{u},\mathbf{v};\Psi)^N \prod_{s=1}^N \pi_{\mathbf{Y}_s^o|\mathbf{U},\mathbf{V},\Psi}(\mathbf{y}_s^o) \Bigg(\sum_{t=1}^N \underbrace{\int_{\mathbb{R}^{d_m}} \pi_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U},\mathbf{V},\Psi}(\mathbf{y}_t^m) w(\mathbf{y}_t,\mathbf{u},\mathbf{v};\Psi,\Psi^*)\, d\mathbf{y}_t^m}_{v(\mathbf{y}_t,\mathbf{u},\mathbf{v};\Psi,\Psi^*)} \Bigg) \Bigg\} d\mathbf{u}\, d\mathbf{v}
\end{aligned}
$$

for the function $v : \mathbb{R}^{d_o} \times \mathbb{R}_+^d \times \mathbb{R}_+^k \longrightarrow \mathbb{R}$ defined in Lemma A15. Let us denote for the convenience of the notation

$$
\begin{aligned}
h(\mathbf{y}_{1:N}^o,\mathbf{u},\mathbf{v},\Psi) &:= \pi_{\mathbf{U},\mathbf{V}|\Psi}(\mathbf{u},\mathbf{v})\, C(\mathbf{u},\mathbf{v};\Psi)^N \left(\prod_{s=1}^N \pi_{\mathbf{Y}_s^o|\mathbf{U},\mathbf{V},\Psi}(\mathbf{y}_s^o)\right) = \pi_{\mathbf{U},\mathbf{V}|\Psi}(\mathbf{u},\mathbf{v}) \left(\prod_{s=1}^N \pi_{\mathbf{Y}_s^o|\mathbf{U},\mathbf{V},\Psi}(\mathbf{y}_s^o)\right) \\
&\times (\sigma^2)^{\frac{kN}{2}} \left|\mathbf{D}_\epsilon\right|^{\frac{N}{2}} \left|\mathbf{D}_{x,t}\right|^{\frac{N}{2}} \left|\mathbf{M}\right|^{-\frac{N}{2}} \left|\mathbf{N}\right|^{-\frac{N}{2}} \exp\Big\{ -\frac{N}{2}\delta_x \left(\mathbf{D}_x^{-1} - \sigma^2\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\mathbf{N}^{-1}\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1} - \sigma^2\mathbf{M}^{-1}\right)\delta_x^T \Big\}.
\end{aligned}
$$

The E-step of the Generalized Skew t-Student ID PPCA is the following

$$
\tilde{Q}_{GStS,id}(\Psi,\Psi^*) = \frac{1}{\pi_{\mathbf{Y}_{1:N}^o|\Psi}(\mathbf{y}_{1:N}^o)} \int_{\mathbb{R}_+^d}\int_{\mathbb{R}_+^k} \Big\{ \Big(\log\left(\pi_{\mathbf{U},\mathbf{V}|\Psi^*}(\mathbf{u},\mathbf{v}) + \sum_{t=1}^N v(\mathbf{y}_t^o,\mathbf{u},\mathbf{v};\Psi,\Psi^*)\right) h(\mathbf{y}_{1:N},\mathbf{u},\mathbf{v};\Psi) \Big\} d\mathbf{u}\, d\mathbf{v}. \quad \text{(A18)}
$$

We follow the same steps as in the proof of Theorem 15 in Appendix A.4.3.1 and recall the definitions the quantile functions T_ϵ and $T_x : [0,1]$. The formulation of the E-step of Generalized Skew t-Student ID PPCA can be reduced to the 2-dimensional integration on the unit hypercube $[0,1]^2$, that is

$$\tilde{Q}_{GStS,id}(\Psi,\Psi^*) = \frac{1}{\pi_{\mathbf{Y}^o_{1:N}|\Psi}(\mathbf{y}^o_{1:N})}\int_0^1\int_0^1\left\{\left(-\log\left|\frac{\partial T_\epsilon(s_\epsilon)}{\partial s_\epsilon}\right|-\log\left|\frac{\partial T_x(s_x)}{\partial s_x}\right|+\sum_{t=1}^N\tilde{v}(\mathbf{y}^o_t,s_\epsilon,s_x;\Psi,\Psi^*)\right)m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\right\}ds_\epsilon\,ds_x$$

for the function $m:\mathbb{R}^{N\times d_o}\times[0,1]^2\longrightarrow\mathbb{R}$ such that

$$m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi) = \left(\prod_{t=1}^N\pi_{\mathbf{Y}^o_t|s_\epsilon,s_x,\Psi}(\mathbf{y}^o_t)\right)(\sigma^2)^{\frac{k}{2}}\left|\mathbf{D}_\epsilon\right|^{\frac{N}{2}}\left|\mathbf{D}_x\right|^{\frac{N}{2}}\left|\mathbf{M}\right|^{-\frac{N}{2}}\left|\mathbf{N}\right|^{-\frac{N}{2}}e^{-\frac{N}{2}\delta_x\left(\mathbf{D}_x^{-1}-\sigma^2\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\mathbf{N}^{-1}\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}_i^{-1}-\sigma^2\mathbf{M}^{-1}\right)\delta_x^T}$$

and $\tilde{v}(\mathbf{y}^o_t,s_\epsilon,s_x;\Psi,\Psi^*) := v\Big(\mathbf{y}^o_t,T_\epsilon(s_\epsilon),T_x(s_x);\Psi,\Psi^*\Big)$. □

A.4.4.2. Proof of Theorem 28

Proof. The M-step of the EM algorithm evaluates the maximizers of the function $\tilde{Q}_{GStS,id}(\Psi,\Psi^*)$ defined in Theorem 27 with respect to the vector of static parameters $\Psi^* = \left[\boldsymbol{\mu}^*,\mathbf{W}^*,\sigma^{*2},\delta^*_\epsilon,\delta_x\right]$. Recall that the only element of the function $\tilde{Q}_{GStS,id}$ which contains components of the vector Ψ^* is the function $\tilde{v}$ which is a component of the integral $I_1(\mathbf{y}^o_{1:N};\Psi,\Psi^*)$. The maximizers of the function $\tilde{Q}_{GStS,id}$ with respect to the vector of static parameters Ψ^* are the solution to the set of equations $\nabla_{\Psi^*}\tilde{Q}_{GStS,id}(\Psi,\Psi^*)=\mathbf{0}$ and therefore require specification of the gradient $\frac{\partial}{\partial\Psi^*}\tilde{v}(\mathbf{y}^o_t,s_{\epsilon,t},s_{x,t};\Psi,\Psi^*)$. The gradient $\frac{\partial}{\partial\Psi^*}\tilde{v}(\mathbf{y}_t s_\epsilon,s_x;\Psi,\Psi^*)$ is similar to the equivalent gradient of the function $\tilde{v}$ from the proof of Theorem 26 in Appendix A.4.1.3 for Generalized Skew t-Student ID PPCA model. Therefore, its sum over all realisations of the observation vector $\mathbf{Y}_t$ is derived as following

$$\begin{aligned}\sum_{t=1}^N\frac{\partial}{\partial\boldsymbol{\mu}^*}\tilde{v}(\mathbf{y}^o_t,s_\epsilon,s_x;\Psi,\Psi^*) &= \frac{N}{\sigma^{*2}}\Big[\bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon-\boldsymbol{\mu}^*\mathbf{D}_\epsilon-\boldsymbol{\delta}^*_\epsilon-\bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon\\&\quad+\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon+\boldsymbol{\delta}_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon-\sigma^2\boldsymbol{\delta}_x\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon\Big]\end{aligned}$$

$$\begin{aligned}\sum_{t=1}^N\frac{\partial}{\partial\mathbf{W}^*}\tilde{v}(\mathbf{y}^o_t,s_\epsilon,s_x;\Psi,\Psi^*) &= \frac{N}{\sigma^{*2}}\Big[\mathbf{D}_\epsilon\bar{\mathbf{S}}^m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}-(\mathbf{D}_\epsilon\boldsymbol{\mu}^{*T}+\boldsymbol{\delta}^{*T}_\epsilon)\bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\\&\quad+(\bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon-\boldsymbol{\mu}^*\mathbf{D}_\epsilon-\boldsymbol{\delta}^*_\epsilon)^T(-\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}-\boldsymbol{\delta}_\epsilon\mathbf{W}+\sigma^2\boldsymbol{\delta}_x)\mathbf{M}^{-1}\\&\quad-\mathbf{D}_\epsilon\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\bar{\mathbf{S}}^m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\\&\quad-\mathbf{D}_\epsilon\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)^T(-\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}-\boldsymbol{\delta}_\epsilon\mathbf{W}+\sigma^2\boldsymbol{\delta}_x)\mathbf{M}^{-1}\\&\quad-\mathbf{D}_\epsilon\mathbf{W}^*\mathbf{M}^{-1}(-\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}-\boldsymbol{\delta}_\epsilon\mathbf{W}+\sigma^2\boldsymbol{\delta}_x)^T\bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\\&\quad-\mathbf{D}_\epsilon\mathbf{W}^*\mathbf{M}^{-1}\Big(-\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}-\boldsymbol{\delta}_\epsilon\mathbf{W}+\sigma^2\boldsymbol{\delta}_x\Big)^T\Big(-\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}-\boldsymbol{\delta}_\epsilon\mathbf{W}+\sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}^{-1}-\sigma^2\mathbf{D}_\epsilon\mathbf{W}^*\mathbf{M}^{-1}\Big]\end{aligned}$$

$$\begin{aligned}\sum_{t=1}^N\frac{\partial}{\partial\sigma^{*2}}\tilde{v}(\mathbf{y}^o_t,s_\epsilon,s_x;\Psi,\Psi^*) &= -\frac{dN}{2\sigma^{*2}}-\frac{N}{2(\sigma^{*2})^2}\Big[-\operatorname{Tr}\Big\{\bar{\mathbf{S}}^m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\Big\}+2\bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)(\mathbf{D}_\epsilon\boldsymbol{\mu}^{*T}+\boldsymbol{\delta}^{*T}_\epsilon)\\&\quad-\boldsymbol{\mu}^*\mathbf{D}_\epsilon\boldsymbol{\mu}^{*T}-2\boldsymbol{\mu}^*\boldsymbol{\delta}^{*T}_\epsilon-\boldsymbol{\delta}^*_\epsilon\mathbf{D}_\epsilon^{-1}\boldsymbol{\delta}^{*T}_\epsilon+2\operatorname{Tr}\Big\{\bar{\mathbf{S}}^m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon\Big\}\\&\quad-2\Big(\bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}-\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}-\boldsymbol{\delta}_\epsilon\mathbf{W}+\sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}^{-1}\mathbf{W}^{*T}(\boldsymbol{\mu}^*\mathbf{D}_\epsilon+\boldsymbol{\delta}^*_\epsilon)^T\\&\quad+2\bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}^*\mathbf{M}^{-1}(-\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}-\boldsymbol{\delta}_\epsilon\mathbf{W}+\sigma^2\boldsymbol{\delta}_x)^T-\sigma^2\operatorname{Tr}\Big\{\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon\mathbf{W}^*\Big\}\\&\quad-\operatorname{Tr}\Big\{\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\bar{\mathbf{S}}^m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon\mathbf{W}^*\Big\}\\&\quad-\operatorname{Tr}\Big\{\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)^T\Big(-\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}-\boldsymbol{\delta}_\epsilon\mathbf{W}+\sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon\mathbf{W}^*\Big\}\\&\quad-\operatorname{Tr}\Big\{\mathbf{M}^{-1}\Big(-\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}-\boldsymbol{\delta}_\epsilon\mathbf{W}+\sigma^2\boldsymbol{\delta}_x\Big)^T\bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon\mathbf{W}^*\Big\}\\&\quad-\operatorname{Tr}\Big\{\mathbf{M}^{-1}\Big(-\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}-\boldsymbol{\delta}_\epsilon\mathbf{W}+\sigma^2\boldsymbol{\delta}_x\Big)^T\Big(-\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}-\boldsymbol{\delta}_\epsilon\mathbf{W}+\sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon\mathbf{W}^*\Big\}\Big]\end{aligned}$$

$$\sum_{t=1}^{N}\frac{\partial}{\partial\delta_\epsilon^*}\tilde{v}\Big(\mathbf{y}_t^o,s_\epsilon,s_x;\Psi,\Psi^*\Big)=\frac{N}{\sigma^{*2}}\Big[\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)-\boldsymbol{\mu}^*-\delta_\epsilon^*\mathbf{D}_\epsilon^{-1}-\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}$$
$$+\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}+\delta_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}-\sigma^2\delta_x\mathbf{M}^{-1}\mathbf{W}^{*T}\Big]$$

$$\sum_{t=1}^{N}\frac{\partial}{\partial\delta_x^*}\tilde{v}(\mathbf{y}_t^o,s_\epsilon,s_x;\Psi,\Psi^*)=N\Big[-\delta_x^*\mathbf{D}_x^{-1}+\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}-\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}-\delta_\epsilon\mathbf{W}\mathbf{M}^{-1}+\sigma^2\delta_x\mathbf{M}^{-1}\Big]$$

where $\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)_{1\times d}=\frac{1}{N}\sum_{t=1}^{N}\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_\epsilon,S_x,\Psi}[\mathbf{Y}_t]$ and $\bar{\mathbf{S}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)_{d\times d}=\frac{1}{N}\sum_{t=1}^{N}\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,S_\epsilon,S_x,\Psi}[\mathbf{Y}_t^T\mathbf{Y}_t]$. □

A.4.4.3. Proof of Corollary 22

Proof. Given the system of equations specified in Theorem 28, the maximizers of the function $\tilde{Q}_{GStS,id}$ with respect to $\boldsymbol{\mu}^*$,σ^{*2}, δ_ϵ^* and δ_x^* have closed form solutions defined by sequences of specific 2-dimensional integration problems determined as follows

$$\frac{\partial}{\partial\boldsymbol{\mu}^*}\tilde{Q}_{GStS,id}(\Psi,\Psi^*)=\mathbf{0}\Longleftrightarrow\int_0^1\int_0^1\Big(\sum_{t=1}^{N}\frac{\partial}{\partial\boldsymbol{\mu}^*}\tilde{v}(\mathbf{y}_t,s_\epsilon,s_x;\Psi,\Psi^*)\Big)\,m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x=\mathbf{0}$$
$$\Longleftrightarrow\int_0^1\int_0^1\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\,m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x-\boldsymbol{\mu}^*\int_0^1\int_0^1\mathbf{D}_\epsilon\,m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x$$
$$-\delta_\epsilon^*\int_0^1\int_0^1 m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x-\int_0^1\int_0^1\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\,\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon\,m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x$$
$$+\boldsymbol{\mu}\int_0^1\int_0^1\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon\,m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x+(\delta_\epsilon\mathbf{W}-\sigma^2\delta_x)\int_0^1\int_0^1\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon\,m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x=\mathbf{0}$$

$$\frac{\partial}{\partial\delta_\epsilon^*}\tilde{Q}_{GStS,id}(\Psi,\Psi^*)=\mathbf{0}\Longleftrightarrow\int_0^1\int_0^1\Big(\sum_{t=1}^{N}\frac{\partial}{\partial\delta_\epsilon^*}\tilde{v}(\mathbf{y}_t,s_\epsilon,s_x;\Psi,\Psi^*)\Big)\,m(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x=\mathbf{0}$$
$$\Longleftrightarrow\int_0^1\int_0^1\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x-\boldsymbol{\mu}^*\int_0^1\int_0^1 m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x$$
$$-\delta_\epsilon^*\int_0^1\int_0^1\mathbf{D}_\epsilon^{-1}\,m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x-\int_0^1\int_0^1\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\,m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x\,\mathbf{W}^{*T}$$
$$+\boldsymbol{\mu}\int_0^1\int_0^1\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\,m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x\,\mathbf{W}^{*T}+(\delta_\epsilon\mathbf{W}-\sigma^2\delta_x)\int_0^1\int_0^1\mathbf{M}^{-1}\,m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x\,\mathbf{W}^{*T}=\mathbf{0}$$

$$\frac{\partial}{\partial\delta_x^*}\tilde{Q}_{GStS,id}(\Psi,\Psi^*)=\mathbf{0}\Longleftrightarrow\int_0^1\int_0^1\Big(\sum_{t=1}^{N}\frac{\partial}{\partial\delta_x^*}\tilde{v}(\mathbf{y}_t,s_\epsilon,s_x;\Psi,\Psi^*)\Big)\,m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x=\mathbf{0}$$
$$\Longleftrightarrow-\delta_x^*\int_0^1\int_0^1\mathbf{D}_x^{-1}\,m(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x+\int_0^1\int_0^1\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\,m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x$$
$$-\boldsymbol{\mu}\int_0^1\int_0^1\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\,m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x-(\delta_\epsilon\mathbf{W}-\sigma^2\delta_x)\int_0^1\int_0^1\mathbf{M}^{-1}\,m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x=\mathbf{0}$$

$$\frac{\partial}{\partial\sigma^{*2}}\tilde{Q}_{GStS,id}(\Psi,\Psi^*)=\mathbf{0}\Longleftrightarrow\int_0^1\int_0^1\Big(\sum_{t=1}^{N}\frac{\partial}{\partial\sigma^{*2}}\tilde{v}(\mathbf{y}_t,s_\epsilon,s_x;\Psi,\Psi^*)\Big)\,m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x=\mathbf{0}$$
$$\Longleftrightarrow-\frac{d}{2\sigma^{*2}}\int_0^1\int_0^1 m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x-\frac{1}{2(\sigma^{*2})^2}\Big[-\mathrm{Tr}\Big\{\int_0^1\int_0^1\bar{\mathbf{S}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\,m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x\Big\}$$
$$+2\int_0^1\int_0^1\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\,m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x\,\boldsymbol{\mu}^{*T}+2\int_0^1\int_0^1\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\,m(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x\,\delta_\epsilon^{*T}$$
$$-\boldsymbol{\mu}^*\int_0^1\int_0^1\mathbf{D}_\epsilon\,m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x\Big\}\boldsymbol{\mu}^{*T}-2\boldsymbol{\mu}^*\delta_\epsilon^{*T}\int_0^1\int_0^1 m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x\Big\}$$
$$-\delta_\epsilon^*\int_0^1\int_0^1\mathbf{D}_\epsilon^{-1}\,m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x\,\delta_\epsilon^{*T}$$
$$+2\,\mathrm{Tr}\Big\{\int_0^1\int_0^1\bar{\mathbf{S}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon\,m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x\Big\}$$
$$-2\int_0^1\int_0^1\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon\,m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x\boldsymbol{\mu}^{*T}$$
$$+2\boldsymbol{\mu}\int_0^1\int_0^1\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon\,m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x\boldsymbol{\mu}^{*T}+2(\delta_\epsilon\mathbf{W}-\sigma^2\delta_x)\int_0^1\int_0^1\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon\,m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x\,\boldsymbol{\mu}^{*T}$$

$$
\begin{aligned}
&- 2\int_0^1\int_0^1 \bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\, m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\, ds_\epsilon\, ds_x\, \mathbf{W}^{*T}\boldsymbol{\delta}_\epsilon^{*T}\\
&+ 2\boldsymbol{\mu}\int_0^1\int_0^1 \mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\, m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\, ds_\epsilon\, ds_x\Big\}\mathbf{W}^{*T}\boldsymbol{\delta}_\epsilon^{*T}\\
&+ 2(\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x)\int_0^1\int_0^1 \mathbf{M}^{-1}\, m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\, ds_\epsilon\, ds_x\mathbf{W}^{*T}\boldsymbol{\delta}_\epsilon^{*T}\\
&- 2\int_0^1\int_0^1 \bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\, m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\, ds_\epsilon\, ds_x\, \boldsymbol{\mu}^T\\
&- 2\int_0^1\int_0^1 \bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}^*\mathbf{M}^{-1}\, m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\, ds_\epsilon\, ds_x(\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x)^T\\
&- \sigma^2\,\mathrm{Tr}\left\{\int_0^1\int_0^1 \mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\, ds_\epsilon\, ds_x\mathbf{W}^*\right\}\\
&- \mathrm{Tr}\left\{\int_0^1\int_0^1 \mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\bar{\mathbf{S}}^m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon\, m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\, ds_\epsilon\, ds_x\mathbf{W}^*\right\}\\
&+ \mathrm{Tr}\left\{\int_0^1\int_0^1 \mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)^T(\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}+\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x)\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon\, m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\, ds_\epsilon\, ds_x\mathbf{W}^*\right\}\\
&+ \mathrm{Tr}\left\{\int_0^1\int_0^1 \mathbf{M}^{-1}(\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}+\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x)^T\bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon\, m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\, ds_\epsilon\, ds_x\mathbf{W}^*\right\}\\
&- \mathrm{Tr}\left\{\int_0^1\int_0^1 \mathbf{M}_t^{-1}(\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}+\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x)^T(\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W}+\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x)\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\, m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\, ds_\epsilon\, ds_x\mathbf{W}^*\right\}\right] = 0.
\end{aligned}
$$

The iterative updates of the parameters $\boldsymbol{\mu}$, $\boldsymbol{\delta}_\epsilon$, $\boldsymbol{\delta}_x$ and σ^2, being the maximizers of the function $Q_{GStS,id}$ with respect to $\boldsymbol{\mu}^*$, $\boldsymbol{\delta}^*_\epsilon$, $\boldsymbol{\delta}^*_x$ and σ^{*2}, are obtained in the closed forms determined by various integration problems specified on 2-dimensional hypercube $[0,1]^2$. Recall that the equations defining these maximizers are dependent on $\mathbf{W}^*$. The maximizer which correspond to the parameter $\mathbf{W}^*$ is more difficult to obtain since the partial derivative of the function $Q_{GStS,id}$ with respect to $\mathbf{W}^*$ requires integrating the matrix products which contain $\mathbf{W}^*$ and cannot be further simplified, that is

$$
\begin{aligned}
&\frac{\partial}{\partial\mathbf{W}^*}\tilde{Q}_{GStS,id}(\Psi,\Psi^*)=\mathbf{0}\iff\int_0^1\int_0^1\Big(\sum_{t=1}^N\frac{\partial}{\partial\mathbf{W}^*}\tilde{v}(\mathbf{y}_t,s_\epsilon,s_x;\Psi,\Psi^*)\Big)\, m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\, ds_\epsilon\, ds_x=\mathbf{0}\\
&\iff\int_0^1\int_0^1\mathbf{D}_\epsilon\bar{\mathbf{S}}^m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\, m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\, ds_\epsilon\, ds_x-\int_0^1\int_0^1\mathbf{D}_\epsilon\,\boldsymbol{\mu}^{*T}\bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N};\Psi)\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\, m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\, ds_\epsilon\, ds_x\\
&-\boldsymbol{\delta}_\epsilon^{*T}\int_0^1\int_0^1\bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\, m(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)\, ds_\epsilon\, ds_x\\
&-\int_0^1\int_0^1\mathbf{D}_\epsilon(\bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)-\boldsymbol{\mu}^*)^T(\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}+\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x)\mathbf{M}^{-1}\, m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\, ds_\epsilon\, ds_x\\
&+\boldsymbol{\delta}_\epsilon^{*T}\boldsymbol{\mu}\int_0^1\int_0^1\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\, m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\, ds_\epsilon\, ds_x\\
&\quad+\boldsymbol{\delta}_\epsilon^{*T}(\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x)\int_0^1\int_0^1\mathbf{M}^{-1}\, m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\, ds_\epsilon\, ds_x\\
&-\int_0^1\int_0^1\mathbf{D}_\epsilon\,\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\bar{\mathbf{S}}^m(\mathbf{y}_{1:N}6o,s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\, m(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)\, ds_\epsilon\, ds_x\\
&+\int_0^1\int_0^1\mathbf{D}_\epsilon\,\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)^T\Big(\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}+\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}^{-1}\, m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\, ds_\epsilon\, ds_x\\
&+\int_0^1\int_0^1\mathbf{D}_\epsilon\,\mathbf{W}^*\mathbf{M}^{-1}\Big(\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}+\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\Big)^T\bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\, m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\, ds_\epsilon\, ds_x\\
&-\int_0^1\int_0^1\mathbf{D}_\epsilon\,\mathbf{W}^*\mathbf{M}^{-1}\Big(\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}+\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\Big)^T\Big(\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W}+\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}^{-1}\, m(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)\, ds_\epsilon\, ds_x\\
&-\sigma^2\int_0^1\int_0^1\mathbf{D}_\epsilon\,\mathbf{W}^*\mathbf{M}^{-1}\, m(\mathbf{y}_{1:N},s_\epsilon,s_x;\Psi)\, ds_\epsilon\, ds_x=\mathbf{0}
\end{aligned}
$$

The maximizer with respect to $\mathbf{W}^*$ requires solving a root finding problem $\frac{\partial Q(\Psi,\Psi^*)}{\partial\mathbf{W}^*}=\mathbf{0}$. Recall that $\frac{\partial Q(\Psi,\Psi^*)}{\partial\mathbf{W}^*}$ contains elements of $\boldsymbol{\mu}^*$ and $\boldsymbol{\delta}^*_\epsilon$. By denoting $f:\mathbb{R}^{d\times k}\to\mathbb{R}^{d\times k}$ such that $f(\mathbf{W}^*;\Psi,\Psi^*):=\frac{\partial Q(\Psi,\Psi^*)}{\partial\mathbf{W}^*}$ equals to

$$
\begin{aligned}
f(\mathbf{W}^*;\Psi,\Psi^*)=\;&\mathbf{A}_{14}(\mathbf{y}_{1:N};\Psi)-\mathbf{A}_{15}(\mathbf{y}_{1:N};\Psi,\Psi^*)-\boldsymbol{\delta}_\epsilon^{*T}\mathbf{A}_8(\mathbf{y}_{1:N};\Psi)-\mathbf{A}_{16}(\mathbf{y}_{1:N};\Psi,\Psi^*)+\boldsymbol{\delta}_\epsilon^{*T}\boldsymbol{\mu}\mathbf{A}_{11}(\mathbf{y}_{1:N};\Psi)\\
&+\boldsymbol{\delta}_\epsilon^{*T}(\boldsymbol{\delta}\mathbf{W}-\sigma^2\mathbf{W})\mathbf{A}_4(\mathbf{y}_{1:N};\Psi)-\mathbf{A}_{17}(\mathbf{y}_{1:N};\Psi,\Psi^*)+\mathbf{A}_{18}(\mathbf{y}_{1:N};\Psi,\Psi^*)+\mathbf{A}_{19}(\mathbf{y}_{1:N};\Psi,\Psi^*)-\mathbf{A}_{20}(\mathbf{y}_{1:N};\Psi,\Psi^*)\\
&-\sigma^2\mathbf{A}_{12}(\mathbf{y}_{1:N};\Psi,\Psi^*),
\end{aligned}
$$

the system of equations $\nabla_{\Psi^*}\tilde{Q}_{GStS,id}(\Psi,\Psi^*)=\mathbf{0}$ reduces to

$$
\begin{cases}
\boldsymbol{\mu}^* = \Big[\mathbf{A}_6(\mathbf{y}^o_{1:N};\Psi) - \mathbf{A}_{10}(\mathbf{y}^o_{1:N};\Psi,\Psi^*) - \boldsymbol{\delta}^*_\epsilon A_0(\mathbf{y}^o_{1:N};\Psi) + \boldsymbol{\mu}\mathbf{A}_{13}(\mathbf{y}^o_{1:N};\Psi,\Psi^*) + \Big(\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\Big)\mathbf{A}_{12}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)^T\Big]\mathbf{A}_1(\mathbf{y}^o_{1:N};\Psi)^{-1}\\
\boldsymbol{\delta}^*_\epsilon = \Big[\mathbf{A}_5(\mathbf{y}^o_{1:N};\Psi) - \boldsymbol{\mu}^* A_0(\mathbf{y}^o_{1:N};\Psi) - \mathbf{A}_8(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}^{*T} + \boldsymbol{\mu}\mathbf{A}_{11}(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}^{*T} + (\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x)\mathbf{A}_4(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}^{*T}\Big]\mathbf{A}_2(\mathbf{y}^o_{1:N};\Psi)^{-1}\\
\boldsymbol{\delta}^*_x = \Big[\mathbf{A}_8(\mathbf{y}^o_{1:N};\Psi) - \boldsymbol{\mu}\mathbf{A}_{11}(\mathbf{y}^o_{1:N};\Psi) - (\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x)\mathbf{A}_4(\mathbf{y}^o_{1:N};\Psi)\Big]\mathbf{A}_3(\mathbf{y}^o_{1:N},\Psi)^{-1}\\
\sigma^{*2} = \frac{1}{dA_0(\mathbf{y}^o_{1:N};\Psi)}\Big[A_{21}(\mathbf{y}^o_{1:N};\Psi) + \boldsymbol{\mu}^*\mathbf{A}_1(\mathbf{y}^o_{1:N};\Psi)\boldsymbol{\mu}^{*T} + \boldsymbol{\delta}^*_\epsilon\mathbf{A}_2(\mathbf{y}^o_{1:N};\Psi)\boldsymbol{\delta}^{*T}_\epsilon + 2\mathbf{A}_9(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\boldsymbol{\mu}^T\\
\qquad +2\Big(\mathbf{A}_{10}(\mathbf{y}^o_{1:N};\Psi,\Psi^*) - \mathbf{A}_6(\mathbf{y}^o_{1:N};\Psi) - \boldsymbol{\mu}\mathbf{A}_{13}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\Big)\boldsymbol{\mu}^{*T} - 2\operatorname{Tr}\Big\{A_{14}(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}^{*T}\Big\} + \sigma^2\operatorname{Tr}\Big\{\mathbf{A}_{12}(\mathbf{y}^o_{1:N};\Psi)^T\mathbf{W}^*\Big\}\\
\qquad +\operatorname{Tr}\Big\{\mathbf{A}_{17}(\mathbf{y}^o_{1:N};\Psi)^T\mathbf{W}^*\Big\} - \operatorname{Tr}\Big\{\mathbf{A}_{18}(\mathbf{y}^o_{1:N};\Psi)^T\mathbf{W}^*\Big\} - \operatorname{Tr}\Big\{\mathbf{A}_{19}(\mathbf{y}^o_{1:N};\Psi)^T\mathbf{W}^*\Big\} + \operatorname{Tr}\Big\{\mathbf{A}_{20}(\mathbf{y}^o_{1:N};\Psi)^T\mathbf{W}^*\Big\}\\
\qquad +2\Big(\boldsymbol{\mu}^* A_0(\mathbf{y}^o_{1:N};\Psi) - \mathbf{A}_5(\mathbf{y}^o_{1:N};\Psi) + \mathbf{A}_8(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}^{*T} - \boldsymbol{\mu}\mathbf{A}_{11}(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}^{*T}\Big)\boldsymbol{\delta}^{*T}_\epsilon\\
\qquad +2\Big(\mathbf{A}_7(\mathbf{y}^o_{1:N},\Psi,\Psi^*) - \boldsymbol{\mu}^*\mathbf{A}_{12}(\mathbf{y}^o_{1:N};\Psi,\Psi^*) - \boldsymbol{\delta}^*_\epsilon\mathbf{W}^*\mathbf{A}_4(\mathbf{y}^o_{1:N};\Psi)\Big)\big(\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\big)^T\Big]\\
f(\mathbf{W}^*;\Psi,\Psi^*) = \mathbf{0}.
\end{cases}
$$

for

$$
\begin{aligned}
&A_0(\mathbf{y}^o_{1:N};\Psi)_{1\times 1} := \int_0^1\int_0^1 m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x\\
&\mathbf{A}_1(\mathbf{y}^o_{1:N};\Psi)_{d\times d} := \int_0^1\int_0^1 \mathbf{D}_\epsilon\, m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x\\
&\mathbf{A}_2(\mathbf{y}^o_{1:N};\Psi)_{d\times d} := \int_0^1\int_0^1 \mathbf{D}_\epsilon^{-1}\, m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x\\
&\mathbf{A}_3(\mathbf{y}_{1:N};\Psi)_{k\times k} := \int_0^1\int_0^1 \mathbf{D}_x^{-1}\, m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x\\
&\mathbf{A}_4(\mathbf{y}^o_{1:N};\Psi)_{k\times k} := \int_0^1\int_0^1 \mathbf{M}_t^{-1}\, m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x\\
&\mathbf{A}_5(\mathbf{y}^o_{1:N};\Psi)_{1\times d} := \int_0^1\int_0^1 \bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x\\
&\mathbf{A}_6(\mathbf{y}^o_{1:N};\Psi)_{1\times d} := \int_0^1\int_0^1 \bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\, m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x\\
&\mathbf{A}_7(\mathbf{y}^o_{1:N};\Psi,\Psi^*)_{1\times k} := \int_0^1\int_0^1 \bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}^*\mathbf{M}^{-1}\, m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x\\
&\mathbf{A}_8(\mathbf{y}^o_{1:N};\Psi)_{1\times k} := \int_0^1\int_0^1 \bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\, m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x\\
&\mathbf{A}_9(\mathbf{y}^o_{1:N};\Psi,\Psi^*)_{1\times d} := \int_0^1\int_0^1 \bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\, m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x\\
&\mathbf{A}_{10}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)_{1\times d} := \int_0^1\int_0^1 \bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon\, m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x\\
&\mathbf{A}_{11}(\mathbf{y}^o_{1:N};\Psi)_{d\times k} := \int_0^1\int_0^1 \mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\, m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x\\
&\mathbf{A}_{12}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)_{d\times k} := \int_0^1\int_0^1 \mathbf{D}_\epsilon\mathbf{W}^*\mathbf{M}^{-1}\, m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x\\
&\mathbf{A}_{13}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)_{d\times d} := \int_0^1\int_0^1 \mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon\, m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x\\
&\mathbf{A}_{14}(\mathbf{y}^o_{1:N};\Psi)_{d\times k} := \int_0^1\int_0^1 \mathbf{D}_\epsilon\bar{\mathbf{S}}^m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\, m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x\\
&\mathbf{A}_{15}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)_{d\times k} := \int_0^1\int_0^1 \mathbf{D}_\epsilon\boldsymbol{\mu}^{*T}\bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\, m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x\\
&\mathbf{A}_{16}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)_{d\times k} := \int_0^1\int_0^1 \mathbf{D}_\epsilon\big(\bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi) - \boldsymbol{\mu}^*\big)^T(\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W} + \boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x)\mathbf{M}^{-1}\, m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x\\
&\mathbf{A}_{17}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)_{k\times d} := \int_0^1\int_0^1 \mathbf{D}_\epsilon\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\bar{\mathbf{S}}^m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}_t^{-1}\, m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x\\
&\mathbf{A}_{18}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)_{k\times d} := \int_0^1\int_0^1 \mathbf{D}_\epsilon\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)^T\Big(\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W} + \boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}^{-1}\, m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x\\
&\mathbf{A}_{19}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)_{k\times d} := \int_0^1\int_0^1 \mathbf{D}_\epsilon\mathbf{W}^*\mathbf{M}^{-1}\Big(\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W} + \boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\Big)^T\bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\, m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x\\
&\mathbf{A}_{20}(\mathbf{y}^o_{1:N};\Psi,\Psi^*)_{k\times d} := \int_0^1\int_0^1 \mathbf{D}_\epsilon\mathbf{W}^*\mathbf{M}^{-1}\Big(\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W} + \boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\Big)^T\Big(\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W} + \boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}^{-1}\, m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x\\
&A_{21}(\mathbf{y}^o_{1:N};\Psi)_{1\times 1} := \operatorname{Tr}\Big\{\int_0^1\int_0^1 \bar{\mathbf{S}}^m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\mathbf{D}_\epsilon\, m(\mathbf{y}^o_{1:N},s_\epsilon,s_x;\Psi)\,ds_\epsilon\,ds_x\Big\}.
\end{aligned}
$$

□

A.4.4.4. Proof of Corollary 26

Proof. Given the formulation of the gradient $\nabla_{\Psi^*}\tilde{Q}_{GStS,id,2}$ from Theorem 28 and the assumptions of Special Case 2 from Subsection 3.4.1.3, the M-step for Generalized Skew t-Student ID PPCA is specified by finding the the maximizers of the function $\tilde{Q}_{GStS,id,2}$ from Corollary 25 with respect to $\boldsymbol{\mu}^*$, σ^{*2}, δ_ϵ^* and δ_x^* as following

$$\begin{aligned}
&\sum_{t=1}^{N}\frac{\partial\tilde{v}(\mathbf{y}_t^o,s_\epsilon,s_x;\Psi,\Psi^*)}{\partial\boldsymbol{\mu}^*}=\frac{N}{\sigma^{*2}}\Big[T_\epsilon(s_\epsilon)\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)-T_\epsilon(s_\epsilon)\boldsymbol{\mu}^*-\delta_\epsilon^*-T_\epsilon(s_\epsilon)^2\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\\
&\qquad+T_\epsilon(s_\epsilon)^2\boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}+T_\epsilon(s_\epsilon)\big(\delta_\epsilon\mathbf{W}-\sigma^2\delta_x\big)\mathbf{M}^{-1}\mathbf{W}^{*T}\Big]\\
&\Longrightarrow\frac{\partial\tilde{Q}_{GStS,id,2}(\Psi,\Psi^*)}{\partial\boldsymbol{\mu}^*}=\mathbf{0}\\
&\Longrightarrow\mathbf{A}_6(\mathbf{y}_{1:N}^o;\Psi)-A_1(\mathbf{y}_{1:N}^o;\Psi)\boldsymbol{\mu}^*-A_0(\mathbf{y}_{1:N}^o;\Psi)\delta_\epsilon^*-\mathbf{A}_{10}(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}+\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{13}(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}+\big(\delta_\epsilon\mathbf{W}+\sigma^2\delta_x\big)\mathbf{A}_{11}(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}=\mathbf{0}
\end{aligned}$$

$$\begin{aligned}
&\sum_{t=1}^{N}\frac{\partial\tilde{v}(\mathbf{y}_t^o,s_\epsilon,s_x;\Psi,\Psi^*)}{\partial\mathbf{W}^*}=\frac{N}{\sigma^{*2}}\Big[T_\epsilon(s_\epsilon)^2\bar{\mathbf{S}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\mathbf{W}\mathbf{M}^{-1}-T_\epsilon(s_\epsilon)^2\boldsymbol{\mu}^{*T}\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\mathbf{W}\mathbf{M}^{-1}-T_\epsilon(s_\epsilon)\delta_\epsilon^{*T}\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\mathbf{W}\mathbf{M}^{-1}\\
&\quad-T_\epsilon(s_\epsilon)^2\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)^T\boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}-T_\epsilon(s_\epsilon)\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)^T\big(\delta_\epsilon\mathbf{W}-\sigma^2\delta_x\big)\mathbf{M}^{-1}\\
&\quad+T_\epsilon(s_\epsilon)^2\boldsymbol{\mu}^{*T}\boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}+T_\epsilon(s_\epsilon)\boldsymbol{\mu}^{*T}\big(\delta_\epsilon\mathbf{W}-\sigma^2\delta_x\big)\mathbf{M}^{-1}+T_\epsilon(s_\epsilon)\delta_\epsilon^{*T}\boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}+\delta_\epsilon^{*T}\big(\delta_\epsilon\mathbf{W}-\sigma^2\delta_x\big)\mathbf{M}^{-1}\\
&\quad-T_\epsilon(s_{\epsilon,t})^3\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{S}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\mathbf{W}\mathbf{M}^{-1}+T_\epsilon(s_\epsilon)^2\mathbf{W}^*\mathbf{M}^{-1}\mathbf{W}^T\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)^T\Big(T_\epsilon(s_\epsilon)\boldsymbol{\mu}\mathbf{W}+\delta_\epsilon\mathbf{W}-\sigma^2\delta_x\Big)\mathbf{M}^{-1}\\
&\quad+T_\epsilon(s_\epsilon)^2\mathbf{W}^*\mathbf{M}_t^{-1}\Big(T_\epsilon(s_\epsilon)\boldsymbol{\mu}\mathbf{W}+\delta_\epsilon\mathbf{W}-\sigma^2\delta_x\Big)^T\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\mathbf{W}^T\mathbf{M}^{-1}\\
&\quad-T_\epsilon(s_\epsilon)\mathbf{W}^*\mathbf{M}^{-1}\Big(T_\epsilon(s_\epsilon)\boldsymbol{\mu}\mathbf{W}+\delta_\epsilon\mathbf{W}-\sigma^2\delta_x\Big)^T\Big(T_\epsilon(s_\epsilon)\boldsymbol{\mu}\mathbf{W}+\delta_\epsilon\mathbf{W}-\sigma^2\delta_x\Big)\mathbf{M}^{-1}-\sigma^2T_\epsilon(s_\epsilon)\mathbf{W}^*\mathbf{M}^{-1}\Big]\\
&\Longrightarrow\frac{\partial\tilde{Q}_{GStS,id,2}(\Psi,\Psi^*)}{\partial\mathbf{W}^*}=\mathbf{0}\\
&\Longleftrightarrow\mathbf{A}_{14}(\mathbf{y}_{1:N}^o;\Psi)-\boldsymbol{\mu}^{*T}\mathbf{A}_{10}(\mathbf{y}_{1:N}^o;\Psi)-\delta_\epsilon^{*T}\mathbf{A}_8(\mathbf{y}_{1:N}^o;\Psi)-\mathbf{A}_{16}(\mathbf{y}_{1:N}^o;\Psi)+\boldsymbol{\mu}^{*T}\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{13}(\mathbf{y}_{1:N}^o;\Psi)+\boldsymbol{\mu}^{*T}\big(\delta_\epsilon-\sigma^2\delta_x\big)\mathbf{A}_{11}(\mathbf{y}_{1:N}^o;\Psi)\\
&\qquad+\delta_\epsilon^{*T}\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}_{1:N}^o;\Psi)+\delta_\epsilon^{*T}\big(\delta_\epsilon\mathbf{W}-\sigma^2\delta_x\big)\mathbf{A}_4(\mathbf{y}_{1:N}^o;\Psi)-\mathbf{W}^*\Big(\mathbf{A}_{17}(\mathbf{y}_{1:N}^o;\Psi)-\mathbf{A}_{18}(\mathbf{y}_{1:N}^o;\Psi)-\mathbf{A}_{18}(\mathbf{y}_{1:N}^o;\Psi)^T+\mathbf{A}_{19}(\mathbf{y}_{1:N}^o;\Psi)+\sigma^2\mathbf{A}_1
\end{aligned}$$

$$\begin{aligned}
&\sum_{t=1}^{N}\frac{\partial\tilde{v}(\mathbf{y}_t^o,s_\epsilon,s_x;\Psi,\Psi^*)}{\partial\delta_\epsilon^*}=\frac{N}{\sigma^{*2}}\Big[\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)-\boldsymbol{\mu}^*-T_\epsilon(s_\epsilon)^{-1}\delta_\epsilon^*-T_\epsilon(s_\epsilon)\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}+T_\epsilon(s_\epsilon)\boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\\
&\qquad+\big(\delta_\epsilon\mathbf{W}-\sigma^2\delta_x\big)\mathbf{M}^{-1}\mathbf{W}^{*T}\Big]\\
&\Longrightarrow\frac{\partial\tilde{Q}_{GStS,id,2}(\Psi,\Psi^*)}{\partial\delta_\epsilon^*}=\mathbf{0}\\
&\Longrightarrow\mathbf{A}_5(\mathbf{y}_{1:N}^o;\Psi)-A_0(\mathbf{y}_{1:N}^o;\Psi)\boldsymbol{\mu}^*-A_2(\mathbf{y}_{1:N}^o;\Psi)\delta_\epsilon^*-\mathbf{A}_8(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}+\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}+\big(\delta_\epsilon\mathbf{W}-\sigma^2\delta_x\big)\mathbf{A}_4(\mathbf{y}_{1:N}^o;\Psi)\mathbf{W}^{*T}=\mathbf{0}
\end{aligned}$$

$$\begin{aligned}
&\sum_{t=1}^{N}\frac{\partial\tilde{v}(\mathbf{y}_t^o,s_\epsilon,s_x;\Psi,\Psi^*)}{\partial\delta_x^*}=N\Big[-T_x(s_x)^{-1}\delta_x^*+T_\epsilon(s_\epsilon)\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\mathbf{W}\mathbf{M}^{-1}-T_\epsilon(s_\epsilon)\boldsymbol{\mu}\mathbf{W}\mathbf{M}^{-1}-\big(\delta_\epsilon\mathbf{W}-\sigma^2\delta_x\big)\mathbf{M}^{-1}\Big]\\
&\Longrightarrow\frac{\partial\tilde{Q}_{GStS,id,2}(\Psi,\Psi^*)}{\partial\delta_x^*}=\mathbf{0}\\
&\Longrightarrow-A_3(\mathbf{y}_{1:N}^o;\Psi)\delta_x^*+\mathbf{A}_8(\mathbf{y}_{1:N}^o;\Psi)-\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}_{1:N}^o;\Psi)-\big(\delta_\epsilon\mathbf{W}-\sigma^2\delta_x\big)\mathbf{A}_4(\mathbf{y}_{1:N}^o;\Psi)=\mathbf{0}
\end{aligned}$$

$$\begin{aligned}
&\sum_{t=1}^{N}\frac{\partial\tilde{v}(\mathbf{y}_t^o,s_\epsilon,s_x;\Psi,\Psi^*)}{\partial\sigma^{*2}}=-\frac{dN}{2\sigma^{*2}}+\frac{N}{2(\sigma^{*2})^2}\Big[T_\epsilon(s_\epsilon)\operatorname{Tr}\Big\{\bar{\mathbf{S}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\Big\}-2\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\Big(T_\epsilon(s_\epsilon)\boldsymbol{\mu}^{*T}+\delta_\epsilon^{*T}\Big)\\
&\qquad+T_\epsilon(s_\epsilon)\boldsymbol{\mu}^*\boldsymbol{\mu}^{*T}+2\boldsymbol{\mu}^*\delta_\epsilon^{*T}+T_\epsilon(s_\epsilon)^{-1}\delta_\epsilon^*\delta_\epsilon^{*T}-2T_\epsilon(s_\epsilon)^2\operatorname{Tr}\Big\{\bar{\mathbf{S}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\Big\}\\
&\qquad+2T_\epsilon(s_\epsilon)\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\Big(T_\epsilon(s_\epsilon)\boldsymbol{\mu}^{*T}+\delta_\epsilon^{*T}\Big)+2T_\epsilon(s_\epsilon)\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\mathbf{W}^*\mathbf{M}^{-1}\Big(\boldsymbol{\mu}T_\epsilon(s_\epsilon)\mathbf{W}+\delta_\epsilon\mathbf{W}-\sigma^2\delta_x\Big)^T\\
&+2\Big(-T_\epsilon(s_\epsilon)\boldsymbol{\mu}\mathbf{W}-\delta_\epsilon\mathbf{W}+\sigma^2\delta_x\Big)\mathbf{M}^{-1}\mathbf{W}^{*T}\Big(T_\epsilon(s_\epsilon)\boldsymbol{\mu}^{*T}+\delta_\epsilon^{*T}\Big)+T_\epsilon(s_\epsilon)^3\operatorname{Tr}\Big\{\mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{S}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{W}^*\Big\}\\
&-T_\epsilon(s_\epsilon)^2\operatorname{Tr}\Big\{\mathbf{M}^{-1}\mathbf{W}^T\bar{\boldsymbol{\mu}}^m(\mathbf{y}_{1:N}^o,s_\epsilon,s_x;\Psi)^T\Big(T_\epsilon(s_\epsilon)\boldsymbol{\mu}\mathbf{W}+\delta_\epsilon\mathbf{W}-\sigma^2\delta_x\Big)\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{W}^*\Big\}
\end{aligned}$$

$$
\begin{aligned}
&- T_\epsilon(s_\epsilon)^2 \operatorname{Tr}\left\{\mathbf{M}^{-1}\left(T_\epsilon(s_\epsilon)\boldsymbol{\mu}\mathbf{W} + \boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\right)^T \bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N}, s_\epsilon, s_x; \Psi)\mathbf{W}^T\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{W}^*\right\} \\
&+ T_\epsilon(s_\epsilon)\operatorname{Tr}\left\{\mathbf{M}^{-1}\left(T_\epsilon(s_\epsilon)\boldsymbol{\mu}\mathbf{W} + \boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\right)^T\left(T_\epsilon(s_\epsilon)\boldsymbol{\mu}\mathbf{W} + \boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\right)\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{W}^*\right\} + \sigma^2 T_\epsilon(s_\epsilon)\operatorname{Tr}\left\{\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{W}^*\right\}\Bigg] \\
\implies &\frac{\partial \tilde{Q}_{GStS,id,2}(\Psi, \Psi^*)}{\partial \sigma^{*2}} = \mathbf{0} \\
\iff &-\frac{d}{2\sigma^{*2}} A_0(\mathbf{y}^o_{1:N}; \Psi) + \frac{1}{2(\sigma^{*2})^2}\Bigg[A_{20}(\mathbf{y}^o_{1:N}; \Psi) - 2\mathbf{A}_6(\mathbf{y}^o_{1:N}; \Psi)\boldsymbol{\mu}^{*T} - 2\mathbf{A}_5(\mathbf{y}^o_{1:N}; \Psi)\boldsymbol{\delta}_\epsilon^{*T} \\
&+ A_1(\mathbf{y}^o_{1:N}; \Psi)\boldsymbol{\mu}^*\boldsymbol{\mu}^{*T} + 2A_0(\mathbf{y}^o_{1:N}; \Psi)\boldsymbol{\mu}^*\boldsymbol{\delta}_\epsilon^{*T} + A_2(\mathbf{y}^o_{1:N}; \Psi)\boldsymbol{\delta}_\epsilon^*\boldsymbol{\delta}_\epsilon^{*T} - 2\operatorname{Tr}\left\{\mathbf{A}_{14}(\mathbf{y}^o_{1:N}; \Psi)\mathbf{W}^{*T}\right\} + 2\mathbf{A}_{10}(\mathbf{y}^o_{1:N}; \Psi)\mathbf{W}^{*T}\boldsymbol{\mu}^{*T} \\
&+ 2\mathbf{A}_8(\mathbf{y}^o_{1:N}; \Psi)\mathbf{W}^{*T}\boldsymbol{\delta}_\epsilon^{*T} + 2\mathbf{A}_9(\mathbf{y}^o_{1:N}; \Psi, \Psi^*)\mathbf{W}^T\boldsymbol{\mu}^T + 2\mathbf{A}_7(\mathbf{y}^o_{1:N}; \Psi, \Psi)\left(\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\right)^T \\
&- 2\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{13}(\mathbf{y}^o_{1:N}; \Psi)\mathbf{W}^{*T}\boldsymbol{\mu}^{*T} - 2\left(\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\right)\mathbf{A}_{11}(\mathbf{y}^o_{1:N}; \Psi)\mathbf{W}^{*T}\boldsymbol{\mu}^{*T} \\
&- 2\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}^o_{1:N}; \Psi)\mathbf{W}^{*T}\boldsymbol{\delta}_\epsilon^{*T} - 2\left(\boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\right)\mathbf{A}_4(\mathbf{y}^o_{1:N}; \Psi)\mathbf{W}^{*T}\boldsymbol{\delta}_\epsilon^{*T} + \operatorname{Tr}\left\{\mathbf{A}_{17}(\mathbf{y}^o_{1:N}; \Psi)\mathbf{W}^{*T}\mathbf{W}^*\right\} \\
&- \operatorname{Tr}\left\{\left(\mathbf{A}_{18}(\mathbf{y}^o_{1:N}; \Psi) + \mathbf{A}_{18}(\mathbf{y}^o_{1:N}; \Psi)^T\right)\mathbf{W}^{*T}\mathbf{W}^*\right\} + \operatorname{Tr}\left\{\mathbf{A}_{19}(\mathbf{y}^o_{1:N}; \Psi)\mathbf{W}^{*T}\mathbf{W}^*\right\} + \sigma^2\operatorname{Tr}\left\{\mathbf{A}_{11}(\mathbf{y}^o_{1:N}; \Psi)\mathbf{W}^{*T}\mathbf{W}^*\right\}\Bigg] = 0
\end{aligned}
$$

for

$$
\begin{aligned}
A_0(\mathbf{y}^o_{1:N}; \Psi)_{1\times 1} &:= \int_0^1\int_0^1 m(\mathbf{y}^o_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x \\
A_1(\mathbf{y}^o_{1:N}; \Psi)_{1\times 1} &:= \int_0^1\int_0^1 T_\epsilon(s_\epsilon)\, m(\mathbf{y}^o_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x \\
A_2(\mathbf{y}^o_{1:N}; \Psi)_{1\times 1} &:= \int_0^1\int_0^1 T_\epsilon(s_\epsilon)^{-1}\, m(\mathbf{y}^o_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x \\
A_3(\mathbf{y}^o_{1:N}; \Psi)_{1\times 1} &:= \int_0^1\int_0^1 T_x(s_x)^{-1}\, m(\mathbf{y}^o_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x \\
\mathbf{A}_4(\mathbf{y}^o_{1:N}; \Psi)_{k\times k} &:= \int_0^1\int_0^1 \mathbf{M}^{-1}\, m(\mathbf{y}^o_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x \\
\mathbf{A}_5(\mathbf{y}^o_{1:N}; \Psi)_{1\times d} &:= \int_0^1\int_0^1 \bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N}, s_\epsilon, s_x; \Psi)\, m(\mathbf{y}^o_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x \\
\mathbf{A}_6(\mathbf{y}^o_{1:N}; \Psi)_{1\times d} &:= \int_0^1\int_0^1 T_\epsilon(s_\epsilon)\bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N}, s_\epsilon, s_x; \Psi)\, m(\mathbf{y}^o_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x \\
\mathbf{A}_7(\mathbf{y}^o_{1:N}; \Psi, \Psi^*)_{1\times k} &:= \int_0^1\int_0^1 T_\epsilon(s_\epsilon)\bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N}, s_\epsilon, s_x; \Psi)\mathbf{W}^*\mathbf{M}^{-1}\, m(\mathbf{y}^o_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x \\
\mathbf{A}_8(\mathbf{y}^o_{1:N}; \Psi)_{1\times k} &:= \int_0^1\int_0^1 T_\epsilon(s_\epsilon)\bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N}, s_\epsilon, s_x; \Psi)\mathbf{W}\mathbf{M}^{-1}\, m(\mathbf{y}^o_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x \\
\mathbf{A}_9(\mathbf{y}^o_{1:N}; \Psi, \Psi^*)_{1\times d} &:= \int_0^1\int_0^1 T_\epsilon(s_\epsilon)^2\bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N}, s_\epsilon, s_x; \Psi)\mathbf{W}^*\mathbf{M}^{-1}\, m(\mathbf{y}^o_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x \\
\mathbf{A}_{10}(\mathbf{y}^o_{1:N}; \Psi)_{1\times d} &:= \int_0^1\int_0^1 T_\epsilon(s_{\epsilon,t})^2\bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N}, s_\epsilon, s_x; \Psi)\mathbf{W}\mathbf{M}^{-1}\, m(\mathbf{y}^o_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x \\
\mathbf{A}_{11}(\mathbf{y}^o_{1:N}; \Psi)_{k\times k} &:= \int_0^1\int_0^1 T_\epsilon(s_\epsilon)\mathbf{M}^{-1}\, m(\mathbf{y}^o_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x \\
\mathbf{A}_{13}(\mathbf{y}^o_{1:N}; \Psi)_{k\times k} &:= \int_0^1\int_0^1 T_\epsilon(s_\epsilon)^2\mathbf{M}^{-1}\, m(\mathbf{y}^o_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x \\
\mathbf{A}_{14}(\mathbf{y}^o_{1:N}; \Psi)_{d\times k} &:= \int_0^1\int_0^1 T_\epsilon(s_\epsilon)^2\bar{\mathbf{S}}^m(\mathbf{y}^o_{1:N}, s_\epsilon, s_x; \Psi)\mathbf{W}\mathbf{M}^{-1}\, m(\mathbf{y}^o_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x \\
\mathbf{A}_{16}(\mathbf{y}^o_{1:N}; \Psi, \Psi^*)_{d\times k} &:= \int_0^1\int_0^1 T_\epsilon(s_\epsilon)\bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N}, s_\epsilon, s_x; \Psi)^T\left(T_\epsilon(s_\epsilon)\boldsymbol{\mu}\mathbf{W} + \boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\right)\mathbf{M}^{-1} \\
&\quad\times m(\mathbf{y}^o_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x \\
\mathbf{A}_{17}(\mathbf{y}^o_{1:N}; \Psi)_{k\times k} &:= \int_0^1\int_0^1 T_\epsilon(s_\epsilon)^3\mathbf{M}^{-1}\mathbf{W}^T\bar{\mathbf{S}}^m(\mathbf{y}^o_{1:N}, s_\epsilon, s_x; \Psi)\mathbf{W}\mathbf{M}^{-1}\, m(\mathbf{y}^o_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x \\
\mathbf{A}_{18}(\mathbf{y}^o_{1:N}; \Psi)_{k\times k} &:= \int_0^1\int_0^1 T_\epsilon(s_\epsilon)^2\mathbf{M}^{-1}\mathbf{W}^T\bar{\boldsymbol{\mu}}^m(\mathbf{y}^o_{1:N}, s_\epsilon, s_x; \Psi)^T\left(T_\epsilon(s_\epsilon)\boldsymbol{\mu}\mathbf{W} + \boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\right)\mathbf{M}^{-1} \\
&\quad\times m(\mathbf{y}^o_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x \\
\mathbf{A}_{19}(\mathbf{y}^o_{1:N}; \Psi)_{k\times k} &:= \int_0^1\int_0^1 T_\epsilon(s_\epsilon)\mathbf{M}^{-1}\left(T_\epsilon(s_\epsilon)\boldsymbol{\mu}\mathbf{W} + \boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\right)^T\left(T_\epsilon(s_\epsilon)\boldsymbol{\mu}\mathbf{W} + \boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\right)\mathbf{M}^{-1}\, m(\mathbf{y}^o_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x \\
A_{20}(\mathbf{y}^o_{1:N}; \Psi)_{1\times 1} &:= \int_0^1\int_0^1 T_\epsilon(s_\epsilon)\operatorname{Tr}\left\{\bar{\mathbf{S}}^m(\mathbf{y}^o_{1:N}, s_\epsilon, s_x; \Psi)\right\}\, m(\mathbf{y}^o_{1:N}, s_\epsilon, s_x; \Psi)\, ds_\epsilon\, ds_x.
\end{aligned}
$$

Hence, the system of equations $\nabla_{\Psi^*}\tilde{Q}_{GStS,id,2}(\Psi,\Psi^*)=\mathbf{0}$ is the following

$$
\begin{cases}
\boldsymbol{\mu}^* = A_1(\mathbf{y}^o_{1:N};\Psi)^{-1}\Big[\mathbf{A}_6(\mathbf{y}^o_{1:N};\Psi) - A_0(\mathbf{y}^o_{1:N};\Psi)\boldsymbol{\delta}^*_\epsilon - \mathbf{A}_{10}(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}^{*T} + \boldsymbol{\mu}\mathbf{W}\mathbf{A}_{13}(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}^{*T} + \big(\boldsymbol{\delta}_\epsilon\mathbf{W}+\sigma^2\boldsymbol{\delta}_x\big)\mathbf{A}_{11}(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}^{*T}\Big] \\
\boldsymbol{\delta}^*_\epsilon = A_2(\mathbf{y}^o_{1:N};\Psi)^{-1}\Big[\mathbf{A}_5(\mathbf{y}_{1:N};\Psi) - A_0(\mathbf{y}_{1:N};\Psi)\boldsymbol{\mu}^* - \mathbf{A}_8(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}^{*T} + \boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}^{*T} + \Big(\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\Big)\mathbf{A}_4(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}^{*T}\Big] \\
\boldsymbol{\delta}^*_x = A_3(\mathbf{y}^o_{1:N};\Psi)^{-1}\Big(\mathbf{A}_8(\mathbf{y}_{1:N};\Psi) - \boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}^o_{1:N};\Psi) - \big(\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\big)\mathbf{A}_4(\mathbf{y}^o_{1:N};\Psi)\Big) \\
\mathbf{W}^* = \Big[\mathbf{A}_{14}(\mathbf{y}^o_{1:N};\Psi) - \boldsymbol{\mu}^{*T}\mathbf{A}_{10}(\mathbf{y}^o_{1:N};\Psi) - \boldsymbol{\delta}^*_\epsilon\mathbf{A}_8(\mathbf{y}^o_{1:N};\Psi) - \mathbf{A}_{16}(\mathbf{y}^o_{1:N};\Psi) + \boldsymbol{\mu}^{*T}\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{13}(\mathbf{y}^o_{1:N};\Psi) + \boldsymbol{\mu}^{*T}\big(\boldsymbol{\delta}_\epsilon-\sigma^2\boldsymbol{\delta}_x\big)\mathbf{A}_{11}(\mathbf{y}^o_{1:N};\Psi) \\
\quad +\boldsymbol{\delta}^{*T}_\epsilon\boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}^o_{1:N};\Psi) + \boldsymbol{\delta}^{*T}_\epsilon\big(\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\big)\mathbf{A}_4(\mathbf{y}^o_{1:N};\Psi)\Big]\Big[\mathbf{A}_{17}(\mathbf{y}^o_{1:N};\Psi) - \mathbf{A}_{18}(\mathbf{y}^o_{1:N};\Psi) - \mathbf{A}_{18}(\mathbf{y}^o_{1:N};\Psi)^T + \mathbf{A}_{19}(\mathbf{y}^o_{1:N};\Psi) \\
\quad +\sigma^2\mathbf{A}_{11}(\mathbf{y}^o_{1:N};\Psi)\Big]^{-1} \\
\sigma^{*2} = \frac{1}{dA_0(\mathbf{y}^o_{1:N};\Psi)}\Big[A_{20}(\mathbf{y}^o_{1:N};\Psi) - 2\mathbf{A}_6(\mathbf{y}^o_{1:N};\Psi)\boldsymbol{\mu}^{*T} - 2\mathbf{A}_5(\mathbf{y}^o_{1:N};\Psi)\boldsymbol{\delta}^{*T}_\epsilon + A_1(\mathbf{y}^o_{1:N};\Psi)\boldsymbol{\mu}^*\boldsymbol{\mu}^{*T} + 2A_0(\mathbf{y}^o_{1:N};\Psi)\boldsymbol{\mu}^*\boldsymbol{\delta}^{*T}_\epsilon \\
\quad +A_2(\mathbf{y}^o_{1:N};\Psi)\boldsymbol{\delta}^*_\epsilon\boldsymbol{\delta}^{*T}_\epsilon - 2\operatorname{Tr}\Big\{\mathbf{A}_{14}(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}^{*T}\Big\} + 2\mathbf{A}_9(\mathbf{y}^o_{1:N};\Psi,\Psi^*)\mathbf{W}^T\boldsymbol{\mu}^T \\
\quad +2\mathbf{A}_7(\mathbf{y}^o_{1:N};\Psi,\Psi)\big(\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\big)^T + 2\Big(\mathbf{A}_{10}(\mathbf{y}^o_{1:N};\Psi) - \boldsymbol{\mu}\mathbf{W}\mathbf{A}_{13}(\mathbf{y}^o_{1:N};\Psi) - \big(\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\big)\mathbf{A}_{11}(\mathbf{y}^o_{1:N};\Psi)\Big)\mathbf{W}^{*T}\boldsymbol{\mu}^{*T} \\
\quad +2\Big(\mathbf{A}_8(\mathbf{y}^o_{1:N};\Psi) - \boldsymbol{\mu}\mathbf{W}\mathbf{A}_{11}(\mathbf{y}^o_{1:N};\Psi) - \big(\boldsymbol{\delta}_\epsilon\mathbf{W}-\sigma^2\boldsymbol{\delta}_x\big)\mathbf{A}_4(\mathbf{y}^o_{1:N};\Psi)\Big)\mathbf{W}^{*T}\boldsymbol{\delta}^{*T}_\epsilon + \operatorname{Tr}\Big\{\mathbf{A}_{17}(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}^{*T}\mathbf{W}^*\Big\} \\
\quad -\operatorname{Tr}\Big\{\Big(\mathbf{A}_{18}(\mathbf{y}^o_{1:N};\Psi) + \mathbf{A}_{18}(\mathbf{y}^o_{1:N};\Psi)^T\Big)\mathbf{W}^{*T}\mathbf{W}^*\Big\} + \operatorname{Tr}\Big\{\mathbf{A}_{19}(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}^{*T}\mathbf{W}^*\Big\} + \sigma^2\operatorname{Tr}\Big\{\mathbf{A}_{11}(\mathbf{y}^o_{1:N};\Psi)\mathbf{W}^{*T}\mathbf{W}^*\Big\}\Big].
\end{cases}
$$

□

A.4.5. Supporting Lemmas for Generelized Skew t-Student IND PPCA

Lemma A8. *Let a d dimensional observation vector $\mathbf{Y}_t$ be modelled as in Equation (2.7) under the Generelized t-Student IND PPCA model defined in Section 3.4 with the latent processes $\mathbf{X}_t$, $\boldsymbol{\epsilon}_t$, $\mathbf{U}_t$ and $\mathbf{V}_t$ following the assumptions defined in Equation (3.27) and Equation (3.28). We can decompose the join probability function of the variables $\mathbf{Y}_t, \mathbf{X}_t, \mathbf{U}_t$ and $\mathbf{V}_t$ into the product of three functions*

$$\pi_{\mathbf{Y}_t,\mathbf{X}_t,\mathbf{U}_t,\mathbf{V}_t|\Psi}(\mathbf{y}_t,\mathbf{x}_t,\mathbf{u}_t,\mathbf{v}_t) = \pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{x}_t)\cdot\pi_{\mathbf{Y}_t|\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{y}_t)\cdot C(\mathbf{u}_t,\mathbf{v}_t;\Psi)$$

where $\pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{x}_t)$ and $\pi_{\mathbf{Y}_t|\mathbf{U}_t\mathbf{V}_t,\Psi}(\mathbf{x}_t)$ are the conditional probability function of the k dimensional random vector $\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U}_t,\mathbf{V}_t \sim \mathcal{N}(\boldsymbol{\mu}_{x,t},\boldsymbol{\Sigma}_{x,t})$ and d dimensional random vector $\mathbf{Y}_t|\mathbf{U}_t,\mathbf{V}_t \sim \mathcal{N}(\boldsymbol{\mu}_{y,t},\boldsymbol{\Sigma}_{y,t})$, respectively, for the first and second central moments of the distributions given by

$$\begin{aligned}
\boldsymbol{\mu}_{x,t} &= \Big(\big(\mathbf{y}_t-\boldsymbol{\mu}-\boldsymbol{\delta}_\epsilon\mathbf{D}^{-1}_{\epsilon,t}\big)\mathbf{D}_{\epsilon,t}\mathbf{W}+\sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}_t^{-1} \text{ and } \boldsymbol{\Sigma}_{x,t} = \sigma^2\mathbf{M}_t^{-1}, \\
\boldsymbol{\mu}_{y,t} &= \boldsymbol{\mu}+\boldsymbol{\delta}_\epsilon\mathbf{D}^{-1}_{\epsilon,t}+\sigma^2\boldsymbol{\delta}_x\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbf{N}_t^{-1} \text{ and } \boldsymbol{\Sigma}_{y,t} = \sigma^2\mathbf{N}_t^{-1}
\end{aligned}$$

for $\mathbf{M}_t = \sigma^2\mathbf{D}_{x,t}+\mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbf{W}$ and $\mathbf{N}_t = \mathbf{D}_{\epsilon,t}-\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}$. The function $C:\mathbb{R}^d_+\times\mathbb{R}^k_+\longrightarrow\mathbb{R}$ is given by

$$\begin{aligned}
C(\mathbf{u}_t,\mathbf{v}_t;\Psi) = \pi_{\mathbf{U}_t|\Psi}(\mathbf{u}_t)\cdot\pi_{\mathbf{V}_t|\Psi}(\mathbf{v}_t)\big(\sigma^2\big)^{\frac{k}{2}}\Big|\mathbf{D}_{\epsilon,t}\Big|^{\frac{1}{2}}\Big|\mathbf{D}_{x,t}\Big|^{\frac{1}{2}}\Big|\mathbf{M}_t\Big|^{-\frac{1}{2}}\Big|\mathbf{N}_t\Big|^{-\frac{1}{2}} \\
\times\exp\Big\{-\frac{1}{2}\boldsymbol{\delta}_x\Big(\mathbf{D}^{-1}_{x,t}-\sigma^2\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbf{N}_t^{-1}\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}-\sigma^2\mathbf{M}_t^{-1}\Big)\boldsymbol{\delta}_x^T\Big\}.
\end{aligned}$$

Proof. Using the Chain Rule of probabilities we obtain the following decomposition of the likelihood function

$$\pi_{\mathbf{Y}_t,\mathbf{X}_t,\mathbf{U}_t,\mathbf{V}_t|\Psi}(\mathbf{y}_t,\mathbf{x}_t,\mathbf{u}_t,\mathbf{v}_t) = \pi_{\mathbf{Y}_t|\mathbf{X}_t,\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{y}_t)\cdot\pi_{\mathbf{X}_t|\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{x}_t)\cdot\pi_{\mathbf{U}_t|\Psi}(\mathbf{u}_t)\cdot\pi_{\mathbf{V}_t|\Psi}(\mathbf{v}_t)$$

$$= \pi_{\mathbf{U}_t|\Psi}(\mathbf{u}_t)\pi_{\mathbf{V}_t|\Psi} \cdot (\mathbf{v}_t) \cdot (2\pi)^{-\frac{d+k}{2}} \left|\sigma^2\mathbf{D}_{\epsilon,t}^{-1}\right|^{-\frac{1}{2}} \left|\mathbf{D}_{x,t}^{-1}\right|^{-\frac{1}{2}} \tag{A19}$$

$$\times \exp\left\{ -\frac{1}{2}\left((\mathbf{x}_t - \boldsymbol{\delta}_x\mathbf{D}_{x,t}^{-1})\mathbf{D}_{x,t}(\mathbf{x}_t - \boldsymbol{\delta}_x\mathbf{D}_{x,t}^{-1})^T + \frac{1}{\sigma^2}(\mathbf{y}_t - \boldsymbol{\mu} - \boldsymbol{\delta}_\epsilon\mathbf{D}_{\epsilon,t}^{-1} - \mathbf{x}_t\mathbf{W}^T)\mathbf{D}_{\epsilon,t}(\mathbf{y}_t - \boldsymbol{\mu} - \boldsymbol{\delta}_\epsilon\mathbf{D}_{\epsilon,t}^{-1} - \mathbf{x}_t\mathbf{W}^T)^T \right) \right\}$$

$$= \pi_{\mathbf{U}_t|\Psi}(\mathbf{u}_t) \cdot \pi_{\mathbf{V}_t|\Psi}(\mathbf{v}_t) \cdot (2\pi)^{-\frac{d+k}{2}} (\sigma^2)^{-\frac{d}{2}} \left|\mathbf{D}_{\epsilon,t}\right|^{\frac{1}{2}} \left|\mathbf{D}_{x,t}\right|^{\frac{1}{2}} \exp\left\{ -\frac{1}{2}\left(\mathbf{x}_t\mathbf{D}_{x,t}\mathbf{x}_t^T - 2\mathbf{x}_t\boldsymbol{\delta}_x^T + \boldsymbol{\delta}_x\mathbf{D}_{x,t}^{-1}\boldsymbol{\delta}_x^T \right.\right.$$

$$\left.\left. + \frac{1}{\sigma^2}(\mathbf{y}_t - \boldsymbol{\mu} - \boldsymbol{\delta}_\epsilon\mathbf{D}_{\epsilon,t}^{-1})\mathbf{D}_{\epsilon,t}(\mathbf{y}_t - \boldsymbol{\mu} - \boldsymbol{\delta}_\epsilon\mathbf{D}_{\epsilon,t}^{-1})^T - \frac{2}{\sigma^2}\mathbf{x}_t\mathbf{W}^T\mathbf{D}_{\epsilon,t}(\mathbf{y}_t - \boldsymbol{\mu} - \boldsymbol{\delta}_\epsilon\mathbf{D}_{\epsilon,t}^{-1})^T + \frac{1}{\sigma^2}\mathbf{x}_t\mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{x}_t^T \right)\right\}$$

$$= \pi_{\mathbf{U}_t|\Psi}(\mathbf{u}_t) \cdot \pi_{\mathbf{V}_t|\Psi}(\mathbf{v}_t)\ (2\pi)^{-\frac{d+k}{2}} (\sigma^2)^{-\frac{d}{2}} \left|\mathbf{D}_{\epsilon,t}\right|^{\frac{1}{2}} \left|\mathbf{D}_{x,t}\right|^{\frac{1}{2}} \exp\left\{ -\frac{1}{2}(\mathbf{x}_t - \boldsymbol{\mu}_{x,t})\boldsymbol{\Sigma}_{x,t}^{-1}(\mathbf{x}_t - \boldsymbol{\mu}_{x,t})^T \right\}$$

$$\times \exp\left\{ -\frac{1}{2}\left(\boldsymbol{\delta}_x\mathbf{D}_{x,t}^{-1}\boldsymbol{\delta}_x^T + \frac{1}{\sigma^2}(\mathbf{y}_t - \boldsymbol{\mu} - \boldsymbol{\delta}_\epsilon\mathbf{D}_{\epsilon,t}^{-1})\mathbf{D}_{\epsilon,t}(\mathbf{y}_t - \boldsymbol{\mu} - \boldsymbol{\delta}_\epsilon\mathbf{D}_{\epsilon,t}^{-1})^T - \boldsymbol{\mu}_{x,t}\boldsymbol{\Sigma}_{x,t}^{-1}\boldsymbol{\mu}_{x,t}^T \right) \right\} \tag{A20}$$

where $\boldsymbol{\Sigma}_{x,t} = \sigma^2\mathbf{M}_t^{-1}$, $\boldsymbol{\mu}_{x,t} = \left((\mathbf{y}_t - \boldsymbol{\mu} - \boldsymbol{\delta}_\epsilon\mathbf{D}_{\epsilon,t}^{-1})\mathbf{D}_{\epsilon,t}\mathbf{W} + \sigma^2\boldsymbol{\delta}_x\right)\mathbf{M}_t^{-1}$ and $\mathbf{M}_t = \sigma^2\mathbf{D}_{x,t} + \mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbf{W}$. Let us denote

$$\pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{x}_t) = (2\pi)^{-\frac{k}{2}} \left|\boldsymbol{\Sigma}_{x,t}\right|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(\mathbf{x}_t - \boldsymbol{\mu}_{x,t})\boldsymbol{\Sigma}_{x,t}^{-1}(\mathbf{x}_t - \boldsymbol{\mu}_{x,t})^T \right\}$$

and remark that it is a probability function of k dimensional Gaussian random variable with the mean vector $\boldsymbol{\mu}_{x,t}$ and the covariance matrix $\boldsymbol{\Sigma}_{x,t}$. The probability function $\pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{x}_t)$ contains all expressions with the vector $\mathbf{x}_t$ present in the probability function $\pi_{\mathbf{Y}_t,\mathbf{X}_t,\mathbf{U}_t,\mathbf{V}_t|\Psi}(\mathbf{y}_t,\mathbf{x}_t,\mathbf{u}_t,\mathbf{v}_t)$. In the next step we show how to obtain the formula for the probability function $\pi_{\mathbf{Y}_t|\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{x}_t)$, which jointly with the function $\pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{x}_t)$, contains all expressions with the vector $\mathbf{y}_t$. Resuming the simplification of the Equation (A19), we obtain the following formulations

$$\pi_{\mathbf{Y}_t,\mathbf{X}_t,\mathbf{U}_t,\mathbf{V}_t|\Psi}(\mathbf{y}_t,\mathbf{x}_t,\mathbf{u}_t,\mathbf{v}_t) = \pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{x}_t)\ \pi_{\mathbf{U}_t|\Psi}(\mathbf{u}_t) \cdot \pi_{\mathbf{V}_t|\Psi}(\mathbf{v}_t)\ (2\pi)^{-\frac{d}{2}} (\sigma^2)^{-\frac{d}{2}} \left|\mathbf{D}_{\epsilon,t}\right|^{\frac{1}{2}} \left|\mathbf{D}_{x,t}\right|^{\frac{1}{2}} \left|\boldsymbol{\Sigma}_{x,t}\right|^{\frac{1}{2}}$$

$$\times \exp\left\{ -\frac{1}{2}\left(\boldsymbol{\delta}_x\mathbf{D}_{x,t}^{-1}\boldsymbol{\delta}_x^T + \frac{1}{\sigma^2}(\mathbf{y}_t - \boldsymbol{\mu} - \boldsymbol{\delta}_\epsilon\mathbf{D}_{\epsilon,t}^{-1})\mathbf{D}_{\epsilon,t}(\mathbf{y}_t - \boldsymbol{\mu} - \boldsymbol{\delta}_\epsilon\mathbf{D}_{\epsilon,t}^{-1})^T - \boldsymbol{\mu}_{x,t}\boldsymbol{\Sigma}_{x,t}^{-1}\boldsymbol{\mu}_{x,t}^T \right) \right\}$$

$$= \pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{x}_t)\ \pi_{\mathbf{U}_t|\Psi}(\mathbf{u}_t) \cdot \pi_{\mathbf{V}_t|\Psi}(\mathbf{v}_t)\ (2\pi)^{-\frac{d}{2}} (\sigma^2)^{-\frac{d-k}{2}} \left|\mathbf{D}_{\epsilon,t}\right|^{\frac{1}{2}} \left|\mathbf{D}_{x,t}\right|^{\frac{1}{2}} \left|\mathbf{M}_t\right|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}\boldsymbol{\delta}_x\mathbf{D}_{x,t}^{-1}\boldsymbol{\delta}_x^T \right\}$$

$$\times \exp\left\{ -\frac{1}{2\sigma^2}(\mathbf{y}_t - \boldsymbol{\mu} - \boldsymbol{\delta}_\epsilon\mathbf{D}_{\epsilon,t}^{-1})\mathbf{D}_{\epsilon,t}(\mathbf{y}_t - \boldsymbol{\mu} - \boldsymbol{\delta}_\epsilon\mathbf{D}_{\epsilon,t}^{-1})^T \right\}$$

$$\times \exp\left\{ \frac{1}{2\sigma^2}\left((\mathbf{y}_t - \boldsymbol{\mu} - \boldsymbol{\delta}_\epsilon\mathbf{D}_{\epsilon,t}^{-1})\mathbf{D}_{\epsilon,t}\mathbf{W} + \sigma^2\boldsymbol{\delta}_x\right)\mathbf{M}_t^{-1}\left((\mathbf{y}_t - \boldsymbol{\mu} - \boldsymbol{\delta}_\epsilon\mathbf{D}_{\epsilon,t}^{-1})\mathbf{D}_{\epsilon,t}\mathbf{W} + \sigma^2\boldsymbol{\delta}_x\right)^T \right\}$$

$$= \pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{x}_t)\ \pi_{\mathbf{U}_t|\Psi}(\mathbf{u}_t)\ \pi_{\mathbf{V}_t|\Psi}(\mathbf{v}_t)\ (2\pi)^{-\frac{d}{2}} (\sigma^2)^{-\frac{d-k}{2}} \left|\mathbf{D}_{\epsilon,t}\right|^{\frac{1}{2}} \left|\mathbf{D}_{x,t}\right|^{\frac{1}{2}} \left|\mathbf{M}_t\right|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}\boldsymbol{\delta}_x\mathbf{D}_{x,t}^{-1}\boldsymbol{\delta}_x^T \right\}$$

$$\times \exp\left\{ -\frac{1}{2}(\mathbf{y}_t - \boldsymbol{\mu} - \boldsymbol{\delta}_\epsilon\mathbf{D}_{\epsilon,t}^{-1}) \underbrace{\left(\frac{1}{\sigma^2}\mathbf{D}_{\epsilon,t} - \frac{1}{\sigma^2}\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\right)}_{\boldsymbol{\Sigma}_{y,t}^{-1}} (\mathbf{y}_t - \boldsymbol{\mu} - \boldsymbol{\delta}_\epsilon\mathbf{D}_{\epsilon,t}^{-1})^T \right\}$$

$$\times \exp\left\{ -\frac{1}{2}\left(-2(\mathbf{y}_t - \boldsymbol{\mu} - \boldsymbol{\delta}_\epsilon\mathbf{D}_{\epsilon,t}^{-1})\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\boldsymbol{\delta}_x^T - \sigma^2\boldsymbol{\delta}_x\mathbf{M}_t^{-1}\boldsymbol{\delta}_x^T \right) \right\}$$

$$= \pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{x}_t)\ \pi_{\mathbf{U}_t|\Psi}(\mathbf{u}_t)\ \pi_{\mathbf{V}_t|\Psi}(\mathbf{v}_t)\ (2\pi)^{-\frac{d}{2}} (\sigma^2)^{-\frac{d-k}{2}} \left|\mathbf{D}_{\epsilon,t}\right|^{\frac{1}{2}} \left|\mathbf{D}_{x,t}\right|^{\frac{1}{2}} \left|\mathbf{M}_t\right|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}\boldsymbol{\delta}_x\mathbf{D}_{x,t}^{-1}\boldsymbol{\delta}_x^T \right\}$$

$$\times \exp\left\{ -\frac{1}{2}(\underbrace{\mathbf{y}_t - \boldsymbol{\mu} - \boldsymbol{\delta}_\epsilon\mathbf{D}_{\epsilon,t}^{-1} - \sigma^2\boldsymbol{\delta}_x\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbf{N}_t^{-1}}_{\mathbf{y}_t - \boldsymbol{\mu}_{y,t}})\boldsymbol{\Sigma}_{y,t}^{-1}(\underbrace{\mathbf{y}_t - \boldsymbol{\mu} - \boldsymbol{\delta}_\epsilon\mathbf{D}_{\epsilon,t}^{-1} - \sigma^2\boldsymbol{\delta}_x\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbf{N}_t^{-1}}_{\mathbf{y}_t - \boldsymbol{\mu}_{y,t}})^T \right\}$$

$$\times \exp\left\{ -\frac{1}{2}\left(-\sigma^2\boldsymbol{\delta}_x\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbf{N}_t^{-1}\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\boldsymbol{\delta}_x^T - \sigma^2\boldsymbol{\delta}_x\mathbf{M}_t^{-1}\boldsymbol{\delta}_x^T \right) \right\}$$

where $\boldsymbol{\Sigma}_{y,t} = \sigma^2\mathbf{N}_t^{-1}$ for $\mathbf{N}_t = \mathbf{D}_{\epsilon,t} - \mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}$. Let us denote $\boldsymbol{\mu}_{y,t} = \boldsymbol{\mu} + \boldsymbol{\delta}_\epsilon\mathbf{D}_{\epsilon,t}^{-1} + \sigma^2\boldsymbol{\delta}_x\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbf{N}_t^{-1}$ and define

$$\pi_{\mathbf{Y}_t|\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{y}_t) = (2\pi)^{-\frac{d}{2}} \left|\boldsymbol{\Sigma}_{y,t}\right|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(\mathbf{y}_t - \boldsymbol{\mu}_{y,t})\boldsymbol{\Sigma}_{y,t}^{-1}(\mathbf{y}_t - \boldsymbol{\mu}_{y,t})^T \right\}. \tag{A21}$$

The probability function $\pi_{\mathbf{Y}_t|\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{y}_t)$ is a density of the d dimensional random variable which follows Gaussian distribution with the mean vector $\boldsymbol{\mu}_{y,t}$ and the covariance matrix $\boldsymbol{\Sigma}_{y,t}$. Using the definition of the density function we obtain that

$$\pi_{\mathbf{Y}_t,\mathbf{X}_t,\mathbf{U}_t,\mathbf{V}_t|\Psi}(\mathbf{y}_t,\mathbf{x}_t,\mathbf{u}_t,\mathbf{v}_t) = \pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{x}_t)\ \pi_{\mathbf{Y}_t|\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{y}_t)\ \pi_{\mathbf{U}_t|\Psi}(\mathbf{u}_t)\ \pi_{\mathbf{V}_t|\Psi}(\mathbf{v}_t)$$
$$\times (\sigma^2)^{-\frac{d-k}{2}} \left|\mathbf{D}_{\epsilon,t}\right|^{\frac{1}{2}} \left|\mathbf{D}_{x,t}\right|^{\frac{1}{2}} \left|\mathbf{M}_t\right|^{-\frac{1}{2}} \left|\sigma^2\mathbf{N}_t^{-1}\right|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\delta_x\left(\mathbf{D}_{x,t}^{-1} - \sigma^2\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbf{N}_t^{-1}\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1} - \sigma^2\mathbf{M}_t^{-1}\right)\delta_x^T\right\}$$

Therefore, denoting

$$C(\mathbf{u}_t,\mathbf{v}_t;\Psi) = \pi_{\mathbf{U}_t|\Psi}(\mathbf{u}_t)\cdot\pi_{\mathbf{V}_t|\Psi}(\mathbf{v}_t)(\sigma^2)^{\frac{k}{2}}\left|\mathbf{D}_{\epsilon,t}\right|^{\frac{1}{2}}\left|\mathbf{D}_{x,t}\right|^{\frac{1}{2}}\left|\mathbf{M}_t\right|^{-\frac{1}{2}}\left|\mathbf{N}_t\right|^{-\frac{1}{2}}$$
$$\times\exp\left\{-\frac{1}{2}\delta_x\left(\mathbf{D}_{x,t}^{-1} - \sigma^2\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbf{N}_t^{-1}\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1} - \sigma^2\mathbf{M}_t^{-1}\right)\delta_x^T\right\}$$

we obtain the following decomposition of the joint probability function

$$\pi_{\mathbf{Y}_t,\mathbf{X}_t,\mathbf{U}_t,\mathbf{V}_t|\Psi}(\mathbf{y}_t,\mathbf{x}_t,\mathbf{u}_t,\mathbf{v}_t) = \pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{x}_t)\cdot\pi_{\mathbf{Y}_t|\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{y}_t)\cdot C(\mathbf{u}_t,\mathbf{v}_t;\Psi).$$

□

Lemma A9. *Let us recall the probability function $\pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{x}_t)$ of a k-dimensional Gaussian random variable $\mathbf{X}_t$ with the covariance matrix $\boldsymbol{\Sigma}_{x,t}$ and the mean vector $\boldsymbol{\mu}_{x,t}$ defined in Lemma A8. Under the assumptions of the Generalized Skew t-Student IND PPCA specified in Section 3.4, the solution to the following integration problem*

$$\int_{\mathbb{R}^k}\log\left(\pi_{\mathbf{Y}_t,\mathbf{X}_t,\mathbf{U}_t,\mathbf{V}_t|\Psi^*}(\mathbf{y}_t,\mathbf{x}_t,\mathbf{u}_t,\mathbf{v}_t)\right)\pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{x}_t)\,d\mathbf{x}_t$$

is equal to the function $w:\mathbb{R}^d\times\mathbb{R}_+^d\times\mathbb{R}_+^k\longrightarrow\mathbb{R}$ defined as following

$$w(\mathbf{y}_t,\mathbf{u}_t,\mathbf{v}_t;\Psi,\Psi^*) = \log\pi_{\mathbf{U}_t|\Psi^*}(\mathbf{u}_t) + \log\pi_{\mathbf{V}_t|\Psi^*}(\mathbf{v}_t) - \frac{k+d}{2}\log 2\pi - \frac{d}{2}\log\sigma^{*2} + \frac{1}{2}\sum_{i=1}^{d}\log u_t^i + \frac{1}{2}\sum_{j=1}^{k}\log v_t^j - \frac{1}{2}\delta_x^*\mathbf{D}_{x,t}^{-1}\delta_x^*$$
$$-\frac{1}{2\sigma^{*2}}\left(\mathbf{y}_t-\boldsymbol{\mu}^*-\delta_\epsilon^*\mathbf{D}_{\epsilon,t}^{-1}\right)\mathbf{D}_{\epsilon,t}\left(\mathbf{y}_t-\boldsymbol{\mu}^*-\delta_\epsilon^*\mathbf{D}_{\epsilon,t}^{-1}\right)^T - \frac{1}{2}Tr\left\{\sigma^2\mathbf{M}_t^{-1}\left(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\mathbf{W}^*+\mathbf{D}_{x,t}\right)\right\}$$
$$+\left(\left(\mathbf{y}_t-\boldsymbol{\mu}-\delta_\epsilon\mathbf{D}_{\epsilon,t}^{-1}\right)\mathbf{D}_{\epsilon,t}\mathbf{W}+\sigma^2\delta_x\right)\mathbf{M}_t^{-1}\left(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\left(\mathbf{y}_t-\boldsymbol{\mu}^*-\delta_\epsilon^*\mathbf{D}_{\epsilon,t}^{-1}\right)^T+\delta_x^{*T}\right)$$
$$-\frac{1}{2}Tr\left\{\mathbf{M}_t^{-1}\left((\mathbf{y}_t-\boldsymbol{\mu}-\delta_\epsilon\mathbf{D}_{\epsilon,t}^{-1})\mathbf{D}_{\epsilon,t}\mathbf{W}+\sigma^2\delta_x\right)^T\left((\mathbf{y}_t-\boldsymbol{\mu}-\delta_\epsilon\mathbf{D}_{\epsilon,t}^{-1})\mathbf{D}_{\epsilon,t}\mathbf{W}+\sigma^2\delta_x\right)\mathbf{M}_t^{-1}\left(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\mathbf{W}^*+\mathbf{D}_{x,t}\right)\right\}$$

where $\mathbf{M}_t=\sigma^2\mathbf{D}_{x,t}+\mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbf{W}$ and $\boldsymbol{\mu}_{x,t}$ being specified in Lemma A8.

Proof. Let us recall the probability function $\pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{x}_t)$ defined in Lemma A8 which is a density of a k-dimensional Gaussian random variable $\mathbf{X}_t$ with the covariance matrix $\boldsymbol{\Sigma}_{x,t}=\sigma^2\mathbf{M}_t^{-1}$ and the mean vector $\boldsymbol{\mu}_{x,t}=\left((\mathbf{y}_t-\boldsymbol{\mu}-\delta_\epsilon\mathbf{D}_{\epsilon,t}^{-1})\mathbf{D}_{\epsilon,t}\mathbf{W}+\sigma^2\delta_x\right)\mathbf{M}_t^{-1}$, where $\mathbf{M}_t=\sigma^2\mathbf{D}_{x,t}+\mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbf{W}$. Hence, the solution to the integration problem is given by

$$w(\mathbf{y}_t,\mathbf{u}_t,\mathbf{v}_t;\Psi,\Psi^*) = \int_{\mathbb{R}^k}\log\left(\pi_{\mathbf{Y}_t,\mathbf{X}_t,\mathbf{U}_t,\mathbf{V}_t|\Psi^*}(\mathbf{y}_t,\mathbf{x}_t,\mathbf{u}_t,\mathbf{v}_t)\right)\pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{x}_t)\,d\mathbf{x}_t$$
$$=\int_{\mathbb{R}^k}\left(\log\pi_{\mathbf{Y}_t|\mathbf{X}_t,\mathbf{U}_t,\mathbf{V}_t|\Psi^*}(\mathbf{y}_t)+\log\pi_{\mathbf{X}_t|\mathbf{U}_t,\mathbf{V}_t,\Psi^*}(\mathbf{x}_t)+\log\pi_{\mathbf{U}_t|\Psi^*}(\mathbf{u}_t)+\log\pi_{\mathbf{V}_t|\Psi^*}(\mathbf{v}_t)\right)\pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{x}_t)\,d\mathbf{x}_t$$
$$=\log\pi_{\mathbf{U}_t|\Psi^*}(\mathbf{u}_t)+\log\pi_{\mathbf{V}_t|\Psi^*}(\mathbf{v}_t)+\int_{\mathbb{R}^k}\left(\log\pi_{\mathbf{Y}_t|\mathbf{X}_t,\mathbf{U}_t,\mathbf{V}_t|\Psi^*}(\mathbf{y}_t)+\log\pi_{\mathbf{X}_t|\mathbf{U}_t,\mathbf{V}_t,\Psi^*}(\mathbf{x}_t)\right)\pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{x}_t)\,d\mathbf{x}_t$$
$$=\log\pi_{\mathbf{U}_t|\Psi^*}(\mathbf{u}_t)+\log\pi_{\mathbf{V}_t|\Psi^*}(\mathbf{v}_t)-\frac{k+d}{2}\log 2\pi+\log\left|\sigma^{*2}\mathbf{D}_{\epsilon,t}^{-1}\right|^{-\frac{1}{2}}+\log\left|\mathbf{D}_{x,t}^{-1}\right|^{-\frac{1}{2}}$$

$$
\begin{aligned}
&- \frac{1}{2}\int_{\mathbb{R}^k} \frac{1}{\sigma^{*2}}\Big(\mathbf{y}_t - \boldsymbol{\mu}^* - \boldsymbol{\delta}_\epsilon^*\mathbf{D}_{\epsilon,t}^{-1} - \mathbf{x}_t\mathbf{W}^{*T}\Big)\mathbf{D}_{\epsilon,t}\Big(\mathbf{y}_t - \boldsymbol{\mu}^* - \boldsymbol{\delta}_\epsilon^*\mathbf{D}_{\epsilon,t}^{-1} - \mathbf{x}_t\mathbf{W}^{*T}\Big)^T \pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{x}_t)\, d\mathbf{x}_t \\
&- \frac{1}{2}\int_{\mathbb{R}^k} \Big(\mathbf{x}_t - \boldsymbol{\delta}_x^*\mathbf{D}_{x,t}^{-1}\Big)\mathbf{D}_{x,t}\Big(\mathbf{x}_t - \boldsymbol{\delta}_x^*\mathbf{D}_{x,t}^{-1}\Big)^T \pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{x}_t)\, d\mathbf{x}_t \\
&= \log \pi_{\mathbf{U}_t|\Psi^*}(\mathbf{u}_t) + \log \pi_{\mathbf{V}_t|\Psi^*}(\mathbf{v}_t) - \frac{k+d}{2}\log 2\pi - \frac{d}{2}\log\sigma^{*2} + \frac{1}{2}\sum_{i=1}^{d}\log u_t^i + \frac{1}{2}\sum_{j=1}^{k}\log v_t^j \\
&- \frac{1}{2\sigma^{*2}}\Big(\mathbf{y}_t - \boldsymbol{\mu}^* - \boldsymbol{\delta}_\epsilon^*\mathbf{D}_{\epsilon,t}^{-1}\Big)\mathbf{D}_{\epsilon,t}\Big(\mathbf{y}_t - \boldsymbol{\mu}^* - \boldsymbol{\delta}_\epsilon^*\mathbf{D}_{\epsilon,t}^{-1}\Big)^T + \frac{1}{\sigma^{*2}}\int_{\mathbb{R}^k}\mathbf{x}_t\pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{x}_t)\, d\mathbf{x}_t\ \mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\Big(\mathbf{y}_t - \boldsymbol{\mu}^* - \boldsymbol{\delta}_\epsilon^*\mathbf{D}_{\epsilon,t}^{-1}\Big)^T \\
&- \frac{1}{2}\int_{\mathbb{R}^k}\mathbf{x}_t\Big(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\mathbf{W}^* + \mathbf{D}_{x,t}\Big)\mathbf{x}_t^T\pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{x}_t)\, d\mathbf{x}_t + \int_{\mathbb{R}^k}\mathbf{x}_t\pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{x}_t)\, d\mathbf{x}_t\boldsymbol{\delta}_x^{*T} - \frac{1}{2}\boldsymbol{\delta}_x^*\mathbf{D}_{x,t}^{-1}\boldsymbol{\delta}_x^{*T} \\
&= \log \pi_{\mathbf{U}_t|\Psi^*}(\mathbf{u}_t) + \log \pi_{\mathbf{V}_t|\Psi^*}(\mathbf{v}_t) - \frac{k+d}{2}\log 2\pi - \frac{1}{2}\log\sigma^{*2} + \frac{1}{2}\sum_{i=1}^{d}\log u_t^i + \frac{1}{2}\sum_{j=1}^{k}\log v_t^j - \frac{1}{2}\boldsymbol{\delta}_x^*\mathbf{D}_{x,t}^{-1}\boldsymbol{\delta}_x^{*T} \\
&- \frac{1}{2\sigma^{*2}}\Big(\mathbf{y}_t - \boldsymbol{\mu}^* - \boldsymbol{\delta}_\epsilon^*\mathbf{D}_{\epsilon,t}^{-1}\Big)\mathbf{D}_{\epsilon,t}\Big(\mathbf{y}_t - \boldsymbol{\mu}^* - \boldsymbol{\delta}_\epsilon^*\mathbf{D}_{\epsilon,t}^{-1}\Big)^T + \boldsymbol{\mu}_{x,t}\Big(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\Big(\mathbf{y}_t - \boldsymbol{\mu}^* - \boldsymbol{\delta}_\epsilon^*\mathbf{D}_{\epsilon,t}^{-1}\Big)^T + \boldsymbol{\delta}_x^{*T}\Big) \\
&- \frac{1}{2}\operatorname{Tr}\Big\{\Big(\boldsymbol{\Sigma}_{x,t} + \boldsymbol{\mu}_{x,t}^T\boldsymbol{\mu}_{x,t}\Big)\Big(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\mathbf{W}^* + \mathbf{D}_{x,t}\Big)\Big\} \\
&= \log \pi_{\mathbf{U}_t|\Psi^*}(\mathbf{u}_t) + \log \pi_{\mathbf{V}_t|\Psi^*}(\mathbf{v}_t) - \frac{k+d}{2}\log 2\pi - \frac{d}{2}\log\sigma^{*2} + \frac{1}{2}\sum_{i=1}^{d}\log u_t^i + \frac{1}{2}\sum_{j=1}^{k}\log v_t^j - \frac{1}{2}\boldsymbol{\delta}_x^*\mathbf{D}_{x,t}^{-1}\boldsymbol{\delta}_x^{*T} \\
&- \frac{1}{2\sigma^{*2}}\Big(\mathbf{y}_t - \boldsymbol{\mu}^* - \boldsymbol{\delta}_\epsilon^*\mathbf{D}_{\epsilon,t}^{-1}\Big)\mathbf{D}_{\epsilon,t}\Big(\mathbf{y}_t - \boldsymbol{\mu}^* - \boldsymbol{\delta}_\epsilon^*\mathbf{D}_{\epsilon,t}^{-1}\Big)^T - \frac{1}{2}\operatorname{Tr}\Big\{\sigma^2\mathbf{M}_t^{-1}\Big(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\mathbf{W}^* + \mathbf{D}_{x,t}\Big)\Big\} \\
&+ \Big(\Big(\mathbf{y}_t - \boldsymbol{\mu} - \boldsymbol{\delta}_\epsilon\mathbf{D}_{\epsilon,t}^{-1}\Big)\mathbf{D}_{\epsilon,t}\mathbf{W} + \sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}_t^{-1}\Big(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\Big(\mathbf{y}_t - \boldsymbol{\mu}^* - \boldsymbol{\delta}_\epsilon^*\mathbf{D}_{\epsilon,t}^{-1}\Big)^T + \boldsymbol{\delta}_x^{*T}\Big) \\
&- \frac{1}{2}\operatorname{Tr}\Big\{\mathbf{M}_t^{-1}\Big(\Big(\mathbf{y}_t - \boldsymbol{\mu} - \boldsymbol{\delta}_\epsilon\mathbf{D}_{\epsilon,t}^{-1}\Big)\mathbf{D}_{\epsilon,t}\mathbf{W} + \sigma^2\boldsymbol{\delta}_x\Big)^T\Big(\Big(\mathbf{y}_t - \boldsymbol{\mu} - \boldsymbol{\delta}_\epsilon\mathbf{D}_{\epsilon,t}^{-1}\Big)\mathbf{D}_{\epsilon,t}\mathbf{W} + \sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}_t^{-1}\Big(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\mathbf{W}^* + \mathbf{D}_{x,t}\Big)\Big\}
\end{aligned}
$$

□

Lemma A10. *Let us consider the partition of the observation vector $\mathbf{Y}_t = [\mathbf{Y}_t^o, \mathbf{Y}_t^m]$ into the the subvector with observed and missing entries, $\mathbf{Y}_t^o$ and $\mathbf{Y}_t^m$, respectively. The observation vector $\mathbf{Y}_t$ is modelled as in Equation (2.7) under the assumptions stated in Subsection 3.4.2.1. The conditional distribution of the random vector $\mathbf{Y}_t|\mathbf{U}_t, \mathbf{V}_t, \Psi$, derived in Lemma A8, is Gaussian with the mean vector $\boldsymbol{\mu}_{y,t} = \boldsymbol{\mu} + \boldsymbol{\delta}_\epsilon\mathbf{D}_{\epsilon,t}^{-1} + \sigma^2\boldsymbol{\delta}_x\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{N}_t^{-1}$ and the covariance matrix $\boldsymbol{\Sigma}_{y,t} = \sigma^2\mathbf{D}_{\epsilon,t}^{-1}\mathbf{N}_t^{-1}$, where $\mathbf{M}_t = \sigma^2\mathbf{D}_{x,t} + \mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbf{W}$ and $\mathbf{N}_t = \Big(\mathbb{I}_d - \mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\Big)$. The conditional distribution $\mathbf{Y}_t^m|\mathbf{Y}_t^o, \mathbf{U}_t, \mathbf{V}_t, \Psi$ is also Gaussian with the mean vector $\tilde{\boldsymbol{\mu}}_{Y^m,t}$ and the covariance matrix $\tilde{\boldsymbol{\Sigma}}_{Y^m,t}$ such that*

$$
\begin{aligned}
\tilde{\boldsymbol{\mu}}_{Y^m,t} &= \boldsymbol{\mu}_{y,t}^m + \big(\mathbf{Y}_t^o - \boldsymbol{\mu}_{y,t}^m\big)\boldsymbol{\Sigma}_{y,t}^{oo\ -1}\boldsymbol{\Sigma}_{y,t}^{om} \\
\tilde{\boldsymbol{\Sigma}}_{Y^m,t} &= \boldsymbol{\Sigma}_{y,t}^{mm} - \boldsymbol{\Sigma}_{y,t}^{mo}\boldsymbol{\Sigma}_{y,t}^{oo\ -1}\boldsymbol{\Sigma}_{y,t}^{om}.
\end{aligned}
$$

In addition, the marginal distribution of $\mathbf{Y}_t^o|\mathbf{U}_t, \mathbf{V}_t, \Psi$ is Gaussian with the mean vector $\boldsymbol{\mu}_{y,t}^o$ and the covariance matrix $\boldsymbol{\Sigma}_{y,t}^{oo}$. The subvectors $\boldsymbol{\mu}_{y,t}^o$ and $\boldsymbol{\mu}_{y,t}^m$ contain elements of the vector $\boldsymbol{\mu}_{y,t}$ which correspond to the observed or missing entries of the observation vector $\mathbf{Y}_t$, respectively, and are d_o and d_m-dimensional. The square matrices $\boldsymbol{\Sigma}_{y,t}^{oo}$ and $\boldsymbol{\Sigma}_{y,t}^{mm}$ contain elements of the matrix $\boldsymbol{\Sigma}_{y,t}$ which correspond by rows and columns to the observed or missing entries of the observation vector $\mathbf{Y}_t$, respectively, and are $d_o \times d_o$ and $d_m \times d_m$-dimensional. The non-square matrix $\boldsymbol{\Sigma}_{y,t}^{mo} = \boldsymbol{\Sigma}_{y,t}^{om\ T}$ contains elements of the matrix $\boldsymbol{\Sigma}_{y,t}$ which correspond by rows to the missing and by columns to the observed entries of the observation vector $\mathbf{Y}_t$, and is $d_m \times d_o$-dimensional.

Proof. Please refer to Lemma A8 for the derivation of the conditional distribution $\mathbf{Y}_t|\mathbf{U}_t, \mathbf{V}_t, \Psi$ and apply the formulas for the standard conditional Gaussian distribution from Lemma A1. □

Lemma A11. *Let us consider the partition of the observation vector* $\mathbf{Y}_t = [\mathbf{Y}_t^o, \mathbf{Y}_t^m]$ *into the the subvector with observed and missing entries,* $\mathbf{Y}_t^o$ *and* $\mathbf{Y}_t^m$*, respectively. The observation vector* $\mathbf{Y}_t$ *is modelled as in Equation (2.7) under the assumptions of Independent Generalized Skew t-Student PPCA stated in Subsection 3.4.2.1. Let us recall the function* $w(\mathbf{y}_t, \mathbf{u}_t, \mathbf{v}_t; \Psi, \Psi^*)$ *defined as in Lemma A9 and the conditional distribution of the random vector* $\mathbf{Y}_t^m|\mathbf{Y}_t^o, \mathbf{U}_t, \mathbf{V}_t, \Psi$ *specified in Lemma A10. Given the realisation of the vector* $\mathbf{Y}_t$ *in time points* $t = 1, \ldots, N$ *with observable components* $\mathbf{y}_{1:N}^o = [\mathbf{y}_1^o, \ldots, \mathbf{y}_N^o]$ *the function* v *defined by the following sum of integration problems*

$$v(\mathbf{y}_t^o, \mathbf{u}_t, \mathbf{v}_t; \Psi, \Psi^*) = \int_{\mathbb{R}^{d_m}} w(\mathbf{y}_t, \mathbf{u}_t, \mathbf{v}_t; \Psi, \Psi^*) \cdot \pi_{\mathbf{Y}_t^m|\mathbf{Y}_t^o, \mathbf{U}_t, \mathbf{V}_t, \Psi}(\mathbf{y}_t^m)\, d\mathbf{y}_t^m \tag{A22}$$

for $\tilde{\boldsymbol{\mu}}_{Y^m,t} = \boldsymbol{\mu}_{y,t}^m + \left(\mathbf{Y}_t^o - \boldsymbol{\mu}_{y,t}^m\right)\boldsymbol{\Sigma}_{y,t}^{oo\ -1}\boldsymbol{\Sigma}_{y,t}^{om}$ *and* $\tilde{\boldsymbol{\Sigma}}_{Y^m,t} = \boldsymbol{\Sigma}_{y,t}^{mm} - \boldsymbol{\Sigma}_{y,t}^{mo}\boldsymbol{\Sigma}_{y,t}^{oo\ -1}\boldsymbol{\Sigma}_{y,t}^{om}$*, where* $\boldsymbol{\mu}_{y,t} = \boldsymbol{\mu} + \delta_\epsilon \mathbf{D}_{\epsilon,t}^{-1} + \sigma^2\delta_x\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{N}_t^{-1}$ *and* $\boldsymbol{\Sigma}_{y,t} = \sigma^2\mathbf{D}_{\epsilon,t}^{-1}\mathbf{N}_t^{-1}$*, given* $\mathbf{M}_t = \sigma^2\mathbf{D}_{x,t} + \mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbf{W}$ *and* $\mathbf{N}_t = \left(\mathbb{I}_d - \mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\right)$*, which are the central moments of the conditional distribution* $\mathbf{Y}_t|\mathbf{U}_t, \mathbf{V}_t, \Psi$ *derived in Lemma A8. Hence, we can express the integral (A22) using the defined moments of the variable* $\mathbf{Y}_t^m|\mathbf{Y}_t^o, \mathbf{U}_t, \mathbf{V}_t, \Psi$*, that is*

$$\begin{aligned}
v(\mathbf{y}_t^o, \mathbf{u}_t, \mathbf{v}_t; \Psi, \Psi^*) &= \log \pi_{\mathbf{U}_t|\Psi^*}(\mathbf{u}_t) + \log \pi_{\mathbf{V}_t|\Psi^*}(\mathbf{v}_t) - \frac{k+d}{2}\log 2\pi - \frac{d}{2}\log\sigma^{*2} + \frac{1}{2}\sum_{i=1}^{d}\log u_t^i + \frac{1}{2}\sum_{j=1}^{k}\log v_t^j \\
&- \frac{1}{2}\delta_x^*\mathbf{D}_{x,t}^{-1}\delta_x^{*T} - \frac{1}{2\sigma^{*2}}Tr\left\{\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U}_t,\mathbf{V}_t,\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right]\mathbf{D}_{\epsilon,t}\right\} + \frac{1}{\sigma^{*2}}\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U}_t,\mathbf{V}_t,\Psi}\left[\mathbf{Y}_t\right]\left(\mathbf{D}_{\epsilon,t}\boldsymbol{\mu}^{*T} + \delta_\epsilon^{*T}\right) \\
&- \frac{1}{2\sigma^{*2}}\boldsymbol{\mu}^*\mathbf{D}_{\epsilon,t}\boldsymbol{\mu}^{*T} - \frac{1}{\sigma^{*2}}\boldsymbol{\mu}^*\delta_\epsilon^{*T} - \frac{1}{2\sigma^{*2}}\delta_\epsilon^*\mathbf{D}_{\epsilon,t}^{-1}\delta_\epsilon^{*T} + \frac{1}{\sigma^{*2}}Tr\left\{\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U}_t,\mathbf{V}_t,\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right]\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\right\} \\
&+ \mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U}_t,\mathbf{V}_t,\Psi}\left[\mathbf{Y}_t\right]\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\left(-\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\boldsymbol{\mu}^{*T} - \frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\delta_\epsilon^{*T} + \delta_x^{*T}\right) \\
&+ \frac{1}{\sigma^{*2}}\left(-\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W} - \delta_\epsilon\mathbf{W} + \sigma^2\delta_x\right)\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U}_t,\mathbf{V}_t,\Psi}\left[\mathbf{Y}_t\right]^T \\
&+ \left(-\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W} - \delta_\epsilon\mathbf{W} + \sigma^2\delta_x\right)\mathbf{M}_t^{-1}\left(-\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\boldsymbol{\mu}^{*T} - \frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\delta_\epsilon^{*T} + \delta_x^{*T}\right) \\
&- \frac{1}{2}Tr\left\{\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U}_t,\mathbf{V}_t,\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right]\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\left(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\mathbf{W}^* + \mathbf{D}_{x,t}\right)\right\} \\
&- Tr\left\{\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U}_t,\mathbf{V}_t,\Psi}\left[\mathbf{Y}_t\right]^T\left(-\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W} - \delta_\epsilon\mathbf{W} + \sigma^2\delta_x\right)\mathbf{M}_t^{-1}\left(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\mathbf{W}^* + \mathbf{D}_{x,t}\right)\right\} \\
&- \frac{1}{2}Tr\left\{\mathbf{M}_t^{-1}\left(-\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W} - \delta_\epsilon\mathbf{W} + \sigma^2\delta_x\right)^T\left(-\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W} - \delta_\epsilon\mathbf{W} + \sigma^2\delta_x\right)\mathbf{M}_t^{-1}\left(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\mathbf{W}^* + \mathbf{D}_{x,t}\right)\right\} \\
&- \frac{\sigma^2}{2}Tr\left\{\mathbf{M}_t^{-1}\left(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\mathbf{W}^* + \mathbf{D}_{x,t}\right)\right\}.
\end{aligned}$$

Using the mean $\tilde{\boldsymbol{\mu}}_{Y^m,t}$ *and the covariance matrix* $\tilde{\boldsymbol{\Sigma}}_{Y^m,t}$ *of the random vector* $\mathbf{Y}_t^m|\mathbf{Y}_t^o, \mathbf{U}_t, \mathbf{V}_t, \Psi$ *from Lemma A10, the first and second non-central moments of the vector* $\mathbf{Y}_t$ *are the following*

$$\begin{aligned}
\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U}_t,\mathbf{V}_t,\Psi}\left[\mathbf{Y}_t\right] &= \left[\mathbf{y}_t^o,\ \tilde{\boldsymbol{\mu}}_{Y^m,t}\right], \\
\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U}_t,\mathbf{V}_t,\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right] &= \begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \tilde{\boldsymbol{\Sigma}}_{Y^m,t}\end{bmatrix} + \mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U}_t,\mathbf{V}_t,\Psi}\left[\mathbf{Y}_t\right]^T\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U}_t,\mathbf{V}_t,\Psi}\left[\mathbf{Y}_t\right],
\end{aligned}$$

Proof. Let us recall, that the observation vector $\mathbf{Y}_t = [\mathbf{Y}_t^o, \mathbf{Y}_t^m]$ is partitioned into the d_o-dimensional subvector with observed entries, $\mathbf{Y}_t^o$ and the d_m-dimensional subvector with unobserved entries, $\mathbf{Y}_t^m$, such that $d_m = d - d_m$. Given the function $w(\mathbf{y}_t, \mathbf{u}_t, \mathbf{v}_t; \Psi, \Psi^*)$ defined in Lemma A9, the solution to the integration problem from Equation (A22) is derived as follows

$$v(\mathbf{y}_t^o, \mathbf{u}_t, \mathbf{v}_t; \Psi, \Psi^*) = \int_{\mathbb{R}^{d_m}} w(\mathbf{y}_t, \mathbf{u}_t, \mathbf{v}_t; \Psi, \Psi^*) \cdot \pi_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{y}_t^m)\, d\mathbf{y}_t^m$$

$$
\begin{aligned}
&= \log \pi_{\mathbf{U}_t|\Psi^*}(\mathbf{u}_t) + \log \pi_{\mathbf{V}_t|\Psi^*}(\mathbf{v}_t) - \frac{k+d}{2}\log 2\pi - \frac{d}{2}\log \sigma^{*2} + \frac{1}{2}\sum_{i=1}^{d}\log u_t^i + \frac{1}{2}\sum_{j=1}^{k}\log v_t^j - \frac{1}{2}\boldsymbol{\delta}_x^* \mathbf{D}_{x,t}^{-1}\boldsymbol{\delta}_x^{*T} \\
&- \frac{1}{2\sigma^{*2}}\int_{\mathbb{R}^{d_m}} \mathbf{y}_t \mathbf{D}_{\epsilon,t}\mathbf{y}_t^T \; \pi_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{y}_t^m)\, d\mathbf{y}_t^m + \frac{1}{\sigma^{*2}}\int_{\mathbb{R}^{d_m}} \mathbf{y}_t \; \pi_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{y}_t^m)\, d\mathbf{y}_t^m \Big(\mathbf{D}_{\epsilon,t}\boldsymbol{\mu}^{*T} + \boldsymbol{\delta}_\epsilon^{*T}\Big) \\
&- \frac{1}{2\sigma^{*2}}\boldsymbol{\mu}^*\mathbf{D}_{\epsilon,t}\boldsymbol{\mu}^{*T} - \frac{1}{\sigma^{*2}}\boldsymbol{\mu}^*\boldsymbol{\delta}_\epsilon^{*T} - \frac{1}{2\sigma^{*2}}\boldsymbol{\delta}_\epsilon^*\mathbf{D}_{\epsilon,t}^{-1}\boldsymbol{\delta}_\epsilon^{*T} + \frac{1}{\sigma^{*2}}\int_{\mathbb{R}^{d_m}} \mathbf{y}_t\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\mathbf{y}_t^T \; \pi_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{y}_t^m)\, d\mathbf{y}_t^m \\
&+ \int_{\mathbb{R}^{d_m}} \mathbf{y}_t \; \pi_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{y}_t^m)\, d\mathbf{y}_t^m \mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\Big(-\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\boldsymbol{\mu}^{*T} - \frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\boldsymbol{\delta}_\epsilon^{*T} + \boldsymbol{\delta}_x^{*T}\Big) \\
&+ \frac{1}{\sigma^{*2}}\int_{\mathbb{R}^{d_m}} \mathbf{y}_t \; \pi_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{y}_t^m)\, d\mathbf{y}_t^m \mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}_t^{-1}\Big(-\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W} - \boldsymbol{\delta}_\epsilon\mathbf{W} + \sigma^2\boldsymbol{\delta}_x\Big)^T \\
&+ \Big(-\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W} - \boldsymbol{\delta}_\epsilon\mathbf{W} + \sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}_t^{-1}\Big(-\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\boldsymbol{\mu}^{*T} - \frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\boldsymbol{\delta}_\epsilon^{*T} + \boldsymbol{\delta}_x^{*T}\Big) \\
&- \frac{1}{2}\operatorname{Tr}\Big\{\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\int_{\mathbb{R}^{d_m}} \mathbf{y}_t^T\mathbf{y}_t \; \pi_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{y}_t^m)\, d\mathbf{y}_t^m \mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\Big(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\mathbf{W}^* + \mathbf{D}_{x,t}\Big)\Big\} \\
&- \operatorname{Tr}\Big\{\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\int_{\mathbb{R}^{d_m}} \mathbf{y}_t^T \; \pi_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{y}_t^m)\, d\mathbf{y}_t^m\Big(-\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W} - \boldsymbol{\delta}_\epsilon\mathbf{W} + \sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}_t^{-1}\Big(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\mathbf{W}^* + \mathbf{D}_{x,t}\Big)\Big\} \\
&- \frac{1}{2}\operatorname{Tr}\Big\{\mathbf{M}_t^{-1}\Big(-\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W} - \boldsymbol{\delta}_\epsilon\mathbf{W} + \sigma^2\boldsymbol{\delta}_x\Big)^T\Big(-\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W} - \boldsymbol{\delta}_\epsilon\mathbf{W} + \sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}_t^{-1}\Big(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\mathbf{W}^* + \mathbf{D}_{x,t}\Big)\Big\} \\
&- \frac{\sigma^2}{2}\operatorname{Tr}\Big\{\mathbf{M}_t^{-1}\Big(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\mathbf{W}^* + \mathbf{D}_{x,t}\Big)\Big\}.
\end{aligned}
$$

The probability function $\pi_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{y}_t^m)$ is specified in Lemma A10 as well as the first and second non-central moments of the distribution $\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U}_t,\mathbf{V}_t,\Psi$. Consequently we have the following solutions to the integrations problems

$$
\begin{aligned}
&\int_{\mathbb{R}^{d_m}} \mathbf{y}_t \; \pi_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{y}_t^m)\, d\mathbf{y}_t^m = \mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U}_t,\mathbf{V}_t,\Psi}\big[\mathbf{Y}_t\big] = \big[\mathbf{Y}_t^o,\; \tilde{\boldsymbol{\mu}}_{Y^m,t}\big], \\
&\int_{\mathbb{R}^{d_m}} \mathbf{y}_t^T\mathbf{y}_t \; \pi_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U}_t,\mathbf{V}_t,\Psi}(\mathbf{y}_t^m)\, d\mathbf{y}_t^m = \mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U}_t,\mathbf{V}_t,\Psi}\big[\mathbf{Y}_t^T\mathbf{Y}_t\big] = \begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \tilde{\boldsymbol{\Sigma}}_{Y^m,t}\end{bmatrix} + \mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U}_t,\mathbf{V}_t,\Psi}\big[\mathbf{Y}_t\big]^T\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U}_t,\mathbf{V}_t,\Psi}\big[\mathbf{Y}_t\big],
\end{aligned}
$$

for $\tilde{\boldsymbol{\mu}}_{Y^m,t} = \boldsymbol{\mu}_{y,t}^m + \big(\mathbf{Y}_t^o - \boldsymbol{\mu}_{y,t}^m\big)\boldsymbol{\Sigma}_{y,t}^{oo\;-1}\boldsymbol{\Sigma}_{y,t}^{om}$ and $\tilde{\boldsymbol{\Sigma}}_{Y^m,t} = \boldsymbol{\Sigma}_{y,t}^{mm} - \boldsymbol{\Sigma}_{y,t}^{mo}\boldsymbol{\Sigma}_{y,t}^{oo\;-1}\boldsymbol{\Sigma}_{y,t}^{om}$, where $\boldsymbol{\mu}_{y,t} = \boldsymbol{\mu} + \boldsymbol{\delta}_\epsilon\mathbf{D}_{\epsilon,t}^{-1} + \sigma^2\boldsymbol{\delta}_x\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{N}_t^{-1}$ and $\boldsymbol{\Sigma}_{y,t} = \sigma^2\mathbf{D}_{\epsilon,t}^{-1}\mathbf{N}_t^{-1}$, given $\mathbf{M}_t = \sigma^2\mathbf{D}_{x,t} + \mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbf{W}$ and $\mathbf{N}_t = \Big(\mathbb{I}_d - \mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\Big)$, which are the central moments of the conditional distribution $\mathbf{Y}_t|\mathbf{U}_t,\mathbf{V}_t,\Psi$ derived in Lemma A8. Hence, we can express the integral (A22) using the moments of the random vector $\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U}_t,\mathbf{V}_t,\Psi$, that is

$$
\begin{aligned}
v(\mathbf{y}_t^o,\mathbf{u}_t,\mathbf{v}_t;\Psi,\Psi^*) &= \log \pi_{\mathbf{U}_t|\Psi^*}(\mathbf{u}_t) + \log \pi_{\mathbf{V}_t|\Psi^*}(\mathbf{v}_t) - \frac{k+d}{2}\log 2\pi - \frac{d}{2}\log \sigma^{*2} + \frac{1}{2}\sum_{i=1}^{d}\log u_t^i + \frac{1}{2}\sum_{j=1}^{k}\log v_t^j \\
&- \frac{1}{2}\boldsymbol{\delta}_x^*\mathbf{D}_{x,t}^{-1}\boldsymbol{\delta}_x^{*T} - \frac{1}{2\sigma^{*2}}\operatorname{Tr}\Big\{\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U}_t,\mathbf{V}_t,\Psi}\big[\mathbf{Y}_t^T\mathbf{Y}_t\big]\mathbf{D}_{\epsilon,t}\Big\} + \frac{1}{\sigma^{*2}}\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U}_t,\mathbf{V}_t,\Psi}\big[\mathbf{Y}_t\big]\Big(\mathbf{D}_{\epsilon,t}\boldsymbol{\mu}^{*T} + \boldsymbol{\delta}_\epsilon^{*T}\Big) \\
&- \frac{1}{2\sigma^{*2}}\boldsymbol{\mu}^*\mathbf{D}_{\epsilon,t}\boldsymbol{\mu}^{*T} - \frac{1}{\sigma^{*2}}\boldsymbol{\mu}^*\boldsymbol{\delta}_\epsilon^{*T} - \frac{1}{2\sigma^{*2}}\boldsymbol{\delta}_\epsilon^*\mathbf{D}_{\epsilon,t}^{-1}\boldsymbol{\delta}_\epsilon^{*T} + \frac{1}{\sigma^{*2}}\operatorname{Tr}\Big\{\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U}_t,\mathbf{V}_t,\Psi}\big[\mathbf{Y}_t^T\mathbf{Y}_t\big]\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\Big\} \\
&+ \mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U}_t,\mathbf{V}_t,\Psi}\big[\mathbf{Y}_t\big]\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\Big(-\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\boldsymbol{\mu}^{*T} - \frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\boldsymbol{\delta}_\epsilon^{*T} + \boldsymbol{\delta}_x^{*T}\Big) \\
&+ \frac{1}{\sigma^{*2}}\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U}_t,\mathbf{V}_t,\Psi}\big[\mathbf{Y}_t\big]\mathbf{D}_{\epsilon,t}\mathbf{W}^*\mathbf{M}_t^{-1}\Big(-\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W} - \boldsymbol{\delta}_\epsilon\mathbf{W} + \sigma^2\boldsymbol{\delta}_x\Big)^T \\
&+ \Big(-\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W} - \boldsymbol{\delta}_\epsilon\mathbf{W} + \sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}_t^{-1}\Big(-\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\boldsymbol{\mu}^{*T} - \frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\boldsymbol{\delta}_\epsilon^{*T} + \boldsymbol{\delta}_x^{*T}\Big) \\
&- \frac{1}{2}\operatorname{Tr}\Big\{\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U}_t,\mathbf{V}_t,\Psi}\big[\mathbf{Y}_t^T\mathbf{Y}_t\big]\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}\Big(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\mathbf{W}^* + \mathbf{D}_{x,t}\Big)\Big\} \\
&- \operatorname{Tr}\Big\{\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U}_t,\mathbf{V}_t,\Psi}\big[\mathbf{Y}_t\big]^T\Big(-\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W} - \boldsymbol{\delta}_\epsilon\mathbf{W} + \sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}_t^{-1}\Big(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\mathbf{W}^* + \mathbf{D}_{x,t}\Big)\Big\} \\
&- \frac{1}{2}\operatorname{Tr}\Big\{\mathbf{M}_t^{-1}\Big(-\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W} - \boldsymbol{\delta}_\epsilon\mathbf{W} + \sigma^2\boldsymbol{\delta}_x\Big)^T\Big(-\boldsymbol{\mu}\mathbf{D}_{\epsilon,t}\mathbf{W} - \boldsymbol{\delta}_\epsilon\mathbf{W} + \sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}_t^{-1}\Big(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\mathbf{W}^* + \mathbf{D}_{x,t}\Big)\Big\} \\
&- \frac{\sigma^2}{2}\operatorname{Tr}\Big\{\mathbf{M}_t^{-1}\Big(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\mathbf{W}^* + \mathbf{D}_{x,t}\Big)\Big\}.
\end{aligned}
$$

□

A.4.6. Supporting Lemmas for Generalized Skew t-Student ID PPCA

Lemma A12. *Let a d dimensional observation vector* $\mathbf{Y}_t$ *be modelled as in Equation (2.7) with the assumptions given in Subsubsection 3.4.1.2, Equation (3.30) and Equation (3.31) for Generalized Skew t-Student ID PPCA. The join probability function of the variables* $\mathbf{Y}_t, \mathbf{X}_t, \mathbf{U}$ *and* $\mathbf{V}$ *can be factorized into the product of three functions*

$$\pi_{\mathbf{Y}_t,\mathbf{X}_t|\mathbf{U},\mathbf{V},\Psi}(\mathbf{y}_t,\mathbf{x}_t) = \pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U},\mathbf{V},\Psi}(\mathbf{x}_t)\cdot\pi_{\mathbf{Y}_t|\mathbf{U},\mathbf{V},\Psi}(\mathbf{y}_t)\cdot C(\mathbf{u},\mathbf{v};\Psi)$$

where $\pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U},\mathbf{V},\Psi}(\mathbf{x}_t)$ *and* $\pi_{\mathbf{Y}_t|\mathbf{U}\mathbf{V},\Psi}(\mathbf{x}_t)$ *are the conditional probability function of the k- dimensional random vector* $\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U},\mathbf{V}\sim\mathcal{N}(\boldsymbol{\mu}_{x,t},\boldsymbol{\Sigma}_{x,t})$ *and d-dimensional random vector* $\mathbf{Y}_t|\mathbf{U},\mathbf{V}\sim\mathcal{N}(\boldsymbol{\mu}_{y,t},\boldsymbol{\Sigma}_{y,t})$*, respectively, for the first and second central moments of the distributions given by*

$$\boldsymbol{\mu}_{x,t} = \Big((\mathbf{y}_t-\boldsymbol{\mu}-\delta_\epsilon\mathbf{D}_\epsilon^{-1})\mathbf{D}_\epsilon\mathbf{W}+\sigma^2\delta_x\Big)\mathbf{M}^{-1} \text{ and } \boldsymbol{\Sigma}_x=\sigma^2\mathbf{M}^{-1},$$
$$\boldsymbol{\mu}_y = \boldsymbol{\mu}+\delta_\epsilon\mathbf{D}_\epsilon^{-1}+\sigma^2\delta_x\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\mathbf{N}^{-1} \text{ and } \boldsymbol{\Sigma}_y=\sigma^2\mathbf{N}^{-1}$$

for $\mathbf{M}=\sigma^2\mathbf{D}_x+\mathbf{W}^T\mathbf{D}_\epsilon\mathbf{W}$ *and* $\mathbf{N}=\mathbf{D}_\epsilon-\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon$*. The function* $C:\mathbb{R}_+^d\times\mathbb{R}_+^k\longrightarrow\mathbb{R}$ *is given by*

$$C(\mathbf{u},\mathbf{v};\Psi)=(\sigma^2)^{\frac{k}{2}}\left|\mathbf{D}_\epsilon\right|^{\frac{1}{2}}\left|\mathbf{D}_x\right|^{\frac{1}{2}}\left|\mathbf{M}\right|^{-\frac{1}{2}}\left|\mathbf{N}^{-1}\right|^{\frac{1}{2}}\exp\left\{-\frac{1}{2}\delta_x\Big(\mathbf{D}_x^{-1}-\sigma^2\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\mathbf{N}^{-1}\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}-\sigma^2\mathbf{M}^{-1}\Big)\delta_x^T\right\}.$$

Proof. Using the Chain Rule of probabilities we obtain the following decomposition of the likelihood function

$$\begin{aligned}
&\pi_{\mathbf{Y}_t,\mathbf{X}_t|\mathbf{U},\mathbf{V},\Psi}(\mathbf{y}_t,\mathbf{x}_t)=\pi_{\mathbf{Y}_t|\mathbf{X}_t,\mathbf{U},\mathbf{V},\Psi}(\mathbf{y}_t)\cdot\pi_{\mathbf{X}_t|\mathbf{U},\mathbf{V},\Psi}(\mathbf{x}_t)\\
&=(2\pi)^{-\frac{d+k}{2}}\left|\sigma^2\mathbf{D}_\epsilon^{-1}\right|^{-\frac{1}{2}}\left|\mathbf{D}_x^{-1}\right|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\mathbf{x}_t-\delta_x\mathbf{D}_x^{-1})\mathbf{D}_x(\mathbf{x}_t-\delta_x\mathbf{D}_x^{-1})^T\right\}\\
&\times\exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}_t-\boldsymbol{\mu}-\delta_\epsilon\mathbf{D}_\epsilon^{-1}-\mathbf{x}_t\mathbf{W}^T)\mathbf{D}_\epsilon(\mathbf{y}_t-\boldsymbol{\mu}-\delta_\epsilon\mathbf{D}_\epsilon^{-1}-\mathbf{x}_t\mathbf{W}^T)^T\right\}\\
&=(2\pi)^{-\frac{d+k}{2}}\left|\sigma^2\mathbf{D}_\epsilon^{-1}\right|^{-\frac{1}{2}}\left|\mathbf{D}_x^{-1}\right|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\Big(\mathbf{x}_t\mathbf{D}_x\mathbf{x}_t^T-2\mathbf{x}_t\delta_x^T+\delta_x\mathbf{D}_x^{-1}\delta_x^T\Big)\right\}\\
&\times\exp\left\{-\frac{1}{2}\Big(\frac{1}{\sigma^2}(\mathbf{y}_t-\boldsymbol{\mu}-\delta_\epsilon\mathbf{D}_\epsilon^{-1})\mathbf{D}_\epsilon(\mathbf{y}_t-\boldsymbol{\mu}-\delta_\epsilon\mathbf{D}_\epsilon^{-1})^T-\frac{2}{\sigma^2}\mathbf{x}_t\mathbf{W}^T\mathbf{D}_\epsilon(\mathbf{y}_t-\boldsymbol{\mu}-\delta_\epsilon\mathbf{D}_\epsilon^{-1})^T+\frac{1}{\sigma^2}\mathbf{x}_t\mathbf{W}^T\mathbf{D}_\epsilon\mathbf{W}\mathbf{x}_t^T\Big)\right\}\\
&=(2\pi)^{-\frac{d+k}{2}}\left|\sigma^2\mathbf{D}_\epsilon^{-1}\right|^{-\frac{1}{2}}\left|\mathbf{D}_x^{-1}\right|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\mathbf{x}_t-\boldsymbol{\mu}_{x,t})\boldsymbol{\Sigma}_x^{-1}(\mathbf{x}_t-\boldsymbol{\mu}_{x,t})^T\right\}\\
&\times\exp\left\{-\frac{1}{2}\Big(\delta_x\mathbf{D}_x^{-1}\delta_x^T+\frac{1}{\sigma^2}(\mathbf{y}_t-\boldsymbol{\mu}-\delta_\epsilon\mathbf{D}_\epsilon^{-1})\mathbf{D}_\epsilon(\mathbf{y}_t-\boldsymbol{\mu}-\delta_\epsilon\mathbf{D}_\epsilon^{-1})^T-\boldsymbol{\mu}_{x,t}\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\mu}_{x,t}^T\Big)\right\}
\end{aligned}\tag{A23}$$

where $\boldsymbol{\Sigma}_x=\sigma^2\mathbf{M}^{-1}$, $\boldsymbol{\mu}_{x,t}=\Big((\mathbf{y}_t-\boldsymbol{\mu}-\delta_\epsilon\mathbf{D}_\epsilon^{-1})\mathbf{D}_\epsilon\mathbf{W}+\sigma^2\delta_x\Big)\mathbf{M}^{-1}$ and $\mathbf{M}=\sigma^2\mathbf{D}_x+\mathbf{W}^T\mathbf{D}_\epsilon\mathbf{W}$. Let us denote

$$\pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U},\mathbf{V},\Psi}(\mathbf{x}_t)=(2\pi)^{-\frac{k}{2}}\left|\boldsymbol{\Sigma}_x\right|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\mathbf{x}_t-\boldsymbol{\mu}_{x,t})\boldsymbol{\Sigma}_x^{-1}(\mathbf{x}_t-\boldsymbol{\mu}_{x,t})^T\right\}$$

and remark that it is a probability function of k dimensional Gaussian random variable with the mean vector $\boldsymbol{\mu}_{x,t}$ and the covariance matrix $\boldsymbol{\Sigma}_x$. The probability function $\pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U},\mathbf{V},\Psi}(\mathbf{x}_t)$ contains all expressions with the vector $\mathbf{x}_t$ present in the probability function $\pi_{\mathbf{Y}_t,\mathbf{X}_t|\mathbf{U},\mathbf{V},\Psi}(\mathbf{y}_t,\mathbf{x}_t)$. In the next step we show how to obtain the formula for the probability function $\pi_{\mathbf{Y}_t|\mathbf{U},\mathbf{V},\Psi}(\mathbf{y}_t)$, which jointly with the function $\pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U},\mathbf{V},\Psi}(\mathbf{x}_t)$, contains all expressions with the vector $\mathbf{y}_t$. Resuming the simplification of the Equation (A23), we obtain the following formulations

$$\pi_{\mathbf{Y}_t,\mathbf{X}_t|\mathbf{U},\mathbf{V},\Psi}(\mathbf{y}_t,\mathbf{x}_t)=\pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U},\mathbf{V},\Psi}(\mathbf{x}_t)\cdot(2\pi)^{-\frac{d}{2}}\left|\sigma^2\mathbf{D}_\epsilon^{-1}\right|^{-\frac{1}{2}}\left|\mathbf{D}_x^{-1}\right|^{-\frac{1}{2}}\left|\boldsymbol{\Sigma}_x\right|^{\frac{1}{2}}$$

$$\times \exp\left\{-\frac{1}{2}\left(\delta_x \mathbf{D}_x^{-1}\delta_x^T + \frac{1}{\sigma^2}(\mathbf{y}_t-\boldsymbol{\mu}-\delta_\epsilon\mathbf{D}_\epsilon^{-1})\mathbf{D}_\epsilon(\mathbf{y}_t-\boldsymbol{\mu}-\delta_\epsilon\mathbf{D}_\epsilon^{-1})^T - \boldsymbol{\mu}_{x,t}\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\mu}_{x,t}^T\right)\right\}$$

$$= \pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U},\mathbf{V},\Psi}(\mathbf{x}_t)\cdot(2\pi)^{-\frac{d}{2}}\left|\mathbf{D}_\epsilon^{-1}\right|^{-\frac{1}{2}}\left|\mathbf{D}_x^{-1}\right|^{-\frac{1}{2}}\left|\mathbf{M}\right|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\delta_x\mathbf{D}_x^{-1}\delta_x^T - \frac{1}{2\sigma^2}(\mathbf{y}_t-\boldsymbol{\mu}-\delta_\epsilon\mathbf{D}_\epsilon^{-1})\mathbf{D}_\epsilon(\mathbf{y}_t-\boldsymbol{\mu}-\delta_\epsilon\mathbf{D}_\epsilon^{-1})^T\right\}$$

$$\times \exp\left\{\frac{1}{2\sigma^2}\left((\mathbf{y}_t-\boldsymbol{\mu}-\delta_\epsilon\mathbf{D}_\epsilon^{-1})\mathbf{D}_\epsilon\mathbf{W}+\sigma^2\delta_x\right)\mathbf{M}^{-1}\left((\mathbf{y}_t-\boldsymbol{\mu}-\delta_\epsilon\mathbf{D}_\epsilon^{-1})\mathbf{D}_\epsilon\mathbf{W}+\sigma^2\delta_x\right)^T\right\}$$

$$= \pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U},\mathbf{V},\Psi}(\mathbf{x}_t)\cdot(2\pi)^{-\frac{d}{2}}\left|\mathbf{D}_\epsilon\right|^{\frac{1}{2}}\left|\mathbf{D}_x\right|^{\frac{1}{2}}\left|\mathbf{M}\right|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\delta_x\mathbf{D}_x^{-1}\delta_x^T\right\}$$

$$\times \exp\left\{-\frac{1}{2}(\mathbf{y}_t-\boldsymbol{\mu}-\delta_\epsilon\mathbf{D}_\epsilon^{-1})\underbrace{\left(\frac{1}{\sigma^2}\mathbf{D}_\epsilon - \frac{1}{\sigma^2}\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\right)}_{\boldsymbol{\Sigma}_y^{-1}}(\mathbf{y}_t-\boldsymbol{\mu}-\delta_\epsilon\mathbf{D}_\epsilon^{-1})^T\right\}$$

$$\times \exp\left\{-\frac{1}{2}\left(-2(\mathbf{y}_t-\boldsymbol{\mu}-\delta_\epsilon\mathbf{D}_\epsilon^{-1})\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\delta_x^T - \sigma^2\delta_x\mathbf{M}^{-1}\delta_x^T\right)\right\}$$

$$= \pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U},\mathbf{V},\Psi}(\mathbf{x}_t)\cdot(2\pi)^{-\frac{d}{2}}\left|\mathbf{D}_\epsilon\right|^{\frac{1}{2}}\left|\mathbf{D}_x\right|^{\frac{1}{2}}\left|\mathbf{M}\right|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\delta_x\mathbf{D}_x^{-1}\delta_x^T\right\}$$

$$\times \exp\left\{-\frac{1}{2}(\underbrace{\mathbf{y}_t-\boldsymbol{\mu}-\delta_\epsilon\mathbf{D}_\epsilon^{-1}-\sigma^2\delta_x\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\mathbf{N}^{-1}}_{\mathbf{y}_t-\boldsymbol{\mu}_y})\boldsymbol{\Sigma}_y^{-1}(\underbrace{\mathbf{y}_t-\boldsymbol{\mu}-\delta_\epsilon\mathbf{D}_\epsilon^{-1}-\sigma^2\delta_x\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\mathbf{N}^{-1}}_{\mathbf{y}_t-\boldsymbol{\mu}_y})^T\right\}$$

$$\times \exp\left\{-\frac{1}{2}\left(-\sigma^2\delta_x\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\mathbf{N}^{-1}\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\delta_x^T - \sigma^2\delta_x\mathbf{M}^{-1}\delta_x^T\right)\right\}$$

where $\boldsymbol{\Sigma}_y = \sigma^2\mathbf{N}^{-1}$ for $\mathbf{N} = \mathbf{D}_\epsilon - \mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon$. Let us denote $\boldsymbol{\mu}_y = \boldsymbol{\mu} + \delta_\epsilon\mathbf{D}_\epsilon^{-1} + \sigma^2\delta_x\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\mathbf{N}^{-1}$ and define

$$\pi_{\mathbf{Y}_t|\mathbf{U},\mathbf{V},\Psi}(\mathbf{y}_t) = (2\pi)^{-\frac{d}{2}}\left|\boldsymbol{\Sigma}_y\right|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(\mathbf{y}_t-\boldsymbol{\mu}_{y,t})\boldsymbol{\Sigma}_y^{-1}(\mathbf{y}_t-\boldsymbol{\mu}_{y,t})^T\right\}.$$

The probability function $\pi_{\mathbf{Y}_t|\mathbf{U},\mathbf{V},\Psi}(\mathbf{y}_t)$ is a density of the d dimensional random variable which follows Gaussian distribution with the mean vector $\boldsymbol{\mu}_y$ and the covariance matrix $\boldsymbol{\Sigma}_y$. Using the definition of the density function we obtain that

$$\pi_{\mathbf{Y}_t,\mathbf{X}_t|\mathbf{U},\mathbf{V},\Psi}(\mathbf{y}_t,\mathbf{x}_t) = \pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U},\mathbf{V},\Psi}(\mathbf{x}_t)\,\pi_{\mathbf{Y}_t|\mathbf{U},\mathbf{V},\Psi}(\mathbf{y}_t)(\sigma^2)^{\frac{k}{2}}\left|\mathbf{D}_\epsilon\right|^{\frac{1}{2}}\left|\mathbf{D}_x\right|^{\frac{1}{2}}\left|\mathbf{M}\right|^{-\frac{1}{2}}\left|\mathbf{N}\right|^{-\frac{1}{2}}\exp^{-\frac{1}{2}\delta_x\left(\mathbf{D}_x^{-1}-\sigma^2\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\mathbf{N}^{-1}\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}-\sigma^2\mathbf{M}^{-1}\right)\delta_x^T}$$

Therefore, denoting

$$C(\mathbf{u},\mathbf{v};\Psi) = (\sigma^2)^{\frac{k}{2}}\left|\mathbf{D}_\epsilon\right|^{\frac{1}{2}}\left|\mathbf{D}_x\right|^{\frac{1}{2}}\left|\mathbf{M}\right|^{-\frac{1}{2}}\left|\mathbf{N}\right|^{-\frac{1}{2}}\exp^{-\frac{1}{2}\delta_x\left(\mathbf{D}_x^{-1}-\sigma^2\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\mathbf{N}^{-1}\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}-\sigma^2\mathbf{M}^{-1}\right)\delta_x^T}$$

we obtain the following factorisation of the probability function

$$\pi_{\mathbf{Y}_t,\mathbf{X}_t|\mathbf{U},\mathbf{V},\Psi}(\mathbf{y}_t,\mathbf{x}_t) = \pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U},\mathbf{V},\Psi}(\mathbf{x}_t)\cdot\pi_{\mathbf{Y}_t|\mathbf{U},\mathbf{V},\Psi}(\mathbf{y}_t)\cdot C(\mathbf{u},\mathbf{v};\Psi).$$

□

Lemma A13. *Let us recall the probability density function $\pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U},\mathbf{V},\Psi}(\mathbf{x}_t)$ of a k-dimensional Gaussian random variable $\mathbf{X}_t$ with the covariance matrix $\boldsymbol{\Sigma}_x$ and the mean vector $\boldsymbol{\mu}_{x,t}$ defined in Lemma A12. Under the assumptions of the Generalized Skew t-Student ID PPCA in Subsubsection 3.4.1.2, the solution to the following integration problem*

$$\int_{\mathbb{R}^k}\log\left(\pi_{\mathbf{Y}_t,\mathbf{X}_t|\mathbf{U},\mathbf{V},\Psi^*}(\mathbf{y}_t,\mathbf{x}_t)\right)\pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U},\mathbf{V},\Psi}(\mathbf{x}_t)\,d\mathbf{x}_t$$

is equal to the function $w:\mathbb{R}^d\times\mathbb{R}_+^d\times\mathbb{R}_+^k\longrightarrow\mathbb{R}$ defined as following

$$w(\mathbf{y}_t,\mathbf{u},\mathbf{v};\Psi,\Psi^*) = -\frac{k+d}{2}\log 2\pi - \frac{d}{2}\log\sigma^{*2} + \frac{1}{2}\sum_{i=1}^{d}\log u^i + \frac{1}{2}\sum_{j=1}^{k}\log v^j - \frac{1}{2}\delta_x^*\mathbf{D}_x^{-1}\delta_x^{*T}$$

$$
\begin{aligned}
&- \frac{1}{2\sigma^{*2}}\Big(\mathbf{y}_t - \boldsymbol{\mu}^* - \boldsymbol{\delta}_\epsilon^*\mathbf{D}_\epsilon^{-1}\Big)\mathbf{D}_\epsilon\Big(\mathbf{y}_t - \boldsymbol{\mu}^* - \boldsymbol{\delta}_\epsilon^*\mathbf{D}_\epsilon^{-1}\Big)^T - \frac{\sigma^2}{2}\,Tr\left\{\mathbf{M}_t^{-1}\Big(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_\epsilon\mathbf{W}^* + \mathbf{D}_x\Big)\right\}\\
&+ \left(\Big(\mathbf{y}_t - \boldsymbol{\mu} - \boldsymbol{\delta}_\epsilon\mathbf{D}_\epsilon^{-1}\Big)\mathbf{D}_\epsilon\mathbf{W} + \sigma^2\boldsymbol{\delta}_x\right)\mathbf{M}^{-1}\left(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_\epsilon\Big(\mathbf{y}_t - \boldsymbol{\mu}^* - \boldsymbol{\delta}_\epsilon^*\mathbf{D}_\epsilon^{-1}\Big)^T + \boldsymbol{\delta}_x^{*T}\right)\\
&- \frac{1}{2}\,Tr\left\{\mathbf{M}^{-1}\Big((\mathbf{y}_t - \boldsymbol{\mu} - \boldsymbol{\delta}_\epsilon\mathbf{D}_\epsilon^{-1})\mathbf{D}_\epsilon\mathbf{W} + \sigma^2\boldsymbol{\delta}_x\Big)^T\Big((\mathbf{y}_t - \boldsymbol{\mu} - \boldsymbol{\delta}_\epsilon\mathbf{D}_{\epsilon,t}^{-1})\mathbf{D}_\epsilon\mathbf{W} + \sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}^{-1}\Big(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_\epsilon\mathbf{W}^* + \mathbf{D}_x\Big)\right\}
\end{aligned}
$$

where $\mathbf{M} = \sigma^2\mathbf{D}_x + \mathbf{W}^T\mathbf{D}_\epsilon\mathbf{W}$ *and* $\boldsymbol{\mu}_{x,t}$ *being specified in Lemma A12.*

Proof. The probability function $\pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U},\mathbf{V},\Psi}(\mathbf{x}_t)$ is defined in Lemma A12 as a density of a k-dimensional Gaussian random variable $\mathbf{X}_t$ with the covariance matrix $\boldsymbol{\Sigma}_x = \sigma^2\mathbf{M}^{-1}$ and the mean vector $\boldsymbol{\mu}_{x,t} = \Big((\mathbf{y}_t - \boldsymbol{\mu} - \boldsymbol{\delta}_\epsilon\mathbf{D}_\epsilon^{-1})\mathbf{D}_\epsilon\mathbf{W} + \sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}^{-1}$, where $\mathbf{M} = \sigma^2\mathbf{D}_x + \mathbf{W}^T\mathbf{D}_\epsilon\mathbf{W}$. Hence, the solution to the integration problem is given by

$$
\begin{aligned}
w(\mathbf{y}_t,\mathbf{u},\mathbf{v};\Psi,\Psi^*) &= \int_{\mathbb{R}^k}\log\Big(\pi_{\mathbf{Y}_t,\mathbf{X}_t|\mathbf{U},\mathbf{V},\Psi^*}(\mathbf{y}_t,\mathbf{x}_t,\mathbf{u},\mathbf{v})\Big)\pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U},\mathbf{V},\Psi}(\mathbf{x}_t)\,d\mathbf{x}_t\\
&= \int_{\mathbb{R}^k}\Big(\log\pi_{\mathbf{Y}_t|\mathbf{X}_t,\mathbf{U},\mathbf{V}|\Psi^*}(\mathbf{y}_t) + \log\pi_{\mathbf{X}_t|\mathbf{U},\mathbf{V},\Psi^*}(\mathbf{x})\Big)\pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U},\mathbf{V},\Psi}(\mathbf{x}_t)\,d\mathbf{x}_t\\
&= -\frac{k+d}{2}\log 2\pi + \log\Big|\sigma^{*2}\mathbf{D}_\epsilon^{-1}\Big|^{-\frac{1}{2}} + \log\Big|\mathbf{D}_x^{-1}\Big|^{-\frac{1}{2}} - \frac{1}{2}\int_{\mathbb{R}^k}\Big(\mathbf{x}_t - \boldsymbol{\delta}_x^*\mathbf{D}_x^{-1}\Big)\mathbf{D}_x\Big(\mathbf{x}_t - \boldsymbol{\delta}_x^*\mathbf{D}_x^{-1}\Big)^T\pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U},\mathbf{V},\Psi}(\mathbf{x}_t)\,d\mathbf{x}_t\\
&- \frac{1}{2}\int_{\mathbb{R}^k}\frac{1}{\sigma^{*2}}\Big(\mathbf{y}_t - \boldsymbol{\mu}^* - \boldsymbol{\delta}_\epsilon^*\mathbf{D}_\epsilon^{-1} - \mathbf{x}_t\mathbf{W}^{*T}\Big)\mathbf{D}_\epsilon\Big(\mathbf{y}_t - \boldsymbol{\mu}^* - \boldsymbol{\delta}_\epsilon^*\mathbf{D}_\epsilon^{-1} - \mathbf{x}_t\mathbf{W}^{*T}\Big)^T\pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U},\mathbf{V},\Psi}(\mathbf{x}_t)\,d\mathbf{x}_t\\
&= -\frac{k+d}{2}\log 2\pi - \frac{d}{2}\log\sigma^{*2} + \frac{1}{2}\sum_{i=1}^{d}\log u^i + \frac{1}{2}\sum_{j=1}^{k}\log v^j - \frac{1}{2\sigma^{*2}}\Big(\mathbf{y}_t - \boldsymbol{\mu}^* - \boldsymbol{\delta}_\epsilon^*\mathbf{D}_\epsilon^{-1}\Big)\mathbf{D}_\epsilon\Big(\mathbf{y}_t - \boldsymbol{\mu}^* - \boldsymbol{\delta}_\epsilon^*\mathbf{D}_\epsilon^{-1}\Big)^T\\
&- \frac{1}{2}\boldsymbol{\delta}_x^*\mathbf{D}_x^{-1}\boldsymbol{\delta}_x^{*T} + \frac{1}{\sigma^{*2}}\int_{\mathbb{R}^k}\mathbf{x}_t\pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U},\mathbf{V},\Psi}(\mathbf{x}_t)\,d\mathbf{x}_t\,\mathbf{W}^{*T}\mathbf{D}_\epsilon\Big(\mathbf{y}_t - \boldsymbol{\mu}^* - \boldsymbol{\delta}_\epsilon^*\mathbf{D}_\epsilon^{-1}\Big)^T + \int_{\mathbb{R}^k}\mathbf{x}_t\pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U},\mathbf{V},\Psi}(\mathbf{x}_t)\,d\mathbf{x}_t\boldsymbol{\delta}_x^{*T}\\
&- \frac{1}{2}\int_{\mathbb{R}^k}\mathbf{x}_t\Big(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_\epsilon\mathbf{W}^* + \mathbf{D}_x\Big)\mathbf{x}_t^T\pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U},\mathbf{V},\Psi}(\mathbf{x}_t)\,d\mathbf{x}_t\\
&= -\frac{k+d}{2}\log 2\pi - \frac{1}{2}\log\sigma^{*2} + \frac{1}{2}\sum_{i=1}^{d}\log u_t^i + \frac{1}{2}\sum_{j=1}^{k}\log v_t^j - \frac{1}{2}\boldsymbol{\delta}_x^*\mathbf{D}_x^{-1}\boldsymbol{\delta}_x^{*T}\\
&- \frac{1}{2\sigma^{*2}}\Big(\mathbf{y}_t - \boldsymbol{\mu}^* - \boldsymbol{\delta}_\epsilon^*\mathbf{D}_\epsilon^{-1}\Big)\mathbf{D}_\epsilon\Big(\mathbf{y}_t - \boldsymbol{\mu}^* - \boldsymbol{\delta}_\epsilon^*\mathbf{D}_\epsilon^{-1}\Big)^T + \boldsymbol{\mu}_{x,t}\left(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\Big(\mathbf{y}_t - \boldsymbol{\mu}^* - \boldsymbol{\delta}_\epsilon^*\mathbf{D}_\epsilon^{-1}\Big)^T + \boldsymbol{\delta}_x^{*T}\right)\\
&- \frac{1}{2}\,\mathrm{Tr}\left\{\Big(\boldsymbol{\Sigma}_x + \boldsymbol{\mu}_{x,t}^T\boldsymbol{\mu}_{x,t}\Big)\Big(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_\epsilon\mathbf{W}^* + \mathbf{D}_x\Big)\right\}\\
&= -\frac{k+d}{2}\log 2\pi - \frac{d}{2}\log\sigma^{*2} + \frac{1}{2}\sum_{i=1}^{d}\log u_t^i + \frac{1}{2}\sum_{j=1}^{k}\log v_t^j - \frac{1}{2}\boldsymbol{\delta}_x^*\mathbf{D}_x^{-1}\boldsymbol{\delta}_x^{*T}\\
&- \frac{1}{2\sigma^{*2}}\Big(\mathbf{y}_t - \boldsymbol{\mu}^* - \boldsymbol{\delta}_\epsilon^*\mathbf{D}_\epsilon^{-1}\Big)\mathbf{D}_\epsilon\Big(\mathbf{y}_t - \boldsymbol{\mu}^* - \boldsymbol{\delta}_\epsilon^*\mathbf{D}_\epsilon^{-1}\Big)^T - \frac{1}{2}\,\mathrm{Tr}\left\{\sigma^2\mathbf{M}_t^{-1}\Big(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_\epsilon\mathbf{W}^* + \mathbf{D}_x\Big)\right\}\\
&+ \left(\Big(\mathbf{y}_t - \boldsymbol{\mu} - \boldsymbol{\delta}_\epsilon\mathbf{D}_\epsilon^{-1}\Big)\mathbf{D}_\epsilon\mathbf{W} + \sigma^2\boldsymbol{\delta}_x\right)\mathbf{M}^{-1}\left(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_\epsilon\Big(\mathbf{y}_t - \boldsymbol{\mu}^* - \boldsymbol{\delta}_\epsilon^*\mathbf{D}_\epsilon^{-1}\Big)^T + \boldsymbol{\delta}_x^{*T}\right)\\
&- \frac{1}{2}\,\mathrm{Tr}\left\{\mathbf{M}^{-1}\Big((\mathbf{y}_t - \boldsymbol{\mu} - \boldsymbol{\delta}_\epsilon\mathbf{D}_\epsilon^{-1})\mathbf{D}_\epsilon\mathbf{W} + \sigma^2\boldsymbol{\delta}_x\Big)^T\Big((\mathbf{y}_t - \boldsymbol{\mu} - \boldsymbol{\delta}_\epsilon\mathbf{D}_\epsilon^{-1})\mathbf{D}_\epsilon\mathbf{W} + \sigma^2\boldsymbol{\delta}_x\Big)\mathbf{M}^{-1}\Big(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_\epsilon\mathbf{W}^* + \mathbf{D}_x\Big)\right\}
\end{aligned}
$$

for the non-central second moment of the $\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U},\mathbf{V},\Psi$ given by

$$
\mathbb{E}_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U},\mathbf{V},\Psi^*}\left[\mathbf{X}_t^T\mathbf{X}_t\right] = \int_{\mathbb{R}^k}\mathbf{x}_t^T\mathbf{x}_t\pi_{\mathbf{X}_t|\mathbf{Y}_t,\mathbf{U},\mathbf{V},\Psi}(\mathbf{x}_t)\,d\mathbf{x}_t = \boldsymbol{\Sigma}_x + \boldsymbol{\mu}_{x,t}^T\boldsymbol{\mu}_{x,t}.
$$

□

Lemma A14. *Let us consider the partition of the observation vector* $\mathbf{Y}_t = [\mathbf{Y}_t^o, \mathbf{Y}_t^m]$ *into the the subvector with observed and missing entries,* $\mathbf{Y}_t^o$ *and* $\mathbf{Y}_t^m$*, respectively. The observation vector* $\mathbf{Y}_t$ *is modelled as in Equation (2.7) under the assumptions stated in Subsubsubsection 3.4.1.2 of Generalized Skew t-Student ID PPCA. The conditional distribution of the random vector* $\mathbf{Y}_t|\mathbf{U},\mathbf{V},\Psi$*, derived in Lemma A12, is Gaussian with the mean vector* $\boldsymbol{\mu}_y = \boldsymbol{\mu} + \boldsymbol{\delta}_\epsilon\mathbf{D}_\epsilon^{-1} + \sigma^2\boldsymbol{\delta}_x\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\mathbf{N}^{-1}$ *and the covariance matrix* $\boldsymbol{\Sigma}_y =$

$\sigma^2\mathbf{N}^{-1}$, *where* $\mathbf{M} = \sigma^2\mathbf{D}_x + \mathbf{W}^T\mathbf{D}_\epsilon\mathbf{W}$ *and* $\mathbf{N} = \mathbf{D}_\epsilon - \mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon$. *The conditional distribution* $\mathbf{Y}_t^m|\mathbf{Y}_t^o, \mathbf{U}, \mathbf{V}, \Psi$ *is also Gaussian with the mean vector* $\tilde{\boldsymbol{\mu}}_{Y^m,t}$ *and the covariance matrix* $\tilde{\boldsymbol{\Sigma}}_{Y^m}$ *such that*

$$\begin{aligned}\tilde{\boldsymbol{\mu}}_{Y^m,t} &= \boldsymbol{\mu}_y^m + \left(\mathbf{Y}_t^o - \boldsymbol{\mu}_y^o\right)\boldsymbol{\Sigma}_y^{oo\ -1}\boldsymbol{\Sigma}_y^{om}\\ \tilde{\boldsymbol{\Sigma}}_{Y^m} &= \boldsymbol{\Sigma}_y^{mm} - \boldsymbol{\Sigma}_y^{mo}\boldsymbol{\Sigma}_y^{oo\ -1}\boldsymbol{\Sigma}_y^{om}.\end{aligned}$$

In addition, the marginal distribution of $\mathbf{Y}_t^o|\mathbf{U}, \mathbf{V}, \Psi$ *is Gaussian with the mean vector* $\boldsymbol{\mu}_y^o$ *and the covariance matrix* $\boldsymbol{\Sigma}_y^{oo}$. *The subvectors* $\boldsymbol{\mu}_y^o$ *and* $\boldsymbol{\mu}_y^m$ *contain elements of the vector* $\boldsymbol{\mu}_y$ *which correspond to the observed or missing entries of the observation vector* $\mathbf{Y}_t$, *respectively, and are* d_o *and* d_m*-dimensional. The square matrices* $\boldsymbol{\Sigma}_y^{oo}$ *and* $\boldsymbol{\Sigma}_y^{mm}$ *contain elements of the matrix* $\boldsymbol{\Sigma}_y$ *which correspond by rows and columns to the observed or missing entries of the observation vector* $\mathbf{Y}_t$, *respectively, and are* $d_o \times d_o$ *and* $d_m \times d_m$*-dimensional. The non-square matrix* $\boldsymbol{\Sigma}_y^{mo} = \boldsymbol{\Sigma}_y^{om\ T}$ *contains elements of the matrix* $\boldsymbol{\Sigma}_y$ *which correspond by rows to the missing and by columns to the observed entries of the observation vector* $\mathbf{Y}_t$, *and is* $d_m \times d_o$*-dimensional.*

Proof. Please refer to Lemma A12 for the derivation of the conditional distribution $\mathbf{Y}_t|\mathbf{U}, \mathbf{V}, \Psi$ and apply the formulas for the standard conditional Gaussian distribution from Lemma A1. □

Lemma A15. *Let us consider the partition of the observation vector* $\mathbf{Y}_t = [\mathbf{Y}_t^o, \mathbf{Y}_t^m]$ *into the the subvector with observed and missing entries,* $\mathbf{Y}_t^o$ *and* $\mathbf{Y}_t^m$, *respectively. The observation vector* $\mathbf{Y}_t$ *is modelled as in Equation (2.7) under the assumptions of Generalized Skew t-Student ID PPCA stated in Subsubsubsection 3.4.1.2. Let us recall the function* $w : \mathbb{R}^d \times \mathbb{R}_+^d \times \mathbb{R}_+^k \longrightarrow \mathbb{R}$ *defined in Lemma A13 and the conditional distribution of the random vector* $\mathbf{Y}_t^m|\mathbf{Y}_t^o, \mathbf{U}, \mathbf{V}, \Psi$ *specified in Lemma A14. Given the N realisations of the vector* $\mathbf{Y}_t$, $\mathbf{y}_{1:N}^o = [\mathbf{y}_1^o, \ldots, \mathbf{y}_N^o]$, *the function* $v : \mathbb{R}^{d_o} \times \mathbb{R}_+^d \times \mathbb{R}_+^k \longrightarrow \mathbb{R}$ *defined by the following integration problems*

$$v(\mathbf{y}_t^o, \mathbf{u}, \mathbf{v}; \Psi, \Psi^*) = \int_{\mathbb{R}^{d_m}} w(\mathbf{y}_t, \mathbf{u}, \mathbf{v}; \Psi, \Psi^*) \cdot \pi_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U},\mathbf{V},\Psi}(\mathbf{y}_t^m)\, d\mathbf{y}_t^m \tag{A24}$$

equals to

$$\begin{aligned}
v(\mathbf{y}_t^o, \mathbf{u}, \mathbf{v}; \Psi, \Psi^*) = &-\frac{k+d}{2}\log 2\pi - \frac{d}{2}\log\sigma^{*2} + \frac{1}{2}\sum_{i=1}^{d}\log u^i + \frac{1}{2}\sum_{j=1}^{k}\log v^j\\
&- \frac{1}{2}\boldsymbol{\delta}_x^*\mathbf{D}_x^{-1}\boldsymbol{\delta}_x^{*T} - \frac{1}{2\sigma^{*2}}\,Tr\left\{\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U},\mathbf{V},\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right]\mathbf{D}_\epsilon\right\} + \frac{1}{\sigma^{*2}}\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U},\mathbf{V},\Psi}\left[\mathbf{Y}_t\right]\left(\mathbf{D}_\epsilon\boldsymbol{\mu}^{*T} + \boldsymbol{\delta}_\epsilon^{*T}\right)\\
&- \frac{1}{2\sigma^{*2}}\boldsymbol{\mu}^*\mathbf{D}_\epsilon\boldsymbol{\mu}^{*T} - \frac{1}{\sigma^{*2}}\boldsymbol{\mu}^*\boldsymbol{\delta}_\epsilon^{*T} - \frac{1}{2\sigma^{*2}}\boldsymbol{\delta}_\epsilon^*\mathbf{D}_\epsilon^{-1}\boldsymbol{\delta}_\epsilon^{*T} + \frac{1}{\sigma^{*2}}\,Tr\left\{\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U},\mathbf{V},\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right]\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon\right\}\\
&+ \mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U},\mathbf{V},\Psi}\left[\mathbf{Y}_t\right]\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\left(-\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_\epsilon\boldsymbol{\mu}^{*T} - \frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\boldsymbol{\delta}_\epsilon^{*T} + \boldsymbol{\delta}_x^{*T}\right)\\
&+ \frac{1}{\sigma^{*2}}\left(-\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W} - \boldsymbol{\delta}_\epsilon\mathbf{W} + \sigma^2\boldsymbol{\delta}_x\right)\mathbf{M}^{-1}\mathbf{W}^{*T}\mathbf{D}_\epsilon\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U},\mathbf{V},\Psi}\left[\mathbf{Y}_t\right]^T\\
&+ \left(-\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W} - \boldsymbol{\delta}_\epsilon\mathbf{W} + \sigma^2\boldsymbol{\delta}_x\right)\mathbf{M}^{-1}\left(-\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_\epsilon\boldsymbol{\mu}^{*T} - \frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\boldsymbol{\delta}_\epsilon^{*T} + \boldsymbol{\delta}_x^{*T}\right)\\
&- \frac{1}{2}\,Tr\left\{\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U},\mathbf{V},\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right]\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\left(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_\epsilon\mathbf{W}^* + \mathbf{D}_x\right)\right\}\\
&+ \frac{1}{2}\,Tr\left\{\mathbf{M}^{-1}\mathbf{W}^T\mathbf{D}_\epsilon\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U},\mathbf{V},\Psi}\left[\mathbf{Y}_t\right]^T\left(\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W} + \boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\right)\mathbf{M}^{-1}\left(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_\epsilon\mathbf{W}^* + \mathbf{D}_x\right)\right\}\\
&+ \frac{1}{2}\,Tr\left\{\mathbf{M}^{-1}\left(\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W} + \boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\right)\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U},\mathbf{V},\Psi}\left[\mathbf{Y}_t\right]\mathbf{D}_\epsilon\mathbf{W}\mathbf{M}^{-1}\left(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_\epsilon\mathbf{W}^* + \mathbf{D}_x\right)\right\}\\
&- \frac{1}{2}\,Tr\left\{\mathbf{M}_t^{-1}\left(\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W} + \boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\right)^T\left(\boldsymbol{\mu}\mathbf{D}_\epsilon\mathbf{W} + \boldsymbol{\delta}_\epsilon\mathbf{W} - \sigma^2\boldsymbol{\delta}_x\right)\mathbf{M}^{-1}\left(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_\epsilon\mathbf{W}^* + \mathbf{D}_x\right)\right\}
\end{aligned}$$

$$-\frac{\sigma^2}{2}\,Tr\left\{\mathbf{M}^{-1}\Big(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_\epsilon\mathbf{W}^*+\mathbf{D}_x\Big)\right\}.$$

Using the mean $\tilde{\boldsymbol{\mu}}_{Y^m,t}$ *and the covariance matrix* $\tilde{\boldsymbol{\Sigma}}_{Y^m}$ *of the random vector* $\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U},\mathbf{V},\Psi$ *from Lemma A14, the first and second non-central moments of the vector* $\mathbf{Y}_t$ *are the following*

$$\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U},\mathbf{V},\Psi}\left[\mathbf{Y}_t\right]=\left[\mathbf{y}_t^o,\ \tilde{\boldsymbol{\mu}}_{Y^m,t}\right],$$
$$\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U},\mathbf{V},\Psi}\left[\mathbf{Y}_t^T\mathbf{Y}_t\right]=\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \tilde{\boldsymbol{\Sigma}}_{Y^m}\end{bmatrix}+\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U},\mathbf{V},\Psi}\left[\mathbf{Y}_t\right]^T\mathbb{E}_{\mathbf{Y}_t^m|\mathbf{Y}_t^o,\mathbf{U},\mathbf{V},\Psi}\left[\mathbf{Y}_t\right],$$

Proof. The proof of Lemma A15 is analogous to the proof of Lemma A11 in Appendix A.4.5. □

A.4.7. Proof of Proposition 1

Lemma A16. *The quantile distribution function of the chi-square variable with* ν *degrees of freedom,* $\chi_\nu^{-1}(s):[0,1]\to[0,\infty)$*, belongs to the differentiable class* $\mathbb{C}^{[\frac{\nu}{2}]}$ *functions as has derivatives of all orders not greater than* $[\frac{\nu}{2}]$ *given by the following differential relation*

$$\frac{\partial^m\chi_\nu^{-1}(s)}{\partial s^m}=2^{-m}\sum_{j=1}^{m}\binom{m}{j}(-1)^{m+j}\partial^m\chi_{\nu-2j}^{-1},$$

where the operator $[c]$ *for any number* c *returns the closest integer number to* c*, which is not greater than* s*. Hence, the quantile distribution function* $\chi_\nu^{-1}(s)$ *is continuous on the closed interval* $[0,1]$ *and has continuous derivatives up to* $[\frac{\nu}{2}]$ *order.*

Proof. Please refer to (Abramowitz and Stegun 1974) and recall that the degrees of freedom belongs to the class of positive numbers and consequently $\nu-2m$ is greater than 0. □

Lemma A17. *The function* $m:\mathbb{R}^d\times[0,1]^2\longrightarrow\mathbb{R}$ *defined in Theorem 13 is of class* $\mathbb{C}^{r_\epsilon,r_x}\left([0,1]^2\right)$ *with respect to* $s_{\epsilon,t}$ *and* $s_{x,t}$ *for* $r_\epsilon=\min\left[\frac{\nu_\epsilon}{2}\right]$ *and* $r_x=\min\left[\frac{\nu_x}{2}\right]$.

Proof. We show that the function $m:\mathbb{R}^d\times[0,1]^2\to\mathbb{R}$

$$m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi):=\pi_{\mathbf{Y}|S_{\epsilon,t},S_{x,t},\Psi}(\mathbf{y}_t)(\sigma^2)^{\frac{k}{2}}\left|\mathbf{D}_{\epsilon,t}\right|^{\frac{1}{2}}\left|\mathbf{D}_{x,t}\right|^{\frac{1}{2}}\left|\mathbf{M}_t\right|^{-\frac{1}{2}}\left|\mathbf{N}_t\right|^{-\frac{1}{2}}e^{-\frac{1}{2}\delta_x\left(\mathbf{D}_{x,t}^{-1}-\sigma^2\mathbf{M}_t^{-1}\mathbf{W}^T\mathbf{D}_{\epsilon,t}\mathbf{N}_t^{-1}\mathbf{D}_{\epsilon,t}\mathbf{W}\mathbf{M}_t^{-1}-\sigma^2\mathbf{M}_t^{-1}\right)\delta_x^T}.$$

has continuous partial derivatives

$$\frac{\partial^{i+j}m(\mathbf{y}_t,s_{\epsilon,t},s_{x,t};\Psi)}{\partial s_{\epsilon,t}^i\,\partial s_{x,t}^j},\tag{A25}$$

for $0\le i\le r_\epsilon$ and $0\le j\le r_x$.

Let $\tilde{m}(\mathbf{y}_t,s_\epsilon,s_x):=m(s_\epsilon,s_x)$. The function m can be formulated as a composite of polynomial p and the function g such that $\tilde{m}(s_\epsilon,s_x)=(p\circ g)(s_\epsilon,s_x)$ for $i=1,\dots,d+k$ where

- $g:[0,1]^2\to\mathbb{R}_+^d\times\mathbb{R}_+^k$, such that $g(s_\epsilon,s_x):=\left(T_\epsilon(s_\epsilon),T_x(s_x)\right)=\left(g_1(s_\epsilon),\dots,g_d(s_\epsilon),g_{d+1}(s_x)\dots,g_{d+k}(s_x)\right)$ for $g_i:[0,1]\to\mathbb{R}_+$ being the following

$$\left(g_1(s),\dots,g_d(s)\right)=\left(\frac{\chi_{\nu_\epsilon^1}^{-1}(s)}{\nu_\epsilon^1},\dots,\frac{\chi_{\nu_\epsilon^d}^{-1}(s)}{\nu_\epsilon^d}\right)\text{ and }\left(g_{d+1}(s),\dots,g_{d+k}(s)\right)=\left(\frac{\chi_{\nu_x^1}^{-1}(s)}{\nu_x^1},\dots,\frac{\chi_{\nu_x^k}^{-1}(s)}{\nu_x^k}\right)$$

given the function $T_\epsilon:[0,1]\to\mathbb{R}_+^d$ and $T_x:[0,1]\to\mathbb{R}_+^k$ defined in Theorem 13.

- $p : \mathbb{R}_+^d \times \mathbb{R}_+^k \to \mathbb{R}$, such that $p(\mathbf{t}_\epsilon, \mathbf{t}_x) = \prod_{k=1}^5 p_k(\mathbf{t}_\epsilon, \mathbf{t}_x)$ for $(\mathbf{t}_\epsilon, \mathbf{t}_x) \in \mathbb{R}_+^d \times \mathbb{R}_+^k$ and $p_k : \mathbb{R}_+^d \times \mathbb{R}_+^k \to \mathbb{R}$ being the following

$$p_1(\mathbf{t}_\epsilon, \mathbf{t}_x) = \Big(\prod_{i=1}^{d} t_{\epsilon,i} \prod_{j=1}^{k} t_{x,j} \Big)^{\frac{1}{2}},\ p_2(\mathbf{t}_\epsilon, \mathbf{t}_x) = \det\Big(M(\mathbf{t}_\epsilon, \mathbf{t}_x)\Big)^{-\frac{1}{2}},\ p_3(\mathbf{t}_\epsilon, \mathbf{t}_x) = \det\Big(N(\mathbf{t}_\epsilon, \mathbf{t}_x)\Big)^{-\frac{1}{2}}$$

$$p_4(\mathbf{t}_\epsilon, \mathbf{t}_x) = \exp\Big\{ -\frac{1}{2}\delta_x \Big(\operatorname{diag}(\mathbf{t}_x)^{-1} - \sigma^2 \mathbf{M}(\mathbf{t}_\epsilon, \mathbf{t}_x)^{-1} \mathbf{W}^T \operatorname{diag}(\mathbf{t}_\epsilon) \mathbf{N}(\mathbf{t}_\epsilon, \mathbf{t}_x)^{-1} \operatorname{diag}(\mathbf{t}_\epsilon) \mathbf{W} \mathbf{M}(\mathbf{t}_\epsilon, \mathbf{t}_x)^{-1} - \sigma^2 \mathbf{M}(\mathbf{t}_\epsilon, \mathbf{t}_x)^{-1} \Big) \delta_x^T \Big\}$$

$$p_5(\mathbf{t}_\epsilon, \mathbf{t}_x) = \Big(2\pi\sigma^2\Big)^{-\frac{d}{2}} \det\Big(\mathbf{N}(\mathbf{t}_\epsilon, \mathbf{t}_x)\Big)^{\frac{1}{2}} \exp\Big\{ -\frac{1}{2\sigma^2} \big(\mathbf{y} - u(\mathbf{t}_\epsilon, \mathbf{t}_x)\big) \mathbf{N}(\mathbf{t}_\epsilon, \mathbf{t}_x)^{-1} \big(\mathbf{y} - u(\mathbf{t}_\epsilon, \mathbf{t}_x)\big)^T \Big\}$$

where

$$M(\mathbf{t}_\epsilon, \mathbf{t}_x)_{d\times d} = \sigma^2 \operatorname{diag}(\mathbf{t}_x) + \mathbf{W}^T \operatorname{diag}(\mathbf{t}_\epsilon) \mathbf{W},$$
$$N(\mathbf{t}_\epsilon, \mathbf{t}_x)_{k\times k} = \operatorname{diag}(\mathbf{t}_\epsilon) - \operatorname{diag}(\mathbf{t}_\epsilon) \mathbf{W} M(\mathbf{t}_\epsilon, \mathbf{t}_x) \mathbf{W}^T \operatorname{diag}(\mathbf{t}_\epsilon),$$
$$u(\mathbf{t}_\epsilon, \mathbf{t}_x) = \boldsymbol{\mu} + \delta_\epsilon \operatorname{diag}(\mathbf{t}_\epsilon)^{-1} + \sigma^2 \delta_x \mathbf{M}(\mathbf{t}_\epsilon, \mathbf{t}_x)^{-1} \mathbf{W}^T \operatorname{diag}(\mathbf{t}_\epsilon) \mathbf{N}(\mathbf{t}_\epsilon, \mathbf{t}_x)^{-1}.$$

Given the above representation of $\tilde{m}$, the differentiability class of the function can be determined by specifying the minimum differentiability class of g and p. It can be seen by applying the chain rule formula for multivariate partial derivatives from (Faa di Bruno 1857) to Equation (A25) which results in

$$\frac{\partial^{i+j}}{\partial s_{\epsilon,t}^i \, \partial s_{x,t}^j} \tilde{m}(s_{\epsilon,t}, s_{x,t}; \Psi) = \frac{\partial}{\partial \mathbf{t}_\epsilon \partial \mathbf{t}_x} p(\mathbf{t}_\epsilon, \mathbf{t}_x) \, \frac{\partial^{i+j}}{\partial s_{\epsilon,t}^i \, \partial s_{x,t}^j} g(s_{\epsilon,t}, s_{x,t})$$

We start by specifying the class of the functions T_ϵ and T_x defined as

$$T_\epsilon(s) := \Big(\frac{\chi_{\nu_\epsilon^1}^{-1}(s)}{\nu_\epsilon^1}, \ldots, \frac{\chi_{\nu_\epsilon^d}^{-1}(s)}{\nu_\epsilon^d} \Big)_{1\times d} \text{ and } T_x(s) := \Big(\frac{\chi_{\nu_x^1}^{-1}(s)}{\nu_x^1}, \ldots, \frac{\chi_{\nu_x^k}^{-1}(s)}{\nu_x^k} \Big)_{1\times k}$$

Using Lemma A16, it is straightforward to show that the functions $T_\epsilon(s)$ and $T_x(s)$ have continuous derivatives up to the order $\min_i[\frac{\nu_\epsilon^i}{2}]$ for $T_\epsilon(s)$ and $\min_j[\frac{\nu_x^j}{2}]$ for $T_x(s)$. It stems from the fact that the differentiability class of the multidimensional functions equals to the minimum differentiability class of its marginals. Consequently, the function g belongs to the class $\mathbb{C}^{r_\epsilon, r_x}([0,1]^2)$ for $r_\epsilon = \min\big[\frac{\nu_\epsilon}{2}\big]$ and $r_x = \min\big[\frac{\nu_x}{2}\big]$ with respect to s_ϵ and s_x, respectively.

Secondly, we show that the functions p is infinitely differentiable under the model assumptions from Subsubsection 3.4.1.1. Since the function p_1 is polynomial, it is infinitely differentiable. The function p_k for $k = 2, \ldots, 5$ are rational polynomials and, consequently, are infinitely differentiable except for regions where $\det\Big(M(\mathbf{t}_\epsilon, \mathbf{t}_x)\Big) = 0$ or $\det\Big(N(\mathbf{t}_\epsilon, \mathbf{t}_x)\Big) = 0$. However, given the distribution assumptions of the GStS IND PPCA model from Subsubsection 3.4.1.1, the determinants of $M(\mathbf{t}_\epsilon, \mathbf{t}_x)\Big)$ and $N(\mathbf{t}_\epsilon, \mathbf{t}_x)$ are always greater than zero. It is due to the fact the matrices represent covariances of random vectors as shown in proven in Lemma A8 and hence, are positive-definite from definition. It implicates that the function $p = \prod_{k=1}^5 p_k$ is smooth.

the above reasoning shows that function $\tilde{m}$ belongs to the class $\mathbb{C}^{r_\epsilon, r_x}\big([0,1]^2\big)$ for $r_\epsilon = \min\big[\frac{\nu_\epsilon}{2}\big]$ and $r_x = \min\big[\frac{\nu_x}{2}\big]$. □

Lemma A18. *The solutions of the integration problem*

$$\int_0^1 \int_0^1 m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi) \, ds_{\epsilon,t} \, ds_{x,t}$$

can be approximated by the 2-dimensional Gaussian-Legendre product quadrature rule on $[0,1]^2$ such that

$$\int_0^1 \int_0^1 m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi) \, ds_{\epsilon,t} \, ds_{x,t} = \frac{1}{4} \sum_{i=0}^{n} \sum_{j=0}^{n'} \phi_i \phi_j \, m\Big(\mathbf{y}_t, \frac{s_{n,i}+1}{2}, \frac{s_{n',j}+1}{2}\Big) + E_{n,n'}\{m\}$$

where ϕ_i are defined as in Equation (3.49), the distinct points $s_{n,i} \in [-1,1]$ for $i = 0, \ldots, n$ and $s_{n',j} \in [-1,1]$ for $j = 0, \ldots, n'$ are the zero roots of the Legendre polynomials P_n and $P_{n'}$, respectively, defined as

in Equation (3.48). If $n = [\frac{r_\epsilon}{2}]$ *and* $n' = [\frac{r_x}{2}]$ *for* $r_\epsilon = \min_i [\frac{\nu_\epsilon^i}{2}]$ *and* $r_x = \min_j [\frac{\nu_x^j}{2}]$*, the quadrature error term* $E_{n,n'}\{m\}$ *is given by the following explicit formula*

$$E_{n,n'}\{m\}(\mathbf{y}_t) = \frac{1}{4}\sum_{j=0}^{n'} \phi_j E_n\Big(\mathbf{y}_t, \frac{s_{n',i}-1}{2}\Big) + \frac{1}{4}\sum_{i=0}^{n} \phi_i E_{n'}\Big(\mathbf{y}_t, \frac{s_{n,i}-1}{2}\Big),$$
$$+ \frac{1}{(2n+1)!(2n'+1)!}\Big[\frac{2^{n'}(n'!)^2}{(2n')!}\,\frac{2^n(n!)^2}{(2n)!}\Big]^2 \frac{\partial^{2n+2n'}m}{\partial s_\epsilon^{2n}\partial s_x^{2n'}}m\Big(\mathbf{y}_t, \frac{\zeta+1}{2}, \frac{\varsigma+1}{2}; \Psi\Big),$$

for $-1 < \zeta < 1$ *and* $-1 < \varsigma < 1$ *where*

$$E_n(\mathbf{y}_t, s_x) = \frac{2}{(2n+1)!}\Big[\frac{2^n(n!)^2}{(2n)!}\Big]^2 \frac{\partial^{2n}}{\partial s_\epsilon^{2n}}m\Big(\mathbf{y}_t, \frac{\zeta+1}{2}, s_x; \Psi\Big),$$
$$E_{n'}(\mathbf{y}_t, s_\epsilon) = \frac{2}{(2n'+1)!}\Big[\frac{2^{n'}(n'!)^2}{(2n')!}\Big]^2 \frac{\partial^{2n'}}{\partial s_x^{2n'}}m\Big(\mathbf{y}_t, s_\epsilon, \frac{\varsigma+1}{2}; \Psi\Big).$$

Proof. Let us recall that the function $m : \mathbb{R}^d \times [0,1]^2 \longrightarrow \mathbb{R}$ defined in Theorem 13 belongs to the class $\mathbb{C}^{r_\epsilon, r_x}([0,1]^2)$ with respect to $s_{\epsilon,t}$ and $s_{x,t}$ for r_ϵ and r_x, respectively, as proven in Lemma A17. For the simplicity of the notation let us denote $\tilde{m}(s_\epsilon, s_x) := m(\mathbf{y}_t, s_\epsilon, s_x; \Psi)$. Given the Gaussian-Legendre formula for univariate function and n being a number of the distinct points of the quadrature rule,

$$\int_0^1 \tilde{m}(s_\epsilon, s_x)\, ds_\epsilon = \frac{1}{2}\sum_{i=0}^{n} \phi_i \tilde{m}\Big(\frac{s_{n,i}+1}{2}, s_x\Big) + \frac{1}{2}E_n(s_x)$$

for $E_n(s_x)$ being an error term of the quadrature rule. If the parameter n is chosen that $n = [\frac{r_\epsilon}{2}]$, we can apply the formula from Equation (3.50) and the error tern E_n can be specified as follows

$$E_n(s_x) = \frac{2}{(2n+1)!}\Big[\frac{2^n(n!)^2}{(2n)!}\Big]^2 \frac{\partial^{2n}}{\partial s_\epsilon^{2n}}\tilde{m}\Big(\frac{\zeta+1}{2}, s_x\Big) \text{ for } -1 < \zeta < 1.$$

Applying the product rule and Gaussian-Legendre quadrature to the integration problem $I\{\tilde{m}\}$, we obtain the following approximation of its solution

$$I\{\tilde{m}\} = \int_0^1\int_0^1 \tilde{m}(s_\epsilon, s_x)\, ds_\epsilon\, ds_x = \frac{1}{2}\int_0^1 \sum_{i=0}^{n} \phi_i \tilde{m}\Big(\frac{s_{n,i}+1}{2}, s_x\Big)\, ds_x + \frac{1}{2}\int_0^1 E_n(s_x)\, ds_x$$
$$= \frac{1}{4}\sum_{i=0}^{n}\sum_{j=0}^{n'} \phi_i\phi_j \tilde{m}\Big(\frac{s_{n,i}+1}{2}, \frac{s_{n',j}+1}{2}\Big) + E_{n,n'}\{\tilde{m}\}$$

for some non-negative integer number n' representing the number of points for the quadrature with respect to s_x. We are interested in determining the magnitude of $E_{n,n'}\{\tilde{m}\}$.

When $n = [\frac{r_\epsilon}{2}]$ and $n' = [\frac{r_x}{2}]$, the error term can be explicitly formulated using Equation (3.50) since the function $\tilde{m} \in \mathbb{C}^{r_\epsilon, r_x}([0,1]^2)$. It means that the partial derivatives of the function $\tilde{m}$ which are of the order r_ϵ and r_x or less with respect to s_ϵ and s_x,respectively, they exist and are continuous. For this case, the error term can be explicitly calculated as

$$E_{n,n'}\{\tilde{m}\} = \frac{1}{2}\int_0^1 E_n(s_x)\, ds_x + \frac{1}{4}\sum_{i=0}^{n} \phi_i E_{n'}\Big(\frac{s_{n,i}+1}{2}\Big)$$

where

$$E_{n'}(s_\epsilon) = \frac{2}{(2n'+1)!}\Big[\frac{2^{n'}(n'!)^2}{(2n')!}\Big]^2 \frac{\partial^{2n'}}{\partial s_x^{2n'}}\tilde{m}\Big(s_\epsilon, \frac{\varsigma+1}{2}\Big) \text{ for } -1 < \varsigma < 1.$$

Then

$$E_{n,n'}\{\tilde{m}\} = \frac{1}{4}\sum_{j=0}^{n'} \phi_j E_n\Big(\frac{s_{n',i}+1}{2}\Big) + \frac{1}{4}\sum_{i=0}^{n} \phi_i E_{n'}\Big(\frac{s_{n,i}+1}{2}\Big) + \frac{2^2}{(2n+1)!(2n'+1)!}\Big[\frac{2^{n'}(n'!)^2}{(2n')!}\,\frac{2^n(n!)^2}{(2n)!}\Big]^2 \frac{\partial^{2n+2n'}}{\partial s_\epsilon^{2n}\partial s_x^{2n'}}\tilde{m}\Big(\frac{\zeta+1}{2}, \frac{\varsigma+1}{2}\Big)$$

for $-1 < \zeta < 1$ and $-1 < \varsigma < 1$. □

Lemma A19. *The function* $\tilde{w} : \mathbb{R}^d \times [0,1]^2 \longrightarrow \mathbb{R}$ *defined in Theorem 13 is of class* $\mathbb{C}^{r_\epsilon, r_x}\left([0,1]^2\right)$ *with respect to* $s_{\epsilon,t}$ *and* $s_{x,t}$ *for* $r_\epsilon = \min\left[\frac{\nu_\epsilon}{2}\right]$ *and* $r_x = \min\left[\frac{\nu_x}{2}\right]$*, that is, for* $0 \leq i \leq r_\epsilon$ *and* $0 \leq j \leq r_x$*, the partial derivative*

$$\frac{\partial^{i+j}\tilde{w}(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi)}{\partial s^i_{\epsilon,t}\, \partial s^j_{x,t}}$$

exists and is continuous.

Proof. The proof of the differentiability class for the function $\tilde{w}$, defined in Theorem 13 as

$$\begin{aligned}
\tilde{w}(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi, \Psi^*) &= \log \pi_{S_{\epsilon,t}|\Psi^*}(s_{\epsilon,t}) + \log \pi_{S_{x,t}|\Psi^*}(s_{x,t}) - \frac{k+d}{2}\log 2\pi - \frac{d}{2}\log\sigma^{*2} + \frac{1}{2}\sum_{i=1}^{d}\log\frac{\chi^{-1}_{\nu^i_\epsilon}(s_{\epsilon,t})}{\nu^i_\epsilon} + \frac{1}{2}\sum_{j=1}^{k}\log\frac{\chi^{-1}_{\nu^j_x}(s_{x,t})}{\nu^j_x} \\
&- \frac{1}{2\sigma^{*2}}\left(\mathbf{y}_t - \boldsymbol{\mu}^* - \delta^*_\epsilon \mathbf{D}^{-1}_{\epsilon,t}\right)\mathbf{D}_{\epsilon,t}\left(\mathbf{y}_t - \boldsymbol{\mu}^* - \delta^*_\epsilon \mathbf{D}^{-1}_{\epsilon,t}\right)^T - \frac{1}{2}\operatorname{Tr}\left\{\sigma^2 \mathbf{M}_t^{-1}\left(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\mathbf{W}^* + \mathbf{D}_{x,t}\right)\right\} \\
&+ \left(\left(\mathbf{y}_t - \boldsymbol{\mu} - \delta_\epsilon \mathbf{D}^{-1}_{\epsilon,t}\right)\mathbf{D}_{\epsilon,t}\mathbf{W} + \sigma^2\delta_x\right)\mathbf{M}_t^{-1}\left(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\left(\mathbf{y}_t - \boldsymbol{\mu}^* - \delta^*_\epsilon \mathbf{D}^{-1}_{\epsilon,t}\right)^T + \delta^{*T}_x\right) \\
&- \frac{1}{2}\operatorname{Tr}\left\{\mathbf{M}_t^{-1}\left((\mathbf{y}_t - \boldsymbol{\mu} - \delta_\epsilon \mathbf{D}^{-1}_{\epsilon,t})\mathbf{D}_{\epsilon,t}\mathbf{W} + \sigma^2\delta_x\right)^T\left((\mathbf{y}_t - \boldsymbol{\mu} - \delta_\epsilon \mathbf{D}^{-1}_{\epsilon,t})\mathbf{D}_{\epsilon,t}\mathbf{W} + \sigma^2\delta_x\right)\mathbf{M}_t^{-1}\left(\frac{1}{\sigma^{*2}}\mathbf{W}^{*T}\mathbf{D}_{\epsilon,t}\mathbf{W}^* + \mathbf{D}_{x,t}\right)\right\},
\end{aligned}$$

following the same steps as in Lemma A17. One needs to recall that the function $\tilde{w}$ can be represented as a composite of a rational polynomial and the function g which is defined in the proof of Lemma A19. The rational polynomial, which define $\tilde{w}$, satisfies the same conditions of smoothness as the rational polynomial p defined in the proof of Lemma A19. Hence, the function $\tilde{w}$ belongs to the class $\mathbb{C}^{r_\epsilon, r_x}\left([0,1]^2\right)$. □

Lemma A20. *Let us define the function* $h(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi) = \tilde{w}(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi, \Psi^*)\, m(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi, \Psi^*)$ *for* m *and* $\tilde{w}$ *defined in Theorem 13. The solutions of the integration problem*

$$\int_0^1 \int_0^1 h(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t}$$

can be approximated by the 2-dimensional Gaussian-Legendre product quadrature rule on $[0,1]^2$ *such that*

$$\int_0^1 \int_0^1 h(\mathbf{y}_t, s_{\epsilon,t}, s_{x,t}; \Psi)\, ds_{\epsilon,t}\, ds_{x,t} = \frac{1}{4}\sum_{i=0}^{n}\sum_{j=0}^{n'}\phi_i\phi_j h\left(\mathbf{y}_t, \frac{s_{n,i}+1}{2}, \frac{s_{n',j}+1}{2}\right) + E_{n,n'}\{m\}$$

where ϕ_i *are defined as in Equation (3.49), the distinct points* $s_{n,i} \in [-1,1]$ *for* $i = 0, \ldots, n$ *and* $s_{n',j} \in [-1,1]$ *for* $j = 0, \ldots, n'$ *are the zero roots of the Legendre polynomials* P_n *and* $P_{n'}$*, respectively, defined in Equation (3.48). If* $n = [\frac{r_\epsilon}{2}]$ *and* $n' = [\frac{r_x}{2}]$ *for* r_ϵ *and* r_x *specified in Lemma A19, the quadrature error term* $E_{n,n'}\{h\}$ *is given by the following explicit formula*

$$\begin{aligned}
E_{n,n'}\{h\}(\mathbf{y}_t) &= \frac{1}{4}\sum_{j=0}^{n'}\phi_j E_n\left(\mathbf{y}_t, \frac{s_{n',i}+1}{2}\right) + \frac{1}{4}\sum_{i=0}^{n}\phi_i E_{n'}\left(\mathbf{y}_t, \frac{s_{n,i}+1}{2}\right) \\
&+ \frac{1}{(2n+1)!(2n'+1)!}\left[\frac{2^{n'}(n'!)^2}{(2n')!}\,\frac{2^n(n!)^2}{(2n)!}\right]^2 \frac{\partial^{2n+2n'}}{\partial s^{2n}_\epsilon \partial s^{2n'}_x} h\left(\mathbf{y}_t, \frac{\zeta+1}{2}, \frac{\varsigma+1}{2}; \Psi\right)
\end{aligned}$$

for $-1 < \zeta < 1$ *and* $-1 < \varsigma < 1$ *where*

$$E_n(\mathbf{y}_t, s_x) = \frac{2}{(2n+1)!}\left[\frac{2^n(n!)^2}{(2n)!}\right]^2 \frac{\partial^{2n}}{\partial s^{2n}_\epsilon} h\left(\mathbf{y}_t, \frac{\zeta+1}{2}, s_x; \Psi\right)$$

$$E_{n'}(\mathbf{y}_t, s_\epsilon) = \frac{2}{(2n'+1)!}\left[\frac{2^{n'}(n'!)^2}{(2n')!}\right]^2 \frac{\partial^{2n'}}{\partial s_x^{2n'}} h\Big(\mathbf{y}_t, s_\epsilon, \frac{\varsigma+1}{2}; \Psi\Big).$$

Proof. The proof of Lemma A20 is analogous to the proof of Lemma A18. We recall the smoothness of function m and $\tilde{w}$ proved in Lemma A17 and Lemma A19, respectively and the fact that differentiability class of the product of two functions equals to the minimum order of differentiability of these functions. □

B. Supplementary Material to Chapter 11

Overview of Appendices

In the following section we provide supplementary information to the studies discussed in the main part of the manuscript,

- in Appendix B.1 we provide the detailed overview of the time series analysed in this study and discussed in Section 11.2 along with their Bloomberg identifiers;
- in Appendix B.3 we detail the supplementary results to the real data case studies.

B.1. Data Description

The following section provides detail discussion on financial and macroeconomic data sets used in this study and briefly overviewed in Section 11.2.

B.1.1. Euro Libor Yield Curve

The Libor rate is the benchmark of interbank interest rates estimated by large global banks in London. It represents an average cost of short-term loans which are charged when one bank borrows funds from other banks for a given period of time. It is calculated by Intercontinental Exchange and published daily by Thomson Reuters. The currently active maturities for which the Libor rates are specified include 1 day, 1 week, 1 month, 2 months, 3 months, 6 months and 12 months. For further details of the calculation methodology please refer to the ICE Libor official website.

Euro Libor denotes Libor rate for interbank deals denominated in Euro currency. The yield curve used in this study is sourced from Bloomberg using the identifiers given in Table A1.

Table A1. The availability and the Bloomberg identifiers (tickers) of Euro Libor yield curves sourced from Bloomberg.

Maturity	Ticker	Availability	Maturity	Ticker	Availability
1W	EUR001W	01/2006 - 08/2017	7Y	EUSW7V3	01/2006 - 08/2017
1M	EUR001M	01/2006 - 08/2017	8Y	EUSW8V3	01/2006 - 08/2017
2M	EUR002M	01/2006 - 08/2017	9Y	EUSW9V3	01/2006 - 08/2017
3M	EUR003M	01/2006 - 08/2017	10Y	EUSW10V3	01/2006 - 08/2017
1Y	EUSW1VC	01/2006 - 08/2017	15Y	EUSW15V3	01/2006 - 08/2017
2Y	EUSW2V3	01/2006 - 08/2017	20Y	EUSW20V3	01/2006 - 08/2017
3Y	EUSW3V3	01/2006 - 08/2017	25Y	EUSW25V3	01/2006 - 08/2017
5Y	EUSW5V3	01/2006 - 08/2017			

The time series span from 01/01/2006 - 14/08/2017 for maturities 1 week (1W), 1 month (1M), 2 months (2M), 3 months (3M), 1 year (1Y), 2 years (2Y), 3 years (3Y), 5 years (5Y), 7 years (7Y), 8 years (8Y), 9 years (9Y), 10 years (10Y), 15 years (15Y), 20 years (20Y) and 25 years (25Y). We use Euro Libor rates provided by Intercontinental Exchange up to 12 months, whereas additional 1Y and longer maturities are represented by swap rates on Euro Libor. The panels in Figure A1 present the time series of the Euro Libor yield curve at fixed tenors over time (left) and indicators of missing values in the data over time (right) denoted by single crosses. The reader may notice

that the rates corresponding to the maturities less than 1 year, and provided by Intercontinental Exchange, are less volatile than the instruments constructed using swap rates. In addition, the number of values which are not available over time (are not observed) is much greater for smaller maturities and their pattern is more regular. Furthermore, one can observe that the swap-related rates have the majority of missing values before 2008 and 1Y having none.

(a) Euro Libor yield curve **(b)** The indicators of missing values

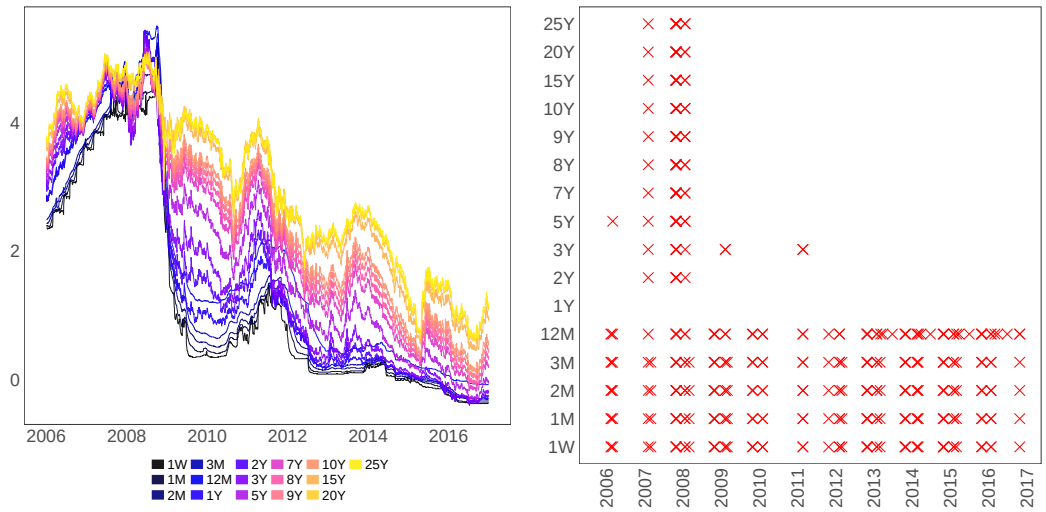


Figure A1. The Euro Libor rates (%) (left panel) and the indicators of missing values (right panel) over time. The different colours of lines in the left panel correspond to the rates with different maturities given in the legend. In the right panel, a single cross indicates daily unavailability of an instrument. The missingness does not correspond to weekends and holidays.

B.1.2. Country-Specific Sovereign Yield Curves

The full sample of country-specific sovereign yield curves consists of 94 time series of interest rates corresponding to different maturities from the 12 countries listed in Table 11.1. The examined sample spans 01/2006 - 12/2016. The sovereign yield curves are downloaded from Bloomberg using the search query '**GGR**' which lists generic rates for international government securities. The yields are comprised of the most recently issued or closest current nominal maturity government security over time. One can locate the underlying benchmark bonds under Blomberg identifiers of sovereign yield curves, given in Table A2.

Table A2. The availability and Bloomberg identifiers (tickers) of the sovereign yield curves and inflation-linked yield curves sourced from Bloomberg. The column Type indicate what are the benchmark instruments of the inflation-linked yields, bond rates (1) or swap rates (2)

Country	Maturity	Sovereign Yield Curve		Inflation-linked Yield Curve	
		Ticker	Availability	Ticker	Availability
AU	3M	GACGB3M Index	05/2009 - 12/2016		
	1Y	GACGB1 Index	01/2006 - 12/2016	AUSWIT1 Index	06/2007 - 12/2016
	2Y	GACGB2 Index	01/2006 - 12/2016	AUSWIT2 Index	06/2007 - 12/2016
	5Y	GACGB5 Index	01/2006 - 12/2016	AUSWIT5 Index	06/2007 - 12/2016
	7Y	GACGB7 Index	01/2006 - 12/2016	AUSWIT7 Index	06/2007 - 12/2016
	10Y	GACGB10 Index	01/2006 - 12/2016	AUSWIT10 Index	06/2007 - 12/2016
	15Y	GACGB15 Index	01/2006 - 12/2016	AUSWIT15 Index	06/2007 - 12/2016
	20Y	GACGB20 Index	11/2013 - 12/2016	AUSWIT20 Index	06/2007 - 12/2016

Continue on the next page

Table A2. The availability and the reference to the Bloomberg terminal (tickers) of the sovereign yield curves and inflation-linked yield curves sourced from Bloomberg (cont.).

Country	Maturity	Sovereign Yield Curve		Inflation-linked Yield Curve	
		Ticker	Availability	Ticker	Availability
	25Y			AUSWIT25 Index	06/2007 - 12/2016
BZ	3M	GEBR03M Index	03/2007 - 12/2016		
	6M	GEBR06M Index	03/2007 - 12/2016		
	1Y	GEBR01Y Index	03/2007 - 12/2016		
	2Y	GEBR02Y Index	03/2007 - 12/2016		
	5Y	GEBR5Y Index	01/2007 - 12/2016		
	7Y	GEBR7Y Index	01/2007 - 12/2016		
	10Y	GEBR10Y Index	01/2007 - 12/2016		
DE	1Y	GDBR1 Index	01/2006 - 12/2016	GRSWIT1 Curncy	04/2006 - 12/2016
	2Y	GDBR2 Index	01/2006 - 12/2016	GRSWIT2 Curncy	04/2006 - 12/2016
	5Y	GDBR5 Index	01/2006 - 12/2016	GRSWIT5 Curncy	04/2006 - 12/2016
		GDBR5 Index	01/2006 - 12/2016	GEIL5Y Index	12/2008 - 12/2015
	7Y	GDBR7 Index	01/2006 - 12/2016	GRSWIT7 Curncy	04/2006 - 12/2016
		GDBR7 Index	01/2006 - 12/2016	GEIL7Y Index	12/2008 - 12/2016
	10Y	GDBR10 Index	01/2006 - 12/2016	GRSWIT10 Curncy	04/2006 - 12/2016
		GDBR10 Index	01/2006 - 12/2016	GEIL10Y Index	06/2009 - 12/2016
	15Y	GDBR15 Index	05/2008 - 12/2016	GRSWIT15 Curncy	04/2006 - 12/2016
	20Y	GDBR20 Index	01/2006 - 12/2016	GRSWIT20 Curncy	04/2006 - 12/2016
	25Y			GRSWIT25 Curncy	04/2006 - 12/2016
ES	3M	GSPG3M Index	01/2006 - 12/2016		
	6M	GSPG6M Index	01/2006 - 12/2016		
	1Y	GSPG12M Index	01/2006 - 12/2016	SPSWIT1 Index	01/2006 - 12/2016
	2Y	GSPG2YR Index	01/2006 - 12/2016	SPSWIT2 Index	01/2006 - 12/2016
	5Y	GSPG5YR Index	01/2006 - 12/2016	SPSWIT5 Index	01/2006 - 12/2016
	7Y	GSPG7YR Index	01/2006 - 12/2016	SPSWIT7 Index	01/2006 - 12/2016
	10Y	GSPG10YR Index	01/2006 - 12/2016	SPSWIT10 Index	01/2006 - 12/2016
	15Y	GSPG15YR Index	01/2006 - 12/2016	SPSWIT15 Index	01/2006 - 12/2016
	20Y			SPSWIT20 Index	01/2006 - 12/2016
	25Y			SPSWIT25 Index	01/2006 - 12/2016
FR	3M	GBTF3MO Index	01/2006 - 12/2016		
	6M	GBTF6MO Index	01/2006 - 12/2016		
	1Y	GBTF1YR Index	01/2006 - 12/2016	FRSWI1 Index	01/2006 - 12/2016
	2Y	GFRN2 Index	01/2006 - 12/2016	GFRGIN02 Index	04/2009 - 06/2016
		GFRN2 Index	01/2006 - 12/2016	FRSWI2 Index	01/2006 - 12/2016
	5Y	GFRN5 Index	01/2006 - 12/2016	GFRGIN05 Index	06/2006 - 12/2016
		GFRN5 Index	01/2006 - 12/2016	FRSWI5 Index	01/2006 - 12/2016
	7Y	GFRN7 Index	01/2006 - 12/2016	FRSWI7 Index	01/2006 - 12/2016
	10Y	GFRN10 Index	01/2006 - 12/2016	GFRGIN10 Index	06/2006 - 12/2016
		GFRN10 Index	01/2006 - 12/2016	FRSWI10 Index	01/2006 - 12/2016
	15Y	GFRN15 Index	01/2006 - 12/2016	GFRGIN15 Index	04/2009 - 12/2016
		GFRN15 Index	01/2006 - 12/2016	FRSWI15 Index	01/2006 - 12/2016
	20Y	GFRN20 Index	01/2006 - 12/2016	FRSWI20 Index	01/2006 - 12/2016
		GFRN20 Index	01/2006 - 12/2016	GFRGIN20 Index	01/2008 - 06/2012
	25Y			FRSWI25 Index	01/2006 - 12/2016
GB	3M	GUKG3M Index	12/2008 - 12/2016		
	6M	GUKG6M Index	08/2007 - 11/2016		
	1Y	GUKG1 Index	01/2006 - 12/2016	BPSWIT1 Index	01/2006 - 12/2016
	2Y	GUKG2 Index	01/2006 - 12/2016	GUKGIN02 Index	01/2007 - 02/2010
		GUKG2 Index	01/2006 - 12/2016	BPSWIT2 Index	01/2006 - 12/2016
	5Y	GUKG5 Index	01/2006 - 12/2016	GUKGIN05 Index	01/2006 - 12/2016
		GUKG5 Index	01/2006 - 12/2016	BPSWIT5 Index	01/2006 - 12/2016
	7Y	GUKG7 Index	01/2006 - 12/2016	BPSWIT7 Index	01/2006 - 12/2016
	10Y	GUKG10 Index	01/2006 - 12/2016	GUKGIN10 Index	01/2006 - 12/2016
		GUKG10 Index	01/2006 - 12/2016	BPSWIT10 Index	01/2006 - 12/2016
	15Y	GUKG15 Index	01/2006 - 12/2016	BPSWIT15 Index	01/2006 - 12/2016
	20Y	GUKG20 Index	01/2006 - 12/2016	GUKGIN20 Index	01/2006 - 12/2016
		GUKG20 Index	01/2006 - 12/2016	BPSWIT20 Index	01/2006 - 12/2016
	25Y			BPSWIT25 Index	01/2006 - 12/2016
IR	3M	GIGB3M Index	10/2013 - 11/2015		
	1Y	GIGB1YR Index	01/2006 - 12/2016		
	2Y	GIGB2YR Index	01/2006 - 12/2016		
	5Y	GIGB5YR Index	01/2006 - 06/2016		
	7Y	GIGB7YR Index	02/2006 - 12/2016		
	10Y	GIGB10YR Index	01/2006 - 12/2016		
	15Y	GIGB15YR Index	01/2006 - 12/2016		
IT	3M	GBOTG3M Index	01/2006 - 12/2016		

Continue on the next page

Table A2. The availability and the reference to the Bloomberg terminal (tickers) of the sovereign yield curves and inflation-linked yield curves sourced from Bloomberg (cont.).

Country	Maturity	Sovereign Yield Curve		Inflation-linked Yield Curve	
		Ticker	Availability	Ticker	Availability
JP	6M	GBOTG6M Index	01/2006 - 12/2016		
	1Y	GBOTG12M Index	01/2006 - 12/2016	GRITIL1 Index	06/2007 - 12/2016
		GBOTG12M Index	01/2006 - 12/2016	ILSWI1 Index	01/2006 - 12/2016
	2Y	GBTP2YR Index	01/2006 - 12/2016	ILSWI2 Index	01/2006 - 12/2016
	5Y	GBTP5YR Index	01/2006 - 12/2016	GRITIL5 Index	06/2007 - 12/2016
		GBTP5YR Index	01/2006 - 12/2016	ILSWI5 Index	01/2006 - 12/2016
	7Y	GBTP7YR Index	01/2006 - 12/2016	ILSWI7 Index	01/2006 - 12/2016
		GBTP10YR Index	01/2006 - 12/2016	GRITIL10 Index	06/2007 - 12/2016
	10Y	GBTP10YR Index	01/2006 - 12/2016	ILSWI10 Index	01/2006 - 12/2016
		GBTP15YR Index	02/2006 - 12/2016	GRITIL15 Index	05/2009 - 12/2016
	15Y	GBTP15YR Index	02/2006 - 12/2016	ILSWI15 Index	01/2006 - 12/2016
	20Y	GBTP20YR Index	02/2006 - 12/2016	ILSWI20 Index	01/2006 - 12/2016
	25Y			ILSWI25 Index	01/2006 - 12/2016
JP	3M	GJTB3MO Index	01/2006 - 12/2016		
	6M	GJTB6MO Index	01/2006 - 12/2016		
	1Y	GJGB1 Index	01/2006 - 12/2016	JYSWIT1 Index	03/2007 - 12/2016
	2Y	GJGB2 Index	01/2006 - 12/2016	JYSWIT2 Index	03/2007 - 12/2016
	5Y	GJGB5 Index	01/2006 - 12/2016	JYSWIT5 Index	03/2007 - 12/2016
		GJGB5 Index	01/2006 - 12/2016	GJYGIN05 Index	07/2009 - 04/2014
	7Y	GJGB7 Index	01/2006 - 12/2016	GJYGIN07 Index	07/2009 - 07/2012
		GJGB7 Index	01/2006 - 12/2016	JYSWIT7 Index	03/2007 - 12/2016
	10Y	GJGB10 Index	01/2006 - 12/2016	JYSWIT10 Index	03/2007 - 12/2016
		GJGB10 Index	01/2006 - 12/2016	GJYGIN10 Index	01/2006 - 12/2016
	15Y	GJGB15 Index	01/2006 - 12/2016	JYSWIT15 Index	03/2007 - 12/2016
	20Y	GJGB20 Index	01/2006 - 12/2016	JYSWIT20 Index	03/2007 - 12/2016
	25Y			JYSWIT25 Index	03/2007 - 03/2015
PO	3M	GSPT3M Index	01/2006 - 12/2016		
	6M	GSPT6M Index	01/2006 - 12/2016		
	1Y	GSPT12M Index	01/2006 - 12/2016		
	2Y	GSPT2YR Index	01/2006 - 12/2016		
	5Y	GSPT5YR Index	01/2006 - 12/2016		
	7Y	GSPT7YR Index	01/2006 - 12/2016		
	10Y	GSPT10YR Index	01/2006 - 12/2016		
	15Y	GSPT15YR Index	02/2009 - 12/2016		
SA	1Y			SASWIT1 Curncy	05/2007 - 12/2016
	2Y	GSAB2YR Index	01/2006 - 12/2016	SASWIT2 Curncy	05/2007 - 12/2016
	5Y	GSAB5YR Index	01/2006 - 12/2016	SASWIT5 Curncy	05/2007 - 12/2016
	7Y	GSAB7YR Index	07/2007 - 12/2016	SASWIT7 Curncy	05/2007 - 12/2016
	10Y	GSAB10YR Index	01/2006 - 12/2016	SASWIT10 Curncy	05/2007 - 12/2016
	15Y	GSAB15YR Index	01/2006 - 12/2016	SASWIT15 Curncy	05/2007 - 12/2016
	20Y	GSAB20YR Index	01/2006 - 12/2016	SASWIT20 Curncy	05/2007 - 12/2016
	25Y			SASWIT25 Curncy	05/2007 - 12/2016
US	6M	USGG6M Index	01/2006 - 12/2016		
	1Y	USGG12M Index	06/2008 - 12/2016	USSWIT1 Index	01/2006 - 12/2016
	2Y	USGG2YR Index	01/2006 - 12/2016	USSWIT2 Index	01/2006 - 12/2016
	5Y	USGG5YR Index	01/2006 - 12/2016	USGGT05Y Index	01/2006 - 12/2016
		USGG5YR Index	01/2006 - 12/2016	USSWIT5 Index	01/2006 - 12/2016
	7Y	USGG7YR Index	02/2009 - 12/2016	USSWIT7 Index	01/2006 - 12/2016
	10Y	USGG10YR Index	01/2006 - 12/2016	USGGT10Y Index	01/2006 - 12/2016
		USGG10YR Index	01/2006 - 12/2016	USSWIT10 Index	01/2006 - 12/2016
	15Y			USSWIT15 Index	01/2006 - 12/2016
	20Y			USSWIT20 Index	01/2006 - 12/2016
				USGGT20Y Index	01/2006 - 12/2016
	25Y			USSWIT25 Index	01/2006 - 12/2016

The Figure A2 illustrates the evolution of sovereign yield curves (left subfigure) and the associate unavailability of the yield over time (right subfigure), where each panel corresponds to country-specific statistics. The yield curves are characterized by both, irregular sets of available maturities and unevenly spaced time series. The full information related to the availability of the instruments and their Blomberg identifiers are given in Table A2.

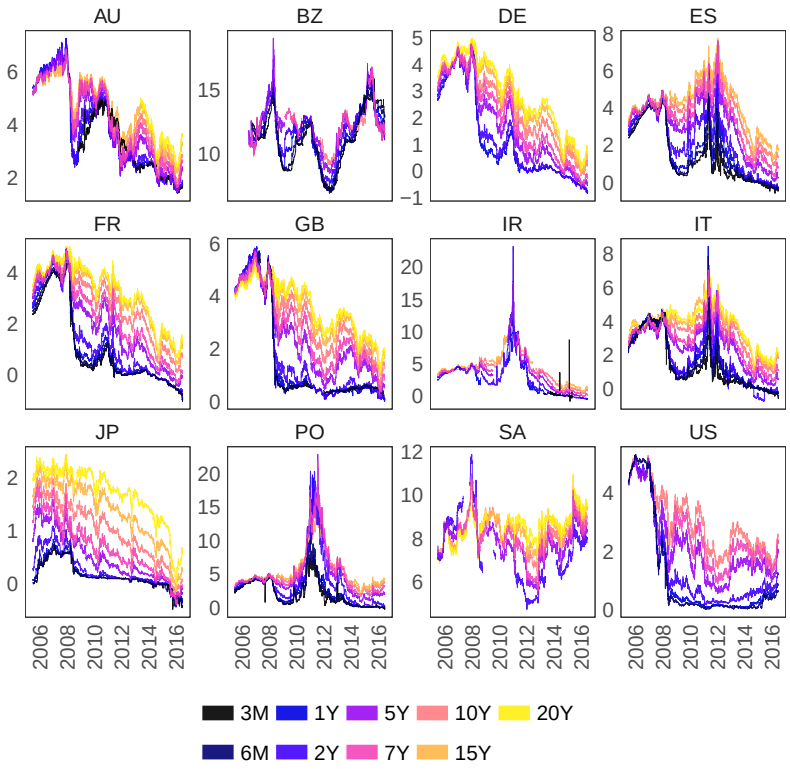


Figure A2. The sovereign yield curves (%) of 12 countries listed in Table 11.1. The colors of lines correspond to the diffrent maturities.

The irregular availability of time of the country-specific yield curves illustrated in the right panel of Figure A2 is manifest in two distinct types of patterns highlighted daily for a given instrument by a single red cross (when missing). The first pattern is related to the different dates when instruments have started to be issued, for instance, the Irish sovereign yield curve at 3M maturity only began to be available from 10/2013. Therefore, we face constantly in time unavailability of this instrument which is indicated on the plots as a solid straight red line (constructed from a dense set of red crosses) which starts from the beginning of the sample. The second pattern appears when we have a sparse set of dates for which the yields are only temporally not available. It can be the result of a temporal lack of financial securities liquid enough to calculate the rate. This pattern is highlighted by single, sparse groups of red crosses. For instance, one may notice short unavailabilities of 3M, 6M and 1Y maturities from the French sovereign yield curve. Figure A4 and Figure A5 summarises the overall proportion of missing values per a sovereign bond rate across different countries and maturities given the sample size of 2 870 days (excluding holidays and weekends).

B.1.3. Measures of Inflation Risk

The data under the category **Inflation** consists of monthly Consumer Price Indexes from 12 countries and 13 daily end-of-the-day inflation-indexed swap or bond based yield curves, from 8 major global economies: Germany, France, Spain, Italy, United Kingdom, Japan, Australia and the United States of America. The country-specific rates are referenced against CPI indexes.

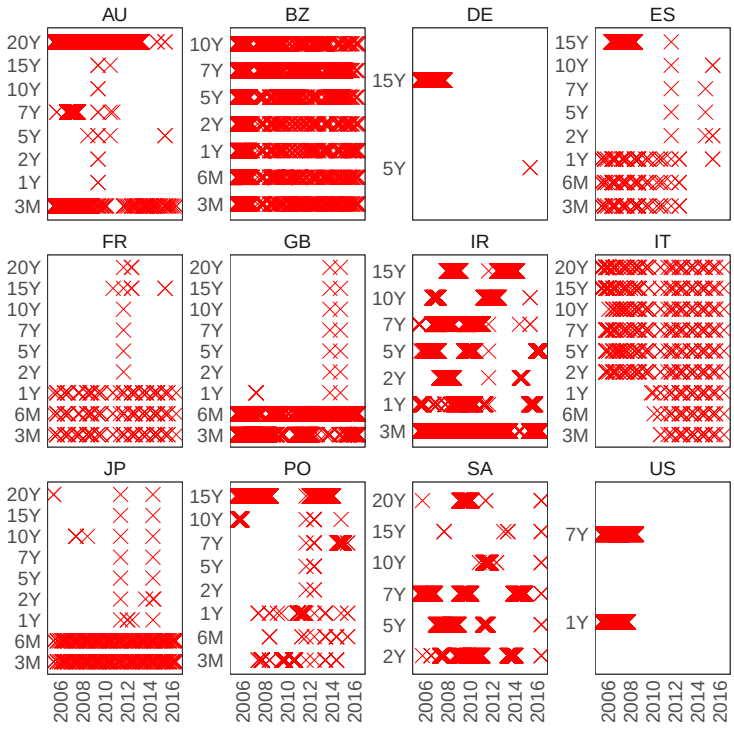


Figure A3. The indicators of missing values for sovereign yields of 12 countries listed in Table 11.1. The red crosses correspond to daily unavailabilities of rates at various maturities (y-axis) over time (x-axis). The missingness does not correspond to weekends and holidays.

B.1.3.1. Country-Specific Consumer Price Index

The nation-wide Consumer Price Index (CPI) is usually tailored to the countries' objectives and therefore differs across different economies with regard to the products included into its construction.

In order to ensure that the CPIs among 12 economies are calculated using a methodology which is unified as much as possible across different countries, we use the Harmonised Index of Consumer Prices (HICP) for European Union members which reflects monthly changes. The monthly time series of the index are calculated by Eurostat on behalf of National Statistics Institutes (NSIs) of the European Union member states and European Free Trade Association (EFTA) which provides the guidelines for calculating HICP, please refer to the official website for details. The term 'harmonised' highlights the fact that all the countries in the European Union follow the same methodology. In the report provided by the Office of National Statistics in (Eidukas 2016), the author exemplifies the most frequent differences between CPIs and HICP, i.e showing that the majority of CPIs excludes the expenditures of non-residents in the economic territory. The HICP and CPI are available from Organisation for Economic Cooperation and Development (OECD) database. The CPIs for Japan, South Africa and the United States of America, are specified on a monthly basis, whereas the CPI of Australian is available quarterly and illustrates yearly changes. The indexes are calculated by national statistical institutions: Australian Bureau of Statistics, Ministry of Internal Affairs and Communications, Statistics South Africa and Bureau of Labor Statistics respectively. As a measure of Brazilian Consumer Price Index we use monthly time series

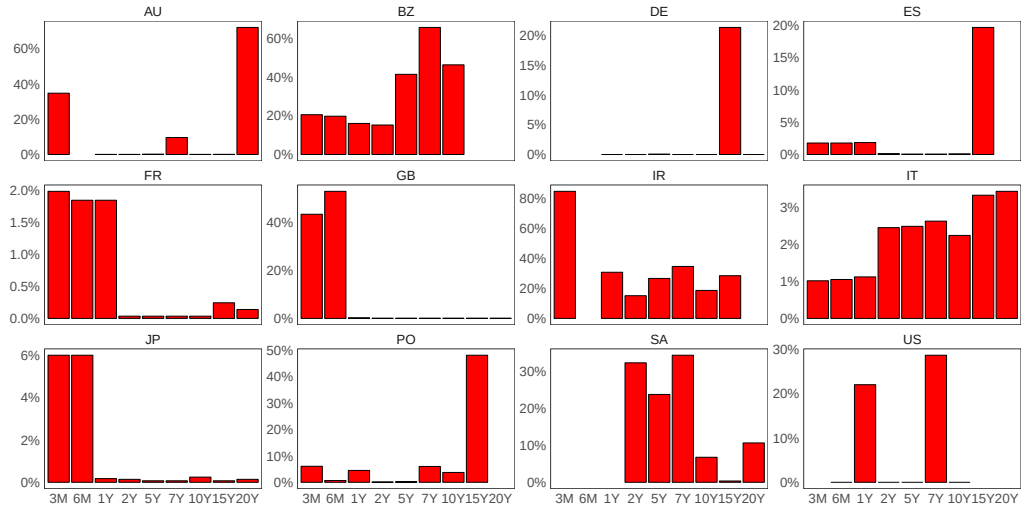


Figure A4. The proportion (%) of missing values present in the time series of sovereign bond rates across different maturities for a sample period 01/2006 - 12/2016 (2870 days excluding weekends and holidays). The blank spaces on the x-axis correspond to instruments which are not available in Bloomberg. The patterns of missingness are illustrated in Figure A3. The missingness does not correspond to weekends and holidays.

of Extended National Consumer Price Index) which covers families living in 11 urban areas of Brazil: Belem, Fortaleza, Recife, Salvador, Belo Horizonte, Rio de Janeiro, Sao Paulo, Curitiba, Porto Alegre, Brasilia and Goiania and whose monthly income is between 1 and 40 times the Brazilian minimum wage. The Bloomberg identifiers of the collected CPIs are listed in Table A3 and the evolution of indexes over time is illustrated in the top left panel of Figure A9.

B.1.3.2. Country-Specific Inflation-Linked Yield Curves

The collected inflation-linked sovereign yield curves were sourced following similar steps adopted for the sovereign yield curves via a search query of Bloomberg generic yield curves, '**GGR**'. The sets of inflation-linked yields are shorter and consists of yield curves for 9 countries as indicated in Table 11.1. The Bloomberg identifiers of elements of the yields which correspond to a single maturity, and their availabilities are listed in Table A2. The collected set of inflation-linked sovereign yields can be partitioned into two groups, the yields which are comprised of inflation-indexed government securities, and those which are comprised of inflation-linked swap rates. The inflation-linked instruments are referenced against a consumer price index calculated monthly by national statistics institutes.

The most commonly used indicator of inflation expectations is derived as the difference between a nominal conventional and inflation-linked bonds yields at similar maturity. However, with the rapid growth of the swap market starting from 2003, it is nowadays much easier to construct comprehensive inflation-linked yield curve which is comprised of inflation-linked swap rates, rather than sovereign bonds, due to their broader availability. Various studies of global market regulators such as (Schulz and Stapf 2014), investigate the quality of the prediction given by swap-based inflation-linked sovereign yield curve. The authors of (Schulz and Stapf 2014) argue that the measures of inflation expectations extracted from inflation linked-swaps bias the indicator of inflation expectations. They indicate the inflation risk premium and liquidity premium which

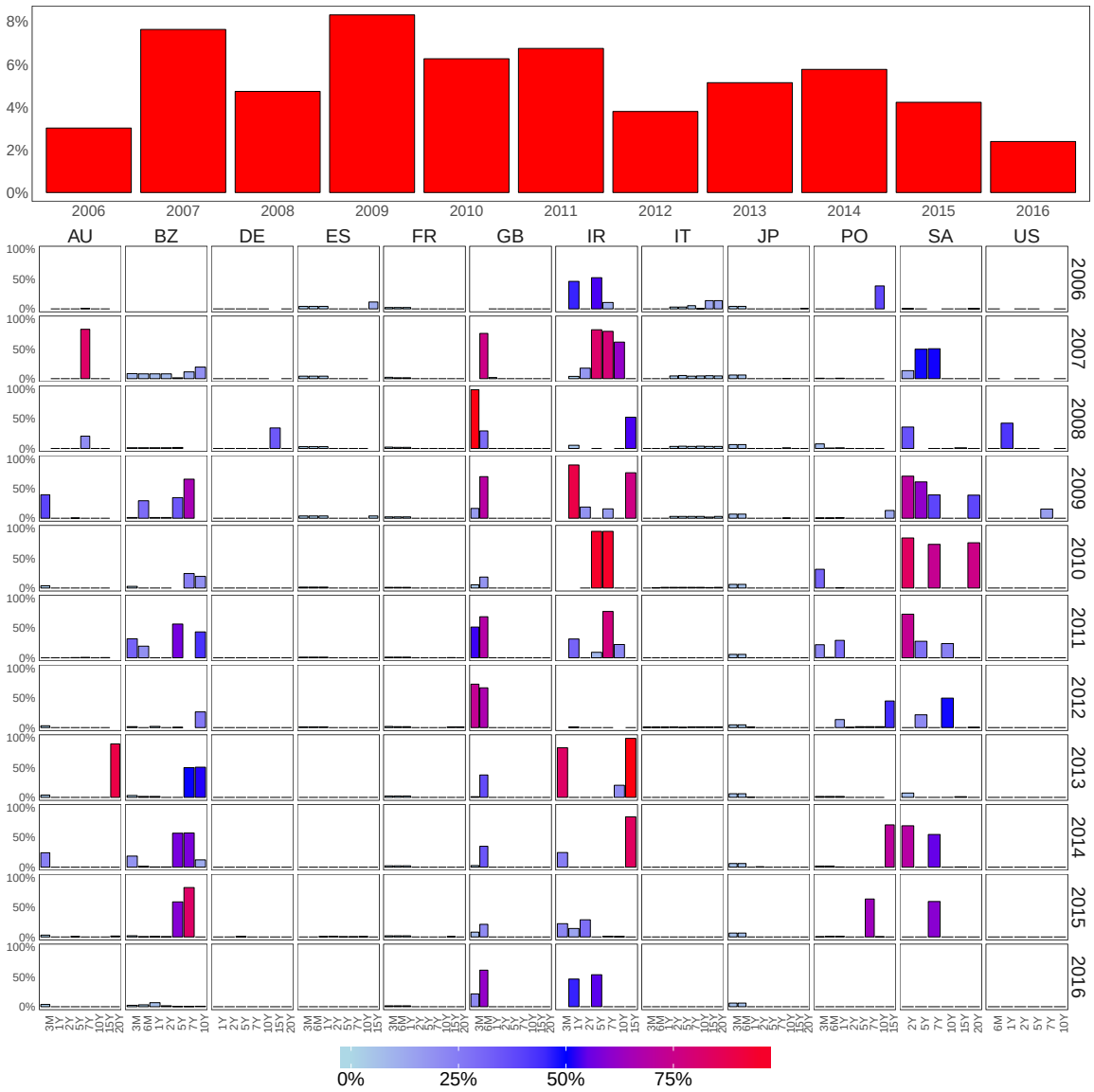


Figure A5. The percentage of the aggregated missing entries per year (top panel) and the percentage of missing entries per year (column-wise) and country (row-wise) of yields across maturities (x-axis) from sovereign yields. The colours of bars in the bottom plot correspond to the proportion of the missingness.

is mitigated from the objectives of the inflation-linked swap market participants. For instance, the corporations, which revenue is linked to inflation, hedge against the risk of low inflation, or pension funds and insurance companies, which hedge against the risk of high inflation and rising prices. The panels in Figure A6 show the evolution of inflation-linked swap (the blue line) and sovereign bonds (the black line) yield curves across different maturities (column-wise) for 8 countries. The corresponding country-specific process of missing values indicator is illustrated in left and right panels of Figures A8, respectively, with proportions of missing values per the whole sample listed in Figure A4.

The reader may notices similarities between the country-specific yields from Figure A2 and the bond-based inflation-linked sovereign yield curves as the sovereign yield curves comprised either of conventional or inflation-linked bonds, share common dynamics which are characteristic among chosen countries and more resemble each other than the corresponding swap rates. On

Table A3. The availability and Bloomberg identifiers (tickers) of the national macroeconomic measures of economic activity: Consumer Price Index (CPI), Gross Domestic Product (GDP), Labour Productivity (LP), Unemployment Rates (UE) and German Bund Open Interest, 2W Euro repo rate, 5Y Markit Intraxx Index and exchange rates between major global economies.

Instrument	Country	Ticker	Availability	Instrument	Country	Ticker	Availability
Consumer Price Index	AU	AUCPIYOY Index	03/2006 - 12/2016	Gross Domestic Product	AU	WGDPAUST Index	12/2006 - 12/2016
	BZ	BZPIIPCY Index	01/2006 - 12/2016		BZ	WGDPBRAZ Index	12/2006 - 12/2016
	DE	ECCPDEMM Index	01/2006 - 12/2016		DE	WGDPGERM Index	12/2006 - 12/2016
	ES	ECCPESMM Index	01/2006 - 12/2016		ES	WGDPSPAI Index	12/2006 - 12/2016
	FR	ECCPFRMM Index	01/2006 - 12/2016		FR	WGDPFRAN Index	12/2006 - 12/2016
	GB	ECCPUKMM Index	01/2006 - 12/2016		GB	WGDPUK Index	12/2006 - 12/2016
	IR	ECCPIEMM Index	01/2006 - 12/2016		IR	WGDPIREL Index	12/2006 - 12/2016
	IT	ECCPITMM Index	01/2006 - 12/2016		IT	WGDPITAL Index	12/2006 - 12/2016
	JP	JNCPIMOM Index	01/2006 - 12/2016		JP	WGDPJAPA Index	12/2006 - 12/2016
	PO	ECCPPTMM Index	01/2006 - 12/2016		PO	WGDPPORT Index	12/2006 - 12/2016
	SA	SACPIYOY Index	01/2006 - 12/2016		SA	WGDPSOUT Index	12/2006 - 12/2016
	US	CPI YOY Index	01/2006 - 12/2016		US	WGDPUS Index	12/2006 - 12/2016
Labour Productivity	AU	OEEOAULP Index	03/2006 - 12/2016	Unemployment Rate	AU	EHUPAU Index	03/2006 - 12/2016
	BZ	OEEOBRLF Index	12/2006 - 12/2016		BZ	EHUPBR Index	03/2006 - 12/2016
	DE	OEEODELP Index	03/2006 - 12/2016		DE	EHUPDE Index	12/2005 - 12/2016
	ES	OEEOESLP Index	12/2006 - 12/2016		ES	EHUPES Index	03/2006 - 12/2016
	FR	OEEOFRLP Index	03/2006 - 12/2016		FR	EHUPFR Index	03/2006 - 12/2016
	GB	OEEOGBLP Index	03/2006 - 12/2016		GB	EHUPGB Index	03/2006 - 12/2016
	IR	OEEOIELP Index	03/2006 - 12/2016		IR	EHUPIE Index	03/2006 - 12/2016
	IT	OEEOITLP Index	03/2006 - 12/2016		IT	EHUPIT Index	03/2006 - 12/2016
	JP	OEEOJPLP Index	12/2005 - 12/2016		JP	EHUPJP Index	03/2006 - 12/2016
	PO	OEEOPTLP Index	03/2006 - 12/2016		PO	EHUPPT Index	03/2006 - 12/2016
	SA	OEEOZALF Index	12/2005 - 12/2016		SA	EHUPZA Index	03/2006 - 12/2016
	US	OEEOUSLP Index	03/2006 - 12/2016		US	EHUPUS Index	03/2006 - 12/2016
German Bund Open Interest		RX1 Comdty	01/2006 - 12/2016				
3M Euro repo rate		EURR003M Index	01/2006 - 12/2014				
3M Euribor rate		EUR003M Index	01/2006 - 12/2016				
5Y Markit Itraxx		ITRX EUR CDSI GEN 5Y Corp	10/2011 - 12/2016				
FX	AUDUSD	AUDUSD Curncy	01/2006 - 12/2016				
	GBPUSD	GBPUSD Curncy	01/2006 - 12/2016				
	USDJPY	USDJPY Curncy	01/2006 - 12/2016				
	EURUSD	EURUSD Curncy	01/2006 - 12/2016				

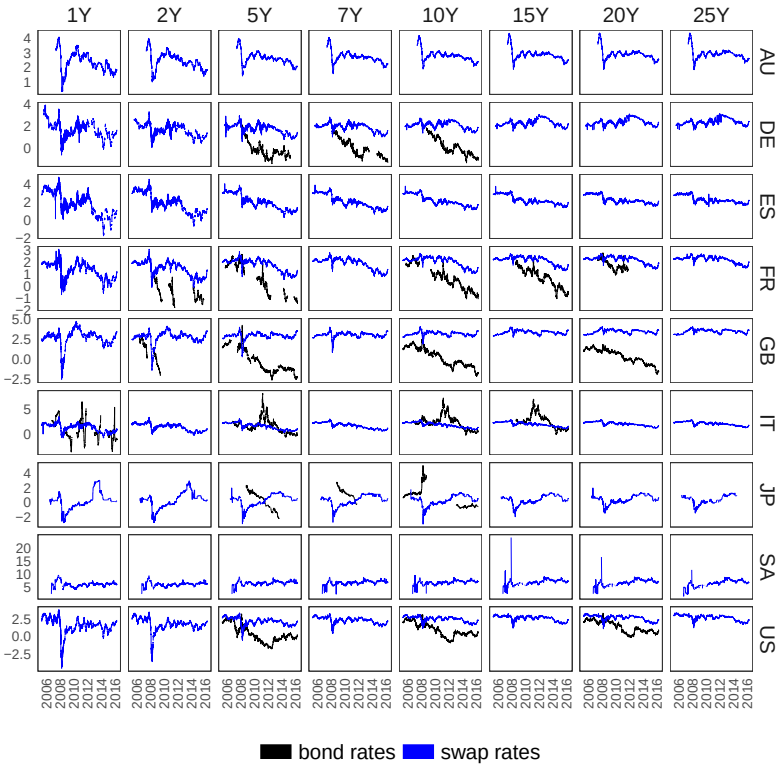


Figure A6. The comparison of inflation-lined yield curves (%) obtained from inflation-linked government bonds (black lines) or inflation-linked swap rates (blue lines) of 8 countries and Euro- zone listed in Table 11.1. The column-wise order of the panels corresponds to the subsequent components of a term structure (maturities), whereas the yields corresponding to particular countries are illustrated by the row-wise order.

the other hand, the availability of the inflation-linked bond-based sovereign yield curves is much more limited both in time and the number of countries, than the swap-based yield curves. The

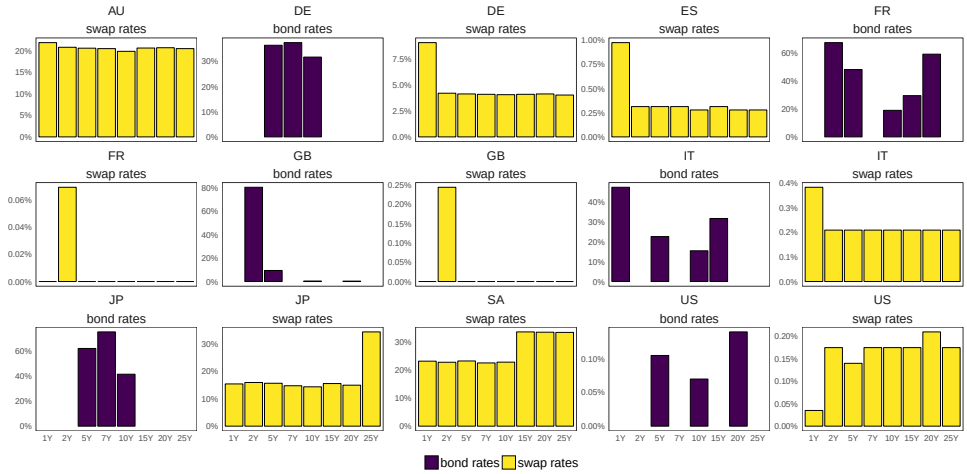
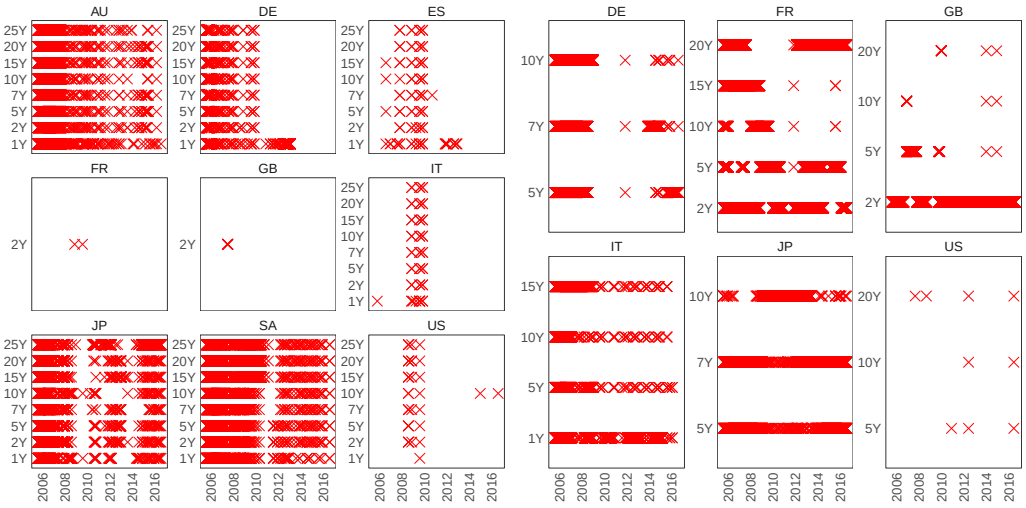


Figure A7. The proportion (%) of missing values present in the time series of inflation-linked sovereign bond or swap rates across different maturities and countries for a sample period 01/2006 - 12/2016 (2'870 days excluding weekends and holidays). The patterns of missingness are illustrated in Figure A8. The blank spaces on the x-axis correspond to instrumented which are not available in Bloomberg.

panels in Figure A7 summarize the proportion of country-specific inflation-linked yield curves which are missing over the sample span, 01/2006-12/2016, which includes 2'870 days.



(a) swap-based rates **(b)** bond-based rates

Figure A8. The indicators of missing values of swap-based (left) and sovereign (right) yield curves comprised of inflation-linked rates at different maturities for countries listed in Table 11.1. The red crosses on the left subfigure correspond to daily lack of availability of rates. The missingness does not correspond to weekends and holidays.

B.1.4. Country-Specific Productivity Proxies

The country-specific productivity is an indicator which measures economic growth, competitiveness, and living standards within an economy. We consider productivity proxies from three indicators: yearly Gross Domestic Product in US dollars obtained from World Bank for 12 economies, quarterly time series of unemployment rates for 12 economies from Bloomberg

and labour productivity for all countries except Brazil and South Africa from the OECD. In order to obtain labour productivity approximation for Brazil and South Africa, we collected quarterly numbers of GDP per labour force from OECD. The time series of GDP, Unemployment rates and labour productivity are presented in Figure A9 whereas the corresponding Bloomberg identifiers are listed in Table A3.

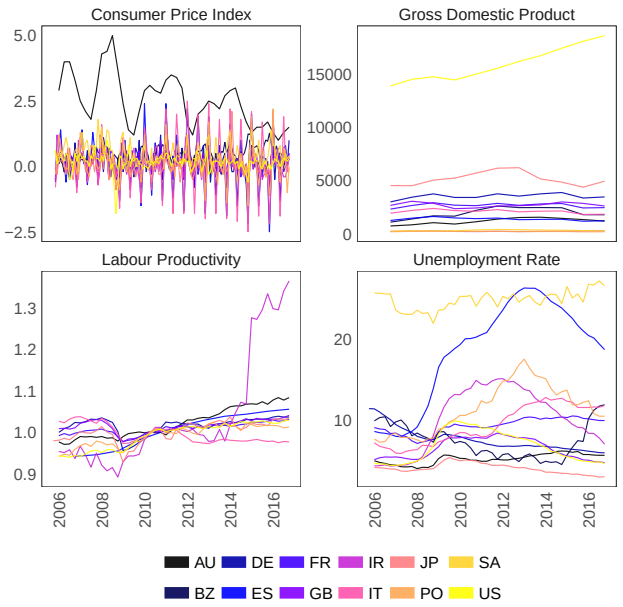


Figure A9. Daily Consumer Price Indexes (top left), quarterly Gross Domestic Products (top right), Unemployment rate in % (bottom left) and Labour Productivity (bottom right) of 12 countries listed in Table 11.1. Different colours of lines refer to different countries.

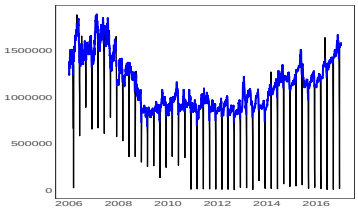
B.1.5. Euro-Zone Liquidity & Credit Proxies and Foreign Exchange

The instruments which we use as a proxy to Euro-Zone liquidity are Open Interests of Euro-Bund futures contracts and 2W Euro repo rate.

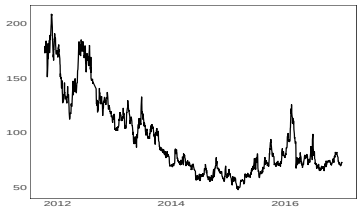
The Euro-Bund futures are the most popular contracts of EUREX, which is the largest exchange for Euro-denominated derivatives and has been ranked within main derivative exchanges by volume. The bond futures contracts are exchange-traded products with an underlying basket of deliverable bonds. The benchmark bonds of Euro-Bund futures are long-term notional debt securities issued by the German Federal Government at reference yield of 6% and with a term of 8.5 to 10.5 years which earliest two delivery months are available for trading. The bonds are considered the benchmark for long-term Euro-denominated government debts. The futures exist for 4 contracts which correspond to the following months: March, June, September and December. In the current study we use Generic 1st 'RX' Future contract provided by Bloomberg, that is, the time series of futures which is comprised of front-month instruments, the contracts with the shortest duration that could be purchased. As a proxy of Euro-zone liquidity, we use the open interest of Euro-Bund futures, the number of the outstanding contracts that exist on a given date. The amount of open interests represents the flow of the money to the futures market and indicates increased interests of market participants. The Figure A10a shows the dynamic of open interest

over time (black solid line) with its smoothed curve. We used median smoothing over a 20-day rolling window in order to remove spikes from the time series which are downward movements of a number of outstanding contracts when the instrument is close to its expiry date.

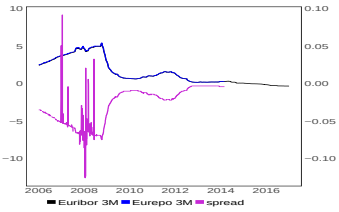
(a) Open Interests



(b) 5Y Markit Itraxx Index



(c) 3M Euribor and 3M Eurepo



(d) Exchange rates against EUR

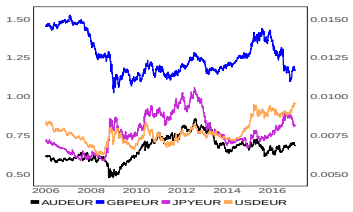


Figure A10. Top left panel (a): The raw (black) and smoothed (blue) open interests of Euro-Bund futures contracts over times. Top right panel (b): 5Y Marikt Itraxx Index over time. The bottom left panel (c): the spread (secondary y-axis) between 3 month Euribor and Eurorepo rates (%) over time. The bottom right panel (d): the exchange rates against EUR for AUD, GBP and USA (left y-axis) and JPY (right y-axis)over time.

In addition, as a proxy for short-term liquidity, we use the spread between the 3 months Euribor rate and the 3 months Eurepo rate illustrated in Figure A10c, both reported by the European Banking Federation. The Euribor rate refers to Euro Interbank Offered Rate which determines, similarly as Euro Libor, the cost of short-term loans on the Euro-zone interbank market. The rate is published daily by European Bank Federation. The concept for the Euribor is similar to Euro Libor. Euribor is an average rate at which the biggest in volume European Union banks borrow funds from one another, whereas Euro Libor is determined by a group of banks which operate in London's money market. The 3 months Eurepo rate refers to a benchmark repurchase rate which is offered by one bank to another bank for Euro-denominated deals. A repurchase agreement is a lending transaction with collateral, where one party agrees to sell securities to another party with the agreement to repurchase the same securities at a specific price in the future. The Bloomberg identifiers of rates are given in Table A3 in Appendix B.1.

The European Markit iTraxx indices trade 3, 5, 7 and 10-year maturities and are determined on the basis of liquidity every six months. The benchmark Markit iTraxx Europe index comprises 125 equally-weighted credit default swaps on investment grade European corporates. We chose as a credit proxy of Euro-zone the most actively traded contract at 5Y maturity. The index is presented in Figure A10d and its Bloomberg identifier is listed in Table A3.

The daily US Dollar (USD), Great Britain Pound (GBP) and Japanese Yen (JPN) exchange against Euro are plotted in Figure A10d. We obtained the rates by converting daily exchange rates

AUDUSD, GBPUSD, USDJPY and EURUSD downloaded from Bloomberg to AUDEUR, GBPEUR, USDEUR and JPYEUR. The identifiers of instruments are listed in Table A3.

B.2. Supplementary Material to Section 11.4

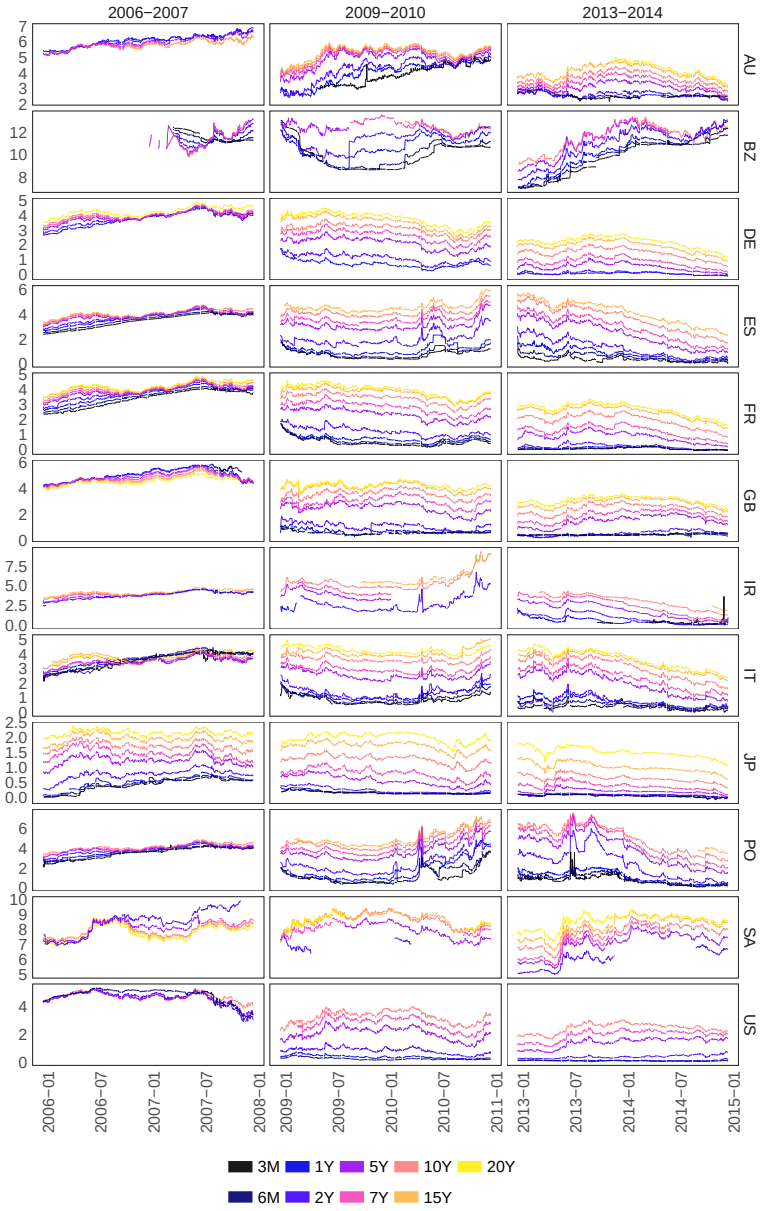


Figure A11. The yield curves of 12 countries (row-wise) from the data set **D2** for selected years (column-wise) across different maturities over time. The colors of lines correspond to the tenors and are listed in the legend.

B.3. Model Fit Statistics

In the following section we provide the supplementary materials to the result of in-sample and out-of-sample analysis from Section 11.5.

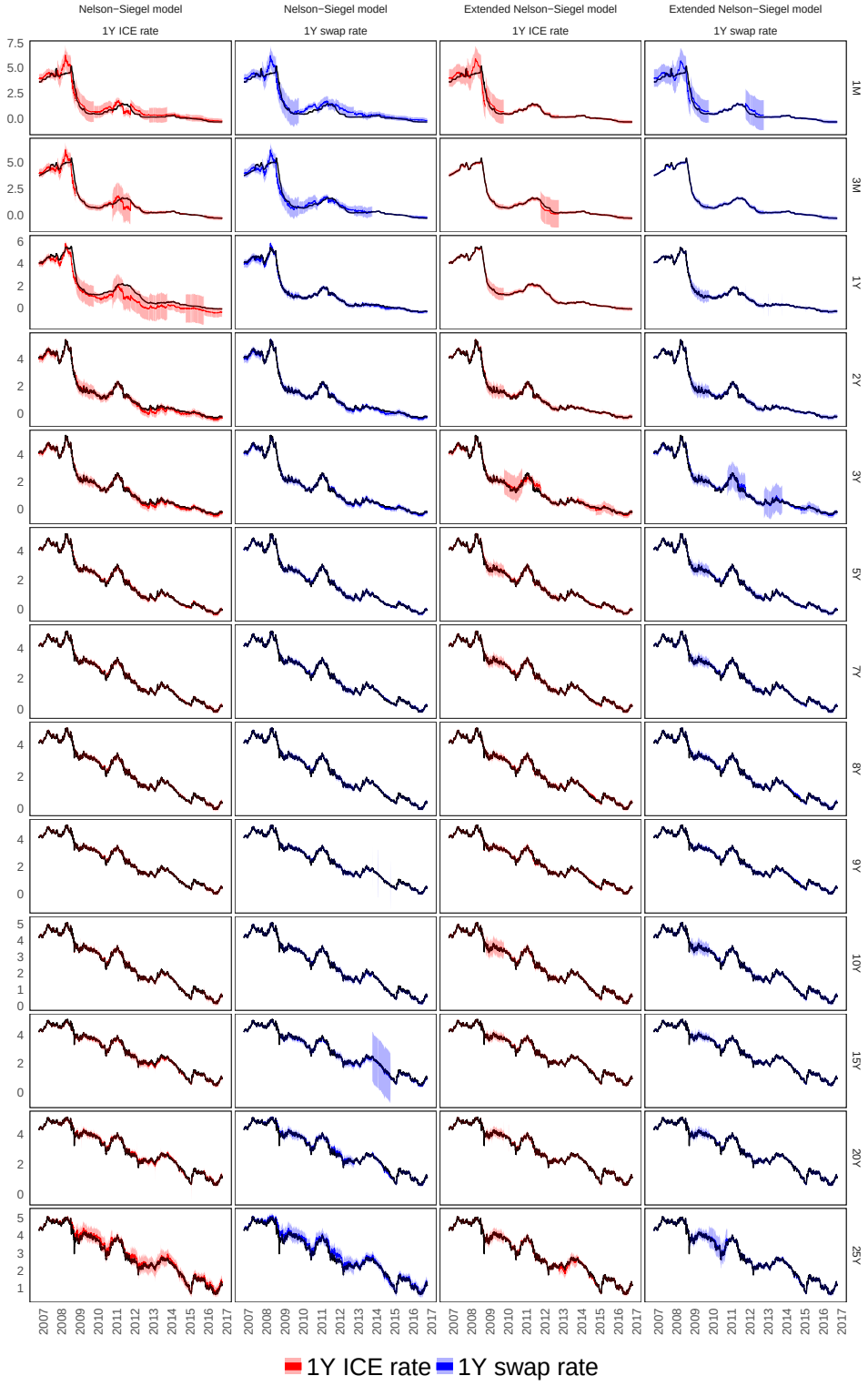


Figure A12. The Euro Libor yield curve (black line) with the one step ahead out-of-sample prediction given by the conditional mean $\mathbf{y}_{T+s+1|T+s}$ with corresponding 97.5% confidence intervals over examined period. The column-wise order of panels correspond to different classes of models, the Nelson-Siegel or Extended Nelson-Siegel model, calibration on two data sets with different 1Y rate (colors of lines and shading). The row-wise order corresponds to the evolution of rates at different maturities over time.

Table A4. The list of yearly features from the data sets **D2-D3** which are chosen as exogenous variables in the calibration of the Extended Nelson-Siegel model introduced in Section 11.3. The choice of the features follows the procedures described in the **Step 3**, Subsection 11.5.1. The analysis of the pairwise correlation between the yearly features is discussed in Subsection 11.4.4. The first column refers to the abbreviations of the data sets outlined in the introduction to the Section 11.4. The second column of the table list all employed PPCA methodologies: the two tStudent PPCA discussed in Section 3.2 and Gaussian PPCA from Section 2.4. The symbols '✓' indicates that the corresponding feature (data set, PPCA methodology and year) was used in the study.

Set	PPCA	Year										
		2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
D2	Gaussian	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	t-Student IND	✓	✓		✓	✓		✓		✓	✓	
	t-Student IID	✓	✓						✓			
D3	Gaussian	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓
	t-Student IND		✓			✓	✓					✓
	t-Student IID			✓	✓		✓				✓	
D3b	Gaussian	✓	✓	✓	✓		✓		✓		✓	✓
	t-Student IND	✓							✓		✓	
	t-Student IID		✓	✓					✓			
D3s	Gaussian	✓	✓		✓		✓	✓	✓	✓		✓
	t-Student IND							✓		✓	✓	
	t-Student IND											

Table A5. The comparison of the optimal models selected using two employed strategies - by minimum AIC (minAIC) and the lowest complexity (diagAll) for the Euro Libor yield with **1Y swap rate**. The first 22 correspond to the description of the yearly (the first column) in-sample and out-of-sample fit results for the Extended Nelson-Siegel model and the latter to the standard Nelson-Siegel model. The second column, **Crit.**, indicates which selection strategy has been used to decide the assumptions about the static parameters of the models. The columns 3-6 provide the information about the chosen assumptions: **Case** - gives information which matrices are assumed to be diagonal according to the cases in Subsection 11.3.4; **Q**- provides information which type of covariance matrix for the observation error term is chosen, heterogeneous (hete) or homogeneous (homo); **R**- provides information which type of covariance matrix for the state equation error term is chosen, heterogeneous (hete), homogeneous (homo) or homogeneous per group (homoP); $\log \lambda$ - provides the value of the shape parameters of the Nelson-Siegel model. The columns 7-10 present the yearly in-sample model fit statistic: **AIC** - Akaike Information Criterion of the models, **BIC** - Bayesian Information Criterion of the models, **MSE** - the logarithm of the mean square errors of the in-sample predictions $\mathbf{y}_{t|t-1}$ obtained by the Kalman Filter, **L** - the logarithm of the Kalman Filter likelihood. The columns 11-12 show the forecasting performance of the models measured by the mean square error of prediction (MSEP) being an averaged squared residuals between the out-of-sample prediction provided by the Kalman Filter, $\mathbf{y}_{t|t-1}$ and observed values: **1D** - the logarithm of 1 day MSEP, **1W** - the logarithm of 1 week MSEP, **1M** - the logarithm of 1 week MSEP. The last two columns refer to the information about the optimal factors chosen for Extended Nelson-Siegel model: **PPCA** - the PPCA methodology used to obtain the factors from data sets, 1- Gaussian, 2- t-Student IND, 3 - t-Student IID; **Model** - the acronyms of the examined Extended Nelson-Siegel models listed in Subsection 11.5.1.

	Year	Crit.	Static Params				In-sample				Out-of-sample			Factors	
			Case	Q	R	$\log \lambda$	AIC	BIC	MSE	L	1D	1W	1M	PPCA	Model
Extended Nelson-Siegel model	2006	diagAll	1	hete	hete	-1.42	-16200.80	-16197.25	-4.36	8101.40	-6.20	-4.96	-3.59	2	M2
		minAIC	2	hete	hete	-2.31	-17898.52	-17894.97	-5.10	8950.26	-6.35	-4.77	-2.74	1	M2
	2007	diagAll	1	hete	hete	-0.96	-14227.19	-14223.63	-3.99	7114.59	-4.92	-3.42	-1.94	1	M2
		minAIC*	4	homo	hete	-0.40	-8220.76	-8217.20	-6.12	4111.38	-4.39	-3.27	-1.75	3	M2
	2008	diagAll	1	hete	hete	-0.48	-10153.47	-10149.90	-3.81	5077.73	-5.46	-3.56	-2.05	3	M3
		minAIC	1	hete	hete	-0.48	-10153.47	-10149.90	-3.81	5077.73	-5.46	-3.56	-2.05	3	M3
	2009	diagAll	1	hete	hete	-0.48	-14214.22	-14210.65	-5.77	7108.11	-5.66	-3.16	-1.53	2	M2
		minAIC*	1	homo	hete	-0.40	-11494.12	-11490.55	-5.73	5748.06	-5.19	-2.47	-0.66	2	M2
	2010	diagAll	1	hete	hete	-0.64	-14898.48	-14894.92	-5.81	7450.24	-6.02	-4.51	-2.68	2	M2
		minAIC*	1	homo	homoP	-2.31	-11994.02	-11990.45	-5.84	5998.01	-5.04	-4.08	-2.52	2	M2
	2011	diagAll	1	hete	hete	-0.56	-13025.47	-13021.91	-5.07	6513.74	-6.02	-3.98	-1.70	1	M3
		minAIC	1	homo	homo	-0.48	-10689.50	-10685.94	-5.61	5345.75	-5.95	-5.18	-4.16	1	M2
	2012	diagAll	1	hete	hete	-0.74	-15068.73	-15065.16	-5.76	7535.36	-7.26	-5.82	-4.17	2	M2
		minAIC	1	homo	hete	-0.64	-11603.07	-11599.51	-6.31	5802.54	-6.96	-5.79	-4.23	1	M2
	2013	diagAll	1	hete	hete	-0.74	-16255.27	-16251.71	-6.72	8128.64	-7.19	-5.88	-3.74	3	M2
		minAIC	1	hete	hete	-0.74	-16255.27	-16251.71	-6.72	8128.64	-7.19	-5.88	-3.74	3	M2
	2014	diagAll	1	hete	hete	-0.96	-17049.46	-17045.90	-6.47	8525.73	-7.01	-5.21	-3.01	2	M2
		minAIC	2	hete	homoP	-0.84	-18140.85	-18137.29	-6.41	9071.43	-7.16	-5.48	-3.32	2	M2
	2015	diagAll	1	hete	hete	-0.96	-17853.02	-17849.45	-6.67	8927.51	-7.69	-6.29	-4.38	1	M2
		minAIC	4	hete	hete	-0.96	-20499.40	-20495.83	-6.71	10250.70	-7.26	-5.15	-2.97	1	M2
	2016	diagAll	1	hete	hete	-1.09	-17814.46	-17810.90	-7.15	8908.23				1	M2
		minAIC	1	homo	hete	-1.09	-15528.66	-15525.09	-6.41	7765.33				1	M2
Nelson-Siegel model	2006	diagAll	1	hete	hete	-1.42	-12327.83	-12324.27	-4.07	6164.91	-4.01	-3.85	-3.26		
		minAIC	1	hete	hete	-1.42	-12327.83	-12324.27	-4.07	6164.91	-4.01	-3.85	-3.26		
	2007	diagAll	1	hete	hete	-0.96	-12511.59	-12508.03	-3.23	6256.79	-2.24	-2.36	-1.75		
		minAIC*	2	hete	homo	-0.96	-12709.18	-12705.63	-3.37	6355.59	-2.18	-2.33	-1.87		
	2008	diagAll	1	hete	hete	-0.48	-7160.47	-7156.90	-2.64	3581.24	-4.43	-3.89	-2.65		
		minAIC	1	hete	hete	-0.48	-7160.47	-7156.90	-2.64	3581.24	-4.43	-3.89	-2.65		
	2009	diagAll	1	hete	hete	-0.48	-11357.15	-11353.59	-4.47	5679.58	-3.80	-2.76	-1.38		
	2009	minAIC*	2	hete	hete	-0.56	-11664.50	-11660.93	-4.49	5833.25	-4.60	-3.14	-1.51		
	2010	diagAll	1	hete	hete	-0.64	-12052.63	-12049.06	-4.31	6027.31	-4.18	-3.62	-2.34		
		minAIC*	2	hete	hete	-2.31	-12145.68	-12142.11	-4.95	6073.84	-3.22	-2.95	-1.91		
	2011	diagAll*	1	hete	hete	-0.56	-10380.08	-10376.52	-4.26	5191.04	-3.49	-3.18	-2.12		
		minAIC*	2	hete	hete	-0.56	-10754.53	-10750.97	-4.29	5378.26	-3.46	-3.23	-2.39		
	2012	diagAll	1	hete	homo	-0.74	-9508.23	-9504.67	-4.21	4755.12	-4.03	-3.97	-3.03		
		minAIC	3	hete	homo	-0.74	-9518.42	-9514.86	-4.35	4760.21	-4.47	-3.86	-1.42		
	2013	diagAll	1	hete	homo	-0.96	-10788.14	-10784.57	-5.13	5395.07	-5.34	-4.87	-3.10		
		minAIC	2	hete	hete	-0.96	-11464.71	-11461.15	-5.37	5733.36	-5.44	-4.85	-2.82		
	2014	diagAll	1	hete	hete	-0.96	-11274.46	-11270.90	-5.40	5638.23	-5.83	-4.93	-2.92		
		minAIC	3	hete	hete	-0.96	-11378.09	-11374.53	-5.37	5690.04	-5.73	-4.88	-3.33		
	2015	diagAll	1	hete	homo	-0.96	-13368.97	-13365.40	-5.85	6685.48	-5.46	-5.09	-3.98		
		minAIC	4	hete	homo	-0.96	-14395.09	-14391.52	-6.06	7198.54	-5.65	-3.95	-1.54		
	2016	diagAll	1	hete	hete	-1.09	-12214.66	-12211.10	-5.54	6108.33					
		minAIC	2	hete	homo	-1.09	-12990.60	-12987.04	-5.85	6496.30					

Table A6. The comparison of the optimal models selected using two employed strategies - by minimum AIC (minAIC) and the lowest complexity (diagAll) for the Euro Libor yield with **1Y ICE rate**. The first 22 correspond to the description of the yearly (the first column) in-sample and out-of-sample fit results for the Extended Nelson-Siegel model and the latter to the standard Nelson-Siegel model. The second column, **Crit.**, indicates which selection strategy has been used to decide the assumptions about the static parameters of the models. The columns 3-6 provide the information about the chosen assumptions: **Case** - gives information which matrices are assumed to be diagonal according to the cases in Subsection 11.3.4; **Q**- provides information which type of covariance matrix for the observation error term is chosen, heterogeneous (hete) or homogeneous (homo); **R**- provides information which type of covariance matrix for the state equation error term is chosen, heterogeneous (hete), homogeneous (homo) or homogeneous per group (homoP); $\log \lambda$ - provides the value of the shape parameters of the Nelson-Siegel model. The columns 7-10 present the yearly in-sample model fit statistic: **AIC** - Akaike Information Criterion of the models, **BIC** - Bayesian Information Criterion of the models, **MSE** - the logarithm of the mean square errors of the in-sample predictions $\mathbf{y}_{t|t-1}$ obtained by the Kalman Filter, **L** - the logarithm of the Kalman Filter likelihood. The columns 11-12 show the forecasting performance of the models measured by the mean square error of prediction (MSEP) being an averaged squared residuals between the out-of-sample prediction provided by the Kalman Filter, $\mathbf{y}_{t|t-1}$ and observed values: **1D** - the logarithm of 1 day MSEP, **1W** - the logarithm of 1 week MSEP, **1M** - the logarithm of 1 week MSEP. The last two columns refer to the information about the optimal factors chosen for Extended Nelson-Siegel model: **PPCA** - the PPCA methodology used to obtain the factors from data sets, 1- Gaussian, 2- t-Student IND, 3 - t-Student IID; **Model** - the acronyms of the examined Extended Nelson-Siegel models listed in Subsection 11.5.1.

	Year	Crit.	Static Params				In-sample				Out-of-sample			Factors	
			Case	Q	R	$\log\lambda$	AIC	BIC	MSE	L	1D	1W	1M	PPCA	Model
Extended Nelson-Siegel model	2006	diagAll	1	hete	hete	-1.42	-15794.25	-15790.70	-4.30	7898.12	-6.35	-4.93	-3.41	2	M2
		minAIC	1	hete	hete	-1.42	-15794.25	-15790.70	-4.30	7898.12	-6.35	-4.93	-3.41	2	M2
	2007	diagAll	1	hete	hete	-0.76	-14609.24	-14605.70	-3.67	7305.62	-4.90	-3.33	-2.01	1	M2
		minAIC	2	hete	homoP	-0.83	-14127.90	-14124.35	-4.43	7064.95	-5.00	-3.49	-2.08	2	M2
	2008	diagAll*	1	hete	hete	-0.52	-9680.15	-9676.59	-3.98	4841.08	-5.70	-3.96	-2.57	3	M3
		minAIC	2	hete	hete	-0.52	-13679.93	-13676.37	-3.93	6840.96	-5.41	-3.47	-1.59	3	M3
	2009	diagAll	1	hete	hete	-0.52	-13527.22	-13523.65	-5.64	6764.61	-5.60	-3.07	-1.45	2	M2
		minAIC	4	hete	hete	-0.52	-16645.06	-16641.49	-4.81	8323.53	-5.32	-3.17	-1.87	3	M3
	2010	diagAll	1	hete	hete	-2.18	-15298.46	-15294.90	-6.20	7650.23	-5.52	-3.69	-2.22	2	M2
		minAIC	4	hete	hete	-2.51	-18306.79	-18303.23	-6.33	9154.40	-5.70	-4.65	-2.88	2	M2
	2011	diagAll	1	hete	hete	-0.52	-13003.63	-13000.07	-5.19	6502.81	-6.16	-4.23	-1.99	1	M3
		minAIC	2	hete	hete	-0.52	-16224.58	-16221.02	-4.99	8113.29	-5.93	-4.01	-1.79	1	M3
	2012	diagAll	1	hete	hete	-0.52	-15544.17	-15540.61	-6.23	7773.09	-7.31	-5.79	-4.13	2	M2
		minAIC	1	hete	hete	-0.52	-15544.17	-15540.61	-6.23	7773.09	-7.31	-5.79	-4.13	2	M2
	2013	diagAll	1	hete	hete	-0.83	-15706.06	-15702.49	-6.55	7854.03	-7.18	-5.74	-3.73	3	M2
		minAIC	2	hete	hete	-0.83	-18964.99	-18961.43	-6.36	9483.50	-7.39	-5.61	-3.38	1	M3
	2014	diagAll	1	hete	hete	-0.91	-16922.96	-16919.39	-6.42	8462.48	-7.05	-5.25	-3.17	2	M2
		minAIC	2	hete	hete	-0.91	-19222.55	-19218.98	-6.60	9612.27	-7.31	-6.14	-4.43	1	M3
	2015	diagAll	1	hete	hete	-0.99	-17770.05	-17766.48	-6.63	8886.02	-7.77	-6.31	-4.49	1	M2
		minAIC	2	hete	hete	-0.99	-20003.81	-20000.24	-6.63	10002.90	-7.75	-6.29	-4.47	2	M2
	2016	diagAll	1	hete	hete	-0.99	-17189.57	-17186.00	-6.98	8595.78				1	M2
		minAIC	2	hete	homoPerGroup	-0.99	-18389.86	-18386.30	-6.93	9195.93				1	M2
Nelson-Siegel model	2006	diagAll	1	hete	hete	-1.42	-11843.46	-11839.91	-3.99	5922.73	-3.93	-3.78	-2.94		
		minAIC	1	hete	hete	-1.42	-11843.46	-11839.91	-3.99	5922.73	-3.93	-3.78	-2.94		
	2007	diagAll	1	hete	homo	-0.83	-12007.74	-12004.19	-3.25	6004.87	-2.07	-2.20	-1.82		
		minAIC	2	hete	homo	-0.83	-12426.62	-12423.07	-3.29	6214.31	-1.91	-2.06	-1.86		
	2008	diagAll*	1	hete	hete	-0.52	-6660.50	-6656.94	-3.03	3331.25	-4.10	-3.74	-2.81		
		minAIC	2	hete	hete	-0.52	-6967.12	-6963.56	-3.09	3484.56	-4.03	-3.27	-1.59		
	2009	diagAll	1	hete	hete	-0.52	-10367.44	-10363.88	-4.27	5184.72	-3.77	-2.77	-1.26		
		minAIC	2	hete	hete	-0.52	-10536.63	-10533.06	-4.28	5269.31	-3.70	-2.75	-1.33		
	2010	diagAll	1	hete	hete	-2.18	-10842.19	-10838.62	-3.66	5422.09	-2.29	-2.29	-1.78		
		minAIC	4	hete	hete	-2.51	-11453.43	-11449.87	-3.83	5727.72	-2.44	-2.37	-1.87		
	2011	diagAll	1	hete	hete	-0.52	-8712.66	-8709.10	-3.40	4357.33	-2.90	-2.90	-2.23		
		minAIC	2	hete	hete	-0.52	-9677.18	-9673.62	-3.39	4839.59	-2.88	-2.88	-2.49		
	2012	diagAll	1	hete	hete	-0.52	-8734.87	-8731.31	-2.90	4368.44	-3.86	-3.69	-3.03		
		minAIC	1	hete	hete	-0.52	-8734.87	-8731.31	-2.90	4368.44	-3.86	-3.69	-3.03		
	2013	diagAll	1	hete	hete	-0.83	-11474.72	-11471.15	-4.69	5738.36	-4.30	-4.24	-2.97		
		minAIC	1	hete	hete	-0.83	-11474.72	-11471.15	-4.69	5738.36	-4.30	-4.24	-2.97		
	2014	diagAll	1	hete	hete	-0.91	-10448.68	-10445.12	-4.43	5225.34	-4.86	-4.42	-2.92		
		minAIC	2	hete	hete	-0.91	-10487.17	-10483.60	-4.49	5244.58	-4.91	-4.48	-3.05		
	2015	diagAll	1	hete	homo	-0.99	-12733.24	-12729.67	-4.94	6367.62	-4.37	-4.28	-3.61		
		minAIC	2	hete	homo	-0.99	-13518.68	-13515.11	-5.08	6760.34	-4.53	-4.43	-3.63		
	2016	diagAll	1	hete	hete	-0.99	-10298.73	-10295.16	-4.69	5150.36					
		minAIC	4	hete	homo	-0.99	-11304.45	-11300.89	-4.71	5653.23					

C. Supplementary Material to Chapter 12

C.1. Observed Patterns of Missingness

Missing data appear when no value is available for a component of the observation vector. The Probabilistic Principal Component Analysis with missing data discussed in Chapter 4 handles missing data by filling them with projection using principal components that are calculated from available information. Thus, the more missing data is present for a given variable, the less impact this variable has on the specification of the projection over the assumptions. The results for variables with missing data are more influenced by our assumption PPCA framework we want to use, as will be discussed in detail below.

In order to handle missing values we need to understand and study how their pattern. In the following part we demonstrate the three patterns of missingness which are present in the analysed data sets:

Type 1: no information about a variable in a few observations for a given country;
Type 2: no information about a variable in all observations for a given country;
Type 3: general unavailability of information about all variables for a country except for a limited set of observations.

The analysis is conducted separately for four data sets and sexes. We show the results for four cases where we segment the data among four proportions of missing entries per total observation with maximal percentage of missing entries equal to: 0%, 25%, 50% and 75%. For instance, a single observation of Birth counts is 31 dimensional. When we analyse the data set with respect to the case 50%, we exclude all observation, when the number of missing entries per observation is greater than $0.5 \times 31 \approx 15$.

In the case of no missing data, that is 0% for Death counts and Death Rates in Females, the number of rows without missing entries is too small for any significant analysis. Due to this fact we drop the minimal number of columns which have the highest number of missing entries in order to collect significant sample for our analysis.

The results of the analysis are similar among Birth counts and Life Expectancy at Birth and Death counts and Death Rates. Hence we discuss the patterns of missingness in Number of Births and Number of Deaths.

C.1.1. Births

The left plot of Figure A1 shows the percentage of missing entries per observation vector of total births over all countries considered versus calendar years for Births counts disaggregated for Females and Males. The sample starts in 1751 (Swedish data) and spans to 2014. Until 1946, the sample has a percentage of missing entries above 50% (the middle vertical red line on the plot). The sample of case 25% starts in 1950. We observe the same missing values pattern between female

and male population (the corresponding lines for populations overlap and only the black line, corresponding to male population, is visible).

Figure A2 indicates the availability of data per country (y axis) versus the calendar year (x axis). The black colour denotes points in time when the data is not available for a given country. The Swedish data is not labelled by any black entry except in 2014 as it is the longest time series. Recall that the missing data pattern which is characteristics for this data set is a limited availability of data for a given country. From the empirical analysis of the data we learn that the missing entries do not appear randomly. However, it does not violate the assumed behaviour of missingness and we can still proceed with the methodology of feature extraction described in the previous parts of the paper.

The red vertical lines correspond to the starting points of the subsamples when the maximum number of missing entries per row (an observation in time) is equal to (from the left side on corresponding plots) 75%, 50% and 25%. The subsamples for the cases 25% and 0% provide with principal components which are determined by almost equal sizes of information from every country included into the analysis for the examined period of time. However, the calculation of the components for the subsample in the case 75% is dominated by the time series of countries which are not available before 1950. Therefore, the calculated features are more prone to be impacted by the distribution assumptions and convey less information about dynamics present in the examined period of time that the components obtained from longer time series.

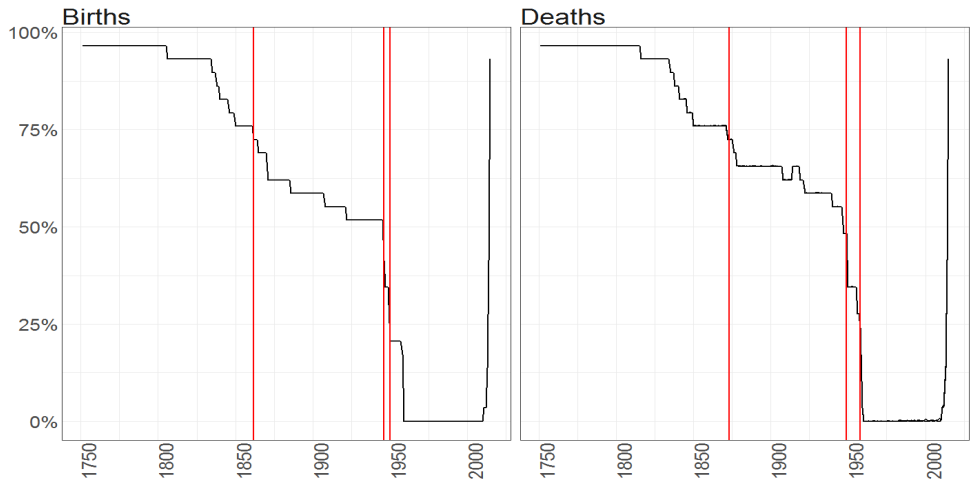


Figure A1. The percentage of missing entries (y axis) per observation vector over time (x axis) for the Births counts (left plot) and Deaths counts (right plot) for female (blue line) and male (black line) population. Red vertical lines correspond to the starting points in time of samples when maximum of missing entries is equal to (from the left side on corresponding plots) 75%, 50% and 25%.

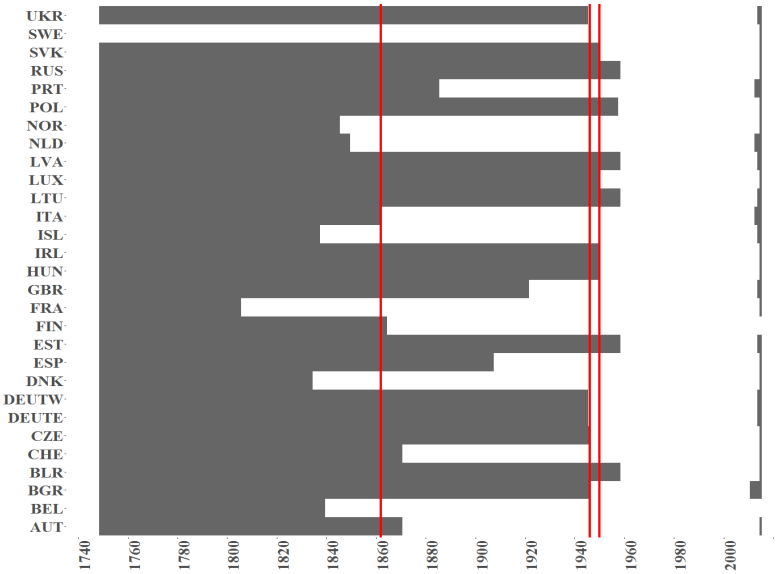


Figure A2. The indicator of a missing value (black colour) per country (y axis) over time (x axis) for the Births counts for female population. Red vertical lines correspond to the starting points in time when samples with maximum of missing entries is equal to (from the left side on corresponding plots) 75%, 50% and 25%.

C.1.2. Deaths

The single observation of Death counts is a 504 dimensional vector which reflect the numbers of death per country and age group in time. The proportion of non missing entries per calendar year, again disaggregated by gender for Death counts is shown on the right plot in Figure A1. The percentage of missing entries decreases slower than for Births counts (the slope of the curve is flattener) which indicates the longer distance between the cases 50% and 25%. We notice the discrepancies in the patterns of missing values between female and male population. Interestingly, the female population data has no observations without missing values.

It is also informative to demonstrate the pattern of missing values per age group, as displayed in Figures A3 and A4 below. Red vertical lines correspond to the starting points in calendar time for the total proportions of missing data corresponding to the cases 50% and 25%. We observe a new pattern of missing values: particular variables have a few missing observations within the subsample. The pattern occurs mainly after 1950 in age groups between 1 and 25 (darker shade of blue for single observations) or in age group 95–100. The other interesting analysis can be done when we display the patterns of missingness which are present by age and per country, which is shown in Figures A5 and A6, again disaggregated by gender. The pattern is present only for the subset of countries. For instance, data related to the Irish population has higher percentage of missing values among only four age groups. Still, the dominant pattern of missingness is the availability of the information for a country which is limited to some period of time.

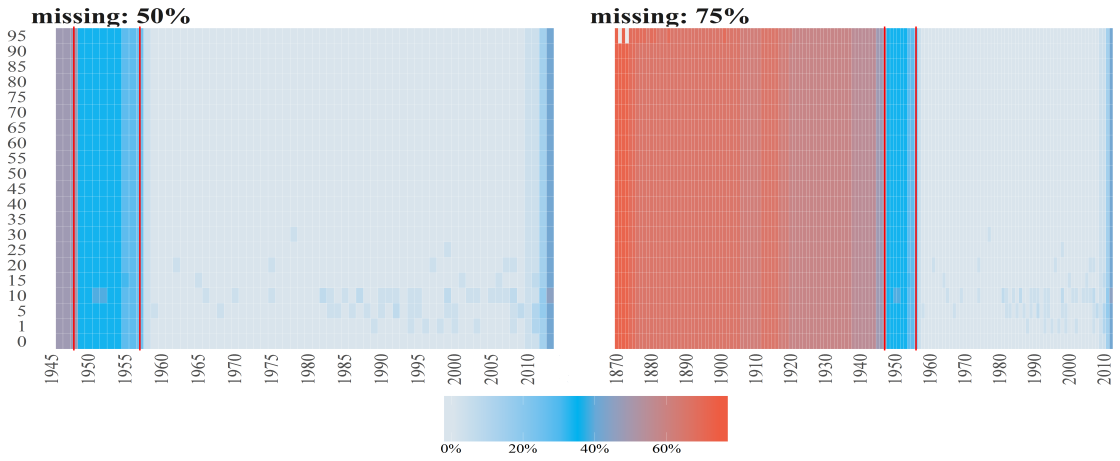


Figure A3. Percentages of missing values (denoted by different colours) per observation for Number of Deaths for Females per age groups (y axis) over time (x axis). The titles of the subplots indicate the case of missing values (50%, 75%). The percentage for a given country and given age group is computed dividing number of missing values by number of countries. Red vertical lines correspond to the starting points in time when the cases 50% and 25% start (from the left to right side on corresponding plots).

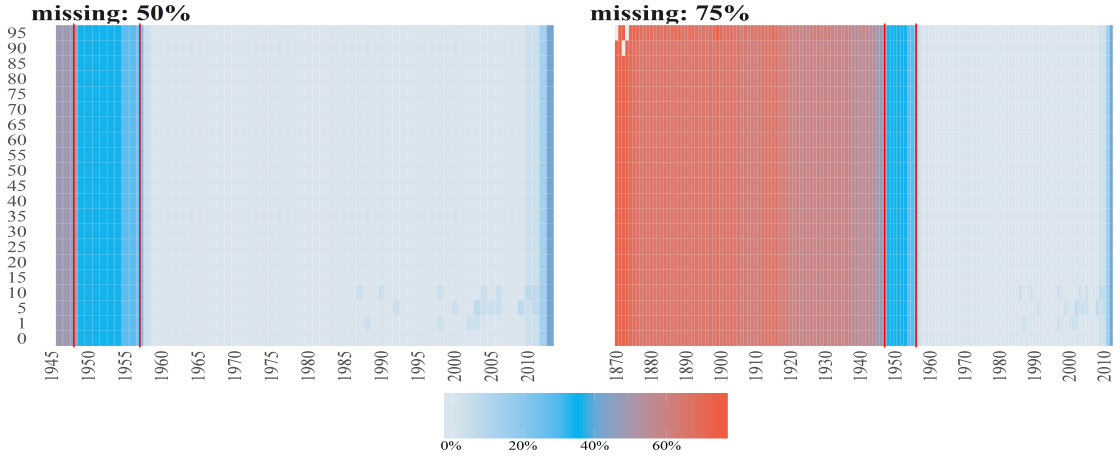


Figure A4. Percentages of missing values (denoted by different colours) per observation for Number of Deaths for Males per age groups (y axis) over time (x axis). The titles of the subplots indicate the case of missing values (50%, 75%). The percentage for a given country and given age group is computed dividing number of missing values by number of countries. Red vertical lines correspond to the starting points in time when the cases 50% and 25% starts (from the left to right side on corresponding plots).

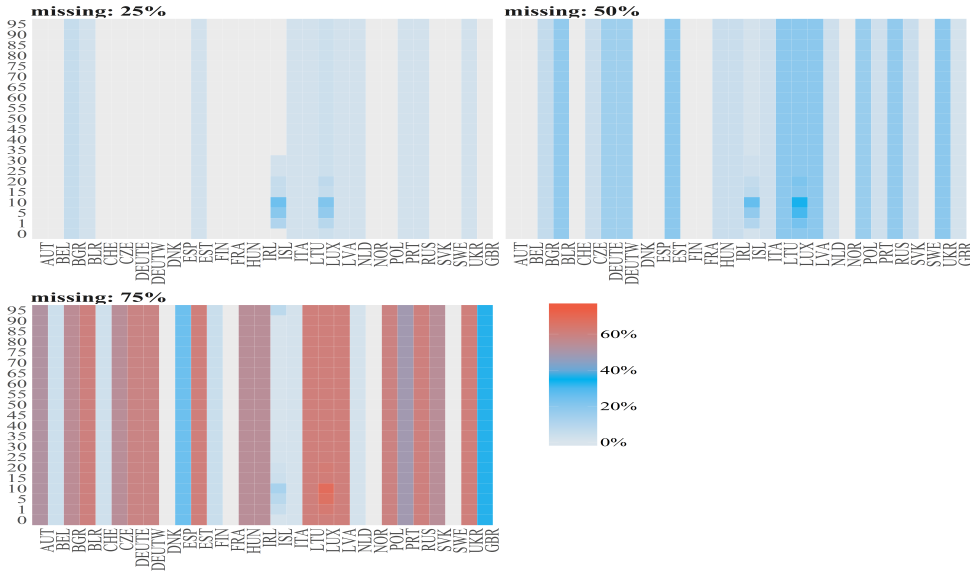


Figure A5. Percentages of missing values (denoted by different colours) for Number of Deaths for Females per country (x axis) and age group (y axis). The titles of the subplots indicate the case of missing values (25%, 50%, 75%). The percentage of missing values for a given country and an age group is calculated dividing number of missing values by length of subsample which is different for different cases.

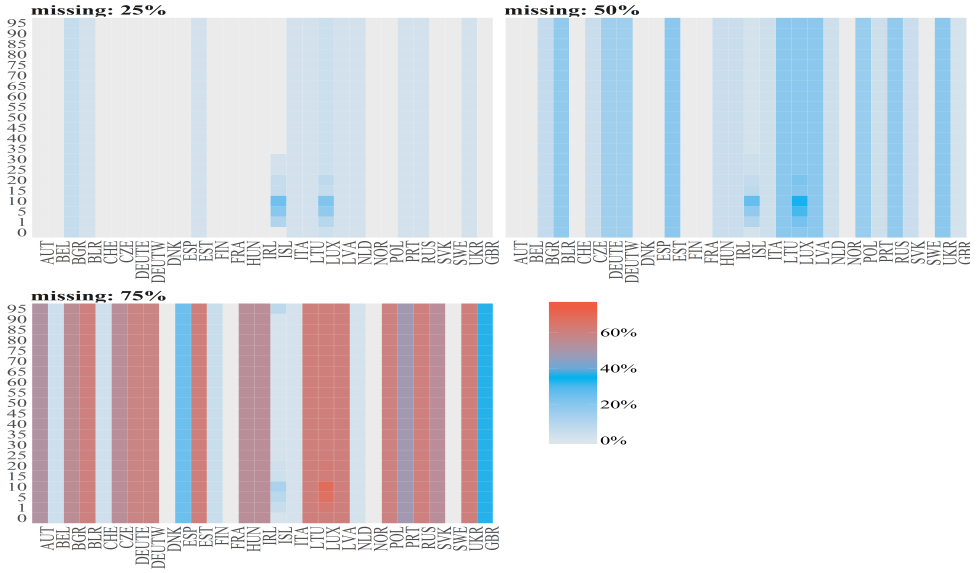


Figure A6. Percentages of missing values (denoted by different colours) for Number of Deaths for Males per country (x axis) and age group (y axis). The titles of the subplots indicate the case of missing values (25%, 50%, 75%). The percentage of missing values for a given country and an age group is calculated dividing number of missing values by length of subsample which is different for different cases.

C.2. Description of Stochastic Mortality Models Utilizing Factor Extraction from European Demographic Data

Recalling the notation from the Equation (12.8), all models which utilize features extracted from Demographic data have the following the state-space representation

$$\boldsymbol{y}_t = \boldsymbol{\alpha} + \tilde{\mathbf{B}}_t \tilde{\boldsymbol{\varphi}}_t + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \overset{iid}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2 \mathbb{I}_{21}), \tag{A1a}$$

$$\tilde{\boldsymbol{\varphi}}_t = \tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\varphi}}_{t-1} + \tilde{\boldsymbol{\Theta}} + \tilde{\boldsymbol{\omega}}_t, \quad \tilde{\boldsymbol{\kappa}}_t \overset{iid}{\sim} \mathcal{N}(0, \tilde{\Upsilon}). \tag{A1b}$$

As we examine only **Case 1** defined in Section 12.3, the corresponding transition matrices of the observation and state equations are equal to

$$\begin{aligned} \tilde{\mathbf{B}}_{t\ 21\times(22+m)} &= \left(\begin{array}{c|c} \mathbf{B}_{21\times 22} & \tilde{\mathbf{F}}_t \end{array} \right) \\ \tilde{\boldsymbol{\Lambda}}_{(22+m)\times(22+m)} &= \left(\begin{array}{c|c} \boldsymbol{\Lambda}_{22\times 22} & \mathbf{0}_{22\times m} \\ \hline \mathbf{0}_{m\times 22} & \boldsymbol{\Omega}_{m\times m} \end{array} \right) \end{aligned} \tag{A2}$$

where m is the dimensionality of the latent process vector $\boldsymbol{\varrho}_t$ which corresponds to the factor matrix $\tilde{\mathbf{F}}_t$. The structure of the matrix depends on the models what is further discussed in the next two subsections.

C.2.1. DFM-PC-B Model

The model is constructed as follow

Step 1: Take the first eigenvector of robustly standardized Birth counts which is of vector-type, country specific

$$[\mathbf{F}]_{\cdot,1} = \begin{bmatrix} f_{1,1}^{AUT} \\ \vdots \\ f_{21,1}^{UKR} \end{bmatrix};$$

Step 2: Take the mean across the countries (components of the first eigenvector vector) which is a scalar $\hat{f}$ and use a one per age group latent process to model additional supplementary information per age group;

Step 3: Incorporate $\hat{f}$ as elements of the diagonal matrix $\tilde{\mathbf{F}}$

$$\tilde{\mathbf{F}} = \begin{bmatrix} \hat{f} & 0 & 0 & \cdots & 0 \\ 0 & \hat{f} & 0 & \cdots & 0 \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \hat{f} \end{bmatrix}_{21\times 21}$$

The corresponding $\boldsymbol{\varrho}_t$ is age group specific vector, that is

$$\boldsymbol{\varrho}_t = \left(\varrho_t^0, \ldots, \varrho_t^{95}\right)_{1\times 21}$$

C.2.2. The Models of the Class DFM-PC-D and DFM-PC-Mx

The models are constructed as follow

Step 1: Take the first eigenvector of robustly standardized corresponding data set which is of matrix-type, age and country specific

$$\mathbf{F} = \begin{bmatrix} f_{1,1}^{AUT} & \cdots & f_{1,28}^{UKR} \\ \vdots & \ddots & \vdots \\ f_{21,1}^{AUT} & \cdots & f_{21,28}^{UKR} \end{bmatrix}_{21\times 28}$$

Step 2: Notice that the matrix $\tilde{\mathbf{F}}$ is equal to $\mathbf{F}$

Step 3: Use a one per country latent process to model the impact of country specific vector;

The corresponding $\boldsymbol{\varrho}_t$ is country specific vector, that is

$$\boldsymbol{\varrho}_t = \left(\varrho_t^{AUT}, \ldots, \varrho_t^{UKR}\right)_{1\times 28}$$