

Advances in approximate Bayesian computation and trans-dimensional sampling methodology

by

Gareth William Peters

B.Sc., University of Melbourne

B.Eng (Hons. 1st), University of Melbourne

M.Sc. (by research), Cambridge University

Ph.D. research thesis by publication.

Department of Mathematics and Statistics,
University of NSW.

December 2009



UNSW
THE UNIVERSITY OF NEW SOUTH WALES

Originality

I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgment is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.

Signature:

Date:

Copyright

I hereby grant the University of New South Wales or its agents the right to archive and to make available my thesis or dissertation in whole or part in the University libraries in all forms of media, now or here after known, subject to the provisions of the Copyright Act 1968. I retain all proprietary rights, such as patent rights. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation. I also authorise University Microfilms to use the 350 word abstract of my thesis in Dissertation Abstract International. I have either used no substantial portions of copyright material in my thesis or I have obtained permission to use copyright material; where permission has not been granted I have applied/will apply for a partial restriction of the digital copy of my thesis or dissertation.

Signature:

Date:

Authenticity

I certify that the Library deposit digital copy is a direct equivalent of the final officially approved version of my thesis. No emendation of content has occurred and if there are any minor variations in formatting, they are the result of the conversion to digital format.

Signature: _____

Date: _____

Acknowledgements

During the course of this thesis I have met and interacted with many interesting people, who influenced the path of my research.

I would first like to thank my academic supervisors Dr. Scott Sisson, Dr. Yanan Fan and Dr. Pavel Shevchenko. Their support over the last few years and the academic freedom they afforded me to explore my ideas has been invaluable. Their depth of knowledge, friendship, encouragement, technical insights and constructive criticism have taught me a great deal. They have directly influenced my academic development and knowledge in so many aspects, and for this I am extremely thankful. I would also like to thank them for the numerous instances of financial support that enabled me to give my research work full attention and travel to international conferences.

I would like to thank the following organisations and people for providing financial support throughout my thesis: APA - UNSW, research top up scholarship Mathematics and Statistics UNSW, CSIRO travel and PhD top up, Commonwealth Bank, Boronia Capital hedge fund, ETH RiskLab Zurich, SAMSI North Carolina US, and ISM in Tokyo Japan.

I have had the great pleasure to work with a number of people from whom I have learnt a great deal. To begin with I would like to make special mention of four collaborators and good friends from whom I have especially benefited. Their depth of knowledge and willingness to impart and share with me their thoughts and ideas has resulted in many positive outcomes for me and for this I am very thankful. The people to whom I refer are Dr. Mario Wüthrich of RiskLab ETH, Dr. Ido Nevat of Electrical Engineering at University of NSW, Dr. Arnaud Doucet of the University of British Columbia, Canada and the Institute of Statistical Mathematics, Tokyo, Japan and Dr. Mark Briers of QintQ. Mario's knowledge of Actuarial modelling, probability and risk theory is immense, as is Ido's knowledge of Signal processing and Telecommunications engineering. Arnaud and Mark have both taught me a great deal about filtering and smoothing methodology, including many interesting discussions over the years.

Other colleagues and friends I would like to thank, in chronological order, include: Dr. Pierre del Moral of the University of INRIA, France; Dr. Adam Johansen of Warwick University; Matthew Delasey and David Farmer at Commonwealth Bank of Australia; Dr. Xiaolin Lou and

John Donnelly at CMIS in the Commonwealth Scientific Industrial Research Organisation; Dr Paul Embrechts of RiskLab ETH; Dr. Kannan Balakrishnan and Dr. Ben Lasscock of Boronia Capital hedge fund and Dr. Jinhong Yuan of Electrical Engineering at the University of NSW.

I would also like to thank all my friends and colleagues at various places with whom I have spent time, thanks goes to: Adam, Ainslie, Chen, Chris, Dale, Jie, Julien, Karan, Roman.

Thank you also goes to UNSW staff, CSIRO staff, ETH Staff and SAMSI staff for their help with any technical or administrative issues.

Finally, I would like to thank very special people in my life for their unwavering support throughout the years of this thesis. They have been a constant source of strength and encouragement for me, they include Joanna, my mother and my family.

Publications, proceedings and reports

The following journal publications and conference proceedings were all peer reviewed and completed during the PhD candidature.

Journal Papers - peer reviewed, accepted:

1. **Peters G.W.**, Nevat I., Sisson S.A., Fan Y. and Yuan J. (2010) "Bayesian Symbol Detection in Wireless Relay Networks via Likelihood Free Inference". *IEEE Transactions on Signal Processing*, 58, 5206-5218.
2. **Peters G.W.**, Balkrishnan K. and Lasscock B. (2010) "Model selection and Adaptive Markov Chain Monte Carlo for Bayesian Cointegrated VAR Models". *Bayesian Analysis*, to appear.
3. **Peters G.W.**, Sisson S.A. and Fan Y. (2009) "Likelihood-free Bayesian inference for α -stable models". *Computational Statistics and Data Analysis*, to appear.
4. Cornebise J. and **Peters G.W.** (2009) "Comments on 'Particle Markov Chain Monte Carlo'". *Journal of the Royal Statistical Society Series B - comments on read paper*, to appear.
5. Nevat I., **Peters G.W.** and Yuan J. (2009). "Detection of Gaussian Constellations in MIMO Systems Under Imperfect CSI". *IEEE Transactions of Communications*, to appear.
6. **Peters G.W.**, Shevchenko P. and Wüthrich (2009). "Dynamic Operational Risk: modelling dependence and combining different sources of information". *Journal of Operational Risk*, 4(2), 69-104.
7. **Peters G.W.**, Shevchenko P. and Wüthrich M. (2009) "Model Uncertainty in Claims Reserving within Tweedie's Compound Poisson Models". *ASTIN Bulletin* 39(1), 1-33.
8. **Peters G.W.**, Wüthrich M. and Shevchenko P. (2009) "Chain Ladder Method: Bayesian Bootstrap versus Classical Bootstrap". *Insurance: Mathematics and Economics*, conditionally accepted with revision.
9. Fan Y., **Peters G.W.** and Sisson S.A (2009) "Automating and Evaluating Reversible Jump MCMC Proposal Distributions". *Statistics and Computing*, 19, 401-429.

10. **Peters G.W.**, Nevat I. and Yuan J. (2009). "Channel Estimation in OFDM Systems with Unknown Power Delay Profile using Trans-dimensional MCMC". *IEEE Transactions on Signal Processing, IEEE Trans. on Signal Processing*, 57(9), 3545-3561.
11. Nevat I., **Peters G.W.** and Yuan J. (2008) "A Low Complexity MAP Estimation in Linear Models with a Random Gaussian Mixing Matrix". *IEEE Transactions on Communications*, to appear.
12. **Peters G.W.**, Johansen A. M. and Doucet A. (2007) "Simulation of the Annual Loss Distribution in Operational Risk via Panjer Recursions and Volterra Integral Equations for Value at Risk and Expected Shortfall Estimation". *Journal of Operational Risk*, 2(3).
13. **Peters G.W.** and Sisson S.A. (2006) "Bayesian Inference, Monte Carlo Sampling and Operational Risk". *Journal of Operational Risk*, 1(3).

Conference Publications - peer reviewed proceedings, accepted:

1. Nevat I., **Peters G.W.** and Yuan J. (2009) "Coherent Detection or Cooperative Networks with Arbitrary Relay Functions using 'Likelihood Free' Inference". *Proceedings of NEWCOM-ACorn Workshop*, Barcelona, Spain.
2. Nevat I., **Peters G.W.** and Yuan J. (2009) "Channel Estimation in OFDM Systems with Unknown Power Delay Profile using Trans-dimensional MCMC via Stochastic Approximation". in *Proc. IEEE Vehicular Technology Conference, VTC09*, Barcelona, Spain.
3. **Peters G.W.**, Shevchenko P. and Wüthrich (2009) "Dynamic Operational Risk: modelling dependence and combining different sources of information". *15th International Conference on Computing in Economics and Finance*.
4. **Peters G.W.**, Kannan B., Lasscock B. and Mellen C. (2009) "Rank Estimation and Adaptive Markov chain Monte Carlo for Bayesian Cointegrated VAR Models". *15th International Conference on Computing in Economics and Finance*.
5. **Peters G.W.**, Shevchenko P. and Wüthrich M. (2008) "Model Risk in Claims Reserving within Tweedie's Compound Poisson Models". *Astin Colloquium*, UK.
6. Nevat I., **Peters G.W.** and Yuan J. (2008) "Bayesian Inference in Linear Models With a Random Gaussian Matrix : Algorithms and Complexity". *PIMRC*, France.
7. Nevat I., **Peters G.W.** and Yuan J. (2008) "Maximum A-Posteriori Estimation in Linear Models With a Random Gaussian Model Matrix: a Bayesian-EM Approach". *ICASSP*, Las Vegas, USA.
8. Nevat I., **Peters G.W.** and Yuan J. (2008) "OFDM Channel Impulse Response Estimation with Unknown Length using Bayesian Model Order Selection and Model Averaging". *VTC*, Singapore.

Journal Papers - peer reviewed, submitted:

1. **Peters G.W.**, Fan Y. and Sisson S.A. (2009) "On Sequential Monte Carlo, Partial Rejection Control and Approximate Bayesian Computation". *In review*.
2. Sisson S.A., **Peters G.W.**, Fan Y. and Briers, M. (2009) "Likelihood Free Samplers".

"Out of intense complexities intense simplicities emerge"

Sir Winston Churchill

*Dedicated to two very special people in my life:
my mother who always supported and encouraged me
- among many things you taught me the joy of
scientific enquiry;
to Joanna - your love, support, encouragement and
friendship over all these years has made this possible.*

Abstract

Bayesian statistical models continue to grow in complexity, driven in part by a few key factors: the massive computational resources now available to statisticians; the substantial gains made in sampling methodology and algorithms such as Markov chain Monte Carlo (MCMC), trans-dimensional MCMC (TDMCMC), sequential Monte Carlo (SMC), adaptive algorithms and stochastic approximation methods and approximate Bayesian computation (ABC); and development of more realistic models for real world phenomena as demonstrated in this thesis for financial models and telecommunications engineering. Sophisticated statistical models are increasingly proposed for practical solutions to real world problems in order to better capture salient features of increasingly more complex data. With sophistication comes a parallel requirement for more advanced and automated statistical computational methodologies.

The key focus of this thesis revolves around innovation related to the following three significant Bayesian research questions.

- How can one develop practically useful Bayesian models and corresponding computationally efficient sampling methodology, when the likelihood model is intractable?
- How can one develop methodology in order to automate Markov chain Monte Carlo sampling approaches to efficiently explore the support of a posterior distribution, defined across multiple Bayesian statistical models?
- How can these sophisticated Bayesian modelling frameworks and sampling methodologies be utilized to solve practically relevant and important problems in the research fields of financial risk modeling and telecommunications engineering ?

This thesis is split into three bodies of work represented in three parts. Each part contains journal papers with novel statistical model and sampling methodological development. The coherent link between each part involves the novel sampling methodologies developed in Part I and utilized in Part II and Part III. Papers contained in each part make progress at addressing the core research questions posed.

Part I of this thesis presents generally applicable key statistical sampling methodologies that will be utilized and extended in the subsequent two parts. In particular it presents novel

developments in statistical methodology pertaining to likelihood-free or ABC and TDMCMC methodology. The TDMCMC methodology focuses on several aspects of automation in the between model proposal construction, including approximation of the optimal between model proposal kernel via a conditional path sampling density estimator. Then this methodology is explored for several novel Bayesian model selection applications including cointegrated vector autoregressions (CVAR) models and mixture models in which there is an unknown number of mixture components. The second area relates to development of ABC methodology with particular focus on SMC Samplers methodology in an ABC context via Partial Rejection Control (PRC). In addition to novel algorithmic development, key theoretical properties are also studied for the classes of algorithms developed. Then this methodology is developed for a highly challenging practically significant application relating to multivariate Bayesian α -stable models.

Then Part II focuses on novel statistical model development in the areas of financial risk and non-life insurance claims reserving. In each of the papers in this part the focus is on two aspects: foremost the development of novel statistical models to improve the modeling of risk and insurance; and then the associated problem of how to fit and sample from such statistical models efficiently. In particular novel statistical models are developed for Operational Risk (OpRisk) under a Loss Distributional Approach (LDA) and for claims reserving in Actuarial non-life insurance modelling. In each case the models developed include an additional level of complexity which adds flexibility to the model in order to better capture salient features observed in real data. The consequence of the additional complexity comes at the cost that standard fitting and sampling methodologies are generally not applicable, as a result one is required to develop and apply the methodology from Part I.

Part III focuses on novel statistical model development in the area of statistical signal processing for wireless communications engineering. Statistical models will be developed or extended for two general classes of wireless communications problem: the first relates to detection of transmitted symbols and joint channel estimation in Multiple Input Multiple Output (MIMO) systems coupled with Orthogonal Frequency Division Multiplexing (OFDM); the second relates to co-operative wireless communications relay systems in which the key focus is on detection of transmitted symbols. Both these areas will require advanced sampling methodology developed in Part I to find solutions to these real world engineering problems.

Contents

| | |
|---|----|
| 1. <i>Motivation and structure</i> | 1 |
| <i>Part I Advances in ABC and TDMCMC methodology and theory</i> | 5 |
| 2. <i>Part I Introduction</i> | 7 |
| 2.1 Technical overview for likelihood-free methodology | 7 |
| 2.1.1 Likelihood-free models | 8 |
| 2.1.2 Ingredients of all likelihood-free Bayesian models | 9 |
| 2.2 Contextual literature review for likelihood-free inference | 13 |
| 2.2.1 Methodological likelihood-free developments | 13 |
| 2.2.2 Theoretical likelihood-free developments | 17 |
| 2.2.3 Development of likelihood-free applications | 19 |
| 2.3 Technical overview for advanced Monte Carlo methods | 21 |
| 2.3.1 Sequential Monte Carlo Samplers methodology | 21 |
| 2.3.2 Trans-dimensional Markov chain Monte Carlo | 25 |
| 2.4 Literature review and context of advanced Monte Carlo methods | 26 |
| 2.4.1 Developments in Sequential Monte Carlo Samplers | 26 |
| 2.4.2 Developments in trans-dimensional samplers | 27 |
| 2.5 Contribution Part I | 28 |
| 3. <i>Journal Paper 1</i> | 31 |

| | | |
|-------|--|----|
| 3.1 | Introduction | 33 |
| 3.2 | Models for computationally intractable likelihoods | 33 |
| 3.3 | Sampler ambiguity and validity | 34 |
| 3.3.1 | Rejection samplers | 35 |
| 3.3.2 | Markov chain Monte Carlo samplers | 35 |
| 3.3.3 | Population-based samplers | 37 |
| 3.4 | Discussion | 38 |
| | <i>References</i> | 41 |
| 4. | <i>Journal Paper 2</i> | 43 |
| 4.1 | Abstract | 44 |
| 4.2 | Introduction | 45 |
| 4.3 | Sequential Monte Carlo and partial rejection | 46 |
| 4.3.1 | Sequential Monte Carlo sampler | 46 |
| 4.3.2 | Incorporating partial rejection control | 47 |
| 4.3.3 | Estimation of the normalizing constant | 48 |
| 4.4 | SMC Sampler PRC algorithm analysis | 49 |
| 4.4.1 | Variance of the incremental weights | 49 |
| 4.4.2 | A central limit theorem | 50 |
| 4.4.3 | Connections to an existing SMC algorithm | 51 |
| 4.4.4 | Analysis of the number of rejection attempts | 52 |
| 4.5 | Approximate Bayesian computation | 52 |
| 4.5.1 | Simulation study | 54 |
| 4.6 | A stochastic claims reserving analysis | 57 |
| 4.6.1 | Analysis and results | 58 |
| 4.7 | Discussion | 59 |
| | <i>References</i> | 65 |
| 5. | <i>Journal Paper 3</i> | 69 |

| | | |
|-------|--|-----|
| 5.1 | Introduction | 71 |
| 5.2 | Likelihood-free models | 72 |
| 5.3 | Bayesian α -stable models | 73 |
| 5.3.1 | Univariate α -stable Models | 74 |
| 5.3.2 | Multivariate α -stable Models | 76 |
| 5.4 | Evaluation of model and sampler performance | 78 |
| 5.4.1 | Univariate summary statistics and samplers | 78 |
| 5.4.2 | Multivariate samplers | 79 |
| 5.5 | Analysis of exchange rate daily returns | 80 |
| 5.6 | Discussion | 81 |
| | <i>References</i> | 89 |
| 6. | <i>Journal Paper 4</i> | 93 |
| 6.1 | Abstract | 94 |
| 6.2 | Introduction | 94 |
| 6.3 | Moving between models | 96 |
| 6.3.1 | Trans-dimensional move types | 96 |
| 6.3.2 | Fan et al (2006)'s marginal density estimator | 97 |
| 6.4 | Automating between-model moves | 98 |
| 6.4.1 | Construction of proposal distribution | 98 |
| 6.4.2 | Use of partial derivatives | 99 |
| 6.4.3 | Obtaining samples | 99 |
| 6.5 | Examples | 100 |
| 6.5.1 | An autoregression with unobserved initial states | 101 |
| 6.5.2 | Mixture of univariate Gaussians | 104 |
| 6.6 | Discussion | 111 |
| | <i>References</i> | 115 |
| 7. | <i>Journal Paper 5</i> | 117 |

| | | |
|-------|--|-----|
| 7.1 | Abstract | 119 |
| 7.2 | Introduction | 120 |
| 7.2.1 | Contribution and structure | 121 |
| 7.3 | CVAR model under ECM framework | 122 |
| 7.4 | Bayesian CVAR models conditional on Rank (r) | 123 |
| 7.4.1 | Prior and Posterior Model | 124 |
| 7.5 | Sampling and Estimation Conditional on Rank r | 124 |
| 7.5.1 | Algorithm 1 | 125 |
| 7.5.2 | Algorithm 2: Adaptive Metropolis within Gibbs | 127 |
| 7.6 | Rank Estimation for Bayesian VAR Cointegration models | 128 |
| 7.6.1 | Posterior Model Probabilities for Rank r via Bayes Factors | 128 |
| 7.6.2 | Model Selection, Model Averaging and Prediction | 129 |
| 7.7 | Simulation Experiments | 131 |
| 7.7.1 | Synthetic Experiments | 131 |
| 7.7.2 | Financial Example 1 - US mini indexes | 133 |
| 7.7.3 | Financial Example 2 - US notes | 133 |
| 7.7.4 | Financial Example 3 | 134 |
| 7.8 | Conclusions | 135 |
| 7.9 | Acknowledgements | 135 |
| 7.10 | Appendix 1 | 136 |

| | |
|-----------------------------|-----|
| <i>References</i> | 143 |
|-----------------------------|-----|

| | |
|--|-----|
| <i>Part II Advances in Bayesian financial risk and non-life insurance models</i> | 147 |
|--|-----|

| | |
|--|-----|
| 8. <i>Part II Introduction</i> | 149 |
|--|-----|

| | | |
|-------|---|-----|
| 8.1 | Operational risk modelling and Basel II | 149 |
| 8.1.1 | Executive Summary: quantifying bank Operational Risk | 150 |
| 8.1.2 | Background and context within Australias financial industry | 151 |
| 8.1.3 | The Advanced Measurement Approach (AMA) | 155 |

| | | |
|--------|--|-----|
| 8.1.4 | Model frameworks for Operational Risk | 159 |
| 8.1.5 | Issues associated with modelling Operational Risk | 159 |
| 8.1.6 | Modelling methodology for OpRisk and the LDA | 160 |
| 8.1.7 | Modelling different data sources and expert elicitation | 163 |
| 8.1.8 | Survey data and scenario analysis | 164 |
| 8.1.9 | Internal loss data and external data | 166 |
| 8.1.10 | Managing Operational Risk | 167 |
| 8.2 | Non-life insurance claims reserving | 168 |
| 8.3 | Contribution Part II | 170 |
| 9. | <i>Journal Paper 6</i> | 173 |
| 9.1 | Abstract | 174 |
| 9.2 | Introduction | 175 |
| 9.3 | Bayesian Inference | 176 |
| 9.3.1 | Bayesian Inference and Operational Risk | 176 |
| 9.3.2 | Bayesian Parameter Estimation and Operational Risk | 177 |
| 9.3.3 | Bayesian Model Selection and Operational Risk | 178 |
| 9.4 | Non-Conjugate Distributions for Modelling Operational Risk | 180 |
| 9.5 | Simulation in Bayesian Models for Operational Risk | 183 |
| 9.5.1 | Simulation Technique 1: Markov chain Monte Carlo | 184 |
| 9.5.2 | Simulation Technique 2: Approximate Bayesian Computation | 185 |
| 9.5.3 | Simulation Technique 3: Simulated Annealing | 186 |
| 9.6 | Simulation Results and Discussion | 187 |
| 9.6.1 | Parameter estimation and simulation - GB2 model | 187 |
| 9.6.2 | Parameter estimation and simulation - g-and-h model | 189 |
| 9.7 | Discussion | 192 |
| | <i>References</i> | 197 |
| 10. | <i>Journal Paper 7</i> | 199 |
| 10.1 | Abstract | 200 |

| | | |
|--|---|-----|
| 10.2 | Introduction | 201 |
| 10.3 | Panjer Recursions and the Volterra Integral Equation. | 202 |
| 10.4 | Importance Sampling using TD-MCMC | 206 |
| 10.4.1 | Simple Importance Sampling Solution | 207 |
| 10.4.2 | Optimal Importance Sampling | 209 |
| 10.5 | Simulation Results and Analysis | 213 |
| 10.5.1 | Example 1: Poisson-Lognormal compound process. | 217 |
| 10.5.2 | Example 2: Poisson-GB2 compound process. | 220 |
| 10.6 | Discussion | 221 |
| 10.7 | Acknowledgements | 221 |
| 10.8 | Appendix | 222 |
| <i>References</i> | | 227 |
| <i>11. Journal Paper 8</i> | | 231 |
| 11.1 | abstract | 232 |
| 11.2 | Introduction | 233 |
| 11.3 | Model | 236 |
| 11.4 | Simulation Study - Bivariate Case | 239 |
| 11.5 | Bayesian Inference: combining different data sources | 240 |
| 11.5.1 | Modelling frequencies for a single risk cell | 241 |
| 11.5.2 | Modelling frequencies for multiple risk cells | 245 |
| 11.6 | Simulation Methodology - Slice sampler | 246 |
| 11.7 | Bayesian Parameter Estimation | 247 |
| 11.7.1 | Conditional on a priori knowledge of copula parameter | 247 |
| 11.7.2 | Joint inference of marginal and copula parameters. | 248 |
| 11.8 | Slice sampling | 248 |
| 11.8.1 | Extensions | 250 |
| 11.9 | Results | 251 |
| 11.9.1 | Estimation of model if copula parameter is known. | 251 |

| | |
|---|-----|
| 11.9.2 Joint estimation of marginal and copula parameters | 253 |
| 11.10 Discussion | 255 |
| 11.11 Appendix A: Simulation of annual losses | 256 |
| 11.12 Appendix B: Full conditional posterior distributions | 257 |
| 11.13 Appendix C: Slice sampler algorithm. | 259 |
| <i>References</i> | 265 |
| <i>12. Journal Paper 9</i> | 269 |
| 12.1 abstract | 270 |
| 12.2 Claims reserving | 271 |
| 12.3 Tweedie's compound Poisson model | 273 |
| 12.4 Parameter estimation | 275 |
| 12.4.1 Likelihood function | 275 |
| 12.4.2 Maximum likelihood estimation | 277 |
| 12.4.3 Bayesian inference | 279 |
| 12.4.4 Random walk Metropolis Hastings-algorithm within Gibbs | 281 |
| 12.4.5 Markov chain results and analysis | 283 |
| 12.5 Variable selection via posterior model probabilities | 284 |
| 12.6 Calculation of the claims reserves | 287 |
| 12.6.1 Results: average over p | 288 |
| 12.6.2 Results: conditioning on p | 289 |
| 12.6.3 Overdispersed Poisson and Gamma models | 291 |
| 12.7 Discussion | 291 |
| <i>References</i> | 299 |
| <i>13. Journal Paper 10</i> | 301 |
| 13.1 abstract | 302 |
| 13.2 Motivation | 303 |
| 13.3 Claims development triangle and DFCL model | 304 |

| | | |
|---------|---|-----|
| 13.3.1 | Classical chain ladder algorithm | 305 |
| 13.3.2 | Bayesian DFCL model | 306 |
| 13.4 | DFCL model estimators | 308 |
| 13.5 | Bootstrap and mean square error of prediction | 309 |
| 13.5.1 | Non-parametric classical bootstrap (conditional version) | 310 |
| 13.5.2 | Frequentist bootstrap estimates | 311 |
| 13.5.3 | Bayesian estimates | 312 |
| 13.5.4 | Credibility Estimates | 313 |
| 13.6 | ABC for intractable likelihoods and numerical Markov chain sampler | 313 |
| 13.6.1 | ABC methodology | 314 |
| 13.6.2 | Technical justification for MCMC-ABC algorithm | 316 |
| 13.7 | Example 1: Analysis of MCMC-ABC bootstrap methodology on synthetic data | 317 |
| 13.7.1 | Generation of synthetic data | 317 |
| 13.7.2 | Sensitivity analysis and convergence assessment | 318 |
| 13.7.3 | Convergence diagnostics | 318 |
| 13.7.4 | Bayesian parameter estimates | 319 |
| 13.8 | Example 2: Real Claims Reserving data | 320 |
| 13.9 | Discussion | 322 |
| 13.10 | Appendix | 324 |
| 13.10.1 | Section 1 | 324 |
| 13.10.2 | Section 2 | 325 |
| 13.10.3 | ABC algorithmic choices for the time series DFCL model | 327 |
| 13.10.4 | Section 3 | 330 |
| 13.10.5 | Section 4 | 331 |

| | | |
|-------------------|-----------|-----|
| <i>References</i> | | 341 |
|-------------------|-----------|-----|

| | | |
|--|--|-----|
| <i>Part III Advances in Bayesian models for telecommunications engineering</i> | | 343 |
|--|--|-----|

| | | |
|----------------------------------|-----------|-----|
| <i>14. Part III Introduction</i> | | 345 |
|----------------------------------|-----------|-----|

| | | |
|---------|--|-----|
| 14.1 | Motivating advanced statistical modelling in wireless communications | 345 |
| 14.2 | Multiple Input Multiple Output antenna systems | 349 |
| 14.2.1 | Uncertainty models for the Channel State Information (CSI) | 350 |
| 14.3 | Orthogonal Frequency Division Multiplexing (OFDM) | 351 |
| 14.4 | Contribution Part III | 354 |
| 15. | <i>Journal Paper 11</i> | 357 |
| 15.1 | Abstract | 358 |
| 15.2 | Introduction | 359 |
| 15.3 | System Description | 360 |
| 15.4 | Problem Statement | 361 |
| 15.5 | Generic TDMCMC Algorithm for Channel Estimation | 363 |
| 15.5.1 | Within-Model moves | 367 |
| 15.5.2 | Between-model moves | 368 |
| 15.6 | Design of <i>Between Model Birth and Death Proposal</i> | 368 |
| 15.6.1 | Algorithm 1: <i>Between Model Birth and Death Moves</i> (BD-TDMCMC) | 369 |
| 15.6.2 | Algorithm 2: Stochastic Approximation TDMCMC (SA-TDMCMC) | 369 |
| 15.6.3 | Algorithm 3: Conditional Path Sampling TDCMC (CPS-TDMCMC) | 372 |
| 15.7 | Complexity Analysis | 376 |
| 15.8 | Estimator Efficiency via Bayesian Cramér Rao Type Bounds | 376 |
| 15.9 | Simulation Results | 379 |
| 15.9.1 | System Configuration and Algorithms Initialization | 379 |
| 15.9.2 | Model Sensitivity Analysis | 380 |
| 15.9.3 | Comparative Performance of Algorithms | 382 |
| 15.9.4 | Algorithm Performance | 383 |
| 15.10 | Conclusions | 384 |
| 15.11 | acknowledgement | 384 |
| 15.11.1 | Adaptive Grid Placement centred on Estimated Posterior Mode | 385 |
| | <i>References</i> | 391 |

| | |
|---|---------|
| <i>16. Journal Paper 12</i> | 395 |
| 16.1 Abstract | 396 |
| 16.2 Introduction | 397 |
| 16.3 System Description | 398 |
| 16.4 Pilot Aided Maximum Likelihood Channel Estimation | 399 |
| 16.5 Bayesian Detection without Considering Channel Uncertainty | 401 |
| 16.6 Bayesian Detection under Channel Uncertainty | 402 |
| 16.6.1 Optimal MAP detection | 402 |
| 16.6.2 Linear MMSE detection | 403 |
| 16.6.3 Hidden Convexity Based Near-Optimal MAP Detector | 404 |
| 16.6.4 Near-Optimal MAP Detector using Bayesian EM | 405 |
| 16.6.5 Hidden Convexity Vs. BEM approach | 407 |
| 16.7 Simulation Results | 407 |
| 16.7.1 System Configuration | 407 |
| 16.7.2 Constellation Design | 408 |
| 16.7.3 Comparison of Detection Techniques | 408 |
| 16.7.4 Affect of Training | 409 |
| 16.8 Conclusions | 409 |
| <i>References</i> | 415 |
| <i>17. Journal Paper 13</i> | 417 |
| 17.1 Abstract | 418 |
| 17.2 Problem formulation | 419 |
| 17.3 MAP Estimation using Hidden Convexity Optimization | 420 |
| 17.3.1 Complexity Analysis | 421 |
| 17.4 Simulation Results | 422 |
| 17.5 Conclusions | 423 |
| <i>References</i> | 425 |

| | |
|---|-----|
| <i>18. Journal Paper 14</i> | 427 |
| 18.1 Abstract | 428 |
| 18.2 Background | 429 |
| 18.3 Bayesian system model and detection | 430 |
| 18.3.1 Model and assumptions | 431 |
| 18.3.2 Inference and MAP sequence detection | 433 |
| 18.4 Likelihood-free methodology | 434 |
| 18.4.1 Approximate Bayesian computation MCMC approach | 435 |
| 18.5 Auxiliary variable MCMC approach | 438 |
| 18.6 Alternative MAP detectors and lower bound performance | 439 |
| 18.7 Sub-optimal exhaustive search Zero Forcing approach | 439 |
| 18.7.1 Lower bound MAP detector performance | 440 |
| 18.8 Results | 441 |
| 18.8.1 Analysis of mixing and convergence of MCMC-ABC methodology | 441 |
| 18.8.2 Analysis of ABC model specifications | 442 |
| 18.8.3 Comparisons of detector performance | 443 |
| 18.9 Conclusions | 444 |
| <i>References</i> | 451 |
| <i>19. Summary and future work</i> | 453 |
| 19.1 Conclusions | 453 |
| 19.1.1 Outcomes of Part I | 454 |
| 19.1.2 Outcomes of Part II | 456 |
| 19.1.3 Outcomes of Part III | 458 |
| 19.2 Future work | 460 |
| <i>References</i> | 463 |

1

Motivation and structure

“You don’t write because you want to say something, you write because you have something to say.”

F. Scott Fitzgerald

The experiences we each face on a daily basis can be summarized as a series of decisions to take actions which manipulate our environment in some way or another. Such decisions are based on predictions or inferences of quantities that are arrived at via models of what we expect to observe. Often the models we strive to create are designed to capture salient trends or regularities in the observed data with the intention of then making predictions for future outcomes. In real world settings we are content with reduction and approximation in models which still help us to shed light on the underlying stochastic processes we are interested in understanding. The necessity for approximation can arise for many reasons, for example the data being modelled is too complex and the statistical model being estimated using such data fails to capture all features in a single model. Alternatively, it may be that the underlying stochastic processes giving rise to the data are not yet well enough understood and therefore postulating a statistical model will be at best a useful approximation. In such situations we can hope only to design models that are simplifying approximations of the true processes that generated the data. This thesis will concentrate significantly on one such class of approximate models, within a Bayesian modelling context, known as approximate Bayesian computation (ABC) models.

Bayesian inference is the main statistical framework focused upon in this thesis when constructing models for both the financial and wireless telecommunications engineering applications. A Bayesian framework provides the mathematical machinery that can be used for modelling systems, where the uncertainties of the system are taken into account and the decisions are made according to rational principles. The tools of this machinery are the probability distributions and the rules of probability calculus. In a Bayesian analysis, all the quantities have a probability distribution associated with them. Bayes rule provides a means of updating the distribution over parameters from the prior to the posterior distribution in light of observed data. In theory, the posterior distribution captures all information inferred from the data about the parame-

ters. This posterior is then used to make optimal decisions or predictions, or to select between models. The Bayesian framework depends on the existence of *a priori* distributions for model parameters. These priors reflect the user's knowledge about the quantities of interest before any data have been considered. When non standard, i.e. non conjugate priors are considered, it often becomes a real statistical challenge to efficiently obtain samples from the resulting posterior distribution. Making inference, estimation and prediction also challenging. In this thesis we also consider the additional difficulties that can arise in settings in which the likelihood model does not admit a closed form density or may not be evaluated pointwise. This is the nature of models in the family of approximate Bayesian computation. In such settings we develop efficient means to still obtain samples from the posterior distribution.

In general the ability to make inferences and predictions in a Bayesian framework relies on the ability to efficiently sample from the posterior distribution developed as the Bayesian model. Tackling these challenging methodological and numerical sampling problems, often in high dimensions is a modern statistical challenge. The development of Monte Carlo methods over the past decades has seen an explosion in the use of high fidelity models within the engineering, finance, and bioinformatics disciplines, to name but a few. As model fidelity grows, so does the requirement for more advanced Monte Carlo techniques that are able to efficiently explore the (probability) space of interest. It is the intention of this thesis to provide several novel methodological contributions to the scientific literature so that previously encountered problems are now rigorously circumvented, whilst careful consideration is made to the computational efficiency of all proposed approaches.

In this thesis we consider Monte Carlo methods to be the class of algorithms which fundamentally make use of random samples from some distribution to achieve their simulation result. Of course, in practice, almost all computer simulations make use of pseudo-random number generators, this aspect is not the focus of this thesis. Additionally, it is worth noting that until recently Monte Carlo methods were only of specialist interest. However, with the onset of immense computational resources and the development of new sampling methodologies, this domain of statistical research has exploded. Monte Carlo methods are now one of the most broadly used computational techniques. Evidence of this claim is seen when one considers the Metropolis algorithm which has been named one of the ten most influential algorithms of the twentieth century, see [Cipira 2006] (24) for discussion.

This thesis is primarily concerned with development of methodology for nonstandard Bayesian modelling and associated novel Monte Carlo sampling algorithms. The applications focus on Bayesian models developed particularly in the settings of financial risk modelling and insurance mathematics and then in a separate area of wireless communications engineering. The impact that Bayesian modelling has had in these two disciplines has already been substantial. At the same time there is an ever increasing set of expectations from a practical perspective to continue to extend Bayesian modelling features. This may involve developing more flexible models to capture complexities observed in practice. Then in the process of developing these new models questions related to automation of the associated numerical sampling methodol-

ogy, typically required to work with such Bayesian models in practice, arise.

The structure of the thesis is presented as three bodies of journal publications. As such the figures, acronyms, notations, definitions and references will be contained and defined within each journal paper. Additionally, due to the different style files of each journal, the citation style will also alter between chapters. The English grammar utilized is uniformly set according to the Oxford English Dictionary American edition, as this reflects the style used in the majority of the journal papers contained in the thesis. Each chapter of the thesis will be comprised of a journal paper. The thesis is presented in three parts each comprising publications with a particular focus. Though each part can be read individually and still maintain a clear focus on the problems being addressed, there are however common threads binding the parts together. These common themes involve the development of novel sampling methodology in two important Bayesian modelling settings. The first setting concerns approximate Bayesian computation or "Likelihood-free" models. The second deals with trans-dimensional models involving Bayesian posterior distributions defined over multiple models and multiple associated sets of parameters.

Hence, the papers contained in each part develop statistical models which incorporate either: a likelihood model which for some reason is intractable; or they contain aspects of model uncertainty, which is manifest in the form of several possible Bayesian models being developed and one would like to perform model selection or model averaging.

The specific structure followed in each part comprises an introduction chapter providing first an advanced overview pertaining to statistical background material. This is followed by contextual discussion and a relevant literature survey. The remaining chapters in each part correspond to journal publications. Finally, a concluding chapter for the thesis is provided along with a bibliography which is only relevant to the introductory chapters containing literature survey found at the beginning of each part.

Part I introduces the contributions made to the development of statistical methodology in two areas of likelihood-free modelling and trans-dimensional sampling and the associated theoretical studies performed. In particular Part I introduces the following topics: approximate Bayesian computation methodology; Sequential Monte Carlo Samplers; Bayesian model selection, model averaging and associated sampling algorithms generically termed trans -dimensional Markov chain Monte Carlo (TDMCMC). This will aid the understanding of the context of the statistical models and sampling methodology, developed in this thesis.

Part II then develops advanced Bayesian models for financial risk and non-life insurance. Part III develops advanced Bayesian models in the setting of wireless telecommunications systems. Therefore both Part II and Part III utilize and extend the novel sampling methodology developed in Part I. Hence, in addition to proposing novel models, these sections also demonstrate in several settings how the methodological work in Part I can be applied to real world problems.

Part I

ADVANCES IN APPROXIMATE BAYESIAN
COMPUTATION AND TRANS-DIMENSIONAL
MARKOV CHAIN MONTE CARLO

METHODOLOGY AND THEORY

Part I Introduction

“All things are ready, if our minds be so”

William Shakespeare

This introduction chapter of the thesis for Part I comprises five sections. The first section presents specific technical background relating to likelihood-free Bayesian methodology. Leading on from this technical review the second section then presents a contextual literature review relating to likelihood-free methodology. This literature review particularly focuses on aspects of the methodology directly pertinent to the journal papers contained in the remaining chapters of the thesis. As such it is separated into three subsections: the first covering a review of methodological developments in likelihood-free models; the second covering relevant theoretical results for such likelihood-free methodology; and the third considering application of Bayesian likelihood-free models divided into biologically inspired versus non-biologically inspired settings.

The third section then presents a brief technical background for particularly relevant methodology relating to advanced Monte Carlo sampling algorithms. This section is split into two subsections the first focusing on advanced Sequential Monte Carlo type algorithms and the second subsection focusing on trans-dimensional Markov chain Monte Carlo methodology. Then the fourth section provides a brief literature review related to these particular advanced Monte Carlo sampling methodologies. Finally, the fifth section summarizes the novelty and contribution introduced in each of the journal papers contained in Part I of the thesis.

2.1 *Technical overview for likelihood-free methodology*

Bayesian inference proceeds via the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{y}) \propto \pi(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$, the updating of prior information $\pi(\boldsymbol{\theta})$ for a parameter $\boldsymbol{\theta} \in \Theta$ through the likelihood $\pi(\mathbf{y}|\boldsymbol{\theta})$ after observing data $\mathbf{y} \in \mathcal{X}$. This section primarily deals with the class of Bayesian statistical models which

involve intractability in the likelihood model. These classes of algorithms are typically referred to as either likelihood-free or approximate Bayesian computation (ABC), and these terms will be used interchangeably throughout. The term "intractability" will be used broadly to refer to settings in which the likelihood: can not be expressed in a closed analytic form; can only be written down analytically as a function, operation or an integral expression, which can not be solved analytically; can not be directly evaluated point-wise; or evaluation point-wise involves a prohibitive computational cost.

The primary focus of developing methodology for this purpose is that one can now work with a significantly richer class of Bayesian statistical models. The methodology developed for ABC allows one to solve previously intractable problems.

2.1.1 Likelihood-free models

The development of ABC methodology requires several components which include: an intractable target posterior distribution that the ABC posterior will approximate; a technique to simulate data from the intractable model, given a set of parameters; summary statistics for the actual data and the synthetically simulated data; a distance metric to quantify the difference between the two sets of summary statistics; a tolerance level to specify an accuracy level for distance metric quantifications; and a numerical sampling algorithm such as rejection, Markov chain Monte Carlo (MCMC), Sequential Monte Carlo (SMC) or Population Monte Carlo (PMC).

This section demonstrates the ABC model approximation to the true posterior in its most general form, more details can be found in [Sisson *et al.* (2009)](111) and [Peters *et al.* (2009)](92). It describes how evaluation of the intractable likelihood is circumvented. Of course the basic concept of working with a posterior distribution with elements of intractability, such as the normalizing constant, is not a new concept, see e.g. [Moller *et al.* (2006)] (80). What separates likelihood-free methodology from this literature is its generality. Under ABC methodology one can simulate from the posterior when the entire likelihood function is computationally intractable or evaluation is prohibitive.

Simulation from the target posterior is achieved in likelihood-free methodology via embedding it within an augmented model,

$$\pi(\boldsymbol{\theta}, \boldsymbol{x}|\boldsymbol{y}) \propto \pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}). \quad (2.1.1)$$

Here $\boldsymbol{x} \in \mathcal{X}$ is a vector of auxiliary variables, on the same space as data \boldsymbol{y} , which is considered an artificial dataset termed throughout as a "synthetic data" set which is generated from the model likelihood, $\boldsymbol{x} \sim \pi(\boldsymbol{x}|\boldsymbol{\theta})$. The function $\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$ is considered a weighting of the intractable posterior, with high weight in regions where summary statistics $\boldsymbol{T}(\boldsymbol{y}) \approx \boldsymbol{T}(\boldsymbol{x})$ of \boldsymbol{x} and \boldsymbol{y} are similar, low weight in regions where $\boldsymbol{T}(\boldsymbol{y})$ and $\boldsymbol{T}(\boldsymbol{x})$ are not similar and uniquely maximized when $\boldsymbol{T}(\boldsymbol{y}) = \boldsymbol{T}(\boldsymbol{x})$.

The resulting target distribution is obtained as the marginal distribution of the expression in

Equation (2.1.1), denoted by

$$\pi_{ABC}(\boldsymbol{\theta}|\mathbf{y}) \propto \int_{\mathcal{X}} \pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})\pi(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\mathbf{x}. \quad (2.1.2)$$

Clearly, performing this marginalization in the likelihood-free setting is done via approximation. To demonstrate this is an approximation we label the marginal approximation to the target posterior by $\pi_{ABC}(\boldsymbol{\theta}|\mathbf{y})$ which is given by,

$$\begin{aligned} \pi_{ABC}(\boldsymbol{\theta}|\mathbf{y}) &\propto \pi(\boldsymbol{\theta}) \int_{\mathcal{X}} \pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})\pi(\mathbf{x}|\boldsymbol{\theta})d\mathbf{x} \\ &= \pi(\boldsymbol{\theta})\mathbb{E}_{\pi(\mathbf{x}|\boldsymbol{\theta})}[\pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})] \\ &\approx \frac{\pi(\boldsymbol{\theta})}{S} \sum_{s=1}^S \pi(\mathbf{y}|\mathbf{x}^s, \boldsymbol{\theta}), \end{aligned} \quad (2.1.3)$$

where $\mathbf{x}^1, \dots, \mathbf{x}^S \sim \pi(\mathbf{x}|\boldsymbol{\theta})$ are draws of S sets of synthetic data from the (intractable) likelihood. The expectation given in Equation (2.1.3) was first stated explicitly by [Marjoram *et al.* (2003)] (77) and has since been examined by several authors including [Peters *et al.* (2008)] (90); [Reeves and Pettitt (2005)] (99); [Ratmann *et al.* (2009)] (98) and [Toni *et al.* (2009)] (117).

Under this likelihood-free methodology there have been two approaches to posterior simulation from $\pi_{ABC}(\boldsymbol{\theta}|\mathbf{y})$ considered in the literature. The first involves simulation via the augmented model $\pi(\boldsymbol{\theta}, \mathbf{x}|\mathbf{y})$, in which one obtains joint samples $(\boldsymbol{\theta}, \mathbf{x}) \in \Theta \times \mathcal{X}$ before *a posteriori* marginalization over \mathbf{x} . The second approach works on a marginal space of the parameters of interest by working directly on $\pi_{ABC}(\boldsymbol{\theta}|\mathbf{y})$ via Monte Carlo integration, shown in Equation (2.1.3), in lieu of each (ABC) likelihood evaluation, this is discussed in detail in Section 1, Journal Paper 1.

Typically the majority of literature on likelihood-free methodology has made a further simplifying approximation, by working with the hierarchical model $\boldsymbol{\theta} \rightarrow \mathbf{x} \rightarrow \mathbf{y}$ proposed in [Reeves and Pettitt (2005)] (99). This reduces the embedded joint posterior $\pi(\boldsymbol{\theta}, \mathbf{x}|\mathbf{y})$ to the form $\pi(\boldsymbol{\theta}, \mathbf{x}|\mathbf{y}) \propto \pi(\mathbf{y}|\mathbf{x})\pi(\mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$.

Before proceeding, we note that some likelihood-free methods trade computationally intensive ABC posterior simulation for a further level of approximation in the ABC model. After drawing samples from $\pi(\boldsymbol{\theta}, \mathbf{x}|\mathbf{y})$ with a diffuse weighting function $\pi(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$, these methods aim to make regression-based location adjustments to the samples, so that they are drawn approximately from the joint posterior under a much less diffuse weighting [Beaumont *et al.* (2002)] (10) and [Blum and Francois (2009)] (13). Aspects of these approximations are discussed in more detail below in the literature review.

2.1.2 Ingredients of all likelihood-free Bayesian models

This section briefly presents the key ingredients that must be considered in any likelihood-free methodology and discusses their development through the statistical literature. More details

and explicit examples can be found in papers contained throughout this thesis.

Generation of a synthetic data set.

Likelihood-free methodology involves a key assumption, required for this methodology to be applicable, that is the ability to efficiently generate realizations from the model. This generation of "synthetic data" is used to replace the need to evaluate point-wise the likelihood. This is clear in the likelihood-free methodology, as presented in Equation (2.1.1), where the embedded posterior distribution and the associated likelihood-free marginal target posterior distribution approximation, given in Equation (2.1.3) involve simulation of S realizations of the synthetic data set denoted generically by, $\mathbf{x}^{(s)}$.

The exception to this assumption is presented in the paper of [Peters *et al.* (2009)] (92) in which the model can also not be simulated from directly. To overcome this complication, a novel approach is introduced to likelihood-free methodology, involving a specialized bootstrap procedure to generate the synthetic data \mathbf{x} .

Summary statistics.

Having generated a set of synthetic data \mathbf{x} , likelihood free methodology then involves calculation of summaries of the actual and synthetic data, denoted respectively by vectors $\mathbf{T}(\mathbf{y})$ and $\mathbf{T}(\mathbf{x})$.

Many choices for the summary statistics have been considered. The earliest ABC methods used the choice of $\mathbf{T}(\mathbf{y}) = \mathbf{y}$, which compared directly the simulated data and actual data. It was soon realized that for observations defined on a continuous domain, this comparison was overly restrictive. Ideally, one would like to consider a vector of summary statistics \mathbf{T} which are comprised of sufficient statistics. Then according to the Neymann-Fisher Factorization Theorem this choice of summary for the observations would capture all the information about the model. Therefore, under this choice the ABC methodology compares the evaluation of the sufficient statistics on the synthetic and actual data, without bias.

In the majority of practical likelihood-free models, sufficient statistics can not be obtained. In general the vector of summary statistics \mathbf{T} is commonly assumed to provide some degree of near-sufficiency for θ through $\pi(\mathbf{x}|\theta)$. Several alternatives to sufficient statistics have been proposed for the vector of summary statistics \mathbf{T} , see [Peters *et al.* (2009)] (93) for examples and discussion. Popular examples include generally applicable empirical summaries such as: moments; quantiles and percentiles; empirical characteristic function estimates; orthogonal statistics; transformation of the model parameters θ to ϕ for which estimators are known. Note in the ABC context, one can compare the data from the transformed parameters, as long as the synthetic data is obtained via the transformed model. As shown in several examples in [Peters *et al.* (2009)] (93), this fact is very useful in terms of computational efficiency and in generalizing the methodology.

Recently, there has been a further development by [Ratman *et al.* (2009)] (98) who develop an alternative use for \mathbf{T} concerning the assessment of model adequacy.

Distance metrics.

Having obtained summary statistic vectors $\mathbf{T}(\mathbf{y})$ and $\mathbf{T}(\mathbf{x})$, likelihood-free methodology then measures the distance between these vectors using a distance metric, denoted generically by $\rho(\mathbf{T}(\mathbf{y}), \mathbf{T}(\mathbf{x}))$. The most popular example involves the basic Euclidean distance metric which sums up the squared error between each summary statistic as follows:

$$\rho(\mathbf{T}(\mathbf{y}), \mathbf{T}(\mathbf{x})) = \sum_i^{\dim(T)} (\mathbf{T}_i(\mathbf{y}) - \mathbf{T}_i(\mathbf{x}))^2. \quad (2.1.4)$$

Recently more advanced choices have been proposed and their impact on the methodology has been assessed, see for example in this thesis [Peters *et al.* (2009)](91) and [Peters *et al.* (2009)](92). The extensions analyzed include: scaled Euclidean distance given by

$$\rho(\mathbf{T}(\mathbf{y}), \mathbf{T}(\mathbf{x})) = \sum_i^{\dim(T)} W_i (\mathbf{T}_i(\mathbf{y}) - \mathbf{T}_i(\mathbf{x}))^2; \quad (2.1.5)$$

Mahalanobis distance,

$$\rho(\mathbf{T}(\mathbf{y}), \mathbf{T}(\mathbf{x})) = \sqrt{(\mathbf{T}(\mathbf{y}) - \mathbf{T}(\mathbf{x})) \Sigma^{-1} (\mathbf{T}(\mathbf{y}) - \mathbf{T}(\mathbf{x}))^T}; \quad (2.1.6)$$

L^p norm,

$$\rho(\mathbf{T}(\mathbf{y}), \mathbf{T}(\mathbf{x})) = \sum_i^{\dim(T)} [|\mathbf{T}_i(\mathbf{y}) - \mathbf{T}_i(\mathbf{x})|^p]^{\frac{1}{p}}; \quad (2.1.7)$$

and city block distance,

$$\rho(\mathbf{T}(\mathbf{y}), \mathbf{T}(\mathbf{x})) = \sum_i^{\dim(T)} |\mathbf{T}_i(\mathbf{y}) - \mathbf{T}_i(\mathbf{x})|. \quad (2.1.8)$$

In particular we note that distance metrics which include information regarding correlation between the summary statistics produce estimates of the marginal posterior $\pi_{ABC}(\boldsymbol{\theta}|y)$ which are, for a finite computational budget, typically more accurate and involve greater efficiency in the simulation algorithms utilized.

In theory the form of the distance function ρ is not important as long as it satisfies the condition that $\rho(\mathbf{T}(\mathbf{y}), \mathbf{T}(\mathbf{x})) = 0$ iff $\mathbf{T}(\mathbf{y}) = \mathbf{T}(\mathbf{x})$. This condition ensures that the limiting distribution is correct. This can be generalized by considering a comparative tolerance ϵ , between the summary of the synthetic and actual data, such that as the tolerance $\epsilon \rightarrow 0$ one has

$$\lim_{\epsilon \rightarrow 0} p(\boldsymbol{\theta} | \rho(\mathbf{T}(\mathbf{y}), \mathbf{T}(\mathbf{x})) \leq \epsilon) = p(\boldsymbol{\theta} | \mathbf{T}(\mathbf{y})) \quad (2.1.9)$$

In practice, as ϵ decreases the computational cost increases. For a given set of summary statistics, the increase in computation is highly dependent on the choice of distance function. This

in turn affects the accuracy of the estimate for $\pi_{ABC}(\boldsymbol{\theta}|y)$. Hence, one must identify a distance function which provides a good approximation $p(\boldsymbol{\theta}|\rho(\mathbf{T}(\mathbf{y}), \mathbf{T}(\mathbf{x})) \leq \epsilon) = p(\boldsymbol{\theta}|\mathbf{T}(\mathbf{y}))$ for as large a ‘tolerance’ ϵ as possible, to ensure reasonable computational cost. Discussion about this can be found in several papers included in this thesis, see [Peters *et al.* (2008)] (90); [Peters *et al.* (2009)] (92); and [Nevat *et al.* (2009)] (82).

Decision or Weighting function.

Having quantified the distance between the summary statistics of the actual and synthetic data through $\rho(\mathbf{T}(\mathbf{y}), \mathbf{T}(\mathbf{x}))$, the next stage is to produce a weighting as specified by $\pi(y|x^s, \boldsymbol{\theta})$ in Equation (2.1.3). Several choices have been proposed for the weighting function. The original papers in this field used a weighting function, which compared directly the summary of the synthetic data and actual data, as follows:

$$\pi(y|x, \boldsymbol{\theta}) \propto \begin{cases} 1 & \text{if } \rho(T(\mathbf{y}), T(\mathbf{x})) = 0 \\ 0 & \text{else} \end{cases} \quad (2.1.10)$$

This was then generalized to include the concept of near sufficiency through a tolerance level ϵ , which in the ideal case of sufficient statistics directly controls the degree of ABC approximation. Again, these papers resorted to the simplistic "hard" decision weighting function, now given by

$$\pi(y|x, \boldsymbol{\theta}) \propto \begin{cases} 1 & \text{if } \rho(T(\mathbf{y}), T(\mathbf{x})) \leq \epsilon \\ 0 & \text{else} \end{cases} \quad (2.1.11)$$

This weighting function is still commonly utilized, but it has now been shown in several of the papers in this thesis to be an inefficient weighting function. Alternative forms of weighting function have been proposed and studied, these include the Gaussian and Epanechnikov kernels, see [Peters *et al.* (2009)] (91) and [Beaumont *et al.* (2002)] (10) respectively.

Tolerance schedule.

The choice of tolerance is an important consideration as it directly affects the performance in terms of accuracy of the estimated approximation $\pi_{ABC}(\boldsymbol{\theta}|y)$. Several choices have been proposed, generally depending on the algorithm used to perform the sampling from the posterior. Rejection algorithms typically use a deterministic fixed tolerance value. More sophisticated versions of the Markov chain Monte Carlo and Sequential Importance Sampling based approaches tend to use alternatives such as deterministic annealed or tempered tolerance schedules. Recently, state of the art work involves designing adaptive and dynamic annealing schedules. Details of these approaches can be found in the papers of [Peters *et al.* (2009)] (93) and [Del Moral *et al.* (2008)] (32).

2.2 Contextual literature review for likelihood-free inference

The intention of this section is to provide a literature review for the developments of likelihood-free methodology, from a statistical perspective. It is widely regarded, that ABC methodology originated not in the statistics literature, but rather in the population genetics literature just over a decade ago, see [Beaumont *et al.* (2002)] (10); [Tavare *et al.* (1997)] (113). Since, these original papers the advancement in both the understanding of ABC methodology, models and the numerical simulation algorithms has resulted in the propagation of ABC models through an ever-increasing range of disciplines.

The following section is separated into three subsections dealing with methodological, theoretical and application based developments. In particular it will contain a brief chronological review of the key papers contributing to the ABC literature, with preference given to papers of particular relevance to this thesis, as such it is not exhaustive. It is also worth noting, that a frequentist type approach, which is similar to, but not the same as likelihood-free methodology, can be found in the Econometrics literature. The research field of Indirect Inference has parallels and links with likelihood-free methodology, see for example [Gourieroux *et al.* (1993)] (56) and the references therein. A common feature of this methodology is the ability to circumvent the intractability of a likelihood, in this case for performing maximum likelihood parameter estimation also via simulation from the likelihood model. In this thesis we focus on a Bayesian framework.

2.2.1 Methodological likelihood-free developments

The basic concept of likelihood-free methodology in the Bayesian literature arose initially in the paper of [Tavare *et al.* (1997)](113). The authors of this paper developed two basic rejection sampling ABC algorithms. These algorithms were used to tackle a problem in statistical genetics, which is related to inference about coalescence times, based on DNA sequence data. This was the first rudimentary specification of a simple ABC methodology. Following this paper, advances to the rejection sampling framework are presented in papers such as [Fu and Li (1997)] (47), [Weiss and von Haeseler (1998)] (120), [Pritchard *et al.* (1999)] (95), [Tishkoff *et al.* (2001)] (116) and [Estoup *et al.* (2002)] (41).

Subsequently, a significant paper by [Beaumont *et al.* (2002)] (10) introduced a novel development to the ABC literature in which improvements could be made to the approximation of the ABC posterior density estimate. This was achieved by fitting a local-linear regression of simulated parameter values on simulated summary statistics. Then the observed, summary statistics were substituted into the regression equation. This method was found to produce a significant efficiency gain relative to previous approaches, which were simply based on data.

The next important methodological development came from [Marjoram *et al.* (2003)] (77) and [Plagnol and Tavare (2004)] (94). They realized the restrictions associated with the rejection algorithm in the likelihood-free setting were prohibitive from a computational perspective.

The rejection probability is directly related to the tolerance level ϵ . Instead they developed an MCMC alternative, and were able to demonstrate marked improvements in computational efficiency. The particular aspects, proposed to improve the ABC methodology, included replacing a direct comparison, between the simulated data and the actual data, in the hard decision weighting function, with the comparison of summary statistics, as in Equation (2.1.11). The second aspect, they introduced, was the concept of using an approximation, by introducing the tolerance level as in Equation (2.1.11). In this case the samples only come from the true target posterior, once the stationary regime is reached by the MCMC sampler and asymptotically in the tolerance, $\epsilon \rightarrow 0$.

These developments have resulted in the MCMC-ABC approach becoming a popular alternative and there have been several extensions exploring the MCMC-ABC sampler. Examples of these include the papers of [Peters and Sisson (2006)] (89) and [Bortot *et al.* (2007)] (17), who develop a novel extension to the MCMC-ABC sampler each based on the concept of simulated annealing, modified for the setting of the likelihood-free framework. In particular the paper of [Peters and Sisson (2006)] (89), develops a novel simulated annealing version of the MCMC-ABC algorithm to perform MAP estimation in a financial modelling application and also mentions the possibility of an extension in which the tolerance level ϵ is learnt adaptively on line.

The paper of [Bortot *et al.* (2007)] (17) makes this concept explicit by developing a novel joint sampler for both the parameters of the posterior distribution $\boldsymbol{\theta}$ and the tolerance ϵ . The prior for the tolerance, $p(\epsilon)$, was specified to favor low values. A key aspect of this proposed tempering based approach was the post-simulation filtering of the joint samples to only keep those samples with tolerance below a specified tolerance level, $\epsilon \leq \epsilon_T$. In this way, the approximation error, obtained by using non-zero tolerance, as demonstrated in line 2 of Equation (2.1.3), is controlled. The approximate ABC posterior distribution from which N Markov chain samples are used to approximate, $\pi_{ABC}(\boldsymbol{\theta}|y)$ is given by,

$$\pi_{ABC}(\boldsymbol{\theta}|y) \approx \frac{\pi(\boldsymbol{\theta})}{N} \sum_{n=1}^N \pi(y|x^n, \boldsymbol{\theta}, \epsilon^n) \pi(\epsilon^n) \mathbb{I}(\epsilon^n \leq \epsilon^T). \quad (2.2.1)$$

Alternatively [Nevat *et al.* (2008)] (82), developed a deterministic annealing schedule for the tolerance level ϵ . This was used instead of the approach of sampling the tolerance jointly with the parameters ϵ of the model and imposing a prior to ensure posterior preference for small ϵ . In doing this [Nevat *et al.* (2008)] (82) are able to avoid the additional Monte Carlo variance in estimates, due to numerically integrating out the tolerance random variable from the posterior.

The tolerance is decreased according to a deterministic schedule during the burn-in period of the MCMC-ABC algorithm and stopped at a final value of ϵ_T after a given mixing criteria based on the acceptance probability of the MCMC-ABC sampler is satisfied.

Even with these advances in likelihood-free methodology, when the MCMC-ABC algorithm is utilized, one must pay particular attention to the mixing properties of the chain created.

This is obvious, when one considers the simple indicator decision function, given in Equation (2.1.11). This indicator condition enters in a multiplicative fashion into the acceptance probability. Therefore, when the Markov chain moves into a region of the parameter space with low posterior mass, even under a moderate ϵ , there will be a low probability of generating a synthetic data set to result in non-zero acceptance probability. This phenomenon of long autocorrelation tails for the resulting simulated Markov chain has been termed "sticking". It is studied and discussed in detail in [Peters and Sisson (2006)] (89).

As a result [Peters *et al.* (2009)] (92) combined the idea of a deterministically annealed tolerance schedule with the application of three convergence criteria, to monitor and design the MCMC-ABC sampler to reduce the impact of "sticking". The three approaches were based on auto-correlation time, the [Geweke (1991)] (51) time series diagnostic and the [Gelman and Rubin (1992)] (50) R statistic diagnostic. These were used to quantify the suitability of the MCMC-ABC samples obtained. They were calculated post tuning of the tolerance schedule and final choice of ϵ_T obtained during the burn-in period of the sampler. In addition, this paper introduced a novel concept of working with likelihood-free models, in which not only is the likelihood intractable, but it cannot be generated from to produce the synthetic data \mathbf{x} . Instead of generating directly from the model, conditional and unconditional bootstrap procedures were introduced. Finally, though not developed in the paper, there are comments about the extension of utilizing the work of [Gramacy *et al.* (2008)] (57) on "Importance Tempering". This could be combined either in the setting of a deterministic annealed tolerance schedule, or in a setting based on the work of [Bortot *et al.* (2007)] (17), to avoid the approximation of the integral given in Equation (2.2.1) for ϵ . That is, one can avoid wasting all the samples which are above the tolerance level ϵ_T . Instead, all samples would be included with an associated importance weight.

At the same time that these developments were occurring, a paper by [Reeves and Pettitt (2005)] (99) presented an initial theoretical framework for approximate Bayesian computation models. This paper explicitly described the exact nature of the ABC approximation to the true posterior distribution. It linked the concepts of Indirect Inference with the related ideas found in likelihood-free methodology. A recent paper by [Wilkinson (2008)] (121), follows along these same lines, though it extends the representation of [Reeves and Pettitt (2005)] (99) and demonstrates, that ABC approach will produce exact results under an assumption of model error. In particular two assumptions are made, the first that sufficient statistics are used in the ABC model. The second assumes existence of a uniform additive model error term in an ABC framework. These assumptions are sufficient to demonstrate, that exact results for the ABC posterior distribution are obtained. The authors state, that by representing ABC models as misspecified additive error models, then analyzing the distribution of the error term, one can obtain a relevant metric to understand the possible choices, that can be made for the distance function and the weighting function.

There is also a recent paper by [Blum (2009)] (14), dealing with the situation, in which statistics used are not sufficient. It looks at ABC methodology from a non-parametric perspective,

considering ABC models in which non-parametric statistics are used for the summary statistics. The asymptotic bias and variance of standard estimators for the posterior distribution are studied for the rejection sampler. One result of this paper is that the asymptotic results derived display a curse of dimensionality in ABC methods. That is, as the number of summary statistics increases, the quality of estimators, for the posterior distribution, deteriorates. Finally, the recent paper by [Leuenberger *et al.* (2009)] (74) deals with regression-based approaches to improve the ABC posterior approximation. The novelty introduced involves interpretation of the ABC methodologies that sample from the prior as being equivalent to obtaining samples from a mixed model, ie. a truncated prior mixed with a truncated model.

The next significant developments in the methodological aspects of likelihood-free simulation techniques is found in the papers of [Sisson *et al.* (2007); rejoinder (2009)] (110) and [Peters *et al.* (2008)] (90). They introduced respectively, novel Population based and Sequential Monte Carlo-based approaches for sampling from the target ABC posterior. They developed numerical algorithms involving aspects of partial rejection control from [Liu (2001)] (75), and SMC Samplers methodology from [Del Moral *et al.* (2006)] (32) and [Peters (2005)] (88). It was found that in general a sequential Monte Carlo approach provides a significantly more efficient alternative to perform ABC inference, in comparison to rejection sampling and MCMC methods.

The SMC Samplers methodology, discussed in detail in Section 2.3.1, requires a sequence of target distributions to be developed. These papers create the artificial sequence of target distributions based on Equation (2.1.3) via a sequence of decreasing tolerance levels, $\epsilon_1, \dots, \epsilon_T$ with $\epsilon_i < \epsilon_j$ for all $i > j$. For each value of ϵ_t in the sequence, one gets a target distribution in the SMC Samplers algorithm. Additionally, a mutation kernel, based on partial rejection control (PRC), was required to maintain particle diversity. Significant extensions to this original proposed algorithm have since been made and are presented in the papers contained in this thesis, [Peters *et al.* (2009)] (91) and [Sisson *et al.* (2009)] (111). These extensions include theoretical studies and development of a general class of algorithms, which incorporate the approach of [Sisson *et al.* (2007)] (110). The framework developed provides greater insight into these approaches and more efficient alternatives to the original proposed algorithm.

Recent extensions based on likelihood-free SMC Samplers, include the work of [Del Moral *et al.* (2008)] (32). In this paper the authors present an alternative SMC Sampler based approach which avoids the partial rejection stage by sampling from the likelihood-free posterior distribution on the joint space given in Equation (2.1.1). Again, a sequence of distributions is constructed as in [Sisson *et al.* (2007)] (110). However, the other significant development they include is an adaptive automated methodology for setting the tolerance schedule in the sequence of target distributions. In particular this work is an ABC-application of the more general "expected auxiliary variable" approach of [Andrieu *et al.* 2009] (1).

An alternative approach, which adaptively aims to sample from a sequence of target posteriors, is found in the paper of [Beaumont *et al.* (2009)] (11). The approach of this paper is based on the Population Monte Carlo (PMC) algorithm of [Cappe *et al.* (2004)] (20) adapted to the likelihood-free methodology. The relationship between these approaches and the general class

of SMC Samplers PRC-ABC algorithms is discussed in [Sisson *et al.* (2009)] (111). As such it will not be discussed further in this introduction.

Other recent methodological developments in the literature of likelihood-free modelling include the paper of [Ratman *et al.* (2009)] (98). This paper presents a different perspective on ABC methodology, it does not focus on improving the approximation to the target posterior, given in Equation (2.1.3). Instead, this paper introduces the insight that the degree to which the likelihood could reproduce each of the individual observed summary data could be estimated as part of the model-fitting process. Then this information provides a natural diagnostic for informed model-criticism.

There have been several recent papers focusing on improving the approximation to the ABC posterior, given in Equation (2.1.3). The paper of [Blum *et al.* (2009)] (13) extends the work of [Beaumont *et al.* (2002)] (10) to develop a nonlinear and heteroscedastic regression model to improve the likelihood-free approximation. Other approaches to improving the approximation include the work of [Joyce *et al.* (2008)] (67). They recognize the significance of the choice of summary statistics in the approximation of the ABC posterior distribution. As a result they develop a sequential methodology, which considers, whether including additional summary statistics to construct the ABC approximation will substantially improve the quality of inference. The important drawback of this approach, is that the criteria, developed to analyze the improvement, are sensitive to the ordering in which the additional summary statistics are added to the model.

Another aspect that is beginning to be explored in the ABC context involves dynamic time series modelling. The paper of [Peters *et al.* (2008)] (90) contains a time series model constructed in a likelihood-free methodology. Papers, that take this idea into a setting of dynamical systems model, are the papers of [Toni *et al.* (2009)] (117) and [Cornebise and Peter (2009)](27). These papers directly utilizes the methodology of [Sisson *et al.* (2007)] (110), [Peters *et al.* (2008)](90), [Del Moral *et al.* (2006)] (30), [Peters (2005)] (88) and apply it to popular dynamic state space models to estimate the static parameters and study model sensitivity. The authors also utilize standard model selection criteria to perform model selection in the ABC models they investigate.

2.2.2 Theoretical likelihood-free developments

The theoretical developments in the likelihood-free literature have largely focused on two aspects:

1. the first aspect involves analysis of the approximation error associated with the ABC posterior, the papers associated with this analysis have been covered in the above section on methodological developments;
2. the second aspect has involved establishing asymptotic properties of sampling methodology used to obtain samples from the ABC posterior distribution.

It is the second aspect that is of particular significance to this thesis and therefore will form the focus of the literature review in this section.

To begin with it is worth noting that the paper of [Sisson *et al.* (2008)] (111) provides a summary of the recent significant algorithmic and theoretical developments in the SMC based methodology for likelihood-free inference. The first paper to appear, studying in detail aspects of the SMC Samplers PRC-ABC algorithms, was that of [Peters *et al.* (2009)] (90). In particular, it contains an explicit representation of the SMC Samplers PRC mutation kernel. After developing this PRC mutation kernel the paper goes on to study, firstly the SMC Samplers PRC algorithm relative to basic SMC Samplers algorithm. This is performed under different choices for the forward and backward mutation kernels. The results include asymptotic analysis of the variance in the incremental importance sampling weights under a PRC stage relative to non-inclusion of a PRC stage. Then a CLT is obtained and finally an expression and bounds for the asymptotic variance of the SMC Samplers algorithm on the path space are obtained, based on earlier work from [Del Moral, *et al.* (2006)] (30), [Del Moral, *et al.* (2007)] (31) and [Liu (2001)] (75).

These theoretical results and insights are then placed in the context of likelihood-free inference and several relevant algorithmic developments and observations are obtained from the resulting analysis. For example it provides clear advice about computational efficiency in terms of specification of the forward and backward mutation kernels to avoid added computation associated with estimation of the normalizing constant of the PRC mutation kernel.

A paper by [Del Moral *et al.* (2008)] (32) following on from [Peters *et al.* (2008)] (90) is also significant in terms of contribution to the theoretical developments of the methodology associated with sampling from an ABC posterior distribution. In particular this paper discusses the substantial computational cost associated with SMC based samplers in the context of a likelihood-free framework. It discusses the fact that all current importance sampling (IS) based methodologies have a computational complexity that is quadratic in the number of Monte Carlo samples, when one works with the marginal model presented in Equation (2.1.3). This point is also discussed in detail in [Peters *et al.* (2008)] (90). They then formulate a novel alternative SMC Samplers based algorithm which works directly on the joint space formulation and obtain computational complexity which is linear in the number of samples. However, as discussed in both [Peters *et al.* (2008)] (90) and [Sisson *et al.* (2008)] (111), one must be very careful to ensure the SMC based sampler in this setting does not result in early termination due to "death" of all particles at a given tolerance level. That is it is critical in this joint space formulation to ensure the sequence of distributions is chosen carefully to ensure that the particle system can not at any stage result in all particles with zero importance weight as a result of the likelihood weighting function entering into the incremental importance sampling weight. They overcome this problem by automating the selection of the tolerance schedule, in a restricted fashion. It is also worth noting that under their joint space formulation they have not utilized the partial rejection control mechanism, hence the requirement for restrictions on the tolerance schedule.

In terms of theoretical contributions, the work of [Beaumont *et al.* (2009)] (11) presents a basic analysis of bias for the SMC Samplers based algorithm of [Sisson *et al.* (2007)] (110). Detailed

discussion of this analysis is deferred to Part 1 Chapter 2 and the papers therein.

2.2.3 Development of likelihood-free applications

The level of model complexity, required to motivate the use of a computationally intractable likelihood, is now widespread through a range of diverse disciplines. The intention of this section is to present a brief overview of the applied ABC literature, separated by area of application. In particular the separation between biologically and non-biologically motivated applications will be made.

Biologically inspired applications

It is fitting to begin this section with some of the papers found in the areas of statistical genetics, in which much of the original methodology and popularity for ABC methods arose. As pointed out in [Sisson (2006)] (109), the adoption of ABC methodology in biological modelling can be attributed to the fact, that in many biological applications, the model typically consists of large numbers of nuisance parameters. This is especially the case for models in the area of population genetics. The result of a large number of parameters, in combination with complex models, can result in likelihoods, that are computationally prohibitive or even impossible to evaluate. This sentiment about the complexity of work with statistical models in genetics is also echoed in the papers of [Beaumont *et al.* (2002)] (10); [Marjoram *et al.* (2003)] (77); [Estoup *et al.* (2004)] (41) and [Hamilton *et al.* (2005)] (62).

The paper of [Marjoram *et al.* (2006)] (78) provides a well presented survey of population genetics and the methods, that are used to answer biological questions within species, rather than between species. It discusses particularly relevant questions relating to approximate models, summary statistics of data and computational methods such as ABC samplers. In this regard it presents clearly, for a general audience, rejection based algorithms, MCMC samplers, importance sampling and ABC models.

In [Sisson (2006)] (109) a review is also presented, concentrating on the impact of statistical modelling on computational genetics. In particular, it focuses on numerical Monte Carlo based sampling methodologies, including ABC methodology. It sets out to present to a general audience, how and why naive implementation of popular sampling methodology, such as MCMC, to solve sophisticated and highly complex statistical genetics problems, can often, and will be, problematic. It then goes on to describe how methodologies, such as those focused on ABC samplers, can help to overcome some of these problems.

The paper of [Estoup *et al.* (2004)] (42) considers models, which aim to make inference with regard to the spatial expansion dynamics of an invading species via molecular data. The paper considers five Bayesian demographic models of expansion. The use of ABC methodology was justified as a result of the complexity present in the demographic history. The rejection-regression algorithm was utilized to formally test the relative likelihoods of the five models of expansion and to infer demographic parameters. Also studying demographic models is the

paper of [Hamilton *et al.* (2005)] (62). This paper considers the utility of ABC methodology when assessing Bayesian demographic expansion models.

There are demographic models that have focused on the properties of disease spread via analysis of genotype data, instead of migration of species. The paper of [Tanaka *et al.* (2006)] (112) considers the analysis of transmission rate of Tuberculosis. This is achieved through the use of an ABC computational method combined with a stochastic model of tuberculosis transmission and mutation of a molecular marker to estimate the net transmission rate, the doubling time, and the reproductive value of the pathogen.

Other papers, not covered here in detail, that also utilize ABC methodology involve applications in: ecology [Butler *et al.* (2006)] (19); modelling of drug-resistance [Luciani *et al.* (2009)] (76); and protein networks [Ratmann *et al.* (2009)] (98) and [Ratmann *et al.* (2007)] (97). The next section will consider applications which are not biologically motivated.

Non-biologically inspired applications

The first non-biologically inspired application paper is found in this thesis, see [Peters *et al.* 2006] (89). Briefly it addresses an important class of models for financial mathematics involving the modelling of OpRisk. This is a new and important risk category in banking that arose as a result of the regulatory specifications given in Basel II. The only other financially motivated paper currently utilizing ABC methodology is also contained in this thesis, see [Peters *et al.* (2009)] (91). This paper addresses an important modelling question, regarding the means square error of prediction in a particular popular class of stochastic models for non-life insurance claims reserving.

Other areas of application include the statistical modelling of stereological extremes, see [Bortot *et al.* (2007)] (17). This paper considers models for inclusions, ie. imperfections, in clean steel production. It combines models from extreme value theory (EVT) with a likelihood-free methodology. In particular, measurements of imperfections on planar slices are obtained, leading to an EVT version of the stereological problem of how to make inference on large three dimensional inclusions based on observations of planar slices. The models typically considered in this setting are extended to include elliptical inclusions, resulting in an intractable likelihood.

In terms of physics models, the paper of [Grelaud *et al.* (2009)] (60) considers model choice in the setting of Gibbs random fields. The paper focuses on Ising models under different possible neighborhood structures. It shows, that sufficient statistics can be obtained across the models considered to allow for an exact simulation from the posterior distribution of the models. To clarify, it will be exact in the sense that as the tolerance, $\epsilon \rightarrow 0$, the use of sufficient statistics ensures exact samples drawn from the target posterior.

State space modelling and filtering applications have begun to be explored from the perspective of likelihood-free inference, see [Toni *et al.* (2008)] (117) and [Cornebise and Peters (2009)](27). These papers consider the application of the basic SMC-ABC methodology, when one wants to perform the estimation of parameters in a dynamic model. In the first case the authors also

assess the sensitivity of parameter estimates in dynamic models such as the Lotka-Volterra model; the stochastic repressilator model; and disease transmission models such as the SIR model.

The other area of application, utilizing ABC methodology to overcome significant modelling issues, is in the area of cooperative communications for wireless relay telecommunications networks, see [Nevat *et al.* (2008)] (82) and [Peters *et al.* (2009)] (91). These papers are presented in Part 3 of the thesis and so discussion is deferred until later.

2.3 Technical overview for advanced Monte Carlo methods

This section presents a brief background technical overview of several advanced Monte Carlo simulation methodologies of direct significance to the papers presented in each part of this thesis. It begins with the methodology of Sequential Monte Carlo Samplers and finishes with an overview of Trans-dimensional MCMC.

2.3.1 Sequential Monte Carlo Samplers methodology

This section provides a technical overview for Sequential Monte Carlo Samplers, a particular class of Sequential Monte Carlo algorithms.

Basic SMC algorithms involve finding a numerical solution to a set of filtering recursions. These filtering recursions naturally introduce a sequence of distributions, we denote by $\tilde{\pi}_n(x_{1:n})$, indexed by n . The SMC numerical solution involves obtaining samples from these distributions in a sequential manner. These samples then form a weighted dirac mass estimate of each of the distributions in the sequence up to time n . Each distribution in the sequence is defined on the support $E^n = E \times E \times \dots \times E$ and SMC methodology allows one to recursively construct a set of samples from the distributions on this path space at each time n .

The past fifteen years have witnessed extensive development of literature devoted to this methodology. See the section below related to literature overview. Much of this development has focused on classical non-linear or non-Gaussian state space model filtering or smoothing algorithms. As such much of the original methodological and algorithmic literature on this subject has been in the domain of engineering and target tracking. This thesis does not focus on these state space models. Instead the focus is on developing SMC algorithms as an alternative to popular MCMC algorithms to perform static Bayesian inference.

In this manner SMC Samplers algorithms generalize the standard filtering framework to the setting in which one wishes to sample from a sequence of distributions $\{\pi_n\}$ all defined a fixed common support E . This class of algorithms was initially developed as part of the author's Masters of Science thesis at Cambridge University [Peters (2005)](88) and the associated publication of [Del Moral *et al.* (2006)] (30). Theoretical aspects of this class of algorithm are also discussed in the paper of [Del Moral *et al.* (2007)] (31).

The relevant extension to these techniques, to include partial rejection control (PRC), see [Liu (2001)](75) and in the ABC context [Sisson *et al.* (2007)](110), is deferred until the second chapter of Part I in the thesis. It is an integral part of the paper [Peters *et al.* (2008)](90). We simply note here, that inclusion of PRC is critical to the successful implementation of any approximate Bayesian computation SMC algorithm on the marginal space posterior model given in Equation (2.1.3).

The cost associated with the generality, introduced by the SMC Samplers setting, is that one no longer has a set of filtering recursions or dynamic equations to induce a natural sequence of distributions. Instead, the original work of [Peters (2005)](88) and [Del Moral *et al.* (2006)](30) demonstrates how to sensibly introduce an artificial sequence of distributions, $\tilde{\pi}_n(x_{1:n}) = \pi_n(x_n) \prod_{k=1}^{n-1} L_k(x_{k+1}, x_k)$, which take the required support E^n , where L is an arbitrary backward kernel admitting the correct marginal distributions $\int \tilde{\pi}_n(x_{1:n}) dx_{1:n-1} = \pi_n(x_n)$.

The numerical approximation, established by an SMC Samplers methodology, involves a set of samples or "particles" drawn from an arbitrary initial distribution $\mu_1(x_1)$. Then one may define the sampling distribution on the path space as $\mu_n(x_{1:n}) = \mu_1(x_1) \prod_{k=2}^n M_k(x_{k-1}, x_k)$, where $M_k(x_{k-1}, x_k)$ is a mutation kernel at time k . If one were to perform simple importance sampling on the path space then the importance sampling weight with respect to the target distribution $\tilde{\pi}_n(x_{1:n})$ is given by

$$w(x_{1:n}) = \frac{\tilde{\pi}_n(x_{1:n})}{\mu_n(x_{1:n})} = \frac{\pi_n(x_n) \prod_{k=1}^{n-1} L_k(x_{k+1}, x_k)}{\mu_1(x_1) \prod_{k=2}^n M_k(x_{k-1}, x_k)}.$$

In a sequential importance sampling framework, we assume that at time $n - 1$ an empirical particle approximation, denoted by $\tilde{\pi}_{n-1}^N(x_{1:n-1})$, is available to approximate $\tilde{\pi}_{n-1}(x_{1:n-1})$, where N is the number of particles. Under this construction the weights W_n at time n can be defined recursively in terms of weights W_{n-1} from time $n - 1$, so that $W_n = W_{n-1} w_n(x_{1:n})$, where the incremental weight $w_n(x_{1:n})$ is given by

$$w_n(x_{1:n}) = \frac{\tilde{\pi}_n(x_{1:n})}{\tilde{\pi}_{n-1}(x_{1:n-1}) M_n(x_{n-1}, x_n)} = \frac{\pi_n(x_n) L_{n-1}(x_n, x_{n-1})}{\pi_{n-1}(x_{n-1}) M_n(x_{n-1}, x_n)}.$$

While the construction of the sequence of distributions through the L kernel is arbitrary, several forms have been proposed (e.g. [Peters (2005)](88) and [Del Moral *et al.* (2006)](30)). The choice of the L kernel has a direct influence on algorithm performance in terms of the asymptotic variance of the algorithm. This is discussed extensively in both [Peters *et al.*](90) and [Sisson *et al.*](111).

The SMC sampler algorithm, presented in Part I, journal papers 1 and 2, involves three stages: mutation, correction and selection. Moving from target distribution $\pi_{n-1}(x_{n-1})$ to $\pi_n(x_n)$ at iteration n of the algorithm involves taking the particle estimate of the target distribution at

iteration $n - 1$ given, after resampling by,

$$\tilde{\pi}_{n-1}^N(x_{n-1}) = \sum_{i=1}^N \frac{1}{N} \delta_{X_{n-1}^{(i)}}(x_{n-1}) \quad (2.3.1)$$

and then mutating each particle. The set of N particles at each iteration, denoted by $\{X_{n-1}\}$, is often used in some manner to aid in the construction of the mutation kernel.

The new particles at time n are sampled from this mutation kernel to obtain N new particles $\{X_n\}$. See e.g. [Doucet *et al.* (2001)](34) and [Del Moral (2004)] (29) for discussions on choices of M_n . The new particles are then corrected with respect to the appropriate target distribution π_n via calculation of the importance sampling weights W_n , given by

$$W_n^{(i)} \propto \frac{1}{N} w(X_{n-1}^{(i)}, X_n^{(i)}) . \quad (2.3.2)$$

The corrected particles are then potentially resampled with respect to their weights in the selection stage. Many alternative resampling methodologies have been developed and discussed, see for example (e.g. [Künsch (2001)](71), [Kitigawa (1996)](69); [Cappé (21)] (21) and [Doucet *et al.*](35)).

The diagram in Figures 2.3.1 below demonstrates the stages of an SMC algorithm. It graphically shows how the three stages described above are performed. The next diagram in Figure 2.3.2 shows a simple graphic to demonstrate the application of SMC Samplers in the likelihood-free context. In particular the application of the SMC Samplers methodology to the setting of ABC models is ideal since the ABC framework lends itself naturally to the specification of a sequence of distributions. A natural setup is to consider an annealed sequence of distributions, annealed based on the ABC tolerance level ϵ_n . Each distribution in the sequence is constructed in such a way the $\epsilon_1 \geq \epsilon_2 \dots \geq \epsilon_{n-1} \geq \epsilon_n \dots$. Defining a sequence of ABC posterior distributions in this manner fits into the intended framework for SMC Samplers. Each target distribution in the sequence will be defined on a common support. This is developed further in the second chapter of Part I of the thesis.

We note that it is well known that the performance of SMC methods is strongly dependent on the mutation kernel, see discussion in [Cornebise *et al.* (2008)] (26), denoted generically at stage n by M_n . If M_n is selected poorly at time n , then it will not mutate the next stage of particles to regions where the next target distribution π_n has high mass. A result of this will be that many importance sampling weights will be close to zero. This in turn leads to sample degeneracy, resulting in large variance in estimates or predictions made using these samples. The consequences of this degeneracy when it arises in the context of likelihood-free methodology can result in all particles having zero importance sampling weights and the termination of the SMC Samplers algorithm prematurely. This is easily observed, when an indicator function is used for the weighting function. It is for this reason that the (PRC) stage was introduced to ensure the particle system will survive through each stage of the sequence of distributions. That is, the addition of a PRC stage, at each time n , effectively modifies the mutation kernel

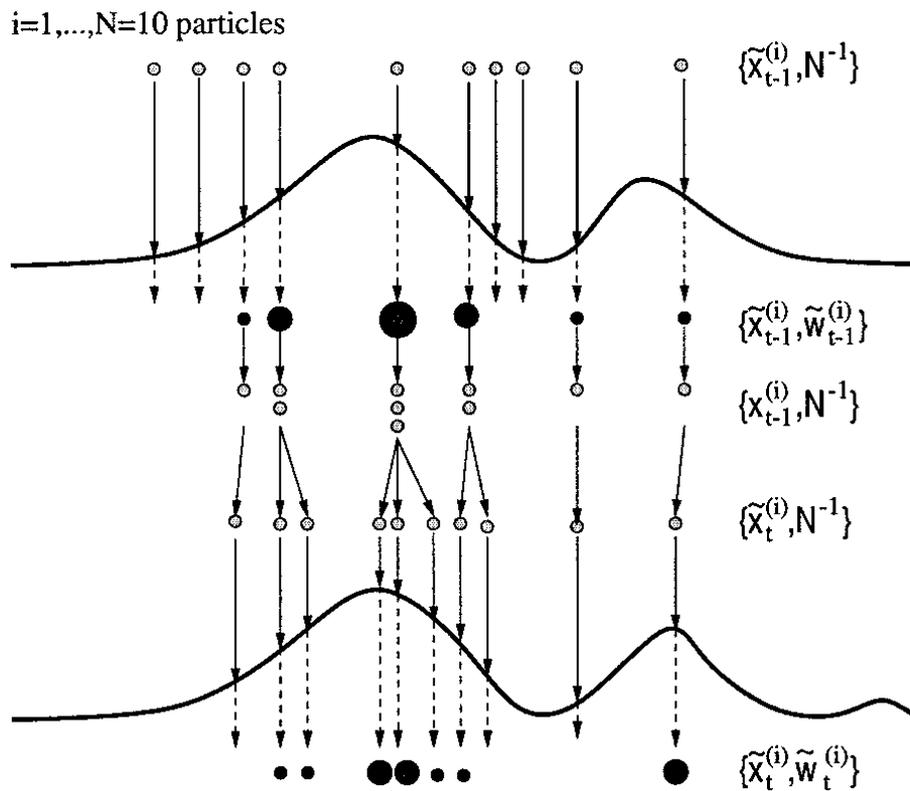


Fig. 2.3.1: The basic stages of an SMC algorithm.

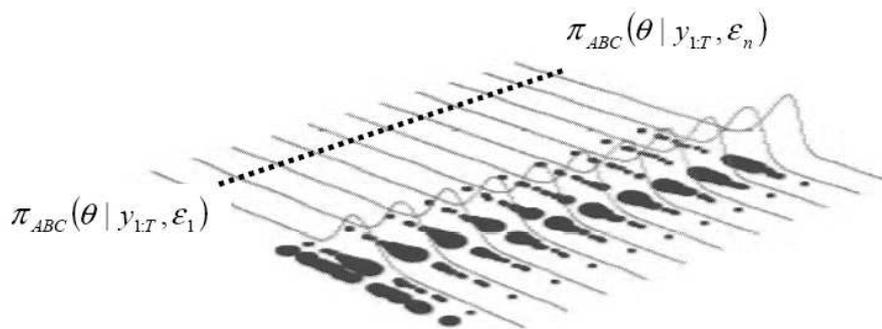


Fig. 2.3.2: Example of how to define the sequence of distributions in the context of likelihood-free Bayesian inference.

M_n to combat sample degeneracy resulting from a less than optimal choice for M_n . It does this through a rejection mechanism, resulting in a new mutation kernel which we denote by M_n^* . As part of the PRC stage we define a sequence of threshold rejection levels, c_n , that the particle weights must pass at each time n to remain in the sample. Presentation and discussion of this is a key part of the papers in the first part of this thesis and so it is deferred to Part I.

2.3.2 Trans-dimensional Markov chain Monte Carlo

This section presents a sampling methodology specifically designed to create a reversible Markov chain with stationary distribution defined over several possible model subspaces. Simultaneous inference on the model choice and the parameter space is a significant aspect of modern statistical practice. In general, for an observed dataset $\mathbf{y} = (y_1, \dots, y_n)$, one may consider a countable set of models $\mathcal{M} = \{G_1, G_2, \dots\}$ indexed by a parameter $k \in \mathcal{K}$, each with a parameter vector $\boldsymbol{\theta}_k \in \Theta_k$ of length n_k . When developing a Bayesian framework for such a setting it is natural to consider the joint posterior distribution of the model and model indicator pair $\pi(\boldsymbol{\theta}_k, k | \mathbf{y})$. This is particularly relevant for settings in which one is performing Bayesian model selection (BMOS) or model averaging (BMA). Discussion on aspects of model selection and model averaging in a Bayesian context are presented in several papers in Parts I, II and III of this thesis. As such the focus of this section will involve a brief overview of Trans-dimensional MCMC methodology.

It is fair to say that in the majority of cases, analytical computations on this joint model and parameter posterior are generally unavailable. Therefore, numerical sampling algorithms are commonly utilized. Designing valid numerical algorithms in this setting has typically focused on Markov chain Monte Carlo methodology, with the exception of more recent SMC based approaches, see [Jasra *et al.* (2008)](65), [Jasra *et al.* (2007)](65), [Peters (2005)](88) and [Godsill *et al.* (2004)](54).

This general extended framework, including multiple models, means that one must now work with a target probability measure, denoted generically by $\pi(d\boldsymbol{\theta})$, and a proposal kernel, $q(\boldsymbol{\theta}, d\phi)$, since comparing densities in different dimensions has no real meaning. Working with distributions instead of densities ensures that one only makes comparisons under the same volume measure. As a result of this realization, a sequence of key papers in this literature written by [Green (1995)](58) and [Green (2003)](59), introduce a general framework on which to consider developing Markov chain based samplers for such general state spaces and posterior distributions.

A key idea that Green's (1995) paper and subsequent works has pioneered is the disjoint union of subspaces formulation. Instead of doing the model search in the full product space that would arise if one sampled over the model indicator and the parameters, alternatively one could focus on spaces of the form $\{(k, \boldsymbol{\theta}_k)\} = \cup_{k \in \mathbf{K}} (\{k\} \times \Theta_k)$. The target distribution defined on such a space is then given by,

$$\pi(d\boldsymbol{\theta}, k | \mathbf{y}) = \sum_{m=1}^M \pi(d\boldsymbol{\theta}_m, m | \mathbf{y}) \mathbb{I}_{\Theta_m \times \{m\}}(\boldsymbol{\theta}, k), \quad (2.3.3)$$

where the size of set \mathbf{K} is denoted by M .

In the remainder of this section the simplest version of TDMCMC samplers, the popular reversible jump MCMC (RJ-MCMC) algorithm of [Green (1995)](58), is briefly discussed. The more advanced discussion presenting explicit details and the automation of such algorithms

is deferred until their presentation in research papers in this thesis. These research papers will consider approaches, which attempt to automate choices for this sampler, to improve the mixing rate of the Markov chain created. This in turn will allow the Markov chain to explore efficiently different model subspaces under consideration.

The RJ-MCMC sampler is effectively a restatement of the Metropolis-Hastings algorithm in terms of the general state space structure. It permits exploration of the joint parameter and model indicator space via a single Markov chain. In the same manner as a standard MCMC algorithm, the transition kernel, containing a proposal, is designed to propose a move from the current state to a new state. The new proposed state of the Markov chain is then accepted according to a Metropolis-Hastings acceptance probability. Different formulations and the exact form of each of their acceptance probabilities and proposal mechanisms is found in papers contained in Parts I and III of this thesis.

The primary difficulty with the reversible jump sampler lies with the efficient construction of transitions between differing models. In general, moving between G_i and G_j is problematic as Θ_i and Θ_j may differ in dimensionality and interpretation, and so designing an efficient proposal can be difficult.

Efficiency of the reversible jump sampler is highly dependent on the choice of proposal distribution. Several papers contained in Part I and III of this thesis directly address this topic.

2.4 Literature review and context of advanced Monte Carlo methods

The intention of this section is to provide a brief high level literature review for the developments of generic SMC algorithms, referring to several key overview papers for more details. The focus will be particularly on the SMC Samplers algorithm and its development as it is relevant to this thesis. In addition to this there will also be a separate section detailing the literature associated with TDMCMC algorithms.

2.4.1 Developments in Sequential Monte Carlo Samplers

There is now a vast literature on non-linear, non-Gaussian filtering methodology, models and theory. The most popular numerical approach and methodology involves Sequential Monte Carlo (SMC) techniques. Sequential Monte Carlo (SMC) algorithms have originally been popularized and extensively developed in the fields of engineering, stochastic control, telecommunications, computer vision and target tracking literatures. The algorithm appears under the names of sequential Monte Carlo, particle filtering and interacting particle systems.

Due to the popularity of these algorithms and the extensive effort made to extend the basic methodology of the early paper of [Gordon *et al.* (1993)] (55) on Sequential Importance Sampling (SIS) and Sequential Importance Resampling (SIR) filters, there has been a range of excellent literature reviews. The interested reader is referred to the following papers and books,

[Arulampalam *et al.* (2002)] (6); [Cappe *et al.* (2007)] (21); [Fearnhead (2008)] (43); [Künsch (2001)] (71) and the book length discussions of [Liu (2001)] (75) and [Doucet *et al.* (2001)] (34).

In parallel with many of the developments in the engineering literature there were extensive theoretical developments regarding this class of algorithms. These were made under the name of Feynman-Kac Interacting Particle Systems, see the text [Del Moral (2004)] (29) for an excellent presentation.

The development of SMC Samplers began with the methodological work of [Peters (2005)] and [Del Moral *et al.* (2006)] (30). The original SMC Samplers ideas were then extended in papers such as [Vermaak *et al.* (2003)] (118) who developed methodology combining SMC Samplers and a trans-dimensional sampler. Applications of SMC Samplers methodology include the rare event modelling methodology of [Johansen *et al.* (2006)] (66) and the application to Bayesian optimal nonlinear design presented in [Kueck *et al.* (2007)] (70). In terms of SMC Samplers and their application in likelihood-free methodology, the paper of [Sisson *et al.* (2007)] (110) originated this connection and this has since been extended significantly, see the papers contained in Part 1 of this thesis.

2.4.2 *Developments in trans-dimensional samplers*

There has been a range of papers presenting methodology and application of TDMCMC, since the invention of the RJMCMC algorithm of [Green (1995)] (58). The formalization of a Markov chain, that moves across models and parameter spaces, allowed for the Bayesian processing of a wide variety of new models. This in turn added to its success in applications.

Alternative TDMCMC approaches include the work of [Gelfand *et al.* (1994)], [Carlin *et al.* (1995)] (22) and [Grenander *et al.* (1994)] (61). The generalized balance condition on cross-model Markov kernels in [Green (1995)] (58) gives a generic setup for exploring variable dimension spaces, even when the number of models under comparison is infinite. Some papers, that provide overviews, extensions and insights to TDMCMC methodology, can be obtained from papers such as [Sisson (2005)] (108), [Green (2003)] (59), [Robert *et al.* (2008)] (100), [Godsill (2003)] (53), [Godsill (2001)] (52) and [Brooks *et al.* (2003)] (18). There are many other excellent papers dealing with both applications and methodological extensions. Many of these are discussed and mentioned in either the papers in Parts 1 and 3 dealing with TDMCMC or in the review papers mentioned.

2.5 Contribution Part I

Journal Papers:

Paper 1: **Sisson S.A., Peters G.W., Fan Y. and Briers M. (2008) "Likelihood Free Samplers". In review.**

This paper presents a detailed overview of likelihood-free Bayesian methodology. In particular it is the first paper to clearly enunciate the different approaches to specification of a Bayesian likelihood-free framework, either on the joint space or on the marginal space. It establishes a mathematical basis for this dual representation and then makes the appropriate links to other papers in the literature explaining clearly how each fits into such a framework. In doing this, it clears up misconceptions and ambiguity present in the existing literature on ABC methodology and the Monte Carlo algorithms developed for such models.

In addition, the paper makes explicit the direct relationships between several sequential Monte Carlo samplers in the likelihood-free setting. These links have not previously been made explicit. This helps lead to identification and understanding regarding a number of factors affecting different sampler performances in the ABC context.

Paper 2: **Peters G.W., Fan Y. and Sisson S.A. (2008) "On Sequential Monte Carlo, Partial Rejection Control and Approximate Bayesian Computation". In review.**

This paper presents the technical details for a novel class of Sequential Monte Carlo Samplers algorithms, in which the mutation kernel is modified to incorporate a Partial Rejection Control (PRC) stage. The theoretical and methodological aspects of such a modification are explored. These include detailed specification of the mutation kernel under PRC, including practical and methodological details to efficiently use such a mutation kernel in SMC Samplers methodology. This is particularly relevant to avoiding the calculation of the mutation kernels normalizing constant in the calculation of the incremental importance sampling weights. In addition, specification of the incremental importance sampling weights and analysis of the asymptotic properties of this SMC Samplers PRC algorithm are explored. This includes, analysis of the impact of PRC versus no PRC stage on the variance of the incremental importance sampling weights under different choices in the SMC Samplers algorithms mutation and backward kernels. Analysis of the number of rejection attempts in the PRC stage is studied. A Central Limit Theorem is established after re-interpretation of the SMC Samplers PRC algorithm in terms of an existing rare-event sampling particle algorithm. Finally, a recursive expression for the path space asymptotic variance in the CLT is obtained which clearly demonstrates the role played by the PRC stage of the mutation kernel. Bounds are provided for the asymptotic variance under PRC versus no PRC stage.

Next, the SMC Samplers PRC algorithm is extended to the ABC context. This provides a mathematical justification for the algorithm developed in [Sisson *et al.* (2007)](110). The

paper then demonstrates how the approach of [Sisson *et al.* (2007)] relates to a proper SMC Samplers PRC framework. Finally, a family of SMC Samplers PRC-ABC algorithms is presented along with analysis of their theoretical properties, leading to practical guidelines for implementation and specification of algorithm settings. These findings are demonstrated on both a synthetic and an actual data example.

Paper 3: **Peters G.W., Sisson S.A. and Fan Y. (2009) "Bayesian Alpha Stable models via Likelihood Free Inference". In review.**

This paper sets out to develop a general Bayesian model and sampling framework to work with the family of univariate and multivariate α -stable distributions. This involves development of likelihood-free inference methodology in this setting, which is not dependent on the parameterization of the α -stable family that one is working with, only that simulation of data from this parameterization is efficient. This is the first time a general Bayesian model has been developed for both the univariate and multi-variate setting. Additionally, it is a novel development of likelihood-free modelling methodology. The sampling approach developed involves a SMC Samplers PRC-ABC algorithm for both the univariate and multivariate models. In the univariate setting the SMC Samplers PRC-ABC approach is compared to alternative univariate samplers based on MCMC.

In developing this framework, several aspects of α -stable models are explored, with particular attention paid to summary statistics in univariate setting. In the multi-variate setting this will be the first time a general multi-variate Bayesian model is produced with novel developments for the summary statistics combining the projection techniques with univariate summary statistics. Finally, the models and sampling methodology are explored on both synthetic and actual data sets.

Paper 4: **Fan Y., Peters G.W., Sisson S.A. (2009) "Automating and Evaluating Reversible Jump MCMC Proposal Distributions". *Statistics and Computing*, 19, 401-429.**

This paper explores the design of a probabilistic proposal mechanism for the transition kernel of a TD-MCMC algorithm. The intention is twofold, firstly development of an automated TDMCMC proposal mechanism using conditional path sampling (CPS) is explored and developed. Particular attention is paid to exploration of the efficiency of a CPS proposal mechanism, this involves studying several aspects involved in designing such a proposal. Secondly, such a CPS proposal mechanism is examined as a tool for analysis of between model proposal mechanism design. Finally, the paper demonstrates the methodology compared to two popular algorithms in the literature for a Autoregressive time series model and a mixture model example. The CPS proposal results are compared to RJ-MCMC alternatives on real data sets.

Paper 5: **Peters G.W., Balakrishnan K., Lasscock B. and Mellen C. (2009) "Model selection and adaptive Markov chain Monte Carlo for Bayesian cointegrated VAR models". In review.**

This paper develops a matrix-variate adaptive Markov chain Monte Carlo (MCMC) methodology for Bayesian Cointegrated Vector Auto Regressions (CVAR). We replace the popular approach to sampling Bayesian CVAR models, involving griddy Gibbs, with an automated efficient alternative, based on the Adaptive Metropolis algorithm. Developing the adaptive MCMC framework for Bayesian CVAR models allows for efficient estimation of posterior parameters in significantly higher dimensional CVAR series than previously possible with existing griddy Gibbs samplers. For a n -dimensional CVAR series, the matrix-variate posterior is in dimension $3n^2 + n$, with significant correlation present between the blocks of matrix random variables. Hence, utilising a griddy Gibbs sampler for large n becomes computationally impractical as it involves approximating a $n \times n$ full conditional posterior using a spline over a high dimensional $n \times n$ grid. The adaptive MCMC approach is demonstrated to be ideally suited to learning on-line a proposal to reflect the posterior correlation structure, therefore improving the computational efficiency of the sampler.

We also treat the rank of the CVAR model as a random variable and perform joint inference on the rank and model parameters. This is achieved with a Bayesian posterior distribution defined over both the rank and the CVAR model parameters, and inference is made via Bayes Factor analysis of rank.

Practically the adaptive sampler also aids in the development of automated Bayesian cointegration models for algorithmic trading systems considering instruments made up of several assets, such as currency baskets. Previously the literature on financial applications of CVAR trading models typically only considers pairs trading ($n=2$) due to the computational cost of the griddy Gibbs. We are able to extend under our adaptive framework to $n \gg 2$ and demonstrate an example with $n = 10$, resulting in a posterior distribution with parameters upto dimension 310. By also considering the rank as a random quantity we can ensure our resulting trading models are able to adjust to potentially time varying market conditions in a coherent statistical framework.

3

Journal Paper 1

"Science is built up of facts, as a house is built up of stones; but an accumulation of facts is no more a science than a heap of stones is a house."

Henri Poincaré

Sisson S.A., Peters G.W., Fan Y. and Briers M. (2008) "Likelihood Free Samplers". In review.

This work was instigated by the first three authors. The second author on this major paper can claim around 50% of the credit for the contents. His work included developing large amounts of the methodology contained, in particular related to the Sequential Monte Carlo Samplers PRC-ABC algorithm. This paper has been submitted for review and is expected to have a significant impact in the statistics community and is already receiving feedback whilst it was a technical report and in revision for the journal. It has been revised three times to make it more appropriate for the audience and to address reviewer comments in this rapidly changing field. Permission from all the co-authors has been granted for submission of this paper as part of the thesis.

Likelihood-free samplers

S. A. Sisson (*corresponding author*)

UNSW Mathematics and Statistics Department, Sydney, 2052, Australia

Gareth W. Peters

CSIRO Mathematical and Information Sciences, Locked Bag 17, North Ryde, NSW, 1670, Australia;

UNSW Mathematics and Statistics Department, Sydney, 2052, Australia

Y. Fan

UNSW Mathematics and Statistics Department, Sydney, 2052, Australia

M. Briers

QinetiQ Ltd., Malvern, Worcestershire, WR14 3PS, UK

Submitted: 16 November 2009

Abstract Methods for Bayesian simulation in the presence of computationally intractable likelihood functions are of growing interest. Termed *likelihood-free samplers*, standard simulation algorithms have been adapted for this setting. In this article we demonstrate that for likelihood-free samplers, there is a previously unnoticed ambiguity over the form of target distribution: in particular whether samples are obtained from the joint distribution of model parameters and auxiliary datasets, or from the marginal distribution of model parameters only. We consider sampler validity under each target distribution. While we show that this ambiguity does not lead to different algorithms, many samplers are not strictly valid under the marginal target interpretation.

Keywords: Approximate Bayesian computation; Likelihood-free computation; Rejection Sampling; Markov chain Monte Carlo; Sequential Monte Carlo.

3.1 Introduction

Bayesian inference proceeds via the posterior distribution $\pi(\theta|y) \propto \pi(y|\theta)\pi(\theta)$, the updating of prior information $\pi(\theta)$ for a parameter $\theta \in \Theta$ through the likelihood function $\pi(y|\theta)$ after observing data $y \in \mathcal{X}$. Numerical algorithms, such as importance sampling, Markov chain Monte Carlo (MCMC) and sequential Monte Carlo (SMC), are commonly employed to draw samples from the posterior $\pi(\theta|y)$.

There is growing interest in posterior simulation in situations where the likelihood function is computationally intractable i.e. $\pi(y|\theta)$ may not be numerically evaluated pointwise. As a result, sampling algorithms based on repeated likelihood evaluations require modification for this task. Collectively known as *likelihood-free samplers* these methods have been developed across multiple disciplines and literatures. However a previously unnoticed ambiguity over the form of the target distribution has resulted in a lack of clarity over the validity of many likelihood-free samplers.

In this article we present general forms of two likelihood-free models, and extend earlier likelihood-free samplers (based on rejection sampling and MCMC) to these models. In doing so, we demonstrate that likelihood-free samplers are, in general, ambiguous over the exact form of their target distribution. In Section 3.2 we establish the notation and models underlying likelihood-free methods. In Section 3.3 we consider importance sampling, MCMC and SMC algorithms in turn, and discuss sampler validity and algorithm equivalence under both target distributions. We conclude with a summary and discussion in Section 3.4.

3.2 Models for computationally intractable likelihoods

In essence, likelihood-free methods embed the intractable target posterior $\pi(\theta|y)$ within an augmented model from which sampling is viable. Specifically the joint posterior of the model parameters θ , and auxiliary data x given observed data y is

$$\pi(\theta, x|y) \propto \pi(y|x, \theta)\pi(x|\theta)\pi(\theta), \quad (3.2.1)$$

where $x \in \mathcal{X}$ is may be interpreted as a dataset simulated according to the model $x \sim \pi(x|\theta)$. Assuming such simulation is possible, data-generation under the model forms the basis of computation in the likelihood-free setting – see Section 3.3. The target marginal posterior $\pi_M(\theta|y)$ for the parameters θ , is then obtained as

$$\pi_M(\theta|y) \propto \int_{\mathcal{X}} \pi(y|x, \theta)\pi(x|\theta)\pi(\theta)dx \quad (3.2.2)$$

(e.g. 11; 16). The function $\pi(y|x, \theta)$ weights the intractable posterior, with high weight in regions $x \approx y$ where auxiliary and observed datasets are similar (or more usually, where $T(x) \approx T(y)$ are similar, where $T(\cdot)$ denotes a vector of sufficient or summary statistics). As

such, $\pi_M(\theta|y) \approx \pi(\theta|y)$ forms an approximation to the intractable posterior. In the case where $\pi(y|x, \theta)$ is a point mass at $T(y) = T(x)$ and is zero elsewhere, if $T(\cdot)$ is sufficient for θ then the intractable posterior marginal $\pi_M(\theta|y) = \pi(\theta|y)$ is recovered exactly. Typically, $\pi(y|x, \theta)$ is a standard smoothing kernel, and various choices have been examined (e.g. 7; 2; 8; 3).

For our discussion on likelihood-free samplers, it is convenient to consider a generalization of the joint distribution (3.2.1) incorporating $S \geq 1$ auxiliary datasets

$$\pi_J(\theta, x_{1:S}|y) \propto \pi(y|x_{1:S}, \theta)\pi(x_{1:S}|\theta)\pi(\theta)$$

where $x_{1:S} = (x^1, \dots, x^S)$, and $x^1, \dots, x^S \sim \pi(x|\theta)$ are S independent datasets generated from the (intractable) model. As the auxiliary datasets are, by construction, conditionally independent given θ , we have $\pi(x_{1:S}|\theta) = \prod_{s=1}^S \pi(x^s|\theta)$. One choice of $\pi(y|x_{1:S}, \theta)$ (5) produces the joint posterior

$$\pi_J(\theta, x_{1:S}|y) \propto \left[\frac{1}{S} \sum_{s=1}^S \pi(y|x^s, \theta) \right] \left[\prod_{s=1}^S \pi(x^s|\theta) \right] \pi(\theta), \quad (3.2.3)$$

where in (3.2.3) we generalize the uniform choice of $\pi(y|x^s, \theta)$ by (5) to the general case. It is easy to see that $\int_{\mathcal{X}^S} \pi_J(\theta, x_{1:S}|y) dx_{1:S} = \pi_M(\theta|y)$ admits the distribution (3.2.2) as a marginal distribution. The case $S = 1$ with $\pi_J(\theta, x_{1:S}|y) = \pi(\theta, x|y)$ corresponds to the more usual joint posterior (3.2.1) in the likelihood-free setting.

There are two obvious approaches to posterior simulation from $\pi_M(\theta|y) \approx \pi(\theta|y)$. The first approach proceeds by sampling directly on the augmented model $\pi_J(\theta, x_{1:S}|y)$, realizing joint samples $(\theta, x_{1:S}) \in \Theta \times \mathcal{X}^S$ before *a posteriori* marginalization over $x_{1:S}$ (i.e. by discarding the x^s or $T(x^s)$ realizations from the sampler output). The second approach is to sample from $\pi_M(\theta|y)$ directly by approximating the integral (3.2.2) via Monte Carlo integration in lieu of each posterior evaluation of $\pi_M(\theta|y)$. In this case

$$\pi_M(\theta|y) \propto \pi(\theta) \int_{\mathcal{X}} \pi(y|x, \theta)\pi(x|\theta)dx \approx \frac{\pi(\theta)}{S} \sum_{s=1}^S \pi(y|x^s, \theta) := \hat{\pi}_M(\theta|y), \quad (3.2.4)$$

where $x^1, \dots, x^S \sim \pi(x|\theta)$. This expression, examined by various authors (e.g. 7; 11; 12; 10; 14), requires multiple generated datasets x^1, \dots, x^S , for each evaluation of the marginal posterior $\pi_M(\theta|y)$. As with standard Monte Carlo approximations, $\text{Var}[\hat{\pi}_M(\theta|y)]$ reduces as S increases, with $\lim_{S \rightarrow \infty} \text{Var}[\hat{\pi}_M(\theta|y)] = 0$.

We now examine the relationships between, and validity of, likelihood-free samplers constructed with $\pi_M(\theta|y)$ and $\pi_J(\theta, x_{1:S}|y)$ as the target distribution.

3.3 Sampler ambiguity and validity

In this section we examine each of the basic sampler types: rejection sampling, MCMC and population-based methods. We extend the first two of these algorithms to multiple data gen-

erations ($S \geq 1$). We will examine sampler validity with respect to the two target posterior distributions $\pi_M(\theta|y)$ and $\pi_J(\theta, x_{1:S}|y)$, and demonstrate algorithm equivalence under both the joint and marginal distributional targets.

3.3.1 Rejection samplers

Rejection-based likelihood-free samplers were developed in the population genetics literature (13; 9; 7). Table 3.1 presents a generalization of the rejection sampling algorithm. The specific case of $S = 1$ is the original implementation of the sampler. We now demonstrate that this algorithm has both $\pi_M(\theta|y)$ and $\pi_J(\theta, x_{1:S}|y)$ as valid target distributions.

LF-REJ Algorithm

1. Generate $\theta \sim \pi(\theta)$ from the prior.
2. Generate $x^1, \dots, x^S \sim \pi(x|\theta)$ independently from the model.
3. Accept θ with probability proportional to $\frac{1}{S} \sum_{s=1}^S \pi(y|x^s, \theta)$

Tab. 3.1: The generalized likelihood-free rejection sampling (LF-REJ) algorithm.

Assuming the joint model target $\pi_J(\theta, x_{1:S}|y)$, under the LF-REJ algorithm a sample $(\theta, x_{1:S})$ is first drawn from the prior predictive distribution $\pi(\theta, x_{1:S}) = \pi(\theta) \prod_{s=1}^S \pi(x^s|\theta)$ (steps 1 and 2). The acceptance probability (step 3) for $(\theta, x_{1:S})$ is

$$\frac{\pi_J(\theta, x_{1:S}|y)}{\pi(\theta, x_{1:S})} = \frac{1}{S} \sum_{s=1}^S \pi(y|x^s, \theta)$$

as indicated in Table 3.1. *A posteriori* marginalization over $x_{1:S} \in \mathcal{X}^S$ (by discarding the $x_{1:S}$ realizations) then provides draws from $\pi_M(\theta|y)$.

Assuming the marginal model target, $\pi_M(\theta|y)$, a sample θ is first drawn from the prior (step 1). The acceptance probability for this sample is then approximated as

$$\frac{\pi_M(\theta|y)}{\pi(\theta)} \approx \frac{\hat{\pi}_M(\theta|y)}{\pi(\theta)} = \frac{1}{S} \sum_{s=1}^S \pi(y|x^s, \theta),$$

using the draws x^1, \dots, x^S from the model (steps 2 and 3). Note that while $\hat{\pi}_M(\theta|y)/\pi(\theta)$ is an approximation of the acceptance probability, it is unbiased for all $S \geq 1$. Hence, while smaller S will result in more variable estimates of the acceptance probability, the accepted samples will still correspond to a draws from $\pi_M(\theta|y)$ for all $S \geq 1$. Hence the LF-REJ algorithm is valid for both targets $\pi_J(\theta, x_{1:S}|y)$ and $\pi_M(\theta|y)$.

3.3.2 Markov chain Monte Carlo samplers

MCMC-based likelihood-free samplers were introduced to avoid rejection sampling inefficiencies when the posterior and prior were sufficiently different (7; 4). The generalized likelihood-

free MCMC algorithm for $S \geq 1$ is presented in Table 3.2. Again, $S = 1$ is the original implementation of this sampler.

LF-MCMC Algorithm

- Initialize θ_0 (and $x_{1:S,0} = (x_0^1, \dots, x_0^S)$ with $x_0^i \sim \pi(x|\theta_0)$ drawn from the model)
 At stage $t \geq 1$
 1. Generate $\theta \sim q(\theta_t, \theta)$ from a proposal distribution.
 2. Generate $x_{1:S} = (x^1, \dots, x^S)$ with $x^i \sim \pi(x|\theta)$ independently from the model.
 3. With probability $\min \left\{ 1, \frac{\frac{1}{S} \sum_s \pi(y|x^s, \theta) \pi(\theta) q(\theta, \theta_t)}{\frac{1}{S} \sum_s \pi(y|x_t^s, \theta_t) \pi(\theta_t) q(\theta_t, \theta)} \right\}$ accept $\theta_{t+1} = \theta$, ($x_{1:S,t+1} = x_{1:S}$)
 otherwise set $\theta_{t+1} = \theta_t$, ($x_{1:S,t+1} = x_{1:S,t}$).
 4. Increment $t = t + 1$ and go to 1.
-

Tab. 3.2: The generalized likelihood-free MCMC (LF-MCMC) algorithm. Statements in parentheses involving $x_{1:S}$ relate to sampler with target $\pi_J(\theta, x_{1:S}|y)$.

The LF-MCMC sampler was introduced in the context of targeting the marginal model $\pi_M(\theta|y)$. The papers (7) and (15) (for $S = 1$) present variations on proofs of detailed balance under this assumption. We now demonstrate that for finite (i.e. practical values of) S , the LF-MCMC sampler is only theoretically valid under the joint target posterior $\pi_J(\theta, x_{1:S}|y)$, but is practically unbiased under both $\pi_M(\theta|y)$ and $\pi_J(\theta, x_{1:S}|y)$.

Implementing the LF-MCMC sampler assuming the marginal posterior target $\pi_M(\theta|y)$, and a proposal density $q(\theta_t, \theta)$ for θ , the probability of accepting the move from θ_t at time t to a proposed value $\theta \sim q(\theta_t, \theta)$ is given by

$$\min \left\{ 1, \frac{\pi_M(\theta|y)q(\theta, \theta_t)}{\pi_M(\theta_t|y)q(\theta_t, \theta)} \right\} \approx \min \left\{ 1, \frac{\frac{1}{S} \sum_s \pi(y|x^s, \theta) \pi(\theta) q(\theta, \theta_t)}{\frac{1}{S} \sum_s \pi(y|x_t^s, \theta_t) \pi(\theta_t) q(\theta_t, \theta)} \right\}.$$

Unlike rejection sampling, where the acceptance probability is proportional to an unbiased estimate $\hat{\pi}_M(\theta|y)/\pi(\theta)$, the above Markov chain acceptance probability consists of a ratio of two unbiased estimates $\hat{\pi}_M(\theta|y)/\hat{\pi}_M(\theta_t|y)$. As such, the estimate of the acceptance probability (involving this ratio) is biased. Only as $S \rightarrow \infty$ so that the bias of the ratio diminishes, can this algorithm target the marginal posterior $\pi_M(\theta|y)$. Many authors (e.g. 7; 4; 15 and others) implement the LF-MCMC algorithm with $S = 1$, which is apparently too small to result in an unbiased sampler targeting $\pi_M(\theta|y)$.

Within an MCMC algorithm targeting the joint posterior $\pi_J(\theta, x_{1:S}|y)$, the probability of accepting a proposed move from $(\theta_t, x_{1:S,t})$ at time t to

$$(\theta, x_{1:S}) \sim q[(\theta_t, x_{1:S,t}), (\theta, x_{1:S})] = q(\theta_t, \theta) \prod_{s=1}^S \pi(x^s|\theta)$$

at time $t + 1$, is then

$$\min \left\{ 1, \frac{\pi_J(\theta, x_{1:S}|y)q[(\theta, x_{1:S}), (\theta_t, x_{1:S,t})]}{\pi_J(\theta_t, x_{1:S,t}|y)q[(\theta_t, x_{1:S,t}), (\theta, x_{1:S})]} \right\} = \min \left\{ 1, \frac{\frac{1}{S} \sum_s \pi(y|x^s, \theta) \pi(\theta) q(\theta, \theta_t)}{\frac{1}{S} \sum_s \pi(y|x_t^s, \theta_t) \pi(\theta_t) q(\theta_t, \theta)} \right\}.$$

Hence, an LF-MCMC sampler targeting $\pi_J(\theta, x_{1:S}|y)$ results in the same algorithm as a LF-MCMC sampler targeting $\pi_M(\theta|y)$, for any $S \geq 1$. Thus, all applications of marginal LF-MCMC samplers are in practice unbiased for $S \geq 1$, in that the sampler produces correlated draws from $\pi_M(\theta|y)$. However, this validity is strictly only available through that conveyed by the equivalent sampler targeting $\pi_J(\theta, x_{1:S}|y)$.

3.3.3 Population-based samplers

Population-based likelihood-free samplers were introduced to circumvent poor mixing in MCMC samplers (12; 14; 1; 8; 5). These samplers propagate a population of *particles*, $\theta^{(1)}, \dots, \theta^{(N)}$, with associated importance weights $W(\theta^{(i)})$, through a sequence of related densities $\phi_1(\theta_1), \dots, \phi_n(\theta_n)$, defining a smooth transition from ϕ_1 , from which direct sampling is available, to ϕ_n the target distribution. For likelihood-free samplers, ϕ_k is defined by allowing $\pi_k(y|x, \theta)$ to place greater density on regions for which $T(y) \approx T(x)$ as k increases. Hence, we denote $\pi_{J,k}(\theta, x_{1:S}|y) \propto \pi_k(y|x_{1:S}, \theta)\pi(x_{1:S}|\theta)\pi(\theta)$ and $\pi_{M,k}(\theta|y) \propto \pi(\theta) \int_{\mathcal{X}^S} \pi_k(y|x_{1:S}, \theta)\pi(x_{1:S}|\theta)dx_{1:S}$ for $k = 1, \dots, n$, under the joint and marginal posterior models respectively.

Sequential Monte Carlo-based samplers

Under the sequential Monte Carlo samplers algorithm (6) the particle population θ_{k-1} at time $k-1$ is mutated to $\phi_k(\theta_k)$ by the kernel $M_k(\theta_{k-1}, \theta_k)$. The weights for the mutated particles θ_k may be obtained as $W_k(\theta_k) = W_{k-1}(\theta_{k-1})w_k(\theta_{k-1}, \theta_k)$ where, for the marginal model sequence $\pi_{M,k}(\theta_k|y)$, the incremental weight is

$$w_k(\theta_{k-1}, \theta_k) = \frac{\pi_{M,k}(\theta_k|y)L_{k-1}(\theta_k, \theta_{k-1})}{\pi_{M,k-1}(\theta_{k-1}|y)M_k(\theta_{k-1}, \theta_k)} \approx \frac{\hat{\pi}_{M,k}(\theta_k|y)L_{k-1}(\theta_k, \theta_{k-1})}{\hat{\pi}_{M,k-1}(\theta_{k-1}|y)M_k(\theta_{k-1}, \theta_k)}. \quad (3.3.1)$$

Here $L_{k-1}(\theta_k, \theta_{k-1})$ is a reverse-time kernel describing the mutation of particles from time k to time $k-1$. Under the joint model $\pi_{J,k}(\theta, x_{1:S}|y)$, with the natural mutation kernel factorization $M_k[(\theta_{k-1}, x_{1:S}^{k-1}), (\theta_k, x_{1:S}^k)] = M_k(\theta_{k-1}, \theta_k) \prod_{s=1}^S \pi(x_k^s|y)$ (and similarly for L_{k-1}), following the form of (3.3.1), the incremental weight is exactly

$$w_k[(\theta_{k-1}, x_{1:S}^{k-1}), (\theta_k, x_{1:S}^k)] = \frac{\frac{1}{S} \sum_s \pi(y|x_k^s, \theta_k)\pi(\theta_k)L_{k-1}(\theta_k, \theta_{k-1})}{\frac{1}{S} \sum_s \pi(y|x_{k-1}^s, \theta_{k-1})\pi(\theta_{k-1})M_k(\theta_{k-1}, \theta_k)}. \quad (3.3.2)$$

Hence, as the incremental weights (3.3.1, 3.3.2) are equal, they induce identical SMC algorithms for both marginal and joint models $\pi_M(\theta|y)$ and $\pi_J(\theta, x_{1:S}|y)$. As a result, while applications of the marginal sampler targeting $\pi_M(\theta|y)$ are theoretically biased for finite $S \geq 1$ (as for the LF-MCMC algorithm, the incremental weights consist of the ratio $\hat{\pi}_{M,k}(\theta_k|y)/\hat{\pi}_{M,k-1}(\theta_{k-1}|y)$), as before, they are in practice unbiased through association with the equivalent sampler on joint space targeting $\pi_J(\theta, x_{1:S}|y)$.

However, a theoretically valid sampler targeting $\pi_M(\theta|y)$, for all $S \geq 1$, can be obtained by

careful choice of the kernel $L_{k-1}(\theta_k, \theta_{k-1})$. For example, (8) use the approximate optimal kernel (6)

$$L_{k-1}(\theta_k, \theta_{k-1}) = \frac{\pi_{M,k-1}(\theta_{k-1}|y)M_k(\theta_{k-1}, \theta_k)}{\int \pi_{M,k-1}(\theta_{k-1}|y)M_k(\theta_{k-1}, \theta_k)d\theta_{k-1}}, \quad (3.3.3)$$

from which the incremental weight (3.3.1) is approximated by

$$\begin{aligned} w_k(\theta_{k-1}, \theta_k) &= \pi_{M,k}(\theta_k|y) / \int \pi_{M,k-1}(\theta_{k-1}|y)M_k(\theta_{k-1}, \theta_k)d\theta_{k-1} \\ &\approx \hat{\pi}_{M,k}(\theta_k|y) / \sum_{i=1}^N W_{k-1}(\theta_{k-1}^{(i)})M_k(\theta_{k-1}^{(i)}, \theta_k). \end{aligned} \quad (3.3.4)$$

Thus under this construction, the weight calculation is now unbiased for all $S \geq 1$, since the approximation $\hat{\pi}_{M,k-1}(\theta|y)$ in the denominator is no longer needed.

In practice, application of SMC samplers in the likelihood-free setting requires the avoidance of severe particle depletion. targeting $\pi_J(\theta, x_{1:S}|y)$, (5) use a standard MCMC kernel in combination with a large number of slowly changing distributions $\pi_{J,k}(\theta, x_{1:S}|y)$ to maintain particle diversity. In an alternative approach targeting $\pi_M(\theta|y)$, (8) probabilistically reject particles with weight below a given threshold. The final form of the weight including the rejection mechanism involves the form (3.3.4), and so is unbiased for all $S \geq 1$.

Alternative population-based samplers

The papers (12), (14) and (1) propose alternative population-based likelihood-free algorithms. While deriving from different sampling frameworks, they are essentially the same sampler and utilize importance-sampling weights of the form (3.3.4). Following the arguments in Section 3.3.1, such samplers validly target $\pi_M(\theta|y)$ and $\pi_J(\theta, x_{1:S}|y)$ for all $S \geq 1$, and produce identical algorithms.

3.4 Discussion

In this article, we have extended some existing likelihood-free samplers to incorporate multiple ($S > 1$) auxiliary data generations, $x_{1:S} \in \mathcal{X}^S$. In doing so, we have established an ambiguity over the target distribution of such samplers, which is problematic from an interpretive perspective. Those algorithms targeting $\pi_M(\theta|y)$, and requiring estimates of likelihood ratios within acceptance probabilities or importance weights, require the number of Monte Carlo draws $S \rightarrow \infty$ to avoid a theoretical bias (see summary in Table 3.3). Fortunately, through an equivalence with a likelihood-free sampler targeting $\pi_J(\theta, x_{1:S}|y)$, inferences performed with the marginal posterior sampler are in practice unbiased. However, this practical unbiasedness does not justify the sampler targeting $\pi_M(\theta|y)$. Such samplers can only be theoretically justified from the perspective of the joint posterior $\pi_J(\theta, x_{1:S}|y)$ given by (3.2.3). Alternative forms of $\pi_J(\theta, x_{1:S}|y)$ may not offer support for the marginal posterior samplers.

| Sampler | Parameter Space | Theoretical Requirements |
|-----------|-----------------|--|
| Rejection | Joint | Any $S \geq 1$ |
| | Marginal | Any $S \geq 1$ |
| MCMC | Joint | Any $S \geq 1$ |
| | Marginal | $S \rightarrow \infty$ |
| SMC | Joint | Any $S \geq 1$ |
| | Marginal | $S \rightarrow \infty$ |
| | Marginal | Any $S \geq 1$ with optimal L_{k-1} kernel |

Tab. 3.3: Summary of valid likelihood-free samplers. Joint space samplers target $\pi_J(\theta, x_{1:S}|y)$ whereas marginal space samplers target $\pi_M(\theta|y)$.

Acknowledgments

SAS and YF are supported by the ARC-DP scheme (DP0664970 and DP0877432). GWP is supported by APAS and CSIRO CMIS. GWP thanks M. Wüthrich for useful discussion, and ETH FIM and P. Embrechts for financial assistance. MB would like to thank the UK MoD for funding through the DIF Defence Technology Centre. This work was partially supported by the NSF under Grant DMS-0635449 to SAMSI.

References

- [1] Beaumont, M. A.; Cornuet, J.-M.; Marin, J.-M. Robert, C. P. Adaptivity for ABC algorithms: the ABC-PMC scheme *Biometrika*, 2009, In press.
- [2] Beaumont, M. A.; Zhang, W. Balding, D. J. Approximate Bayesian computation in population genetics *Genetics*, 2002, 162, 2025 - 2035
- [3] Blum, M. G. B. Approximate Bayesian computation: a non-parametric perspective Université Joseph Fourier, Grenoble, France, 2009
- [4] Bortot, P.; Coles, S. G. Sisson, S. A. Inference for stereological extremes *Journal of the American Statistical Association*, 2007, 102, 84-92.
- [5] Moral, P. D.; Doucet, A. Jasra, A. Adaptive sequential Monte Carlo samplers University of Bordeaux, 2008
- [6] Moral, P. D.; Doucet, A. Jasra, A. Sequential Monte Carlo samplers *J. R. Statist. Soc. B*, 2006, 68, 411 - 436
- [7] Marjoram, P.; Molitor, J.; Plagnol, V. Tavaré, S. Markov chain Monte Carlo without likelihoods *Proc. Natl. Acad. Sci. USA*, 2003, 100, 15324 - 15328
- [8] Peters, G. W.; Fan, Y. Sisson, S. A. On sequential Monte Carlo, partial rejection control and approximate Bayesian computation. Tech. report, UNSW. UNSW, 2008
- [9] Pritchard, J. K.; Seielstad, M. T.; Perez-Lezaun, A. Feldman, M. W. Population growth of human Y chromosomes: A study of Y chromosome microsatellites *Molecular Biology and Evolution*, 1999, 16, 1791-1798
- [10] Ratmann, O.; Andrieu, C.; Hinkley, T.; Wiuf, C. Richardson, S. Model criticism based on likelihood-free inference, with an example in protein network evolution *Proc. Natl. Acad. Sci. USA*, Imperial College London, 2009, 106, 10576-10581
- [11] Reeves, R. W. Pettitt, A. N. A theoretical framework for approximate Bayesian computation 2005
- [12] Sisson, S. A.; Fan, Y. Tanaka, M. M. Sequential Monte Carlo without likelihoods *Proc. Natl. Acad. Sci.*, 2007, 104, 1760-1765. Errata (2009), 106, 16889.

-
- [13] Tavaré, S.; Balding, D. J.; Griffiths, R. C. Donnelly, P. Inferring coalescence times from DNA sequence data *Genetics*, 1997, 145, 505 - 518
 - [14] Toni, T.; Welch, D.; Strelkowa, N.; Ipsen, A. Stumpf, M. P. H. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems *J. R. Soc. Interface*, 2009, 6, 187-202
 - [15] Wegmann, D.; Leuenberger, C. Excoffier, L. Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood *Genetics*, 2009, 182, 1207-1218
 - [16] Wilkinson, R. D. Approximate Bayesian computation (ABC) gives exact results under the assumption of model error *Univ. of Sheffield*, 2008

4

Journal Paper 2

"Anyone who attempts to generate random numbers by deterministic means is, of course, living in a state of sin."

John von Neumann

Peters G.W., Fan Y. and Sisson S.A. (2008) "On Sequential Monte Carlo, Partial Rejection Control and Approximate Bayesian Computation". In review.

This work was instigated by the first author and he can claim around 80% of the credit for the contents. His work included developing large amounts of the theory included, in particular related to the Sequential Monte Carlo Samplers PRC-ABC algorithm. Whilst the first version of this paper was rejected, this paper has now been reworked and is in submission to a journal with a more suitable audience. We strongly believe it has a high chance of acceptance as the technical report has already been cited multiple times and the paper has been discussed by a few working groups. Qintq a company in the UK is investigating the use of these novel techniques developed in target tracking applications. Permission from all the co-authors has been granted for submission of this paper as part of the thesis.

On sequential Monte Carlo, partial rejection control and approximate Bayesian computation

Gareth W. Peters (*corresponding author*)

CSIRO Mathematical and Information Sciences, Locked Bag 17, North Ryde, NSW, 1670, Australia;

UNSW Mathematics and Statistics Department, Sydney, 2052, Australia

Y. Fan

UNSW Mathematics and Statistics Department, Sydney, 2052, Australia

S. A. Sisson

UNSW Mathematics and Statistics Department, Sydney, 2052, Australia

Submitted: 30 September 2009

4.1 Abstract

We present a sequential Monte Carlo sampler variant of the partial rejection control algorithm, introduced by (26), and show that this variant can be considered as a sequential Monte Carlo sampler with a modified mutation kernel. We prove that the new sampler can reduce the variance of the incremental importance weights when compared with standard sequential Monte Carlo samplers. We provide a study of theoretical properties of the new algorithm, and make connections with some existing algorithms. Finally, the sampler is adapted for application under the challenging “likelihood free,” approximate Bayesian computation modelling framework, where we demonstrate superior performance over existing likelihood-free samplers.

Keywords: Approximate Bayesian computation; Bayesian computation; Likelihood-free inference; Sequential Monte Carlo samplers; Partial rejection control.

4.2 Introduction

Sequential Monte Carlo (SMC) methods have emerged out of the fields of engineering, probability and statistics in recent years. Variants of the methods sometimes appear under the names of particle filtering or interacting particle systems (e.g. 2; 1; 10; 14), and their theoretical properties have been extensively studied (9; 10; 21).

The standard SMC algorithm involves finding a numerical solution to a set of filtering recursions, such as filtering problems arising from non-linear / non-Gaussian state space models. Under this framework, the SMC algorithm samples from a (often naturally occurring) sequence of distributions π_t , indexed by $t = 1, \dots, T$. Each distribution is defined on the support $E^t = E \times E \times \dots \times E$. (12) (see also 29) generalize the SMC algorithm to the case where the distributions π_t are all defined on the same support E . This generalization, termed the SMC *sampler*, adapts the SMC algorithm to the more popular setting in which the state space E remains static.

In short, the SMC sampler generates weighted samples (termed *particles*) from a sequence of distributions π_t , for $t = 1, \dots, T$, where π_T may be of particular interest. We refer to π_T as the target distribution. Procedurally, particles obtained from an arbitrary initial distribution π_1 , with a set of corresponding initial weights, are sequentially propagated through each distribution π_t in the sequence via three processes, involving mutation (or move), correction (or importance weighting) and selection (or resampling). The final weighted particles at distribution π_T are considered weighted samples from the target distribution π . The mechanism is similar to sequential importance sampling (resampling) (26; 14), with one of the crucial differences being the framework under which the particles are allowed to move, resulting in differences in the calculation of the weights of the particles.

One of the major difficulties with SMC-type algorithms is particle depletion, in which the weights of the majority of the particles gradually decrease to zero, while a few particle weights dominate the population. This severely increases the variability of Monte Carlo estimates of expectations under π . In this article, we develop an algorithm which incorporates the partial rejection control (PRC) strategy of (26) into the SMC sampler framework. A particular motivation for this stems from the recent developments in “likelihood-free” (or approximate Bayesian) computation (2; 28; 43), where an extremely high proportion of mutated particles are expected to have very small, or exactly zero, posterior weights.

In this article, we develop the SMC samplers PRC algorithm, in which the partial rejection control mechanism is built directly into the mutation kernel of the SMC sampler. In this manner, a particle mutation may be rejected if the resulting importance weight is below a certain threshold. We begin with a brief introduction to the standard sequential Monte Carlo sampler in Section 4.3, and then present the SMC sampler PRC algorithm. We also discuss implementational issues arising from the inclusion of the PRC stage, including estimation for the resultant kernel normalizing constant. Section 4.4 provides some theoretical results that justify the addition of PRC in terms of improvements in the variance of the incremental importance weights. We also

discuss a central limit theorem and derive a recursive expression for the asymptotic variance of our algorithm. In addition, we make a novel connection between the SMC sampler PRC algorithm and the AliveSMC algorithm from the rare-event literature developed in (23). In Section 4.5 we adapt the SMC sampler PRC algorithm for application in the likelihood-free modelling framework, and demonstrate the computational gains achieved over existing likelihood-free algorithms via a simulated example. Finally, we present a stochastic claims reserving analysis using the developed methods in Section 4.6, and conclude with a discussion.

4.3 Sequential Monte Carlo and partial rejection

4.3.1 Sequential Monte Carlo sampler

(12) introduced a modification of the sequential Monte Carlo algorithm, termed the sequential Monte Carlo *sampler*. Consider a sequence of distributions $\pi_t(x)$, $t = 1, \dots, T$, with $x \in E$, where the final distribution π_T is the distribution of interest. By introducing a sequence of backward kernels L_k , a new distribution $\tilde{\pi}_t(x_1, \dots, x_t) = \pi_t(x_t) \prod_{k=1}^{t-1} L_k(x_{k+1}, x_k)$ may be defined for the *path* of a particle $(x_1, \dots, x_t) \in E^t$ through the sequence π_1, \dots, π_t . The only restriction on the backward kernels is that the correct marginal distributions $\int \tilde{\pi}_t(x_1, \dots, x_t) dx_1, \dots, dx_{t-1} = \pi_t(x_t)$ are available.

Within this framework, one may then work with the sequence of distributions, $\tilde{\pi}_t$, under the standard SMC algorithm. In summary, the SMC sampler algorithm involves three stages: *mutation*, whereby the particles are moved from x_{t-1} to x_t via a mutation kernel $M_t(x_{t-1}, x_t)$ as described below (15; 12); *correction*, where the particles are reweighted with respect to π_t via the incremental importance weight (4.3.1); and *selection*, where according to some measure of particle diversity, commonly the effective sample size (ESS, 21; 69; 14; 24), the weighted particles may be resampled in order to reduce the variability of the importance weights.

In more detail, suppose that at time $t - 1$, the distribution $\tilde{\pi}_{t-1}$ can be approximated empirically by $\tilde{\pi}_{t-1}^N$ using N weighted particles. These particles are first propagated to the next distribution $\tilde{\pi}_t$ using a mutation kernel $M_t(x_{t-1}, x_t)$, and then assigned new weights $W_t = W_{t-1} w_t(x_1, \dots, x_t)$, where W_{t-1} is the weight of a particle at time $t - 1$ and w_t is the incremental weight given by

$$w_t(x_1, \dots, x_t) = \frac{\tilde{\pi}_t(x_1, \dots, x_t)}{\tilde{\pi}_{t-1}(x_1, \dots, x_{t-1}) M_t(x_{t-1}, x_t)} = \frac{\pi_t(x_t) L_{t-1}(x_t, x_{t-1})}{\pi_{t-1}(x_{t-1}) M_t(x_{t-1}, x_t)}. \quad (4.3.1)$$

The resulting particles are now weighted samples from $\tilde{\pi}_t$. Consequently from (4.3.1), under the SMC sampler framework, one may work directly with the marginal distributions $\pi_t(x_t)$ such that $w_t(x_1, \dots, x_t) = w_t(x_{t-1}, x_t)$. While the choice of the backward kernels L_{t-1} is essentially arbitrary, their specification can strongly affect the performance of the algorithm. See (12) for detailed discussion.

4.3.2 Incorporating partial rejection control

It is well known that the performance of SMC methods are strongly dependent on the mutation kernel (8). If M_t is poorly chosen, such that it does not place particles in regions of the support of π_t with high density, then many importance sampling weights will be close to zero. This leads to sample degeneracy, as a few well located particles with large weights dominate the particle population, resulting in large variance for estimates made using these samples.

(26) (see also 25) introduced the partial rejection control strategy to overcome particle degeneracy in a sequential importance sampling setting. Under this mechanism, when the weight of a particle at distribution π_t falls below a finite threshold, $c_t \geq 0$, the particle is probabilistically discarded. It is replaced with a particle drawn from the previous distribution π_{t-1} which is then mutated to π_t . This new particle's weight is then compared to the threshold, with this process repeating until a particle is accepted. This approach is termed *partial rejection*, as the replacement particle is drawn from π_{t-1} , not π_1 (25).

Under the SMC sampler framework we modify this approach and incorporate the partial rejection mechanism directly within the mutation kernel. Hence at time $t - 1$, the particle x_{t-1} is moved via the mutation kernel $M_t(x_{t-1}, x_t)$ and weighted according to (4.3.1). This particle is accepted with probability p , determined by the particle's weight and the weight threshold c_t . If rejected, a new particle is obtained via the mutation kernel M_t , until a particle is accepted.

For the sequence of distributions π_t , $t = 1, \dots, T$, the mutation and backward kernels M_t and L_{t-1} , a sequence of weight thresholds c_t , and PRC normalizing constants $r(c_t, x_{t-1})$ (defined below), the SMC sampler PRC algorithm is given by:

SMC sampler PRC algorithm

Initialization: Set $t = 1$.

For $i = 1, \dots, N$, sample $x_1^{(i)} \sim \pi_1(x)$, and set weights $W_1(x_1^{(i)}) = \frac{1}{N}$.

Resample: Normalize the weights $\sum_i W_t(x_t^{(i)}) = 1$. If $[\sum_i W_t(x_t^{(i)})^2]^{-1} < H$ resample N particles with respect to $W_t(x_t^{(i)})$ and set $W_t(x_t^{(i)}) = \frac{1}{N}$, $i = 1, \dots, N$.

Mutation and correction: Set $t = t + 1$ and $i = 1$:

(a) Sample $x_t^{(i)} \sim M_t(x_{t-1}^{(i)}, x_t)$ and set weight for $x_t^{(i)}$ to

$$W_t(x_t^{(i)}) = W_{t-1}(x_{t-1}^{(i)}) \frac{\pi_t(x_t^{(i)}) L_{t-1}(x_t^{(i)}, x_{t-1}^{(i)})}{\pi_{t-1}(x_{t-1}^{(i)}) M_t(x_{t-1}^{(i)}, x_t^{(i)})}.$$

(b) With probability $1 - p^{(i)} = 1 - \min\{1, W_t(x_t^{(i)})/c_t\}$, reject $x_t^{(i)}$ and go to (a).

(c) Otherwise, accept $x_t^{(i)}$ and set $W_t(x_t^{(i)}) = W_t(x_t^{(i)}) r(c_t, x_{t-1}^{(i)})/p^{(i)}$.

(d) Increment $i = i + 1$. If $i \leq N$, go to (a).

(e) If $t < T$ go to Resample.

The above algorithm without the mutation and correction steps (b) and (c) is equivalent to the standard SMC sampler algorithm (12). In the resample stage, the degeneracy of the par-

title approximation is quantified through the usual estimate of the effective sample size, $1 \leq [\sum_i W_t(x_t^{(i)})^2]^{-1} \leq N$ (24). We discuss the choice of the thresholds, c_t , in later Sections.

The addition of a rejection step at each time t effectively modifies the mutation kernel M_t . We denote the resulting kernel by M_t^* , where

$$M_t^*(x_{t-1}, x_t) = r(c_t, x_{t-1})^{-1} \min \left[\left\{ 1, W_{t-1}(x_{t-1}) \frac{w_t(x_{t-1}, x_t)}{c_t} \right\} M_t(x_{t-1}, x_t) \right]. \quad (4.3.2)$$

The quantity $r(c_t, x_{t-1})$ denotes the normalizing constant for particle x_{t-1} , given by

$$r(c_t, x_{t-1}) = \int \min \left\{ 1, W_{t-1}(x_{t-1}) \frac{w_t(x_{t-1}, x_t)}{c_t} \right\} M_t(x_{t-1}, x_t) dx_t. \quad (4.3.3)$$

Note that $0 < r(c_t, x_{t-1}) \leq 1$ if (w.l.o.g.) the mutation kernel M_t is normalized, so that $\int M_t(x_{t-1}, x_t) dx_t = 1$, and if the PRC threshold $0 \leq c_t < \infty$ is finite. Thus the SMC sampler PRC algorithm can be considered as an SMC sampler algorithm with the mutation kernel $M_t^*(x_{t-1}, x_t)$, and the correction weight

$$W_t(x_t) = W_{t-1}(x_{t-1}) \frac{\pi_t(x_t) L_{t-1}(x_t, x_{t-1})}{\pi_{t-1}(x_{t-1}) M_t^*(x_{t-1}, x_t)}. \quad (4.3.4)$$

4.3.3 Estimation of the normalizing constant

As the normalizing constant $r(c_t, x_{t-1})$ in the weight calculation (4.3.4) in general depends on x_{t-1} , it must be evaluated. Where no analytic solution can be found, approximating (4.3.3) may be achieved by, for example, quadrature methods if the sample space E is relatively low dimensional or Monte Carlo methods if E is high dimensional. For example, for $j = 1, \dots, m$ independent samples $x_t^{*(j)}$ sampled from $M_t(x_{t-1}, x_t)$

$$\hat{r}(c_t, x_{t-1}) \approx \frac{1}{m} \sum_{j=1}^m \min \left\{ 1, W_{t-1}(x_{t-1}) \frac{w(x_{t-1}, x_t^{*(j)})}{c_t} \right\}.$$

An alternative, computationally more efficient approach is to select kernels M_t and L_{t-1} such that $r(c_t, x_{t-1}) = r(c_t)$ will be constant for all particles x_{t-1} . In this case, the value of $r(c_t)$ may be absorbed into the proportionality constant of the weights, and safely ignored. Equation (4.3.3) suggests that this can be achieved if $M_t(x_{t-1}, x_t)$, $W_{t-1}(x_{t-1})$ and $w(x_{t-1}, x_t)$ are independent of x_{t-1} .

Specifying mutation kernels M_t such that $M_t(x_{t-1}, x_t) = M_t(x_t)$ amounts to choosing a *global* kernel which is the same for all particles x_{t-1} . This is common in practice (e.g. 39). The particle dependent weight $W_{t-1}(x_{t-1})$ can be set to $1/N$ for all particles following a resampling (or preselection) step. Finally, consider for a moment the backward kernel of the form

$$L_{t-1}^{opt}(x_t, x_{t-1}) = \frac{\pi_{t-1}(x_{t-1}) M_t(x_{t-1}, x_t)}{\int \pi_{t-1}(x_{t-1}) M_t(x_{t-1}, x_t) dx_{t-1}}. \quad (4.3.5)$$

This backward kernel is an approximation of the optimal backward kernel, in the sense of the choice of L_{t-1} that minimizes the variance of the importance sampling weights (12). Under the backward kernel (4.3.5), the incremental weight can be approximated by

$$\begin{aligned} w_t(x_{t-1}, x_t) &= \pi_t(x_t) / \int \pi_{t-1}(x_{t-1}) M_t(x_{t-1}, x_t) dx_{t-1} \\ &\approx \pi_t(x_t) / \sum_{i=1}^N W_{t-1}(x_{t-1}^{(i)}) M_t(x_{t-1}^{(i)}, x_t). \end{aligned}$$

Under a global mutation kernel $M_t(x_t)$, and following a resampling step, then the incremental weight under this backward kernel reduces to $w_t(x_{t-1}, x_t) = \pi_t(x_t) / M_t(x_t)$, which is independent of x_{t-1} . Thus, the weight calculation in (4.3.4) becomes

$$\begin{aligned} W_t(x_t) &\propto \pi_t(x_t) / \left[\min \left\{ 1, \frac{w(x_{t-1}, x_t)}{Nc_t} \right\} M_t(x_t) \right] \\ &= \begin{cases} \pi_t(x_t) / M_t(x_t) & \text{if } \min \left\{ 1, \frac{w(x_{t-1}, x_t)}{Nc_t} \right\} = 1 \\ Nc_t & \text{otherwise.} \end{cases} \end{aligned}$$

Note that under this setting, the SMC sampler PRC algorithm can be considered as a sequence of importance sampling strategies with partial rejection control.

4.4 SMC Sampler PRC algorithm analysis

In this section we study theoretical properties of the SMC sampler PRC algorithm. We firstly bound the variance of the importance weights, and then present a central limit theorem for the sampler with a recursive expression for the asymptotic variance. Finally, via a connection with an existing algorithm, we establish a condition for which the number of rejection steps under the PRC mechanism is almost surely finite.

4.4.1 Variance of the incremental weights

We begin this section by establishing a bound on the variance of the importance weights of the SMC sampler PRC algorithm.

Theorem 4.4.1. *Let $W_t(x_t)$ denote the importance sampling weight at time t from a standard SMC sampler with mutation kernel M_t , and let $W_t^*(x_t)$ denote the equivalent weight following a partial rejection control step under the SMC sampler PRC algorithm, with resulting mutation kernel M_t^* . Then*

$$\text{Var}_{M_t^*} [W_t^*(x_t)] \leq \text{Var}_{M_t} [W_t(x_t)].$$

Proof: See Appendix A.1. Hence, applying partial rejection control within the SMC sampler framework will not worsen, and may improve the variance of the importance weights, by reducing the χ^2 -distance between the sampling and target distributions at each stage, t .

In the case where $\min \left\{ 1, \frac{W_t(x_t)}{c_t} \right\} = 1$ for all x_t , which is achieved when $c_t \leq \inf_{x_t} \{W_t(x_t)\}$, then from (4.3.3) we have $r(c_t, x_{t-1}) = 1$ for all x_{t-1} . From (4.3.2), this results in $M_t^*(x_{t-1}, x_t) = M_t(x_{t-1}, x_t)$ and hence $\text{Var}_{M_t^*} [W_t^*(x_t)] = \text{Var}_{M_t} [W_t(x_t)]$. That is, the SMC sampler PRC algorithm reduces to the standard SMC sampler when $c_t \leq \inf_{x_t} \{W_t(x_t)\}$, and in this case, the variance of the importance weights is maximized. When $W_t(x_t) \in [0, \infty)$ this is realized for $c_t = 0$ where we define $0/0 := 1$.

4.4.2 A central limit theorem

Central Limit Theorems (CLTs) for SMC and particle filtering algorithms have been derived in various literatures (10; 21; 7; 12; 17). They are based on the observation that an SMC algorithm introduces local errors (fluctuations) as a result of the approximations introduced by sampling numerically from the transitions. Hence, at each stage t , one can decompose the error between the target distribution π_t and the N -particle approximation π_t^N . This turns out to be a sum of the local sampling fluctuations at each discrete time in the past, propagated forwards in time to t .

In the setting of the SMC sampler algorithm, the existence of a CLT is established by (10). Explicitly, under the assumption of multinomial resampling at each stage of the algorithm, and the integrability conditions given in (7) [Theorem 1] and (10) [Section 9.4, pp.300-306], then for a suitable continuous and bounded test function $\varphi \in C_b(E)$ we have

$$N^{1/2} \left(\mathbb{E}_{\pi_t^N}(\varphi) - \mathbb{E}_{\pi_t}(\varphi) \right) \rightarrow \mathcal{N}(0, V_{SMC,t}(\varphi)) \quad (4.4.1)$$

as $N \rightarrow \infty$, for each $t = 1, \dots, T$. (12) obtain a recursive expression for the asymptotic variance $V_{SMC,t}(\varphi)$ as an explicit function of the backward kernels L_{t-1} and the sequence of distributions on path space, $\tilde{\pi}_t$.

Following (12), we obtain an analogous result for the SMC sampler PRC algorithm. Under the same assumptions as the above, we have the CLT (4.4.1) with asymptotic variance given by

$$\begin{aligned} V_{SMC-PRC,t}(\varphi) &= \int I_1 \frac{\tilde{\pi}_1^2(x_1)}{\pi_1(x_1)} \left(\int \varphi(x_1) \tilde{\pi}_1(x_t|x_1) dx_t - \mathbb{E}_{\pi_t}(\varphi) \right)^2 dx_1 \\ &+ \sum_{k=2}^{t-1} \int I_k \frac{(\tilde{\pi}_t(x_k)L_{k-1}(x_k, x_{k-1}))^2}{\pi_{k-1}(x_{k-1})M_k(x_{k-1}, x_k)} \left(\int \varphi(x_t) \tilde{\pi}_t(x_t|x_k) dx_t - \mathbb{E}_{\pi_t}(\varphi) \right)^2 dx_{k-1} dx_k \\ &+ \int I_t \frac{(\pi_t(x_t)L_{t-1}(x_t, x_{t-1}))^2}{\pi_{t-1}(x_{t-1})M_t(x_t, x_{t-1})} (\varphi(x_t) - \mathbb{E}_{\pi_t}(\varphi))^2 dx_{t-1} dx_t \end{aligned} \quad (4.4.2)$$

with $I_k = \left[r(c_k, x_{k-1})^{-1} \min \left\{ 1, \frac{1}{N} \frac{\pi_k(x_k)L_{k-1}(x_k, x_{k-1})}{\pi_{k-1}(x_{k-1})M_k(x_{k-1}, x_k)c_k} \right\} \right]^{-1}$ and $I_1 = 1$.

The contribution of the PRC stage to the asymptotic variance is encapsulated in the I_k terms. Under the standard SMC sampler algorithm we have $c_k \leq \inf_{x_t} \{W_t(x_t)\}$ so that $I_k = 1$ for all $k = 1, \dots, t$. In this setting, (4.4.2) reduces to the asymptotic variance expression obtained by (12).

4.4.3 Connections to an existing SMC algorithm

In rare event applications there is a high probability of generating particles with exactly zero weights. The AliveSMC algorithm (23; 22) was developed to ensure that a particle population of a desired size persists at each iteration of a standard SMC algorithm (see 13; 19 for related methods). In this setting, the number of particles at each time t is considered as a random variable N_{c_t} . That is, N_{c_t} is the number of particles required to generate exactly N non-zero weighted particles. In this Section we reinterpret the SMC sampler PRC algorithm in terms of the AliveSMC algorithm. As a consequence, in Section 4.4.4 we are able to establish a condition under which the PRC resampling stage will require a finite number of rejection attempts.

In (22), at iteration t , a fitness function is applied to select particles satisfying a desired criteria. Those particles not satisfying the criteria receive a zero weight. In an SMC sampler PRC setting, the fitness function can be interpreted as selecting those particles with a weight that is immediately accepted under the PRC acceptance probability. As such, we may rewrite the SMC sampler PRC algorithm with a modified mutation and correction step:

Reinterpreted SMC sampler PRC/AliveSMC algorithm

Mutation and correction: Set $t = t + 1$ and $j = 1$. For $i = 1, \dots, N_{c_t}$:

- (a) Sample $x_t^{(i)} \sim M_t(x_{t-1}^{(j)}, x_t)$ and calculate $W = W_{t-1}(x_{t-1}^{(j)}) \frac{\pi_t(x_t^{(i)})L_{t-1}(x_t^{(i)}, x_{t-1}^{(j)})}{\pi_{t-1}(x_{t-1}^{(j)})M_t(x_{t-1}^{(j)}, x_t^{(i)})}$.
- (b) Set weight for $x_t^{(i)}$ as

$$W_t(x_t^{(i)}) \propto \begin{cases} Wr(c_t, x_{t-1}^{(j)})/p^{(i)} & \text{with probability } p^{(i)} = \min\{1, W/c_t\} \\ 0 & \text{otherwise.} \end{cases}$$
- (c) If $W_t(x_t^{(i)}) \neq 0$, increment $j = j + 1$.

If $t < T$ go to Resample.

Note that $j = 1, \dots, N$ indexes the particles $x_{t-1}^{(j)}$ at time $t - 1$ such that particle mutations from $x_{t-1}^{(j)}$ generate a non-zero weight exactly once. Also, under the fitness function, particle $x_t^{(i)}$ has a probability $1 - p^{(i)} = 1 - \min\{1, W/c_t\}$ of being exactly zero. Given that it is possible to express the SMC sampler PRC algorithm within the AliveSMC framework, we may adapt the results of (23) and (22), to obtain a condition under which the SMC sampler PRC algorithm is guaranteed to require a finite number of attempts, $N_{c_t} < \infty$, to obtain exactly N non-zero weighted particles.

4.4.4 Analysis of the number of rejection attempts

Following (23) and (22), we define the random variable N_{c_t} as

$$N_{c_t} \triangleq \inf \left\{ N^* \geq 1 : \sum_{i=1}^{N^*} W_t^{(i)}(x_t) \geq N \sup_{x_t \in E} W_t(x_t) \right\}$$

(22) proved for the AliveSMC algorithm that the random number of particles N_{c_t} is almost surely finite with $N_{c_t} \geq N$, under the condition that

$\langle \pi_{t-1} M_t, W_t \rangle = \frac{\int W_t(x_t) \int \pi_{t-1}(x_{t-1}) M_t(x_{t-1}, x_t) dx_{t-1} dx_t}{\int \pi_{t-1}(x_{t-1}) M_t(x_{t-1}, x_t) dx_{t-1}} > 0$. A sufficient condition for this to hold is $\mathbb{E}_{M_t} [W_t(x_t) | x_{t-1} = x] > 0$, for all $x \in E$. Thus for the SMC sampler PRC algorithm, we have the following theorem:

Theorem 4.4.2. *Under the SMC sampler PRC algorithm, the number of rejection attempts at each stage of the algorithm, $N_{c_t} \geq N$, is almost surely finite if $c_t < \infty$.*

Proof: See Appendix A.2.

Corollary 5.1 *The following convergence in probability holds, with a rate of $1/\sqrt{N}$:*

$$\frac{N_{c_t}}{N} \rightarrow \frac{\sup_{x_t \in E} W_t(x_t)}{\langle \pi_{t-1} M_t, W_t \rangle} < \infty.$$

See (22) for further details. Hence, the SMC sampler PRC algorithm possesses an almost surely finite number of rejection attempts if the PRC threshold c_t is finite, with the above rate of convergence.

4.5 Approximate Bayesian computation

With the aim of posterior simulation from $\pi(x|\mathcal{D}) \propto \pi(\mathcal{D}|x)\pi(x)$ for parameters x and observed data \mathcal{D} , "likelihood-free," approximate Bayesian computation (ABC) methods are often utilized when the likelihood function, $\pi(\mathcal{D}|x)$, is computationally intractable or when its evaluation is computationally prohibitive. ABC methods can be based on rejection sampling (45; 2), Markov chain Monte Carlo (28; 6; 39) and SMC-type samplers (43; 37; 3; 11). While currently among the most efficient ABC methods, the underlying practical issue with SMC-type algorithms is in avoiding sample degeneracy through extreme numbers of particles with low or exactly zero weights. In this Section, we will demonstrate that the SMC sampler PRC algorithm applied within the ABC framework can achieve significant performance gains and greater modelling flexibility over existing SMC-type ABC methods.

The underlying approach of ABC methods is to augment the (intractable) posterior to $\pi(x, \mathcal{D}'|\mathcal{D}) \propto \pi(\mathcal{D}|\mathcal{D}', x)\pi(\mathcal{D}'|x)\pi(x)$, where the auxiliary parameter is an artificial data set distributed according to the model $\mathcal{D}' \sim \pi(\cdot|x)$. An approximation of the target posterior $\pi(x|\mathcal{D})$ is then given by

$$\pi_{ABC}(x|\mathcal{D}) \propto \int \pi(\mathcal{D}|\mathcal{D}', x)\pi(\mathcal{D}'|x)\pi(x)d\mathcal{D}'. \quad (4.5.1)$$

The weighting function $\pi(\mathcal{D}|\mathcal{D}', x)$ takes high weight in regions where the datasets \mathcal{D} and \mathcal{D}' are similar, and low weight otherwise. Comparison of the datasets is usually achieved through low-dimensional summary statistics $T(\cdot)$, so that, for example

$$\pi(\mathcal{D}|\mathcal{D}', x) \propto \begin{cases} 1 & \text{if } \rho(T(\mathcal{D}), T(\mathcal{D}')) \leq \epsilon \\ 0 & \text{else,} \end{cases} \quad (4.5.2)$$

for some small tolerance value $\epsilon > 0$ and distance measure ρ . If $T(\cdot)$ are sufficient statistics, and $\epsilon \rightarrow 0$ so that $\pi(\mathcal{D}|\mathcal{D}', x)$ reduces to a point mass at $T(\mathcal{D}) = T(\mathcal{D}')$ then $\pi_{ABC}(x|\mathcal{D}) = \pi(x|\mathcal{D})$ is recovered exactly, otherwise the ABC approximation to $\pi(x|\mathcal{D})$ is of the form (4.5.1), with greater accuracy for smaller ϵ . The computational overhead of all ABC samplers increases as ϵ decreases, producing a trade off between computation and accuracy. ABC methods either sample from the joint density $\pi(x, \mathcal{D}'|\mathcal{D})$ by arranging to cancel out the intractable likelihood in a weight or acceptance probability, or sample from $\pi_{ABC}(x|\mathcal{D})$ directly via Monte Carlo integration

$$\pi_{ABC}(x|\mathcal{D}) \approx \frac{\pi(x)}{S} \sum_{s=1}^S \pi(\mathcal{D}|\mathcal{D}'_s, x), \quad (4.5.3)$$

where $\mathcal{D}'_1, \dots, \mathcal{D}'_S \sim \pi(\mathcal{D}'|x)$ are draws from the likelihood given x . Almost all current ABC methods have the weighting density (4.5.2) written directly into the algorithm.

We apply the SMC sampler PRC algorithm in the ABC framework as follows: The target $\pi_t(x_t) = \pi_{ABC,t}(x_t|\mathcal{D})$ is given by (4.5.1), with the weighting function $\pi_t(\mathcal{D}|\mathcal{D}', x)$ parameterized by a different scaling parameter ϵ_t for each t , where $\infty = \epsilon_1 \geq \dots \geq \epsilon_T$, produces increasing accuracy at each step, t . The ϵ_t sequence and its length, T , may be determined *a priori* or dynamically. Evaluation of $\pi_t(x_t)$ is defined by (4.5.3) through S Monte Carlo draws. Given the high computational overheads of ABC methods, we avoid evaluating the PRC normalizing constant (as demonstrated in Section 4.3.3), through a global mutation kernel $M_t(x_t)$, the backward kernel L_{t-1}^{opt} (c.f. 4.3.5) and enforced resampling.

SMC sampler PRC-ABC algorithm

Initialization: Set $t = 1$.

For $i = 1, \dots, N$, sample $x_1^{(i)} \sim \mu(x)$, and set weights $W_t(x_1^{(i)}) = \pi_{ABC,1}(x_1^{(i)}|\mathcal{D})/\mu(x_1^{(i)})$.

Resample: Resample N particles with respect to $W_t(x_t^{(i)})$ and set $W_t(x_t^{(i)}) = \frac{1}{N}$,
 $i = 1, \dots, N$.

Mutation and correction: Set $t = t + 1$ and $i = 1$:

- (a) Sample $x_t^{(i)} \sim M_t(x_t)$ and set weight for $x_t^{(i)}$ to

$$W_t(x_t^{(i)}) = \pi_{ABC,t}(x_t^{(i)}|\mathcal{D})/M_t(x_t^{(i)}).$$
 - (b) With probability $1 - p^{(i)} = 1 - \min\{1, W_t(x_t^{(i)})/c_t\}$, reject $x_t^{(i)}$ and go to (a).
 - (c) Otherwise, accept $x_t^{(i)}$ and set $W_t(x_t^{(i)}) = W_t(x_t^{(i)})/p^{(i)}$.
 - (d) Increment $i = i + 1$. If $i \leq N$, go to (a).
 - (e) If $t < T$ then go to Resample.
-

The density $\mu(x)$ is an initial sampling distribution, from which direct sampling is available. As with the tolerance ϵ_t , the PRC thresholds c_t may also be determined dynamically (see below for an illustration). Note that as the resampled particles in the Resample step play no subsequent part in the sampler, in practice this step can be omitted. The path of each particle $(x_1^{(i)}, \dots, x_T^{(i)}) \in E^T$ can be reconstructed post-simulation, if required, by resampling the recorded marginal populations $(W_t(x_t^{(i)}), x_t^{(i)})$.

The above algorithm has a number of benefits over existing SMC-type ABC samplers (43; 37; 3; 11). Firstly, the weighting density $\pi(\mathcal{D}|\mathcal{D}', x)$ can take any form – we suggest any smoothing kernel, following (4). Existing samplers in the literature are restricted to the uniform function (4.5.2). Secondly, there is complete control over the PRC threshold, c_t , unlike (43) who impose a specific value. Thirdly, in estimating $\pi_{ABC}(x_t|\mathcal{D})$, as long as the L_{t-1}^{opt} backward kernel is used, any number $S \geq 1$ of Monte Carlo draws can be used in (4.5.3). Existing samplers only use $S = 1$, and so there is less control over the variability of the weights. Finally, providing that the computation required to estimate the PRC normalizing constants, $r(c_t, x_{t-1})$, is acceptable, a form of the SMC sampler PRC-ABC sampler may be constructed which uses arbitrary mutation and backward kernels, allowing the user to select the most appropriate tools for a given problem.

4.5.1 Simulation study

We now demonstrate the superior performance of the SMC sampler PRC-ABC algorithm through a controlled study. Specifically, we specify the true posterior $\pi(x|\mathcal{D})$ as $N(0, 1)$ by defining the likelihood and prior as $\mathcal{D} \sim N(x, 1)$ and $\pi(x) \propto 1$, with a single observed datum, $\mathcal{D} = 0$. For this model, a sufficient statistic is $T(\mathcal{D}) = \mathcal{D}$. From (4.5.1), for the uniform weighting density (4.5.2) with $\rho(a, b) = |a - b|$ and $\epsilon = \epsilon_u$, or for $\pi(\mathcal{D}' | \mathcal{D}, x) = N(\mathcal{D}, \epsilon_g^2)$, then $\pi_{ABC}(x|\mathcal{D})$ may be obtained in closed form as

$$\pi_{ABC}(x|\mathcal{D}) \propto \frac{\Phi(\epsilon_u - x) - \Phi(-\epsilon_u - x)}{2\epsilon_u} \quad \text{or} \quad \pi_{ABC}(x|\mathcal{D}) = N(0, 1 + \epsilon_g^2)$$

respectively, where $\Phi(\cdot)$ denotes the standard Gaussian CDF. In both cases $\pi_{ABC}(x|\mathcal{D}) \rightarrow N(0, 1)$ as $\epsilon \rightarrow 0$. In order to directly compare the two approximate posteriors we impose equal variances on the two weighting functions, so that $\epsilon_g = \sqrt{3}\epsilon_u$

We adopt the following sampler specifications: A particle population of size $N = 1000$ was

drawn from the initial sampling distribution $\mu(x) \sim U(-5, 5)$, and the sequence of distributions, π_1, \dots, π_{10} , is defined by $\{\epsilon_t\} = \{\infty, 10, 5, 2, 1, 0.5, 0.2, 0.1, 0.05, 0.05\}$, on the ϵ_g scale. The mutation kernel $M_t(x_t) = \sum_{j=1}^N W_{t-1}^{(j)} \psi(x_t | x_{t-1}^{(j)}, \tau^2)$ is taken as a Normal kernel density estimate of $\pi_{t-1}(x_{t-1})$, with $\tau^2 = 1$ and where $\psi(x | \mu, \sigma^2)$ denotes the PDF of a $N(\mu, \sigma^2)$ distribution evaluated at x . We initially use $S = 1$ Monte Carlo draws to approximate $\pi_{ABC}(x | \mathcal{D})$ (c.f. 4.5.3).

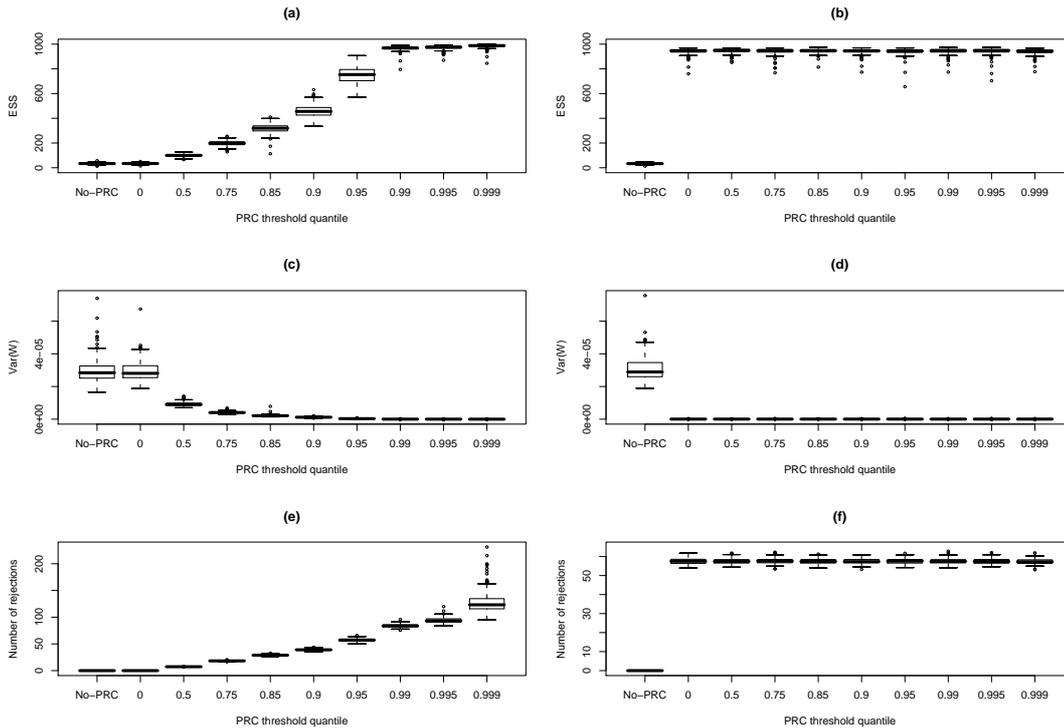


Fig. 4.5.1: Effective sample size (a,b), variance of normalized importance weights (c,d) and mean number of rejections per particle (e,f) as functions of PRC threshold c_t . PRC threshold is defined dynamically as a quantile of the non-zero weights at time t (x -axis). Left plots (a,c,e) and right plots (b,d,f) are obtained under the Gaussian and uniform weighting densities $\pi(\mathcal{D} | \mathcal{D}', x)$ respectively. Boxplots are based on 250 sampler replications.

Figure 4.5.1 examines the effect of PRC on the effective sample size (ESS), the variance of the importance weights and the mean number of rejections per particle. The PRC threshold was determined dynamically at each iteration as $c_t = Q(W_t^+(x_t), q)$, the q -th quantile of the *non-zero* weights at time t (obtained by mutating all x_{t-1} particles under M_t before implementing the PRC stage), for $q = 0, 0.5, 0.75, 0.85, 0.9, 0.95, 0.99, 0.995, 0.999$. Results are shown using the Gaussian (left plots) and uniform (right plots) weighting density, based on 250 sampler replications. Note that the PRC threshold with $q = 0$ approximately corresponds to a standard SMC sampler ("No-PRC") only for the Gaussian weighting function, as the uniform weighting function permits exactly zero importance weights. Setting $q = 0$ for the uniform weighting density corresponds to existing SMC-type ABC samplers (43; 37; 3).

For both weighting densities, the effective sample size increases as c_t increases, and the variance

of the importance weights decreases. Naturally, the higher the PRC threshold, the more rejections occur, quantifying the extra computation required for the gains in sampler performance. However, there is a notable difference in the transition from poor (no PRC) to improved (under PRC) performance between the two different weighting densities. This occurs as the uniform weighting density only permits 0/1 weights, compared to the smoother scale under the Gaussian. As a result, the uniform weighting density (which is the only choice under existing ABC samplers) has a fixed, albeit strong, performance gain over not implementing PRC, but at a very high computational cost (Figure 4.5.1,f). Comparison with panels (a,c,e) suggests that considerable computational gains can be achieved with alternative weighting functions, without sacrificing sampler performance. This is easily permitted under the SMC sampler PRC framework.

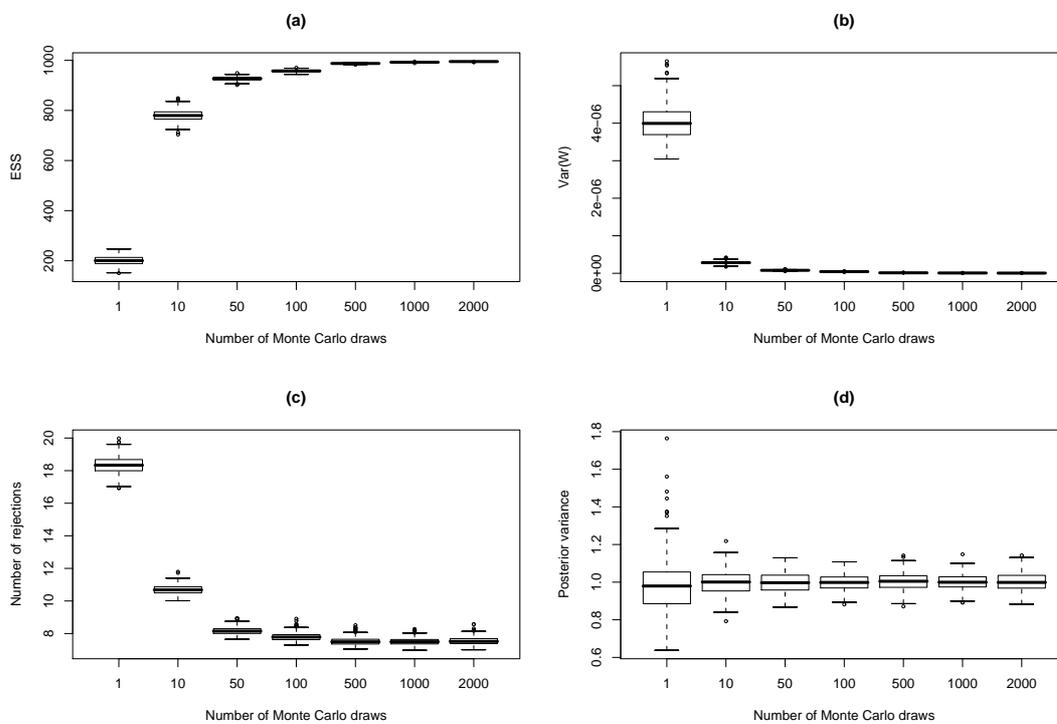


Fig. 4.5.2: The effect of the number of Monte Carlo draws, S , on sampler performance. Panels show (a) effective sample size (ESS), (b) variance of the importance weights, (c) the mean number of rejection attempts, and (d) estimates of the posterior variance (true value ≈ 1), as a function of the number of Monte Carlo draws in the estimation of $\pi_{ABC}(x|\mathcal{D})$.

When using the L_{t-1}^{opt} backward kernel (4.3.5), any number $S \geq 1$ of Monte Carlo draws may be used to approximate $\pi_{ABC}(x|\mathcal{D})$ via (4.5.3). While $S = 1$ is near universal under existing ABC algorithms, one would expect to realize less variable importance weights for $S > 1$. Figure (4.5.2) illustrates the effect of increasing S , under the Gaussian density function, based on the PRC threshold $c_t = Q(W_t^+(x_t), 0.95)$. An increase in the effective sample size (panel a) is reflected by the reduction in the variability of the importance weights (panel b), as is the variability in the estimates of the posterior variance (panel d). This in turn results in lower numbers

of rejections at the PRC stage (panel c). Of course, these performance gains are again balanced by the strong increases in computation required for $S > 1$. It would appear that unless the data-generation procedure $\mathcal{D}'_s \sim \pi(\mathcal{D}|x)$ is computationally inexpensive, $1 \leq S \leq 10$ would seem to be the most useful choice in practice. Regardless, the greatest gains in sampler performance under the SMC sampler PRC algorithm are achieved for $S = 1$.

4.6 A stochastic claims reserving analysis

We present an analysis of an important and popular class of statistical models in actuarial science using stochastic claims reserving. We consider a time series formulation of the distribution-free chain ladder model (27; 16; 30). For a claim on an insurance company for an accident in year i , $C_{i,j}$ denotes the cumulative claim in subsequent years $j \geq i$. Cumulative claims can refer to payments, claims incurred and other expenses. At time I , we have observations $\mathcal{D}_I = \{C_{i,j}; i + j \leq I\}$, and for reserving against future claims we wish to predict $\mathcal{D}_I^c = \{C_{i,j}; i + j > I, i \leq I\}$. One such dataset is illustrated in Table 4.1.

Under a time series formulation, cumulative claims $C_{i,j}$ in different accident years i are independent and satisfy, for $j = 0, \dots, I - 1$,

$$C_{i,j+1} = f_j C_{i,j} + \sigma_j \sqrt{C_{i,j}} \varepsilon_{i,j+1}, \quad (4.6.1)$$

where $\mathbf{f} = (f_0, \dots, f_{I-1})$ and $\boldsymbol{\sigma} = (\sigma_0, \dots, \sigma_{I-1})$ are respectively the chain ladder factors and standard deviations, and the residuals $\varepsilon_{i,j}$ are i.i.d. with mean 0 and variance 1. The model is constrained such that $P(C_{i,j} > 0 | \{C_{k,0}\}_{k=1}^j, \mathbf{f}, \boldsymbol{\sigma}) = 1$ for all i, j (see 30). If distributional assumptions are made on the residuals $\varepsilon_{i,j}$ (e.g. 40), the posterior distribution can be made computationally tractable. However, a primary intention of this model is to work with distribution-free assumptions on the residuals, and therefore on the cumulative claims. Within this distribution-free context one wishes to quantify popular risk metrics such as value-at-risk and expected-shortfall to be calculated for the predicted claims distribution, both of which are highly relevant to regulatory reporting. Alternative approaches, based on credibility results, can relax such distributional assumptions, but can only provide statements on posterior first and second moments in limiting cases (16).

Previously, actuaries have proceeded by predicting claims via a deterministic model known as the classical chain ladder algorithm. This approach predicts unobserved future cumulative claims by the recursion $\widehat{C}_{i,I-i} = C_{i,I-i}$, and for $j > I - i$

$$\widehat{C}_{i,j} = \widehat{C}_{i,j-1} \widehat{f}_{j-1}^{(CL)} \quad \text{where} \quad \widehat{f}_{j-1}^{(CL)} = \frac{\sum_{i=0}^{I-j} C_{i,j}}{\sum_{i=0}^{I-j} C_{i,j-1}}, \quad (4.6.2)$$

and where, in the time series formulation, the variances are estimated by

$$\widehat{\sigma}_j^{2(CL)} = \frac{1}{I-j-1} \sum_{i=0}^{I-j-1} C_{i,j} \left(\frac{C_{i,j+1}}{C_{i,j}} - \widehat{f}_j^{(CL)} \right)^2.$$

See (27) for an estimator of $\hat{\sigma}_{I-1}^{2(CL)}$. As this algorithm is deterministic there is strong interest in stochastic chain ladder models, which naturally allow the quantification of uncertainty, such as the mean square error of prediction. In the claims reserving setting the most popular stochastic models are those with estimators which recover the classical chain ladder estimators. We consider one such Bayesian stochastic model which has the property that as the diffusivity of the priors $\pi(\mathbf{f}, \boldsymbol{\sigma})$ tends to infinity $\hat{\mathbf{f}}^{(MMSE)} \rightarrow \hat{\mathbf{f}}^{(CL)}$ where $MMSE$ denotes the posterior mean (16). Hence by (4.6.2) the posterior mean $E[C_{i,J}|\mathcal{D}_I] = \hat{C}_{i,J}$ recovers the classical estimators, thereby justifying the classical model. We sample from the intractable posterior $\pi_{ABC}(\mathbf{f}, \boldsymbol{\sigma}|\mathcal{D}_I)$ using the SMC sampler PRC-ABC algorithm.

4.6.1 Analysis and results

This model is interesting as the intractability of the likelihood directly impacts the ability to generate synthetic data sets, \mathcal{D}'_I , from the model. That is, if the distributional form of the residuals were known, data-generation from the model would be trivial. To retain a distribution-free setting we alternatively utilize a conditional bootstrap approach (30). Conditional upon proposed parameters \mathbf{f} and $\boldsymbol{\sigma}$, the residuals $\tilde{\epsilon}_{i,j}|\mathbf{f}, \boldsymbol{\sigma}$ are iteratively obtained by inversion of (4.6.1). Then, by independently drawing resampled residuals from the empirical conditional residual distribution, a bootstrap sample of the cumulative claims \mathcal{D}'_I is then available through recursion on (4.6.1).

In analyzing the real claims reserving data in Table 4.1 we specify independent priors $f_j \sim \text{Gamma}(\alpha_j, \beta_j)$ with mean $\alpha_j\beta_j = \hat{f}_j^{(CL)}$ and $\sigma_j \sim \text{IGamma}(a_j, b_j)$ with mean $b_j/(a_j - 1) = \hat{\sigma}_j^{(CL)}$ for $j = 0, \dots, I - 1$, each with large variance. For summary statistics we adopt $T(\mathcal{D}') = (\mathcal{D}'_I, \mu'(\tilde{\epsilon}), s'(\tilde{\epsilon}))$ where $\mu'(\tilde{\epsilon})$ and $s'(\tilde{\epsilon})$ denote the sample mean and standard deviation of the conditionally resampled residuals $\tilde{\epsilon}'_{i,j}|\mathbf{f}, \boldsymbol{\sigma}$. The observed summary statistics are given by $T(\mathcal{D}) = (\mathcal{D}_I, 0, 1)$ following the zero mean and unit variance assumptions on the true residuals.

We implement the SMC sampler PRC-ABC algorithm with uniform weighting density (4.5.2) and $\rho(T(\mathcal{D}_I), T(\mathcal{D}'_I)) = [(T(\mathcal{D}_I) - T(\mathcal{D}'_I))^\top \Sigma^{-1}(T(\mathcal{D}_I) - T(\mathcal{D}'_I))]^{1/2}$ defined as Mahalanobis distance, where the covariance Σ is estimated following (30). We use $N = 5000$ particles, PRC threshold $c_t = Q(W_t^+(x_t), 0)$ and a deterministic distribution schedule $\{\epsilon_t\} = \{\infty, 10, \dots, 0.00001\}$ with $T = 22$. The mutation kernel $M_t(x_t) = \sum_{i=1}^N W_{t-1}^{(i)}(x_{t-1}^{(i)})\text{Gamma}(a(x_{t-1}^{(i)}), b(x_{t-1}^{(i)}))$ is a mixture of gamma densities, with mean $a(x_{t-1}^{(i)})b(x_{t-1}^{(i)}) = x_{t-1}^{(i)}$ and large variance.

Table 4.2 presents a comparison of the parameter estimates $\hat{\mathbf{f}}$ and $\hat{\boldsymbol{\sigma}}$, and predicted cumulative claims, $\hat{C}_{i,j}$ under classical and Bayesian models. Given the uninformative priors, the posterior mean estimates and resulting predicted claims agree well with those obtained under the classical model. This provides some validation for the deterministic classical model estimates under the Bayesian stochastic interpretation.

Perhaps more usefully for inference, the full posterior $\pi_{ABC}(\mathbf{f}, \boldsymbol{\sigma}|\mathcal{D}_I)$ is available. Figure 4.7.1 illustrates how the estimated marginal densities of the first chain ladder factor, $\pi_{ABC,t}(f_0|\mathcal{D}_I)$, and the associated standard deviation, $\pi_{ABC,t}(\sigma_0|\mathcal{D}_I)$, evolve as ϵ_t decreases. The precision of

the densities clearly improves, as decreasing ϵ_t imposes stricter restrictions on the permissible deviations of the ABC approximate posterior $\pi_{ABC}(\mathbf{f}, \boldsymbol{\sigma}|\mathcal{D})$ from the target posterior $\pi(\mathbf{f}, \boldsymbol{\sigma}|\mathcal{D})$. A full predictive analysis may now follow, including upper and lower credible bounds on predicted future claims.

4.7 Discussion

When used in challenging settings, sequential Monte Carlo samplers often suffer from severe particle degeneracy. In this article we have provided a practical approach to tackling this problem, by incorporating the partial rejection control mechanism of (26) directly into the mutation kernel of the SMC sampler. The resulting sampler will not worsen, and can improve the variance of the importance weights, sometimes substantially so. By establishing clear relationships with existing samplers (12; 23), many theoretical properties may be extended to the SMC sampler PRC algorithm, including a central limit theorem, and a proof of an almost sure finite number of PRC rejection attempts.

There is much opportunity for the specification of the sequence of PRC thresholds to be further automated, if desired. For example, by dynamically determining $c_t > c_{t-1}$ if the effective sample size of the particle population at time $t - 1$ falls too low, and conversely allowing $c_t < c_{t-1}$ if the level of PRC resampling is too high, in order to reduce computational overheads.

As the SMC sampler PRC algorithm allows practical inference in challenging situations in which particle weights are highly variable, we anticipate that a primary application of the sampler will be within the rapidly developing “likelihood-free” approximate Bayesian computation framework. The presented sampler is more flexible and efficient than existing SMC-type ABC samplers, allowing a previously unavailable degree of control over the computation utilized for a given analysis. Perhaps more importantly, the extra flexibility achieved by allowing arbitrary weighting densities (unlike existing ABC samplers) enables the analysis of improved models within the ABC framework, in line with recent non-parametric interpretations (4).

Appendix

*A.1: Proof of Theorem 4.4.1

The proof follows the arguments presented by (25). In particular we study the SMC sampler PRC algorithm in terms of χ^2 distance between the sampling distribution and the target distribution at stage t . Let

$$W_t(x_t) \propto W_{t-1}(x_{t-1}) \frac{\pi_t(x_t) L_{t-1}(x_t, x_{t-1})}{\pi_{t-1}(x_{t-1}) M_t(x_{t-1}, x_t)} \quad \text{and} \quad W_t^*(x_t) \propto W_{t-1}(x_{t-1}) \frac{\pi_t(x_t) L_{t-1}(x_t, x_{t-1})}{\pi_{t-1}(x_{t-1}) M_t^*(x_{t-1}, x_t)}.$$

Recall that the normalizing constant for the mutation kernel M_t^* at time t is

$$r(c_t, x_{t-1}) = \int \min \left\{ 1, \frac{W_t(x_t)}{c_t} \right\} M_t(x_{t-1}, x_t) dx_t = \frac{1}{c_t} \mathbb{E}_{M_t} [\min \{c_t, W_t(x_t)\}].$$

The variance of the importance weight at time t from a standard SMC sampler, with respect to M_t , is given by

$$\mathbb{V}ar_{M_t}[W_t(x_t)] = \int [W_t(x_t)]^2 M_t(x_{t-1}, x_t) dx_t - \mu^2,$$

and similarly, the variance of the equivalent importance weight at time t following a PRC step under the SMC sampler PRC algorithm, with respect to M_t^* , is given by

$$\mathbb{V}ar_{M_t^*}[W_t^*(x_t)] = \int [W_t^*(x_t)]^2 M_t^*(x_{t-1}, x_t) dx_t - \mu^2,$$

where

$$\mu = \mathbb{E}_{M_t^*}[W_t^*(x_t)] = \mathbb{E}_{M_t}[W_t(x_t)].$$

We also have that

$$\begin{aligned} \int [W_t^*(x_t)]^2 M_t^*(x_{t-1}, x_t) dx_t &= \int \left[W_{t-1}(x_{t-1}) \frac{\pi(x_t) L_{t-1}(x_t, x_{t-1})}{\pi_{t-1}(x_{t-1}) M_t^*(x_{t-1}, x_t)} \right]^2 M_t^*(x_{t-1}, x_t) dx_t \\ &= r(c_t, x_{t-1}) \int \frac{W_{t-1}^2(x_{t-1})}{\min\{1, \frac{W_t(x_t)}{c_t}\}} \frac{\pi_t^2(x_t) L_{t-1}^2(x_t, x_{t-1})}{\pi_{t-1}^2(x_{t-1}) M_t^2(x_{t-1}, x_t)} M_t(x_{t-1}, x_t) dx_t \\ &= r(c_t, x_{t-1}) \int \max\{W_t^2(x_t), c_t W_t(x_t)\} M_t(x_{t-1}, x_t) dx_t \\ &= r(c_t, x_{t-1}) \mathbb{E}_{M_t} [\max\{W_t(x_t), c_t\} W_t(x_t)] \\ &= \frac{1}{c_t} \mathbb{E}_{M_t} [\min\{c_t, W_t(x_t)\}] \mathbb{E}_{M_t} [\max\{W_t(x_t), c_t\} W_t(x_t)] \\ &\leq \frac{1}{c_t} \mathbb{E}_{M_t} [\min\{c_t, W_t(x_t)\} \max\{W_t(x_t), c_t\} W_t(x_t)] \\ &= \frac{1}{c_t} \mathbb{E}_{M_t} [c_t W_t^2(x_t)] = \mathbb{E}_{M_t} [W_t^2(x_t)]. \end{aligned}$$

The above inequality holds since the random variables $\min\{c_t, W_t(x_t)\}$ and $\max\{W_t(x_t), c_t\} W_t(x_t)$ are positively correlated (see 26), and so

$$\begin{aligned} &\mathbb{E}_{M_t} [\min\{c_t, W_t(x_t)\} \max\{W_t(x_t), c_t\} W_t(x_t)] \\ &\quad - \mathbb{E}_{M_t} [\min\{c_t, W_t(x_t)\}] \mathbb{E}_{M_t} [\max\{W_t(x_t), c_t\} W_t(x_t)] \geq 0. \end{aligned}$$

Hence

$$\mathbb{V}ar_{M_t^*}[W_t^*(x_t)] \leq \mathbb{E}_{M_t} [W_t^2(x_t)] - \mu^2 = \mathbb{V}ar_{M_t}[W_t(x_t)].$$

□

*A.2: Proof of Theorem 4.4.2

To satisfy the condition $\mathbb{E}_{M_t} [W_t(x_t) | x_{t-1} = x] > 0, \forall x \in E$, for the SMC sampler PRC algorithm, we require

$$\int_E W_{t-1}(x)w_t(x, x_t)r(c_t, x) \left[\min \left\{ 1, \frac{W_{t-1}(x)w_t(x, x_t)}{c_t} \right\} \right]^{-1} M_t(x, x_t) dx_t > 0. \quad (4.7.1)$$

For a particle $x_{t-1} = x$, the proposed state x_t can take values in a support which can be split into two regions, $A(x)$ and $A^c(x)$, such that $A(x) \cup A^c(x) = E$. These respectively correspond to when $\min \left\{ 1, \frac{W_{t-1}(x)w_t(x, x_t)}{c_t} \right\} = 1$ (i.e. particle acceptance probability under PRC is 1) and when $\min \left\{ 1, \frac{W_{t-1}(x)w_t(x, x_t)}{c_t} \right\} = W_{t-1}(x)w_t(x, x_t) / c_t$ (i.e. the particle may be rejected under PRC). Note that in the extreme cases, $c_t \leq W_{t-1}(x)w_t(x, x_t) \Rightarrow A^c(x) = \emptyset$ reduces to the SMC sampler algorithm, and $c_t > \sup_{x_t} W_{t-1}(x)w_t(x, x_t) \Rightarrow A(x) = \emptyset$. More generally, (4.7.1) may be expanded as

$$\frac{r(c_t, x)W_{t-1}(x)}{\pi_{t-1}(x)} \int_{A(x)} \pi_t(x_t) L_{t-1}(x_t, x) dx_t + c_t r(c_t, x) \int_{A^c(x)} M_t(x, x_t) dx_t > 0$$

which is always greater than zero for finite $c_t < \infty$ as $0 < r(c_t, x) \leq 1$.

□

| Accident Year, i | Development Year, j | | | | | | | | | |
|--------------------|-----------------------|----------|---------|---------|---------|---------|--------|--------|--------|--------|
| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| 0 | 594.6975 | 372.1236 | 89.5717 | 20.7760 | 20.6704 | 6.2124 | 6.5813 | 1.4850 | 1.1130 | 1.5813 |
| 1 | 634.6756 | 324.6406 | 72.3222 | 15.1797 | 6.7824 | 3.6603 | 5.2752 | 1.1186 | 1.1646 | |
| 2 | 626.9090 | 297.6223 | 84.7053 | 26.2768 | 15.2703 | 6.5444 | 5.3545 | 0.8924 | | |
| 3 | 586.3015 | 268.3224 | 72.2532 | 19.0653 | 13.2976 | 8.8340 | 4.3329 | | | |
| 4 | 577.8885 | 274.5229 | 65.3894 | 27.3395 | 23.0288 | 10.5224 | | | | |
| 5 | 618.4793 | 282.8338 | 57.2765 | 24.4899 | 10.4957 | | | | | |
| 6 | 560.0184 | 289.3207 | 56.3114 | 22.5517 | | | | | | |
| 7 | 528.8066 | 244.0103 | 52.8043 | | | | | | | |
| 8 | 529.0793 | 235.7936 | | | | | | | | |
| 9 | 567.5568 | | | | | | | | | |

Tab. 4.1: A claims development triangle. Upper triangle denotes observed annual claims $Y_{i,j}$ from which $C_{i,j} = \sum_{k=0}^j Y_{i,k} \in \mathcal{D}_I$ may be obtained; lower triangle denotes annual $Y_{i,j}$ and cumulative claims $C_{i,j} \in \mathcal{D}_I^c$ to be predicted. Data are real insurance figures in units of \$10,000 (c.f. Wüthrich and Merz, 2008). The triangle assumes that the number of accident years is equal to the number of observed development periods.

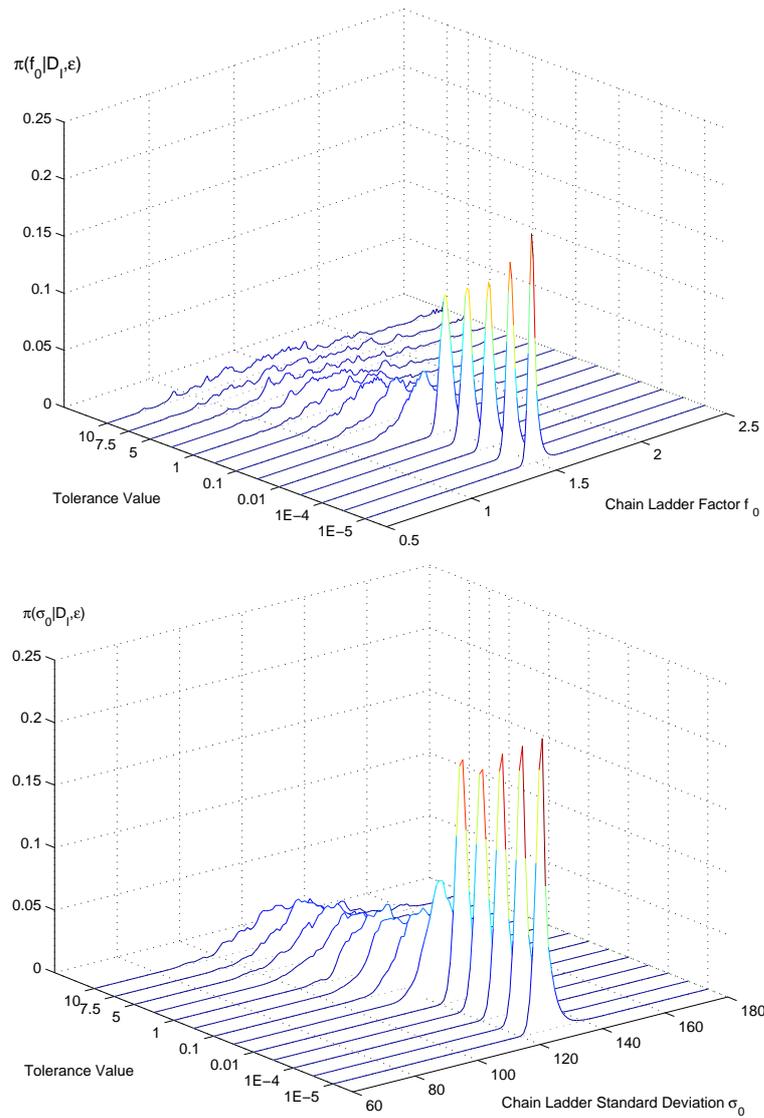


Fig. 4.7.1: Evolution of the marginal posterior density estimates of the chain ladder factor $\pi_t(f_0 | \mathcal{D}_I)$ (left) and the associated standard deviation $\pi_t(\sigma_0 | \mathcal{D}_I)$ (right) as a function of ϵ_t .

| Parameters | Year | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | $\widehat{C}_{i,I} - C_{i,I-i}$ |
|--------------------------|------|---------|-----------|-----------|-----------|-----------|------------|------------|------------|------------|------------|---------------------------------|
| $(\mathbf{f}^{(CL)})$ | 0 | | | | | | | | | | | 0 |
| $(\mathbf{f}^{(MMSE)})$ | | | | | | | | | | | | 0 |
| $(\mathbf{f}^{(CL)})$ | 1 | | | | | | | | | | 10,663,318 | 15,126 |
| $(\mathbf{f}^{(MMSE)})$ | | | | | | | | | | | 10,664,164 | 15,972 |
| $(\mathbf{f}^{(CL)})$ | 2 | | | | | | | | | 10,646,884 | 10,662,008 | 26,257 |
| $(\mathbf{f}^{(MMSE)})$ | | | | | | | | | | 10,645,322 | 10,661,290 | 25,540 |
| $(\mathbf{f}^{(CL)})$ | 3 | | | | | | | | 9,734,574 | 9,744,764 | 9,758,606 | 34,538 |
| $(\mathbf{f}^{(MMSE)})$ | | | | | | | | | 9,736,710 | 9,745,473 | 9,760,092 | 36,023 |
| $(\mathbf{f}^{(CL)})$ | 4 | | | | | | | 9,837,277 | 9,847,906 | 9,858,214 | 9,872,218 | 85,302 |
| $(\mathbf{f}^{(MMSE)})$ | | | | | | | | 9,840,743 | 9,853,536 | 9,862,404 | 9,877,198 | 90,283 |
| $(\mathbf{f}^{(CL)})$ | 5 | | | | | | 10,005,044 | 10,056,528 | 10,067,393 | 10,077,931 | 10,092,247 | 156,494 |
| $(\mathbf{f}^{(MMSE)})$ | | | | | | | 10,019,212 | 10,074,318 | 10,087,415 | 10,096,493 | 10,111,638 | 175,886 |
| $(\mathbf{f}^{(CL)})$ | 6 | | | | | 9,419,776 | 9,485,469 | 9,534,279 | 9,544,580 | 9,554,571 | 9,568,143 | 286,121 |
| $(\mathbf{f}^{(MMSE)})$ | | | | | | 9,422,181 | 9,501,327 | 9,553,584 | 9,566,004 | 9,574,613 | 9,588,975 | 306,953 |
| $(\mathbf{f}^{(CL)})$ | 7 | | | | 8,445,057 | 8,570,389 | 8,630,159 | 8,674,568 | 8,683,940 | 8,693,030 | 8,705,378 | 449,167 |
| $(\mathbf{f}^{(MMSE)})$ | | | | | 8,448,582 | 8,576,155 | 8,648,195 | 8,695,760 | 8,707,065 | 8,714,901 | 8,727,973 | 471,761 |
| $(\mathbf{f}^{(CL)})$ | 8 | | | 8,243,496 | 8,432,051 | 8,557,190 | 8,616,868 | 8,661,208 | 8,670,566 | 8,679,642 | 8,691,971 | 1,043,242 |
| $(\mathbf{f}^{(MMSE)})$ | | | | 8,229,268 | 8,421,009 | 8,548,167 | 8,619,971 | 8,667,381 | 8,678,649 | 8,686,460 | 8,699,489 | 1,050,760 |
| $(\mathbf{f}^{(CL)})$ | 9 | | 8,470,989 | 9,129,696 | 9,338,521 | 9,477,113 | 9,543,206 | 9,592,313 | 9,602,676 | 9,612,728 | 9,626,383 | 3,950,814 |
| $(\mathbf{f}^{(MMSE)})$ | | | 8,477,596 | 9,121,045 | 9,333,566 | 9,474,503 | 9,554,088 | 9,606,636 | 9,619,125 | 9,627,782 | 9,642,223 | 3,966,655 |
| $\widehat{f}_j^{(CL)}$ | | 1.4925 | 1.0778 | 1.0229 | 1.0148 | 1.0070 | 1.0051 | 1.0011 | 1.0010 | 1.0014 | | 6,047,061 |
| $\sigma_j^{(CL)}$ | | 135.253 | 33.803 | 15.760 | 19.847 | 9.336 | 2.001 | 0.823 | 0.219 | 0.059 | | |
| $\widehat{f}_j^{(MMSE)}$ | | 1.4937 | 1.0759 | 1.0233 | 1.0151 | 1.0084 | 1.0055 | 1.0013 | 1.0009 | 1.0015 | | 6,139,834 |
| $\sigma_j^{(MMSE)}$ | | 132.917 | 34.566 | 14.742 | 21.972 | 8.547 | 2.736 | 0.789 | 0.159 | 0.061 | | |

Tab. 4.2: Predicted parameter estimates, $\widehat{\mathbf{f}}$, $\widehat{\sigma}$, cumulative chain ladder claims, $\widehat{C}_{i,j}$, and estimated chain ladder reserves under the classical (CL) and Bayesian (MMSE) models.

References

- [1] Andrieu C., Freitas N.D., Doucet A. and Jordan M. "An introduction to MCMC for machine learning". *Machine Learning*, 2003, 50, 5-43.
- [2] Arulampalam S., Maskell S., Gordon N. and Clapp T. "A tutorial on particle filters for on-line non-linear/non-Gaussian Bayesian tracking". *IEEE Transactions on Signal Processing*, 2002, 50, 174 - 188.
- [3] Beaumont M.A., Cornuet J.-M., Marin J.-M. and Robert C. P. "Adaptivity for ABC algorithms: the ABC-PMC scheme". *Biometrika*, 2009, in press.
- [4] Beaumont M.A., Zhang W. and Balding D. J. "Approximate Bayesian computation in population genetics". *Genetics*, 2002, 162, 2025 - 2035.
- [5] Blum M.G.B. "Approximate Bayesian computation: a non-parametric perspective". *Université Joseph Fourier, Grenoble, France*, 2009.
- [6] Bortot P., Coles S.G. and Sisson S.A. "Inference for stereological extremes". *Journal of the American Statistical Association*, 2007, 102, 84-92.
- [7] Chopin N. "Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference". *Annals of Statistics*, 2004, 32, 2385 - 2411.
- [8] Cornebise J., Moulines E. and Olsson J. "Adaptive methods for sequential importance sampling with application to state space models". *Proceedings of the 16th European Signal Processing Conference, Lausanne*, 2008.
- [9] Crisan D. and Doucet A. "A survey of convergence results on particle filtering for practitioners". *IEEE Transactions on Signal Processing*, 2002, 50, 736-746.
- [10] Moral P.D. *Feynman-Kac formulae: Genealogical and interacting particle systems with applications*. Springer, New York, 2004.
- [11] Moral P.D., Doucet A. and Jasra A. "Adaptive sequential Monte Carlo samplers". *University of Bordeaux*, 2008.
- [12] Moral P.D., Doucet A. and Jasra A. "Sequential Monte Carlo samplers". *Journal of the Royal Statistical Society, Series B*, 2006, 68, 411 - 436.

- [13] Moral P.D., Jacod J. and Protter P. "The Monte Carlo method for filtering with discrete-time observations". *Probability Theory and Related Fields*, 2001, 120, 346-368.
- [14] Doucet A., de Freitas N. and Gordon N. (ed.) *Sequential Monte Carlo Methods in Practice*, Springer-Verlag, New York, 2001.
- [15] Doucet A. and Johansen A.M. "A Tutorial on Particle filtering and smoothing: Fifteen years later". In *Oxford Handbook of Nonlinear Filtering*, D. Crisan and B. Rozovsky (eds.), Oxford University Press, 2009, To appear.
- [16] Gisler A. and Wüthrich M.V. "Credibility for the chain ladder reserving method". *Astin Bulletin*, 2008, 38, 565-600.
- [17] Johansen A.M. and Doucet A. "A note on auxiliary particle filters". *Statistics and Probability Letters*, 2008, (12) 78, 1498-1504.
- [18] Jasra A., Stephens D.A. and Holmes C.C. "On population-based simulation for static inference". *University of Cambridge*, 2006.
- [19] Johansen A.M., Moral P.D. and Doucet A. "Sequential Monte Carlo samplers for rare events". *Proceedings of the 6th International Workshop on Rare Event Simulation*, 2006.
- [20] Kitigawa G. "Monte Carlo filter and smoother for non-Gaussian, non-linear state space models". *Journal of Computational and Graphical Statistics*, 1996, 5, 1-25.
- [21] Kunsch H.R. "Recursive Monte Carlo filters: Algorithms and theoretical analysis". *Annals of Statistics*, 2005, 33, 1983 - 2021.
- [22] Gland F.L. and Oudjane N. "A sequential particle algorithm that keeps the particle system alive". *Proceedings of the 13th European Signal Processing Conference, Antalya*, 2005.
- [23] Gland F.L. and Oudjane N. "Stability and uniform approximation of nonlinear filters using the Hilbert metric, and applications to particle filters". *The Annals of Applied Probability*, 2004, 14, 144-187.
- [24] Liu J. and Chen R. "Sequential Monte Carlo for dynamic systems". *Journal of the American Statistical Association*, 1998, 93, 1032 - 1044.
- [25] Liu J., Chen R. and Wong W. "Rejection control and sequential importance sampling". *Journal of American Statistical Association*, 1998, 93, 1022-1031.
- [26] Liu J.S. *Monte Carlo Strategies in Scientific Computing*, Springer-Verlag, New York, 2001.
- [27] Mack T. "Distribution-free calculation of the standard error of chain ladder reserve estimates". *Astin Bulletin*, 1993, 23, 213-225.
- [28] Marjoram P., Molitor J., Plagnol V. and Tavaré S. "Markov chain Monte Carlo without likelihoods". *Proceedings of the Natational Academy of Science USA*, 2003, 100, 15324 - 15328.
- [29] Peters G.W. "Topics in Sequential Monte Carlo Samplers". *University of Cambridge*, 2005.
- [30] Peters G.W., Wüthrich M.V. and Shevchenko P.V. "Chain ladder method: Bayesian bootstrap versus classical bootstrap". *University of New South Wales*, 2008.

-
- [31] Peters G.W., Sisson S.A. and Fan, Y. "Design efficiency for 'likelihood free' sequential Monte Carlo algorithms". *University of New South Wales*, 2008.
- [32] Ratmann O., Andrieu C., Hinkley T., Wiuf C. and Richardson S. "Model criticism based on likelihood-free inference, with an application to protein network evolution". *Proceedings of the Natational Academy of Science, USA*, 2009, 106, 10576-10581.
- [33] Reeves R.W. and Pettitt A.N.A *Theoretical framework for approximate Bayesian computation*, 2005.
- [34] Ridgeway G. and Madigan D. "A sequential Monte Carlo method for Bayesian analysis of massive data sets". *Data Mining and Knowledge Discovery*, 2002, 7, 301-319.
- [35] Sisson S.A., Fan Y. and Tanaka M.M. "Sequential Monte Carlo without likelihoods". *Proceedings of the Natational Academy of Science*, 2007, 104, 1760-1765.
- [36] Tavaré S., Balding D.J., Griffiths R.C. and Donnelly P. "Inferring coalescence times from DNA sequence data". *Genetics*, 1997, 145, 505-518.
- [37] Toni T., Welch D., Strelkowa N., Ipsen A. and Stumpf M.P.H. "Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems". *Journal of the Royal Statistical Society, Interface*, 2009, 6, 187-202.
- [38] Wüthrich M.V. and Merz M. *Stochastic claims reserving methods in insurance*. Wiley, 2008.
- [39] West M. "Approximating posterior distributions by mixtures". *Journal of the Royal Statistical Society, Series B*, 1993, 55, 409-422.
- [40] Yao J. "Bayesian approach for prediction error in chain ladder claims reserving". *38th International ASTIN Colloquium, July*, 2008.

Journal Paper 3

"Mathematics, rightly viewed, possesses not only truth, but supreme beauty - a beauty cold and austere, like that of sculpture."

Bertrand Russell

Peters G.W., Sisson S.A. and Fan Y. (2009) "Likelihood-free Bayesian inference for α -stable models". In review.

This work was instigated by the first author who can claim around 80% of the credit for the contents. His work included developing large amounts of the theory included, in particular related to the Sequential Monte Carlo Samplers PRC-ABC algorithm. This work is very novel in terms of analysis of this complicated class of multivariate distributions. It has already been presented as a poster session and an invited seminar at two international conferences and has received positive reviews from leading academics in this field. It is expected this paper will be accepted for publication. Permission from all the co-authors has been granted for submission of this paper as part of the thesis.

Likelihood-free Bayesian inference for α -stable models

Gareth W. Peters (*corresponding author*)

CSIRO Mathematical and Information Sciences, Locked Bag 17, North Ryde, NSW, 1670, Australia;

UNSW Mathematics and Statistics Department, Sydney, 2052, Australia

S. A. Sisson

UNSW Mathematics and Statistics Department, Sydney, 2052, Australia

Y. Fan

UNSW Mathematics and Statistics Department, Sydney, 2052, Australia

Submitted: 08 December 2009

Abstract α -stable distributions are utilised as models for heavy-tailed noise in many areas of statistics, finance and signal processing engineering. However, in general, neither univariate nor multivariate α -stable models admit closed form densities which can be evaluated point-wise. This complicates the inferential procedure. As a result, α -stable models are practically limited to the univariate setting under the Bayesian paradigm, and to bivariate models under the classical framework. In this article we develop a novel Bayesian approach to modelling univariate and multivariate α -stable distributions based on recent advances in “likelihood-free” inference. We present an evaluation of the performance of this procedure in 1, 2 and 3 dimensions, and provide an analysis of real daily currency exchange rate data. The proposed approach provides a feasible inferential methodology at a moderate computational cost.

Key words: α -stable distributions; Approximate Bayesian computation; Likelihood-free inference; Sequential Monte Carlo samplers.

5.1 Introduction

Models constructed with α -stable distributions possess several useful properties, including infinite variance, skewness and heavy tails (47; 1; 42; 33). α -stable distributions provide no general analytic expressions for the density, median, mode or entropy, but are uniquely specified by their characteristic function, which has several parameterizations. Considered as generalizations of the Gaussian distribution, they are defined as the class of location-scale distributions which are closed under convolutions. α -stable distributions have found application in many areas of statistics, finance and signal processing engineering as models for impulsive, heavy tailed noise processes (23; 13; 14; 31; 16; 28).

The univariate α -stable distribution is typically specified by four parameters: $\alpha \in (0, 2]$ determining the rate of tail decay; $\beta \in [-1, 1]$ determining the degree and sign of asymmetry (skewness); $\gamma > 0$ the scale (under some parameterizations); and $\delta \in \mathbb{R}$ the location (20). The parameter α is termed the characteristic exponent, with small and large α implying heavy and light tails respectively. Gaussian ($\alpha = 2, \beta = 0$) and Cauchy ($\alpha = 1, \beta = 0$) distributions provide the only analytically tractable sub-members of this family. In general, as α -stable models admit no closed form expression for the density which can be evaluated pointwise (excepting Gaussian and Cauchy members), inference typically proceeds via the characteristic function.

This paper is concerned with constructing both univariate and multivariate Bayesian models in which the likelihood model is from the class of α -stable distributions. This is known to be a difficult problem. Existing methods for Bayesian α -stable models are limited to the univariate setting (6; 15; 16; 22; 8; 41).

Inferential procedures for α -stable models may be classified as auxiliary variable methods, inversion plus series expansion approaches and density estimation methods. The auxiliary variable Gibbs sampler (6) increases the dimension of the parameter space from 4 (α, β, γ and δ) to $n + 4$, where n is the number of observations. As strong correlations between parameters and large sample sizes are common in the α -stable setting, this results in a slowly mixing Markov chain since Gibbs moves are limited to moving parallel to the axes (e.g. 30). Other Markov chain Monte Carlo (MCMC) samplers (12; 22) adopt inversion techniques for numerical integration of the characteristic function, employing inverse Fourier transforms combined with a series expansion (3) to accurately estimate distributional tails. This is performed at each iteration of the Markov chain to evaluate the likelihood, and is accordingly computationally intensive. In addition the quality of the resulting approximation is sensitive to the spacing of the fast Fourier transform grid and the point at which the series expansion begins (22).

Univariate density estimation methods include integral representations (47), parametric mixtures (32; 26) and numerical estimation through splines and series expansions (35; 34). 26 approximates symmetric stable distributions using a mixture of Gaussian and Cauchy densities. 11 and 29 approximate the α -stable density through spline polynomials, and 19 via a mixture of Gaussian distributions. Parameter estimation has been performed by an expectation-maximization (EM) algorithm (21) and by method of (partial) moments (38; 46). Implemented

within an MCMC sampler, such density estimation methods would be highly computational.

None of the above methods easily generalize to the multivariate setting. It is currently only practical to numerically evaluate two dimensional α -stable densities via inversion of the characteristic function. Here the required computation is a function of α and the number and spread of masses in the discrete spectral representation (35; 32). Beyond two dimensions this procedure becomes untenably slow with limited accuracy.

In this article we develop practical Bayesian inferential methods to fit univariate and multivariate α -stable models. To the best of our knowledge, no practical Bayesian methods have been developed for the multivariate model as the required computational complexity increases dramatically with model dimension. The same is true of classical methods beyond two dimensions. Our approach is based on recent developments in “likelihood-free” inference, which permits approximate posterior simulation for Bayesian models without the need to explicitly evaluate the likelihood.

In Section 5.2 we briefly introduce likelihood-free inference and the sampling framework used in this article. Section 5.3 presents the Bayesian α -stable model, with a particular focus on summary statistic specification, a critical component of likelihood-free inference. We provide an evaluation of the performance of the proposed methodology in Section 5.4, based on controlled simulation studies in 1, 2 and 3 dimensions. Finally, in Section 5.5 we demonstrate an analysis of real daily currency data under both univariate and multivariate settings, and provide comparisons with existing methods. We conclude with a discussion.

5.2 Likelihood-free models

Computational procedures to simulate from posterior distributions, $\pi(\theta|y) \propto \pi(y|\theta)\pi(\theta)$, of parameters $\theta \in \Theta$ given observed data $y \in \mathcal{X}$, are well established (e.g. 5). However when pointwise evaluation of the likelihood function $\pi(y|\theta)$ is computationally prohibitive or intractable, alternative procedures are required. Likelihood-free methods (also known as *approximate Bayesian computation*) permit simulation from an approximate posterior model while circumventing explicit evaluation of the likelihood function (45; 2; 77; 43; 39).

Assuming data simulation $x \sim \pi(x|\theta)$ under the model given θ is easily obtainable, likelihood-free methods embed the posterior $\pi(y|\theta)$ within an augmented model

$$\pi_{LF}(\theta, x|y) \propto \pi_{\epsilon}(y|x, \theta)\pi(x|\theta)\pi(\theta), \quad (5.2.1)$$

where $x \sim \pi(x|\theta)$, $x \in \mathcal{X}$, is an auxiliary parameter on the same space as the observed data y . The function $\pi_{\epsilon}(y|x, \theta)$ is typically a standard smoothing kernel (e.g. 4) with scale parameter ϵ , which weights the intractable posterior with high values in regions when the observed data y and auxiliary data x are similar. For example, uniform kernels are commonplace in likelihood-free models (e.g. 77; 43), although alternatives such as Epanechnikov (2) and Gaussian kernels

(36) provide improved efficiency. The resulting approximation to the true posterior target distribution

$$\pi_{LF}(\theta|y) \propto \int_{\mathcal{X}} \pi_{\epsilon}(y|x, \theta) \pi(x|\theta) \pi(\theta) dx = \pi(\theta) \mathbb{E}_{\pi(x|\theta)}[\pi_{\epsilon}(y|x, \theta)] \quad (5.2.2)$$

improves as ϵ decreases, and exactly recovers the target posterior as $\epsilon \rightarrow 0$, as then $\lim_{\epsilon \rightarrow 0} \pi_{\epsilon}(y|x, \theta)$ becomes a point mass at $y = x$ (99).

Posterior simulation from $\pi_{LF}(\theta|y)$ can then proceed via standard simulation algorithms, replacing pointwise evaluations of $\pi_{LF}(\theta|y)$ with Monte Carlo estimates through the expectation (5.2.2), based on draws $x^1, \dots, x^P \sim \pi(x|\theta)$ from the model (e.g. 77). Alternatively, simulation from the joint posterior $\pi_{LF}(\theta, x|y)$ is available by contriving to cancel the intractable likelihoods $\pi(x|\theta)$ in sample weights or acceptance probabilities. For example, importance sampling from the prior predictive distribution $\pi(\theta, x) = \pi(x|\theta)\pi(\theta)$ results in an importance weight of $\pi_{LF}(\theta, x|y)/\pi(\theta, x) \propto \pi_{\epsilon}(y|x, \theta)$, which is free of likelihood terms. See 44 for a discussion of marginal and joint-space likelihood-free samplers.

In general, the distribution of $\pi(x|\theta)$ will be diffuse, unless x is discrete and $\dim(x)$ is small. Hence, generating $x \sim \pi(\cdot|\theta)$ with $x \approx y$ is improbable for realistic datasets y , and as a result the degree of computation required for a good likelihood-free approximation $\pi_{LF}(\theta|y) \approx \pi(\theta|y)$ (i.e. with small ϵ) will be prohibitive. In practice, the function $\pi_{\epsilon}(y|x, \theta)$ is expressed through low dimensional vectors of summary statistics, $S(\cdot)$, such that $\pi_{\epsilon}(y|x, \theta)$ weights the intractable posterior through (5.2.1) with high values in regions where $S(y) \approx S(x)$.

If $S(\cdot)$ is sufficient for θ , then letting $\epsilon \rightarrow 0$ recovers $\pi_{LF}(\theta|y) = \pi(\theta|y)$ as before, but with more acceptable computational overheads, as $\dim(S(x)) \ll \dim(x)$. As sufficient summary statistics are generally unavailable, the use of non-sufficient statistics is commonplace. The effect of less efficient estimators of θ in (5.2.2) is a more diffuse approximation of $\pi(\theta|y)$. Hence the choice of summary statistics in any application is critical, with the ideal being low-dimensional, efficient and near-sufficient.

In this article, we implement the likelihood-free sequential Monte Carlo sampler of 36, detailed in Appendix A. As the class of particle-based algorithms is the most efficient currently available in likelihood-free computation (e.g. 27), and within this class, the sampler of 36 is the only one to allow non-uniform functions $\pi_{\epsilon}(y|x, \theta)$, this sampler provides the best combination of efficient simulation and flexible modelling.

5.3 Bayesian α -stable models

We now develop univariate and multivariate Bayesian α -stable models. Unlike existing methods, likelihood-free inference is independent of model parameterization.

5.3.1 Univariate α -stable Models

Denote the characteristic function of n i.i.d. univariate α -stable distributed random variables X_1, \dots, X_n by $\Phi_X(t)$. A popular and convenient parameterization is

$$\Phi_X(t) = \begin{cases} \exp\left(i\delta t - \gamma^\alpha |t|^\alpha \left[1 + i\beta \tan\frac{\pi\alpha}{2} \operatorname{sgn}(t) \left(|\gamma t|^{1-\alpha} - 1\right)\right]\right) & \text{if } \alpha \neq 1 \\ \exp\left(i\delta t - \gamma |t| \left[1 + i\beta \frac{2}{\pi} \operatorname{sgn}(t) \ln(\gamma |t|)\right]\right) & \text{if } \alpha = 1, \end{cases} \quad (5.3.1)$$

where $\operatorname{sgn}(t) = \frac{t}{|t|}$ and $i^2 = -1$ (e.g. 42). Many alternative parameterizations are detailed in 33 and 47. Under (5.3.1), the intractable stable density function is continuous and unimodal, taking support on $(-\infty, 0)$ if $\alpha < 1, \beta = -1$; $(0, \infty)$ if $\alpha < 1, \beta = 1$ and $(-\infty, \infty)$ otherwise.

Efficient simulation of auxiliary data, $x \sim \pi(x|\theta)$, under the model is critical for the performance of likelihood-free methods (Section 5.2). Here, it is straightforward to generate α -stable variates under the model defined by the characteristic function (5.3.1) (e.g. 10; 33). This approach is provided in Appendix B.

Summary statistics

A key component of likelihood-free inference is the availability of low-dimensional, efficient and near-sufficient summary statistics. Since α -stable models can possess infinite variance ($\alpha > 1$) and infinite mean ($\alpha < 1$), this choice must be made with care. Here we present several candidate summary vectors, S_1 – S_5 , previously utilized for parameter estimation in the univariate α -stable model. In Section 5.4 we evaluate the performance of these vectors, and provide informed recommendations for the choice of summary statistics under the likelihood-free framework.

S_1 McCulloch's Quantiles

25 and 26 estimate model parameters based on sample quantiles, while correcting for estimator skewness due to the evaluation of $\hat{q}_p(x)$, the p^{th} quantile of x , with a finite sample. Here, the data $x_{(i)}$ are arranged in ascending order and matched with $\hat{q}_{s(i)}(x)$, where $s(i) = \frac{2i-1}{2n}$. Linear interpolation to p from the two adjacent $s(i)$ values then establishes $\hat{q}_p(x)$ as a consistent estimator of the true quantiles. Inversion of the functions

$$\hat{v}_\alpha = \frac{\hat{q}_{0.95}(\cdot) - \hat{q}_{0.05}(\cdot)}{\hat{q}_{0.75}(\cdot) - \hat{q}_{0.25}(\cdot)}, \quad \hat{v}_\beta = \frac{\hat{q}_{0.95}(\cdot) + \hat{q}_{0.05}(\cdot) - 2\hat{q}_{0.5}(\cdot)}{\hat{q}_{0.95}(\cdot) - \hat{q}_{0.05}(\cdot)}, \quad \hat{v}_\gamma = \frac{\hat{q}_{0.75}(\cdot) - \hat{q}_{0.25}(\cdot)}{\gamma}$$

then provides estimates of α, β and γ . Note that from a computational perspective, inversion of v_α, v_β or v_γ is not required under likelihood-free methods. Finally, we estimate δ by the sample mean \bar{X} (when $\alpha > 1$). Hence $S_1(x) = (\hat{v}_\alpha, \hat{v}_\beta, \hat{v}_\gamma, \bar{x})$.

S₂ *Zolotarev's Transformation*

Based on a transformation of data from the α -stable family $X \rightarrow Z$, 47 (p.16) provides an alternative parameterization of the α -stable model $(\alpha, \beta, \gamma, \delta) \leftrightarrow (\nu, \eta, \tau)$ with a characteristic function of the form

$$\log \Phi_Z(t) = -\exp \left\{ \nu^{-\frac{1}{2}} \left[\log |t| + \tau - i \frac{\pi}{2} \eta \operatorname{sgn}(t) \right] + \mathbb{C} \left(\nu^{-\frac{1}{2}} - 1 \right) \right\}, \quad (5.3.2)$$

where \mathbb{C} is Euler's constant, and where $\nu \geq \frac{1}{4}$, $|\eta| \leq \min\{1, 2\sqrt{\nu} - 1\}$ and $|\tau| < \infty$. This parameterization has the advantage that logarithmic moments have simple expressions in terms of parameters to be estimated. For a fixed constant $0 < \xi \leq \frac{1}{2}$ (47 recommends $\xi = 0.25$) and for integer $n/3$, the transformation is

$$Z_j = X_{3j-2} - \xi X_{3j-1} - (1 - \xi) X_{3j}, \quad j = 1, 2, \dots, n/3.$$

Defining $V_j = \log |Z_j|$ and $U_j = \operatorname{sgn}(X_j)$, estimates for ν, η and τ are then given by

$$\hat{\nu} = \max\{\tilde{\nu}, (1 + |\hat{\eta}|)^2 / 4\}, \quad \hat{\eta} = \mathbb{E}(U), \quad \hat{\tau} = \mathbb{E}(V),$$

where $\tilde{\nu} = \frac{6}{\pi^2} S^2(V) - \frac{3}{2} S^2(U) + 1$, using sample variances $S^2(V)$ and $S^2(U)$. As before, δ is estimated by \bar{X} (for $\alpha > 1$), and so $S_2(x) = (\hat{\nu}, \hat{\eta}, \hat{\tau}, \bar{x})$.

S₃ *Press's Method Of Moments*

For $\alpha \neq 1$ and unique evaluation points t_1, t_2, t_3, t_4 , the method of moments equations obtained from $\log \Phi_X(t)$ can be solved to obtain (38; 46)

$$\begin{aligned} \log(\hat{\gamma}) &= \frac{\log |t_1| \log(-\log |\Phi(t_2)|) - \log |t_2| \log(-\log |\Phi(t_1)|)}{\log |t_1/t_2|}, & \hat{\alpha} &= \frac{\log \frac{\log |\Phi(t_1)|}{\log |\Phi(t_2)|}}{\log |t_1/t_2|} \\ \hat{\beta} \Big|_{\hat{\alpha}, \hat{\gamma}} &= \frac{\hat{u}(t_4)/t_4 - \hat{u}(t_3)/t_3}{(|t_4|^{\hat{\alpha}-1} - |t_3|^{\hat{\alpha}-1}) \hat{\gamma} \hat{\alpha} \tan\left(\frac{\hat{\alpha}\pi}{2}\right)}, & \hat{\delta} \Big|_{\hat{\alpha}} &= \frac{|t_4|^{\hat{\alpha}-1} \hat{u}(t_3)/t_3 - |t_3|^{\hat{\alpha}-1} \hat{u}(t_4)/t_4}{|t_4|^{\hat{\alpha}-1} - |t_3|^{\hat{\alpha}-1}} \end{aligned}$$

where $\hat{u}(t) = \tan^{-1}[\sum_{i=1}^n \cos(tx_i) / \sum_{i=1}^n \sin(tx_i)]$. We adopt the evaluation points $t_1 = 0.2, t_2 = 0.8, t_3 = 0.1$ and $t_4 = 0.4$ as recommended by 18, and accordingly obtain $S_3(x) = (\hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{\delta})$.

S₄ *Empirical Characteristic Function*

The empirical characteristic function, $\hat{\Phi}_X(t) = \frac{1}{n} \sum_{j=1}^n e^{itX_j}$ for $t \in (-\infty, \infty)$, can be used as the basis for summary statistics when standard statistics are not available. E.g. this may occur through the non-existence of moment generating functions. Hence, we specify $S_4(x) = (\hat{\Phi}_X(t_1), \dots, \hat{\Phi}_X(t_{20}))$ where $t_i \in \{\pm 0.5, \pm 1, \pm 1.5, \dots, \pm 5\}$.

S₅ *Mean, Quantiles and Kolmogorov-Smirnov Statistic*

The Kolmogorov-Smirnov statistic is defined as $KS(X) = \sup_z |F_n^X(z) - F_n^Y(z)|$, the largest absolute deviation between the empirical cumulative distribution functions of auxiliary (X)

and observed (Y) data, where $F_n^X(z) = \frac{1}{n} \sum_{i=1}^n I_{(X_i \leq z)}$ and $I_{(X_i \leq z)} = 1$ if $X_i \leq z$ and 0 otherwise. We specify $S_5(x) = (\bar{x}, \{\hat{q}_p(x)\}, KS(x))$, where the set of sample quantiles $\{\hat{q}_p(x)\}$ is determined by $p \in \{0.01, 0.05, 0.1, 0.15, \dots, 0.9, 0.95, 0.99\}$.

Likelihood-free inference may be implemented under any parameterization which permits data generation under the model, and for which the summary statistics are well defined. From the above S_1 – S_5 are jointly well defined for $\alpha > 1$. Hence, to complete the specification of the univariate α -stable model we adopt the independent uniform priors $\alpha \sim U[1.1, 2]$, $\beta \sim U[-1, 1]$, $\gamma \sim U[0, 300]$ and $\delta \sim U[-300, 300]$ (e.g. 6). Note that the prior for α has a restricted domain, reflecting the use of sample moments in S_1 – S_3 and S_5 . For S_4 we may adopt $\alpha \sim U(0, 2]$.

5.3.2 Multivariate α -stable Models

Bayesian model specification and simulation in the multivariate α -stable setting is challenging (35; 34; 42). Here we follow (34), who defines the multivariate model for the random vector $\mathbf{X} = (X_1, \dots, X_d) \in \mathbb{R}^d$ through the functional equations

$$\begin{aligned} \sigma^\alpha(\mathbf{t}) &= \int_{S_d} |\langle \mathbf{t}, \mathbf{s} \rangle|^\alpha \Gamma(ds); & \beta(\mathbf{t}) &= \sigma^{-\alpha}(\mathbf{t}) \int_{S_d} \text{sgn}(\langle \mathbf{t}, \mathbf{s} \rangle) |\langle \mathbf{t}, \mathbf{s} \rangle|^\alpha \Gamma(ds) \\ \mu(\mathbf{t}) &= \begin{cases} \langle \mathbf{t}, \boldsymbol{\mu}^0 \rangle & \alpha \neq 1 \\ \langle \mathbf{t}, \boldsymbol{\mu}^0 \rangle - \frac{2}{\pi} \int_{S_d} \langle \mathbf{t}, \mathbf{s} \rangle \ln |\langle \mathbf{t}, \mathbf{s} \rangle| \Gamma(ds) & \alpha = 1 \end{cases} \end{aligned}$$

where $\mathbf{t} = (t_1, \dots, t_d)$, $\mathbf{s} = (s_1, \dots, s_d)$, $\langle \mathbf{t}, \mathbf{s} \rangle = \sum_{i=1}^d t_i s_i$, \mathbb{S}^d denotes the unit d -sphere, $\Gamma(ds)$ denotes the unique spectral measure, and $\sigma^\alpha(\mathbf{t})$ represents scale, $\beta(\mathbf{t})$ skewness and $\mu(\mathbf{t})$ location (through the vector $\boldsymbol{\mu}^0 = (\mu_1^0, \dots, \mu_d^0)$). Scaling properties of the functional equations, e.g. $\sigma^\alpha(r\mathbf{t}) = r\sigma^\alpha(\mathbf{t})$ for $r > 0$, mean that it is sufficient to consider them on \mathbb{S}^d (32). The corresponding characteristic function is

$$\Phi_{\mathbf{X}}(\mathbf{t}) = \mathbb{E} \exp(i \langle \mathbf{X}, \mathbf{t} \rangle) = \exp(-I_{\mathbf{X}}(\mathbf{t}) + i \langle \boldsymbol{\mu}^0, \mathbf{t} \rangle)$$

with $I_{\mathbf{X}}(\mathbf{t}) = \int_{\mathbb{S}^d} \psi_\alpha(\langle \mathbf{t}, \mathbf{s} \rangle) \Gamma(ds)$, where the function ψ_α is given by

$$\psi_\alpha(u) = \begin{cases} |u|^\alpha (1 - i \text{sgn}(u) \tan(\frac{\pi\alpha}{2})) & \alpha \neq 1 \\ |u|^\alpha (1 - i \frac{2}{\pi} \text{sgn}(u) \ln |u|) & \alpha = 1. \end{cases}$$

The spectral measure $\Gamma(\cdot)$ and location vector $\boldsymbol{\mu}^0$ uniquely characterize the multivariate distribution (42) and carry essential information relating to the dependence between the elements of \mathbf{X} . The continuous spectral measure is typically well approximated by a discrete set of k Dirac masses $\Gamma(\cdot) = \sum_{j=1}^k w_j \delta_{\mathbf{s}_j}(\cdot)$ (e.g. 7) where w_j and $\delta_{\mathbf{s}_j}(\cdot)$ respectively denote the weight and Dirac

mass of the j^{th} spectral mass at location $\mathbf{s}_j \in \mathbb{S}^d$. By simplifying the integral in $I_{\mathbf{X}}(\mathbf{t})$, computation with the characteristic function $\Phi_{\mathbf{X}}^*(\mathbf{t}) = \exp\{-\sum_{j=1}^k w_j \psi_\alpha(\langle \mathbf{t}, \mathbf{s}_j \rangle)\}$ becomes tractable and data generation from the distribution defined by $\Phi_{\mathbf{X}}^*(\mathbf{t})$ is efficient (Appendix B). As with the univariate case (5.3.1), standard parameterizations of $\Phi_{\mathbf{X}}^*(\mathbf{t})$ will be discontinuous at $\alpha = 1$, resulting in poor estimates of location and $\Gamma(\cdot)$. In the multivariate setting this is overcome by Zolotarev's M-parameterization (35; 34). Although likelihood-free methods are parameterization independent, it is sensible to work with models with good likelihood properties.

In a Bayesian framework we parameterize the model via the spectral mass, which involves estimation of the weights $\mathbf{w} = (w_1, \dots, w_k)$ and locations $\mathbf{s}_{1:k} \in \mathbb{S}^{d \times k}$ of $\Gamma(\cdot)$. For $k = 2$ this corresponds to $\mathbf{s}_i = (\cos \phi_i, \sin \phi_i) \in \mathbb{S}^2$. More generally, for $\mathbf{s}_i = (s_i^1, \dots, s_i^d) \in \mathbb{S}^d$, we use hyperspherical coordinates $\phi_i = (\phi_1^i, \dots, \phi_{d-1}^i)$, where

$$\phi_{d-1}^i = \tan^{-1} \left(\frac{s_i^d}{s_i^{d-1}} \right), \quad \dots, \quad \phi_1^i = \tan^{-1} \left(\frac{\sqrt{(s_i^d)^2 + (s_i^{d-1})^2 + \dots + (s_i^2)^2}}{s_i^1} \right).$$

We define priors for the parameters $(\mathbf{w}, \phi_{1:k}, \boldsymbol{\mu}^0, \alpha)$, with $\phi_{1:k} = (\phi_1, \dots, \phi_k)$, as $\mathbf{w} \sim \text{Dirichlet}(s, \dots, s)$, $\phi_{j'}^i \sim \text{U}(0, 2\pi)$, $\mu_j^0 \sim \text{N}(\xi, \kappa^{-1})$, $\alpha \sim \text{U}(0, 2)$ for $i = 1, \dots, k$, $j = 1, \dots, d$, $j' = 1, \dots, d-1$ and impose the ordering constraint $s_{i-1}^1 \leq s_i^1$ for all i on the first element of each vector \mathbf{s}_i .

Note that by treating the weights and locations of the spectral masses as unknown parameters, these may be identified with those regions of the spectral measure with significant posterior mass. This differs with the approach of 34 where the spectral mass is evaluated at a large number of deterministic grid locations. In estimating $\Gamma(\cdot) \mid \mathbf{X}$ a significant reduction in the number of required projection vectors is achieved (Sections 5.3.2 and 5.4.2). Further, the above prior specification does not penalize placement of spectral masses in close proximity. While this proved adequate for the presented analyses, alternative priors may usefully inhibit spectral masses at similar locations (37).

Summary statistics

S₆ Nolan, Panorska & McCulloch Projection Method

For the d -variate α -stable observations, $\mathbf{X}_i, i = 1, \dots, n$, we take projections of \mathbf{X}_i onto a unit hypersphere in the direction $\mathbf{t} \in \mathbb{S}^d$. This produces a set of n univariate values, $\mathbf{X}^{\mathbf{t}} = (X_1^{\mathbf{t}}, \dots, X_n^{\mathbf{t}})$, where $X_i^{\mathbf{t}} = \langle \mathbf{X}_i, \mathbf{t} \rangle$. The information in $\mathbf{X}^{\mathbf{t}}$ can then be summarized by any of the univariate summary statistics $S_1(\mathbf{X}^{\mathbf{t}}), \dots, S_5(\mathbf{X}^{\mathbf{t}})$. This process is repeated for multiple projections over $\mathbf{t}_1, \dots, \mathbf{t}_\tau$. With the location parameter $\boldsymbol{\mu}^0$ estimated by $\bar{\mathbf{x}}$, for sufficient numbers of projection vectors, τ , the summary statistics $S_6(\mathbf{X}) = (\bar{\mathbf{x}}, S_s(\mathbf{X}^{\mathbf{t}_1}), \dots, S_s(\mathbf{X}^{\mathbf{t}_\tau}))$ for some $s \in \{1, 2, 3, 4, 5\}$, will capture much of the information contained in the multivariate data, if S_s is itself informative. The best choice of univariate summary vector S_s will be determined in Section 5.4.1. We adopt a randomized approach to the selection of the projection vectors $\mathbf{t}_1, \dots, \mathbf{t}_\tau$, avoiding curse of dimensionality issues as the dimension d increases (34).

5.4 Evaluation of model and sampler performance

We now analyze the performance of the Bayesian α -stable models and likelihood-free sampler in a sequence of simulation studies. For the univariate model, we evaluate the capability of the summary statistics S_1, \dots, S_5 , and contrast the results with the samplers of 6 and 22. The performance of the multivariate model under the statistics S_6 is then considered for two and three dimensions.

In the following, we implement the likelihood-free sequential Monte Carlo algorithm of 36 (Appendix A) in order to simulate from the likelihood-free approximation to the true posterior $\pi_{LF}(\theta|y) \approx \pi(\theta|y)$ given by (5.2.2). This algorithm samples directly from $\pi_{LF}(\theta|y)$ using density estimates based on $P \geq 1$ Monte Carlo draws $x^1, \dots, x^P \sim \pi(x|\theta)$ from the model. We define $\pi_\epsilon(y|x, \theta)$ as a Gaussian kernel so that the summary statistics $S(y) \sim N(S(x), \epsilon^2 \Sigma)$ for a suitably chosen Σ . All inferences are based on $N = 1000$ particles drawn from $\pi_{LF}(\theta|y)$. Detail of algorithm implementation is removed to Appendix A for clarity of exposition.

5.4.1 Univariate summary statistics and samplers

We simulate $n = 200$ observations, y , from a univariate α -stable distribution with parameter values $\alpha = 1.7$, $\beta = 0.9$, $\gamma = 10$ and $\delta = 10$. We then implement the likelihood-free sampler targeting $\pi_{LF}(\theta|y)$ for each of the univariate summary statistics S_1 - S_5 described in Section 5.3.1, with uniform priors for all parameters (Section 5.3.1). Alternative prior specifications were investigated (22; 32), with little impact on the results.

Posterior minimum mean squared error (MMSE) estimates for each parameter, averaged over 10 sampler replicates are detailed in Table 5.1. Monte Carlo standard errors are reported in parentheses. The results indicate that all summary vectors apart from S_5 estimate α and δ parameters well, and for γ , S_3 and S_4 perform poorly. Only S_1 gives reasonable results for β and for all parameters jointly. Figure 5.6.1 illustrates a progression of the MMSE estimates of each parameter using S_1 , from the likelihood-free SMC sampler output for each sampler replicate. As the sampler progresses, the scale parameter ϵ decreases, and the MMSE estimates identify the true parameter values as the likelihood-free posterior approximation improves.

The results in Table 5.1 are based on using $P = 1$ Monte Carlo draws from the model to estimate $\pi_{LF}(\theta|y)$ (c.f. 5.2.2) within the likelihood-free sampler. Repeating the above study using $P \in \{5, 10, 20\}$ produced very similar results, and so we adopt $P = 1$ for the sequel as the most computationally efficient choice.

For comparison, we also implement the auxiliary variable Gibbs sampler of 6 and the MCMC inversion and series expansion sampler of 22, based on chains of length 100,000 iterations (10,000 iterations burnin), and using their respective prior specifications. The Gibbs sampler performed poorly for most parameters. The MCMC method performed better, but has larger standard errors than the likelihood-free sampler using S_1 .

The MCMC sampler (22) performs likelihood evaluations via inverse Fast Fourier transform (FFT) with approximate tail evaluation using Bergstrom expansions. This approach is sensitive to α , which determines the threshold between the FFT and the series expansion. Further, as the tail becomes fatter, a finer spacing of FFT abscissae is required to control the bias introduced outside of the Bergstrom series expansion, significantly increasing computation. Overall, this sampler worked reasonably for α close to 2, though with deteriorating performance as α decreased. The Gibbs sampler (6) performed extremely poorly for most settings and datasets, even when using their proposed change of variables transformations. As such, the results in Table 5.1 represent simulations under which both Gibbs and MCMC samplers performed credibly, thereby typifying their best case scenario performance.

5.4.2 Multivariate samplers

We consider varying numbers of discrete spectral masses, k , in the approximation to the spectral measure $\Gamma(\cdot) = \sum_{j=1}^k w_j \delta_{s_j}(\cdot)$. We assume that the number of spectral masses is known *a priori*, and denote the d -variate α -stable distribution by $\mathcal{S}_\alpha(d, k, \mathbf{w}, \phi_{1:k}, \boldsymbol{\mu}^0)$. Priors are specified in Section 5.3.2. Following the analysis of Section 5.4.1, we incorporate S_1 within the summary vector S_6 .

For datasets of size $n = 200$, we initially consider the performance of the bivariate α -stable model, $\mathcal{S}_\alpha(2, k, \mathbf{w}, \phi_{1:k}, \mathbf{0})$, for $k = 2$ and 3 spectral masses, with respect to parameter estimation and the impact of the number of projection vectors $\mathbf{t}_1, \dots, \mathbf{t}_\tau$.

The true and mean MMSE estimates of each parameter, placing projection vectors at the true locations of the spectral masses, are presented in Table 5.2. In addition, results are detailed using $\tau = 2, 5, 10$ and 20 randomly (uniformly) placed projection vectors, in order to evaluate the impact of spectral mass location uncertainty. The likelihood-free sampler output results in good MMSE parameter estimates, even for 2 randomly placed projection vectors. The parameter least accurately estimated is the location vector, $\boldsymbol{\mu}^0$. Directly summarized by a sample mean in $S_6(\cdot)$, estimation of location requires a large number of observations when the data have heavy tails.

Figure 5.6.2 illustrates progressive sampler performance for the $\mathcal{S}_\alpha(2, 2, \mathbf{w}, \phi_{1:2}, \mathbf{0})$ model, with $\alpha = 1.7$, $\mathbf{w} = (\pi/4, \pi)$ and $\phi_{1:2} = (\pi/4, \pi)$. Each circular scatter plot presents MMSE estimates of weight (radius) and angles (angle) of the two spectral masses, based on 10 sampler replicates. The sequence of plots (a)–(d) illustrates the progression of the parameter estimates as the scale parameter ϵ (of $\pi_\epsilon(y|x, \theta)$) decreases (and hence the accuracy of the likelihood-free approximation $\pi_{LF}(\theta|y) \approx \pi(\theta|y)$ improves). As ϵ decreases, there is a clear movement of the MMSE estimates towards the true angles and weights, indicating appropriate sampler performance.

With simulated datasets of size $n = 400$, we extend the previous bivariate study to 3 dimensions, with $k = 2$ discrete spectral masses. The true parameter values, and posterior mean MMSE estimates and associated standard errors, based on 10 sampler replicates, are presented

in Table 5.3. Again, reasonable parameter estimates are obtained (given finite data), with location (μ^0) again the most difficult to estimate.

In analogy with Figure 5.6.2, progressive sample performance for the first spectral mass (with $w_1 = 0.7$ and $\phi_1 = (\pi/4, \pi)$) for decreasing scale parameter ϵ is shown in Figure 5.6.3. Based on 200 replicate MMSE estimates (for visualization purposes), the shading of each point indicates the value of w_1 as a percentage (black=0%, white=100%), and the location on the sphere represents the angles ϕ_1 . For large ϵ , the MMSE estimates for location are uniformly distributed over the sphere, and the associated weight takes the full range of possible values, 0 – 100%. As ϵ decreases, the estimates of spectral mass location and weight become strongly localized and centered on the true parameter values, again indicating appropriate sampler performance. Similar images are produced for the second discrete spectral mass.

5.5 Analysis of exchange rate daily returns

Our data consist of daily exchange rates for 5 different currencies recorded in GBP between 1 January 2005 and 1 December 2007. The data involve 1065 daily-averaged LIBOR (London interbank offered rate) observations y'_1, \dots, y'_{1065} . The standard log-transform generates a log returns series $y_t = \ln(y'_{t+1}/y'_t)$. cursory examination of each returns series reveals clear non-Gaussian tails and/or skewness (Table 5.4, bottom).

We initially model each currency series as independent draws from a univariate α -stable distribution. Posterior MMSE parameter estimates for each currency are given in Table 5.4, based on 10 replicate likelihood-free samplers using the S_1 summary vector. For comparison, we also compute McCulloch's sample quantile based estimates (derived from S_1 , c.f. Section 5.3.1), and maximum likelihood estimates using J. P. Nolan's *STABLE* program (available online), using the direct search SPDF option with search domains given by $\alpha \in (0.4, 2]$, $\beta \in [-1, 1]$, $\gamma \in [0.00001, 1]$ and $\delta \in [-1, 1]$. Overall, there is good agreement between Bayesian, likelihood- and sample-based estimators. All currency returns distributions are significantly different from Gaussian ($\alpha = 2, \beta = 0$), and exhibit similar family parameter (α) estimates over this time period. However, the GBP to YEN conversion demonstrates a significant asymmetry (β) compared to the other currencies.

An interesting difference between the methods of estimation, is that McCulloch's estimates of α differ considerably from the posterior MMSE estimates, even though the latter are constructed using McCulloch's estimates directly as summary statistics, S_1 . One reason that the Bayesian estimates are more in line with the MLE's, is that likelihood-free methods largely ignore bias in estimators used as summary statistics (comparing the closeness between biased or unbiased estimators will produce similar results – consider comparing sample and maximum likelihood estimators of variance).

The multivariate α -stable distribution assumes that its marginal distributions, which are also α -stable, possess identical shape parameters. This property implies important practical limita-

tions, one of which is that it is only sensible to jointly model data with similar marginal shape parameters. Accordingly, based on Table 5.4, we now consider a bivariate analysis of AUD and EURO currencies. Restricting the analysis to the bivariate setting also permits comparison with the bivariate frequentist approach described in 32 using the *MVSTABLE* software (available online).

A summary illustration of the discrete approximations to the underlying continuous spectral mass is shown in Figure 5.6.4. Assuming $k = 3$ discrete spectral masses and based on 10 likelihood-free sampler replicates, the mean MMSE posterior estimates (solid black line) with mean 3σ posterior credibility intervals (dotted line), identify regions of high spectral mass located at 2.7, 3.9 and 5.6, with respective weights 0.45, 0.2 and 0.35. Broken lines in Figure 5.6.4 denote the frequentist estimates of 32, based on the identification of mass over an exhaustive mesh grid using 40 (dashed line) and 80 (dash-dot line) prespecified grid locations (projections).

Overall, both approaches produce comparable summary estimates of the spectral mass approximation, although the likelihood-free models generate full posterior distributions, compared to Nolan's frequentist estimates. The assumption of $k = 3$ discrete spectral masses provides a parsimonious representation of the actual spectral mass. For example, the spectral mass located at 2.7 accounts for the first two/three masses based on Nolan's estimates (80/40 projections). While the frequentist approach is computationally restricted to bivariate inference, the likelihood-free approach may naturally be applied in much higher dimensions.

5.6 Discussion

Statistical inference for α -stable models is challenging due to the computational intractability of the density function. In practice this limits the range of models fitted, to univariate and bivariate cases. By adopting likelihood-free Bayesian methods we are able to circumvent this difficulty, and provide approximate, but credible posterior inference in the general multivariate case, at a moderate computational cost. Critical to this approach is the availability of informative summary statistics for the parameters. We have shown that multivariate projections of data onto the unit hypersphere, in combination with sample quantile estimators, are adequate for this task.

Overall, our approach has a number of advantages over existing methods. There is far greater sampler consistency than alternative samplers, such as the auxiliary Gibbs or MCMC inversion plus series expansion samplers (6; 22). It is largely independent of the complexities of the various parameterizations of the α -stable characteristic function. The likelihood-free approach is conceptually straightforward, and scales simply and is easily implemented in higher dimensions (at a higher computational cost). Lastly, by permitting a full Bayesian multivariate analysis, the component locations and weights of a discrete approximation to the underlying continuous spectral density are allowed to identify those regions with highest posterior density in a parsimonious manner. This is a considerable advantage over highly computational frequentist approaches, which require explicit calculation of the spectral mass over a determin-

istic and exhaustive grid (e.g. 32).

Each analysis in this article used many millions of data-generations from the model. While computation for likelihood-free methods increases with model dimension and desired accuracy of the model approximation (through ϵ), much of this can be offset through parallelization of the likelihood-free sampler (36).

Finally, while we have largely focused on fitting α -stable models in the likelihood-free framework, extensions to model selection through Bayes factors or model averaging are immediate. One obvious candidate in this setting is the unknown number of discrete spectral masses, k , in the approximation to the continuous spectral density.

Acknowledgments

YF and SAS are supported by the Australian Research Council Discovery Project scheme (DP0877432 & DP1092805). GWP is supported by an APA scholarship and by the School of Mathematics and Statistics, UNSW and CSIRO CMIS. We thank M. Lombardi for generously providing the use of his code (22), and M. Briers, S. Godsill, Xiaolin Lou, P. Shevchenko and R. Wolpert, for thoughtful discussions.

Appendix A

SMC sampler PRC-ABC algorithm (36)

Initialization: Set $t = 1$ and specify tolerance schedule $\epsilon_1, \dots, \epsilon_T$.

For $i = 1, \dots, N$, sample $\theta_1^{(i)} \sim \pi(\theta)$, and set weights $W_1(\theta_1^{(i)}) = \pi_{LF,1}(\theta_1^{(i)}|y)/\pi(\theta_1^{(i)})$.

Resample: Resample N particles with respect to $W_t(\theta_t^{(i)})$ and set $W_t(\theta_t^{(i)}) = \frac{1}{N}$,
 $i = 1, \dots, N$.

Mutation and correction: Set $t = t + 1$ and $i = 1$:

- (a) Sample $\theta_t^{(i)} \sim M_t(\theta_t)$ and set weight for $\theta_t^{(i)}$ to

$$W_t(\theta_t^{(i)}) = \pi_{LF,t}(\theta_t^{(i)}|y)/M_t(\theta_t^{(i)}).$$
- (b) With probability $1 - p^{(i)} = 1 - \min\{1, W_t(\theta_t^{(i)})/c_t\}$, reject $\theta_t^{(i)}$ and go to (a).
- (c) Otherwise, accept $\theta_t^{(i)}$ and set $W_t(\theta_t^{(i)}) = W_t(\theta_t^{(i)})/p^{(i)}$.
- (d) Increment $i = i + 1$. If $i \leq N$, go to (a).
- (e) If $t < T$ then go to Resample.

This algorithm samples N weighted *particles* from a sequence of distributions $\pi_{LF,t}(\theta|y)$ given by (5.2.2), where t indexes a sequence of scale parameters $\epsilon_1 \geq \dots \geq \epsilon_T$. The final particles $\{(W_T(\theta_T^{(i)}), \theta_T^{(i)}) : i = 1, \dots, N\}$, form a weighted sample from the target $\pi_{LF,T}(\theta|y)$ (e.g. 36). The densities $\pi_{LF,t}(\theta|y)$ are estimated through the Monte Carlo estimate of the expectation (5.2.2) based on P draws $x^1, \dots, x^P \sim \pi(x|\theta)$.

For the simulations presented we implement the following specifications: for univariate α -stable models $\theta = (\alpha, \beta, \gamma, \delta)$ and for multivariate models $\theta = (\mathbf{w}, \phi_{1:k}, \boldsymbol{\mu}^0, \alpha)$; we use $N = 1000$ particles, initialized with samples from the prior; the function $\pi_\epsilon(y|x, \theta)$ is defined by $S(y) \sim N(S(x), \epsilon^2 \hat{\Sigma})$ where $\hat{\Sigma}$ is an estimate of $\text{Cov}(S(x)|\hat{\theta})$ based on 1000 draws $x^1, \dots, x^{1000} \sim \pi(x|\hat{\theta})$ given an approximate maximum likelihood estimate $\hat{\theta}$ of θ (17); the mutation kernel $M_t(\theta_t) = \sum_{i=1}^N W_{t-1}^{(i)}(\theta_{t-1}^{(i)})\phi(\theta_t; \theta_{t-1}, \Lambda)$ is a density estimate of the previous particle population $\{(W_{t-1}^{(i)}(\theta_{t-1}^{(i)}), \theta_{t-1}^{(i)}) : i = 1, \dots, N\}$, with a Gaussian kernel density ϕ with covariance Λ ; for univariate α -stable models $\Lambda = \text{diag}(0.25, 0.25, 1, 1)$, and for multivariate models $\Lambda = \text{diag}(1, \dots, 1, 0.25)$ (with Dirichlet proposals and kernel density substituted for \mathbf{w}); the sampler particle rejection threshold is adaptively determined as the 90th quantile of the weights $c_t = \hat{q}_{0.9}(\{W_t^{(i)}(\theta_t^{(i)})\})$, where $\{W_t^{(i)}(\theta_t^{(i)})\}$ are the N particle weights *prior* to particle rejection (steps (b) and (c)) at each sampler stage t (see 36).

For each analysis we implement 10 independent samplers (in order to monitor algorithm performance and Monte Carlo variability), each with the deterministic scale parameter sequence: $\epsilon_t \in \{1000, 900, \dots, 200, 100, 99, \dots, 11, 10, 9.5, 9, \dots, 5.5, 5, 4.95, \dots, 3.05, 3, 2.99, 2.98, \dots, 0.01, 0\}$. However, we adaptively terminate all samplers at the largest ϵ value such that the effective sample size (estimated by $[\sum_{i=1}^N [W_t^{(i)}(\theta_t^{(i)})]^2]^{-1}$) consistently drops below $0.2N$ over all replicate sampler implementations.

Appendix B: Data generation

Simulation of univariate α -stable data (12, 9)

1. Sample $W \sim \text{Exp}(1)$ to obtain w
2. Sample $U \sim \text{Uniform}[-\pi/2, \pi/2]$ to obtain u
3. Apply transformation to obtain sample \bar{y}

$$\bar{y} = \begin{cases} S_{\alpha,\beta} \frac{\sin \alpha(u+B_{\alpha,\beta})}{(\cos u)^{\alpha/2}} \left[\frac{\cos(u-\alpha(u+B_{\alpha,\beta}))}{w} \right]^{\frac{1-\alpha}{\alpha}} & \text{if } \alpha \neq 1 \\ \frac{2}{\pi} \left[\left(\frac{\pi}{2} + \beta u \right) \tan u - \beta \ln \frac{\frac{\pi}{2} w_i \cos u}{\frac{\pi}{2} + \beta u} \right] & \text{if } \alpha = 1 \end{cases}$$

with $S_{\alpha,\beta} = (1 + \beta^2 \tan^2(\frac{\pi\alpha}{2}))^{-1/2\alpha}$ and $B_{\alpha,\beta} = \frac{1}{\alpha} \arctan(\beta \tan(\frac{\pi\alpha}{2}))$. In this case \bar{y} will have distribution defined by $\Phi_X(t)$ with parameters $(\alpha, \beta, 1, 0)$.

4. Apply transformation to obtain sample $y = \gamma\bar{y} + \delta$ with parameters $(\alpha, \beta, \gamma, \delta)$.

Simulation of d -dimensional, multivariate α -stable data (33)

1. Generate Z_1, \dots, Z_k i.i.d. random variables from the univariate α -stable distribution with parameters $(\alpha, \beta, \gamma, \delta) = (\alpha, 1, 1, 0)$.
2. Apply the transformation

$$Y = \begin{cases} \sum_{j=1}^k w_j^{1/\alpha} Z_j \mathbf{s}_j + \boldsymbol{\mu}^0 & \alpha \neq 1 \\ \sum_{j=1}^k w_j \left(Z_j + \frac{2}{\pi} \ln(w_j) \right) \mathbf{s}_j + \boldsymbol{\mu}^0 & \alpha = 1 \end{cases}$$

with $\mathbf{s}_1, \dots, \mathbf{s}_k, \boldsymbol{\mu}^0 \in \mathbb{S}^d$. Note that while the complexity for generating realizations from a multivariate α -stable distribution is linear in the number of point masses (k) in the spectral representation per realization, this method is strictly only exact for discrete spectral measures.

| | Buckle | Lombardi | S_1 | S_2 | S_3 | S_4 | S_5 |
|-----------------|--------------|-------------|--------------|-------------|--------------|--------------|--------------|
| α (1.7) | 1.77 (0.18) | 1.62 (0.10) | 1.69 (0.06) | 1.65 (0.07) | 1.70 (0.06) | 1.71 (0.04) | 1.56 (0.05) |
| β (0.9) | 0.54 (0.21) | 0.86 (0.18) | 0.86 (0.10) | 0.65 (0.13) | 0.31 (0.09) | 0.38 (0.12) | 0.49 (0.11) |
| γ (10.0) | 18.17 (6.19) | 9.59 (2.16) | 9.79 (0.21) | 10.44(0.56) | 38.89 (6.34) | 39.12 (5.92) | 9.34 (0.14) |
| δ (10.0) | 12.30 (4.12) | 9.70 (2.19) | 10.64 (0.83) | 9.31 (0.86) | 10.25 (0.98) | 10.83 (1.34) | 11.18 (1.05) |

Tab. 5.1: Means and standard errors (in parentheses) of posterior MMSE estimates of α, β, γ and δ under the univariate α -stable model, based on 10 sampler replicates. Parameter values used for data simulation are given in the left column. Comparisons are between the auxiliary variable Gibbs sampler method of Buckle (1995), the inversion MCMC method of Lombardi (2007), and the likelihood-free method, using summary statistics S_1 – S_5 .

| | α | μ_1^0 | ϕ^1 | ϕ^2 | ϕ^3 | w_1 | w_2 | w_3 |
|--|----------------|---------------------|--------------------|--------------------|--------------------|-----------------|-----------------|-----------------|
| True: $k = 2$ | 1.7 | 0 | $\frac{\pi}{4}$ | π | – | 0.6 | 0.4 | – |
| True: $k = 3$ | 1.7 | 0 | $\frac{\pi}{4}$ | π | $\frac{3\pi}{2}$ | 0.3 | 0.25 | 0.45 |
| k | $\hat{\alpha}$ | $\widehat{\mu}_1^0$ | $\widehat{\phi}^1$ | $\widehat{\phi}^2$ | $\widehat{\phi}^3$ | \widehat{w}_1 | \widehat{w}_2 | \widehat{w}_3 |
| Projection vectors at locations of true spectral masses. | | | | | | | | |
| 2 | 1.66 (0.04) | 0.16 (0.19) | 0.81 (0.65) | 3.19 (0.46) | – | 0.55 (0.06) | 0.45 (0.05) | – |
| 3 | 1.79 (0.02) | 0.36 (0.18) | 0.84 (0.27) | 3.18 (0.29) | 4.91 (0.24) | 0.35 (0.05) | 0.25 (0.04) | 0.40 (0.05) |
| 2 projection vectors | | | | | | | | |
| 2 | 1.67 (0.06) | -0.13 (0.16) | 0.73 (0.55) | 3.58 (0.57) | – | 0.58 (0.09) | 0.42 (0.10) | – |
| 3 | 1.76 (0.05) | -0.16 (0.26) | 0.91 (0.66) | 3.65 (0.62) | 4.85 (0.55) | 0.36 (0.10) | 0.24 (0.09) | 0.40 (0.08) |
| 5 projection vectors | | | | | | | | |
| 2 | 1.71 (0.05) | 0.08 (0.14) | 0.71 (0.60) | 3.80 (0.67) | – | 0.60 (0.07) | 0.40 (0.09) | – |
| 3 | 1.75 (0.04) | 0.29 (0.17) | 0.86 (0.62) | 3.68 (0.52) | 4.82 (0.41) | 0.35 (0.09) | 0.20 (0.07) | 0.42 (0.09) |
| 10 projection vectors | | | | | | | | |
| 2 | 1.72 (0.02) | 0.21 (0.21) | 0.75 (0.32) | 3.31 (0.32) | – | 0.59 (0.05) | 0.41 (0.07) | – |
| 3 | 1.73 (0.03) | 0.25 (0.16) | 0.76 (0.44) | 3.31 (0.48) | 4.78 (0.19) | 0.34 (0.07) | 0.24 (0.04) | 0.42 (0.05) |
| 20 projection vectors | | | | | | | | |
| 2 | 1.71 (0.03) | -0.14 (0.14) | 0.76 (0.36) | 3.21 (0.23) | – | 0.63 (0.03) | 0.37 (0.03) | – |
| 3 | 1.72 (0.02) | 0.18 (0.23) | 0.77 (0.32) | 3.25 (0.31) | 4.75 (0.15) | 0.34 (0.04) | 0.23 (0.03) | 0.43 (0.03) |

Tab. 5.2: Mean MMSE parameter estimates (and standard errors) for the bivariate α -stable $\mathcal{S}_\alpha(2, k, \mathbf{w}, \phi_{1:k}, \mu^0)$ model, for $k = 2, 3$ discrete spectral masses, calculated over 10 replicate samplers. Projections vectors are placed at the true, and 2, 5, 10 and 20 randomly selected spectral mass locations. The true value of μ^0 is the origin.

| | α | μ_1^0 | ϕ_1^1 | ϕ_2^1 | ϕ_1^2 | ϕ_2^2 | w_1 | w_2 |
|-----------------------|----------------|---------------------|----------------------|----------------------|----------------------|----------------------|-----------------|-----------------|
| True: $k = 2$ | 1.7 | 0 | $\frac{\pi}{4}$ | π | $\frac{\pi}{2}$ | $\frac{3\pi}{2}$ | 0.3 | 0.7 |
| k | $\hat{\alpha}$ | $\widehat{\mu}_1^0$ | $\widehat{\phi}_1^1$ | $\widehat{\phi}_2^1$ | $\widehat{\phi}_1^2$ | $\widehat{\phi}_2^2$ | \widehat{w}_1 | \widehat{w}_2 |
| 20 projection vectors | | | | | | | | |
| 2 | 1.71 (0.02) | 0.53 (0.89) | 1.12 (0.34) | 3.81 (0.45) | 1.84 (0.54) | 4.24 (0.69) | 0.28 (0.06) | 0.72 (0.05) |

Tab. 5.3: Mean MMSE parameter estimates (and standard errors) for the trivariate α -stable $\mathcal{S}_\alpha(3, 2, \mathbf{w}, \phi_{1:2}, \mu^0)$ model, with $k = 2$ discrete spectral masses, calculated over 10 replicate samplers. The true value of μ^0 is the origin.

| | | Currency Exchange from GBP to | | | | |
|--------------------------------|----------------|-------------------------------|----------------|---------------|--------------|--------------|
| | | AUD | CNY | EURO | YEN | USD |
| Likelihood free | $\hat{\alpha}$ | 1.56 (0.03) | 1.57 (0.02) | 1.62 (0.04) | 1.51 (0.04) | 1.53 (0.02) |
| | $\hat{\beta}$ | 0.06 (0.03) | 0.01 (0.009) | -0.007 (0.08) | -0.26 (0.09) | -0.04 (0.03) |
| | $\hat{\gamma}$ | 0.004 (4e-4) | 0.003 (2e-4) | 0.004 (1e-4) | 0.003 (1e-4) | 0.004 (3e-4) |
| | $\hat{\delta}$ | 0.02 (0.01) | 0.001 (0.0006) | -0.03 (0.09) | -0.06 (0.08) | -0.02 (0.07) |
| MLE | $\hat{\alpha}$ | 1.61 (0.05) | 1.50 (0.05) | 1.65 (0.05) | 1.66 (0.04) | 1.57 (0.05) |
| | $\hat{\beta}$ | 0.08 (0.11) | -0.01 (0.10) | -0.10 (0.12) | -0.46 (0.11) | -0.01 (0.11) |
| | $\hat{\gamma}$ | 0.002 (7e-5) | 0.002 (6e-5) | 0.001 (4e-5) | 0.002 (4e-5) | 0.002 (1e-4) |
| | $\hat{\delta}$ | -2e-4 (1e-4) | -2e-5 (1e-4) | 8e-5 (7e-5) | 6e-4 (1e-4) | 5e-5 (9e-5) |
| McCulloch's quantile estimates | $\hat{\alpha}$ | 1.39 | 1.38 | 1.47 | 1.38 | 1.39 |
| | $\hat{\beta}$ | 0.08 | -0.003 | -0.04 | -0.18 | 0.001 |
| | $\hat{\gamma}$ | 0.002 | 0.002 | 0.001 | 0.002 | 0.002 |
| | $\hat{\delta}$ | -4e-5 | 1e-6 | 1e-5 | 2e-4 | 5e-7 |
| Kurtosis | | 8.39 | 9.11 | 15.60 | 6.29 | 4.98 |
| Skewness | | 0.69 | -0.42 | -0.03 | -0.79 | 0.11 |
| Std. dev. | | 0.004 | 0.004 | 0.003 | 0.004 | 0.003 |
| Mean | | -4e-5 | -4e-5 | -9e-6 | 1e-4 | 7e-5 |

Tab. 5.4: Posterior MMSE estimates (Monte Carlo errors) from the likelihood-free model, and maximum likelihood estimates (standard deviation). MLE's, parameter estimates using McCulloch's quantile (McCulloch, 1998), and sample statistics (mean, standard deviation, skewness and kurtosis) obtained from J. P. Nolan's *STABLE* software, available at academic2.american.edu/~jpnolan.

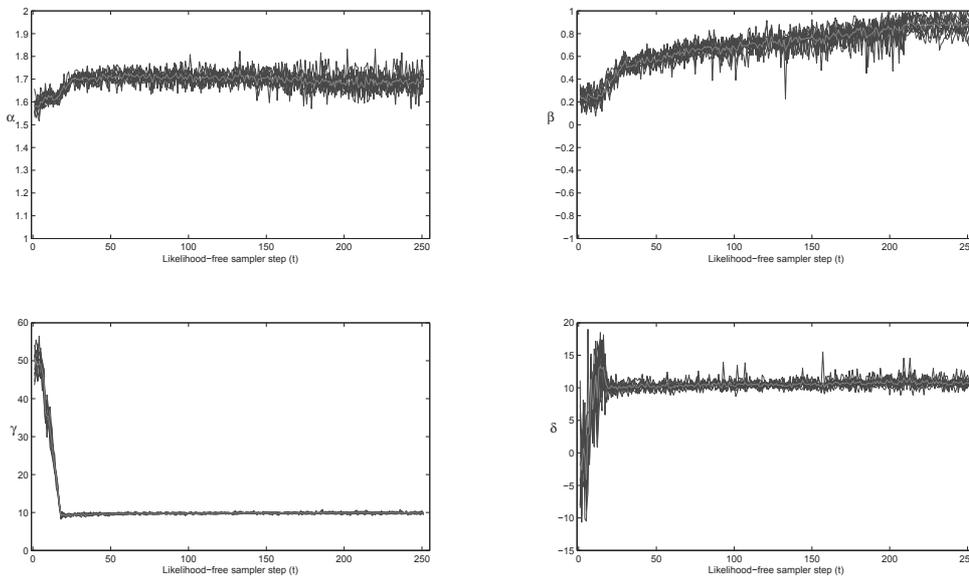


Fig. 5.6.1: Traces of posterior MMSE estimates of α , β , γ and δ under the univariate α -stable model and likelihood-free sampler (summary statistics S_1), based on 10 sampler replicates. Traces are shown as a function (x -axis) of sampler progression (t) and scale parameter reduction $\epsilon_t < \epsilon_{t-1}$. Parameter values used for data generation are $\alpha = 1.7$, $\beta = 0.9$, $\gamma = 10$ and $\delta = 10$.

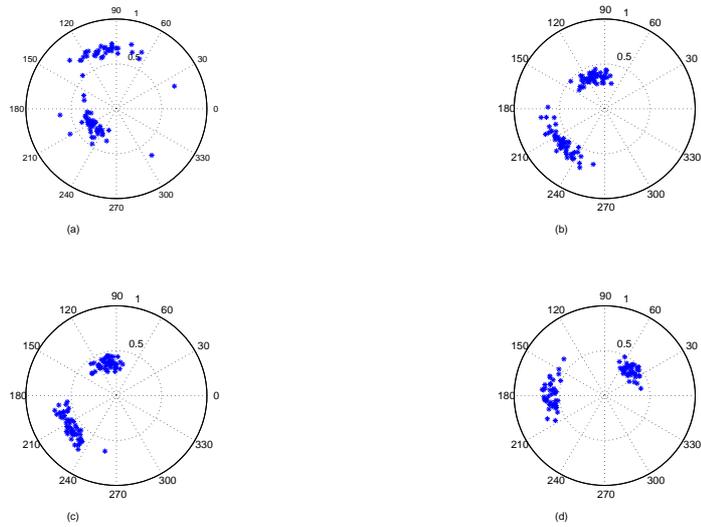


Fig. 5.6.2: Circular scatter plot of MMSE estimates for $k = 2$ spectral mass angles (angle) and weights (radius) for bivariate α -stable $S_\alpha(2, \mathbf{w}, \phi_{1:2}, \mathbf{0})$ model, with $\alpha = 1.7$, $\mathbf{w} = (0.4, 0.6)$ and $\phi_{1:2} = (\pi/4, \pi)$. Plots (a)–(d) demonstrate evolution of the estimates for decreasing scale parameter values ϵ , based on 10 sampler replicates.

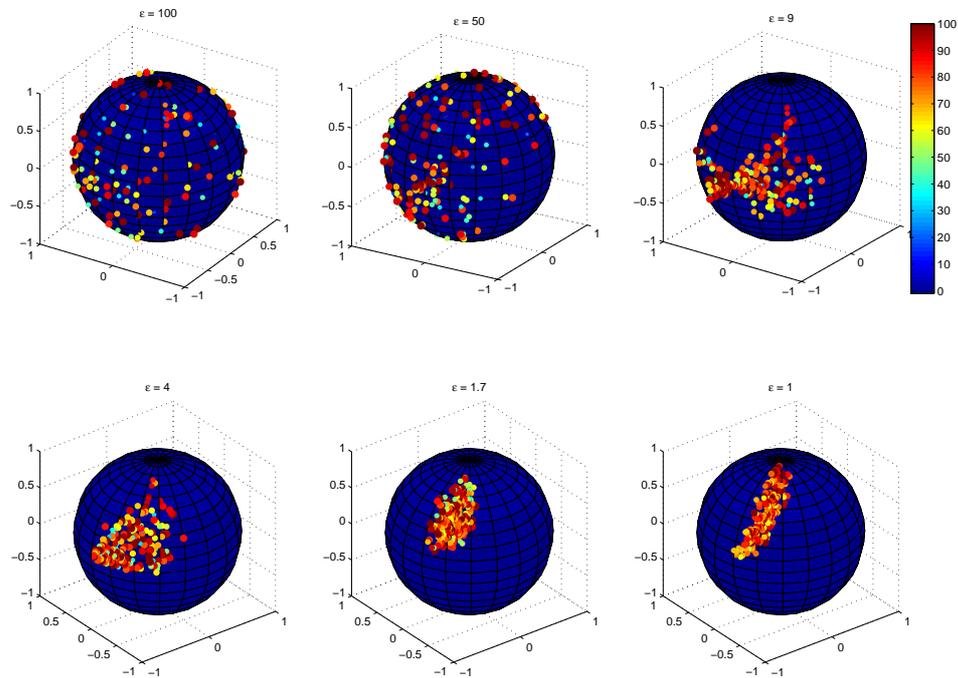


Fig. 5.6.3: Spherical heat map of MMSE estimates for the first of $k = 2$ discrete spectral masses, ϕ_1 , for the trivariate α -stable $S_\alpha(3, 2, \mathbf{w}, \phi_{1:2}, \mathbf{0})$ model. True values of the first spectral mass are $w_1 = 0.7$ (70%) and $\phi_1 = (\pi/4, \pi)$. Point shading indicates MMSE value of w_1 as a percentage. The plots demonstrate the evolution of the estimates for decreasing scale parameter values ϵ , based on 200 sampler replicates.

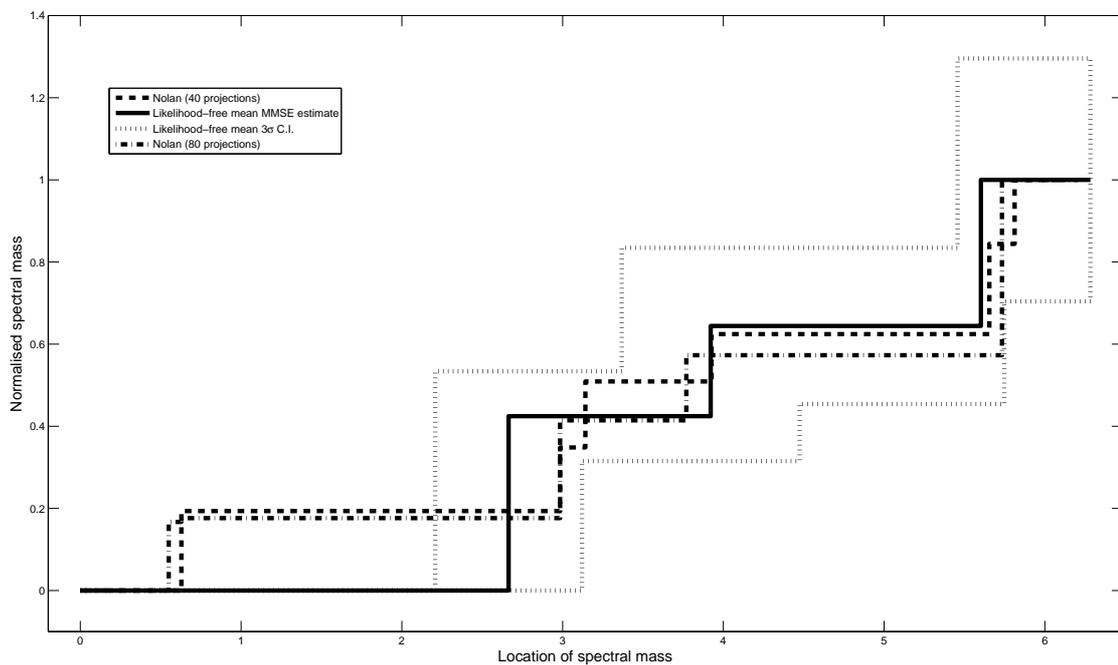


Fig. 5.6.4: Estimates of spectral mass location (x-axis) and cumulative weight (y-axis) for AUD and EURO currencies data. Solid line denotes mean posterior MMSE estimates of likelihood-free SMC sampler output, and dotted line illustrates mean 3σ posterior credibility intervals, based on $k = 3$ discrete spectral masses, 20 randomly placed projection vectors and 10 replicate samplers. Broken lines denote estimate of spectral mass using J. P. Nolan's *MVSTABLE* software, available at academic2.american.edu/~jpnolan, with (dashed line) 40 deterministic projection locations and (dash-dot line) 80 projection locations.

References

- [1] Alder, R., R. Feldman, and M. S. Taqqu (1998). *A practical guide to heavy-tails: Statistical techniques for analysing heavy-tailed distributions*. Birkhäuser.
- [2] Beaumont, M. A., W. Zhang, and D. J. Balding (2002). Approximate Bayesian computation in population genetics. *Genetics* 162, 2025 – 2035.
- [3] Bergstrom, H. (1953). On some expansions of stable distributional functions. *Ark. Mat.* 2, 375–378.
- [4] Blum, M. G. B. (2009). Approximate Bayesian computation: a non-parametric perspective. Technical report, Université Joseph Fourier, Grenoble, France.
- [5] Brooks, S. P., A. Gelman, G. Jones, and X.-L. Meng (2010). *Handbook of Markov chain Monte Carlo*. Chapman and Hall/CRC.
- [6] Buckle, D. J. (1995). Bayesian inference for stable distributions. *Journal of the American Statistical Association* 90, 605–613.
- [7] Byczkowski, T., J. P. Nolan, and B. Rajput (1993). Approximation of multidimensional stable densities. *J. Multiv. Anal.* 46, 13–31.
- [8] Casarin, R. (2004). Bayesian inference for mixtures of stable distributions. *Working paper No. 0428, CEREMADE, University Paris IX*.
- [9] Chambers, J., C. Mallows, and B. Stuck (1976). A method for simulating stable random variables. *J. Am. Stat. Assoc.* 71, 340–334. Correction (1987), 82, 704.
- [10] Devroye, L. (1986). An automatic method for generating random variates with a given characteristic function. *SIAM Journal on Applied Maths* 46, 698–719.
- [11] Doganoglu, T. and S. Mittnik (1998). An approximation procedure for asymmetric stable paretian densities. *Computational Statistics* 13, 463–475.
- [12] DuMouchel, W. (1975). Stable distributions in statistical inference: Information from stably distributed samples. *J. Am. Stat. Assoc.* 70, 386–393.
- [13] Fama, E. (1965). The behaviour of stock market prices. *J. Business* 38, 34–105.
- [14] Fama, E. and R. Roll (1968). Some properties of symmetric stable distributions. *Journal of the American Statistical Association* 63, 817–83.

- [15] Godsill, S. (1999). MCMC and EM-based methods for inference in heavy-tailed processes with alpha stable innovations. In *Proc. IEEE Signal Processing Workshop on Higher Order Statistics*.
- [16] Godsill, S. (2000). Inference in symmetric alpha-stable noise using MCMC and the slice sampler. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Volume VI*, pp. 3806–3809.
- [17] Jiang, W. and B. Turnbull (2004). The indirect method: Inference based on intermediate statistics – A synthesis and examples. *Statistical Science* 19, 238–263.
- [18] Koutrouvelis, I. (1980). Regression type estimation of the parameters of stable laws. *Journal of the American Statistical Association* 75, 918–928.
- [19] Kuruoglu, E. E., C. Molina, S. J. Godsill, and W. J. Fitzgerald (1997). A new analytic representation for the alpha stable probability density function. In H. Bozdogan and R. Soyer (Eds.), *AMS Proceedings, Section on Bayesian Statistics*.
- [20] Levy, P. (1924). Theorie des erreurs. La loi de Gauss et les lois exceptionnelles. *Bulletin-Societe Mathematique de France* 52, 49–85.
- [21] Lombardi, M. and S. Godsill (2006). On-line Bayesian estimation of AR signals in symmetric alpha-stable noise. *IEEE Transactions on Signal Processing*.
- [22] Lombardi, M. J. (2007). Bayesian inference for alpha stable distributions: A random walk MCMC approach. *Comp. Statist. and Data Anal.* 51, 2688–2700.
- [23] Mandelbrot, B. (1960). The Pareto-Levey law and the distribution of income. *International Economic Review* 1, 79–106.
- [24] Marjoram, P., J. Molitor, V. Plagnol, and S. Tavaré (2003). Markov chain Monte Carlo without likelihoods. *Proc. Nat. Acad. Sci. USA* 100, 15324–15328.
- [25] McCulloch, J. H. (1986). Simple consistent estimators of stable distribution parameters. *Comm. Stat. Simulation and computation.* 15, 1109–1136.
- [26] McCulloch, J. H. (1998). Numerical approximation of the symmetric stable distribution and density. In E. Adler, R. R. Feldman, and T. M. Birkhauser (Eds.), *A practical guide to heavy tails: statistical techniques and applications*.
- [27] McKinley, T., A. R. Cook, and R. Deardon (2009). Inference in epidemic models without likelihoods. *The International Journal of Biostatistics* 5: article 24.
- [28] Melchiori, M. R. (2006). Tools for sampling multivariate archimedean copulas. www.YieldCurve.com.
- [29] Mittnik, S. and S. T. Rachev (1991). Alternative multivariate stable distributions and their applications to financial modelling. In E. Cambanis, G. Samorodnitsky, and M. S. Taqqu (Eds.), *Stable processes and related topics*, pp. 107–119. Birkhauser, Boston.
- [30] Neal, R. (2003). Slice sampling. *Annals of Statistics* 31, 705–767.
- [31] Nikias, C. and M. Shao (1995). *Signal processing with alpha stable distributions and applications*. Wiley, New York.

- [32] Nolan, J. P. (1997). Numerical computation of stable densities and distributions. *Comm. Statisti. Stochastic Models* 13, 759–774.
- [33] Nolan, J. P. (2007). *Stable distributions: Models for heavy-tailed data*. Birkhäuser.
- [34] Nolan, J. P. (2008). Stable distributions: Models for heavy-tailed data. Technical report, Math/Stat Department, American University.
- [35] Nolan, J. P., A. K. Panorska, and McCulloch (2001). Estimation of stable spectral measures, stable non-Gaussian models in finance and econometrics. *Math. Comput. Modelling* 34, 1113–1122.
- [36] Peters, G. W., Y. Fan, and S. A. Sisson (2008). On sequential Monte Carlo, partial rejection control and approximate Bayesian computation. Tech. rep. UNSW.
- [37] Pievatolo, A. and P. Green (1998). Boundary detection through dynamic polygons. *Journal of the Royal Statistical Society, B* 60(3), 609–626.
- [38] Press, S. (1972). Estimation in univariate and multivariate stable distributions. *Journal of the Americal Statistical Association* 67, 842–846.
- [39] Ratmann, O., C. Andrieu, T. Hinkley, C. Wiuf, and S. Richardson (2009). Model criticism based on likelihood-free inference, with an application to protein network evolution. *Proc. Natl. Acad. Sci. USA* 106, 10576–10581.
- [40] Reeves, R. W. and A. N. Pettitt (2005). A theoretical framework for approximate Bayesian computation. In A. R. Francis, K. M. Matawie, A. Oshlack, and G. K. Smyth (Eds.), *Proc. 20th Int. Works. Stat. Mod., Australia, 2005*, pp. 393–396.
- [41] Salas-Gonzalez, D., E. E. Kuruoglu, and D. P. Ruiz (2006). Estimation of mixtures of symmetric alpha-stable distributions with an unknown number of components. *IEEE Int. Conf. on Acoustics, Speech and Sig. Proc. Toulouse, France*.
- [42] Samorodnitsky, G. and M. S. Taqqu (1994). *Stable non-Gaussian random processes: Stochastic models with infinite variance*. Chapman and Hall/CRC.
- [43] Sisson, S. A., Y. Fan, and M. M. Tanaka (2007). Sequential Monte Carlo without likelihoods. *Proc. Nat. Acad. Sci.* 104, 1760–1765. Errata (2009), 106, 16889.
- [44] Sisson, S. A., G. W. Peters, M. Briers, and Y. Fan (2009). Likelihood-free samplers. Technical report, University of New South Wales.
- [45] Tavaré, S., D. J. Balding, R. C. Griffiths, and P. Donnelly (1997). Inferring coalescence times from DNA sequence data. *Genetics* 145, 505–518.
- [46] Weron, R. (2006). *Modeling and forecasting electricity loads and prices: A statistical approach*. Wiley.
- [47] Zolotarev, V. M. (1986). *One-Dimensional Stable Distributions*. Translations of Mathematical Monographs. American Mathematical Society.

6

Journal Paper 4

"The essence of mathematics is not to make simple things complicated, but to make complicated things simple"

Stanly Gudder

Fan Y., Peters G.W., Sisson S.A. (2009) "Automating and Evaluating Reversible Jump MCMC Proposal Distributions". *Statistics and Computing*, 19, 401-429.

This work was instigated by Scott Sisson and Yanan Fan. The second author can claim around 70% of the credit for the contents. His work included developing large amounts of the methodology contained, developing the applications, implementation and comparison to alternative approaches, writing of significant sections of the drafts of the paper and revisions. This paper has been accepted and is already cited by 3 other papers and the ideas contained have been successfully developed and use in Electrical Engineering applications published. Permission from all the co-authors has been granted for submission of this paper as part of the thesis.

Automating and evaluating reversible jump MCMC proposal distributions

Y. Fan (*corresponding author*)

UNSW Mathematics and Statistics Department, Sydney, 2052, Australia

Gareth W. Peters

CSIRO Mathematical and Information Sciences, Locked Bag 17, North Ryde, NSW, 1670, Australia;

UNSW Mathematics and Statistics Department, Sydney, 2052, Australia

S. A. Sisson

UNSW Mathematics and Statistics Department, Sydney, 2052, Australia

6.1 Abstract

The reversible jump Markov chain Monte Carlo (MCMC) sampler (58) has become an invaluable device for Bayesian practitioners. However, the primary difficulty with the sampler lies with the efficient construction of transitions between competing models of possibly differing dimensionality and interpretation. We propose use of a marginal density estimator to construct between-model proposal distributions. This provides both a step towards black-box simulation for reversible jump samplers, and a tool to examine the utility of common between-model mapping strategies. We compare the performance of our approach to well established alternatives in both time series and mixture contexts.

Keywords: Between-model mappings; Density estimation; Markov chain Monte Carlo; Path sampling; Reversible jump; Trans-dimensional MCMC.

6.2 Introduction

Simultaneous inference on both model and parameter space is fundamental to modern statistical practice. In general, for an observed dataset $\mathbf{y} = (y_1, \dots, y_n)$ we might consider a countable set of models $\mathcal{M} = \{M_1, M_2, \dots\}$ indexed by a parameter $k \in \mathcal{K}$, each with a parameter vector $\boldsymbol{\theta}_k \in \Theta_k$ of length n_k . Under a Bayesian framework, inference proceeds through the joint posterior distribution of the model and model indicator pair $\pi(\boldsymbol{\theta}_k, M_k \mid \mathbf{y})$, typically in terms of model averaging (7; 19) or model selection (6; 13). Analytical computations on the posterior are generally unavailable, and so numerical approximation methods are commonly utilized.

The most accessible of these methods is the reversible jump sampler (58), a restatement of

the Metropolis-Hastings algorithm in terms of general state spaces. This permits exploration of joint parameter and model indicator space via a single Markov chain. Proposed between-model chain transitions from (θ_i, M_i) to (θ_j, M_j) are accepted with probability $\min\{1, \alpha_{ij}(\theta_i, \theta_j)\}$, where

$$\alpha_{ij}(\theta_i, \theta_j) = \frac{\pi(\theta_j, M_j | \mathbf{y}) r_{ji}(\theta_j) q_{ji}(\theta_j, \theta_i)}{\pi(\theta_i, M_i | \mathbf{y}) r_{ij}(\theta_i) q_{ij}(\theta_i, \theta_j)}. \quad (6.2.1)$$

Here, $r_{ij}(\theta_i)$ denotes the probability of proposing to move from model M_i to M_j , and $q_{ij}(\theta_i, \theta_j)$ is the density from which the proposed parameter θ_j is drawn.

The primary difficulty with the reversible jump sampler lies with the efficient construction of transitions between differing models. In general, moving between M_i and M_j is problematic as Θ_i and Θ_j may differ in dimensionality and interpretation, and so specifying the joint distributions $q_{ij}(\theta_i, \theta_j)$ can be difficult.

In practice the problem is considered in two stages. Firstly, a random vector $\mathbf{u}_i \sim g_i(\mathbf{u} | \phi_i)$ is drawn from a known density with parameters ϕ_i . The proposed new state $\theta_j = h_{ij}(\theta_i, \mathbf{u}_i)$ is then formed as a deterministic function h_{ij} of the current state θ_i and the random vector. Separation of the random innovation \mathbf{u}_i and the deterministic mapping h_{ij} permits great flexibility for the user in defining the between-model transition. If the reverse move from M_j to M_i is constructed in the same manner, and the between-model mapping satisfies $\theta_i = h_{ji}(h_{ij}(\theta_i, \mathbf{u}_i), \mathbf{u}_j)$, the move from (θ_i, M_i) to (θ_j, M_j) is accepted with probability $\min\{1, \alpha_{ij}(\theta_i, \theta_j)\}$, where

$$\alpha_{ij}(\theta_i, \theta_j) = \frac{\pi(\theta_j, M_j | \mathbf{y}) r_{ji}(\theta_j) g_j(\mathbf{u}_j | \phi_j)}{\pi(\theta_i, M_i | \mathbf{y}) r_{ij}(\theta_i) g_i(\mathbf{u}_i | \phi_i)} \left| \frac{\partial h_{ij}(\theta_i, \mathbf{u}_i)}{\partial(\theta_i, \mathbf{u}_i)} \right|. \quad (6.2.2)$$

Efficiency of the reversible jump sampler is highly dependent on the choice of proposal distribution q_{ij} in (6.2.1) or equivalently the joint specification of $g_i(\cdot | \phi_i)$ and mapping function h_{ij} in (6.2.2). In most applications obvious choices of mappings may be far from optimal. In many applications even identifying a potential mapping may be difficult.

Various approaches have been proposed to both automate, and improve the efficiency of between-model proposals. (15) and (18) developed a reversible jump analogy of the random walk Metropolis-Hastings sampler, based on multivariate Gaussian proposals for q_{ij} . The mean and covariance parameters of the proposals were estimated a priori by independent Metropolis-Hastings simulations within each model. The effectiveness of this sampler strongly depends on how closely each model $\pi(\theta_k, M_k | \mathbf{y})$ matches the the proposal, and the amount of computation required to estimate the proposal parameters. Given both $g_i(\cdot | \phi_i)$ and the mapping function h_{ij} (3) demonstrate how to choose the parameters ϕ_i which are likely to generate acceptance values close to one within a region by constraining derivatives of the acceptance probability. (3) alternatively propose augmenting the parameter state to $(\theta_k, M_k, \mathbf{u}_k)$, and then induce temporal dependence between the random vector components to retain a form of memory of previously accepted \mathbf{u}_i which may then be re-used in future between-model moves. Other approaches accept that the proposed between-model move to M_j is likely to be poor, and seek to improve upon the initial mapping e.g. by appending a number of within-model Metropolis-Hastings steps to the between-model move to better locate higher density areas (1);

or by attempting a modified between-model move, conditional upon the rejection of an initial proposal (17). Further details can be found in (22).

The above methods either presume knowledge of the proposal distribution $g_i(\cdot | \phi_i)$ and the mapping function h_{ij} , or simply propose generic moves using particular forms of q_{ij} with potentially limited success. In practice, between model-moves can hugely influence acceptance probabilities and sampler efficiency, while their specification is almost arbitrary.

In this article we propose the use of a convenient marginal path-sampling density estimator (10) to construct between-model proposal distributions. Accordingly we avoid both the difficulties in specifying both the mapping function, h_{ij} , and the associated proposal distribution, $g_i(\cdot | \phi_i)$ (and their separation), and the inflexibility of a fixed or parametric form for q_{ij} . While the resulting class of proposal distributions require moderate computation to construct, they provide further steps towards the goal of an automated generic reversible jump sampler. The constructed proposals also provide a tool with which to examine the relative efficiencies of existing between-model mapping strategies on a per application basis.

The structure of this article is as follows: In Section 6.3 we consider generic between-model move types and summarize the path-sampling density estimator. In Section 6.4 we provide details on the construction of densities in the setting of between-model mappings. In Section 6.5 we examine the performance of two reversible jump samplers specialized for their respective models, and evaluate the utility of their between-model proposal mechanisms. We conclude with a discussion.

6.3 Moving between models

6.3.1 Trans-dimensional move types

The most naturally conceived moves between models M_i and M_j of possibly differing dimension may be loosely classified as ‘global’ or ‘local.’ Global moves (e.g. 15; 18) tend to propose a vector θ_j which is independent of the current vector θ_i , but which uses some knowledge about M_i and/or M_j to locate regions of high density. In this setting, the proposal density may be either represented by a multivariate parametric family e.g. $q_{ij} \sim N(\mu_j, \Sigma_j)$ (15), or more generally via the one-sided conditionals

$$q_{ij}(\theta_i, \theta_j) = p_1(\theta_j^1 | \theta_j^2, \dots, \theta_j^{n_j}) \dots p_{n_j-1}(\theta_j^{n_j-1} | \theta_j^{n_j}) p_{n_j}(\theta_j^{n_j}), \quad (6.3.1)$$

where $\theta_j = (\theta_j^1, \theta_j^2, \dots, \theta_j^{n_j})$ and where p denotes the appropriate conditional or marginal density. In contrast, local moves exploit perceived structural similarities between models; for example, in a regression model where M_i contains a subset of the predictors of M_j . Here it is argued that a subset of θ_j can be held constant at the respective values in θ_i , with the remaining parameters sampled from the proposal. In this setting, assuming for clarity that $n_i \leq n_j$, we move from $\theta_i = (\theta_i^c, \theta_i^{-c})$ to $\theta_j = (\theta_j^c, \theta_j^{-c})$, where $\theta_i^c = \theta_j^c$ is the subvector of length $n_j - r$ held

constant during the proposed move, and θ_j^{-c} is the subvector of r parameters to be sampled. The proposal density now becomes

$$q_{ij}(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = p_1(\theta_j^1 | \theta_j^2, \dots, \theta_j^r, \boldsymbol{\theta}_j^c) \dots p_{r-1}(\theta_j^{r-1} | \theta_j^r, \boldsymbol{\theta}_j^c) p_r(\theta_j^r | \boldsymbol{\theta}_j^c), \quad (6.3.2)$$

where (6.3.2) equals (6.3.1) if $r = n_j$ and $\boldsymbol{\theta}_j^c$ is of dimension zero. Common local moves include birth/death and split/merge strategies (e.g. 21). Global moves can have the advantage of producing samplers with fast mixing properties, although this assumes that the proposals are adequate to describe the posterior of the target model M_j . Construction of the proposals may be computationally inefficient if the number of parameters is large. In comparison, local moves perform very low-dimensional modifications to the current state in order to obtain $\boldsymbol{\theta}_j$, incorporating perceived structural relationships between models through the functions h_{ij} . Accordingly such moves can be very inefficient given heavy reliance on the accuracy of the between-model mappings, and consequently produce chains that may mix more slowly than those with global moves.

In practice, due to computational feasibility, the vast majority of the analysis implements local moves. Draws from $q(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)$ given by (6.3.2) are usually implemented via the functions $g_i(\cdot | \phi_i)$ and the mapping function h_{ij} , as direct simulation via one-sided conditionals is generally infeasible. Potential mis-specification of the mapping function h_{ij} could be avoided by simulating from (6.3.2) directly, if the conditionals p_s for $s = 1, \dots, r$ could be approximated at moderate cost. Further, circumventing the need for mapping functions would be a step closer to automating reversible jump samplers. We now discuss a flexible marginal density estimator that may be used for this purpose.

6.3.2 Fan et al (2006)'s marginal density estimator

A number of parametric marginal density estimators based on MCMC output have been proposed (e.g. 11; 24; 5; 4). The density estimator of (10) is a simple and effective marginal estimator based on path-sampling (12) and gradients of the log-posterior. Unlike some density estimators, the use of the log-posterior provides some robustness against samples obtained from the wrong regions of the target density. Consequently, non-MCMC samples may also be used to estimate the density; see Section 6.4.3. In addition, this density estimator is easily adapted for the construction of conditional densities (c.f. (6.3.2)), which is crucial for sampler computational and mixing efficiency.

Suppose we have samples $\boldsymbol{\psi}^1, \dots, \boldsymbol{\psi}^n$ from a distribution $\pi(\boldsymbol{\psi})$ known up to some normalization constant, and we are interested in estimating the k -th marginal distribution, $\pi_k(\psi_k)$. If $\lambda(\psi_k) = \log \pi_k(\psi_k)$ denotes the log of the (unnormalized) marginal posterior distribution, (10) show that

$$\frac{\partial}{\partial \psi_k} \lambda(\psi_k) = \mathbb{E}_{\boldsymbol{\psi}_{-k}} [U_k(\boldsymbol{\psi})] \quad \text{and} \quad U_k(\boldsymbol{\psi}) = \frac{\partial}{\partial \psi_k} \log \pi(\boldsymbol{\psi}), \quad (6.3.3)$$

where $\mathbb{E}_{\boldsymbol{\psi}_{-k}}$ denotes an expectation under all elements of $\boldsymbol{\psi}$ excluding ψ_k . An estimator for

$\mathbb{E}_{\psi_{-k}}[U_k(\psi)]$ can be obtained by first ordering the samples from the k -th margin $\psi_k^{(1)} < \psi_k^{(2)} < \dots < \psi_k^{(m)}$, $m \leq n$, ignoring any duplicates e.g. which may arise in the context of the Metropolis algorithm. If $n^{(i)}$ is the number of replicates for each $\psi_k^{(i)}$, then the estimator $\bar{U}_k(i)$ is obtained as

$$\bar{U}_k(i) = n^{(i)^{-1}} \sum_{j=1}^{n^{(i)}} U_k(\psi_j). \quad (6.3.4)$$

The $\bar{U}_k(i)$ approximate the gradient of $\log \pi_k(\psi_k)$ at each of the points in the sample, and may be utilized to derive a density estimate. Of the density estimation methods presented by (10) we implement the simplest and most efficient approach, since interest is in an approximate distribution rather than a perfect density estimate. A stepwise linear approximation to $\pi_k(\psi_k)$ is obtained by arbitrarily setting $\hat{\lambda}(\psi_k^{(1)}) = 0$ and defining

$$\hat{\lambda}(\psi_k^{(i)}) = \hat{\lambda}(\psi_k^{(i-1)}) + \left(\psi_k^{(i)} - \psi_k^{(i-1)} \right) \times \left(\bar{U}_k(i) + \bar{U}_k(i-1) \right) / 2 \quad \text{for } i = 2, \dots, m. \quad (6.3.5)$$

Hence for all points $\psi \in [\psi_k^{(1)}, \psi_k^{(m)}]$, the unnormalized density estimate is given by

$$\hat{\pi}_k(\psi_k) = \exp[\hat{\lambda}(\psi_k^{(i)})] \quad \text{for } \psi \in [\psi_k^{(i)}, \psi_k^{(i+1)}] \quad (6.3.6)$$

and $\hat{\pi}_k(\psi_k) = 0$ otherwise. The corresponding distribution function is obtained by integrating over $\hat{\pi}_k(\psi_k)$, with the normalizing constant given by the value of the distribution function at the point $\psi_k^{(m)}$. Simulation from $\hat{\pi}_k(\psi_k)$ may proceed via inversion methods.

6.4 Automating between-model moves

It is trivial to use (10)'s estimator to construct a marginal density estimate which is conditioned upon some subset of the remaining parameters ψ_{-k} , e.g. $\pi_k(\psi_k \mid \psi_1, \dots, \psi_{k-1})$. This is achieved by fixing the appropriate sample margins at their conditioned values and estimating the density as before. As a result it becomes feasible to approximate a between-model proposal decomposition of the form (6.3.2). This may proceed by first estimating and sampling from the density p_r , then estimating and sampling from p_{r-1} conditioning upon the previously sampled point, and so on. We refer to this proposal as the conditional path sampling (CPS) proposal and now consider in more detail how this might be used in a between-model mapping.

6.4.1 Construction of proposal distribution

By construction, the density estimator (6.3.6) is truncated at the lower and upper endpoints of the sampled points. Where appropriate, we propose to append Gaussian tails to the CPS density estimator in order to cover the support of the parameter space. Specifically, suppose we are interested in sampling the k -th parameter, ψ_k , from some joint distribution $\pi(\psi)$. For

ordered samples $\psi_k^{(1)}, \dots, \psi_k^{(m)}$ (ignoring ties) we define

$$\hat{\pi}_k(\psi) = \begin{cases} \phi(\psi | \psi_k^{(1)}, \sigma_1) & \psi \leq \psi_k^{(1)} \\ \exp[\hat{\lambda}(\psi_k^{(i)})] & \psi \in (\psi_k^{(i)}, \psi_k^{(i+1)}] \text{ for } i = 1, \dots, m-1 \\ \phi(\psi | \psi_k^{(m)}, \sigma_2) & \psi \geq \psi_k^{(m)} \end{cases} \quad (6.4.1)$$

as the (unnormalized) proposal distribution for ψ_k , where $\hat{\lambda}(\psi_k^{(i)})$ is given by (6.3.5) and $\phi(\psi | \mu, \sigma)$ denotes the Normal density for ψ with mean μ and standard deviation σ . That is, the Gaussian distributions are centered at $\psi_k^{(1)}$ and $\psi_k^{(m)}$ respectively, with standard deviations defined by satisfying the continuity constraints

$$\sigma_1 = (2\pi)^{-1/2} \exp[-\hat{\lambda}(\psi_k^{(1)})] \quad \text{and} \quad \sigma_2 = (2\pi)^{-1/2} \exp[-\hat{\lambda}(\psi_k^{(m)})]$$

so that the Gaussian densities meet the CPS density estimates at $\psi_k^{(1)}$ and $\psi_k^{(m)}$. The tails may be truncated for bounded parameters. Normalization and simulation from this density is performed as before.

6.4.2 Use of partial derivatives

The construction of CPS proposals require the calculation of partial derivatives through the quantity U_k in (6.3.3). While derivation of analytic expressions for the gradient of $\pi(\psi_k)$ is possible in many settings, in general it may not be straightforward. In such circumstances one may adopt a simple difference approximation

$$U_k(\psi_k) \approx \frac{1}{\pi(\psi_k)} \frac{\pi(\psi_k + \Delta/2) - \pi(\psi_k - \Delta/2)}{\Delta}, \quad (6.4.2)$$

although sensitivity to the value of Δ should be established. We implement both analytical and difference approximation derivatives in the examples in Section 6.5.

6.4.3 Obtaining samples

As a density estimator, (10)'s estimator was shown to work well using MCMC samples from the joint posterior distribution. However, if non-posterior samples are utilized, where all samples are in low posterior density regions, a density in proportion to the true density will be constructed in this region. Similarly, if some samples overlap with high posterior density regions, (6.4.1) will produce a distribution with high density corresponding to these samples, and low density elsewhere. The latter case, arising through samples *approximately* from the high density regions, provides a relatively poor density estimate. However, through the robustness of the (10) estimator, these estimates will still be sufficiently accurate to implement as a good proposal in a Markov chain sampler, with higher proposal probabilities near samples with relatively higher posterior densities.

There are number of strategies available to obtain samples approximately from the posterior in order to construct between-model proposals. There is also an obvious efficiency trade-off between accuracy of the density estimate and computational cost. Perhaps the simplest option is the a priori drawing of samples from each model via pilot chains (15). This approach assumes adequate within-model sampler mixing and a practically small number of candidate models. It also requires a relatively large number of samples from each model, especially for complex posteriors, in order to have sufficient samples to estimate all conceivable conditional distributions. An alternative requiring far fewer, and lower-dimensional samples is to draw them approximately from each required conditional distribution as the move is attempted. This might be performed relatively quickly by propagating samples from the (conditional) prior to the (conditional) posterior via a few steps of simulated annealing. The disadvantage is that the simulation must be repeatedly performed for each between-model proposal, increasing computational overheads. A further alternative to pilot chains might be to appeal to the augmented state space approach of (3), and retain the (say) L most recently visited chain states in each model as part of the current state of the chain, and use these to construct the CPS proposal (with the first proposal to each model utilizing simulated annealing to obtain samples). However, this approach would likely prove unwieldy in terms of book-keeping.

When the dimension of the parameter vector is small with bounded support, or where information regarding the location/scale of the (conditional) posterior is available, one could place samples on grids. Another potentially quick and viable option would be to construct CPS proposals based on draws from the prior. If the priors are reasonably coincident with the posterior, while sampling directly from priors will produce realizations with low posterior density (as may also be the case for constructed grids), to prevent samples from regions of low posterior density dominating the calculation of \bar{U}_k (6.3.4), one could either discard such samples (retaining the most a-posteriori likely samples) or calculate a weighted sum, where the weights are proportional to the posterior density.

For the examples in Section 6.5, we have found that using a combination of grids and prior sampling, and retaining the most probable samples, produces acceptable accuracy. However, for more challenging systems, we propose that obtaining samples through simulated annealing will be the most practicable generic solution. Finally, we note that in a sequential Monte Carlo setting (8), a population of particles in the target model will likely already be available from which the CPS proposal can be constructed. Accordingly the population sampler would be a very natural framework in which to implement the CPS proposal.

6.5 Examples

We consider two model classes: an autoregressive (AR) time series setting, and a univariate Gaussian mixture analysis. In each case we detail our sampler and compare its performance to alternative approaches in the literature. Comparisons are performed on both synthetic and real benchmark data sets. We utilize the CPS approach as both an automated and efficient al-

ternative trans-dimensional sampler, in addition to a diagnostic tool to assess the performance of between-model mapping strategies.

6.5.1 An autoregression with unobserved initial states

Several approaches have been developed for fitting AR(k) models of unknown order (e.g. 9; 2). Here, we consider the hierarchical model and sampler of (25). We do not enforce stationarity of the AR process as this can complicate standard sampler implementation and the fairness of sampler comparison. Stationarity enforced through constraints on the autoregressive coefficient priors implies that unconstrained proposals suffer an automatic rejection rate. Other popular standard approaches (typically based on moment matching considerations; 25) complicate implementation. In this setting CPS proposals provide superior performance as prior constraints are automatically satisfied. We take this advantage, and consequent improvement in acceptance rates as a given, and do not enforce stationarity. This is more favourable ground for standard samplers, and hence a more informative comparison.

If $\mathbf{a}_k = (a_1, \dots, a_k)^\top$ is a vector of autoregressive coefficients, the observations $\mathbf{y} = (y_1, \dots, y_n)$ may be represented by the AR(k) process

$$M_k : \quad y_t = Y_k \mathbf{a}_k + \sigma_k e_t \quad k \in \{0, 1, \dots, k_{\max}\}, \quad t = 1, \dots, n$$

where $e_t \sim N(0, 1)$ and σ_k is the noise standard deviation. The Gaussian likelihood for this model, $p(\mathbf{y} \mid k, \mathbf{a}_k, \sigma_k^2, \mathbf{y}_{0,k})$, also depends on the vector of unobserved initial states $\mathbf{y}_{0,k} = (y_0, y_{-1}, \dots, y_{-k+1})$ which are incorporated into the design matrix Y_k . Writing the parameter vector for model M_k as $\boldsymbol{\theta}_k = (\mathbf{a}_k, \sigma_k^2, \delta^2, \Lambda, \xi^2, \mathbf{y}_{0,k})$ a hierarchical prior is given by

$$p(k, \boldsymbol{\theta}_k) = p(\mathbf{x}_{0,k} \mid k, \sigma_k^2, \xi^2) p(\mathbf{a}_k, \sigma_k^2 \mid k, \delta^2) p(k \mid \Lambda) p(\delta^2) p(\Lambda) p(\xi^2)$$

where δ^2, Λ and ξ^2 are hyperparameters. Here, (25) adopt priors to exploit marginalization through conditional conjugacy, so that sampling for \mathbf{a}_k and σ_k^2 is done from the full conditionals. A reversible jump MCMC step is only required for the initial states $\mathbf{y}_{0,k}$ through the joint posterior conditional $p(k, \mathbf{y}_{0,k} \mid \delta^2, \Lambda, \xi^2, \mathbf{y})$. The birth/death proposal $(k, \mathbf{y}_{0,k}) \leftrightarrow (k+1, \mathbf{y}_{0,k+1})$ is achieved using the reversibility of the AR process, simulating from $N(\hat{\mathbf{a}}_{k+1}^{\text{mle}}, \hat{\sigma}_{k+1}^2)$, where the parameters are maximum likelihood estimates of the AR($k+1$) process with initial conditions set to zero. For brevity, we refer to this maximum likelihood proposal distribution as MLE1.

To evaluate between-model move efficiency we consider an identical sampler but replace the MLE1 proposals with the conditional path-sampling proposal of Section 6.3.2, denoted by CPS1, with all remaining sampler details unchanged. The MLE1 proposal may be extended to a birth/death move involving two components $(k, \mathbf{y}_{0,k}) \leftrightarrow (k+2, \mathbf{y}_{0,k+2})$, termed MLE2, although this is not explicitly discussed by (25). Accordingly we also implement a sampler utilizing an equally weighted mixture of MLE1 and MLE2 proposals, denoted MLE1/2, and a corresponding sampler using a mixture of CPS1 and CPS2 proposals; CPS1/2.

Tab. 6.1: Posterior probability estimates of $\pi(M_3 | \mathbf{y})$ for simulated AR(3) data, under different samplers.

| n: | 35 | 50 | 100 | 150 | 200 | 500 |
|--------|--------|--------|--------|--------|--------|--------|
| CPS1 | 0.1540 | 0.5060 | 0.7812 | 0.7141 | 0.8291 | 0.8915 |
| CPS1/2 | 0.6688 | 0.7608 | 0.8312 | 0.8950 | 0.8629 | 0.9121 |
| MLE1 | 0.0031 | 0.0101 | 0.2987 | 0.4135 | 0.5167 | 0.7091 |
| MLE1/2 | 0.0839 | 0.1915 | 0.2585 | 0.4850 | 0.5701 | 0.7196 |

For the birth/death move of the CPS1 sampler, the proposal distribution for $y_{0,-k}$ (from $\mathbf{y}_{0,k+1}$) was constructed over 81 equally spaced grid values over $(-10,10)$. For the two-component CPS2 move, for each similarly gridded $y_{0,-k}$ value, 100 $y_{0,-k-1}$ (from $\mathbf{y}_{0,k+2}$) values are drawn from the prior, and proposal construction is based on those 50 samples with highest posterior density. Derivation of an analytic expressions for e.g. $U_k(y_{0,-k})$ (c.f. 6.3.3), the gradient of the density at $y_{0,-k}$, is not straightforward. Hence we implemented the numerical approximation (6.4.2) with $\Delta = 0.1y_{0,-k}$. The sensitivity of this approach was examined but did not have a significant impact on either proposal estimation or sampler mixing properties.

Results

We now compare the performance of the four samplers on synthetic and real data. In a similar vein to a study by (25) we analyze a synthetic dataset of length $n = 500$, generated from an AR(3) model with parameters $\mathbf{a}_3 = (-0.9, 0.2, 0.63)^\top$, $\sigma_3 = 0.1$, and truncated to form (nested) datasets of lengths $n = 35, 50, 100, 150, 200, 500$. Each sampler was identically initialized in an AR(1) model and run on each dataset for 100,000 iterations to obtain stationary between-model mixing, with the initial 20,000 iterations discarded.

Table 6.1 shows the posterior model probability estimates for an AR(3) as the length of the dataset increases. Both approaches have estimated posterior model probabilities for the correct model order which converge to one asymptotically as the number of observations $n \rightarrow \infty$, though clearly at different rates. When the data are comparatively weakly informative (small n), the MLE proposals generate far poorer between model mixing than the CPS proposals. If (much) longer chains are implemented all sampler outputs agree, confirming that mixing is the measured effect. (We note that the study by 25 was based on multiple and small data sets, and so our results are not directly comparable.)

For a real data comparison we consider the Southern Oscillation Index (SOI) time series, consisting of 540 monthly recordings between 1950–1995. These data represent the difference of the departure from the long-term monthly mean sea level pressures at Tahiti in the South Pacific and Darwin in Northern Australia. Previous comprehensive analysis fitting ARIMA models gave maximum posterior probability to an ARMA(1,1), with decreasing support for AR(3), AR(4), ARMA(2,1) and ARMA(4,1) models and low probabilities to remaining models (9). A further search over ARMA models performed using both AIC and BIC criteria, also selected ARMA(1,1) as the best model. The model of (25) induces a more parsimonious interpretation of the SOI data when compared to alternatives (a potential advantage of this model), though still giving the AR(3) maximum posterior probability.

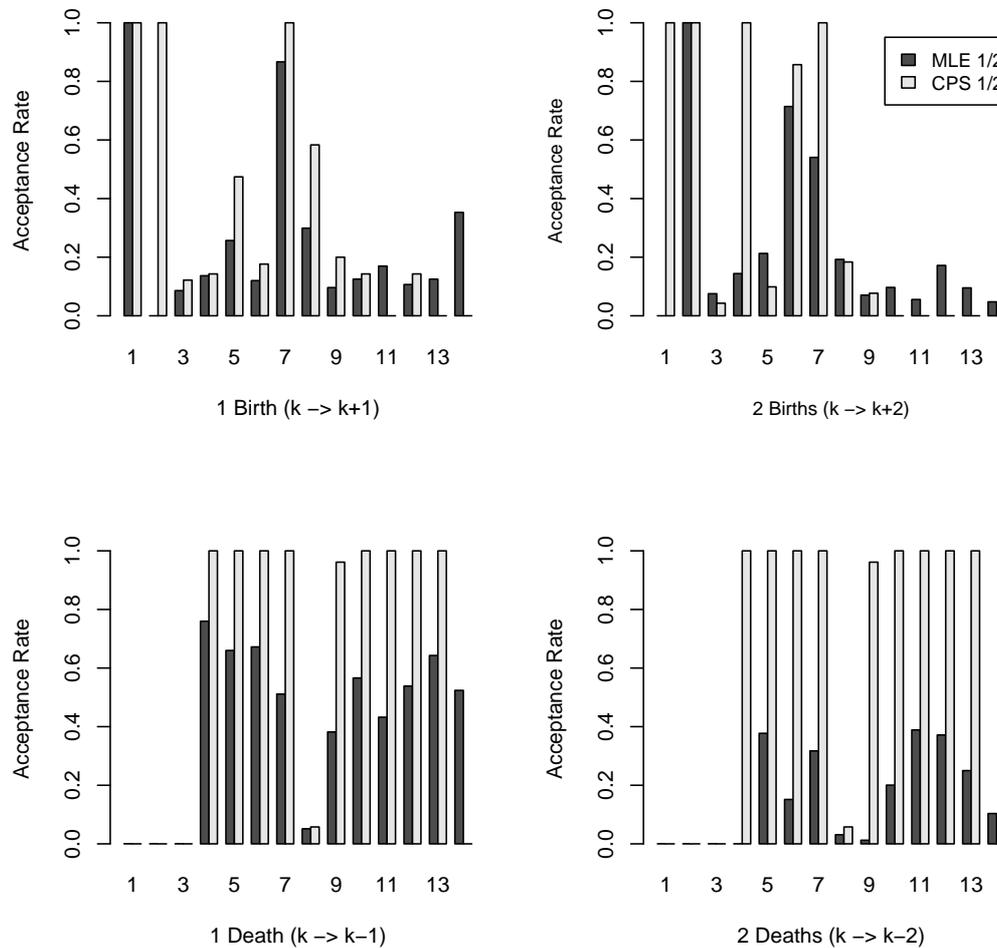


Fig. 6.5.1: Sampler acceptance rates for birth(top)/death(bottom) moves using (left) MLE1 and CPS1, and (right) MLE2 and CPS2 proposal distributions in the MLE1/2 and CPS1/2 samplers.

We implement the MLE1/2 and CPS1/2 samplers for 20,000 iterations with the initial 5,000 samples discarded. Good agreement was found between the sampler outputs for these data, given the size of the dataset and the zeroes of the autoregressive polynomial lying away from the unit circle.

Figure 6.5.1 illustrates acceptance rates of birth/death moves from an $AR(k)$ model for each sampler and move type. Overall, the CPS proposals consistently produce higher acceptance rates (for the larger order models, rates are based on very few move attempts). In particular, excluding $k = 8$, the death move acceptance rates are almost unity for the CPS proposals when moving to an inferior model (recall $AR(3)$ is the MAP model), reflecting the benefits of more accurately constructed proposals. Figure 6.5.2 graphically illustrates typical differences between CPS2 (top) and MLE2 (bottom) proposals for $(y_{0,-k}, y_{0,-k-1})$. The CPS2 distributions are estimated using extra points not used in the sampler for visualization purposes. Scale, orientation (correlation) and location disparities are common.

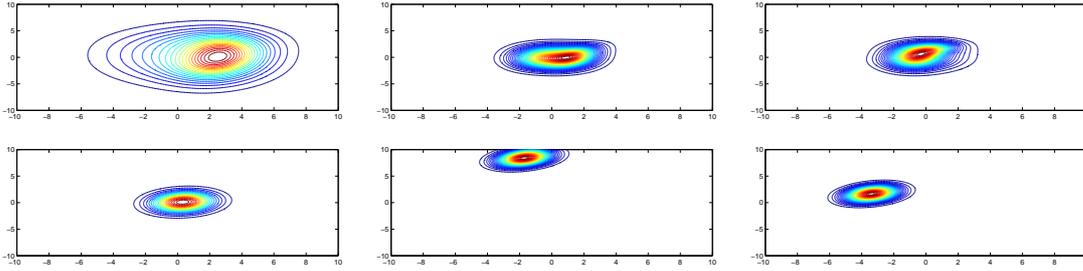


Fig. 6.5.2: Comparison of CPS2 (top) and MLE2 (bottom) proposal distribution pairs for $(y_{0,-k}, y_{0,-k-1})$.

Both sampler codes were written in Matlab and run on a Intel Core 2 Duo processor at 2.4GHz with 4GB RAM. Under these same conditions, clock time for the MLE1/2 sampler was faster than the presented CPS1/2 sampler by approximately 30%. However, this should be taken more as a guide since the number of points used in constructing the CPS proposal sampler will affect this comparison.

6.5.2 Mixture of univariate Gaussians

As an alternative to automating between-model transitions, we now use CPS proposals to evaluate the efficacy of a commonly used between-model mapping strategy: moment matching. In particular, we consider the mixture of univariate Gaussians model and sampler of (21), along with their latent variable and hierarchical prior specification. Here,

$$y_j \sim \sum_{i=1}^k w_i \phi(\cdot \mid \mu_i, \sigma_i) \quad j = 1, \dots, n$$

where $\phi(\cdot \mid \mu_i, \sigma_i)$ denotes the i -th Gaussian mixture component with mean μ_i and standard deviation σ_i , and where $\sum_i w_i = 1$. The (21) sampler consists of a birth/death proposal which creates/deletes components with no allocated observations, and a split/merge proposal which reversibly maps one Gaussian component, j^* , to two components, j_1 and j_2 , and redistributes the allocated observations. The split/merge mapping maintains the 0th, 1st and 2nd moments via the constraints:

$$\begin{aligned} w_{j^*} &= w_{j_1} + w_{j_2} \\ w_{j^*} \mu_{j^*} &= w_{j_1} \mu_{j_1} + w_{j_2} \mu_{j_2} \\ w_{j^*} (\mu_{j^*}^2 + \sigma_{j^*}^2) &= w_{j_1} (\mu_{j_1}^2 + \sigma_{j_1}^2) + w_{j_2} (\mu_{j_2}^2 + \sigma_{j_2}^2). \end{aligned}$$

The remaining three parameters (excluding component allocation variables) are formed as functions (h_{ij}) of independent draws from Beta distributions, and the final proposal is rejected if the ordering constraint $\mu_{j_1} < \mu_{j_2}$, with no other μ_j in the interval $[\mu_{j_1}, \mu_{j_2}]$, is not satisfied. For the following, we compare the moment matching (MM) sampler of (21) to an identical sampler but substituting a CPS proposal for the split/merge move.

In common with many other between-model moves, MM has two weaknesses. Firstly, the or-

dering constraint on the means $\mu_1 < \dots < \mu_k$ for mixture component “identifiability” (20) immediately introduces a non-zero rejection rate under MM. We argue that avoiding this in non-trivial examples through mapping function adjustments to account for arbitrarily constrained regions increases implementational and analytic difficulties. Secondly, the random vector \mathbf{u} is drawn independently of both model and data. Hence, beyond the above moment constraints the proposal mechanism is proposing blindly, with little ability to draw new samples from regions of high posterior density. CPS proposals are both model and data based, and naturally incorporate parameter constraints, and so we expect them to outperform MM in this setting.

We construct the CPS proposal by sampling the split component parameters via their one-sided conditionals (6.3.2) in the following order: $\mu_{j_1}, \mu_{j_2}, \sigma_{j_1}, \sigma_{j_2}, w_{j_1}$. Observation allocations are subsequently performed according to the full conditional distribution of the latent indicator variables. Each univariate density for μ_{j_1}, μ_{j_2} and w_{j_1} is estimated from a grid of 50 equally-spaced points, with the unconditioned variables integrated out by sampling 30 observations from their respective priors and retaining the 20 with highest posterior density. Integrations over the indicator variables were again performed using the full conditionals. Bounds for the grids were $[0, w_{j^*}]$ for w_{j_1} , and the means μ are bound by the identifiability ordering and the convex hull of the data. A grid for $\sigma_{j_1}^2, \sigma_{j_2}^2$ was based on a bivariate mixture of Gaussian densities centered on an estimate of the mode of the true conditional (see the Appendix for the estimation detail) with the grid points taken from a low (0.01) and high (0.1) standard deviation component in equal proportions. Densities for both means and standard deviations possessed Gaussian tails as appropriate. For the merge step we similarly construct the joint proposal for μ_{j^*}, σ_{j^*} , with $w_{j^*} = w_{j_1} + w_{j_2}$ set deterministically. All observations from components j_1 and j_2 were then allocated to component j^* . The full posterior form and the analytic forms for U_k (c.f. 6.3.3) are given in the Appendix.

Sensitivity of the CPS proposal construction was assessed by considering 10, 50, 150 and 1000 grid points and 5, 20, 150 and 500 samples from the prior for unconditioned parameters. Extra points improved proposal accuracy, and thereby sampler mixing, but also increased computation time. The following results, based on 50 grid points and the best 20 prior samples (from 30), represent good accuracy for moderate computation.

Results

We evaluate the performance of MM and CPS proposals under a number of settings for the well-known Enzyme data set, believed to have between 3 and 6 mixture components (21). Firstly we examine the mean probability of accepting a between-move transition. For each model $k = 1, \dots, 14$, we implement a fixed-model sampler for 1000 iterations burn-in. For each of the subsequent 500 iterations, following a within-model update we attempt both split ($k \rightarrow k + 1$) and merge ($k \rightarrow k - 1$) moves, but do not accept them. The proportion of moves “accepted” is an estimate of the mean probability of a between-model transition.

Table 6.2 displays these proportions, plus estimates of posterior model probabilities (based on a longer chain run) for both MM and CPS proposals. As with the autoregressive example, coincident posterior model probabilities are supported by identical within-model density estimates.

Tab. 6.2: Posterior model probabilities $\pi(M_k | \mathbf{y})$ and split/merge move acceptance rates under MM and CPS samplers. Acceptance rates denote proportion of 500 “attempted” split/merge moves accepted under fixed-model samplers. Model probabilities are based on 100K iterations with 20K iterations burn-in (CPS) and 200K with 100K burnin (MM, c.f. Richardson and Green, 1997).

| | | | | | | | | | |
|-------------------------------------|-------|-------|--------|--------|--------|-------|-------|-------|-------|
| $k:$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| CPS Split | 0.278 | 0.224 | 0.213 | 0.269 | 0.347 | 0.321 | 0.335 | 0.275 | 0.312 |
| CPS Merge | – | 0.080 | 0.113 | 0.151 | 0.171 | 0.466 | 0.394 | 0.452 | 0.454 |
| MM Split | 0.135 | 0.046 | 0.074 | 0.078 | 0.066 | 0.065 | 0.062 | 0.068 | 0.094 |
| MM Merge | – | 0.010 | 0.014 | 0.132 | 0.096 | 0.391 | 0.217 | 0.335 | 0.321 |
| $\pi(M_k \mathbf{y}, \text{CPS})$ | 0.000 | 0.041 | 0.241 | 0.300 | 0.207 | 0.105 | 0.067 | 0.021 | 0.009 |
| $\pi(M_k \mathbf{y}, \text{MM})$ | 0.000 | 0.024 | 0.290 | 0.317 | 0.206 | 0.095 | 0.041 | 0.017 | 0.007 |
| $k:$ | 10 | 11 | 12 | 13 | 14 | | | | |
| CPS Split | 0.404 | 0.386 | 0.425 | 0.496 | 0.411 | | | | |
| CPS Merge | 0.430 | 0.586 | 0.633 | 0.558 | 0.530 | | | | |
| MM Split | 0.074 | 0.094 | 0.102 | 0.089 | 0.111 | | | | |
| MM | 0.305 | 0.394 | 0.426 | 0.373 | 0.388 | | | | |
| | 0.003 | 0.003 | >0.001 | >0.001 | >0.001 | | | | |
| | 0.002 | 0.001 | >0.001 | >0.001 | >0.001 | | | | |

However we observe significantly higher split/merge acceptance rates for the CPS sampler in all models. The difference is particularly marked in higher order models, where the CPS proposal has greater success in moving between models with more diffuse components.

Figure 6.5.3 examines the accuracy of both CPS and MM proposals for a specific between-model move involving two high posterior probability models. We run a sampler in a $k = 3$ component model and then attempt to move to a 4-component mixture by splitting the highly weighted component $j = 1$ ($\mu_1 = 0.2, \sigma_1^2 = 0.01, w_1 = 0.78$) conditional on components 2 ($\mu_2 = 1.1, \sigma_2^2 = 0.5, w_2 = 0.21$) and 3 ($\mu_3 = 1.5, \sigma_3 = 1.0, w_3 = 0.01$) being common to both models. Here the CPS proposal is formed using 200 grid points for each conditional, for illustrative purposes. This particular example was examined (i.e. the splitting of a highly weighted component with the smallest mean) as this was where MM and CPS proposals were closest in terms of acceptance rates (discussed in further detail below), providing the most favourable comparison for MM.

The contours in Figure 6.5.3 illustrate the posterior densities of the first two components (under the mean ordering) estimated from the full, *unconditioned* $k = 4$ component density, indicating broadly where the target density for the between-model move lies. There are two well-defined modes for the component means and two (not-corresponding) modes for the variances. There is wide variation in the conditional distribution of (w_{j_1}, w_{j_2}) but with highest posterior density around $(0.5, 0.1)$. We re-emphasize that the contours represent the unconditioned density. Accordingly the target conditional density will be a restricted version of that displayed. Superimposed are 150 proposed moves to model $k = 4$ under both MM (open circles) and CPS (asterisks) proposals. At a glance, there are clearly differences between the two proposal den-

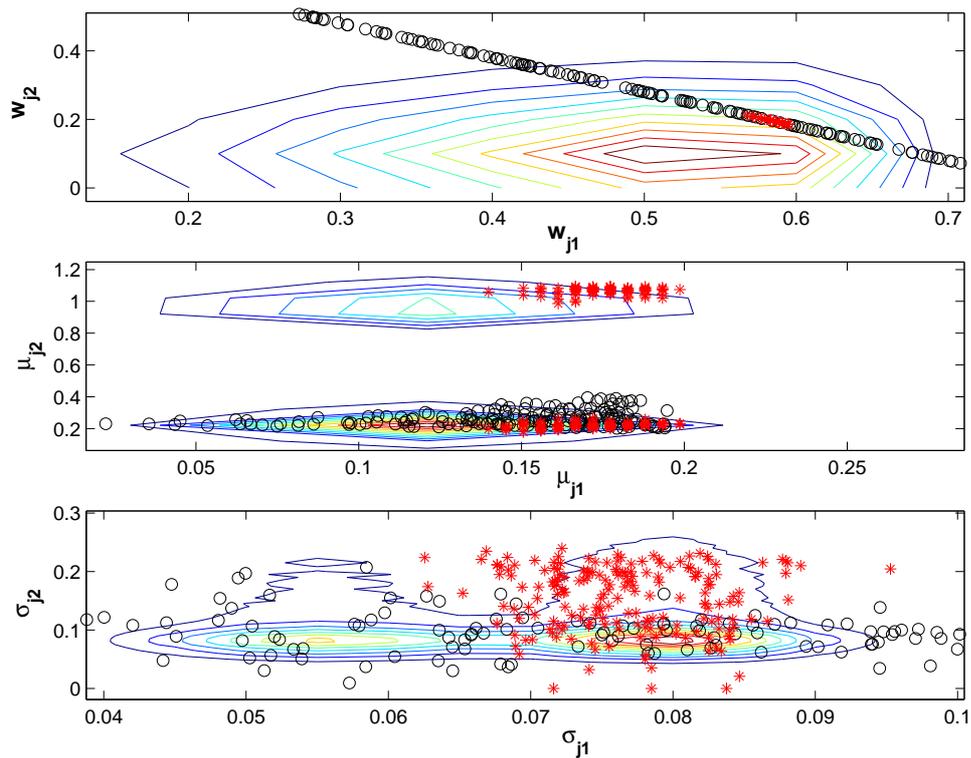


Fig. 6.5.3: Repeated parameter proposals for a fixed split move from a 3→4 component model under CPS (asterisks) and MM (open circles) proposals. Superimposed is the full, *unconditional* density for the first two components (under the mean ordering) of a 4 component model.

sities.

Stepping through the sequence of the CPS proposal, we first draw μ_{j_1} from a near-normal distribution between 0.15 and 0.2. MM also places plenty of proposed points in this region (and elsewhere). However the greatest difference occurs when sampling $\mu_{j_2} \mid \mu_{j_1}$ where a clear bimodal density is present under CPS (bound above by $\mu_2 = 1.1$). MM is restricted by the utilization of Beta-generated random variables and its proposal form (see 21), and so will accordingly find it difficult to reach this portion of the true conditional density. Additionally, MM frequently proposes means beyond any reasonable range given the data, such as large negative values (not shown in Figure 6.5.3).

The distribution of $\sigma_{j_1}^2 \mid \mu_{j_1}, \mu_{j_2}$ is as strongly defined as μ_{j_1} . Similarly, the bimodality exhibited by μ_{j_2} translates into a bimodality for $\sigma_{j_2}^2$, as seen in the rightmost contour-mode. (Note that the leftmost contour mode is present in the full, unconditional density for model $k = 4$, but not in the conditional proposal density, as is clear from reference points placed in this region during CPS density estimation.) In contrast, the MM proposals are spread over a wider range. Finally, the CPS proposal for w_{j_1} is very precise in contrast to the MM proposal which is exactly Beta(2,2) distributed over the range $[0, 0.78]$ (not shown in full). We note that the region of CPS density is very close to the mode of the unconditional (w_{j_1}, w_{j_2}) density, indicating proposal placement in regions of high posterior support.

While the CPS proposal is clearly outperforming MM in this instance, we now examine a wider

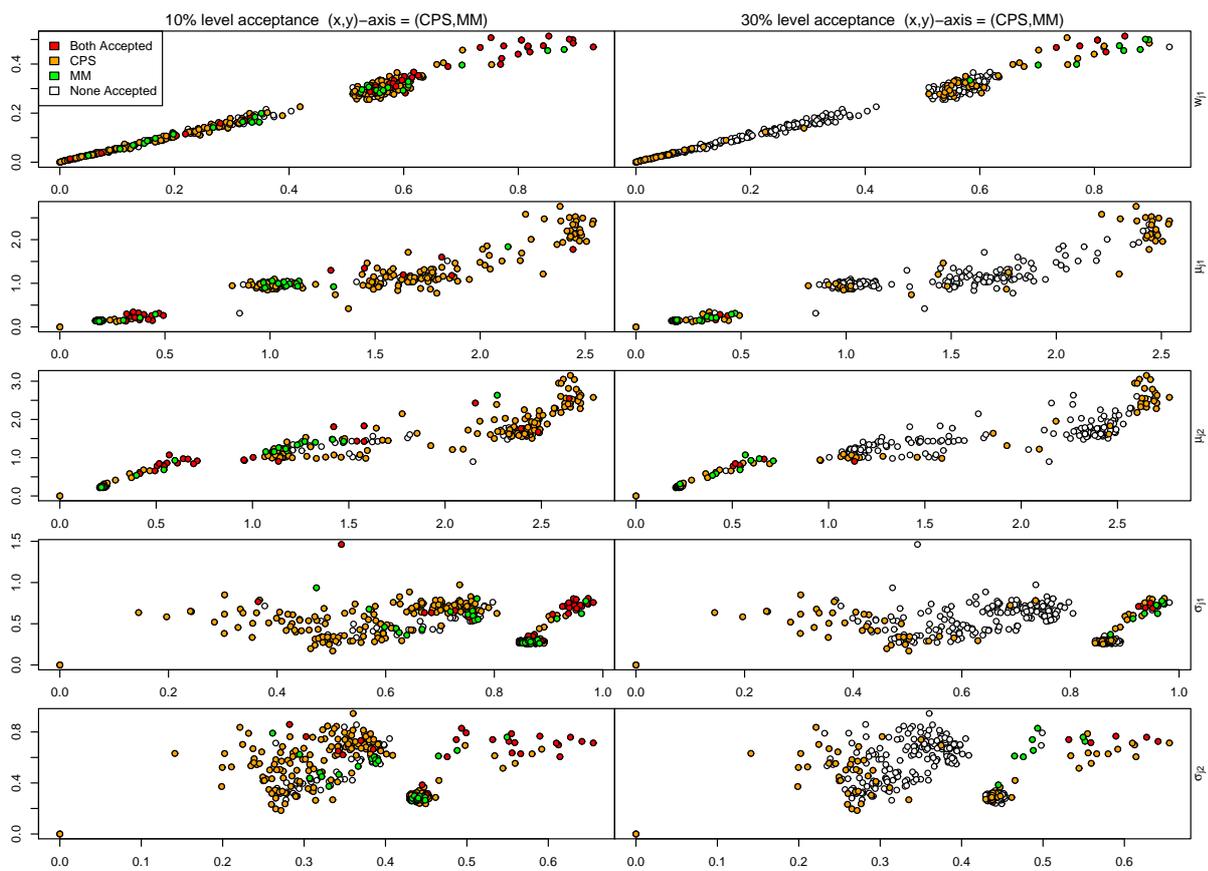


Fig. 6.5.4: Comparison of empirical means of proposals under CPS (x-axis) and MM (y-axis) schemes, for random split moves from a 3→4 component model. Shading indicates that more than 10% (left panels) and 30% (right) of proposals were accepted under CPS (orange), MM (green) or both (red).

range of proposal instances. We implement a sampler in a $k = 3$ component model and now attempt to move to the 4-component mixture by splitting a random component at each update, followed by a full within-model move. The CPS proposal is formed using 50 grid points for each conditional. We implement 150 different between-model moves. Marginal empirical means of the joint proposal were calculated as a measure of proposal location. Scatterplots of these means for CSP versus MM proposals are displayed in Figure 6.5.4. As a measure of the utility of each proposal distribution, each point is coloured (i.e. non-white) if the proportion of the proposals “accepted” at that iteration exceeds a threshold of 0.1 (left panels) or 0.3 (right panels) under CPS (orange), MM (green) or both proposals (red). As might be expected from Table 6.2, acceptance rates are clearly higher under CPS proposals, with large proportions of orange/red-coloured points over green/red at both acceptance levels. This continues to be true if other between model moves are considered e.g. moving from $k = 10$ to 11-component models.

Clearly the mean locations of proposal distributions under the MM and CPS proposals are not identical, as they tend to lie below the $y = x$ axis. This is mostly explained by the independence of MM proposals from model and data (beyond the MM constraints). For example, under MM, w_{j_1} is symmetrically distributed over $[0, w_{j^*}]$, and so the distribution of the marginal proposal mean is reasonably continuously distributed over the range $[0, w_{j^*}/2]$. Whereas CPS proposals place the full density where required on $[0, w_{j^*}]$, and so the marginal proposal mean is likely a good estimate of where the true posterior mass lies (e.g. Figure 6.5.3). Similarly, the ability of CPS proposals to construct multi-modal distributions will more accurately raise the marginal means of CPS over MM proposals.

One natural presumption regarding between-model moves, is that a proposal which better adapts according to the current state of the chain and the target density will outperform one which does this poorly. Monitoring between-model acceptance rates (e.g. Table 6.2) averages this information over the parameter space, and potentially useful information about between-model moves is lost. Fortunately, Figure 6.5.4 displays sufficient information in which to identify under what circumstances MM performs relatively well or badly. The CPS proposal’s good approximation to the true conditional target density translates directly into the clustering of the marginal proposal means on the CPS axis in Figure 6.5.4. In comparison MM proposals means are not nearly so discriminatory (i.e. less clustering), unless the data are highly informative for a parameter, as seen in the MM distribution for μ_{j_1} . Accordingly we can identify clusters where one move proposal does well compared to the other.

The smallest weights under CPS proposals ($w_{j_1} < 0.2$) correspond directly to the largest means ($2.3 < \mu_{j_1} < \mu_{j_2}$) and smaller-to-medium values of $\sigma_{j_1}, \sigma_{j_2}$. These relate to the splitting of components with low posterior mass and fewer associated observations in the upper tail of the observed data. Most of these means have average acceptance rates above 0.3 under CPS proposals, so there is strong evidence that utilizing the information in low-probability components for extreme observations is crucial in the formulation of accurate between-model proposals. Accordingly, MM is weak in this area. It follows that as moves between models with larger

numbers of components are more likely to involve splitting low-probability components, CPS proposals should substantially outperform MM for higher order models. This effect is strongly observed in Table 6.2.

At the other end of the scale, the largest weights under CPS proposals ($w_{j_1} > 0.65$) are associated with the smallest means ($\mu_{j_1} < 0.5, \mu_{j_2} < 0.7$) and the larger cluster of standard deviations. This corresponds to a dominant, well-defined mode in the observed data. Here, acceptance rates for both between-model proposals are higher than 0.3, and there is some evidence that both proposals are performing at near-comparable levels, on average.

From inspection of the w_{j_1} plots in Figure 6.5.4, it is apparent that the successful CPS proposal splits the high posterior probability component into a high (w_{j_1}) and low probability ($w_{j_2} = w_{j^*} - w_{j_1}$) component. Under the moment constraints, if MM proposes a high and low probability component split, the mean μ_{j_1} associated with the high probability component will remain close to the original mean μ_{j^*} , and the mean μ_{j_2} will move far from μ_{j^*} . Given that the dominant mode of the Enzyme dataset is largely separate from the remainder of the dataset, it becomes relatively unimportant how far away the mean μ_{j_2} is from μ_{j^*} , as most observations will subsequently be allocated to the high probability component. Hence, the overall MM move is reasonably likely to be accepted as the MM proposal then only depends on broadly proposing a high/low probability component split, which itself depends on the single random variate u_1 rather than the full random vector component $\mathbf{u} = (u_1, u_2, u_3)$. As it is reasonably likely that such a high/low probability component split will occur under the independent $u_1 \sim \text{Beta}(2, 2)$ proposal for $w_{j_1} = u_1 w_{j^*}$, MM is seen to perform relatively well.

However, the split of lower-probability components ($w_{j_1} < 0.2$) is accepted less often under MM than CPS proposals. This can be attributed to the lower probability (uppermost) modes of the Enzyme dataset overlapping and being accordingly less well defined than the dominant data mode. As a consequence, split moves require more accuracy and precision in order to locate (conditionally) high posterior density regions. The previous insensitivity of the proposal to (u_2, u_3) for well defined modes is lost as a result, and so the acceptance rate falls dramatically. However, we reiterate that while MM performs at its best for separate, well-defined components, it can remain far from optimum, as illustrated in Figure 6.5.3.

In summary, MM incorporates partial information from model and data in the formation of between-model moves through moment constraints. When applied to clearly defined observed data modes MM can perform adequately, but this ability deteriorates as the data modes overlap and become less distinct. The strength of CPS proposals in this setting is the capacity to adaptively and accurately capture crucial conditional posterior aspects such as location and multi-modality leading to tangible performance gains.

6.6 Discussion

Eliciting mapping functions between models of possibly differing dimensionality and interpretation is at best a challenging procedure. This is barely less daunting when local mappings are considered, and the practitioner is an expert, so progress towards the goal of black-box reversible jump MCMC simulation has been slow. Through use of a marginal density estimator we have demonstrated that it is feasible to ignore the myriad intricacies of between-model mappings, and approximate the true density in the target model directly. Beyond automation, this approach is useful in providing commentary on possible mapping function strategies (e.g. moment matching) in terms of potential strengths and weaknesses.

The between-model transition strategies considered in this article may in some sense be considered extreme opposites. Moment matching applies expert intuition that maintaining the first three moments (in the case of a mixture of Gaussians) could be sensible, but effectively proposes blindly within these constraints. In contrast, the density estimation approach eschews expert knowledge in favour of simply identifying regions of high posterior density. Under this simpler but more computational strategy, more ambitious mappings with improved mixing properties are easily entertained – for example split/merge moves between $2 \leftrightarrow 3$ components – provided one is willing to pay the computational price tag.

However, it may be that the most efficient strategy results from a fusion of expert intuition with the location of high posterior density regions, which combines the best features of both approaches. For example in the mixture of Gaussians setting, one may implement moment matching coupled with density estimation for the remaining parameters, resulting in improved mapping accuracy for small computational cost.

Acknowledgment

YF and SAS are supported by the Australian Research Council through the Discovery Project scheme (DP0664970). GWP is supported by an Australian Postgraduate Award and by CSIRO Mathematical and Information Sciences. The authors thank M. Briers and P. Shevchenko for useful discussions.

Appendix

CPS proposal details

The mixture of Gaussians example between-model move reversibly maps component j^* under a k -component model to two components j_1, j_2 under a $(k + 1)$ -component model. Here we describe further details of this process in the CPS framework. Following the joint posterior of (21), the required joint conditional is

$$\begin{aligned} & p(w_{j_1}, w_{j_2}, \mu_{j_1}, \mu_{j_2}, \sigma_{j_1}^2, \sigma_{j_2}^2, z_{\{j_1, j_2\}} \mid k + 1, w_{-\{j_1, j_2\}}, \theta_{-\{j_1, j_2\}}, z_{-\{j_1, j_2\}}, \mathbf{y}, \lambda, \delta, \eta) \\ \propto & \left[\prod_{i=1}^n \prod_{j=j_1, j_2} \left(\frac{w_j}{\sqrt{2\pi}\sigma_j} \exp \left\{ -\frac{(y_i - \mu_j)^2}{2\sigma_j^2} \right\} \right)^{\mathbb{I}(z_i=j)} \right] \left[\prod_{j=j_1, j_2} w_j^{\delta_j-1} \delta irac \left(1 - \sum_{i=1}^{k+1} w_i \right) \right] \\ & \times \left[\prod_{j=j_1, j_2} \frac{1}{\sqrt{2\pi}\kappa^{-1}} \exp \left\{ -\frac{(\mu_j - \xi)^2}{2\kappa^{-2}} \right\} \left(\sigma_j^{2(-\alpha-1)} \frac{\beta^\alpha e^{-\beta/\sigma_j^2}}{\Gamma(\alpha)} \right) \right] \left[\prod_{i=1}^n \prod_{j=j_1, j_2} (w_j)^{\mathbb{I}(z_i=j)} \right] \end{aligned}$$

where, for example, $w_{-\{j_1, j_2\}}$ denotes the vector of weights in the $(k + 1)$ -component model excluding weights w_{j_1} and w_{j_2} , and $z_{\{j_1, j_2\}}$ denotes the allocation variables associated with components j_1 and j_2 .

A 5-dimensional CPS proposal is constructed on the above $(n + 6)$ dimensional density by sampling in the following order $\mu_{j_1}, \mu_{j_2}, \sigma_{j_1}^2, \sigma_{j_2}^2, w_{j_1}$. The weights $w_{j_1} + w_{j_2} = w_{j^*}$ are constrained, and the allocation variables are sampled (and integrated out) from their full conditionals $p(z_i = j \mid k + 1, \dots) \propto \frac{w_j}{\sigma_j} \exp[-(y_i - \mu_j)^2 / (2\sigma_j^2)]$, for $j = j_1, j_2$, thereby reducing the proposal dimensionality by $(n + 1)$ dimensions.

We denote the gradients of the log-posterior for the CPS “split” proposal by \vec{U} . Then, for $\vec{U}_1(\mu_{j_1})$ and $\vec{U}_2(\mu_{j_2})$, with $k = 1, 2$, we have

$$\vec{U}_k(\mu_{j_k}) \propto \sum_{i=1}^n (y_i - \mu_{j_k}) \mathbb{I}(z_i = j_k) / \sigma_{j_k}^2 - \kappa^2 (\mu_{j_k} - \xi),$$

for $\vec{U}_3(\sigma_{j_1}^2)$ and $\vec{U}_4(\sigma_{j_2}^2)$, with $k = 1, 2$, we have

$$\vec{U}_{k+2}(\sigma_{j_k}^2) \propto \frac{1}{2} \sum_{i=1}^n \left[(y_i - \mu_{j_k})^2 \sigma_{j_k}^{-4} - \sigma_{j_k}^{-2} \right] \mathbb{I}(z_i = j_k) - (\alpha + 1) / \sigma_{j_k}^2 + \beta / \sigma_{j_k}^4,$$

and

$$\vec{U}_5(w_{j_1}) \propto 2 \sum_{j=1}^n \mathbb{I}(z_j = j_1) / w_{j_1} + (\delta_{j_1} - 1) / w_{j_1}.$$

The complementary “merge” gradients, $\overleftarrow{U}(\mu_{j^*})$ and $\overleftarrow{U}(\sigma_{j^*}^2)$, are functionally identical to the above, with the obvious notational changes (e.g. $j_k \rightarrow j^*$ and $\mathbb{I}(z_i = j_k) \rightarrow 1$).

While μ_{j_k} and w_{j_1} are bounded above and below, and so their density estimates are constructed over grids, $\sigma_{j_k}^2$ is unbounded above. However the mode of the CPS proposal for σ_j^2 may be located by setting $\vec{U}_{k+2}(\sigma_{j_k}^{2*}) = 0$, with $k = 1, 2$, and solving for $\sigma_{j_k}^{2*}$, giving

$$\sigma_{j_k}^{2\text{mode}} = \left[\frac{1}{2} \sum_{i=1}^n (y_i - \mu_{j_k})^2 \mathbb{I}(z_i = j_k) + \beta \right] / (n_{j_k}/2 + \alpha + 1)$$

where n_{j_k} is the number of observations allocated to component j_k . Under the CPS proposal for $\sigma_{j_k}^2$, μ_{j_k} is fixed, but n_{j_k} and the allocation variables such that $z_i = j_k$ require numerically integrating out. Accordingly we consider the mode of the CPS proposal as the mean of $\sigma_{j_k}^{2\text{mode}}$ taken over each sampled $\{n_{j_k}, z_i = j_k\}$.

References

- [1] Al-Awadhi, F.; Hurn, M. A. Jennison, C. Improving the acceptance rate of reversible jump MCMC proposals *Statistics and Probability Letters*, 2004, 69, 189-198
- [2] Barnett, G.; Kohn, R. Sheather, S. Bayesian estimation of an autoregressive model using Markov chain Monte Carlo *Journal of Econometrics*, Australian School of Management UNSW, 1996, 74, 237-254
- [3] Brooks, S. P.; Giudici, P. Roberts, G. O. Efficient construction of reversible jump MCMC proposal distributions *Journal of the Royal Statistical Society, B*, 2003, 65, 3-55
- [4] Chen, M. Importance weighted marginal Bayesian posterior density estimation *Journal of the American Statistical Association*, 1994, 89, 818-824
- [5] Chib, S. Marginal likelihood from the Gibbs output *Journal of the American Statistical Association*, 1995, 90, 1313-1321
- [6] Chipman, H.; George, E. McCulloch, R. E. Lahiri, P. (ed.) *Model Selection The practical implementation of Bayesian model selection* Institute of Mathematical Statistics, 2001, 38, 67-134
- [7] Clyde, M. A. George, E. I. Model uncertainty *Statistical Science*, 2004, 19, 81-94
- [8] Moral, P. D.; Doucet, A. Jasra, A. Sequential Monte Carlo samplers *J. R. Statist. Soc. B*, 2006, 68, 411 - 436
- [9] Ehlers, R. Brooks, S. P. *Bayesian analysis of order uncertainty for ARIMA models* university of Cambridge, 2006
- [10] Fan, Y.; Brooks, S. P. Gelman, A. Output assessment for Monte Carlo simulations via the score statistic *J. Comp. Graph. Stat.*, 2006, 15, 178-206
- [11] Gelfand, A. E.; Smith, A. F. M. Lee, T. M. Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling *Journal of the American Statistical Association*, 1992, 87, 523-532
- [12] Gelman, A. Meng, X.-L. Simulating normalising constants: From importance sampling to bridge sampling to path sampling *Statistical Science*, 1998, 13, 163-185

-
- [13] George, E. I. McCulloch, R. E. Variable search via Gibbs sampling *Journal of the American Statistical Association*, 1993, 88, 881-889
- [14] Gilks, W. R. Wild, P. Adaptive rejection sampling for Gibbs sampling *Applied Statistics*, 1992, 41, 337-348
- [15] Green, P. J. Trans-dimensional Markov chain Monte Carlo *Highly Structured Stochastic Systems*, OUP, 2003, 179-198
- [16] Green, P. J. Reversible jump MCMC computation and Bayesian model determination *Biometrika*, 1995, 82, 711-732
- [17] Green, P. J. Mira, A. Delayed rejection in reversible jump Metropolis-Hastings *Biometrika*, 2001, 88, 1035-1053
- [18] Hastie, D. *Developments in Markov chain Monte Carlo* University of Bristol, 2004
- [19] Hoeting, J. A.; Madigan, D.; Raftery, A. E. Volinsky, C. T. Bayesian model averaging: A tutorial (with discussion) *Statistical Science*, 1999, 14, 382-417
- [20] Jasra, A.; Holmes, C. Stephens, D. A. Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modelling *Statistical Science*, 2005, 20, 50-67
- [21] Richardson, S. Green, P. J. On Bayesian analysis of mixtures with an unknown number of components *Journal of Royal Statistical Society, Series B*, 1997, 59, 731-792
- [22] Sisson, S. A. Trans-dimensional Markov chains: A decade of progress and future perspectives *Journal of the American Statistical Association*, 2005, 100, 1077-1089
- [23] Stephens, M. Bayesian analysis of mixtures with an unknown number of components — an alternative to reversible jump methods *Annals of Statistics*, 2000, 28, 40-74
- [24] Verdinelli, I. Wasserman, L. Computing Bayes factors using a generalisation of the Savage-Dickey density ratio *Journal of the American Statistical Association*, 1995, 90, 614-618
- [25] Vermaak, J.; Andrieu, C.; Doucet, A. Godsill, S. J. Reversible jump Markov chain Monte Carlo strategies for Bayesian model selection in autoregressive processes *Journal of Time Series Analysis*, 2004, 25, 785-809

Journal Paper 5

"The essence of mathematics is its freedom."

George Cantor

Peters G.W., Balikrishnan K., Lasscock B. and Mellen C. (2009) "Model selection and adaptive Markov chain Monte Carlo for Bayesian cointegrated VAR models." In review.

This work was instigated by the first author and he is the lead author on this paper. He can claim around 85% of the credit for the contents. His work included developing the methodology contained, developing the applications, (80% of the implementation) and comparison to alternative approaches, writing of the drafts of the paper and undertaking revisions. The work has already been presented at two seminars and one international computational economics conference. Permission from all the co-authors has been granted for submission of this paper as part of the thesis.

Model Selection and Adaptive Markov chain Monte Carlo for Bayesian Cointegrated VAR model.

Gareth W. Peters

CSIRO Mathematical and Information Sciences, Locked Bag 17, North Ryde, NSW, 1670, Australia;
UNSW Mathematics and Statistics Department, Sydney, 2052, Australia

Balakrishnan B. Kannan

Baronia Capital Pty. Ltd., 12 Holtermann St., Crows Nest, NSW 2065, Australia.
e-mail:Balakrishnan.Kannan@boroniacapital.com.au

Ben Lasscock

Baronia Capital Pty. Ltd., 12 Holtermann St., Crows Nest, NSW 2065, Australia.
e-mail:ben.lasscock@boroniacapital.com.au

Chris Mellen

Baronia Capital Pty. Ltd., 12 Holtermann St., Crows Nest, NSW 2065, Australia.
e-mail: Chris.Mellen@boroniacapital.com.au

Submitted: 10 November 2009

7.1 Abstract

This paper develops a matrix-variate adaptive Markov chain Monte Carlo (MCMC) methodology for Bayesian Cointegrated Vector Auto Regressions (CVAR). We replace the popular approach to sampling Bayesian CVAR models, involving griddy Gibbs, with an automated efficient alternative, based on the Adaptive Metropolis algorithm of Roberts and Rosenthal, (2009). Developing the adaptive MCMC framework for Bayesian CVAR models allows for efficient estimation of posterior parameters in significantly higher dimensional CVAR series than previously possible with existing griddy Gibbs samplers. For a n -dimensional CVAR series, the matrix-variate posterior is in dimension $3n^2 + n$, with significant correlation present between the blocks of matrix random variables. Hence, utilising a griddy Gibbs sampler for large n becomes computationally impractical as it involves approximating a $n \times n$ full conditional posterior using a spline over a high dimensional $n \times n$ grid. The adaptive MCMC approach is demonstrated to be ideally suited to learning on-line a proposal to reflect the posterior correlation structure, therefore improving the computational efficiency of the sampler.

We also treat the rank of the CVAR model as a random variable and perform joint inference on the rank and model parameters. This is achieved with a Bayesian posterior distribution defined over both the rank and the CVAR model parameters, and inference is made via Bayes Factor analysis of rank.

Practically the adaptive sampler also aids in the development of automated Bayesian cointegration models for algorithmic trading systems considering instruments made up of several assets, such as currency baskets. Previously the literature on financial applications of CVAR trading models typically only considers pairs trading ($n=2$) due to the computational cost of the griddy Gibbs. We are able to extend under our adaptive framework to $n \gg 2$ and demonstrate an example with $n = 10$, resulting in a posterior distribution with parameters upto dimension 310. By also considering the rank as a random quantity we can ensure our resulting trading models are able to adjust to potentially time varying market conditions in a coherent statistical framework.

Keywords: Cointegrated Vector Auto Regression, Adaptive Markov chain Monte Carlo, Bayesian Inference, Bayes Factors.

7.2 Introduction

Bayesian analysis of Cointegrated Vector Auto Regression (CVAR) models has been addressed in several papers, see Koop *et al.* (2006) for an overview. In a Bayesian CVAR model, specification of the matrix-variate model parameters priors, to ensure the posterior is not improper, must be done with care, see Koop *et al.* (2006). This has significant implications on the Bayesian model structure, in particular one can not make a blind specification of priors on the VAR model coefficients as it may result in improper posterior distributions. For this reason it is common to consider the Error Correction Model (ECM) framework, see for example p.141-142 of Reinsel and Velu (1998). In this paper we do not aim to address the issue of prior choice or prior distortions and we adopt the model of Sugita (2002) and Geweke (1996) which admits desirable conjugacy properties. The resulting posterior for a n -dimensional CVAR series, is matrix-variate in dimension upto $3n^2 + n$ for full rank models, with significant correlation present between and within the blocks of matrix random variables. This presents a challenge to efficiently sample from the posterior distribution when n is large.

The focus of the paper and novelty introduced involves developing a Bayesian adaptive MCMC sampling, based on the proposed algorithm of Roberts and Rosenthal (2009), to allow us to significantly increase the dimension, n , of the CVAR series that can be estimated. Typically in the cointegration literature the sampling approach adopted is a gridgy Gibbs sampling framework, see Bauwens and Lubrano (1996), Geweke (1996), Kleibergen and van Dijk (1994) and Sugita (2002). The conjugacy properties of the Bayesian model we consider result in exact sampling of two of the matrix-variate random variables corresponding to the unknown error covariance matrix and the combined matrix random variable containing the cointegration equilibrium reversion rates α and the mean level μ of the CVAR series. However, the third unknown matrix-variate random variable corresponding to the cointegration vectors β has a marginal posterior distribution with support in dimension $n \times r$. When the cointegration rank r and the dimension of the CVAR series n is large ($n > 5$) then the standard gridgy Gibbs based samplers are no longer computationally viable samplers. Alternative samplers which may attempt to deconstruct the full conditional distribution of the posterior for the cointegration vectors β into components of this matrix, updating them one at a time will run into significant difficulties with efficiency in the mixing properties of the resulting Markov chain. The reason for this is due directly to two factors: the identification normalisation constraint of the matrix β ; and the strong correlation present in the full conditional posterior distribution for the matrix random variable β . Hence, utilising a gridgy Gibbs sampler for large n becomes computationally impractical as it involves approximating upto a $n \times n$ matrix-variate full conditional posterior using a spline constructed over a high dimensional space with d knot points per dimension, creating a requirement for d^n total grid points. The sampler we develop overcomes these difficulties utilizing an adaptive MCMC approach. We demonstrate that it is ideally suited to learning on-line a proposal to reflect the posterior correlation in the matrix-variate random variable, ensuring that updating this $n \times r$ matrix at each stage of the adaptive MCMC algorithm results in a non-trivial acceptance probability.

Adaptive MCMC is a new methodology to learn on-line the 'optimal' proposal distribution for an MCMC algorithm, see Atachade and Rosenthal (2005), Haario, Saksman and Tamminen (2001; 2007) and Andrieu and Moulines (2006) and more recently Giordani and Kohn, (2006) and Silva *et al.*, (2009), of which there are several different versions of adaptive MCMC and Particle MCMC algorithms. Basically adaptive MCMC algorithms aim to allow the Markov chain to adapt the Markov proposal distribution online throughout the simulation in such a way that the correct stationary distribution is still preserved, even though the Markov transition kernel of the chain is changing throughout the simulation. Clearly, this requires careful constraints on the type of adaption mechanism and the adaption rate to ensure that stationarity is preserved for the resulting Markov chain.

To summarise, this paper extends the matrix-variate block Gibbs sampling framework typically used in Bayesian Cointegration models by replacing the computational $n \times n$ dimensional griddy Gibbs sampler with two possible automated alternatives which are based on matrix-variate adaptive Metropolis-within-Gibbs samplers. Additionally, we consider rank estimation for reduced rank Cointegration models. From a Bayesian perspective we tackle this via Bayes Factor (BF) analysis for posterior "model" probabilities of the rank. Then we demonstrate estimation and predictive performance under a Bayesian setting for both Bayesian Model Selection (BMS) and Bayesian Model Averaging (BMA).

The models and algorithms developed allow for estimation of either the rank r , i.e. the model index, and the lag p of the CVAR model jointly with the model parameters. For simplicity we shall assume the lag is fixed and known.

In this paper the following notation will be used: $'$ denotes transpose, I_d is the $d \times d$ identity matrix, $p(\cdot)$ denotes a density and $P(\cdot)$ a distribution, Ω will be the space on which densities will take their support and it will be assumed throughout that we are working with Lebesgue measure. The operator \otimes denotes the Kronecker product, $\|\cdot\|$ denotes the total variation norm and Δ denotes the unit vector difference operator. We denote generically the state of a Markov chain at time j by random variable $\Theta^{(j)}$ and the transition kernel from realized state $\Theta^{(j-1)} = \theta$ to $\Theta^{(j)}$ by $Q(\theta, \Theta^{(j)})$. In the case of an adaptive transition kernel we will also assume that there is a sequence of transition kernels denoted by $Q_{\Gamma_j}(\theta, \Theta^{(j)})$, where Γ_j is the sequence index.

7.2.1 Contribution and structure

In section 7.3 we present the matrix-variate posterior distribution for the CVAR model formulated under an Error Correction Model (ECM) model framework. Next in section 7.4 we discuss the Bayesian CVAR model conditional on knowledge of the co-integration rank. This includes discussing and summarizing properties of the Bayesian CVAR model including identification, the justification of the ECM framework and issues to consider when selecting matrix priors for Bayesian CVAR models with respect to prior distortions. At this stage we make explicit the justification for why the Bayesian model decomposes the cointegration matrix $\Pi = \alpha\beta'$ under the ECM framework, since working directly with Π precludes direct use of Monte Carlo samples

for inference in the VAR model setting. As pointed out in Geweke (1995) and Sugita (2002), conditional on matrix β the nonlinear ECM model becomes linear and therefore under the informative priors we utilise, we can once again apply standard Bayesian analysis to the VAR model, this turns out to be a very useful property widely used in the cointegration literature. Then in section 7.5 we present the two algorithms developed based on Adaptive MCMC to obtain samples from the target posterior, followed by section 7.6 which presents the framework for rank estimation we utilise, along with discussion of model selection and model averaging, with respect to the unknown rank of the CVAR system. We conclude with both synthetic simulation examples with n ranging from 4 to 10, resulting in posteriors defined in dimensions between 52 and 310 dimensions. We also provide analysis on two real data examples from pairs and triples trading typically considered in real world financial algorithmic trading models.

7.3 CVAR model under ECM framework

We note that a well presented representation to co-integration models is provided by Engle and Granger (1987), Sugita, (2009) and for the original error correction representation of a co-integrated series, see Granger (1981) and Granger and Weiss (1983). Working with the definitions they provide for a co-integrated series, we denote the vector observation at time t by \mathbf{x}_t . Furthermore, we assume \mathbf{x}_t is an integrated of order 1, I(1), n -dimensional vector with r linear cointegrating relationships. The error vector at time t , ϵ_t are assumed time independent and zero mean multivariate Gaussian distributed, with covariance Σ . The Error Correction Model (ECM) representation is given by,

$$\Delta \mathbf{x}_t = \boldsymbol{\mu} + \boldsymbol{\alpha} \boldsymbol{\beta}' \mathbf{x}_{t-1} + \sum_{i=1}^{p-1} \Psi_i \Delta \mathbf{x}_{t-i} + \boldsymbol{\epsilon}_t \quad (7.3.1)$$

where $t = p, p+1, \dots, T$ and p is the number of lags. Furthermore, the matrix dimensions are: $\boldsymbol{\mu}$ and $\boldsymbol{\epsilon}_t$ are $(n \times 1)$, Ψ_i and Σ are $(n \times n)$, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are $(n \times r)$.

We can now re-express the model in equation (7.3.1) in a multivariate regression format, as follows

$$Y = X\Gamma + Z\boldsymbol{\beta}\boldsymbol{\alpha}' + E = WB + E, \quad (7.3.2)$$

where,

$$Y = \begin{pmatrix} \Delta \mathbf{x}'_p & \Delta \mathbf{x}'_{p+1} & \dots & \Delta \mathbf{x}'_T \end{pmatrix}', Z = \begin{pmatrix} \mathbf{x}'_{p-1} & \mathbf{x}'_p & \dots & \mathbf{x}'_{T-1} \end{pmatrix}'$$

$$E = \begin{pmatrix} \boldsymbol{\epsilon}'_p & \boldsymbol{\epsilon}'_{p+1} & \dots & \boldsymbol{\epsilon}'_T \end{pmatrix}', \Gamma = \begin{pmatrix} \boldsymbol{\mu}' & \Psi'_1 & \dots & \Psi'_{p-1} \end{pmatrix}'$$

$$X = \begin{pmatrix} 1 & \Delta \mathbf{x}'_{p-1} & \dots & \Delta \mathbf{x}'_1 \\ 1 & \Delta \mathbf{x}'_p & \dots & \Delta \mathbf{x}'_2 \\ \vdots & \vdots & \dots & \vdots \\ 1 & \Delta \mathbf{x}'_{T-1} & \dots & \Delta \mathbf{x}'_{T-p+1} \end{pmatrix}, W = \begin{pmatrix} X & Z\boldsymbol{\beta} \end{pmatrix}, B = \begin{pmatrix} \Gamma & \boldsymbol{\alpha}' \end{pmatrix}'$$

Here, we let t be the number of columns of Y , hence $t = T - p + 1$, producing X with dimension $t \times (1 + n(p - 1))$, Γ with dimension $((1 + n(p - 1)) \times n)$, W with dimension $t \times k$ and B with dimension $(k \times n)$, where $k = 1 + n(p - 1) + r$, see Sugita, (2002) for additional details regarding this parameterization. The parameters μ represents the trend coefficients, and Ψ_i is the i^{th} matrix of autoregressive coefficients and the long run multiplier matrix is given by $\Pi = \alpha\beta'$.

The long run multiplier matrix is an important quantity of this model, its properties include: if Π is a zero matrix, the series x_t contains n unit roots; if Π has full rank then each univariate series in x_t are (trend-)stationary; and co-integration occurs when Π is of rank $r < n$. The matrix β contains the co-integration vectors, reflecting the stationary long run relationships between the univariate series within x_t and the α matrix contains the adjustment parameters, specifying the speed of adjustment to equilibria $\beta'x_t$.

This results in a likelihood model, when the parameters of interest are B , Σ and β , given by

$$L(Y|B, \Sigma, \beta) = (2\pi)^{-0.5nt} |\Sigma \otimes I_t|^{-0.5} \exp(-0.5 \text{Vec}(Y - WB)'(\Sigma^{-1} \otimes I_t^{-1}) \text{Vec}(Y - WB)) \\ \propto |\Sigma|^{-0.5t} \exp(-0.5 \text{tr}[\Sigma^{-1}(\hat{S} + R)]),$$

where $\Sigma = \text{Cov}(E)$ and

$$R = (B - \hat{B})'W'W(B - \hat{B}), \hat{S} = (Y - W\hat{B})'(Y - W\hat{B}), \hat{B} = (W'W)^{-1}W'Y$$

7.4 Bayesian CVAR models conditional on Rank (r)

The assumptions and restrictions of our Bayesian CVAR model include:

1. **Identification Issue:** For any non-singular matrix A , the matrix of long run multipliers $\Pi = \alpha\beta'$ is indistinguishable from $\Pi = \alpha AA^{-1}\beta'$, see Koop, Strachan, Dijk and Villani (2006) or Reinsel and Velu (1998). We use a standard approach to globally overcome this problem by incorporating a non unique identification constraint. We impose r^2 restrictions as follows $\beta = [I_r, \beta_*']'$, where I_r denotes the $r \times r$ identity matrix. However, as noted by Kleibergen and van Dijk (1994) and discussed in Koop, Strachan, van Dijk and Villani (2006) this can still result in local identification issues at the point $\alpha = \mathbf{0}$, when β does not enter the model. Hence, one must be careful to ensure that the Markov chain generated by the matrix-variate block Gibbs sampler is not invalidated by the terminal absorbing state. As is standard we monitor the performance of the sampler to ensure this has not occurred.
2. **Error Correction Model:** The ECM framework complicates Bayesian analysis since products, $\alpha\beta'$, preclude direct use of Monte Carlo samples for inference in the VAR model setting. However, conditional on β the nonlinear ECM model becomes linear and therefore under the informative priors used by Geweke (1995) and Sugita (2002), we can once again apply standard Bayesian analysis to the VAR model.

3. **Prior Choices:** We do not consider the issue of prior distortions illustrated by Kleibergen and van Dijk (1994). This is not the focus of the present paper. Alternative prior models in the cointegration setting include Jeffrey's priors, Embedding approach and a focus on the cointegration space.

7.4.1 Prior and Posterior Model

Here we present the model for estimation of β , B and Σ conditional the rank r . As in Sugita (2002), we use a conjugate hierarchical prior.

- $\beta \sim N(\bar{\beta}, Q \otimes H^{-1})$ where $N(\bar{\beta}, Q \otimes H^{-1})$ is the matrix-variate Gaussian distribution with prior mean $\bar{\beta}$, Q is a $(r \times r)$ positive definite matrix, H a $(n \times n)$ matrix.
- $\Sigma \sim IW(S, h)$ where $IW(S, h)$ is the Inverse Wishart distribution with h degrees of freedom and S is an $(n \times n)$ positive definite matrix.
- $B|\Sigma \sim N(P, \Sigma \otimes A^{-1})$ where $N(P, \Sigma \otimes A^{-1})$ is the matrix-variate Gaussian distribution with h degrees of freedom and S is an $(n \times n)$ positive definite matrix.

Combining the priors and likelihood produce matrix-variate conditional posterior distributions (derivation details provided in Sugita, (2002)):

- Inverse Wishart distribution for $p(\Sigma|\beta, Y) \propto |S_*|^{(t+h)/2} |\Sigma|^{-(t+h+n+1)/2} \exp(-0.5tr(\Sigma^{-1}S_*))$ which is trivial to sample exactly;
- Matrix-variate Gaussian for $p(B|\beta, \Sigma, Y) \propto |A_*|^{n/2} |\Sigma|^{-k/2} \exp(-0.5tr(\Sigma^{-1}(B - B_*)'A_*(B - B_*)))$ (or alternatively matrix-variate student-t distribution form for $p(B|\beta, Y)$), both trivial to sample exactly;
- The marginal matrix-variate posterior for the cointegration vectors, $\beta|Y$, is not well studied and is given by

$$p(\beta|Y) \propto p(\beta)|S_*|^{-(t+h+1)/2}|A_*|^{-n/2}. \quad (7.4.1)$$

where we define $A_* = A + W'W$, $B_* = (A + W'W)^{-1}(AP + W'W\hat{B})$ and $S_* = S + \hat{S} + (P - \hat{B})'[A^{-1} + (W'W)^{-1}]^{-1}(P - \hat{B})$.

7.5 Sampling and Estimation Conditional on Rank r

Here we focus on obtaining samples from the posterior distribution which can be used to obtain Bayesian parameter estimates (MMSE, MAP). The complication in sampling arises with the full conditional posterior 7.4.1 which can not be sampled from via straight forward inversion sampling.

In this paper we outline novel algorithms to sample from the posterior distribution $p(\beta|Y, B, \Sigma, r)$, providing an alternative automated approach to the griddy Gibbs sampler algorithm made popular in this Bayesian co-integration setting by Bauwens and Lubrano (1996).

The matrix-variate griddy Gibbs sampler numerically approximates the target posterior on a grid of values and then performs numerical inversion to obtain samples from 7.4.1 at each stage of the MCMC algorithm. Such a grid based procedure will suffer from the curse of dimensionality when n is large ($n > 5$) after which it becomes highly inefficient. Note, alternative approaches such as Importance Sampling will also be problematic once n becomes too large. It is difficult to optimize the choice of the Importance Sampling distribution which will minimize the variance in the importance weights.

Instead we propose alternative samplers using adaptive matrix-variate MCMC methodology. They do not suffer from the curse of dimensionality and are simple to implement and automate.

- **Algorithm 1 - Random Walk (mixture local & global moves):** Involves an offline adaptively pretuned mixture proposal containing a combination of local and global Random Walk (RW) moves. The proposal for the local RW moves have standard deviation tuned to produce average acceptance probabilities between [0.3, 0.5]. The independent global matrix-variate proposal updates all elements of β via a multivariate Gaussian proposal centered on Maximum Likelihood parameter estimates for β and the Fisher information matrix for the covariance of the global proposal. This is similar to the approach adopted in Vermaak et al. (2004) and Fan et al. (2008).
- **Algorithm 2 - Adaptive Random Walk:** Involves an online matrix-variate adaptive Metropolis algorithm based on methodology presented in Roberts and Rosenthal (2009).

Proceeding sections denote the algorithmic 'time' index by j and the current state of a Markov chain for generic parameter θ at time j by $\theta^{(j)}$. The length of the Markov chain is J .

Note, since we have imposed r^2 restrictions in the form of I_r , any proposal for $\beta = [I_r, \tilde{\beta}]$ will only correspond to the unrestricted elements of β denoted by $\tilde{\beta}$. In our case, these correspond to those in locations $(n - r) \times r$.

7.5.1 Algorithm 1

In Algorithm 1 the mixture proposal distribution for parameters $\tilde{\beta}$ will be given by,

$$q\left(\tilde{\beta}^{(t-1)}, \cdot\right) = w_1 N\left(\tilde{\beta}; \tilde{\beta}^{ML}, \Sigma^{ML}\right) + (1 - w_1) \prod_{i=1}^{(n-r) \times r} N\left(\tilde{\beta}_{i,k}; \tilde{\beta}_{i,k}^{(t-1)}, \sigma_{i,k}^2\right). \quad (7.5.1)$$

The Maximum Likelihood parameters are obtained off-line, see (p. 286 Lutkepohl (2007)). The local random walk proposal variances $\sigma_{i,k}^2$ for each element of $\tilde{\beta}$ are obtained via pre-tuning.

Algorithm 1 MH within Gibbs sampler for fixed rank r via a pretuned mixture of global and local moves.

Input Initial Markov chain state $(\Sigma^{(0)}, B^{(0)}, \beta^{(0)})$.

Output Markov chain samples $\{\Sigma^{(j)}, B^{(j)}, \beta^{(j)}\}_{j=1:J} \sim p(\Sigma, B, \beta|Y)$.

Begin

1. Set initial state $(\Sigma^{(0)}, B^{(0)}, \beta^{(0)})$ deterministically or by sampling the priors.
2. Calculate Maximum Likelihood parameters $\tilde{\beta}^{ML}$ and Σ^{ML} .
3. Initialize w_1 and $w_2 = 1 - w_1$ and index $j = 1$.

Repeat while $j < J$

4. Sample Σ via inversion to obtain $\Sigma^{(j)}$.
5. Sample B via inversion to obtain $B^{(j)}$.
6. Sample realization $U = u$ where $U \sim U[0, 1]$

If $u \geq w_1$ (*perform a local random walk move*)

- 7a. Sample uniformly index (i, k) from set of $n - r \times r$ elements.
- 7b. Sample the (i, k) -th component $\tilde{\beta}_{i,k}^* \sim N(\tilde{\beta}_{i,k}; \tilde{\beta}_{i,k}^{(j-1)}, \sigma_{i,k}^2)$.
- 7c. Construct proposal $\beta^* = [I_{r \times r}, \tilde{\beta}^*]$, where $\tilde{\beta}^*$ is $\tilde{\beta}^{(j-1)}$ with the (i, k) -th element given by $\tilde{\beta}_{i,k}^*$.

else (*perform a global independent move*)

- 7a. Sample proposal $\tilde{\beta}^* \sim N(\tilde{\beta}; \tilde{\beta}^{ML}, \Sigma^{ML})$.
- 7b. Construct proposal $\beta^* = [I_{r \times r}, \tilde{\beta}^*]$.

end

8. Calculate Metropolis Hastings Acceptance Probability:

$$A(\beta^{(j-1)}, \beta^*) = \frac{p(\Sigma^{(j)}, B^{(j)}, \beta^*|Y) q(\beta^* \rightarrow \beta^{(j-1)})}{p(\Sigma^{(j)}, B^{(j)}, \beta^{(j-1)}|Y) q(\beta^{(j-1)} \rightarrow \beta^*)} \quad (7.5.2)$$

Accept $\beta^{(j)} = \beta^*$ via rejection using A , otherwise $\beta^{(j)} = \beta^{(j-1)}$.

9. $j = j + 1$
-

7.5.2 Algorithm 2: Adaptive Metropolis within Gibbs for CVAR model given rank r

There are several classes of adaptive MCMC algorithms, see Roberts and Rosenthal (2009). The distinguishing feature of adaptive MCMC algorithms, compared to standard MCMC, is generation of the Markov chain via a sequence of transition kernels. Adaptive algorithms utilize a combination of time or state inhomogeneous proposal kernels. Each proposal in the sequence is allowed to depend on the past history of the Markov chain generated, resulting in many variants.

Due to the inhomogeneity of the Markov kernel used in adaptive algorithms, it is particularly important to ensure the generated Markov chain is ergodic, with the appropriate stationary distribution. Several recent papers proposing theoretical conditions that must be satisfied to ensure ergodicity of adaptive algorithms include, Atchade and Rosenthal (2005), Roberts and Rosenthal (2009), Haario et al. (2007), Andrieu and Moulines (2006) and Andrieu and Atchade (2007).

Haario et al. (2001) developed an adaptive Metropolis algorithm with proposal covariance adapted to the history of the Markov chain. The original proof of ergodicity of the Markov chain under such an adaption was overly restrictive. It required a bounded state space and a uniformly ergodic Markov chain.

Roberts and Rosenthal (2009) proved ergodicity of adaptive MCMC under simpler conditions known as *Diminishing Adaptation* and *Bounded Convergence*. As in Roberts and Rosenthal (2009) we assume that each fixed kernel in the sequence Q_γ has stationary distribution $P(\cdot)$. Define the convergence time for kernel Q_γ when starting from state θ as $M_\epsilon(\theta, \gamma) = \inf\{j \geq 1 : \|Q_\gamma^j(\theta, \cdot) - P(\cdot)\| \leq \epsilon\}$. Under these assumptions, they derive the sufficient conditions;

- **Diminishing Adaptation:** $\lim_{n \rightarrow \infty} \sup_{\theta \in E} \|Q_{\Gamma_{j+1}}(\theta, \cdot) - Q_{\Gamma_j}(\theta, \cdot)\| = 0$ in probability. Note, Γ_j are random indices.
- **Bounded Convergence:** $\{M_\epsilon(\Theta^{(j)}, \Gamma_j)\}_{j=0}^\infty$ is bounded in probability, $\epsilon > 0$.

which guarantee asymptotic convergence in two senses,

- Asymptotic convergence: $\lim_{j \rightarrow \infty} \|\mathcal{L}(\Theta^{(j)}) - P(\cdot)\| = 0$
- WLLN: $\lim_{j \rightarrow \infty} \frac{1}{j} \sum_{i=1}^j g(\Theta^{(i)}) = \int g(\theta)p(\theta)d\theta$ for all bounded $g : E \rightarrow \mathbb{R}$.

It is non-trivial to develop adaption schemes which can be verified to satisfy these two conditions. We develop a matrix-variate adaptive MCMC methodology in the CVAR setting, using a proposal kernel known to satisfy these two ergodicity conditions for unbounded state spaces and general classes of target posterior distribution, see Roberts and Rosenthal (2009) for details.

In Algorithm 2 the mixture proposal distribution for parameters $\tilde{\beta}$ which is $d = (n - r) \times r$

dimensional and is given at iteration j by,

$$q_j \left(\tilde{\beta}^{(t-1)}, \cdot \right) = w_1 N \left(\tilde{\beta}; \tilde{\beta}^{(t-1)}, \frac{(2.38)^2}{d} \Sigma_j \right) + (1 - w_1) N \left(\tilde{\beta}; \tilde{\beta}^{(t-1)}, \frac{(0.1)^2}{d} I_{d,d} \right). \quad (7.5.3)$$

Here, Σ_j is the current empirical estimate of the covariance between the parameters of $\tilde{\beta}$ estimated using samples from the Markov chain up to time j . The theoretical motivation for the choices of scale factors 2.38, 0.1 and dimension d are all provided in Roberts and Rosenthal (2009) and are based on optimality conditions presented in Roberts et al. (1997) and Roberts and Rosenthal (2001). The adaptive MCMC Algorithm 2 is identical to Algorithm 1 except we replace step 7 with the following alternative;

Algorithm 2: matrix-variate adaptive MH within Gibbs sampler for fixed rank r .

If $u \geq w_1$ (perform an adaptive random walk move)

7a. Estimate Σ_j the empirical covariance of β for elements in $(n - r) \times r$ using samples $\{\tilde{\beta}^{(i)}\}_{i=1:j}$.

7b. Sample proposal $\tilde{\beta}^* \sim N \left(\tilde{\beta}; \tilde{\beta}^{(t-1)}, \frac{(2.38)^2}{d} \Sigma_j \right)$.

7c. Construct proposal $\beta^* = [I_{r \times r}, \tilde{\beta}^*]$.

else (perform a non-adaptive random walk move)

7a. Sample proposal $\tilde{\beta}^* \sim N \left(\tilde{\beta}; \tilde{\beta}^{(t-1)}, \frac{(0.1)^2}{d} I_{d,d} \right)$.

7b. Construct proposal $\beta^* = [I_{r \times r}, \tilde{\beta}^*]$.

end

7.6 Rank Estimation for Bayesian VAR Cointegration models

Here we discuss the Bayes Factor approach to rank estimation, noting that it is computationally inefficient, since it involves running $n+1$ Markov chains, one for each model (rank r). For a sophisticated alternative which presents a novel TD-MCMC based approach, requiring a single Markov chain to obtain samples from the posterior distribution $p(B, \Sigma, \beta, r|Y)$, see Peters *et al.* (2009).

7.6.1 Posterior Model Probabilities for Rank r via Bayes Factors

In Sugita (2002) and Kleibergen and Paap (2002) the rank is estimated via Bayes factors, a popular approach to Bayesian model selection in Bayesian cointegration literature. We note that

alternative approaches to rank estimation include Strachan and van Dijk (2004). Sugita (2002) works with a conjugate prior on α which will not produce a problem with Bartlett's paradox, posterior probabilities of the rank are well defined.

Bayes Factors

The earlier work of Sugita, (2002) compares the rank of the unrestricted α to the 0 rank setting. Note, Kleibergen and Pap (2002) have a slightly different approach in that they compared each rank r to the full rank case for the unrestricted α parameter. Recently, Sugita, (2009) revisits the important question of rank estimation via Bayes Factors also comparing the Schwarz BIC approximation and Chib's (1995) approach for the marginal likelihood.

Under a rank 0 comparison, the posterior model probabilities are given by,

$$Pr(r|Y) = \frac{BF_{r|0}}{\sum_{j=0}^n BF_{j|0}}, \quad (7.6.1)$$

with $BF_{0|0}$ defined as 1.

In the calculation of $BF_{r|0}$, Sugita (2002) recommends an approach first introduced by Verdinelli and Wasserman (1995) for nested model structure Bayes factors, which results in

$$BF_{r|0} = \frac{p(\alpha' = \mathbf{0}_{r \times n})}{C_r^{-1} p(\alpha' = \mathbf{0}_{r \times n} | Y)} = \frac{\int p(\alpha, \beta, \Gamma, \Sigma | Y) d\alpha d\beta d\Gamma d\Sigma}{C_r^{-1} \int p(\alpha, \beta, \Gamma, \Sigma | Y) |_{rank(\alpha)=0} d\alpha d\beta d\Gamma d\Sigma} \quad (7.6.2)$$

where the correction factor for the reduction in dimension C_r is given by,

$$C_r = \int p(\alpha, \beta, \Gamma, \Sigma) |_{rank(\alpha)=0} d\beta d\Gamma d\Sigma. \quad (7.6.3)$$

We note that Sugita (2002) does not comment on numerical complications that can arise when implementing this estimator for the CVAR model. We detail in Appendix 1, Section 7.10 steps that were critical to the calculation of the Bayes Factors when handling potential numerical overflows. The numerical issues arise as t increases, for example the term $|S_*^{(i)}|^{\frac{t+h}{2}}$ will explode numerically. This will result in incorrect numerical results for the Bayes Factors if not handled appropriately.

7.6.2 Model Selection, Model Averaging and Prediction

With samples from $p(\beta, B, \Sigma, r|Y)$ one can consider either model selection or model averaging. In a survey of the literature on rank selection, the most common form of inference performed involves model selection. In this paper we note that model averaging should also be considered, especially when it is probable that given the realized data, two different ranks are highly probable according to their posterior model probabilities. We argue that by adopting the Bayesian model averaging framework one is able to reduce potential model risk associated with selection

of the rank from several choices, which may all be fairly probable under the posterior. This in turn should reduce the associate model risk involved in the popular application of CVAR models in algorithmic trading strategies based on these co-integration frameworks and estimation of the rank.

In this case one can use the samples from $p(\beta, B, \Sigma|Y, r)$ in each model r to form a weighted model averaged estimate through the direct knowledge of the estimated model probabilities given by $p(r|Y)$. There is discussion on model averaging in the CVAR context found in Koop, Strachan, van Dijk and Villani (2006).

Bayesian Model Order Selection (BMOS)

In BMOS we select the most probable model corresponding to the maximum *a posteriori* (MAP) estimate from $p(r|Y)$, denoted r_{MAP} . Conditional on r_{MAP} , we then take the samples of $\{\beta^{(j)}, B^{(j)}, \Sigma^{(j)}\}_{j=1:M}$ corresponding to Markov chain simulated for the r_{MAP} model and we estimate point estimates for the parameters.

These point estimates typically include posterior means or modes, though one should be careful. We note that it was demonstrated by Kleiberger and van Dijk (1994) or Bauwens and Lubrano (1996) that in many popular CVAR Bayesian models, certain choices of prior result in a proper posterior yet it may not have finite moments of any order. Some alternatives are proposed by Strachan and Inder (2004).

Bayesian Model Averaging (BMA)

In this section we consider the problem of estimating for example an integral of a quantity or function of interest, $\phi(\{\beta, B, \Sigma\})$, with respect to the posterior distribution of the parameters, e.g. moments of the posterior. Since we have chosen to work with a posterior distribution $p(\beta, B, \Sigma, r|Y)$ we can estimate this integral quantity whilst removing the model risk associated with rank uncertainty. This is achieved by approximating

$$\sum_{r=1}^n \int \phi(\{\beta, B, \Sigma|r\})p(\beta, B, \Sigma|Y, r)p(r|Y)d\beta dB d\Sigma \approx \sum_{r=1}^n \sum_{j=1}^M \phi(\{\beta^j, B^j, \Sigma^j|r^j\})p(r^j|Y). \quad (7.6.4)$$

Prediction Incorporating Model Risk

Here we perform prediction whilst removing model uncertainty related to the rank. This is possible under a Bayesian Model Averaging (BMA) framework using,

$$p(Y^*|Y) = \sum_{r=1}^n \int p(Y^*|\beta, B, \Sigma, r)p(\beta, B, \Sigma|Y, r)p(r|Y)d\beta dB d\Sigma. \quad (7.6.5)$$

We will compare the predictive performance of the MMSE estimate or mean of the estimated distribution for $p(Y^*|Y)$ under the BMA versus BMOS approach which involves,

$$p(Y^*|Y) = \int p(Y^*|\beta, B, \Sigma, \hat{r}^{MAP})p(\beta, B, \Sigma|Y, \hat{r}^{MAP})d\beta dB d\Sigma. \quad (7.6.6)$$

7.7 Simulation Experiments

Analysis of the methodology developed is in three parts: the first part contains simulations performed on synthetic data sets, comparing performance of the proposed model sampling methodology; the second part contains two real data set examples; and the third part involves analysis of predictive performance BMOS and BMA using real data.

7.7.1 Synthetic Experiments

In this section the intention will be to develop a controlled setting in which the true model parameters are known and the data is generated from the true model. This will allow us to assess performance of each of the proposed estimation procedures. In doing this we take an identical model to the simple model studied in Sugita (2002; 2009) [p.4] for our analysis.

Analysis of samplers

The first analysis is to compare the performance of the two adaptive samplers. To achieve this we generate 20 realizations of data sets of length $T = 100$ from the rank $r = 2$ model. Then conditional on knowledge of the rank $r = 2$ we sample $J = 20,000$ samples from the joint posterior $p(B, \Sigma, \beta | Y, r = 2)$ and discard the first 10,000 samples as burnin. We perform this analysis for each of the data realizations under both of the proposed samplers, Algorithm 1 and Algorithm 2, and then we present average MMSE estimates and average posterior standard deviations from each sampler in Table 1. In particular we present the averaged posterior point estimates for: the unrestricted β parameters; the average trace of the posterior estimate of the covariance Σ ; the average of each of the intercept terms; and the averaged first element of the unrestricted α .

Note, the pre-tuning of the local random walk proposal standard deviation for Algorithm 1 is performed offline using an MCMC run of length 20,000. Additionally, the prior parameters were set to be: for $B|\Sigma$ the prior parameters were set as $P = (\widehat{W}'\widehat{W})^{-1} \widehat{W}Y$, $A = \lambda (\widehat{W}'\widehat{W}) / T$ with $\lambda = 1$, $\widehat{W} = (XZ\widehat{\beta})$ and $\widehat{\beta} = [I_r, \mathbf{0}]$; for β the prior parameters were set as $E[\beta] = (I_r, \mathbf{0})$, $Q = I_n$, $H = \tau Z'Z$ and $\tau = 1/T$; for Σ the prior parameters were set as $S = \tau Y'Y$ and $h = n + 1$.

These results demonstrate that both Algorithm 1 and Algorithm 2 perform well. The MMSE estimates produced by both algorithms are accurate compared to the true parameter values used to generate the data. Algorithm 1 which involved the mixture of pretuned local moves and a Global move centered on the Maximum Likelihood parameter estimates required more computational effort than the adaptive MCMC approach of Algorithm 2. Additionally, we point out that as discussed in Rosenthal (2008), the sampler we developed in Algorithm 2 actually achieves optimal performance as $n \rightarrow \infty$. Therefore it will be a far superior algorithm to the griddy Gibbs sampler approach which will not be feasible in high dimensions. Hence, for an automated and computationally efficient alternative to the griddy Gibbs sampler typically used

we would recommend the use of Algorithm 2. In the following studies, we utilize Algorithm 2, the adaptive MCMC algorithm. To conclude, we also present the trace plots of the sample paths under the adaptive MCMC algorithm, see Figure 1. This plot demonstrates that rapid convergence of the MMSE estimates of the parameters in the posterior, even when initialized far from the true values. Additionally, one can see the behaviour of the adaptive proposal, learning the appropriate proposal variance.

Analysis of Adaptive MCMC sampler in high dimension.

In this example, we work with the Adaptive MCMC algorithm we developed for the Bayesian CVAR model. In particular we consider the case in which $n = 10$, which is a setting in which the standard approach of the gridy Gibbs sampler will become excessively computational, due to the curse of dimensionality, since there are now several hundred parameters to be sampled from the posterior.

All coefficients except for the cointegrating vectors are generated by uniform distributions with a range between -0.4 to 0.4, and the error covariance was set to the identity. We generate a realizations of data of length $T = 100$ from the true rank $r = 5$ model in which the cointegration vector has all terms in the matrix of β which are unrestricted set to be 0, other than the last row, which is -1. Then conditional on knowledge of the rank $r = 5$ we sample $J = 20,000$ samples from the joint posterior $p(B, \Sigma, \beta | Y, r = 5)$ and discard the first 10,000 samples as burnin.

The sample paths of the cointegration vector parameters randomly selected to be presented were $\beta_{10,1}, \beta_{10,4}$ which are shown in Figure 2. Clearly, again in this high dimensional setting (310 dimensions), the adaptive MCMC algorithm performs suitably. Even, though the Markov chain is initialised far from the true parameter values of cointegration vector, we see the rapid convergence of our sampler. This is illustrated for the two arbitrarily selected parameters which had true values of of -1 and -1. Note, in this high dimensional setting, the algorithm was implemented in Matlab and took only 132sec to complete the simulation on an Intel Core 2 Duo at 2.40GHz, with 3.56Gb of RAM.

Analysis of model selection in the Bayesian CVAR model

In this section we study on synthetic data the performance of the Bayes Factor estimator applied to estimate posterior model probabilities for the rank. To perform this analysis we consider the model from Sugita, (2002; 2009 [p.4]) and we take data series of length $T = 100$ and we simulate 50 independent data realizations for each possible model rank $r = 1, \dots, 4$. Then for each rank r we count the number of times each model is selected as the MAP estimate out of the total of the 50 simulations, one simulation per generated data set. Note, the algorithm was run for 20,000 iterations with 10,000 samples used as burnin. The results of this analysis are presented in Table 2.

We note that the results of this section demonstrated the following interesting properties:

1. When the true rank used to generate the observations data was small, the BF methodology was clearly able to detect the true model order as the MAP estimate in a high proportion of the tested data sets.
2. In all cases the averaged actual posterior model probabilities were very selective of the correct model, indicating that at least under this synthetic data scenario, there would not be great benefit in performing model averaging. However, we will demonstrate later examples with actual data in which there is significant ambiguity between possible model ranks, in these cases we also study the model averaging results.

7.7.2 Financial Example 1 - US mini indexes

Having assessed the proposed algorithms developed in this paper for synthetic data generated from a CVAR model, we now work with a practical financial example. In this example we will consider data series comprised of US indexes S&P mini, Nasdaq mini and Dow Jones mini. The data obtained for each of these data series consists of 774 values corresponding to the close of market daily price from the 31-Aug-2005 through to 30-Sep-2008. The time series data is presented in Figure 3.

We analyze this data using Algorithm 2 (adaptive MCMC) and estimate the rank via Bayes Factor analysis, the results are presented in Table 3. We run 20 independent samplers with different initializations, for each possible rank. This is performed for each data set, and the total series is split into increasing subsets, each taking subsets of the data from 50 data points through to 400 data points, in increases of 50 data points. This allows us to study the change in the estimated rank as a function of time for each of these time series. Clearly, if the true rank of our model was fixed, then as the total amount of data we include increases, then we should see the posterior model probability of the rank converge to 1 for one of the possible ranks. What we observed after doing this analysis was that there was a clear variability in the predicted rank as we included more data. In particular the model estimates showed preference most often to rank 1, suggesting that 2 common stochastic trends are present in the series. Additionally, the fact that in several cases, the model is less likely to distinguish between rank 1 and 2, suggests it may be prudent to also perform a model averaging analysis. Especially in the popular application of CVAR models in practice to perform algorithmic trading.

7.7.3 Financial Example 2 - US notes

Here we repeat the same procedure performed in Financial Example 1, for a different data set. This time we consider data series comprised of Bond data for US 5 year, 10 year and 30 year notes over the same time period as the US mini index data. The time series data is presented in Figure 4.

We analyze this data using Algorithm 2 (adaptive MCMC) and estimate the rank via Bayes Factor analysis, the results are presented in Table 4. We set up this second data analysis in the same

way as Financial Examp1 1, with 20 independent samplers, each with different initializations, for each possible rank. This allows us to study the change in the estimated rank as a function of time for each of these time series. Again, we observed that with this data, the model gave preference most often to rank 1, suggesting that 2 common trends are present in the series we are analyzing. However, there was much stronger evidence for a single co-integrating relationship over time in this data, compared to the analysis of the US mini index data over the same period. This suggests that the US bond data series is a more stable series to fit the CVAR model too when assuming a constant number of co-integrating relationships over time.

7.7.4 Financial Example 3

In this section we perform a predictive performance comparison using Bayesian Model Selection versus Averaging. We take 2 series for the US bonds, 5 years and 10 years, and we combine these series over the same period with the S& P 500 mini index. We compare the MMSE estimate of the predicted series over 10 steps ahead which is obtained from the distribution of the predicted data $p(Y^*|Y)$, after we have integrated out parameter and rank uncertainties. We demonstrate that in this actual data example, the performance obtained by Bayesian Model Averaging represents the uncertainty in the prediction more accurately than the Bayesian Model Order Selection setting.

This study is performed as follows. We begin by selecting randomly, with replacement, 100 segments of the vector time series, each containing 50 days of data. For each segment of the time series we fit our Bayesian model for each possible rank, also estimating via Bayes Factors the posterior model probability for each rank. Then we calculate the predictive posterior mean, corresponding to the MMSE estimate of the predicted data series for the following 5 days, Y^* . Finally, we take the squared difference between the actual data series over the proceeding 10 days post the 50 days for the given segment and the posterior mean of the predicted data Y^* .

In Figure 5 we present for each prediction day a histogram of the squared difference between the actual data over the random sets of 5 days and the predictive posterior MMSE estimators for the same 5 days. We compare here the performance under Bayesian model selection and averaging. When performing Bayesian model averaging we are integrating out uncertainty in the prediction due to the prediction of the unknown rank.

Clearly, the Bayesian model averaging approach will result in a greater uncertainty in the prediction when compared to the Bayesian model selection. This is reflected especially in the distribution of the prediction at 5 days where the model averaging approach box-whisker plot covers a noticeably wider range than the model selection equivalent. Though not presented here, we also assessed and confirmed this would occur out to longer predictions of 10 days and 20 days.

7.8 *Conclusions*

We have developed and demonstrated how one can utilize state of the art adaptive MCMC methodology to solve a challenging high dimensional econometrics problem based on cointegrated vector autoregressions. The challenging application involved a posterior distribution which was matrix-variate and very high dimensional. We compared the performance of the Adaptive Metropolis algorithm with an alternative based on a mixture proposal of local and global moves centred on the the Maximum Likelihood parameters. We then formulated the rank estimation in as a Bayesian model selection problem and performed analysis of the Bayes factors using our adaptive MCMC algorithm. We concluded with analysis of real market data and performed Bayesian model selection and model averaging, with respect to the unknown rank. In conclusion, the adaptive MCMC methodology developed clearly allowed us to extend significantly the dimension of the estimation problem in the Bayesian CVAR literature. It was shown to be highly efficient and accurate.

From the perspective of developing a Bayesian CVAR model for algorithmic trading we found that historically the US bond data we considered is a more stable series to fit the CVAR model too when assuming a constant number of co-integrating relationships over time. This will therefore impact the stability of trading performance under such models. In addition when considering trading triples made up of the US bond data series and the S&P mini index, it is beneficial to perform Bayesian model averaging for the rank, rather than just selecting the most probable co-integration rank. The adaptive MCMC based framework allows this to be done efficiently and in an automated fashion.

7.9 *Acknowledgements*

We would like to thank Prof. Robert Kohn for useful feedback and also we thank the two anonymous referees and associate editor for their very helpful comments which have significantly improved the exposition of the paper. The first author thanks the Statistics Department of the University of NSW for financial support and Boronia Managed Funds.

7.10 Appendix 1

We begin by calculating the log posterior model probabilities,

$$\log (Pr (r|Y)) = \log (BF_{r|0}) + \log (BF_{max|0}) - \log \left(\sum_{j=0}^n \exp (\log (BF_{j|0}) - \log (BF_{max|0})) \right), \quad (7.10.1)$$

where $BF_{max|0} = \max\{BF_{0|0}, \dots, BF_{n|0}\}$. Additionally, we now consider the log of the Bayes Factor for rank r and we apply the same numerical trick.

$$\log (BF_{r|0}) = \log (p(\boldsymbol{\alpha}' = 0_{r \times n})) + \log (C_r) - \log (p(\boldsymbol{\alpha}' = 0_{r \times n}|Y)) \quad (7.10.2)$$

Now, considering each of the terms:

- $\log (p(\boldsymbol{\alpha}' = 0_{r \times n})) = -\frac{nr}{2} \log (\pi) + \frac{h}{2} \log (|S|) + \frac{n}{2} \log (|A_{22.1}|) + \sum_{j=1}^n \log \left(\frac{\Gamma(\frac{h+r+1-j}{2})}{\Gamma(\frac{h+1-j}{2})} \right) - \frac{h+r}{2} \log (|S|)$
- $\log (p(\boldsymbol{\alpha}' = 0_{r \times n}|Y)) = -\log (N) + \log (L_{max}^{(1)}) - \log \left(\exp \left(\sum_{i=1}^N \log (L_i^{(1)}) - \log (L_{max}^{(1)}) \right) \right)$,
where $L_i^{(1)} = \pi^{-\frac{nr}{2}} |S_*^{(i)}|^{\frac{t+h}{2}} |A_{*22.1}^{(i)}|^{\frac{n}{2}} \prod_{i=1}^n \frac{\Gamma(\frac{t+h+r+1-i}{2})}{\Gamma(\frac{t+h+1-i}{2})} |S_*^{(i)} + B_{*2}^{(i)'} A_{*22.1}^{(i)} B_{*2}^{(i)}|^{-\frac{t+r}{2}}$ and $L_{max}^{(1)} = \max\{L_1^{(1)}, \dots, L_N^{(1)}\}$.
- $\log (C_r) = -\log (N) + \log (L_{max}^{(2)}) \log \left(\exp \left(\sum_{i=1}^N \log (L_i^{(2)}) - \log (L_{max}^{(2)}) \right) \right)$,
where $L_i^{(2)} = \frac{p(\boldsymbol{\alpha}=0, \Gamma^{(i)}|\Sigma^{(i)})}{p(\Gamma^{(i)}|\Sigma^{(i)})}$ and $L_{max}^{(2)} = \max\{L_1^{(2)}, \dots, L_N^{(2)}\}$. Note this sum evaluated using samples from the Markov chain run in model r where, $p(\boldsymbol{\alpha} = 0, \Gamma^{(i)}|\Sigma^{(i)})$ and $p(\Gamma^{(i)}|\Sigma^{(i)})$ are obtained using knowledge of the specified prior, $p(B|\Sigma) = p(\Gamma, \boldsymbol{\alpha}|\Sigma) = p(\mu, \Psi_{1:p-1}, \boldsymbol{\alpha}|\Sigma)$.

| Parameter Estimates | Algorithm 1 | Algorithm 2 | Truth |
|---|----------------|----------------|-------|
| Ave. MMSE $\beta_{1,r+1}$ | -0.002 (0.001) | -0.034 (0.002) | 0 |
| Ave. Posterior Stdev. $\beta_{1,r+1}$ | 0.018 (0.006) | 0.010 (0.003) | - |
| Ave. MMSE $\beta_{2,r+1}$ | -0.819 (0.051) | -0.862 (0.045) | -1 |
| Ave. Posterior Stdev. $\beta_{2,r+1}$ | 0.032 (0.005) | 0.020 (0.003) | - |
| Ave. MMSE $\beta_{1,n}$ | 0.033 (0.025) | -0.024 (0.023) | 0 |
| Ave. Posterior Stdev. $\beta_{1,n}$ | 0.030 (0.012) | 0.026 (0.010) | - |
| Ave. MMSE $\beta_{2,n}$ | -0.752 (0.098) | -0.774 (0.082) | -1 |
| Ave. Posterior Stdev. $\beta_{2,n}$ | 0.038 (0.013) | 0.028 (0.006) | - |
| Ave. Mean acceptance probability β | 0.352 (0.010) | 0.232 (0.029) | - |
| Ave. MMSE $\text{tr}(\Sigma)$ | 4.945 (0.331) | 4.432 (0.332) | 4 |
| Ave. Posterior Stdev. $\text{tr}(\Sigma)$ | 0.420 (0.049) | 0.416 (0.048) | - |
| Ave. MMSE μ_1 | 0.07 (0.051) | 0.065 (0.043) | 0.1 |
| Ave. Posterior Stdev. μ_1 | 0.236 (0.028) | 0.226 (0.026) | - |
| Ave. MMSE μ_2 | -0.027 (0.041) | -0.034 (0.024) | 0.1 |
| Ave. Posterior Stdev. μ_2 | 0.183 (0.041) | 0.181 (0.010) | - |
| Ave. MMSE μ_3 | -0.080 (0.084) | -0.061 (0.045) | 0.1 |
| Ave. Posterior Stdev. μ_3 | 0.199 (0.020) | 0.187 (0.015) | - |
| Ave. MMSE μ_4 | 0.024 (0.049) | 0.030 (0.029) | 0.1 |
| Ave. Posterior Stdev. μ_4 | 0.184 (0.010) | 0.185 (0.011) | - |
| Ave. MMSE $\alpha_{1,1}$ | -0.223 (0.015) | -0.224 (0.016) | -0.2 |
| Ave. Posterior Stdev. $\alpha_{1,1}$ | 0.070 (0.006) | 0.068 (0.005) | - |
| Ave. MMSE $\alpha_{1,2}$ | 0.201 (0.013) | 0.202 (0.013) | 0.2 |
| Ave. Posterior Stdev. $\alpha_{1,2}$ | 0.053 (0.002) | 0.052 (0.002) | - |

Tab. 7.1: **Sampler Analysis** - Algorithm 1 is the pretuned mixture proposal of Global ML move and local pretuned MCMC move; Algorithm 2 is the Global adaptively learnt MCMC proposal. Averages and a standard error are taken for the Bayesian point estimators over 20 data sets, the standard errors are presented in brackets (\cdot). Note in all simulations the initial Markov chain is started very far away from the true parameter values.

| Model Rank | Bayes Factors |
|---------------------------|----------------------|
| $r = 0$ | 3 (0.84) |
| $r = 1$ | 16 (0.93) |
| $r = 2$ | 2 (0.92) |
| $r = 3$ | 0 (-) |
| $r = 4$ | 0 (-) |
| $r = 0$ | 0 (-) |
| $r = 1$ | 5 (0.89) |
| $r = 2$ | 13 (0.91) |
| $r = 3$ | 0 (-) |
| $r = 4$ | 2 (0.92) |
| $r = 0$ | 0 (-) |
| $r = 1$ | 0 (-) |
| $r = 2$ | 4 (0.89) |
| $r = 3$ | 6 (0.90) |
| $r = 4$ | 10 (0.94) |
| $r = 0$ | 0 (-) |
| $r = 1$ | 0 (-) |
| $r = 2$ | 0 (-) |
| $r = 3$ | 2 (0.87) |
| $r = 4$ | 18 (0.89) |

Tab. 7.2: Between Model Analysis - The true model rank used to generate the data is presented in bold. TDMCMC is the Trans-dimensional Markov chain Monte Carlo algorithm utilizing adaptive MH within model moves and the global Independent between model moves. The results represent the total number of times a given rank is selected as the MAP estimate out of the 20 independent data sets, each of length $T=100$, analyzed. Additionally, the average posterior model probability for these cases is presented in brackets.

| Rank \ T | 50 | 100 | 150 | 200 | 250 | 300 | 350 | 400 |
|----------|---------------|--------------|---------------|---------------|---------------|---------------|------------------------------|-------------------------------|
| $r = 0$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $r = 1$ | 8.09 (0.78) | 3.77 (0.29) | 7.01 (0.50) | 11.51 (0.90) | 1.69 (0.54) | 2.71 (3.55) | 3.14 (1.11) | 7.77 (0.83) |
| $r = 2$ | 2.91 (1.24) | 2.33 (1.26) | 4.61 (0.63) | 25.36 (7.19) | -5.33 (1.17) | -5.80 (0.97) | 4.92 (1.06) | -3.88 (1.11) |
| $r = 3$ | -26.03 (1.06) | -8.45 (0.27) | -37.25 (1.08) | -55.79 (1.70) | -14.61 (0.03) | -62.60 (3.31) | $8.88 (0.88) \times 10^{-3}$ | $-2.06 (2.48 \times 10^{-2})$ |

Tab. 7.3: **Log Bayes Factors:** Analysis of VAR series of US mini indexes as a function of data size. Average log Bayes Factors and standard deviation of log Bayes Factors over 20 independent Markov chains each of chain length 20,000.

| Rank \ T | 50 | 100 | 150 | 200 | 250 | 300 | 350 | 400 |
|----------|----------------|---------------|---------------|----------------|-------------------------------|---------------|----------------|----------------|
| $r = 0$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $r = 1$ | 4.81 (1.00) | 3.14 (0.43) | 5.36 (1.06) | 5.92 (0.84) | 3.32 (0.75) | 1.30 (0.37) | 7.30 (0.59) | 3.10 (0.48) |
| $r = 2$ | -1.67 (12.78) | 3.66 (3.87) | -3.75 (3.04) | -1.83 (2.61) | -6.02 (3.22) | 0.14 (6.51) | -2.93 (1.96) | -7.73 (2.46) |
| $r = 3$ | -42.44 (12.38) | -48.58 (2.85) | -33.12 (0.14) | -100.42 (4.82) | $-25.91(6.52 \times 10^{-2})$ | -10.33 (0.72) | -142.89 (3.31) | -195.47 (4.71) |

Tab. 7.4: **Log Bayes Factors:** Analysis of VAR series of US Bonds (5,10,30 Year Notes) as a function of data size. Average log Bayes Factors and standard deviation over 20 independent Markov chains each of chain length 20,000.

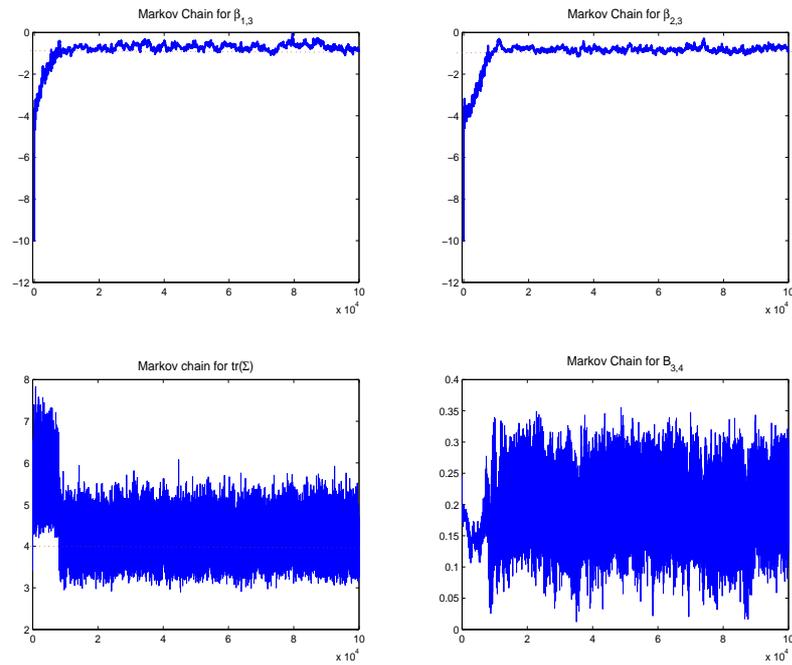


Fig. 7.10.1: Sample paths for posterior parameters, using 100 data points, true rank of $r = 2$ known and an adaptive MCMC algorithm.

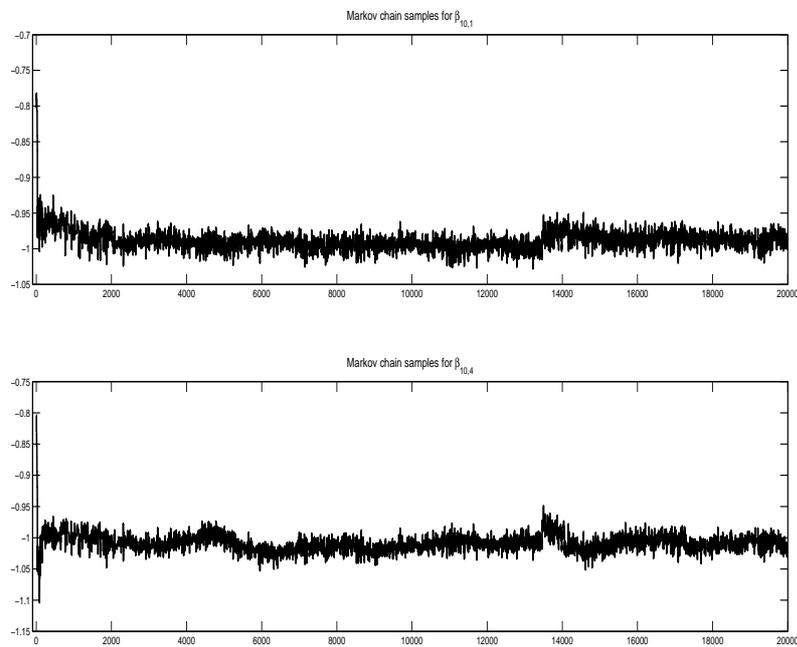


Fig. 7.10.2: Sample paths for posterior parameters, using 100 data points, true rank of $r = 5$ known and an adaptive MCMC algorithm.

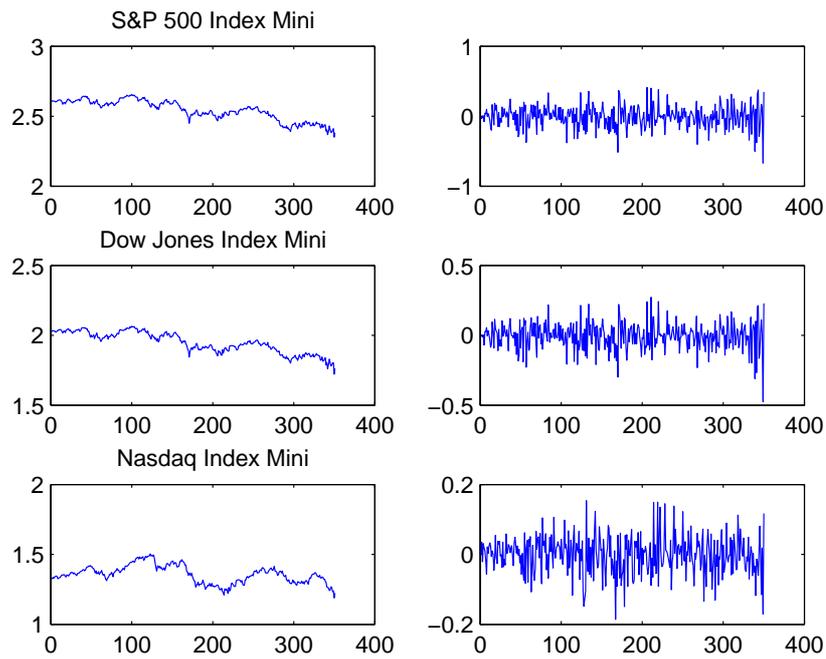


Fig. 7.10.3: S&P 500, Dow Jones and Nasdaq mini Index daily close price data between 01-May-08 to 18-Sep-08. Left column plots represent scaled raw prices; Right plots represent difference data series.

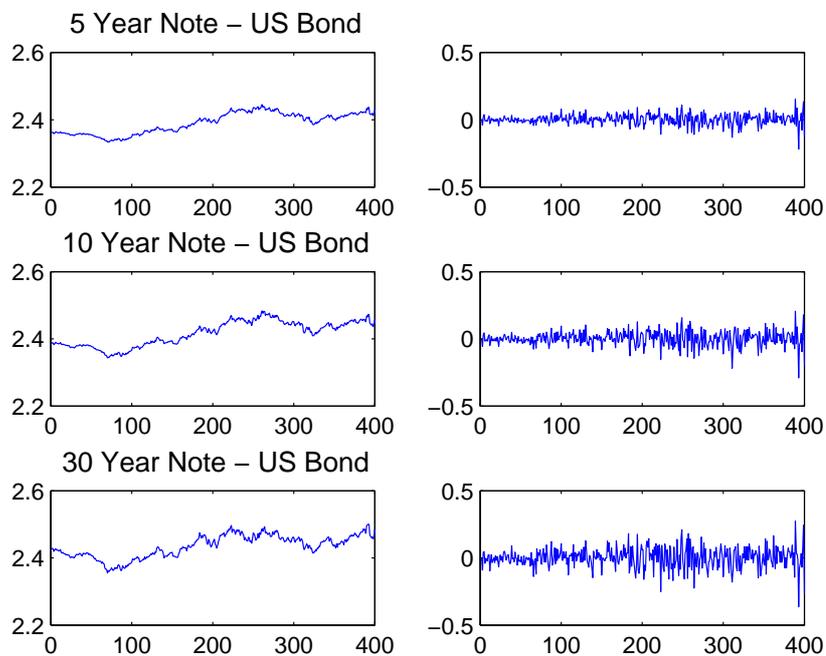


Fig. 7.10.4: 5, 10, 30 Year Notes - daily close price data between 01-May-08 to 18-Sep-08. Left column plots represent scaled raw prices; Right plots represent difference data series.

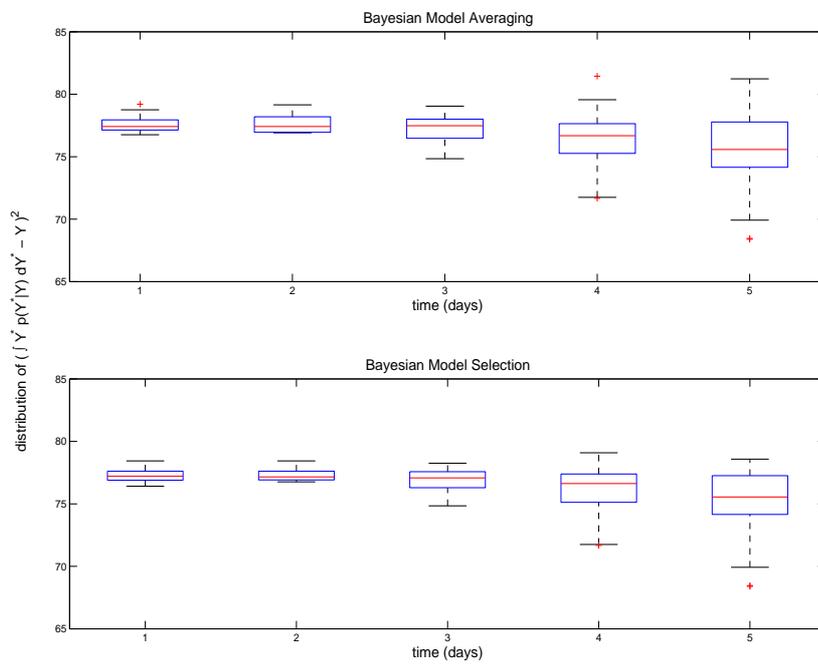


Fig. 7.10.5: Empirical distribution of the Bayesian Model Averaging and Bayesian Model Order Selection, predictive performance for a combination of mini-index and bond data, taken over random intervals.

References

- [1] Ackert, L., Racine, M.D. (1998). "Stochastic trends and cointegration in the market for equities". *Federal Reserve Bank of Atlanta, Working Paper* 98-13.
- [2] Andrieu, C., Moulines, E. (2006). "On the ergodicity properties of some adaptive MCMC algorithms". *Annals of Applied Probability*, 16, (3).
- [3] Andrieu, C., Atchade, Y. (2007). "On the efficiency of adaptive MCMC algorithms". *Electronic Communications in Probability*, 12, 336-349.
- [4] Atchade, Y.F., Rosenthal, J.S. (2005). "On adaptive Markov chain Monte Carlo algorithms". *Bernoulli*, 11(5), 815-828.
- [5] Bauwens, L., Lubrano, M., Richard, J.F. (1999). "Bayesian Inference in Dynamic Econometric Models: Advanced Texts in Econometrics". Oxford University Press.
- [6] Bauwens, L., Lubrano, M. (1996). "Identification restrictions and posterior densities in cointegrated Gaussian VAR systems". *Advances in Econometrics* 11, Part B, (JAI Press, Greenwich) 3-28.
- [7] Bauwens, L., Giot, P. (1997). "A Gibbs sampling approach to cointegration". *Computational Statistics*, 13, 339-368.
- [8] Chib, S. (1995). "Marginal likelihood from the Gibbs output". *Journal of the American Statistical Association*, 90(432), Theory and Methods, 1313-1321.
- [9] Engle, R.F., Granger, C.W.J. (1987). "Co-Integration and Error Correction: Representation, Estimation, and Testing". *Econometrica*, 55(2), 251-276.
- [10] Fan, Y., Peters, G.W., Sisson, S.A. (2009). "Automating and evaluating reversible jump MCMC proposal distributions". *Statistics and Computing*, 19, 409-421.
- [11] Geweke J. (1996). "Bayesian reduced rank regression in econometrics". *Journal of Econometrics*, 75, 121-146.
- [12] Giordani, P. and Kohn, R. (2006). "Adaptive independent Metropolis-Hastings by fast estimation of mixtures of normals." Preprint.
- [13] Granger, C.W.J. (1981). "Some properties of time series data and their use in econometric model specification". *Journal of Econometrics*, 121-130.

- [14] Granger, C.W.J., Weiss, A.A. (1983). "Time series analysis of error-correcting models". *Studies in econometrics, time series, and multivariate statistics*. New York: Academic Press, 255-278.
- [15] Haario, H., Saksman, E., Tamminen, J. (2007). "Componentwise adaptation for high dimensional MCMC". *Computational Statistics*, 20(2).
- [16] Haario, H., Saksman, E., Tamminen, J. (2001). "An adaptive metropolis algorithm". *Bernoulli*, 7, 223-242.
- [17] Kleibergen, F., Paap, R. (2002). "Priors, posteriors and Bayes factors for a Bayesian analysis of cointegration". *Journal of Econometrics*, Elsevier, 111(2), 223-249.
- [18] Kleibergen, F., van Dijk, H.K. (1994). "On the Shape of the Likelihood/Posterior in Cointegration Models". *Econometric Theory*, Cambridge University Press, 10(3-4), 514-551.
- [19] Koop, G., Strachan, R., van Dijk, H., Villani, M. (2006). "Bayesian Approaches to Cointegration". In T.C.Mills and K. Patterson (Ed.), *Palgrave Handbook of Econometrics Volume 1 Econometric Theory 1st ed.* (pp. 871-898) UK: Palgrave Macmillan.
- [20] Krolzig, H. M. (1996). "Statistical analysis of cointegrated VAR processes with Markovian regime shifts". *Institute of Economics and Statistics, Oxford Technical Report*.
- [21] Lütkepohl, H. (2007). *New Introduction to Multiple Time Series Analysis*. Springer
- [22] Peters, G.W., Kannan, B.K., Lasscock, B., Mellen, C. (2009). "Trans-dimensional and adaptive Markov chain Monte Carlo for Bayesian co-integrated VAR models". Technical Report: University of NSW, Statistics Department.
- [23] Reinsel, G.C., Velu, R.P. (1998). *Multivariate Reduced-Rank Regression, Theory and Applications*. Lectuer Notes in Statistics Springer.
- [24] Roberts, G.O., Gelman, A., Gilks, W.R. (1997). "Weak convergence and optimal scaling of random walk Metropolis algorithms". *Annals of Applied Probability*, 7, 110-120.
- [25] Roberts, G.O., Rosenthal, J.S. (2009). "Examples of Adaptive MCMC". *Journal of Computational and Graphical Statistics*, 18(2), p.349-367.
- [26] Roberts, G.O., Rosenthal, J.S. (2001). "Optimal scaling for various Metropolis-Hastings algorithms". *Statistical Science*, 16, 351-367.
- [27] Rosenthal, J.S. (2008). "Optimal Proposal Distributions and Adaptive MCMC". Chapter for MCMC Handbook, Brooks S., Gelman A., Jones G. and Meng X.L. eds.
- [28] Silva, R., Giordani, P., Kohn, R., Pitt, M. (2009). "Particle filtering within adaptive Metropolis Hastings sampling." Arxiv preprint arXiv:0911.0230.
- [29] Strachan, R.W., van Dijk, H.K. (2004). "Bayesian model selection with an uninformative prior". *Keele Economics Research Papers*
- [30] Strachan, R.W., Inder, B. (2004). "Bayesian analysis of the error correction model". *Journal of Econometrics*, 123(2), 307-325.
- [31] Sugita, K. (2002). "Testing for cointegration rank using Bayes factors". *Royal Economic Society Annual Conference*, Royal Economic Society.

-
- [32] Sugita, K. (2009). "A Monte Carlo comparison of Bayesian testing for cointegration rank". *Economics Bulletin*, 29(3), 2145-2151.
- [33] Vermaak, J., Andrieu, C., Doucet, A., Godsill, S.J. (2004). "Reversible jump Markov chain Monte Carlo strategies for Bayesian model selection in autoregressive processes". *Journal of Time Series Analysis*, 25(6), 785-809.
- [34] Villani, M. (2005). "Bayesian reference analysis of cointegration". *Econometric Theory*, (21) 326-357.

Part II

ADVANCES IN BAYESIAN FINANCIAL RISK AND NON-LIFE INSURANCE MODELS

METHODOLOGY AND APPLICATION

Part II Introduction

“Don’t judge each day by the harvest you reap, but by the seeds you plant.”

Robert Louis Stevenson

The aim of this chapter is to provide context and background for the development of Bayesian models for the areas of financial modelling of OpRisk and non-life insurance claims reserving. The technical details relating to likelihood-free methodology and TDMCMC samplers in these contexts is deferred to the journal papers presented in the subsequent chapters of Part II. This chapter establishes the context and background related to the need to develop advanced Bayesian modelling and sampling methodology in order to enhance two areas of financial mathematics. Two key practical motivations discussed involve recent regulatory standards Basel II and Sarbanes-Oxley. This section starts by motivating OpRisk modelling and then it considers claims reserving in non-life insurance actuarial modelling.

The specific structure of this chapter involves three sections. The first section presents relevant background, literature review and contextual understanding to the statistical modelling and methodological developments undertaken in the journal papers in Part II, in the context of Operational risk. The second section of this introduction chapter for Part II presents relevant background for the need to develop more sophisticated statistical models and sampling methodology in the context on non-life insurance actuarial claims reserving. The third section then summarizes the novelty and contribution introduced in each of the journal papers contained in Part II of the thesis.

8.1 Operational risk modelling and Basel II

The practical aspects of Part II of the thesis are developed as a result of five years of work as a quantitative analyst in the Banking industry in Australia. This section utilizes aspects of a commissioned report, by the Australian Centre of Excellence for Risk Analysis (46), in to

the state of OpRisk modelling in the Australian financial industry. This provides important regulatory background to understand the statistical challenges involved in modelling OpRisk in the context of the regulatory guidelines and rules. The Australian banking sector is unique as it leads the world in development and implementation of OpRisk statistical models in practice. Some of these statistical developments can be found in the second chapter of Part II of this thesis.

8.1.1 Executive Summary: quantifying bank Operational Risk

The modelling of OpRisk has taken a prominent place in financial quantitative measurement, this has occurred as a result of Basel II regulatory requirements. The basic framework of Basel II is summarized in Figure 8.1.1.

This report details the modelling of extreme and rare events in the context of OpRisk, with particular focus on the Australian financial sector. Initially the regulatory environment in banking within Australia is discussed in the context of OpRisk. There is a focus on quantification requirements for OpRisk and why such quantification is important in relation to regulatory standards. Definitions and discussion of OpRisk and the associated risk categories introduced by Basel are detailed.

Then the different methodological and modelling approaches, that are allowed for in the regulatory guidelines, are discussed, from the most basic to most advanced approaches, including associated regulatory requirements. This is followed by an overview of the different phases that may be required to implement such a methodological and cultural change in a financial institution, embarking on modelling OpRisk.

Following this is a section, which discusses the issues and difficulties associated with modelling OpRisk. In particular aspects of OpRisk, which make modelling difficult at the most fundamental level, are detailed. A strong focus in this section involves discussion regarding the different forms of data, that can be incorporated in OpRisk quantification. This includes an overview of issues associated with data collection and the analysis of data prior to modelling. This will provide strong motivation for several of the Bayesian modelling papers contained in the second chapter of Part II which specifically develop statistical models to combine each of the data sources present in an OpRisk context.

Proceeding this is a section addressing the industry standard modelling framework, Loss Distributional Approach LDA. This section includes discussion of popular statistical models utilized to model the annual loss distributions of risk profiles that fall under the banner of OpRisk. To finish off the discussion of quantitative approaches, a section on statistical models and methodology, for particular data sources, is presented.

Finally a section on the management aspects of OpRisk is presented, which relates the regulatory requirements for dealing with assessed and modelled risk profiles.

8.1.2 Background and context within Australia's financial industry

In January 2001 the Basel Committee on Banking Supervision proposed a New Basel Accord known as Basel II, which was to replace the 1988 Capital Accord. This proposal considers three pillars, which, by their very nature, emphasize the importance of assessing, modelling and understanding OpRisk profiles. These 3 pillars are; minimum capital requirements (refining and enhancing risk modelling frameworks), supervisory review of an institution's capital adequacy and internal assessment processes and market discipline, which deals with disclosure of information, see Figure 8.1.1. Since this time, the discipline of OpRisk and its quantification have grown in prominence in the financial sector.

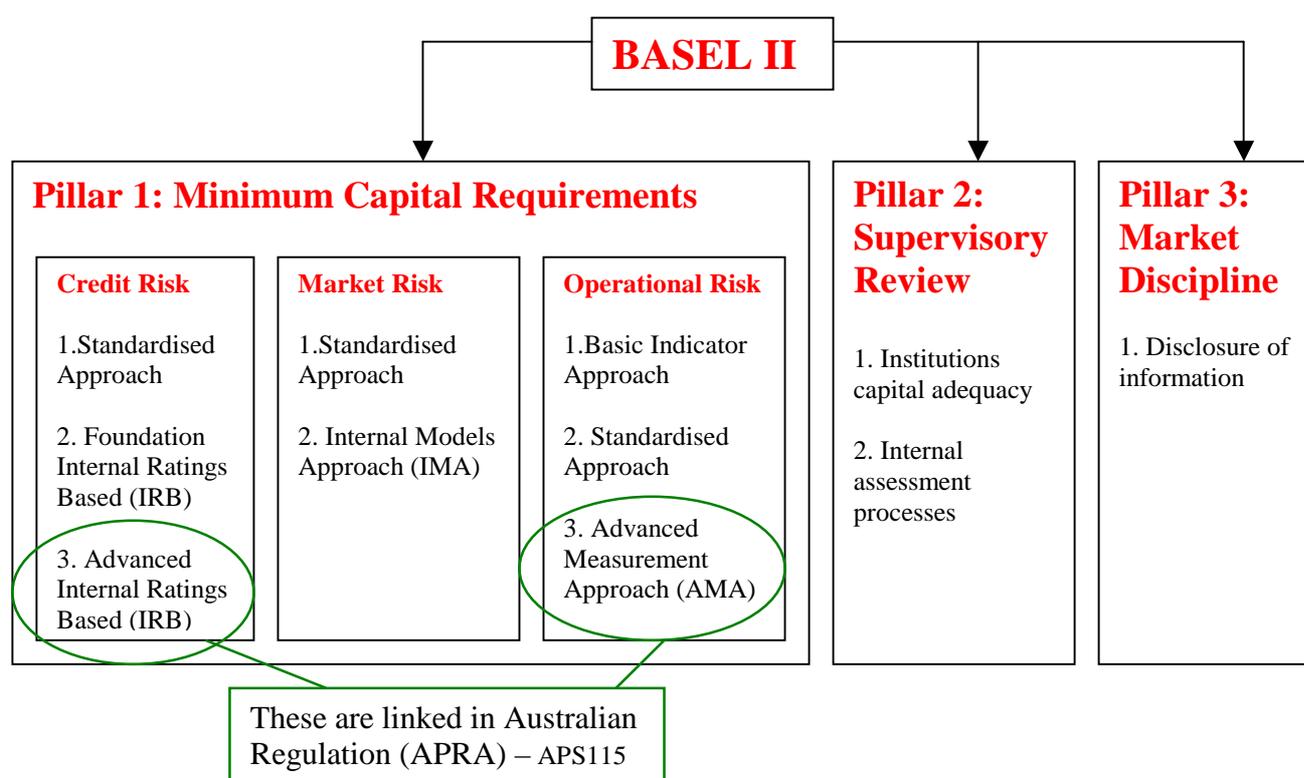


Fig. 8.1.1: BASEL II ACCORD - (2004)

OpRisk, for a business or organization, may broadly be defined as the risk involved in such an entity carrying out its normal operations. For a bank, the Basel Committee on Banking Supervision ("the Committee") defines OpRisk to be "the risk of loss resulting from inadequate or failed internal processes, people and systems or from external events." (Basel Committee on Banking Supervision, 2006, p144)

So OpRisk is indeed a broad category. The Committee gives a further classification into seven types of OpRisk (Basel Committee on Banking Supervision, 2006, Annex 9): Internal Fraud; External Fraud; Employment Practices and Workplace Safety; Clients, Products and Business Practices; Damage to Physical Assets; Business Disruption and System Failure; Execution, Delivery and Process Management; which serves to further illustrate the disparate nature of events in this class, see an overview in Figure 8.1.2. Reputational and strategic risk do not fall

under the OpRisk umbrella, and market and credit risks are treated separately, but almost any other event that may result in a loss to a bank, including legal action, may be termed OpRisk. To demonstrate the diverse range of OpRisk consider the following categories and specific examples: Internal fraud and human error (rogue traders); External fraud (credit card fraud); Acute physical hazards (Tsunami, hail, fire); Terrorism (Bombing, Internet attack); through to Collapse of an individual major partner (Enron, Worldcom). The following figure demonstrates the categories of OpRisk in a typical financial institution, Figure 8.1.2.

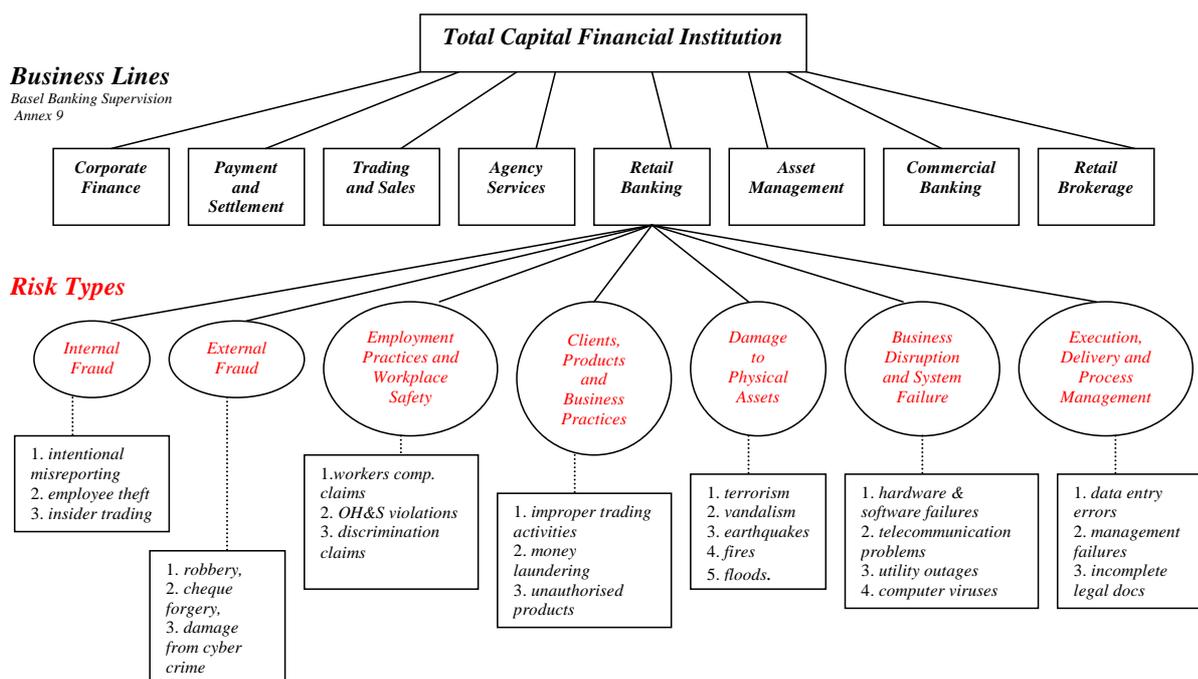


Fig. 8.1.2: BASEL II - Business Unit and Risk Type Categories

To illustrate just how significant OpRisk can be to a financial institution, consider the following OpRisk related events:

1. 1995: Barings Bank (loss GBP 1.3 billion)
2. 1996: Sumitomo Corporation (loss USD 2.6 billion)
3. 2001: September 11
4. 2001: Enron (loss USD 2.2 billion)
5. 2002: Allied Irish (loss GBP 450m)
6. 2004: National Australia Bank (AUD 360m)
7. Societe Generale (Euro 4.9 billion)

⋮

The impact that such significant losses have had on the financial industry and its perceived stability combined with the Basel II regulatory requirements have significantly changed the view, that financial institutions have regarding OpRisk. Under the three pillars of the Basel II agreement, set out in the framework¹, internationally active banks are required to set aside capital reserves against risk, to implement risk management frameworks and processes for their continual review, and to adhere to certain disclosure requirements. These regulatory requirements, which are overseen and enforced in Australia by APRA, have encouraged many banks to deploy significant resources to the task of quantifying OpRisk. Whilst many OpRisk events occur frequently and with low impact (indeed, are 'expected losses'), others are rare, and their impact may be as extreme as the total collapse of the bank. In any case, most institutions will not have sufficient internal data to accurately model their OpRisks, especially with respect to extreme rare losses. The modelling and development of methodology to capture, classify and understand properties of operational losses is a new research area in the banking and finance sector.

Accordingly, the Basel II agreement incorporates a high degree of modelling flexibility. The Committee itself is made up of representatives of both central banks and banking supervisory authorities from each of the G10 countries, and the framework has been developed and revised in consultation with the authorities and the industry in member and non-member countries. Basel II prescribes three different methods by which OpRisk capital may be calculated. In order to implement each of the two more sophisticated methods, a bank must meet certain qualifying criteria: in essence, it must prove to the regulator, that it has sufficient resources and systems in place to properly carry out and audit/review the more sophisticated calculations.

It is important to understand, where OpRisk fits into the overall risk picture within Australian institutions. In Australian retail banking, the largest profit center typically revolves around consumer credit and lending. The mortgage and home loan products represent the majority of profit. Other profit centers include markets and trading on the Australian stock exchange or global markets. Modelling of credit portfolios has been developed over many years and is reasonably well established in the banking sector. There are large databases, there are credit rating agencies, standards and rules for assessing and scoring credit ratings. The modelling of annual loss from a credit perspective, including rare event modelling, is well established, and quantiles of the annual loss distribution are used to produce risk measures such as Value at Risk (VaR) figures, which provide capital estimates. So in both methodology and in systems and process development, including accountability and incentives to report and maintain business processes, this area of modelling for extreme losses is highly developed.

At the other end of the spectrum one has OpRisk. The infancy of modelling of OpRisk relative to other risk disciplines was recognized by APRA early in the introduction of this new risk discipline: "...measuring and managing operational risk is still very much an emerging discipline" [Laker (2006), p6] (72) . . . "Unfortunately there is neither a history, nor broad agreement on the methodologies, for modelling OpRisk." [Egan (2005), p4] (37). As a response one of the key

¹ Note that the agreement covers market, credit and OpRisk, however we shall restrict our consideration here to OpRisk.

drivers put in place by APRA in Australia, in order to push the effort in developing a methodological framework for OpRisk and implementation of this framework in an integrated manner throughout a financial institution, is the fact they have tied the accreditation of an advanced approach to credit modelling with an advanced approach to OpRisk, typically termed Advanced Measurement Approach [APS 115] (4).

The diagram in Figure 8.1.3 presents a visual representation of how the risk metrics, typically considered for reporting or reserving of OpRisk, are related to the annual loss distribution, [Shevchenko (2009)] (107). The important questions from a statistical perspective pertain to how best to model and simulate such an annual loss distribution. This will be discussed in great detail throughout Part II.

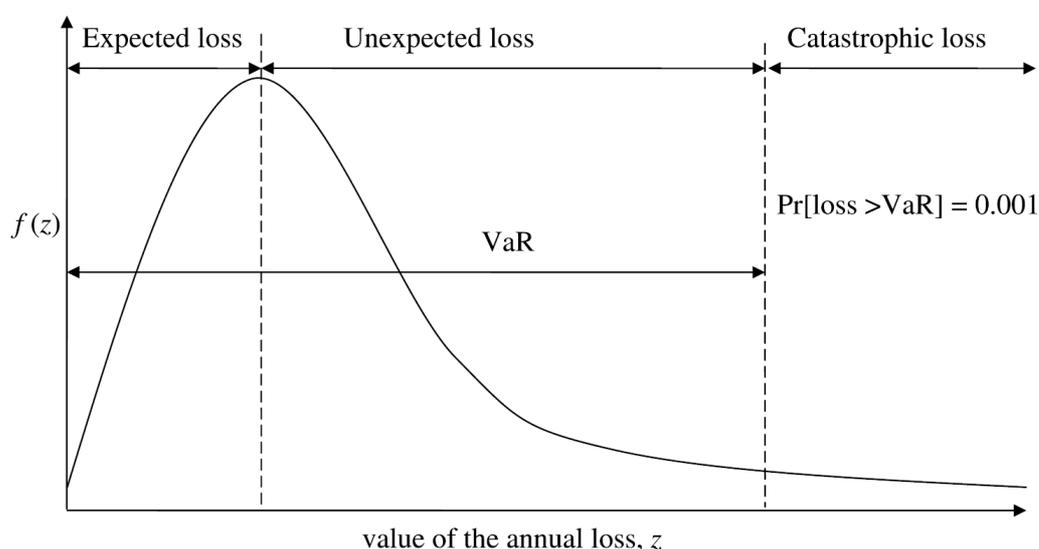


Fig. 8.1.3: Annual loss distribution and relevant risk measures.

The quantities in the above figure are denoted by EL for expected loss, UL for unexpected loss and VaR for value at risk. Their definitions are provided in the context of OpRisk, over a 1 year time horizon for a 1 in 1000 year event, as follows

$$\text{Unexpected Loss : } UL = \text{VaR}_{0.999} - EL$$

$$\text{Risk Measure : } \text{VaR}_q = \inf\{x : \text{Pr}[\text{Loss} > x] \leq 1 - q\}.$$

The general perception is that these advanced approaches are expected to improve understanding of the loss process and lower required capital reserves, typically quantified risk metrics such as VaR, relative to the traditional over conservative approaches. This would therefore free up capital to be used to grow the business. This is a key monetary incentive for the financial industry to develop sophisticated statistical models for OpRisk. Looking at the picture from another perspective, banks should be very prudent in modelling rare events and developing understanding of the processes - such as system failure, infra-structure failure and rogue trading - that lead to massive losses, all of which have the potential to debilitate a financial institution

or its subsidiaries. This is particularly relevant given the catastrophic losses recently resulting from the current financial crisis and subsequent collapse of many financial institutions, largely due to significant OpRisk losses.

The guidelines presented by APRA, as with those in Basel II, are not prescriptive in terms of implementation and methodological development. In particular, from a quantitative perspective, they do not advocate particular models for extreme or rare events. The most important quantitative guideline to date from APRA is APS 115. The key requirements specified in this standard are that a bank must have a “framework to manage, measure and monitor OpRisk commensurate with the nature, scale and complexity of the institution’s operations” and “approval from APRA to use an Advanced Measurement Approach to OpRisk for determining the institution’s OpRisk regulatory capital requirements”.

There are three broad approaches, that a bank may use to calculate its minimal capital reserve, as specified in the first pillar of the Basel II agreement. They are known as Basic Indicator Approach, Alternative Standardized Approach and Advanced Measurement Approach (AMA). In this section the details are provided for the approach most relevant to the papers contained in this thesis, the AMA. AMA is of interest since it is the most advanced framework with regards to statistical modelling.

8.1.3 *The Advanced Measurement Approach (AMA)*

A bank adopting the AMA must develop a comprehensive internal risk quantification system. This approach is the most flexible from a quantitative perspective, as banks may use any methods and models, they believe are most suitable for their operating environment and culture. However, this is the most restricted approach in that banks must gain supervisory approval before beginning to implement the AMA, and their models must satisfy further stringent qualitative and quantitative criteria outlined in the Basel II agreement. To start with, a bank is required to have an independent OpRisk management section, that is responsible for the measurement of OpRisk as well as the development of strategies for its management and mitigation. It must have an embedded ‘risk culture’, where the day to day operations of the bank integrate risk control, measurement and reporting. All risk management processes must be well documented and subject to regular internal and external audits... and so on. [Basel Committee on Banking Supervision, 2006, p150-2]. The key quantitative criteria are that a bank’s models must sufficiently account for potentially high-impact rare events, and incorporate the use of each of: internal data; external data; scenario analysis; and business, environment and control factors.

Implementing these guidelines and meeting these requirements for a bank typically involves a strong interplay between the bank and the supervisory authority. For example representatives from APRA would regularly visit banks applying for the AMA approach to assess and provide feedback on all aspects of OpRisk models and management frameworks being developed. The process also typically involves, in Australia, one or more outside independent parties, such as

KPMG, Ernst & Young, PricewaterhouseCoopers and Deloitte. These act as intermediaries, providing for APRA assurances and validations of approaches, models and implementations of business and data frameworks developed within banks.

This is an important part of the process in applying to the regulator for approval to use the AMA approach, since APRA is interested in external validation reports [see point 20 in attachment A of APS 115]. The reason for this will become clearer in subsequent sections discussing modelling approaches and data management. Additionally, in October 2006 all banks applying to use the AMA in OpRisk were required to take part in an exercise termed QIS5. This included producing a report on:

- the modelling approach implemented to date (including data management and recording systems covered in IT departments)
- the road map for the following year leading up to deadlines for accreditation in the first round
- (most importantly) VaR numbers for OpRisk in Australia and any subsidiary holdings, including both standardized and AMA figures.

We note that the calculation of Value at Risk as a risk measure is hotly debated in the academic community, especially relating to issues such as coherency, see [Artzner *et al.* (2000)] (5) and the difficulty of estimating accurately the quantile level of the annual loss distribution reported ($Q_{0.999}^2$). For such rare events as terrorist attacks, natural disasters and so on, these figures may not be sensible or stable over time. This was highlighted recently by a senior member of the German Financial Supervisory Authority. Dr. Gerhard Stahl commented on concerns that an ADI should have over the level of accuracy, that is attainable for reporting at a 0.999 quantile level, and how stable this will be over time.

It is also worth considering, what is involved in the practical implementation of an OpRisk framework in a financial institution, as it is a massive undertaking. Crudely, the process can be separated into four phases. We shall briefly describe these phases before moving on to the core section of this case study, which revolves around one of these phases “methodological developments”. By understanding how this framework needs to be integrated with the business, one gets a sense of the significance of developing models, which will at best possibly capture the behavior of these rare events in OpRisk. The business is now becoming accountable since managers will need to actively assess and manage their OpRisk profile, which is passed to them from the models developed. Clearly this provides a significant incentive to ensure, models are transparent and well understood by the risk community. Further, this knowledge needs to flow on through the business managers, who are being assessed on how well they actively manage such losses and events occurring from OpRisks.

Phase 1 – The first step in the process is typically to build a core team for the development and implementation of the entire framework. This may include business representatives, risk

² APS 115 - Point 20

specialists and quantitative analysts, policy developers and database experts, business analysts, auditors and validators.

Phase 2 – A key question faced by many financial institutions regards the development of an “in-house” framework versus an “off the shelf” or “plug and play” solution, which would be modified for the given business model or hierarchy. The framework includes

areas of database design and the set up for the capturing of the Internal Loss Data;

choosing the desired modelling methodology and modelling of the actual annual loss distribution under this approach; and

reporting and integration of results from modelling into other sections of the institution, including education and assessment of risk profiles. [Cruz (2002)] (28)

Key reports and information flows in this space include the reporting of economic capital (an internal measure of capital – typically at a different quantile level to regulatory capital) and profit after capital to the bank’s risk committee. In a truly integrated OpRisk framework one could even go as far as assessing individual managers’ Key Performance Indicators (KPIs) according to the performance of the OpRisk capital charge on individual business units. Clearly this impacts the entire institution. Thus another aspect to consider, from a practical perspective, when developing these models, is how to obtain substantial “buy-in” from business units, who will want to understand how they are exposed to different levels of extreme events relative to other business units. This has important implications for modelling correlation between different loss processes, an important aspect of models developed in the journal papers in the second chapter of Part II of this thesis.

Phase 3 – Development of the model methodology. In this area APRA has given significant flexibility to Australian financial institutions. This is reflected in the many varied approaches implemented throughout Australia. In Australia the key requirements from the regulator are set out in a series of documents, which include draft prudential standards, draft prudential practice guides, response to industry progress and discussion and guidelines. At the time of this report the most recent versions of the draft prudential standards released for Australia’s financial industry are APS114, APS115 and APG115.

Of these documents the one, which is directly relevant to OpRisk quantitative methodology is APS115. In this document the first section outlines the process a bank must undertake to obtain approval for an AMA. From the perspective of modelling rare events, points 18 through to 26 provide the guidelines; the allocation of capital charge to business units according to their risk profile is then covered in point 27. Point 18 gives an indication of the level of detail provided: “the [bank’s] OpRisk measurement system, must be sufficiently comprehensive to capture all material sources of OpRisk across the bank, *including those events that can lead to rare or severe operational losses.*” In addition to this note, which refers briefly to the nature of the modelling required, another important point to be addressed regards soundness standards over a universal annual modelling period. Statements such as “This soundness standard provides significant flexibility for [a bank] to develop an OpRisk measurement system that best suits the

nature and complexity of the [bank's] activities" and "Given the subjectivity and uncertainty of OpRisk measurement modelling, [a bank] must be conservative in the assumptions used in its OpRisk measurement model, *including assessment and incorporation of severe loss events*", illustrate the significant challenge involved in constructing appropriate methodology. Further discussion on the soundness standards and the ten principles that underpin them can be found in [KPMG (2005)] (68).

Even before models can be developed, questions, such as *how best to understand the nature of rare and extreme events that may lead to large losses*, must be asked. The answers, that a bank provides to these questions, will dictate many of the modelling assumptions, that can be made. Questions related to data sufficiency and validity, in addition to likely sources of information, and how best to integrate and fuse information on rare events, become critical to the process. In this context one needs to carefully assess how useful different data sources are for a given institution. Typically this requires thorough understanding of sources of bias present in data. OpRisk is inherently an area where data is still scarce and precious. Thus incorporation of expert opinion in many cases becomes a key driver in the measurement models.

The level of a business hierarchy, at which relevant operational risk information can be extracted (and suitably modelled), directly affects the approaches many banks take in modelling, including the granularity of modelling for a given business hierarchy. In this respect, granularity is a term used to refer to the number of levels of the business unit risk type hierarchy used in modelling. For example a model, which is not granular could model at the bank level by collecting all the loss data for a given risk type and combine it together then fit a statistical model. In Australia many banks model data at different levels of granularity, ranging from assessment and modelling of internal loss data or external loss data at an institutional level through to survey and scenario analysis at sub Business Unit and Risk Type (internal fraud, external fraud etc.) levels.

This then influences how easily expert opinion on extreme losses from different business units can be extracted, and how comprehensive this information is for a given business unit's risk profile locally within the business hierarchy. In turn, this affects how efficiently a business unit manager can understand, monitor and improve on her OpRisk performance.

It should be noted that recent years have seen the emergence of typically three different data sources, combinations of which are used in different models implemented in banks. These data sources are scenario analysis or survey data; internal loss data collected to date (can be very scarce and typically does not contain any truly large losses); and external data which comes from external companies such as FITCH.

However, the use of external data is severely hampered by the fact that many providers do not have complete records, and do not release institutional information. Hence scaling of loss amounts according to institution size - which is critical if data from external sources is to be combined with internal data - is very difficult, if not impossible in many cases. This in a sense compounds the problem, since many of the actual events recorded in these external data bases are the truly large or extreme losses, that have been witnessed in the industry.

Once these questions are understood within the context of the bank's business framework, then the models for measurement of OpRisk can be developed. The approaches taken will be elaborated on in future sections of the report.

Phase 4 – Calibration, sensitivity analysis and improvements to scenario analysis approaches. Again, this phase requires a lot of quantitative attention on how best to calibrate the model and how sensitive different modelling approaches are to key assumptions, inputs and approximations.

8.1.4 Model frameworks for Operational Risk

This section presents an overview of modelling aspects of OpRisk.

8.1.5 Issues associated with modelling Operational Risk

It is relevant to start the consideration of OpRisk models with an understanding of what makes developing quantitative models and methodology difficult in this context. There are many reasons. Firstly the sheer size of financial institutions and their subsidiaries makes co-ordination and understanding of approaches to operational risk a practical challenge. This raises issues such as the need for different business units located in different sections of Australia and overseas to understand requirements of assessment, and to act to establish management frameworks. This is important as the line managers of such business units need to actively assess and manage risk according to the behavior of their reported "modelled" risk profile. In this regard there is typically an information asymmetry, with much of the expertise in understanding the models developed - and therefore the key assumptions made in the process - located in center functions, physically far away from many of the business units actually affected by OpRisks.

The second issue is whether a bank is to implement in their models a "top-down" or a "bottom-up" approach. A top-down approach will do the mathematical modelling of the risk profile at a high level, for example the Bank level. All the loss data for the bank will be assumed homogeneous in terms of truncation and threshold levels and will be modelled as one set of data. This makes explicit assumptions about properties of the collected loss data, however it has the advantage of plenty of data for statistical modelling. For mathematical details see [Panjer (2006)] (86). Once modelled at the top level of the hierarchy the capital results will be allocated to business units according to some weighting factors. A bottom-up approach will model data and expert opinion at much lower levels of the hierarchy. For example individual business units will assess material risk types for their business and any loss data associated with this business unit and risk type will be modelled at this level. Then an aggregation process will be performed to combine all the business unit risk type loss profiles to a bank level.

This again will significantly influence the types of models, and in particular, how data is used in such models. The chosen approach is usually dependent on how well a bank believes they

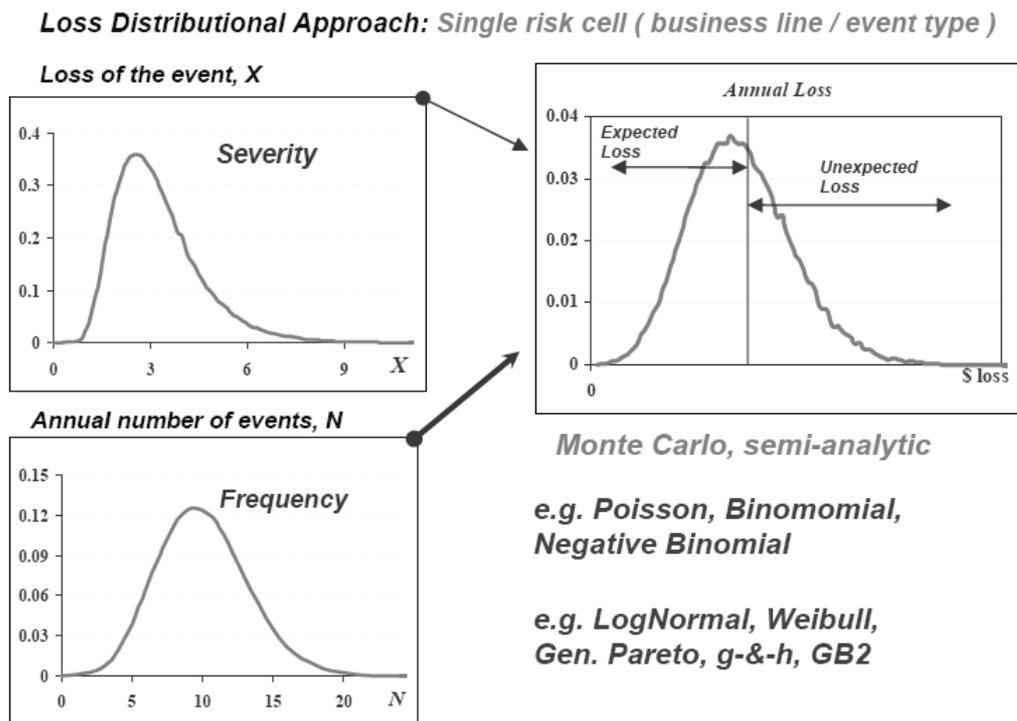


Fig. 8.1.4: LDA model framework.

can capture information from expert judgment and then integrate this with other loss data in the quantification process. Largely this process also involves significant business interaction, “buy-in”, to develop a team of experts in the business unit, who actively assess the local risk profile and take part in risk assessment exercises. This will be discussed in more detail in the next section, where we discuss the modelling of the individual data sources in OpRisk.

OpRisk can borrow ideas from insurance mathematics in the area of methodological development. Many models and approaches, which are based around the mature field of insurance mathematics, have been advocated by researchers in academic institutions [Cruz (2002)] (28); [Panjer (2006)] (86). However, there are several key differences, which will be explored in the context of OpRisk. The most significant is the fact, that OpRisk is still a very new “science” and is inherently an inexact science, where model assumptions and expert opinions are critically important to capture. Understanding the implications for a model of such judgments and assumptions is also a key part of the model development journey.

8.1.6 Modelling methodology for Operational Risk and the Loss Distributional Approach

Once the level of framework granularity is decided for the OpRisk model (as a function of the relevant data that is obtainable for each level of the hierarchy), the next step in the process is to apply a modelling framework. Of the methods developed to model OpRisk, the majority follow the Loss Distributional Approach (LDA), see Figure 8.1.4. The idea of the LDA is to fit severity and frequency distributions over a predetermined time horizon, typically annual as

specified in the APS115 section on soundness standards.

The fitting of frequency and severity distributions, as opposed to simply fitting a single parametric annual loss distribution, involves making the mathematical choice of working with compound distributions. This would seem to complicate the matter, since it is well known, that for most situations, analytical expressions for the distribution of a compound random variable are not attainable. The reason for modelling severity and frequency distributions separately then constructing a compound process is summarized in detail in [Panjer (2006)] (86). Some of the key points relating to why this is important in most practical settings are;

The expected number of operational losses will change as the company grows. Typically growth needs to be accounted for in forecasting the number of OpRisk losses in future years, based on previous years. This can easily be understood, when modelling is performed for frequency and severity separately. Economic inflationary effects can be directly factored into size of losses through scaling of the severity distribution. Insurance and the impacts of altering policy limits and excesses are easily understood by directly altering severity distributions. Changing recording thresholds for loss events and the impact this will have on the number of losses required to be recorded is transparent.

The most popular choices for frequency distributions are Poisson, binomial and negative binomial. The typical choices of severity distribution include exponential, Weibull, lognormal, generalized pareto, and recently in academic literature the g -and- h family of distributions [Dutta *et al.* (2006)] (36), [Peters *et al.* (2006)] (89). On the other side of the methodological divide there is a set of models being developed, utilizing concepts and ideas from Extreme Value Theory EVT [Embrechts *et al.* (2006)] (40). This divide mainly concerns approaches taken to fit such distributions and is discussed in detail in [Embrechts *et al.* (2006)] (40).

A key point to mention is that the most important processes to model accurately are those, which have relatively infrequent losses. However, when these losses do occur they are distributed as a very heavy-tailed severity distribution. These processes are by their very nature the most difficult to model, due to scarcity of data. From a practical perspective, this is where the importance of eliciting expert opinion and performing surveys or scenario analysis becomes critical.

The reason why these simple parametric models are widely used is that from a practical perspective they are relatively simple to fit, and to apply goodness of fit tests to (for purposes of model selection). Additionally, given the scarcity of most data sources, the fitting of parametric distributions with more than two parameters can quickly become problematic and unreliable. This is a practical issue, however there is also the theoretical issue of whether this class of distributions adequately captures the true behavior of the extreme events lying deep in the tails of these severity distributions. Industry consensus tends to suggest many of the extreme events, at least in the Australian financial sector, can be adequately modelled by lognormal and generalized pareto distributions. Returning again to EVT, in this space one can fit heavy tailed distributions for the severity distribution. Typically, fitting these models can be performed using either Points Over Threshold (POT) techniques of block maxima [Embrechts *et al.* (2005)]

(39). There has also been some literature about fitting EVT models from a Bayesian perspective [Brotot *et al* (2007)] (17). This approach will be discussed in another section of the report.

There are many approaches, which can be used to fit these parametric distributions and the approach adopted by a bank will depend on the data source being modelled and how much confidence one has in the data source. This is highly subjective. Techniques commonly adopted to fit frequency and severity models include extreme value theory [Cruz (2002)] (28), Bayesian inference [Schevchenko *et al.* (2006)] (106); [Cruz (2002)] (28), dynamic Bayesian networks [Ramamurthy *et al.* (2005)] (96), maximum likelihood [Dutta *et al.* (2006)] (36) and EM algorithms [Bee (2006)] (12). (In the next section we present a framework for modelling and a set of statistical tools, which can be used to fit these distributions to different data sources and then select between the different proposed models.) After the best-fitting models are selected, these are combined to produce a compound process for the annual loss distribution:

$$Y = \sum_{i=1}^N X_i, \quad (8.1.1)$$

where the random variable $X_i \sim f(x)$ follows the fitted severity distribution. The random variable $N \sim g(n)$, the fitted frequency distribution, is commonly modelled by Poisson, binomial and negative binomial distributions [Dutta *et al.* (2006)] (36). From this compound process, VaR and capital estimates may be derived.

Once compound processes have been fitted for each business unit and risk type, the next step in the process is to aggregate these annual loss random variables for each individual {business unit-risk type} combination, and thus to obtain the institution-wide annual loss distribution. This report will not discuss the issues associated with correlation and dependence modelling. For more information on typical approaches to introducing correlation in an aggregation process, including copula methods, correlation of frequency, severity or annual losses, see [Cruz (2002)] (28).

At a given level of the hierarchy structure, (which we may call a {business unit-risk type} tree), if there are M {business unit-risk type} combinations present³, this process of determining the distribution of the annual loss involves an M-fold convolution:

$$Y_{levelM} = \sum_{i=1}^M Y_i,$$

Then the distribution of such an annual loss random variable will be given by, $f_{levelM}(y) = f_{BuRT(1)}(y_1) * \dots * f_{BuRT(M)}(y_M)$, Since each of these distributions $f_{BuRT(i)}$ for each {business unit-risk type}, at the lowest level of the business unit risk type tree, takes the form of a compound process developed from the LDA model framework, solving these convolution integrals for an analytic expression is not possible [Panjer (2006)] (86). Hence, typically in practice, dif-

³ This number M will depend on the level of granularity of the model being used by the bank.

ferent forms of simulation are used to estimate these compound distributions. Then the convolved institutional level annual loss distribution, and finally the regulatory capital estimate are obtained (typically by using a VaR at the specified Q0.999).

An aside on approaches, that have been used to simulate such compound processes to estimate the annual loss distribution, can now be presented. The reason the compound distribution of Y has no general closed form is that it involves an infinite sum over all possible values of N , where the n^{th} term in the sum is weighted by the probability $Pr(N=n)$ and involves an n -fold convolution of the chosen severity distribution, conditional on $N = n$. Actuarial research has considered the distribution function of Y for insurance purposes through Panjer recursions [Panjer (2006)] (86). Other approaches utilize inversion techniques such as inverse Fourier transforms to approximate annual loss distributions, although they typically require assumptions such as independence between frequency and severity random variables [Embrechts *et al.* (2003)] (38).

8.1.7 Modelling the different data sources, elicitation of expert judgment and models to fit this information

Before mentioning some techniques, that can be used to fit parametric severity and frequency distributions to actual loss data or expert elicited judgments in the form of survey or scenario analysis, it is worth noting some recent theoretical results regarding the aggregation of compound processes.

Recent work found in “Multivariate models for OpRisk” [Bocker *et al* (2005)] (16) provides some analytical results for the asymptotic quantiles of an annual loss distribution constructed by aggregating several compound processes. This is useful as it provides mathematical insight on bounds for the VaR at high quantiles (such as those required for OpRisk capital reporting) after aggregation of several different types of compound process. In this paper, the key findings of the authors can be interpreted as stating that in the independent compound process case, the combined VaR measure at the next level of the hierarchy, after aggregation, will be asymptotically in the quantile level, convergent to the single compound process expression for the quantile with dominating VaR.

That is, if the VaR values of the individual ranked compound processes are dominated by one particular processes VaR then this will be the asymptotic VaR of the aggregated annual loss distribution at the next level, in the independent case. This study proves these results for classes of sub-exponential severity distributions which comprise Poisson processes. A subexponential distribution F satisfies, for $(X_i)_{i \in N}$ i.i.d. random variables,

$$\lim_{x \rightarrow \infty} \frac{P(X_1 + \dots + X_n > x)}{P(\max(X_1, \dots, X_n) > x)} = 1$$

for some value of n . So without concern over the mathematical technicalities presented above, broadly ‘this translates into a statement that for sub-exponential distributions the sum of the random variables will be dominated by one single large loss and not by the summation of several small losses’.

This is particularly relevant to OpRisk analysis. In [Bocker *et al.* (2005)] (16) the authors provide an analytic result for an example of a bi-variate VaR, that is, one calculated from the aggregate of two compound processes. The frequency distribution for both processes is chosen to be Poisson; for the severity distributions the selected distributions are Weibul for one process and Lognormal for the other. When convolution of the two compound annual loss random variables is achieved, the VaR for the annual loss distribution at the aggregated level is dominated by the compound process produced by the Lognormal and Poisson distributions.

Additionally, in earlier work, the same authors demonstrate, that in an LDA setting, tail quantiles of a compound process, with sub-exponential severity distributions, will be simply a multiplicative function of the mean of the frequency distribution and the quantile of the severity distribution. This asymptotically shows, that the quantiles of the compound process, will therefore be independent of the over-dispersion effects that can be added when including for example Negative Binomial processes. This is important from a modelling perspective, as it indicates, that, when concern is in estimation of VaR, one can stick to fitting Poisson processes, which come with well understood properties. These include independent modelling increments, exponentially distributed inter-arrival times for loss events both of which makes fitting such models to actual data considerably simpler.

These results clearly have implications, which are yet to be realized and studied for the dependent processes case, which is typically considered relevant in practical settings. An example of this is where frequency random variables for different business unit risk type processes are dependent on each other. The dynamic modelling of dependence in OpRisk is developed in one of the papers contained in Section 2 of this thesis.

8.1.8 Survey data and scenario analysis

From a practical perspective the most important data source in the Australian financial sector comes from survey or scenario analysis. This is largely a result of scarce data on rare events for processes such as terrorist attack, natural disasters, rogue trading and infra-structure failures. There are two broad approaches to dealing with expert opinions, Scenario Analysis and Survey Data. Scenario Analysis typically involves setting up workshops with each business unit for which OpRisk is being assessed and going through a sequence of exercises to assess potential loss amounts for each non-negligible risk type.

The term scenario analysis is used, since a workshop facilitator will extract loss information from the business expert participants. This is achieved via a sequence of questions relating to internal and external events. Both actual and hypothetical events are considered in the form of scenarios. An example of this would be to ask questions if assessing for example 'rogue trading'; What is the expected exposure and what is the worst case of possible exposure? What systems are in place to set limits? What are the known and potentially unknown flaws in such systems? How are these being managed? How does the scale of operations in this bank

compare to other known incidents in the financial sector, that the bank is operating? What management frameworks are in place?

If an LDA approach is utilized then the information extracted from such scenario analysis workshops is considered by all participants. Then the questions are recast with the explicit aim of extracting information around the severity and frequency of such events. Examples include; What is the typical exposure? What is the expected exposure? What is the worst possible loss? What is the 1 in 10 or 1 in 20 year loss? How often does the loss occur per year ? etc...

These answers are then used to fit frequency and severity distributions. One way to do this is to use extracted measures of location and quantiles in dollar values to fit the severity distribution by solving the simultaneous equations, relating the parameters to the quantiles or summary statistics elicited in the scenario analysis. The Frequency distribution can be fitted to rates of occurrence. Typically the fitted severity and frequency distributions and the simulated annual loss distributions would be played back to the business experts for each of the possible severity and frequency models considered. Then a feedback and refinement process is undertaken until there is comfort from the business and facilitators, that the risk profile adequately captures the behavior of the exposure for the assessed risk.

Other approaches include eliciting a sequence of quantiles or relative probabilities for different loss intervals. In general the following broad distributional summaries can be used as frameworks for developing scenarios and survey questions: Probabilities – extract individual probabilities of loss amounts based on actual industry losses; Quantiles – qth quantiles such as median, 1 in 10 loss or 0.9 quantile; Intervals – probability of losses above some threshold or in some specified dollar range; Location Measures - typical or representative measures of dollar losses (median, mode, mean); Scale and Dispersion Measures – how far from the (mean, median, mode) the loss might be; Measures of Shape – describing the density as unimodal, bimodal or multimodal, skewed left or right and kurtosis in the form of questions relating to tail behaviors.

The merits of each approach and an excellent discussion of such elicitation processes and the sources of inherent bias are presented in great detail in [O'Hagan (2006)] (84). This text considers a very wide cross section of literature from psychological expert elicitation and perception, practical elicitation and facilitation, statistical bias, survey development and modelling.

As pointed out in [O'Hagan (2006)] (84) it is important to understand that “the subjective perceptions and sensations are, in principle, measurable – and with some precision – but such measurements can only be interpreted relatively not absolutely”. In this regard one needs to consider the possible impact of forcing the business experts to conform to a certain summary of the severity and frequency distributions. Additionally, O'Hagan points out, that when capturing expert opinion about some uncertain quantity in the form of a distribution it is important to recognize the two different forms of uncertainty, aleatory and epistemic. ‘Aleatory uncertainty is induced by randomness such as when modelling uncertainty in one or more instances of a random process’. ‘Epistemic uncertainty is due to imperfect knowledge about something that is not itself random and is in principle knowable.’ Hence when developing models based

on this survey and scenario analysis it is important to somehow consider separate variables or behavior as a result of these two different uncertainties.

The second framework involves a Bayesian paradigm [Bayes (1763)] (9). This approach from the perspective of operational risk is captured in [Peters *et al.* (2006)] (89), [Shevchenko *et al.* (2006)] (106). To understand the difference between the Bayesian approach and the scenario analysis approach it is important to realize, that, typically, scenario analysis makes the assumption, that the parameters of the severity and frequency distributions are deterministic. It then aims to extract what is equivalent to point estimates of the parameters required for the LDA approach. The Bayesian approach treats the problem from a different paradigm. The parameters are treated from a mathematical perspective as random variables and the survey and elicitation process now involves extracting information on the prior distribution for these parameters. This prior, coupled with the likelihood model for the severity or frequency distribution, is combined under Bayes law to produce a posterior distribution on the parameters. Hence, from this perspective the elicitation of prior information should follow a different route to the typical scenario analysis. More information on prior elicitation procedures is found in [O'Hagan (2006)] (84) and [O'Hagan (1998)] (83).

8.1.9 Internal loss data and external data

Typically the process involved in internal and external loss data is to firstly study the properties of the data in each risk category. This involves histograms, box plots, time series plots, all of which are used to identify and question trends present in the data, which may be artificial. This could include misclassification of loss events, censoring and truncation etc.

Once the data is investigated, typically a maximum likelihood approach is used to fit the severity distributions. Other approaches could involve generalized moment matching or quantile matching. When mixtures of distributions are used then the popular approaches include Expectation Maximization algorithm [Bee (2006)] (12). This is particularly relevant, when truncation is known to be present. For a review of each approach and the properties see [Panjer (2006)] (86).

If a Bayesian approach is used, typically this loss data would enter into the modelling through the evaluation of the likelihood, when simulating from the posterior distribution of the LDA severity and frequency parameters. The simulation procedure in these cases typically involves development of sophisticated procedures such as Markov chain Monte Carlo (MCMC), importance sampling (IS) and sequential Monte Carlo (SMC) algorithms [Doucet and Johansen (2007)] (35); [Peters (2005)] (88).

Once the models for frequency and severity have been fitted, it is important to introduce some criteria to select the "best model". Typically this involves Kolmogorov-Smirnov or Anderson Darling tests for goodness of fit. Alternatively if a Bayesian approach is adopted one would consider Bayesian Information Criterion BIC or Deviance Information Criterion DIC as statistics to choose between different fitted frequency and severity models, which best represent the

data in the most parsimonious manner.

In summary, there are a number of pertinent issues in fitting models to OpRisk data: the combination of data sources from expert opinions and observed loss data; the elicitation of information from subject matter experts, which incorporates survey design considerations; sample biases in loss data collection, such as survival bias, censoring, incomplete data sets, truncation and, since rare events are especially important, small data sets.

8.1.10 Managing Operational Risk

With so much effort going into the complex task of quantifying a bank's OpRisk, it is important to emphasize that this is just one component of the overall task of managing operational risk. To be specific, the 'management' of OpRisk means the "identification, assessment, monitoring and control/mitigation of risk" [Basel Committee on Banking Supervision, 2003, p3]. As set out in the previous section, in order to use a quantification approach more sophisticated than the basic approach, a bank must fulfil certain requirements, many of which pertain to its risk management systems. The second pillar of the Basel II agreement ['Supervisory Review Process'] sets out a framework under which supervisors (of individual banks, and of the industry as a whole) must implement this process.

Many of the principles underpinning OpRisk management have already been touched on. It is required that banks have a dedicated OpRisk management unit, and much focus is on embedding a thorough awareness of OpRisk in all levels of the bank's operations. Many of the aspects of risk management are a straightforward precursor to risk quantification: a risk can not be quantified until it is identified; a risk cannot hope to be accurately quantified unless it is appropriately monitored and all incidences are reported. This is true within an individual bank and externally as well: banks are required to make "sufficient public disclosure" to allow other banks to compare and assess their OpRisk [Basel Committee on Banking Supervision, 2003, p5], and supervisors are directed to compare the operational risk calculations of similar banks in their domain [Basel Committee on Banking Supervision, 2006, p217]. It is up to a bank to justify to the supervisory authority, that its management systems, as well as its quantification processes, are sufficient, and industry-wide disclosure requirements can help a bank to ensure, that it is of the required standard.

Other aspects of risk management have a less straightforward relationship with risk quantification. In many cases, it is desirable to reduce exposure to an identified risk. (Other risks may be taken on intentionally, as part of a wider strategy to reap certain rewards.) A mitigation strategy such as insurance against a particular risk, will reduce the risk itself, but in itself introduce further risk, which must be measured, quantified, reported and so on. Thus there is a constant interaction between risk measurement and management. A bank will be constantly refining its risk models due to these internal interactions, as well as due to judgments and directives from the supervisory authorities.

So to summarize OpRisk involves the following modelling aspects, diagramatised in Figure

8.1.5.

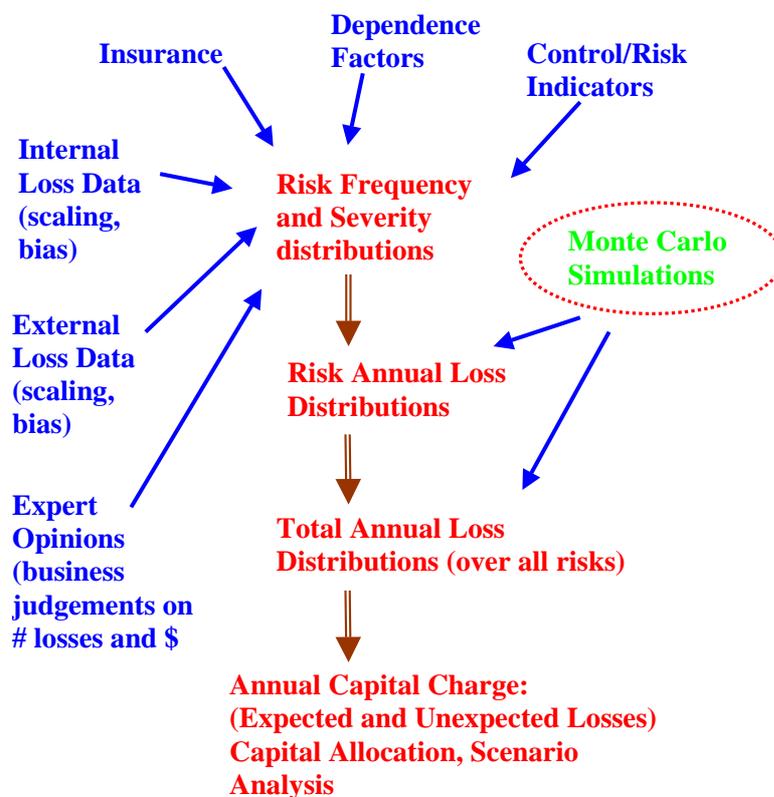


Fig. 8.1.5: Summary of modelling framework for OpRisk

8.2 Non-life insurance claims reserving

The focus of the insurance modelling contained in this thesis is on non-life claims reserving models. These models have a long history and an excellent coverage of them is found in [Wüthrich *et al.* (2008)] (123) and the papers cited within. As motivation for improving the modelling methodology in non-life insurance actuarial claims reserving one can turn to the recent regulatory standards of Sarbanes-Oxley, [Pub.L. 107-204, 116 Stat. 745, enacted July 30, 2002] (102).

The Basell II regulatory standards discussed above are having a major impact on the statistical modelling of banking risk, in the same manner the Sarbanes-Oxley (SOX) regulations are having a significant impact on the insurance and actuarial professions and their statistical models. It is fair to say, that the SOX has had a significant effect on the actuarial profession. It goes to the heart of financial modelling by aiming to enforce accountability for the sign-off of financial accounting. This includes the modelling and prediction of loss reserves. The intention of this act was to enforce accountability to avoid the corporate and accounting scandals such as Enron, Tyco International, Adelphia, Arthur Anderson and WorldCom, that occurred in the U.S., costing investors billions of dollars and causing a crisis of confidence in investors.

As discussed in [Marlo *et al.* (2006)] (79), the SOX act is the most far reaching piece of securities legislation since the SEC Act of 1934. The act applies directly to companies issuing securities on the US Securities exchange, whether domiciled in the US or not. To quote the SEC Chairman William Donaldson, the intention of the SOX act is " ... to restore public confidence in the accounting profession ... (and) improve the quality of financial reporting - to make the numbers accurate." Put another way, it is aimed at providing reasonable assurance of the reliability of financial reporting. In particular section SOX 404 has had implications for claims reserving modelling as it requires management to answer the question:

- Are there adequate controls in place to ensure, that the company prevents or detects material errors in financial statements on a timely basis?

The result of this statement has been significant, resulting in very expensive changes being implemented in many large companies. These typically involve ensuring the following steps are carried out to meet the SOX 404 requirements:

1. Identify significant accounts
2. Identify "financial statement assertions"
3. Define and document each "process"
4. Define "risks"
5. Define "controls" which mitigate each risk
6. Execute controls
7. Document execution of controls
8. Test effectiveness of controls
9. Remediate "control deficiencies".

Claims reserves are considered as a significant account, hence the role of actuarial modelling is in the spotlight since part of the actuarial process is to perform claims reserving calculations. Section 3.1 of [Marlo *et al.* (2006)] (79) details the implications on quantitative modelling for Actuaries of SOX.

As in the modelling of OpRisk under Basel II, the SOX act requires a particular focus on the following aspects which could result in misstated financial results: Data; Actuarial Valuation Systems; Spreadsheets (systems); Compilation of results; and Management review. Under the definitions of risk in SOX one considers model risk explicitly, examples include risks associated with inappropriate assumptions being used and incorrect calculations being performed.

This thesis aims to further develop understanding around models associated with the claims reserving Chain Ladder Model, to minimize the associated risks in claims reserving and to

improve the understanding of the actual claims process. In particular it studies model selection, parameterization and analysis of the uncertainty through the mean square error of prediction.

The reason this work is useful is that the classical Chain Ladder (CL) claims reserving approach, though basic, is not based on a stochastic model. It is a purely deterministic mechanistic process, which does not provide any insight into the prediction uncertainty, as stated in [Taylor *et al.* (2004)] (114).

" While the CL can be applied to any choice of data set, there is no apparent criterion for reliable choice of that data set. Moreover, the CL's phenomenological treatment of the trends is deeply unsatisfying. These trends must have a cause, that resides somewhere in the detailed mechanics of loss payment. However, the formulaic nature of the CL renders it incurious as to these details. The CL ... provides a sausage machine, a rigid and unenquiring algorithm. This is an advantage in terms of required resources. ... A serious disadvantage, to be set against this, is that it may produce a totally wrong result, that it may give precedence to processes over substance."

Yet, the CL method is used extensively by non-life insurance actuaries. Therefore, to fully comply with SOX requirements relating to risk, it has become important to understand the model assumptions and the uncertainty associated with the Chain Ladder method predictions. To achieve this several stochastic models can be considered, which recover the CL estimates whilst providing a sound stochastic framework from which to study the uncertainty in predictions.

8.3 Contribution Part II

Journal Papers:

Paper 1: **Peters, G.W. and Sisson S.A. (2006) "Bayesian Inference, Monte Carlo Sampling and OpRisk". *Journal of Operational Risk*, 1(3).**

OpRisk is an important quantitative topic as a result of the Basel II regulatory requirements. OpRisk models need to incorporate internal and external loss data observations in combination with expert opinion surveyed from business specialists. Following the Loss Distributional Approach, this article considers three aspects of the Bayesian approach to the modelling of OpRisk. Firstly we provide an overview of the Bayesian approach to OpRisk, before expanding on the current literature through consideration of general families of non-conjugate severity distributions, g-and-h and GB2 distributions. Bayesian model selection is presented as an alternative to popular frequentist tests, such as Kolmogorov-Smirnov or Anderson-Darling. We present a number of examples and develop techniques for parameter estimation for general severity and frequency distribution models from a Bayesian perspective. Finally we introduce and evaluate recently developed stochastic sampling techniques and highlight their application to OpRisk through the models developed.

This was one of the earliest papers to appear in which Bayesian models are proposed to capture expert opinions in OpRisk. In addition, it was the first paper to present non-conjugate Bayesian models for OpRisk modelling, using advanced models found to be suitable for many of the risk types studied.

Paper 2: **Peters, G.W. Johansen A. and Doucet A. (2007) "Simulation of the Annual Loss Distribution in OpRisk via Panjer Recursions and Volterra Integral Equations for Value at Risk and Expected Shortfall Estimation". *Journal of Operational Risk*, 2(3).**

Following the Loss Distributional Approach (LDA), this article develops two procedures for simulation of an annual loss distribution for modelling of OpRisk. First, we provide an overview of the typical compound-process LDA used widely in OpRisk modelling, before expanding upon the current literature on evaluation and simulation of annual loss distributions. We present two novel Monte Carlo simulation procedures. In doing so, we make use of Panjer recursions and the Volterra integral equation of the second kind to reformulate the problem of evaluation of the density of a random sum as the calculation of an expectation. We demonstrate the use of importance sampling and trans-dimensional Markov Chain Monte Carlo algorithms to efficiently evaluate this expectation. We further demonstrate their use in the calculation of Value at Risk and Expected Shortfall.

Paper 3: **Peters, G.W., Shevchenko P. and Wuthrich M. (2009) "Dynamic OpRisk: modelling dependence and combining different sources of information". *Journal of OpRisk*, 4(2).**

In this paper, we model dependence between OpRisks by allowing risk profiles to evolve stochastically in time and to be dependent. This allows for a flexible correlation structure where the dependence between frequencies of different risk categories and between severities of different risk categories as well as within risk categories can be modelled. The model is estimated using the Bayesian inference methodology, allowing for combination of internal data, external data and expert opinion in the estimation procedure. We use a specialized Markov chain Monte Carlo simulation methodology known as Slice sampling to obtain samples from the resulting posterior distribution and estimate the model parameters.

Paper 4: **Peters, G.W., Shevchenko P. and Wuthrich M. (2009) "Model Risk in Claims Reserving within Tweedies Compound Poisson Models". *ASTIN Bulletin*, 39(1).**

In this paper we examine the claims reserving problem using Tweedie's compound Poisson model. We develop the maximum likelihood and Bayesian Markov chain Monte Carlo simulation approaches to fit the model and then compare the estimated models under different scenarios. The key point we demonstrate relates to the comparison of reserving quantities with and without model uncertainty incorporated into the prediction. We consider both the model selection problem and the model averaging solutions for the predicted reserves. As a part of this process we also consider the sub problem of variable selection to obtain a parsimonious representation of the model being fitted.

Paper 5: **Peters, G.W., Wuthrich M. and Shevchenko P. (2009) "Chain Ladder Method: Bayesian Bootstrap versus Classical Bootstrap" in review at journal Insurance: Mathematics and Economics (conditionally accepted).**

The intention of this paper is to analyze the mean square error of prediction (MSEP) under the distribution-free chain ladder (DFCL) claims reserving method. We compare the estimation obtained from the classical bootstrap method with the one obtained from a Bayesian bootstrap. To achieve this in the DFCL model we develop a novel approximate Bayesian computation (ABC) sampling algorithm to obtain the empirical posterior distribution. We need an ABC sampling algorithm because we work in a distribution-free setting. The use of this ABC methodology combined with bootstrap allows us to obtain samples from the intractable posterior distribution without the requirement of any distributional assumptions. This then enables us to calculate the MSEP and other risk measures like Value-at-Risk.

Journal Paper 6

"Imagination is more important than knowledge. For while knowledge defines all we currently know and understand, imagination points to all we might yet discover and create."

Albert Einstein

Peters G.W. and Sisson S.A. (2006) "Bayesian Inference, Monte Carlo Sampling and Operational Risk". *Journal of Operational Risk*, 1(3).

This work was instigated by the first author of this major paper and he can claim around 80% of the credit for the contents. This work is a result of industry experience in the Banking sector, prior to his PhD. This paper has already been cited numerous times by people in industry and in academic research in this field. His work included developing the methodology contained, developing the applications, the implementation and comparison to alternative approaches, writing the drafts of the paper and undertaking revisions. This paper was accepted for publication in the *Journal of Operational Risk*, and has appeared. Permission from all the co-authors is granted for inclusion in the PhD.

This paper appears in the thesis in a modified format to that which finally appeared in the *Journal of Operational Risk*, where it was published.

Final print version available at: <http://www.journalofoperationalrisk.com/>

Bayesian Inference, Monte Carlo Sampling and Operational Risk

Gareth W. Peters (*corresponding author*)

CSIRO Mathematical and Information Sciences, Locked Bag 17, North Ryde, NSW, 1670, Australia;

UNSW Mathematics and Statistics Department, Sydney, 2052, Australia;

email: peterga@maths.unsw.edu.au

S. A. Sisson

UNSW Mathematics and Statistics Department, Sydney, 2052, Australia

9.1 Abstract

Operational risk is an important quantitative topic as a result of the Basel II regulatory requirements. Operational risk models need to incorporate internal and external loss data observations in combination with expert opinion surveyed from business specialists. Following the Loss Distributional Approach, this article considers three aspects of the Bayesian approach to the modelling of operational risk. Firstly we provide an overview of the Bayesian approach to operational risk, before expanding on the current literature through consideration of general families of non-conjugate severity distributions, g-and-h and GB2 distributions. Bayesian model selection is presented as an alternative to popular frequentist tests, such as Kolmogorov-Smirnov or Anderson-Darling. We present a number of examples and develop techniques for parameter estimation for general severity and frequency distribution models from a Bayesian perspective. Finally we introduce and evaluate recently developed stochastic sampling techniques and highlight their application to operational risk through the models developed.

Keywords: Approximate Bayesian Computation; Basel II Advanced Measurement Approach; Bayesian Inference; Compound Processes; Loss Distributional Approach; Markov Chain Monte Carlo; Operational Risk.

First Version: 27 October 2006

Revised Version: 8 November 2008

9.2 Introduction

Operational Risk is an important quantitative topic in the banking world as a result of the Basel II regulatory requirements. Through the Advanced Measurement Approach, banks are permitted significant flexibility over the approaches that may be used in the development of operational risk models. Such models incorporate internal and external loss data observations in combination with expert opinion surveyed from business subject matter experts. Accordingly the Bayesian approach provides a natural, probabilistic framework in which to evaluate risk models.

Of the methods developed to model operational risk, the majority follow the Loss Distributional Approach (LDA). The idea of LDA is to fit severity and frequency distributions over a predetermined time horizon, typically annual. Popular choices include exponential, weibull, lognormal, generalised Pareto, and g -and- h distributions [Dutta *et al.* 2006]. The best fitting models are then used to produce compound processes for the annual loss distribution, from which VaR and capital estimates may be derived. Under the compound process,

$$Y = \sum_{i=1}^N X_i, \quad (9.2.1)$$

where the random variable $X_i \sim f(x)$ follows the fitted severity distribution. The random variable $N \sim g(n)$, the fitted frequency distribution, is commonly modelled by Poisson, binomial and negative binomial distributions [Dutta *et al.* 2006].

The distribution of Y has no general closed form as it involves an infinite sum over all possible values of N , where the n^{th} term in the sum is weighted by the probability $Pr(N=n)$ and involves an n -fold convolution of the chosen severity distribution, conditional on $N = n$. Actuarial research has considered the distribution function of Y for insurance purposes through Panjer recursions [Panjer, 2006]. Other approaches utilize inversion techniques such as inverse Fourier transforms to approximate annual loss distributions, although they typically require assumptions such as independence between frequency and severity random variables [Embrechts *et al.* 2003]. Techniques commonly adopted to fit frequency and severity models include extreme value theory [Cruz, 2002], Bayesian inference [Shevchenko *et al.* 2006; Cruz, 2002], dynamic Bayesian networks [Ramamurthy *et al.* 2005], maximum likelihood [Dutta *et al.* 2006] and EM algorithms [Bee, 2006].

There are a number of pertinent issues in fitting models to operational risk data: the combination of data sources from expert opinions and observed loss data; the elicitation of information from subject matter experts, which incorporates survey design considerations; sample biases in loss data collection, such as survival bias, censoring, incomplete data sets, truncation and, since rare events are especially important, small data sets.

The LDA approach is convenient framework that can be enhanced through application of the Bayesian paradigm. [Shevchenko *et al.* 2006] introduce Bayesian modeling under an LDA

framework but restrict consideration to classes of frequency and severity models to those which admit conjugate forms. This permits simple posterior simulation and parameter estimates can often be derived analytically. Modern Bayesian methods have moved away from restrictive conjugate modelling, which has been made possible by the development of sophisticated simulation procedures such as Markov chain Monte Carlo (MCMC), importance sampling (IS) and sequential Monte Carlo (SMC) algorithms [Doucet *et al.* 2006; Peters, 2005]. As modelling of severity and frequency distributions in operational risk becomes more mature, moving away from conjugate families becomes necessary. For example, [Dutta *et al.* 2006] recently propose the use of g -and- h distributions for the severity distribution, which only admits conjugacy in special cases, such as when it reduces to a lognormal form. In these special cases, the work of [Shevchenko *et al.* 2006] could be applied.

In this article we consider three aspects of Bayesian inference as applied to the modelling of operational risk. We firstly present an exposition of why and how Bayesian methods should be considered as the modeling paradigm of choice, and explore how this approach fits into an LDA setting. We examine applications of the Bayesian approach to model selection and introduce to operational risk the notions of Bayes factors, Bayesian information criterion (BIC) and deviance information criterion (DIC).

Secondly, we extend the current literature on Bayesian modelling as applied to operational risk to incorporate general families of severity and frequency distributions. Finally we investigate modern techniques for sampling and parameter estimation for general severity and frequency distribution models. These include MCMC, simulated annealing and approximate Bayesian computation. Throughout we focus on the development of non-conjugate severity distribution models – initially when the likelihood has an analytic form, and subsequently when the likelihood function may not be written down analytically, such as in the case of the g -and- h distribution.

We conclude with some practical examples of the ideas developed in an operational risk modeling situation, followed by a discussion.

9.3 Bayesian Inference

9.3.1 Bayesian Inference and Operational Risk

Here we briefly review the fundamental ideas required for Bayesian analysis. There are two well-studied approaches to performing probabilistic inference in the analysis of data – the frequentist and Bayesian frameworks. In the classical frequentist approach one takes the view that probabilities may be seen as relative frequencies of occurrence of random variables. This approach is often associated with the work of Neyman and Pearson who described the logic of statistical hypothesis testing. In a Bayesian approach the distinction between random variables and model parameters is artificial, and all quantities have an associated probability distribution, representing a degree of plausibility. By conditioning on the observed data a probability

distribution is derived over competing hypotheses. For in-depth discussion on the details of each approach see, for example, [Bernardo *et al.* 2004].

The Bayesian paradigm is a widely accepted means to implement a modern statistical data analysis involving the distributional estimation of k unknown “parameters,” $\theta_{1:k} = [\theta_1, \dots, \theta_k]$, from a set of n observations $y_{1:n} = [y_1, \dots, y_n]$. In operational risk, the observations could be counts of loss events, loss amounts or annual loss amounts in dollars. Prior knowledge of the system being modeled can be formulated through a prior distribution $p(\theta_{1:k})$. In operational risk this would involve prior elicitation from subject matter experts through surveys and workshops. From the mathematical model approximating the observed physical phenomena, the likelihood $p(y_{1:n}|\theta_{1:k})$ may be derived, thereby relating the parameters to the observed data. In the context of operational risk this reflects the quantitative team’s modeling assumptions, such as the class of severity and frequency distributions representing the loss data observations. The prior and likelihood are then combined through Bayes’ rule [Bayes, 1763]:

$$p(\theta_{1:k}|y_{1:n}) = \frac{p(y_{1:n}|\theta_{1:k}) p(\theta_{1:k})}{\int_{\Theta} p(y_{1:n}|\theta_{1:k}) p(\theta_{1:k}) d\theta_{1:k}} \quad (9.3.1)$$

to give the posterior probability of the parameters having observed the data and elicited expert opinion. The posterior may then be used for purposes of predictive inference. In this manner Bayesian approaches naturally provides a sound and robust approach to combining actual loss data observations and subject matter expert opinions and judgments.

Much literature has been devoted to understanding how to sensibly assign prior probability distributions and their interpretation in varying contexts. There are many useful texts on Bayesian inference – for further details see e.g. [Box *et al.* 1992; Gelman *et al.* 1995; Robert, 2004].

9.3.2 Bayesian Parameter Estimation and Operational Risk

Maximum likelihood is the most common technique for parameter estimation employed by operational risk practitioners, and this approach has several useful properties including asymptotic efficiency and consistency. This is typically applied to an LDA analysis by using an optimization routine to find the maximum of a truncated likelihood distribution surface.

One issue with this approach is that variability inherent in very small and incomplete datasets can produce drastically varying parameter estimates. This is where approaches such as the EM algorithm can be useful. In addition, even given severity and frequency maximum likelihood parameter estimates, the problem of fusing fitted severity and frequency distributions with expert opinion still remains. The Bayesian approach avoids these problems under a single probabilistic framework.

In the Bayesian setting, the full posterior distribution should be used directly for predictive and inferential purposes. However, if the operational risk practitioner wishes to perform estimation

of fixed parameter values, a number of sensible options are available. The two most common estimators are the maximum *a posteriori* (MAP) and the minimum mean square error (MMSE) or minimum variance estimator [Ruanaidh *et al.* 1996]. The MAP estimator maximises the likelihood function weighted by the prior probability as

$$\theta_{1:k,MAP} = \arg \max_{\theta_{1:k}} p(\theta_{1:k}|y_{1:n}) \quad , \quad (9.3.2)$$

whereas the MMSE estimator is given by

$$\theta_{1:k,MMSE} = \int \theta_{1:k} p(\theta_{1:k}|y_{1:n}) d\theta_{1:k}. \quad (9.3.3)$$

Thus in a Bayesian setting the MAP and MMSE estimators respectively equal the posterior mode and mean. It is routine in MCMC literature to estimate the MAP or MMSE estimates using samples from the Markov Chain Monte Carlo algorithm created to sample from the posterior of interest. These estimates are then used for LDA, through simulation, to obtain empirical estimates of the compound process annual loss distribution characteristics, such as quantiles, mean, variance, skew and kurtosis. This simulation of the compound process could be performed easily through a Monte Carlo sampling approach or alternatively through a fast inverse Fourier transform approximation of the compound process characteristic function.

9.3.3 Bayesian Model Selection and Operational Risk

Before the quantitative analyst can estimate annual loss distribution characteristics, choices are required regarding the severity and frequency distributions. We propose three popular Bayesian model selection criteria which may be adopted for frequency and severity distribution model selection in the context of operational risk. These approaches provide Bayesian alternatives to standard tools used by operational risk practitioners who frequently utilize Kolmogorov-Smirnov, Anderson-Darling and chi-squared tests.

The task is therefore to implement Bayesian model selection criteria as a means of selecting between different posterior distributions. In an LDA setting, situations which require decisions between different posterior distributions include where different prior distributions are derived from expert opinion, or when fitting competing severity distribution models to observed data.

The ideal case for choosing between posterior models occurs when integration of the posterior (with respect to the parameters) is possible. Here one can evaluate the posterior odds ratio, which is the Bayesian alternative to classical hypothesis testing. For two severity distribution models M_1 and M_2 , parameterized by $\theta_{1:k}$ and $\alpha_{1:j}$ respectively, the posterior odds ratio of M_1 over M_2 is given by

$$\frac{p(y_{1:n}|M_1)p(M_1)}{p(y_{1:n}|M_2)p(M_2)} = \frac{p(M_1) \int p(y_{1:n}|\theta_{1:k}, M_1)p(\theta_{1:k}|M_1) d\theta_{1:k}}{p(M_2) \int p(y_{1:n}|\alpha_{1:j}, M_2)p(\alpha_{1:j}|M_2) d\alpha_{1:j}} = \frac{p(M_1)}{p(M_2)} BF_{12} \quad (9.3.4)$$

which is the ratio of posterior model probabilities, having observed the data, where $p(M_i)$ is

the prior probability of model M_i . This quantity is the Bayesian version of a likelihood ratio test, and a value greater than 1 indicates that model M_1 is more likely than model M_2 (given the observed data, prior beliefs and choice of the two models). The term BF_{12} is known as the Bayes factor. We note that the application of Bayes factors in operational risk modeling is sensible as the priors elicited from subject matter experts will be proper, rather than vague or improper, which can complicate model comparisons when evaluating (5). If analytic integration is not possible, either numeric or analytic approximations can be made. The latter approach leads to the Bayes and deviance information criteria.

The Bayesian information criterion (BIC) or the Schwarz criterion [Robert, 2004] is given by

$$BIC = -2 \ln(L) + k \ln(n), \quad (9.3.5)$$

where n is the number of observations, k is the number of free parameters in the prior to be estimated and L is the maximized value of the likelihood function for the estimated model. This criterion may be evaluated for each fitted severity or frequency distribution model with the selected model – i.e. the one with the lowest BIC value – corresponding to a model which both fits the observed data well and is parsimonious. The BIC criterion, first derived in [Schwarz, 1978], is obtained by forming a Gaussian approximation of the log posterior distribution around the posterior mode using Laplace’s method. BIC is an asymptotic approximation, so the number of observations should be large.

An alternative model selection criterion for operational risk practitioners, when developing hierarchical Bayesian models, is the deviance information criterion (DIC):

$$DIC = 2\overline{D(\vartheta)} - D(\overline{\vartheta}) \quad (9.3.6)$$

which is a generalization of BIC for Bayesian hierarchical models. The bar denotes an averaging over ϑ and the term $D(\vartheta)$ is the “Bayesian deviance”

$$D(\vartheta) = -2 \log(p(y_{1:n}|\vartheta)) + C,$$

where $p(y_{1:n}|\vartheta)$ is the likelihood function depending on data $y_{1:n}$ and parameters ϑ , and where C is a constant. (This constant is common to all models, and so may be safely ignored). See, e.g., [Spiegelhalter *et al.* 2002]. The DIC is particularly useful in Bayesian model selection problems in which the posterior distributions of the models have been determined through simulation. We shortly detail methods in which this may be achieved as operational risk itself moves away from conjugate families of severity and frequency distributions.

For detailed discussion on model selection criterion and comparison of BIC, Bayes Factors and Aikake information criterion (a popular frequentist model selection criterion) see [Wasserman, 1997]. For a detailed exposition on DIC criterion see [Spiegelhalter *et al.* 2002]. These model selection techniques form an integral part of model development where posterior distributions are only known up to a normalization constant or when no analytic form for the likelihood is obtainable, as we now consider.

9.4 Non-Conjugate Distributions for Modelling Operational Risk

[Shevchenko *et al.* 2006] present conjugate poisson-gamma, lognormal-normal and pareto-gamma classes of Bayesian frequency and severity models, and discuss how to handle these families in the context of truncated data, which is a common occurrence in operational risk. This derives from policies in which institutions only record losses when they are above a defined threshold level.

So what is conjugacy? Prior distributions are conjugate to a likelihood function if the resulting posterior distribution has a standard distributional form. Accordingly, the approach of [Shevchenko *et al.* 2006] is elegant as the posterior distribution of the parameters is analytically known and easy to sample from. However, conjugacy restricts the classes of distributions which may be used to elicit information from subject matter experts, and is also restrictive in that it does not in general accommodate recently analyzed classes of distributions for operational risk such as the *g*-and-*h* or the GB2 distributions [Dutta *et al.* 2006; He *et al.* 2006].

In this section and those that follow, we will demonstrate the application of the Bayesian framework to non-conjugate families and obtain parameter estimates for use in LDA analysis. We note that in allowing for greater modelling flexibility by moving away from conjugate families of distributions, one incurs a greater computational burden to simulate from the posterior distribution. This burden is greatest when the likelihood itself admits no closed form.

The models we consider for the observed loss data are assumed to be from either the *g*-and-*h* distribution family or the generalized Beta distribution of the 2nd kind (GB2) as presented in [Dutta *et al.* 2006]. The methodology we present is not restricted to these choices.

Consideration of the GB2 and the *g*-and-*h* distribution is believed to be directly beneficial to modeling of Operational Risk data. The study performed by [Dutta *et al.* 2006] demonstrates empirical evidence of why these two types of distribution should be considered in the context of Operational Risk modeling, using actual loss data collected from a variety of financial institutions. One of the key reasons for considering these models is their flexibility. It is a heavily debated topic in the banking industry around what the most appropriate class of distributions, parametric or non-parametric, to use for the modeling of actual loss data. This issue is compounded by the fact that modeling of operational risk data can be done at various levels of an organization, ranging from the parent level in which all loss data in the organization is analyzed down to individual business unit and risk type levels. Hence, designing models with flexibility to capture salient features of loss data at multiple levels of an organization is of direct benefit. In addition to these issues, even after developing an appropriate model for loss data observations, a mathematically sound technique is required for combining this loss data model with expert opinion. In this context, this paper provides a robust method in which these flexible models for the actual observed loss data can be combined with expert opinion at different levels of an organization. The requirement in doing this is the ability to encode in a prior distribution the subject matter experts judgments, which are typically elicited through surveys. The design of these surveys is critical, an excellent discussion of constructing and performing

expert judgment elicitation is provided by [O'Hagan, 2006].

The flexibility of the GB2 and the g -and- h distributions lies in the fact that they encompass families of distributions. What this means is that depending on the parameter values used in specification of these distributions, they can recover a range of parametric distributions. This includes the typical severity distribution models which are routinely applied in practical industry modeling of operational risk data, such as Lognormal, Weibull and Gamma. For a review of which parametric distributions belong to the g -and- h family see [Bookstaber et al, 1987]. In addition to this they also allow the model to achieve a wide range of location, scale, skewness and kurtosis, the purpose of such flexibility as mentioned previously is to allow the model being fitted to the data to capture as much of the salient features as possible. The cost of fitting such flexible models is an increase in the number of parameters that are required to be estimated. This can be problematic if modeling of loss data is being performed a granular level in which only small data sample sizes are present and should be considered in practical implementation.

The GB2 distribution has density function given by

$$f(y) = \frac{|a| y^{ap-1}}{b^{ap} B(p, q) [1 + (y/b)^a]^{p+q}} I_{(0, \infty)}(y) \quad , \quad (9.4.1)$$

where $B(p, q)$ is the Beta function and the parameters a , p and q control the shape of the distribution and b is the scale parameter. For discussion of how the parameters are interpreted in terms of location, scale and shape see [Dutta *et al.* 2006]. This distribution poses a challenge to Bayesian modelling as it does not admit a conjugate family of prior distributions, and accordingly necessitates more advanced posterior simulation techniques.

The g -and- h distributional family are obtained through transformation of a standard normal random variable $Z \sim N(0, 1)$ via

$$Y_{g,h}(Z) = (e^{gZ} - 1) \frac{\exp(hZ^2/2)}{g} \quad . \quad (9.4.2)$$

In this article we consider the g -and- h four parameter family of distributions presented in [Dutta *et al.* 2006] given by the linear transformation

$$X_{g,h}(Z) = A + B.Y_{g,h}(Z) \quad , \quad (9.4.3)$$

where g and h are real valued functions of Z and where A and B are location and scale parameters. The h parameter is responsible for kurtosis and elongation (how heavy the tails are) and the g parameter controls the degree of skewness. Positive and negative values for g respectively skew the distribution to the right and left. With examples, [Dutta *et al.* 2006] discuss the shapes of the g -and- h distribution for different settings of parameter values.

For simplicity we adopt constant values for g and h , although relaxation of this assumption does not preclude inference. This distributional family poses a unique challenge to Bayesian

modeling as the likelihood function does not in general admit an analytic closed form, since the density may only be expressed through a highly non-linear transform of a standard normal random variable. Accordingly, very recent state-of-the-art Monte Carlo sampling strategies are required, designed specifically for this purpose. These methods, known as approximate Bayesian computation (ABC), were first introduced in the statistical genetics literature [Beaumont *et al.* 2002; Tavaré *et al.* 2003]. As noted in [Dutta *et al.* 2006], the g -and- h distribution has a domain given by the entire real line, this can be handled in practice by introducing a truncation threshold into the model. This paper only considers the case of deterministic, known threshold situations, though future extensions could incorporate unknown or stochastic thresholds.

In the above models we adopt Gamma priors although, as before, this choice is arbitrary and other selections of prior could be used without difficulty. For example, Gaussian priors can be used if one does not want to restrict the domain of the parameters in the model to \mathbb{R}^+ . The choice of prior parameter values will in general be elicited from a subject matter expert through a survey or workshop. Prior elicitation in general is a challenging process. In the current setting an interactive feedback process may be appropriate, whereby subject matter experts repeatedly modify their prior parameters until they judge a realistic posterior annual loss distribution has been produced. Other parameter extraction processes exist; these include graphical distributional shape demonstrations, and combinations of this with questions about worst case (maxima) and typical loss (quantile) values and frequencies of losses. See, for example, [Garthwaite *et al.* 2000; O'Hagan, 1998].

Model 1 – GB2 likelihood with Gamma prior distributions. The posterior distribution for this model is derived from (9.4.1) and the independent Gamma priors $p(a|\alpha_a, \beta_a)$, $p(b|\alpha_b, \beta_b)$, $p(p|\alpha_p, \beta_p)$, and $p(q|\alpha_q, \beta_q)$, yielding

$$\begin{aligned}
 p(a, b, p, q|y_{1:n}) &= \frac{h(y_{1:n}|a, b, p, q)p(a)p(b)p(p)p(q)}{\int h(x|a, b, p, q)p(a)p(b)p(p)p(q)dadpdq} \\
 &\propto h(y_{1:n}|a, b, p, q)p(a|\alpha_a, \beta_a)p(b|\alpha_b, \beta_b)p(p|\alpha_p, \beta_p)p(q|\alpha_q, \beta_q) \\
 &= \prod_{i=1}^n \frac{|a|y_i^{\alpha_p-1}}{b^{\alpha_p}B(p, q)[1+(y_i/b)^\alpha]^{p+q}} \frac{(a/\beta_a)^{\alpha_a-1}(b/\beta_b)^{\alpha_b-1}(p/\beta_p)^{\alpha_p-1}(q/\beta_q)^{\alpha_q-1}}{\Gamma(\alpha_a)\Gamma(\alpha_b)\Gamma(\alpha_p)\Gamma(\alpha_q)\beta_a\beta_b\beta_p\beta_q} \\
 &\times \exp(-a/\beta_a - b/\beta_b - p/\beta_p - q/\beta_q) I_{(0, \infty)}(y).
 \end{aligned} \tag{9.4.4}$$

Clearly this model does not admit a posterior from which it is easy to simulate from using simple methods such as inversion sampling.

Model 2 – g-and-h likelihood with Gamma prior distributions. The posterior distribution for this model, $p(g, h, A, B|y_{1:n})$, cannot be analytically derived. However, we assume subject matter expert elicited gamma priors $p(A|\alpha_A, \beta_A)$, $p(B|\alpha_B, \beta_B)$, $p(g|\alpha_g, \beta_g)$ and $p(h|\alpha_h, \beta_h)$, and for simplicity of presentation, constant g and h .

9.5 Simulation in Bayesian Models for Operational Risk

The overall simulation approach to generate an empirical estimate of the annual loss distribution under LDA proceeds as follows:

-
1. Simulate N realizations, $\{\theta_{1:k,(i)}\}_{i=1:N}$, from the posterior distribution using an appropriate simulation technique, (MCMC, SA-MCMC, ABC-MCMC).
 2. Using these samples of parameters to form an empirical estimate of posterior distribution one can then obtain Bayesian parameter estimates from the sample $\{\theta_{1:k,(i)}\}_{i=1:N}$ where either MAP or MMSE estimators can be empirically estimated.
 3. Apply the Bayesian model selection criterion of your choice to determine the most appropriate model given the data and the elicited expert opinions.
-

Now the application of an LDA Operational Risk approach given Bayesian parameter estimates (MAP or MMSE) proceeds as follows:

Repeat $j=1, \dots, M$ times {

- 4 Using the Bayesian estimated frequency parameter(s), sample a realization for the number of losses in the j^{th} year, $N(j)$, from the chosen frequency distribution.
- 5 Using the Bayesian estimated severity parameters, sample $N(j)$ realizations for the losses in dollars for the j^{th} year from the chosen severity distribution.
- 6 Calculate $Y(j)$ using equation (9.2.1) to obtain the annual loss for the j^{th} year.

}

For the posterior simulation in step 1, possible methods include Markov chain Monte Carlo (MCMC), importance sampling, annealed MCMC, sequential Monte Carlo (SMC) and approximate Bayesian computation (ABC). In this article we examine the MCMC, annealed MCMC and ABC in the context of Models 1 and 2 above.

9.5.1 Simulation Technique 1: Markov chain Monte Carlo

For most distributions of interest, it will not be feasible to draw samples via inversion or rejection sampling – this includes Model 1. Markov chain Monte Carlo constructs an ergodic Markov chain, $\{\theta_1, \dots, \theta_N\}$, taking values, θ_i , in a measurable space. For Model 1, $\theta = (a, b, p, q)$. The Markov chain is constructed to have the posterior distribution $p(a, b, p, q | y_{1:n})$ as its stationary distribution, thereby permitting the chain realizations to be used as samples from the posterior. These can then be used to produce MAP and MMSE estimates which have asymptotic convergence properties as the length of the chain $N \rightarrow \infty$. An excellent review of the properties of general state space Markov chain theory is given by [Meyn *et al.* 1993].

Under the general framework established by Metropolis and Hastings [Metropolis *et al.* 1953; Hastings *et al.* 1970; Gilks *et al.* 1996], transitions from one element in the chain to the next are determined via a transition kernel $K(\theta_{i-1}, \theta_i)$ which satisfies the detailed balance condition

$$p(\theta_{i-1} | y_{1:n}) K(\theta_{i-1}, \theta_i) = p(\theta_i | y_{1:n}) K(\theta_i, \theta_{i-1}) \quad (9.5.1)$$

where $p(\theta_{i-1} | y_{1:n})$ is the desired stationary distribution. The transition kernel $K(\theta_{i-1}, \theta_i)$ defines the probability of proposing to go from θ_{i-1} to θ_i . The transition kernel contains in its definition a proposal density q , from which the proposed next value in the chain is drawn, and an acceptance probability α , which determines whether the proposed value is accepted or if the chain remains in its present state. The acceptance probability is crucial as it ensures that the Markov chain has the required stationary distribution through (12) [Gilks *et al.* 1996]. In this article, for simplicity, we adopt a Gaussian random walk as the proposal density q , with variance chosen to allow efficient movement around the posterior support.

When evaluating Monte Carlo estimates for an integral, such as the posterior mean (i.e. the MMSE estimator), in order to ensure the variance of the estimate is as small as possible it is common to use only those realizations which one is fairly confident come from the Markov chain once it has reached its stationary regime. This is typically achieved by discarding a certain number of initial samples (known as “burnin”) [Gilks *et al.* 1996].

Metropolis-Hastings Algorithm for Model 1:

1. Initialize $i = 0$ and $\theta_0 = [a, b, p, q]$ randomly sampled from the support of the posterior.
2. Draw proposal θ_{i+1}^* from proposal distribution $q(\theta_i, \cdot)$.
3. Evaluate the acceptance probability

$$\alpha(\theta_i, \theta_{i+1}^*) = \min \left\{ 1, \frac{p(\theta_{i+1}^* | y_{1:n}) q(\theta_i, \theta_{i+1}^*)}{p(\theta_i | y_{1:n}) q(\theta_{i+1}^*, \theta_i)} \right\}.$$

4. Sample random variate $U \sim U[0, 1]$.
 5. If $U \leq \alpha(\theta_i, \theta_{i+1}^*)$ then set $\theta_{i+1} = \theta_{i+1}^*$ Else set $\theta_{i+1} = \theta_i$.
 6. Increment $i = i + 1$.
 7. If $i < N$ go to 2.
-

Note that this sampling strategy only requires the posterior distribution to be known up to its normalization constant. This is important as solving the integral for the normalizing constant in this model is not straightforward.

9.5.2 Simulation Technique 2: Approximate Bayesian Computation

It is not possible to apply standard MCMC techniques for Model 2, as the likelihood function, and therefore the acceptance probability, may not be written down analytically or evaluated. A recent advancement in computational statistics, approximate Bayesian computation methods are ideally suited to situations in which the likelihood is either computationally prohibitive to evaluate or cannot be expressed analytically [Beaumont *et al.* 2002; Tavaré *et al.* 2003; Sisson *et al.* 2006].

Under most ABC algorithms, evaluation of the likelihood is circumvented by the generation of a simulated dataset $y_{1:n}^*$ from the likelihood conditional upon the proposed parameter vector θ_{i+1}^* . Accordingly, this simulation method is appropriate for the g -and- h distribution as there is no general analytic expression available, although simulation from this density is computationally inexpensive. We present the MCMC version of ABC methods, introduced by [Tavaré *et al.* 2003].

Approximate Bayesian computation algorithm for Model 2:

1. Initialize $i = 0$ and $\theta_0 = [A, B, p, q]$ randomly sampled from the support of the posterior.
2. Draw proposal θ_{i+1}^* from proposal distribution $q(\theta_i, \cdot)$.
3. Generate a simulated data set $y_{1:n}^*$ from the likelihood conditional on the proposal parameters θ_{i+1}^* .
4. Evaluate the acceptance probability

$$\alpha(\theta_i, \theta_{i+1}^*) = \min \left\{ 1, \frac{p(\theta_{i+1}^*) q(\theta_{i+1}^*, \theta_i)}{p(\theta_i) q(\theta_i, \theta_{i+1}^*)} I[\rho(S(y_{1:n}^*), S(y_{1:n})) < \varepsilon] \right\}.$$

5. Sample random variate $U \sim U[0,1]$.
 6. If $U \leq \alpha(\theta_i, \theta_{i+1}^*)$ then set $\theta_{i+1} = \theta_{i+1}^*$ Else set $\theta_{i+1} = \theta_i$.
 7. Increment $i = i + 1$.
 8. If $i < N$ go to 2.
-

The difference between ABC-MCMC and standard MCMC, is that the intractable likelihood is replaced with an approximation. This approximation crudely replaces the likelihood evaluation with one when the distance (under a metric ρ) between a vector of summary statistics $S(\cdot)$ acting on both the simulated data $y_{1:n}^*$ and observed loss data $y_{1:n}$ is within an acceptable tolerance ε , and zero otherwise. The Markov chain created through this process has a stationary distribution $p(\theta, y_{1:n}^* | \rho(S(y_{1:n}^*), S(y_{1:n})) \leq \varepsilon)$, from which the desired marginal, $p(\theta | \rho(S(y_{1:n}^*), S(y_{1:n})) \leq \varepsilon)$, is easily derived. For $\varepsilon = 0$ this is exactly $p(\theta | y_{1:n})$.

Various summary statistics could be used, such as mean, variance, skewness, kurtosis and quantiles. For the illustrations in this article, we consider the sample mean, variance and skewness under Euclidean distance. One option for the tolerance level, ε , is to let it be adaptively set by the algorithm [Roberts *et al.* 2006], however, for simplicity we consider the tolerance as constant.

9.5.3 Simulation Technique 3: Simulated Annealing

Developed in statistical physics [Kirkpatrick *et al.* 1983], simulated annealing is a form of probabilistic optimization which may be used to obtain MAP parameter estimates. It implements an MCMC sampler which traverses a sequence of distributions in such a way that one may quickly find the mode of the distribution with the largest mass.

A typical sequence of distributions used is given by $[p(a, b, p, q|y_{1:n})]^{\gamma(i)}$, where $\gamma(i)$ is the “temperature” of the distribution, or the annealing schedule. When $\gamma(i) < 1$ the distribution is flat thereby permitting efficient exploration of the distribution by the Markov chain sampler. As $\gamma(i) > 1$ increases the posterior modes become highly peaked, thereby forcing the MCMC sampler towards the MAP estimate, which is taken as the final chain realization. The approach taken in this article will be to perform annealing with a combination of 100 Markov chain samplers started at random locations in the support of the posterior using a logarithmic annealing schedule.

Simulated Annealing algorithm:

1. Initialise $i = 0$ and $\theta_0 = [a, b, p, q]$ randomly sampled from the support of the posterior.
2. Draw proposal θ_{i+1}^* from proposal distribution $q(\theta_i, \cdot)$.
3. Evaluate the acceptance probability

$$\alpha(\theta_i, \theta_{i+1}^*) = \min \left\{ 1, \frac{p(\theta_{i+1}^*|y_{1:n})^{\gamma(i+1)} q(\theta_{i+1}^*, \theta_i)}{p(\theta_i|y_{1:n})^{\gamma(i+1)} q(\theta_i, \theta_{i+1}^*)} \right\}.$$

4. Sample random variate $U \sim U[0,1]$.
 5. If $U \leq \alpha(\theta_i, \theta_{i+1}^*)$ then set $\theta_{i+1} = \theta_{i+1}^*$ Else set $\theta_{i+1} = \theta_i$.
 6. Increment $i = i + 1$.
 7. If $i < N$ go to 2.
-

9.6 Simulation Results and Discussion

9.6.1 Parameter Estimation and Simulation from Model 1: GB2 severity distribution

It was demonstrated in [Bookstaber et al, 1987] that the GB2 distribution (9.4.1) encapsulates many different classes of distribution for certain limits on the parameter set. These include the lognormal ($a \rightarrow 0, q \rightarrow \infty$), log Cauchy ($a \rightarrow 0$) and Weibull/Gamma ($q \rightarrow \infty$) distributions, all of which are important severity distributions used in operational risk models in practice.

We now construct a synthetic dataset with which to illustrate the above methods. The severity loss data will be generated from a Lomax distribution (Pareto distribution of the 2nd kind or Johnson Type VI distribution) with parameters $[a, b, p, q] = [1, b, 1, q]$, which when reparameterized in terms of $\alpha > 0, \lambda > 0$ and $x > 0$, is given by

$$h(x; \alpha, \lambda) = \frac{\alpha}{\lambda} \left(1 + \frac{x}{\lambda}\right)^{-(\alpha+1)}. \quad (9.6.1)$$

The heavy-tailed Lomax distribution has shape parameter $\alpha (= q)$ and scale parameter $\lambda (=b)$. The Lomax distribution has been selected for this study for two reasons. The first reason is of a practical nature, in practical applications of Operational Risk Modelling, it is believed that heavy tailed severity distributions should be considered in order to capture the characteristics of actual Operational Risk severity observations. It is standard industry practice in the area of Operational Risk modeling in Banking to consider such distributions, and discussed in [Dutta *et al.* 2006]. The second, less significant reason for consideration of the Lomax distribution, is that it provides a practically viable distribution from which to sample the loss data set to be analysed. We generate 150 artificial loss values in thousands of dollars. In order to generate these data we use the relation whereby the 4 parameter GB2 distribution can be obtained through the transformation of $X_1 \sim \text{Gamma}(p,1)$ and $X_2 \sim \text{Gamma}(q,1)$ via [Devroye, 1986]

$$Y = b \left(\frac{X_1}{X_2} \right)^{1/a}. \quad (9.6.2)$$

Accordingly we use a GB2 distribution with parameters [1, 10, 1, 1]. Figure 9.7.1 demonstrates the shape of this density.

We further specify the parameters for the Gamma prior distributions as $\alpha_b = 5, \beta_b = 2, \alpha_a = \alpha_p = \alpha_q = 0.5$ and $\beta_a = \beta_p = \beta_q = 2$.

Results for Markov chain Monte Carlo:

We implement MCMC of length $N=500,000$ with the initial 100,000 iterations discarded as “burnin”, and using a truncated multivariate Gaussian distribution

$MVN([a,b,p,q], \sigma_{prop} I_4) I([a,b,p,q] > [0,0,0,0])$ as the random walk proposal density, with mean given by $[a,b,p,q]$ and covariance matrix $\sigma_{prop}^2 I_4$, where $\sigma_{prop} = 0.1$ and I_4 is the 4 by 4 identity matrix and $I(\dots)$ is the indicator function which is one when the logical condition is met and zero otherwise.

Figure 9.7.2 illustrates sample paths of a, b, p and q . These paths demonstrate that the Markov chain produced through this simulation process is mixing well over the support of the posterior distribution and is exploring the regions of support for the posterior which coincide with the true parameter values, thereby demonstrating correct sampler performance.

The resulting MAP estimates are shown in Figure 9.7.3 as the modes of the histogram density estimates (see results for simulated annealing). The posterior mean (MMSE) estimates and posterior standard deviations were approximated using the Markov chain realizations as, $a_{MMSE} = 1.4249$ (*std.dev.* 0.6612), $b_{MMSE} = 9.6641$ (*std.dev.* 2.7267), $p_{MMSE} = 0.9079$ (*std.dev.* 0.5404) and $q_{MMSE} = 0.8356$ (*std.dev.* 0.4007).

Results for Simulated Annealing

We implement the simulated annealing algorithm with 100 random selected starting points

drawn from a Gaussian distribution $N([a,b,p,q],I)$ $I([a,b,p,q] > [0,0,0,0])$. These starting points are then used to produce 100 parallel MCMC chains which each used the same proposal and prior specifications the MCMC simulation. The annealing schedule length was $N = 1000$, and the schedule γ_i was logarithmic between $\log(e)$ and $\log(1000)$. We note that simulated annealing is appropriate in this setting since the LDA approach to modeling operational risk only requires point estimates for the parameters of the severity and frequency distributions. In this case a probabilistic optimization technique has been used to obtain the point estimates. The mean MAP estimates and the standard deviations of the 100 MAP estimates obtained from the parallel chains are $\bar{a}_{MAP}=1.0181$ (*std.dev.* 0.2908), $\bar{b}_{MAP} = 10.0434$ (*std.dev.* 1.3343), $\bar{p}_{MAP} = 1.3617$ (*std.dev.* 0.6927) and $\bar{q}_{MAP} = 1.2017$ (*std.dev.* 0.6901). Figure 9.7.4 illustrates the 100 MAP estimates obtained for the 100 (random initial starting points) parallel annealed MCMC chains. These estimates were then averaged to get the mean MAP and standard deviation estimates above. More consistent results between chains can be achieved through longer chains and more gradual annealing schedules.

There is a trade-off between using simulated annealing and MCMC in Model 1. Annealing stochastically optimizes the posterior distribution and is typically very fast in terms of simulation time. The total simulation time to produce all the 100 parallel chains took less than 6 minutes in non-optimized Matlab code on a P4 2.4GHz laptop with 480MB of RAM. However, annealing provides MAP estimates of parameters, whereas the MCMC estimates the entire posterior distribution. Similarly, a trade off exists between the length of the annealing schedule and the number of annealing chains.

9.6.2 Parameter Estimation and Simulation from Model 2: *g*-and-*h* severity distribution

We now demonstrate the use of the *g*-and-*h* distribution for modeling the severity distribution in an LDA operational risk framework. Posterior simulation from Model 2 requires the use of approximate Bayesian computation methods. We generate an observed dataset of length 150 using the parameters $A = 1$, $B = 1$, $g = 2$ and $h = 1$, as illustrated in Figure 9.7.5, after truncation is applied. Parameters for the Gamma priors are specified as $\alpha_a = \alpha_b = \alpha_q = 0.5$, $\beta_a = \beta_b = \beta_q = 2$, $\alpha_p = 2$ and $\beta_p = 1$.

The ABC-MCMC algorithm requires specification of both summary statistics $S()$ and tolerance level ε . Accordingly, after performing an initial analysis we examine the effect of varying firstly the tolerance and then the summary statistics.

Initial ABC analysis

For our initial analysis we specify the sample mean as the sole summary statistic, and a tolerance level of $\varepsilon = 0.1$. The Markov chain has length $N=500,000$, discarding the first 100,000 realizations. The Markov chain proposal distribution q was again the truncated multivariate Gaussian distribution, but with $\sigma_{prop} = 1$. The sample paths of the parameters A , B , g and h are illustrated in Figure 9.7.6, indicating that the Markov chain is mixing well for these data.

We note that the ABC-MCMC algorithm may become “stuck” in one part of the parameter

space place for an extended period of time. This occurs as the acceptance rate of any state is proportional to the posterior at that point, and so when the current state is in an area of relatively low probability it may take a while to escape from this point. [Bortot *et al.* 2006; Sisson *et al.* 2006] have examined different methods of permitting the tolerance to vary to mitigate this problem.

The resulting posterior mean (MMSE) and standard deviation estimates are $A_{MMSE} = 1.7489$ (*std.dev.* 1.4265), $B_{MMSE} = 1.3672$ (*std.dev.* 2.0456), $g_{MMSE} = 1.5852$ (*std.dev.* 0.8202) and $h_{MMSE} = 0.4286$ (*std.dev.* 0.4038).

These results are acceptable given that only the sample mean was used as a summary statistic to capture all information from the likelihood. We might anticipate improved estimates if the summary statistics were augmented to include other quantities. We examine this shortly. We firstly consider the effect of varying the tolerance level ε .

ABC Analysis of tolerance levels

We present a summary of results for the same analysis as before, but where the tolerance level ε takes the values 0.001, 0.01, 0.1 and 1 and the Markov chain proposal distribution q was the truncated multivariate Gaussian distribution, but with $\sigma_{prop} = 0.5$. Parameter sample paths for $\varepsilon = 0.001$ are illustrated in Figure 9.7.7. Instances of the chain becoming stuck are now more pronounced. The simplest methods to avoid this problem are to run a much longer chain, or to choose a larger tolerance, although see [Bortot *et al.* 2006; Sisson *et al.* 2006]. For tolerances greater than 0.01, the sticking was sufficiently minor to justify using the Markov chain realizations as a sample from the posterior. In the following simulations the length of the Markov chain was increased to $N = 1,000,000$ with the first 400,000 realizations discarded as burnin.

With $\varepsilon = 0.01$, the resulting MMSE and standard deviation estimates were $A_{MMSE} = 0.9693$ (*std.dev.* 0.7379), $B_{MMSE} = 1.4551$ (*std.dev.* 1.0447), $g_{MMSE} = 1.7619$ (*std.dev.* 0.4545) and $h_{MMSE} = 0.5513$ (*std.dev.* 0.3626).

With $\varepsilon = 0.1$, the resulting MMSE and standard deviation estimates were $A_{MMSE} = 1.6741$ (*std.dev.* 1.4137), $B_{MMSE} = 1.3130$ (*std.dev.* 1.6235), $g_{MMSE} = 1.5651$ (*std.dev.* 0.7915) and $h_{MMSE} = 0.4652$ (*std.dev.* 0.4435).

With $\varepsilon = 1$, the resulting MMSE and standard deviation estimates were $A_{MMSE} = 1.8940$ (*std.dev.* 1.5130), $B_{MMSE} = 1.6580$ (*std.dev.* 1.9256), $g_{MMSE} = 1.5130$ (*std.dev.* 0.7469) and $h_{MMSE} = 0.4887$ (*std.dev.* 0.5373).

These results demonstrate that as the tolerance decreases the precision of the estimates improves – this is to be expected. As the tolerance decreases, fewer “unlikely” parameter realizations generate data which is within the tolerance level ε . Accordingly the precision of the posterior distribution increases. However, on average a proposal θ_{i+1}^* will be accepted with probability $\Pr(\rho(S(y_{1:n}^*), S(y_{1:n})) \leq \varepsilon | \theta_{i+1}^*)$, and this will decrease (rapidly) as the tolerance level ε decreases. That is, chain efficiency is strongly linked to the tolerance – if one is increased, the other decreases, and vice versa. However, commonly an acceptable accuracy and

computational intensity trade off can be found.

ABC analysis of summary statistics

We now examine the effect of using different summary statistics. We consider the combinations {mean}, {mean, variance}, {mean, skewness} and {mean, variance, skewness} to approximate the information in the data through the likelihood. The Markov chain proposal distribution q was again the truncated multivariate Gaussian distribution, this time using $\sigma_{prop} = 0.1$. With tolerance fixed at $\varepsilon=1$, our chain length is $N=1,000,000$ with the initial 200,000 iterations discarded. The resulting MMSE and standard deviation estimates are as follows:

{mean}

$A_{MMSE} = 1.77431$ (std.dev. 1.5265), $B_{MMSE} = 1.1933$ (std.dev. 1.4874), $g_{MMSE} = 1.4837$ (std.dev. 0.781) and $h_{MMSE} = 0.4850$ (std.dev. 0.5354).

{mean, variance}

$A_{MMSE} = 0.7369$ (std.dev. 0.5318), $B_{MMSE} = 1.3678$ (std.dev. 1.0177), $g_{MMSE} = 1.7093$ (std.dev. 0.5202) and $h_{MMSE} = 0.6571$ (std.dev. 0.3807).

{mean, skew}

$A_{MMSE} = 1.4955$ (std.dev. 1.1876), $B_{MMSE} = 1.1724$ (std.dev. 1.0471), $g_{MMSE} = 1.7037$ (std.dev. 0.4674) and $h_{MMSE} = 0.3753$ (std.dev. 0.4105).

{mean, variance, skew}

$A_{MMSE} = 1.0049$ (std.dev. 0.5656), $B_{MMSE} = 1.0211$ (std.dev. 0.4148), $g_{MMSE} = 1.7385$ (std.dev. 0.3202) and $h_{MMSE} = 0.7608$ (std.dev.0.2699).

Clearly the latter case produces the most accurate and precise results (as shown by parameter estimates and standard deviations respectively) as the largest set of summary statistics naturally encapsulates distributional characteristics most precisely. Conversely, the most inaccurate results are obtained when the least restrictive criterion is applied.

Additionally one can relate these results back to the role each parameter plays in the shape of the g -and- h distribution. For example, parameter g relates to the skewness of the distribution and as such, one might expect improved performance in estimates of this parameter when skewness is incorporated as a summary statistic. The results support such a hypothesis as the mean is more precise and standard deviation reduced in the posterior marginal distribution for g when skewness is included. Similarly, h relates to kurtosis and accordingly realizes improved estimates upon inclusion of the variance.

However, as expanding the vector of summary statistics increases the number of restrictions placed on the data, proposed parameter values become less likely to generate data sets which will satisfy the selection criteria for a fixed tolerance. Accordingly the acceptance rate of the Markov chain will decrease, thereby necessitating longer runs to obtain reliable sample estimates.

9.7 *Discussion*

In this article, by introducing existing and novel simulation procedures we have substantially extended the range of models admissible for Bayesian inference under the LDA operational risk modeling framework. We strongly advocate that Bayesian approaches to operational risk modeling should be considered as a serious alternative for practitioners in banks and financial institutions, as it provides a mathematically rigorous paradigm in which to combine observed data and expert opinion. We hope that the presented class of algorithms provides an attractive and feasible approach in which to realize these models.

Future work will consider introduction of correlation in a Bayesian setting, such as correlation between parameters for frequency and severity distributions under an LDA approach.

Acknowledgements

The first author is supported by an Australian Postgraduate Award, through the Department of Statistics at UNSW. The second author is supported by the Australian Research Council through the Discovery Project scheme (DP0664970) and by the Australian Center of Excellence for Risk Analysis.

Appendix

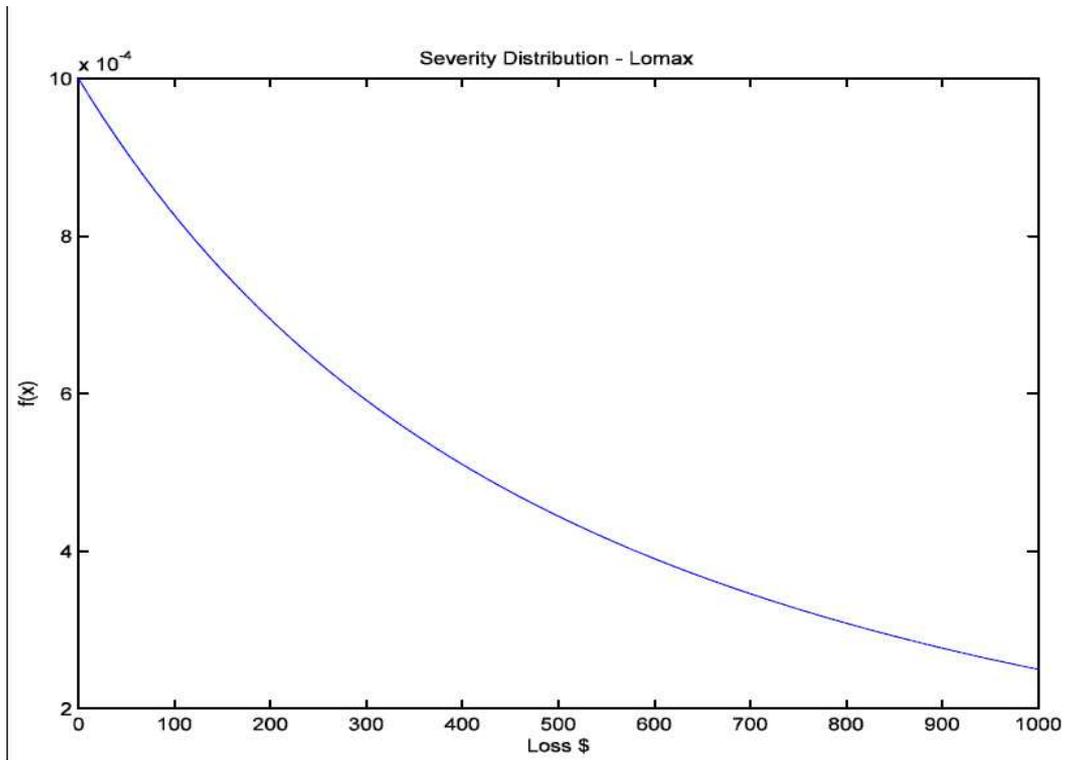


Fig. 9.7.1: GB2 distribution with parameters [1, 10, 1, 1]

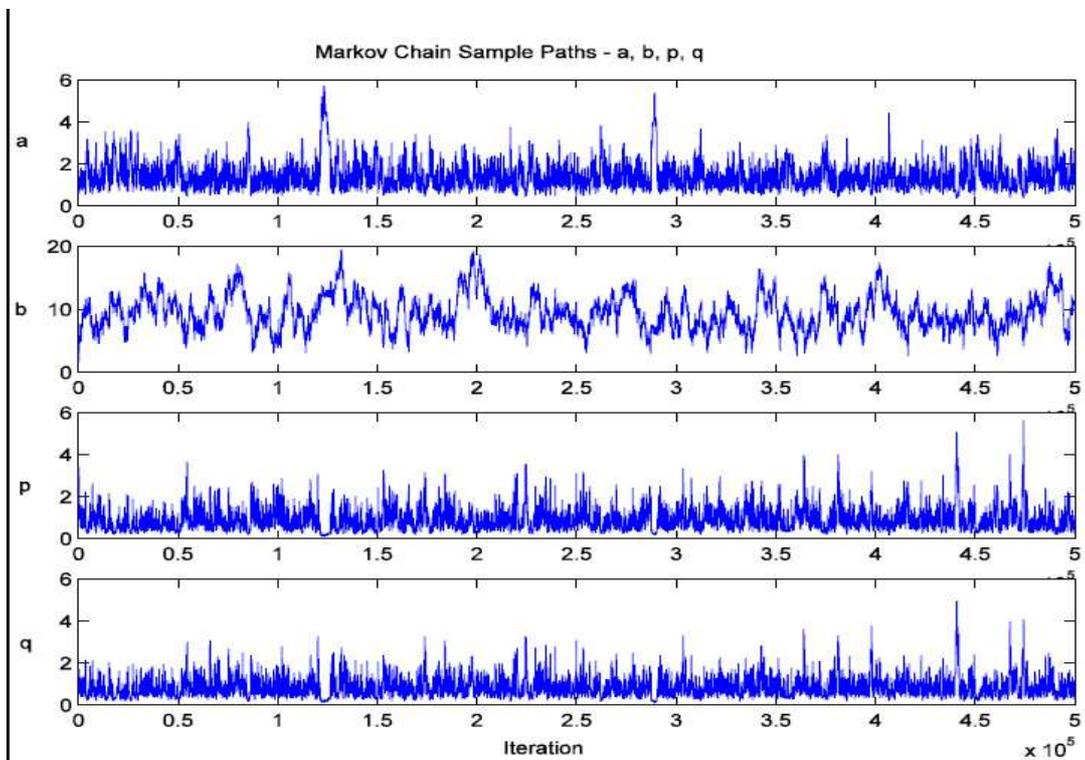


Fig. 9.7.2: Markov chain sample paths for parameters a, b, p and q

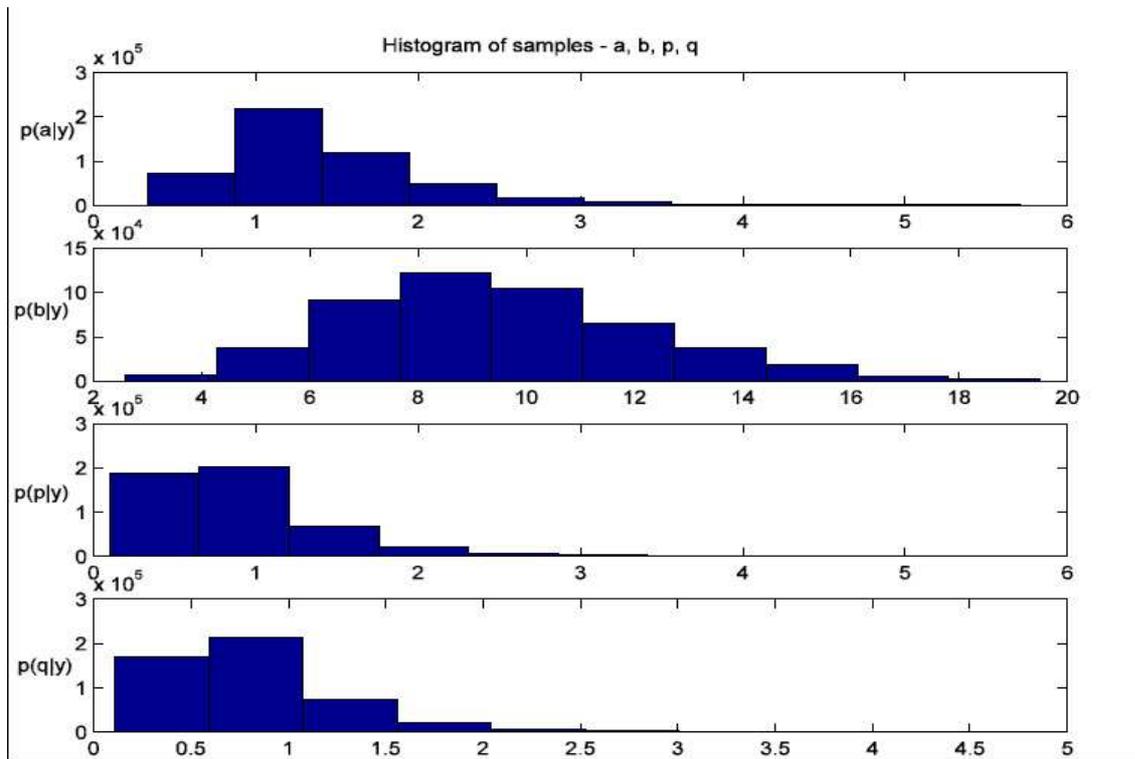


Fig. 9.7.3: Maximum *a posteriori* (MAP) estimates for the mode from Markov chain

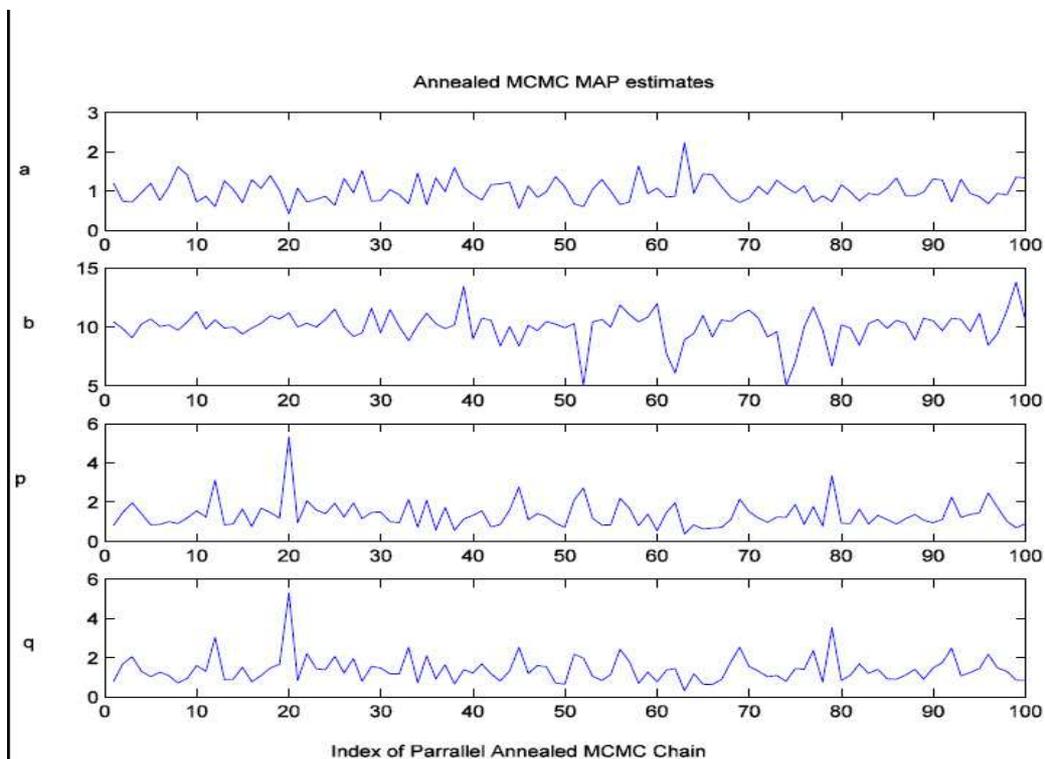


Fig. 9.7.4: Maximum *a posteriori* (MAP) estimates, 100 values from 100 unique starting locations

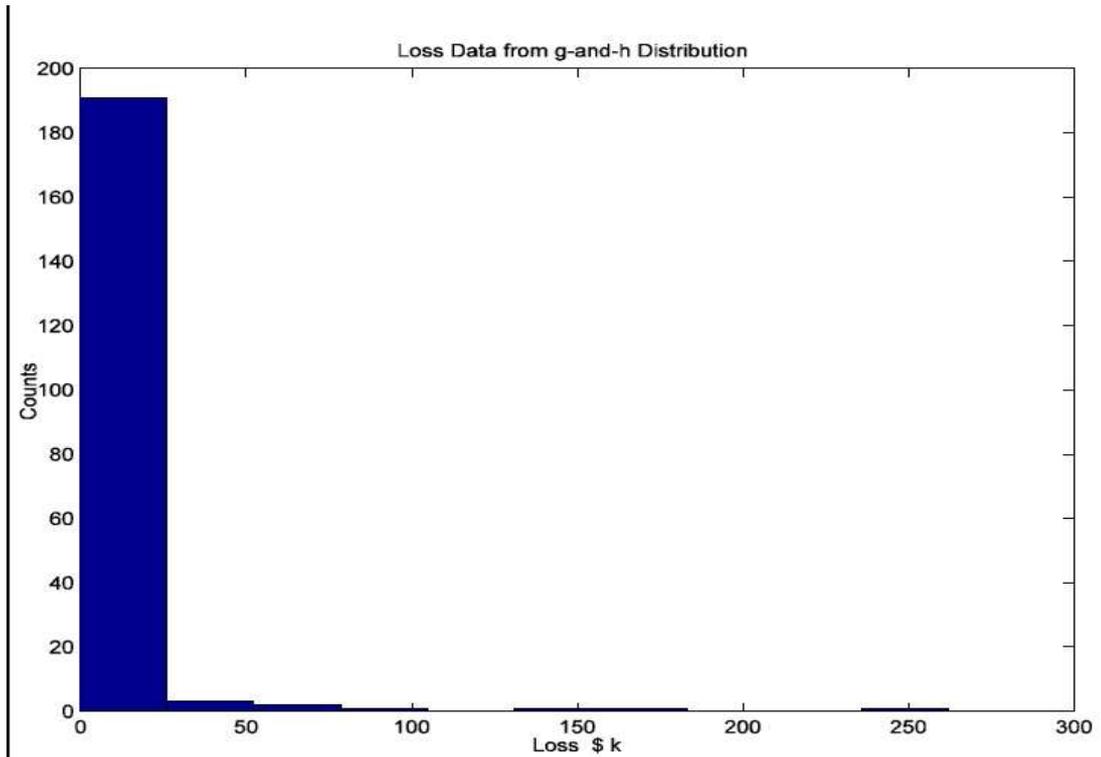


Fig. 9.7.5: g-and-h data set generated with parameters $A = 1$, $B = 1$, $g = 2$ and $h = 1$

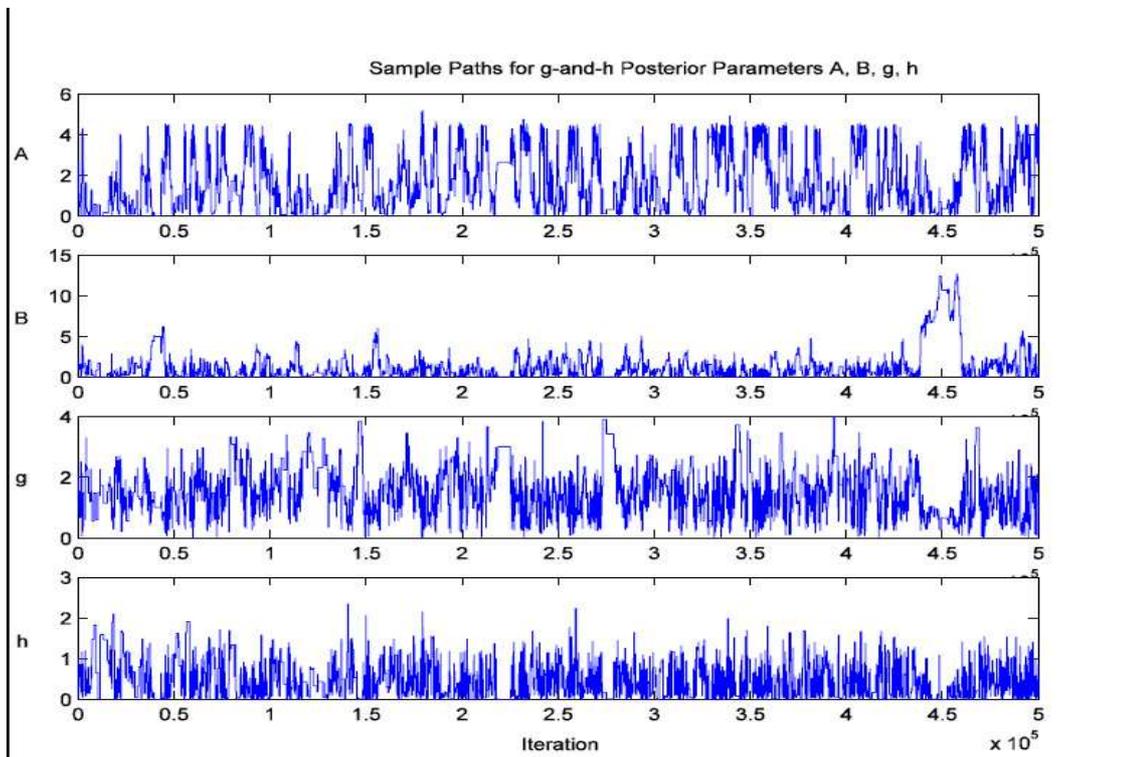


Fig. 9.7.6: Markov chain sample paths for parameters A, B, g and h

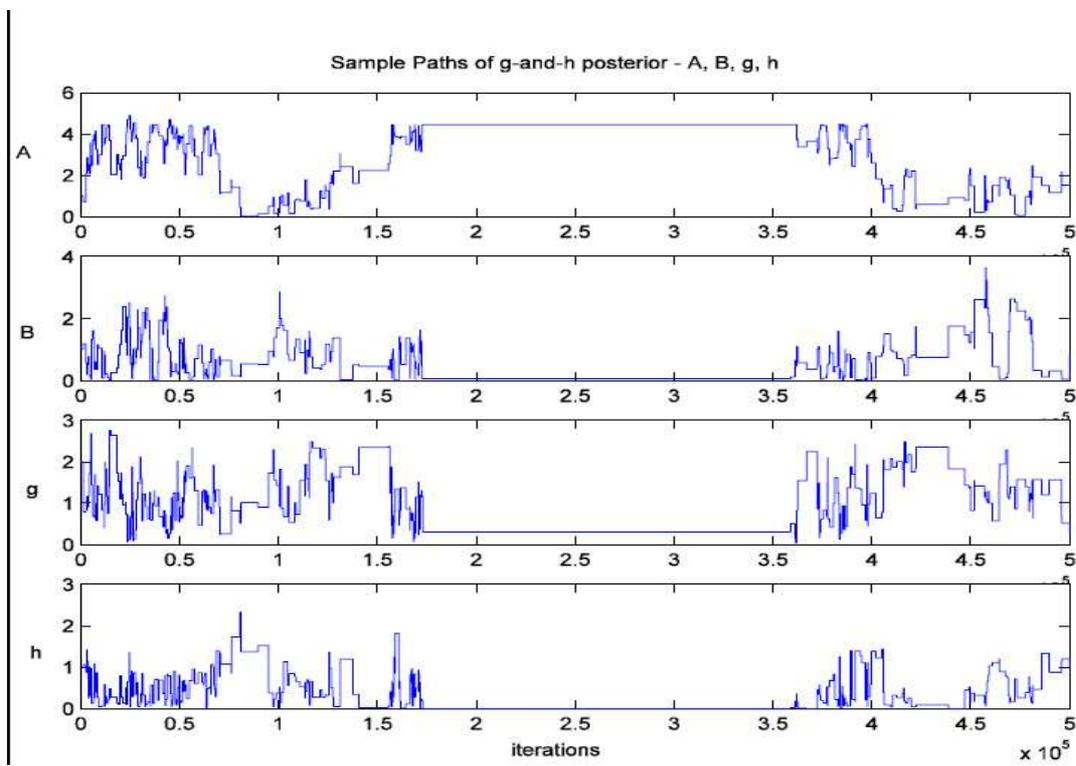


Fig. 9.7.7: Parameter sample paths for tolerance level $\epsilon = 0.001$

References

- [1] Bayes, T. (1763). *An essay towards solving a problem with the doctrine of Chances*. Philos. Trans. R. Soc. London, **53**, 370—418.
- [2] Bee M. (2006). *Estimating and simulating loss distributions with incomplete data*, *Oprisk and Compliance*, **7** (7), 38-41.
- [3] Bernardo J. and A. Smith (1994). *Bayesian Theory*. Wiley Series in Probability and Statistics, Wiley.
- [4] Bookstaber R. and J. McDonald (1987). *A general distribution for describing security price returns*. The Journal of Business, **60**(3), 401-424.
- [5] Bortot P., S. G. Coles and S. A. Sisson (2006). *Inference for stereological extremes*. J. Amer. Stat. Assoc. In press.
- [6] Beaumont M. A., W. Zhang and D. J. Balding (2002). *Approximate Bayesian computation in population genetics*. Genetics, **162**, 2025—2035.
- [7] Box and Tiao (1992) *Bayesian Inference in Statistical Analysis*. Wiley Classics Library.
- [8] Cruz M. (2002). *Modelling, Measuring and Hedging Operational Risk*. John Wiley & Sons, Chapter 4.
- [9] Devroye L. (1986). *Non-Uniform Random Variate Generation*. Springer-Verlag, New York.
- [10] Doucet A., P. Del Moral and A. Jasra (2006). *Sequential Monte Carlo samplers*, J. Roy. Statist. Soc. B, **68**(3), 411-436.
- [11] Dutta K. and J. Perry (2006). *A tale of tails: An empirical analysis of loss distribution models for estimating operational risk capital*. Federal Reserve Bank of Boston, Working Papers No. 06-13.
- [12] Embrechts P., H. Furrer and R. Kaufmann (2003). *Quantifying regulatory capital for operational risk*. Derivatives Use, Trading & Regulation, **9**(3), 217-223.
- [13] Garthwaite P. and A. O'Hagan (2000). *Quantifying expert opinion in the UK water industry: An experimental study*. The Statistician, **49**(4), 455-477.
- [14] Gelman A., J. B. Carlin, H.S. Stern and D.B. Rubin (1995). *Bayesian Data Analysis*. Chapman and Hall.

- [15] Gilks W., S. Richardson and D. Spiegelhalter (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall.
- [16] Hastings W. (1970). *Monte Carlo sampling methods using Markov Chains and their Applications*. *Biometrika*, **57**, 97–109.
- [17] HeY. and Raghunathan T. (2006). *Tukey's gh Distribution for Multiple Imputation*, *The American Statistician*, Vol. 60, **3**
- [18] Kirkpatrick S., C. Gelatt and M. Vecchi (1983). *Optimization by simulated annealing*. *Science*, **220**, 871-680.
- [19] Metropolis N., A. Rosenbluth, M. Rosenbluth, A. Teller, E. and Teller (1953). *Equations of state calculations by fast computing machines*. *J. Chem. Phys.*, **21**, 1087-1091.
- [20] Meyn S. and R. Tweedie (1993). *Markov Chains and Stochastic Stability*, Springer.
- [21] O'Hagan A. (2006). *Uncertain Judgements: Eliciting Expert's Probabilities*, Wiley, *Statistics in Practice*.
- [22] O'Hagan A. (1998). *Eliciting expert beliefs in substantial practical applications*. *The Statistician*, **47**(1), 21-35.
- [23] Panjer H. (2006). *Operational Risk: Modeling Analytics*, Wiley.
- [24] Peters G. (2005). *Topics in Sequential Monte Carlo Samplers*. University of Cambridge, M.Sc. Thesis, Department of Engineering.
- [25] Ramamurthy S., H. Arora and A. Ghosh (2005). *Operational risk and probabilistic networks – An application to corporate actions processing*. Infosys White Paper.
- [26] Robert C. (2004). *The Bayesian Choice, 2nd Edition*. Springer Texts in Statistics.
- [27] Roberts G. O. and J. Rosenthal (2006). *Examples of adaptive MCMC*. Technical Report, Lancaster University.
- [28] Ruanaidh, J. and W. Fitzgerald (1996). *Numerical Bayesian Methods Applied to Signal Processing*. Statistics and Computing, Springer.
- [29] Shevchenko, P. and M. Wuthrich (2006). *The structural modelling of operational risk via Bayesian inference: Combining loss data with expert opinions*. CSIRO Technical Report Series, CMIS Call Number 2371.
- [30] Schwarz, G., (1978). *Estimating the dimension of a model*, *Ann. Statist.*, **6**, 461-464.
- [31] Sisson S. A., Y. Fan and M. M. Tanaka (2006). *Sequential Monte Carlo without likelihoods*. Technical Report, University of New South Wales.
- [32] Spiegelhalter, D., Best, G., Carlin, B. and van der Linde, A., (2002), *Bayesian Measures of Model Complexity and Fit*, *J. R. Statist. Soc. B*, **62**, Part 4, 583-639
- [33] Tavaré S., P. Marjoram, J. Molitor and V. Plagnol (2003). *Markov Chain Monte Carlo without Likelihoods*, *Proc. Natl. Acad. Sci. USA*, **100**, 15324-15328.
- [34] Wasserman, L. (1997), *Bayesian Model Selection and Model Averaging*, CMU Department of Statistics Technical Reports, **666**.

Journal Paper 7

"All things are ready, if our minds be so."

William Shakespeare

Peters G.W. Johansen A. and Doucet A. (2007) "Simulation of the Annual Loss Distribution in Operational Risk via Panjer Recursions and Volterra Integral Equations for Value at Risk and Expected Shortfall Estimation". *Journal of Operational Risk*, 2(3).

This work was instigated by the first author who can claim around 80% of the credit for the contents. This paper has already been cited numerous times by people in industry and in academic research in this field. Additionally, this work was presented at an international Statistical conference in Queensland and received positive feedback from academics after the talk. The first authors work included developing the methodology contained, developing the applications, the implementation and comparison to alternative approaches, writing the drafts of the paper and undertaking revisions. This paper was accepted for publication in the *Journal of Operational Risk* and has appeared. Permission from all the co-authors is granted for inclusion in the PhD.

This paper appears in the thesis in a modified format to that which finally appeared in the *Journal of Operational Risk*, where it was published.

Final print version available at: <http://www.journalofoperationalrisk.com/>

Simulation of the Annual Loss Distribution in Operational Risk via Panjer Recursions and Volterra Integral Equations for Value at Risk and Expected Shortfall Estimation

Gareth W. Peters (*corresponding author*)

CSIRO Mathematical and Information Sciences, Locked Bag 17, North Ryde, NSW, 1670, Australia;

UNSW Mathematics and Statistics Department, Sydney, 2052, Australia;

email: peterga@maths.unsw.edu.au

A. M. Johansen

Department of Mathematics, University of Bristol, Bristol, United Kingdom

A. Doucet

Departments of Statistics & Computer Science, University of British Columbia, Vancouver, Canada

10.1 Abstract

Following the Loss Distributional Approach (LDA), this article develops two procedures for simulation of an annual loss distribution for modeling of Operational Risk. First, we provide an overview of the typical compound-process LDA used widely in Operational Risk modeling, before expanding upon the current literature on evaluation and simulation of annual loss distributions. We present two novel Monte Carlo simulation procedures. In doing so, we make use of Panjer recursions and the Volterra integral equation of the second kind to reformulate the problem of evaluation of the density of a random sum as the calculation of an expectation. We demonstrate the use of importance sampling and trans-dimensional Markov Chain Monte Carlo algorithms to efficiently evaluate this expectation. We further demonstrate their use in the calculation of Value at Risk and Expected Shortfall.

Keywords: Importance Sampling; Trans-dimensional Markov Chain Monte Carlo; Basel II Advanced Measurement Approach; Panjer Recursions; Volterra Integral Equations; Compound Processes; Loss Distributional Approach; Operational Risk; Value at Risk; Expected Shortfall.

10.2 Introduction

Through the Advanced Measurement Approach, financial institutions are permitted significant flexibility over the methodology that may be used in the development of operational risk models. This has led to the consideration of numerous approaches to modeling Operational Risk to satisfy the Basel II regulatory requirements. Such models incorporate internal and external loss data observations in combination with expert opinion surveyed from business subject matter experts. The focus of this paper will be on the popular Loss Distributional Approach (LDA) to modeling Operational Risk.

The idea of LDA is to fit severity and frequency distributions over a predetermined time horizon, typically annual. Popular choices for severity distributions include exponential, Weibull, lognormal, generalised Pareto, GB2 and g -and- h distributions [Dutta *et al.*, 2006; Shevchenko *et al.*, 2006; Peters *et al.* 2006] whilst those commonly used for frequency distributions include Poisson, binomial and negative binomial distributions [Dutta *et al.* 2006]. The fitted models are then used to define a compound process for the annual loss distribution. Value at Risk (VaR), Expected Shortfall (ES) and other capital estimates may then be derived under the compound process,

$$Y = \sum_{i=1}^M X_i, \quad (10.2.1)$$

where the mutually independent random variables, $X_i \sim f_X$ and $M \sim h$, are distributed according to the fitted severity distribution and frequency distribution, respectively.

This paper considers alternative approaches to standard Monte Carlo simulation for the evaluation of the density of Y , which we denote throughout as f_Y . In general, the distribution of Y has no closed analytic form as it involves an infinite sum, whose m^{th} term corresponds to the m -fold convolution of the severity distribution weighted by the probability $Pr(M=m)$ under the chosen frequency distribution. Actuarial research has considered the distribution of Y for insurance purposes through Panjer recursions [Panjer, 2006; Sundt *et al.*, 1981; Willmot *et al.*, 1985]. Other approaches utilize inversion techniques such as inverse Fourier transforms to approximate annual loss distributions, although they typically require assumptions such as independence between frequency and severity random variables [Embrechts *et al.* 2003]. Techniques commonly adopted to fit frequency and severity models in the Operational Risk modeling literature include extreme value theory [Cruz, 2002; Chavez-Demoulin *et al.*, 2006; Neslehova *et al.*, 2006], Bayesian inference [Shevchenko *et al.* 2006; Cruz, 2002; Peters *et al.* 2006], dynamic Bayesian networks [Ramamurthy *et al.* 2005], Maximum Likelihood [Dutta *et al.* 2006] and Expectation Maximization to find Maximum Likelihood parameter estimates [Bee, 2006].

The simulation of an annual loss distribution is critical for measurement of risk estimated from a loss distribution. The current regulatory requirements specify a 0.999 quantile of this loss distribution, for discussion see [Franklin *et al.*, 2007]. Obtaining accurate estimates at this level requires significant computational effort.

The most popular approach to simulation of an annual loss distribution in practice, which we term standard Monte Carlo, is to first sample N realizations $\{m_i\}_{i=1:N}$ of M , the number of annual events, from the fitted frequency distribution. Then for the i^{th} year with m_i loss events, sample each loss severity from the fitted parametric severity distribution and calculate the sum in equation (10.2.1). It is also popular to utilize correlation in such models, typically introduced through a copula transform for the simulated loss values, usually in the form of either frequency or severity correlation. This allows the construction of a histogram estimate of the annual loss distribution and the required quantiles.

In this paper we demonstrate alternative approaches, which provide a more efficient means of simulating tail quantiles and conditional expectations in the tails of an annual loss distribution to estimate quantities such as Expected Shortfall. Our procedure is easy to parallelize which is of critical importance in real practical simulations performed in financial institutions. It utilizes Panjer recursions [Panjer, 2006; Willmot *et al.*, 1985], Importance Sampling [Glasserman, 2003; Doucet *et al.*, 2007] and trans-dimensional Markov Chain Monte Carlo [Green, 1995; Green, 2003]. We will focus on the setting in which the severity and frequency distributions take a parametric form which admits explicit density representation, and will discuss briefly how to extend our approach to settings in which no closed form parametric density is available.

10.3 Panjer Recursions and the Volterra Integral Equation.

The evaluation of distributions of random sums or compound processes has been ubiquitous in actuarial mathematics for many decades [Panjer, 1981; Panjer, 1992; Panjer 2006; Willmot, 1985; Stroter, 1984; Klugman, 2004]. It typically arises as the preferred method of modeling the probability of ruin and the loss distributions for insurance claims that can be assumed to arrive according to, for example, a Poisson or negative binomial compound process. That is, they are typically considered when modeling the distribution of the total claims incurred in a fixed period of time. What makes the explicit computation of these loss distributions difficult is that the conditional distribution of the amount of total claim or loss given a certain number of claims m has occurred involves an m -fold convolution of the severity distribution.

As alluded to in the introduction, Monte Carlo simulation is usually employed to approximate these m -fold convolutions. However, at the extreme values of the mean annual number of events large or small the standard Monte Carlo approach becomes extremely computationally inefficient. Additionally, for any mean arrival rate, trying to achieve a given accuracy for an estimate of Value at Risk or Expected Shortfall, may require a significant computational effort. Below, we propose a simulation procedure which reduces this computational burden.

In Operational Risk one is typically concerned with rare and infrequent events, which if they do occur, can have catastrophic consequences for the annual loss of a given year. This typically corresponds to the situation in which the mean annual loss is small and the mean severity of the given losses is very large. We will focus in this paper on the most important case for Operational Risk and that is the infrequent yet catastrophic event situations. We will demonstrate that

our approach is a very efficient means of accurately estimating VaR (a tail quantile of the loss distribution) and ES deep in the tails of the annual loss distribution. We will additionally point out that given we know the starting point y_s for the $\text{VaR}_\alpha(Y) = y_s$ at which one wants to calculate the expected shortfall,

$$ES_\alpha [Y] = \frac{1}{1 - \alpha} \int_\alpha^1 \text{VaR}_u(Y) du = E_{f_Y} [Y|Y > y_s] = \frac{1}{1 - F_Y(y_s)} \int_{y_s}^\infty y f_Y(y) dy.$$

In such cases our method provides such a solution without computing the annual loss distribution for the domain $[0, y_s]$. In other words our procedure begins the estimation of the annual loss distribution on $[y_s, \infty)$. ES is an important measure of risk since it has the property of coherence, which is not the case for VaR. This provides a new and efficient framework for estimating ES, which as far as the authors can tell has previously been restricted to point-wise estimation of the entire annual loss density up to y_s before the estimation of the distribution in the region of interest could be performed.

In practice we may need to calculate a $\text{VaR}_\alpha(Y)$ to get y_s first before calculation of ES, in these cases we present a fast and efficient algorithm to perform both the calculation of y_s and the calculation of ES. However, there will be some cases in which the value of y_s is known in advance. The initial value of y_s could be known from a previous VaR calculation. Additionally computation of ES without the construction of the entire loss distribution can be valuable for insurance purposes. If a haircut is known to occur after a level y_s then one may be interested in efficiently calculating the Expected loss ignoring the insurance policy and then calculating the Expected loss including the insurance deductions and comparing the excess to the uninsured expected loss to decide between insurance policies.

Our target is to evaluate the compound distribution for the annual loss which is described by equation (10.2.1). In these situations actuarial techniques can prove to be effective means of evaluating an annual loss distribution point-wise. The approach we will review here is the most popular of these known as the Panjer recursion.

If the severity distributions are discrete, then efficient, deterministic techniques based upon the z-transform may be employed but this approach does not generalize to continuous severity distributions. Alternatively, the Panjer recursion provides a recursive expression for evaluation of the coefficients c_k of the probability generating function $P(z)$.

Although, in some settings, discretization of a continuous severity distribution might be justifiable, this is not the preferred approach to most Operational Risk models. However, the Panjer recursion approach mentioned above may also be applied in a continuous setting, leading to the recursion:

$$f_Y(x) = p_1 f_X(x) + \int_0^x \left(a + \frac{by}{x} \right) f_X(y) f_Y(x-y) dy \quad (10.3.1)$$

Where a, b and p_1 parameterize the frequency distribution of the compound process. There are many approaches to evaluate this expression. These numerical techniques will be discussed in the next section along with mention of alternatives to a Panjer recursion, such as inversion

transforms (Fast Fourier Transforms) and series expansions (Bergstrom), see [Cruz, 2002; Panjer, 2006; Menn, 2006].

The property of the Panjer recursion that we will exploit in developing our simulation algorithms is that (10.3.1) can be recognized as a Volterra equation of the second kind [Panjer, 2006; Wilmott *et al.*, 1985]. In general the Volterra integral equation of the second kind takes the form

$$f(x) = g(x) + \int_0^x K(x, x_1, f(x_1)) dx_1.$$

In the case of the Panjer recursion we have a linear Volterra equation in which

$$K(x, x_1, f(x_1)) = k(x, x_1) f(x_1).$$

This gives:

$$f(x) = g(x) + \int_0^x k(x, x_1) f(x_1) dx_1 \tag{10.3.2}$$

allowing us to make explicit the association between the Volterra equation of the second kind and the Panjer recursion. To do this we make the following identifications,

$$\begin{aligned} x_1 &= x - y \\ g(x) &= p_1 f_X(x) \\ k(x, x_1) &= (a + b \frac{x-x_1}{x}) f_X(x - x_1) \\ f(x_1) &= f_Y(x_1). \end{aligned}$$

Working with the Volterra integral equation of the second kind we can obtain the following representation,

$$\begin{aligned} f(x) &= g(x) + \int_0^x K(x, x_1) f(x_1) dx_1 \\ &= g(x) + \int_0^x K(x, x_1) [g(x_1) + \int_0^{x_1} K(x_1, x_2) f(x_2) dx_2] dx_1 \end{aligned}$$

and we recognize that this equation can be represented also as,

$$f(x) = g(x) + \int_0^x r(x, x_1) g(x_1) dx_1$$

in which r is the resolvent kernel for the Volterra equation of the second kind which, under the condition given below, may be expressed as the following von Neumann series expansion – see, for example, [Baker, 2000]:

$$r(x, x_1) = \sum_{n=1}^{\infty} k^n(x, x_1)$$

where $k^1(x, x_1) = k(x, x_1)$ and $k^n(x, x_1) = \int_0^x k(x, u) k^{n-1}(u, x_1) du$ with $n = 2, 3, 4, \dots$

Applying this series expansion to (10.3.2) gives,

$$f(x_0) = g(x_0) + \sum_{n=1}^{\infty} \int_0^{x_0} \dots \int_0^{x_{n-1}} g(x_n) \prod_{l=1}^n k(x_{l-1}, x_l) dx_{1:n},$$

where we use the notation $x_{1:n} = (x_1, \dots, x_n)$.

In order to simplify expressions throughout the paper, it is useful to define the following notation to describe the domain of integration. The conditional one-dimensional domains of integration are defined by $D_k(x_{k-1}) = [0, x_{k-1}]$, and we define the domain of integration of the n^{th} term in the summation as: $D_{1:n}(x_0) = \{(x_1, \dots, x_n) : x_0 > x_1 > \dots > x_n\}$, adopting the convention that $D_{1:0}(x_0) = \{\emptyset\}$.

Doing so allows us to write the previous expression in the form:

$$f(x_0) = g(x_0) + \sum_{n=1}^{\infty} \int_{D_{1:n}(x_0)} \prod_{l=1}^n k(x_{l-1}, x_l) g(x_n) dx_{1:n},$$

with this representation valid whenever the right hand side is finite.

We also define

$$\hat{D}_{0:n}(D_0) = \{(x_0, x_1, \dots, x_n) : D_0 \ni x_0 > x_1 > \dots > x_n\},$$

where D_0 corresponds to a region of values over which we wish to characterize annual loss distribution (typically an interval $[x_a, x_b]$), for later use. In the next section we will make it clear why this representation allows us to develop a novel simulation technique for evaluation of the distribution of a compound process, in our case an annual loss distribution in an LDA model.

To conclude this section, we mention previous approaches for evaluation of Volterra equations of the second kind and related fixed domain problems, the Fredholm integral equation of the second kind. In general this is a large and diverse literature spanning many different disciplines. For well presented primers see [Panjer, 2006; Baker, 2000; Baker, 1977; Doucet *et al.*, 2007; Linz, 1987; Orsi, 1996; Stroter, 1984]. The most commonly used approaches include quadrature methods for solving the integrals in the Panjer recursion, Runge-Kutta methods, Collocation and Galerkin methods which are based on polynomial splines or piecewise-polynomial densely defined approximations and also importance sampling techniques.

Our approach can most easily be associated with the importance sampling approach. In particular we utilize some concepts from [Doucet *et al.*, 2007] to interpret the standard von Neumann expansion of the Panjer recursion as an expectation with respect to a probability distribution defined on a union of subspaces of variable dimension. We then utilize both importance sampling and trans-dimensional Markov Chain Monte Carlo algorithms to simulate from the density with which the expectation is defined. Under this framework we develop two novel algorithms for simulation of an annual loss distribution, we present consistent and unbiased estimators for both point-wise and interval estimates of the annual loss distribution when evaluating VaR and also calculation of ES.

10.4 Importance Sampling using TD-MCMC

In this section we shall build on concepts and algorithms from [Peters *et al.*, 2006] by extending the concepts to trans-dimensional Markov Chain Monte Carlo sampling problems to efficiently evaluate the annual loss distribution when it is framed as an expectation evaluation under the Volterra integral equation representation presented in Section 10.3. If one does not have a closed-form expression for the density of the severity distribution, one can use techniques from [Peters *et al.*, 2006] in the framework we present in this paper.

In order to understand how the importance sampling techniques, which we will introduce subsequently, can be used to evaluate the Panjer recursion, we must first demonstrate how to formulate the problem in the form of an expectation. This follows in essentially the same manner as it does for Fredholm equations of the second kind [Doucet *et al.*, 2007]. We begin with our representation of the Volterra integral equation,

$$f(x_0) = g(x_0) + \sum_{n=1}^{\infty} \int_0^{x_0} \dots \int_0^{x_{n-1}} g(x_n) \prod_{l=1}^n k(x_{l-1}, x_l) dx_{1:n} \quad (10.4.1)$$

and then introducing the notation,

$$f_0(x_0) = g(x_0)$$

and setting

$$f_n(x_{0:n}) = g(x_n) \prod_{l=1}^n k(x_{l-1}, x_l).$$

This allows us to rewrite (10.4.1) as

$$f(x_0) = f_0(x_0) + \sum_{n=1}^{\infty} \int_0^{x_0} \dots \int_0^{x_{n-1}} f_n(x_{0:n}) dx_{1:n}.$$

Now we can frame the quantity of interest as an expectation with respect to some importance sampling distribution, p :

$$f(x) = \frac{f_0(x)}{p(0)} p(0) + \sum_{n=1}^{\infty} \int_{D_{1:n}(x)} \frac{f_n(x, x_{1:n})}{p(n, x_{1:n})} p(n, x_{1:n}) dx_{1:n} = E \left[\frac{f_n(x, X_{1:n})}{p(n, X_{1:n})} \right]. \quad (10.4.2)$$

There are two estimation problems which we are interested in: estimation of $f(x)$ point-wise, and characterizing $f(x)$ over some interval by obtaining samples from its restriction to that interval. We will present algorithms for solving both problems within the framework which we propose, but for definiteness will concentrate on the first in this paper, indicating any substantive changes which must be made in order to consider the other case.

The above representation allows us to perform importance sampling on the space

$\bigcup_{n=0}^{\infty} \{n\} \times D_{1:n}(x)$ in order to estimate $f(x)$ point-wise or on the slightly larger space

$\bigcup_{n=0}^{\infty} \{n\} \times \hat{D}_{0:n}([x_a, x_b])$ to characterize the function over some interval $[x_a, x_b]$.

Although, in the interests of simplicity and clarity of presentation, we will perform importance sampling directly upon the space described above. We note that, when we are interested in es-

timating the function point-wise, as $f_0(x)$ is known it would be more efficient in the sense that variance would be reduced on both a per sample basis and a per unit of computation basis to instead estimate $f(x) - f_0(x)$ by importance sampling on the smaller space $\bigcup_{n=1}^{\infty} \{n\} \times D_{1:n}(x)$ and this approach introduces no further complications.

We have not yet specified the importance sampling distribution, $p(n, x_{1:n})$. We will suggest two choices for this distribution: a simple candidate from which it is easy to obtain samples, and then an optimal selection which will minimize the variance of the importance weights. We will focus on the problem of estimating the density $f(x)$, our annual loss distribution, point-wise. This is challenging because it involves an infinite sum of integrals of increasing dimension. Monte Carlo techniques have been developed to solve complex problems of this sort.

10.4.1 Simple Importance Sampling Solution

Our first proposal distribution arises fairly naturally in the present context and is simple to simulate from, it was originally suggested in the context of solving Fredholm equations [Doucet *et al.*, 2007]. The solution in this setting would involve starting with a Markov Chain from x (or with some initial distribution μ which covers the region of interest if we wish to characterize f over some interval rather than at a point) and a transition kernel for the Markov Chain denoted $M(x, y)$ which will denote the probability density for going from state x to state y . The initial distribution, μ , when it is used, and transition kernel, M , are selected such that $\mu(x) > 0$ over the region of interest and $M(x, y) > 0$ if $k(x, y) \neq 0$, which is important to ensure the importance sampling scheme to be presented is well defined over the domain of interest, avoiding bias in estimates. Additionally the space explored by M is designed to have an absorbing cemetery state we denote by d , where $d \notin [0, \infty)$ and $M(x, d) = P_d$ for any x . The importance sampling approximation of the annual loss density $f_Y(x)$ is given in Algorithm 1. Note, the notation used here is $X_{0:n^{(i)}+1}^{(i)}$ to represent the i^{th} importance sample from $p(n, x_{1:n})$ of Markov Chain length n (additionally, $X_0^{(i)} = x_0$ when we are performing point-wise estimation, and $X_{n^{(i)}+1}^{(i)} = d$).

Algorithm 1- Importance Sampling for Panjer Recursions:

1. Simulate N independent Markov Chain paths $\left\{ X_{0:n^{(i)}+1}^{(i)} \right\}_{i=1:N}$ until absorption, where $X_{n^{(i)}+1}^{(i)} = d$.
2. Calculate Importance Sampling Weights.
If evaluation of the annual loss density at a point is desired, at the value x_0 , then this weight is given by:

$$W \left(X_{0:n^{(i)}}^{(i)} \right) = \begin{cases} \left(\prod_{s=1}^{n^{(i)}} \frac{k(X_{s-1}^{(i)}, X_s^{(i)})}{M(X_{s-1}^{(i)}, X_s^{(i)})} \right) \frac{g(X_{n^{(i)}}^{(i)})}{P_d} & n^{(i)} \geq 1 \\ \frac{g(X_0^{(i)})}{P_d} & n^{(i)} = 0. \end{cases}$$

Whilst, if X_0 is being sampled from some distribution μ in order to characterize f over some interval, then the importance weight function becomes:

$$W\left(X_{0:n^{(i)}}^{(i)}\right) = \begin{cases} \frac{1}{\mu\left(X_0^{(i)}\right)} \left(\prod_{s=1}^{n^{(i)}} \frac{k\left(X_{s-1}^{(i)}, X_s^{(i)}\right)}{M\left(X_{s-1}^{(i)}, X_s^{(i)}\right)} \right) \frac{g\left(X_{n^{(i)}}^{(i)}\right)}{P_d} & n^{(i)} \geq 1 \\ \frac{1}{\mu\left(X_0^{(i)}\right)} \frac{g\left(X_0^{(i)}\right)}{P_d} & n^{(i)} = 0. \end{cases}$$

3. If one is interested only in evaluating the annual loss distribution pointwise at x_0 , then we have the estimate,

$$\hat{f}(x_0) = \frac{1}{N} \sum_{i=1}^N W\left(x_0, X_{1:n^{(i)}}^{(i)}\right). \quad (10.4.3)$$

Otherwise, if approximating the annual loss distribution over some interval, such as when one is interested in calculation of ES, use the empirical estimate given by

$$\hat{f}(x_0) = \frac{1}{N} \sum_{i=1}^N W\left(X_{0:n^{(i)}}^{(i)}\right) \delta_{X_0^{(i)}}(x_0), \quad (10.4.4)$$

where $\delta_{X_0^{(i)}}$ is the Dirac-delta mass located at $X_0^{(i)}$.

When one calculates an expectation of any test function using the estimators constructed in this way and given by (10.4.3) and (10.4.4) these provide unbiased Monte Carlo estimates, either at a point $x_0 = x$ or over an interval D_0 . (7) can be seen to be an importance sampling procedure, (10.4.3) exists on the space $\bigcup_{n=0}^{\infty} \{n\} \times D_{1:n}(x_0)$ and (10.4.4) on the space $\bigcup_{n=0}^{\infty} \{n\} \times \hat{D}_{0:n}(D_0)$. In the first case the importance sampling distribution takes the form $p(n, x_{1:n}) = p(n) p_n(x_{1:n})$ with

$$p(n) = \Pr(X_{1:n} \in D_{1:n}(x_0), X_{n+1} = \{d\}) = (1 - P_d)^n P_d$$

and

$$p_n(x_{1:n}) = \frac{M(x, x_1) \prod_{k=2}^n M(x_{k-1}, x_k)}{(1 - P_d)^n}, \quad (10.4.5)$$

and the changes required to obtain the distribution used in the second case are obvious.

Hence, we have associated the original Panjer recursion with an expectation and then formulated an algorithm to provide an unbiased and consistent approximation of this annual loss distribution given by either (10.4.3) or (10.4.4).

Although we have successfully demonstrated one mechanism for obtaining the expectation of interest via importance sampling, it is known that whenever importance sampling is used, it is important to employ a good proposal distribution which leads to an estimator of low variance. If the Monte Carlo variance of the importance weights is large, then it will not be an efficient means of estimating the integrals comprising the expectation. As argued in [Doucet *et al.*, 2007]

this can be difficult to enforce when using importance sampling on the path space, commonly known as sequential importance sampling [Doucet *et al.*, 2001]. We will consider a principled approach to choosing an importance function, and to obtaining samples with this distribution.

10.4.2 Optimal Importance Sampling using Trans-dimensional Markov Chain Monte Carlo

First we observe that the estimate of (10.4.3) will be unbiased for any importance sampling distribution satisfying, for all $n \geq 1$,

$$\int_{D_{1:n}(x)} f_n(x, x_{1:n}) dx_{1:n} > 0 \Rightarrow p(n, x_{1:n}) > 0 \quad (10.4.6)$$

and $p(0) > 0$, with $f_n(x, x_{1:n})$ absolutely continuous with respect to $p(x_{1:n})$. That of (10.4.4) is unbiased under similar weak conditions. In order to obtain finite variance it is sufficient and recommended for the ratio f_n / p to remain bounded throughout the domain of integration [Robert *et al.*, 2004]. As long as we satisfy these standard conditions, we are free to construct importance distributions of any form. There are many ways to go about doing this; one could consider distributions which are relatively simple to draw samples from, making the techniques fast in the sense that the computational cost of each sample is relatively low. However, this approach comes at the cost of increased, and possibly unbounded, variance in estimates formed for our annual loss distribution.

The criteria we consider for selecting an importance distribution is one which is widely accepted in the importance sampling literature: minimizing the variance of the importance weights. In this regard we utilize a result from [Doucet *et al.*, 2007] which provides the importance sampling distribution which minimizes the variance of the importance weights, for our problem, under mild conditions which are explained in the original derivation. This importance distribution is given by (10.4.7), which we propose to use in place of (10.4.5).

$$p_{opt}(n, x_{1:n}) = p_{opt}(n) p_{n,opt}(x_{1:n}) \quad (10.4.7)$$

with,

$$p_{n,opt}(x_{1:n}) = c_{n,opt}^{-1} f_n(x, x_{1:n})$$

$$c_{0,opt} = f_0(x) \text{ and } c_{n,opt} = \int_{D_n(x)} f_n(x, x_{1:n}) dx_{1:n} \text{ for } n \geq 1,$$

and

$$p_{opt}(n) = c_{n,opt} \left(\sum_{k=0}^{\infty} c_{k,opt} \right)^{-1} \text{ for } n \geq 0.$$

We know $c_{opt} = \sum_{k=0}^{\infty} c_{k,opt} < \infty$ as this follows directly from the assumption required for the existence of a von Neumann series expansion.

We can see that this optimal importance sampling distribution takes support on the same space as our target distribution, a disjoint union of subspaces and if we were able to sample random

variables according to this distribution cheaply, then it would clearly be the best choice of proposal distribution available to us. However, as is often the case, it is not possible to sample directly from this distribution in a computationally efficient manner. Therefore we propose to employ Markov Chain Monte Carlo (MCMC) techniques. The motivation behind this is that the variance reduction obtained by employing this importance distribution will offset the increased computational cost of an individual sample and any additional Monte Carlo error due to correlation within finite sequence of the Markov Chain samples.

Briefly, Markov Chain Monte Carlo techniques involve construction of a ergodic Markov Chain $\{X_1, X_2, \dots, X_N\}$ which has the property that it has a limiting, invariant distribution corresponding to the target distribution one is aiming to produce samples from. That is, we obtain a sequence of statistically dependent samples which have the property that empirical average of any regular function evaluated at the sampled values of X converges, as the sample size increases, to the expectation of that function under the target distribution. Some references that present this material lucidly include [Meyn *et al.*, 1993; Gelman *et al.*, 1995; Gilks *et al.*, 1996; Robert *et al.*, 2004].

In particular we focus on a trans-dimensional methodology that creates a reversible Markov Chain with ergodic distribution given by the optimal importance distribution. The most well known methodology in this space is the Reversible Jump Markov Chain Monte Carlo RJMCMC sampler of [Green, 1995; Green, 2003]. Other approaches include product space formulations or general birth and death processes, for details see [Brooks, 2003; Sisson, 2005]. We utilize the Birth and Death version of RJMCMC. We will not present all the background behind this approach, for detailed references see [Green, 1995; Green, 2003; Brooks, 2003; Sisson, 2005] and for convergence properties see [Gilks *et al.*, 1996; Meyn *et al.*, 1993]. We present next a simple algorithm that can be implemented directly and easily for a broad range of models.

The algorithm we use consists of “within” subspace updates of the Markov Chain utilizing concepts from [Peters *et al.*, 2006] and “trans-dimensional” birth and death moves for traveling between different subspaces of the support of the target distribution. Each algorithmic iteration utilizes a deterministic scan for within and between subspace moves and Algorithm 2 contains the details. Note the probability of birth and death moves will be denoted p_b and p_d , respectively. We set $p_d = 1 - p_b$, and we set $p_b = 1$ whenever $n = 0$.

The following algorithm describes the procedure for obtaining a collection of samples from which to make a point estimate of f at some pre-specified x_0 . In order to obtain an interval estimate instead, it is simply necessary to slightly change the target distribution to take this into account and to allow J to take a value of 0 in step 2 of the update move and step 1 of the birth and death moves.

Algorithm 2: Reversible Jump Markov Chain Monte Carlo for Importance Distribution Initialization:

1. For $i = 1$ set $\left(n^{(1)}, X_{0:n^{(1)}}^{(1)}\right)$ deterministically as $n^{(1)} = 1$ and $X_0^{(1)} = x$, $X_1^{(1)} = x/2$,
Repeat for $i \geq 1$

2. Update Move:

(a) Set $n^{(i)} = n^{(i-1)}$.

(b) Sample uniformly index $J \sim U\{1, \dots, n^{(i)}\}$.

(c) Sample proposed update for J^{th} element, $X_J^* \sim q_u\left(X_J^{(i-1)}, \cdot\right)$.

(d) Evaluate the acceptance probability (note that $X_{0:n^{(i)} \setminus J}^{(i-1)}$ should be interpreted in the natural manner as $(X_{0:J-1}^{(i-1)}, X_{J+1:n^{(i)}}^{(i-1)})$).

$$\begin{aligned} & \alpha\left(\left(n^{(i)}, X_{0:n^{(i)}}^{(i-1)}\right), \left(n^{(i)}, X_{0:n^{(i)} \setminus J}^{(i-1)}, X_J^*\right)\right) \\ &= \min\left\{1, \frac{p_{\text{opt}}\left(n^{(i)}, \left(X_{0:n^{(i)} \setminus J}^{(i-1)}, X_J^*\right)\right) q_u\left(X_J^*, X_J^{(i-1)}\right)}{p_{\text{opt}}\left(n^{(i)}, X_{0:n^{(i)}}^{(i-1)}\right) q_u\left(X_J^{(i-1)}, X_J^*\right)}\right\} \end{aligned}$$

(e) Sample uniform random variate $U \sim U[0, 1]$.

(f) If $U \leq \alpha\left(\left(n^{(i)}, X_{0:n^{(i)}}^{(i-1)}\right), \left(n^{(i)}, X_{0:n^{(i)} \setminus J}^{(i-1)}, X_J^*\right)\right)$ then set

$$\left(n^{(i)}, X_{0:n^{(i)}}^{(i)}\right) = \left(n^{(i)}, X_{0:n^{(i)} \setminus J}^{(i-1)}, X_J^*\right)$$

otherwise set

$$\left(n^{(i)}, X_{0:n^{(i)}}^{(i)}\right) = \left(n^{(i)}, X_{0:n^{(i)}}^{(i-1)}\right).$$

3. Sample uniform random variate $U \sim U[0, 1]$.

If $U < p_b$

4. Birth Move:

(a) Sample an index uniformly to add a new component at $J \sim U\{1, \dots, n^{(i-1)} + 1\}$

(b) Sample new component's value $X_J^* \sim q_b(\cdot)$.

(c) Evaluate the acceptance probability

$$\begin{aligned} & \alpha\left(\left(n^{(i-1)}, X_{0:n^{(i)}}^{(i-1)}\right), \left(n^{(i-1)} + 1, X_{0:n^{(i-1)}}^{(i-1)}, X_J^*\right)\right) \\ &= \min\left\{1, \frac{p_{\text{opt}}\left(n^{(i-1)} + 1, \left(X_{0:J-1}^{(i-1)}, X_J^*, X_{J+1:n^{(i-1)}}^{(i-1)}\right)\right) p_d}{p_{\text{opt}}\left(n^{(i-1)}, X_{0:n^{(i-1)}}^{(i-1)}\right) q_b\left(X_J^*\right) p_b}\right\} \end{aligned}$$

(d) If $U \leq \alpha \left(\left(n^{(i-1)}, X_{0:n^{(i)}}^{(i-1)} \right), \left(n^{(i-1)} + 1, X_{0:n^{(i-1)}}^{(i-1)}, X_J^* \right) \right)$ then set

$$\left(n^{(i)}, X_{0:n^{(i)}}^{(i)} \right) = \left(n^{(i-1)} + 1, \left(X_{0:J-1}^{(i-1)}, X_J^*, X_{J+1:n^{(i-1)}}^{(i-1)} \right) \right)$$

otherwise set

$$\left(n^{(i)}, X_{0:n^{(i)}}^{(i)} \right) = \left(n^{(i-1)}, X_{0:n^{(i-1)}}^{(i-1)} \right).$$

else

5. Death Move:

(a) Sample an index uniformly to delete an existing component $J \sim U \{1, \dots, n^{(i-1)}\}$

(b) Evaluate the acceptance probability

$$\begin{aligned} & \alpha \left(\left(n^{(i-1)}, X_{0:n^{(i)}}^{(i-1)} \right), \left(n^{(i-1)} - 1, X_{0:n^{(i-1)} \setminus J}^{(i-1)} \right) \right) \\ &= \min \left\{ 1, \frac{p_{opt} \left(n^{(i-1)} - 1, \left(X_{0:J-1}^{(i-1)}, X_{J+1:n^{(i-1)}}^{(i-1)} \right) \right) q_b \left(X_J^{(i-1)} \right) p_b}{p_{opt} \left(n^{(i-1)}, X_{0:n^{(i-1)}}^{(i-1)} \right) p_d} \right\} \end{aligned}$$

(c) If $U \leq \alpha \left(\left(n^{(i-1)}, X_{0:n^{(i)}}^{(i-1)} \right), \left(n^{(i-1)} - 1, X_{0:n^{(i-1)} \setminus J}^{(i-1)} \right) \right)$ then set

$$\left(n^{(i)}, X_{0:n^{(i)}}^{(i)} \right) = \left(n^{(i-1)} - 1, \left(X_{0:J-1}^{(i-1)}, X_{J+1:n^{(i-1)}}^{(i-1)} \right) \right)$$

otherwise set

$$\left(n^{(i)}, X_{0:n^{(i)}}^{(i)} \right) = \left(n^{(i-1)}, X_{0:n^{(i-1)}}^{(i-1)} \right).$$

6. If $i < M$ go to 2.

The only two quantities we need to specify to apply Algorithm 2 to approximate a particular annual loss distribution, is the Markov transition kernel that will be used for the within subspace moves $q_u \left(X_J^{(i-1)}, \cdot \right)$ and also the distribution for the birth proposal q_b . Here $q_u \left(X_J^{(i-1)}, x \right)$ denotes the probability of updating the J^{th} element of the current state of the Markov Chain, $X_{1:n^{(i-1)}}^{(i-1)}$, from $X_J^{(i-1)}$ to some value $x \in D_J \left(X_{J-1}^{(i-1)} \right)$. Additionally, the notation q_b denotes the probability density from which the new proposed birth component will be sampled, when proposing to move from subspace $D_{1:n} \left(x_0 \right)$ to $D_{1:n+1} \left(x_0 \right)$.

We will utilize a symmetric Gaussian kernel for the within subspace moves leading to a Random Walk Metropolis (RWM) algorithm. For the birth move we now present the proposal distribution which is optimal in the sense that it minimizes the variance of the ratio within the acceptance probability of a birth and death move. This will be achieved if we recognize that we can always make the following factorization,

$$p_{opt} \left(x_{1:n+1} \right) = p_{opt} \left(x_{n+1} | x_{1:n} \right) p_{opt} \left(x_{1:n} \right).$$

Then we can easily see that the proposal for the birth move minimizing the variance of the ratio appearing in the acceptance probability is given by $p_{opt}(x_{n+1}|x_{1:n})$. In our setting this can be shown to take the form,

$$p_{opt}(x_{n+1}|x_{1:n}) = \frac{f_X(x_{n+1})}{f_X(x_n)} k(x_n, x_{n+1}) \propto f_X(x_{n+1}) \left(a + b \frac{x_n - x_{n+1}}{x_n} \right) f_X(x_n - x_{n+1}), \quad (10.4.8)$$

for $x_{n+1} \in [0, x_n]$. Hence, we propose to sample from an approximation to this distribution using a simple empirical cdf estimate. We construct a fast, crude estimate over a uniform grid using a right end-point rule, using 20 points to construct the piecewise estimate. Alternatively, one could use rejection sampling since $p_{opt}(x_{n+1}|x_{1:n})$ is supported on $[0, x_n]$ and typically the severity distribution will have an analytic bounding distribution.

The approximation of the annual loss distribution, once we have drawn samples $\{n^{(i)}, X_{1:n^{(i)}}^{(i)}\}_{i=1:M}$ approximately distributed as the optimal importance sampling distribution, is given by [Doucet *et al.*, 2007],

$$\hat{f}(x) = f_0(x) + \hat{c}_{opt} \quad . \quad (10.4.9)$$

To perform this calculation one can approximate the optimal normalizing constant as,

$$\hat{c}_{opt} = \frac{\hat{c}_{1,opt}}{\hat{p}_{1,opt}}, \quad (10.4.10)$$

providing that it is possible to estimate $c_{1,opt}$. This can be done in advance and we utilize a right end point trapezoidal rule, but any numerical integration scheme could be used. We obtain the estimate of $p_{1,opt}$ as the proportion of the total number of states explored by the Markov chain which lie in $D_{1:1}(x_0)$.

Summarizing this section, we have developed machinery to allow us to obtain accurate estimates of an annual loss distribution. This was achieved by importance sampling in the case of Algorithm 1 and the methodology to sample from a more principled importance distribution in Algorithm 2. We will next demonstrate the performance of our approach and compare it to standard Monte Carlo simulation in two different models.

10.5 Simulation Results and Analysis

In this section we compare the basic Monte Carlo simulation approach, described in the introduction, to the importance sampling procedure in Algorithm 1 and the trans-dimensional Markov Chain Monte Carlo scheme of Algorithm 2. We will also develop a truncated Gaussian Power Approximation mixture to estimate the annual loss distribution. The quantiles obtained from our mixture will be compared with simulated results from basic Monte Carlo and importance sampling as a further validation of simulation results in the tail of our annual loss distribution. Example 1 considers a simple yet frequently used model in financial institutions, the Poisson-Lognormal compound process. In example 2 we present comparisons for a more

sophisticated model, the Poisson-GB2 compound process, building on models proposed in [Peters *et al.* 2006].

The truncated mixture of Gaussian distributions for example 1 is obtained as follows. First we recognize that the annual loss cumulative distribution is given by (10.5.1),

$$F_Y(x) = \sum_{m=0}^{\infty} h(m) F_X^{*m}(x) \quad (10.5.1)$$

where $F_Y(x)$ denotes the cumulative distribution of our annual loss and $F_X^{*n}(x)$ denotes the m -fold convolution of the selected severity cumulative distribution,

$$F_X^{*m}(x) = \Pr\left(\sum_{i=1}^m X_i < x\right)$$

and $h(n)$ is the selected frequency distribution. We then truncate this sum using a large m , in our case we will use $m_{max}=100$, to obtain the following mixture approximation,

$$F_Y(x) \approx \sum_{m=0}^{m_{max}} \tilde{h}(m) F_X^{*m}(x),$$

where we use $\tilde{h}(m)$ to represent the normalized mixture weights. Finally, instead of calculating the m -fold convolutions exactly, we apply a Gram-Charlier or Edgeworth expansion [Rotar, 2007] for each standardized mixture component,

$$\bar{F}_X^m(x_{C_m}) = F_X^{*m}\left(\frac{x - E(Y_m)}{\sigma(Y_m)}\right),$$

given by

$$\bar{F}_X^m(x_{C_m}) = \Phi(x) + \frac{\gamma_m}{6\sqrt{m}}(1-x^2)\varphi(x) + O\left(\frac{1}{m}\right)$$

and disregard all terms but the first two, thereby correcting for the asymmetry in our true annual loss distribution when approximating with a symmetric Gaussian distribution. The notation x_{C_m} is used to denote the standardized value for the m^{th} mixture component and the terms $\Phi(x)$ and $\varphi(x)$ represent the standard Gaussian cumulative distribution and density, respectively. The term γ_m is the skewness coefficient of our m^{th} mixture component $F_X^{*m}(x)$ and since the X_i 's are i.i.d with distribution $f_X(x)$,

$$\gamma_m = \left[\frac{mE\left[(X - E[X])^3\right]}{(mVar(X))^{\frac{3}{2}}} \right].$$

Hence, we obtain the standardized Gaussian mixture approximation:

$$\bar{F}_Y(x_C) \approx \sum_{m=0}^{m_{max}} \tilde{h}(m) \left[\Phi(x_{C_m}) + \frac{\gamma_m}{6\sqrt{m}}(1-x_{C_m}^2)\varphi(x_{C_m}) \right]. \quad (10.5.2)$$

This provides us with a diagnostic tool to check we are obtaining sensible quantile estimates

in the tails of our annual loss distribution, but it can't be used to easily generate our quantile estimates: the inversion of this mixture distribution can only be performed numerically and at significant computational cost.

In practice there is a set of preliminary steps which must be carried out prior to simulation from the annual loss distribution. These involve estimation of the parameters for the frequency and severity distributions, using a combination of data sources from internal loss data, external loss data and expert elicitation. For non-standard techniques to perform fitting and parameter estimation see [Peters *et al.*, 2006]. However, typically this is achieved via Maximum Likelihood [Cruz, 2002]. This paper now proceeds to the simulation of the annual loss distribution assuming parameter estimates are available.

We now wish to construct point-wise estimates of the annual loss distribution which we showed could be expressed as (10.4.2). To do so when forming estimates for the standard importance sampler in Algorithm 1, we will use the estimator given in (10.4.3). For the Optimal Importance Sampler in Algorithm 2, we will use the estimator given by (10.4.9).

The standard deviation can be calculated for each point estimate of the annual loss distribution. For Algorithm 1, this is given by evaluation of (10.5.3) at each grid point x :

$$\text{Var} \left(\hat{f}_Y(x) \right) = \text{Var} \left(\frac{1}{N} \sum_{i=1}^N W \left(x, X_{1:n}^{(i)} \right) \right). \quad (10.5.3)$$

For Algorithm 2, we calculate the standard deviation of the estimate of the annual loss distribution at each point of evaluation, x , using (10.5.4):

$$\text{Var} \left(\hat{f}(x) \right) = \text{Var} \left(\hat{c}_{opt} \right) = \text{Var} \left(\frac{c_{opt,1}}{\hat{p}_{opt,1}} \right) = \text{Var} \left(\frac{c_{opt,1}N}{\sum_{i=1}^N \delta_1(n^{(i)})} \right). \quad (10.5.4)$$

The method used to evaluate the expressions in (10.5.3) and (10.5.4) is to take the sums and randomly split them into S sub-blocks of length $\lceil N/S \rceil$, then calculate (10.5.3) or (10.5.4) for each of the S sub-blocks and take the variance over the values obtained from each sub-block.

Settings for Algorithm 1:

In the proceeding simulation studies we use $P_d=0.1$ for example 1 and $P_d=0.05$ for example 2. In general, this cannot be set too large as sufficient exploration of the support of the importance sampling distribution will not be achieved, leading to high variance estimates of the annual loss distribution. However, at the other extreme setting, P_d too small leads to excessive computational burden. We do not claim any optimality in selection of this probability: in general it will be problem specific. For Algorithm 1 we have

$$p(n, x_{1:n}) = p(n) p_n(x_{1:n})$$

with

$$p(n) = (1 - P_d)^n P_d$$

$$\begin{aligned}
p_n(x_{1:n}) &= \frac{M(x, x_1) \prod_{l=2}^n M(x_{l-1}, x_l)}{(1-P_d)^n} \\
M(x_{l-1}, x_l) &= (1-P_d) M'(x_{l-1}, x_l) + P_d \delta_d(x) \\
M'(x_{l-1}, x_l) &\propto N(x_l; x_{l-1}, \sigma) I[0 < x_l < x_{l-1}].
\end{aligned}$$

Here we denote the Normal distribution with mean m and standard deviation s by $N(x; m, s)$. In this setting it is straightforward to sample the importance distribution and to evaluate the importance weights. This simplicity comes at the cost of increased variance of estimates of the annual loss distribution at each point of evaluation for a given number of samples.

Settings for Algorithm 2:

For Algorithm 2, the approximately optimal trans-dimensional MCMC importance sampler will have,

$$p_{opt}(n, x_{1:n}) = p_{opt}(n) p_{n,opt}(x_{1:n})$$

with

$$p_{n,opt}(x_{1:n}) = c_{n,opt}^{-1} f_n(x, x_{1:n}) \propto f_X(x_n) \prod_{l=1}^n \left(a + b \frac{x_{l-1} - x_l}{x_{l-1}} \right) f_X(x_{l-1} - x_l)$$

and

$$p_{opt}(n) = c_{n,opt} \left(\sum_{n=0}^{\infty} c_{n,opt} \right)^{-1}.$$

In the algorithm we present, we consider the RWM algorithm for within model moves, this specifies

$$q_u \left(X_J^{(i-1)}, X_J^{*(i-1)} \right) \propto N \left(X_J^{*(i-1)}; X_J^{(i-1)}, \sigma_{RW} \right) I \left[\max(X_{J-1}^{(i-1)}, 0) < X_J^{*(i-1)} < \min(X_{J+1}^{(i-1)}, x) \right].$$

For the birth move we will use an empirical estimate of $q_b(x_{n+1}) = p_{opt}(x_{n+1} | x_{1:n})$ which approximates (10.4.8).

Note, for the sake of being concise, we will only present the standard Monte Carlo results for either $N=50m$ or $N=10m$ simulated annual years which we will consider to be the exact solution. For the importance sampling (IS) approaches we must consider N , the number of importance samples or the length of the Markov Chain after burn-in. Unless stated otherwise we use $N=50k$, the length of the burn-in for the RJMCMC $N_{burn-in}=50k$, the grid spacing for evaluation of the annual loss distribution will be equally spaced unit intervals and the domain over which we will evaluate the annual loss distribution has default $[0,100]$. In general a non-linearly spaced grid can easily be used and this would be expected to further speed up computations. This is a significant point, since if one is interested in evaluation of the entire annual loss distribution using a non-linearly spaced grid can provide significant computational gains compared to techniques such as inversion techniques, like FFT methods which require a uniformly spaced grid. When we calculate expected shortfall we will use domain $[x_s = q_{1-\alpha}, q_{1-\alpha} + 100]$ with evaluation on the integers, where $q_{1-\alpha}$ is the value of the $1-\alpha$ quantile.

10.5.1 Example 1: Poisson-Lognormal compound process.

In this example we utilize LDA model, with Poisson(λ) frequency distribution and lognormal(μ, σ) severity distribution and we make the following definitions and associations:

$$p_1 = \lambda e^{-\lambda}, a = 0, b = \lambda, f_X(x) = LN(x; \mu, \sigma)$$

$$p_{n,opt}(x_{1:n}) \propto LN(x_n; \mu, \sigma) \prod_{l=1}^n \left(\frac{x_{l-1} - x_l}{x_{l-1}} \right) LN(x_{l-1} - x_l; \mu, \sigma).$$

Next we demonstrate results for standard Monte Carlo, standard importance sampling and approximate optimal importance sampling for the following case

$$(\lambda = 2, \mu = 2, \sigma = 0.5).$$

Simulation Results.

In Figure 10.8.1, we compare the simulated histogram estimate of the annual loss distribution and the empirical cdf using standard Monte Carlo, standard importance sampling Algorithm 1 and the MCMC importance sampling procedure of Algorithm 2. For the importance samplers the calculation of the variance in the estimate used $S=50$.

In Figure 10.8.2, we present the standard deviation in our estimate, as a function of N_T . In the case of the standard importance sampler we will use $N_T = E[M] \times N = (P_d)^{-1} N$ and in the case of the trans-dimensional sampler we utilize $N_T = N + \text{burnin}$. Comparing the performance of each algorithm as a function of N_T allows for a fair comparison for a given number of samples. It is not sufficient to simply use the number of importance samples N , since in algorithm 1 each sample has implicitly run a chain with mean length $(P_d)^{-1}$, hence we need to take this into consideration. We compare Algorithm 1 and Algorithm 2 on a log-log axis. The standard deviation is calculated at the points $x=10,15$ and the number of sub-blocks $S=20$.

Figure 10.8.2 confirms that our optimal importance sampling scheme using trans-dimensional MCMC has lower variance than that of standard importance sampling. This is directly due to the fact that we sample approximately from the optimal importance distribution, minimizing the variance of the importance weights. We note that, although we are sampling from this target distribution once the Markov chain becomes stationary, these samples are correlated and only when the samples are truly i.i.d. can we be certain of achieving a minimum variance estimate. This type of analysis is useful as it helps determine the number of importance samples for a given accuracy.

We note that in general relying solely on birth and death steps to make the Markov Chain mix between the different dimensions requires long burn-in times. Increasing the complexity of design and computation of the trans-dimensional moves could reduce these burn-in times. Additionally, the estimate of the normalizing constant could be made more accurate by employing one of the approaches advocated by [Bartolucci *et al.*, 2006].

In Figure 10.8.3, we present a histogram of the $\{n^{(i)}\}_{i=1:N}$ at selected evaluation points of the

annual loss distribution using Algorithm 2, the trans-dimensional sampler. As expected, as the distance from the origin to x , the grid point at which we evaluate the annual loss distribution, increases, this leads to the most frequently visited subspace $D_{1:n}(x_0)$ also increasing in dimension. This is not surprising: we would expect to require more terms in the Panjer recursion to obtain accurate estimation of the density when far away from the origin.

The results we present in Table 10.1 are obtained on a linearly spaced grid of width 1. For comparison between the standard Monte Carlo and the importance sampling estimates, we histogram the standard Monte Carlo using unit length bins. The estimate of $\hat{\sigma}$, used to form intervals in Table 10.1, is calculated as the standard deviations of the importance weights at the point of evaluation and $\hat{\sigma}_T$ is the accumulated error over evaluation points.

We can see from Table 10.1 that as expected the optimal importance sampler performs better than standard importance sampling *for an equal number of samples* and both provide good estimates of tail quantiles and ES. Overall, demonstrating that these algorithms provide accurate means of estimating VaR and ES in a practical LDA model. These will be shown to be superior in terms of computational efficiency in the next section.

In Figure 10.8.4 we provide analysis of the mixing of the trans-dimensional sampler by considering the proportion of samples from the chain which were of length 1, which is used in the estimation of \hat{c}_{opt} . We present these results as a function of the length of the chain N , for several different points at which we evaluate the annual loss distribution, $x=10, 50, 200$. It is clear from these sample paths that as the distance from the origin, at which we choose to evaluate our annual loss distribution $f_Y(x)$ increases, the length of chain N required before this estimate stabilizes will increase.

We also note that the error in the approximation due to forming a piecewise linear approximation of the annual loss distribution is not accounted for in the calculation of $\hat{\sigma}$. However, we know from the Central Limit Theorem that providing that the variance is finite the value of $\hat{\sigma}$ will $\rightarrow 0$ as the number of importance samples $N \rightarrow \infty$ [Kipnis & Varadhan,1986], for any given grid spacing. This does not however ensure our estimate of the quantile is consistent, only that the actual estimates of the annual loss distribution are asymptotically correct at the locations for which we place our grid points. However, as we can control this granularity, we can make this as accurate as we desire. Standard results may be used to bound the additional error introduced by this numerical integration stage [Robert & Casella, 2004]. Additionally, we have not accounted for error associated with the fitting procedure to estimate the parameters of the severity and frequency distributions, this is future work.

Computational Considerations.

To compare the computation time, we consider two evaluations. The first is the time taken to obtain a single point estimate of the annual loss distribution, for roughly the same accuracy. The second is the computation time for expected shortfall. Note we have not made use of the fact that we can easily parallelize our computation, reducing the simulation time significantly

for Algorithm 1 and Algorithm 2. All simulations were implemented in Matlab and performed on an Intel Core 2 Duo processor (2.4GHz) with 4GB RAM.

Table 10.2 demonstrates that the computational cost of using standard importance sampling under Algorithm 1 is a significant improvement in computation time compared to basic Monte Carlo. For more complex evaluation scenarios simple Monte Carlo will be unable to provide solutions for a reasonable computational cost. Additionally, Algorithm 1 relies upon the simple importance distribution matching the target distribution reasonably well; if the function g is sharply peaked then this is unlikely to be the case and in such situations Algorithm 2 should remain an effective and efficient means of evaluating the quantities of interest at a fixed computational cost.

It is clear that evaluation of the annual loss distribution at a fixed point $x \gg$ mode of f_Y when using the fixed importance sampling distribution, Algorithm 1, will degrade in performance as the function g becomes more peaked and hence the distance between x and the peak of g increases. This can be understood since if the importance distribution remains unchanged, then as the annual loss distribution becomes more peaked as a consequence of g becoming more peaked, relative to the importance sampling distribution, then the performance of such an importance sampling distribution will degrade in accuracy and variance. In these scenarios the optimal importance sampling distribution will be important.

What is not so clear is how the performance of each algorithm compares if evaluating the annual loss distribution at different x values as g changes shape. We conclude by demonstrating in Figure 10.8.5 that the performance of MCMC importance sampling algorithm also significantly outperforms the standard importance sampler in these circumstances. To demonstrate this we present analysis of the impact of the shape of the severity distribution on the variance estimate obtained from each technique when performing evaluation of $f_Y(x_m)$, where

$$x_m = \max(1, \arg \max [g(x)]) = \max(1, \arg \max [p_1 f_X(x)]).$$

We only alter the parameters of the severity distribution and N for the standard importance sampling from Algorithm 1. Note, we do not change P_d since in practice, a good value for P_d will not be known *a priori*. We can then demonstrate that this is one of the advantages of trans-dimensional sampling since in Algorithm 2 $p_{opt}(n)$ is “discovered” by the sampler on-line.

Overall, the results we obtained from using Algorithms 1&2, demonstrate that they work well. We show that unless one knows *a priori* that the function g is highly peaked, we advocate the use of the standard importance sampler as it is significantly faster and able to perform well. Only in more complex scenarios, if the standard importance sampler is producing variance in estimates which are not tolerable, should one then consider the trans-dimensional MCMC approach as a variance reduction technique.

10.5.2 Example 2: Poisson-GB2 compound process.

Utilizing the analysis and subsequent advocated advice from example 1, we now build a more sophisticated model based on work presented in [Peters *et al.* 2006]. We develop a Poisson-GB2 compound process, Poisson(λ) frequency and GB2(a, b, p, q) severity distribution. We demonstrate that we again obtain accurate estimates of the annual loss distribution when compared to basic Monte Carlo if using Algorithm 1 as recommended from the analysis presented in Example 1 with significantly less computation time.

The flexibility of the GB2 distribution lies in the fact that it encompasses a family of distributions. Depending on the parameter values fitted using the loss data one can recover a range of parametric severity distributions with a flexible range of location, scale, skewness and kurtosis. We refer the reader to [Dutta *et al.* 2006, Peters *et al.* 2006] for discussion on the merits and properties of the GB2 distribution when used to model the severity distribution in an LDA framework.

The GB2 distribution has density function given by

$$f(x) = \frac{|a|x^{ap-1}}{b^{ap}B(p,q)[1+(x/b)^a]^{p+q}}I_{(0,\infty)}(x) \quad , \quad (10.5.5)$$

where $B(p,q)$ is the Beta function and the parameters a, p and q control the shape of the distribution and b is the scale parameter. For discussion of how the parameters are interpreted in terms of location, scale and shape see [Dutta *et al.* 2006].

It was demonstrated in [Bookstaber et al, 1987] that the GB2 distribution (10.5.5) encapsulates many different classes of distribution for certain limits on the parameter set. These include the lognormal ($a \rightarrow 0, q \rightarrow \infty$), log Cauchy ($a \rightarrow 0$) and Weibull/Gamma ($q \rightarrow \infty$) distributions, all of which are important severity distributions used in operational risk models in practice.

Here we consider the parameter values for the GB2 family corresponding to the Lomax distribution (Pareto distribution of the 2nd kind or Johnson Type VI distribution) with parameters $[a, b, p, q]=[1, b, 1, q]$, which when reparameterized in terms of $q > 0, \lambda > 0$ and $x > 0$, is given by

$$h(x; \alpha, b) = \frac{q}{b} \left(1 + \frac{x}{b}\right)^{-(q+1)} . \quad (10.5.6)$$

The heavy-tailed Lomax distribution has shape parameter q and scale parameter b . Next we present three sets of analysis using our approach for a range of parameter values $[a, b, p, q, \lambda]$.

The first analysis in Figure 10.8.6 demonstrates the basic Monte Carlo versus the standard importance sampling procedure as a function of the number of importance samples $N(=100,500,1k,10k)$. Demonstrating that as the number of importance samples increases, the accuracy improves as expected, we also include simulation time for the entire distribution on $[0,100]$ in the figure.

The second analysis in Figure 10.8.7 demonstrates, the performance as a function of $\lambda(=0.01,1,10)$, *note each plot also contains simulation time for each approach.* In this example we use approximately

the smallest N found to provide acceptable accuracy. Clearly this analysis demonstrates two points, our approach is highly accurate for a range of mean frequencies and secondly the computational time savings using our approach are significant. In the case in which $\lambda=0.01$ the basic Monte Carlo would require more than $N>50\text{mil}$ to obtain the same accuracy obtained by our approach as x increases.

The third analysis in Figure 10.8.8 demonstrates the performance as a function of $\alpha(= q)$ ($=0.1,1,10$), the shape parameter of the severity distribution.

10.6 Discussion

In this article, we have introduced new methodology to the simulation of annual loss distributions under an LDA framework for Operational Risk. We have demonstrated their performance on some real practical problems and compared with standard industry practices. We advocate that practitioners adopt these techniques as they can significantly improve performance for a given computational time budget.

Furthermore, correlation between different Business unit/Risk types can be introduced under our approach in the usual manner used in financial institutions, via the use of Copula transforms in an LDA framework. In such settings, one could sample from each constructed annual loss distribution and apply a copula transform to these samples to obtain correlation at the annual loss level.

Future work will consider developing a richer class of trans-dimensional Markov Chain Monte Carlo moves to reduce the computational effort when using minimum variance estimates. One could also consider extending such approaches to allow for introduction of correlation between severity and frequency during the simulation process. It would also be interesting to compare the performance of the algorithms discussed here when applied to more challenging distributions.

10.7 Acknowledgements

The first author is supported jointly by an Australian Postgraduate Award, through the Department of Statistics at UNSW and a top-up research scholarship through CSIRO Quantitative Risk Management. Thank you goes to Pavel Shevchenko, Xiaolin Luo, Jerzy Sawa and Matthew Delasey for useful discussions.

10.8 Appendix

| Quantile $D_0=[1,100]$ | Truth Standard MC $N=50\text{mil.}$ | Gaussian Mixture Approx. | Standard Importance Sampling Estimate $[\hat{q}_{1-\alpha} - \hat{\sigma}, \hat{q}_{1-\alpha} + \hat{\sigma}]$ $N=50\text{k per grid point}$ $N_T=500\text{k}$ | Optimal Importance Sampling Estimate $[\hat{q}_{1-\alpha} - \hat{\sigma}, \hat{q}_{1-\alpha} + \hat{\sigma}]$ $N=50\text{k per grid point}$ $N_T=100\text{k}$ |
|--|--|--------------------------|---|--|
| | Histogram Bin Width = 1 | | Grid Width = 1 | Grid Width = 1 |
| 0.5 | 14 | $\hat{F}_Y(14) = 0$ | 15 [14,16] | 15 [14,16] |
| 0.8 | 27 | $\hat{F}_Y(27) = 0.5996$ | 25 [26,28] | 27 [25,29] |
| 0.9 | 35 | $\hat{F}_Y(35) = 0.8199$ | 33 [31,35] | 35 [28, 49] |
| 0.95 | 42 | $\hat{F}_Y(42) = 0.9136$ | 40 [38, 43] | 42 [30, 51] |
| 0.99 | 57 | $\hat{F}_Y(57) = 0.9846$ | 55 [54, 56] | 57 [50, 66] |
| 0.999 | 77 | $\hat{F}_Y(77) = 0.9988$ | 73 [68, 79] | 76 [63, 100] |
| 0.9995 | 83 | $\hat{F}_Y(83) = 0.9995$ | 79 [73, 91] | 81 [71, 100] |
| $ES_{0.05}(Y) = E[Y Y > q_{0.95}]$ $D_0=[42,142]$ | $2.85*(1/0.05) = 57$ | | $2.24*(1/0.05) = 44.8$ $[ES - 2\hat{\sigma}_T, ES + 2\hat{\sigma}_T] = [25.0, 64.4]$ | $2.61*(1/0.05) = 52.2$ $[ES - 2\hat{\sigma}_T, ES + 2\hat{\sigma}_T] = [41.6, 62.8]$ |

Tab. 10.1: Comparison for Poisson-Lognormal Model Note - Standard IS = Algorithm 1, Optimal IS = Algorithm 2,

| Algorithm | Computation of $f_Y(10)$ | $E[Y Y > q_{0.95}]$ |
|---|--|--|
| Basic Monte Carlo $N=50\text{mil}$ | $\sim 171\text{min}$ | $\sim 171\text{min}$ |
| Standard Importance Sampling Algorithm 1 * M samples: $f_Y(x = 10)$ * N samples per grid point: $E[Y Y > q_{0.95}]$ | $\sim 0.1\text{min}$ ($N=10\text{k}, N_T=100\text{k}$) $\sim 0.6\text{min}$ ($N=50\text{k}, N_T=500\text{k}$) $\sim 1\text{min}$ ($N=100\text{k}, N_T=1000\text{k}$) $\sim 4.9\text{min}$ ($N=500\text{k}, N_T=5000\text{k}$) | $N=50\text{k per grid point} - \{42,43, \dots, 142\}$ $\sim 45\text{min}$ ($N=50\text{k}, N_T=100\text{k}$) |
| Optimal Importance Sampling Algorithm 2 * N samples + (50k burn-in): $f_Y(10)$ * N samples+(100k burn-in) per grid point: $E[Y Y > q_{0.95}]$ | $\sim 2\text{min}$ ($N=10\text{k}, N_T=60\text{k}$) $\sim 3.5\text{min}$ ($N=50\text{k}, N_T=100\text{k}$) $\sim 5.3\text{min}$ ($N=100\text{k}, N_T=150\text{k}$) $\sim 20.5\text{min}$ ($N=500\text{k}, N_T=550\text{k}$) | $N=50\text{k} + (100\text{k burn-in})$ $N_T=150\text{k}$ per grid point - $\{42,43, \dots, 142\}$ $\sim 13\text{hrs}$ (serial processing) |
| | | $N=50\text{k} + (100\text{k burn-in})$ $N_T=150\text{k}$ per grid point - $\{42,43, \dots, 142\}$ $\sim (100/3)*5.3 \approx 171\text{ min}$ (parallel processing 3 CPUs) |

Tab. 10.2: Comparison of Computation Time

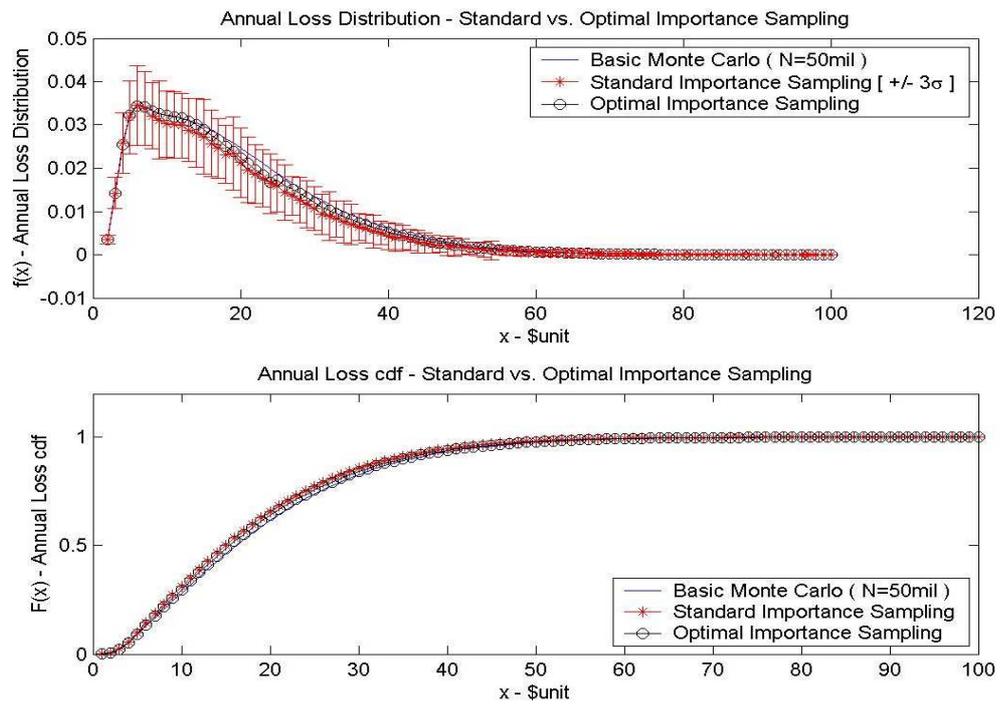


Fig. 10.8.1: Comparison of the Estimated Annual Loss distributions, empirical probability and cumulative probability distributions.

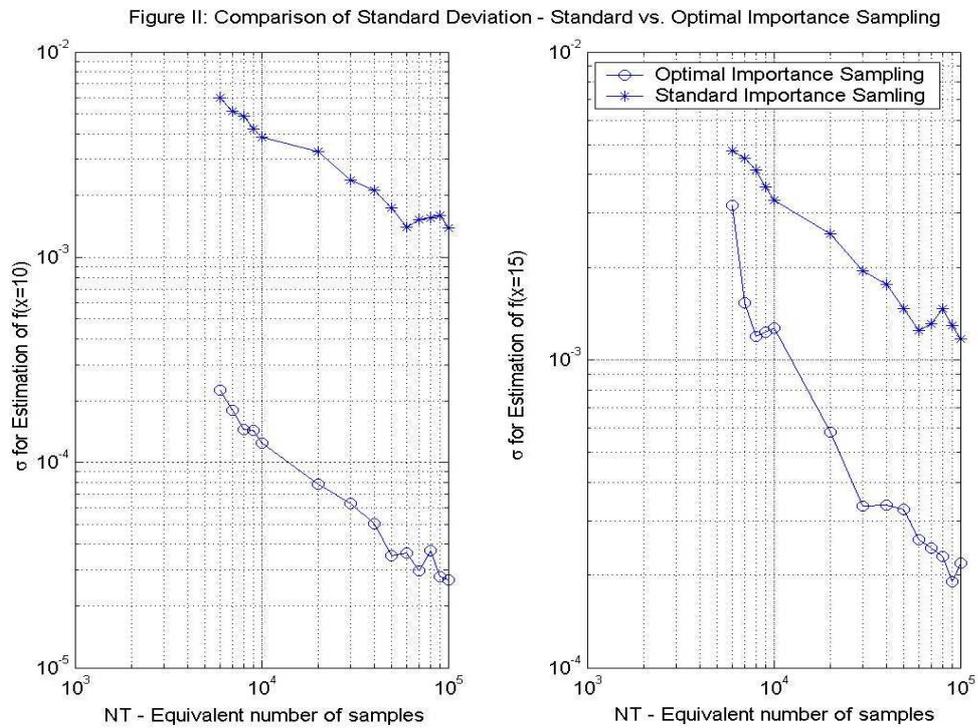


Fig. 10.8.2: Comparison of $\hat{\sigma}$ for the standard IS and MCMC IS versus the number of samples at grid points ($x=10,15$).

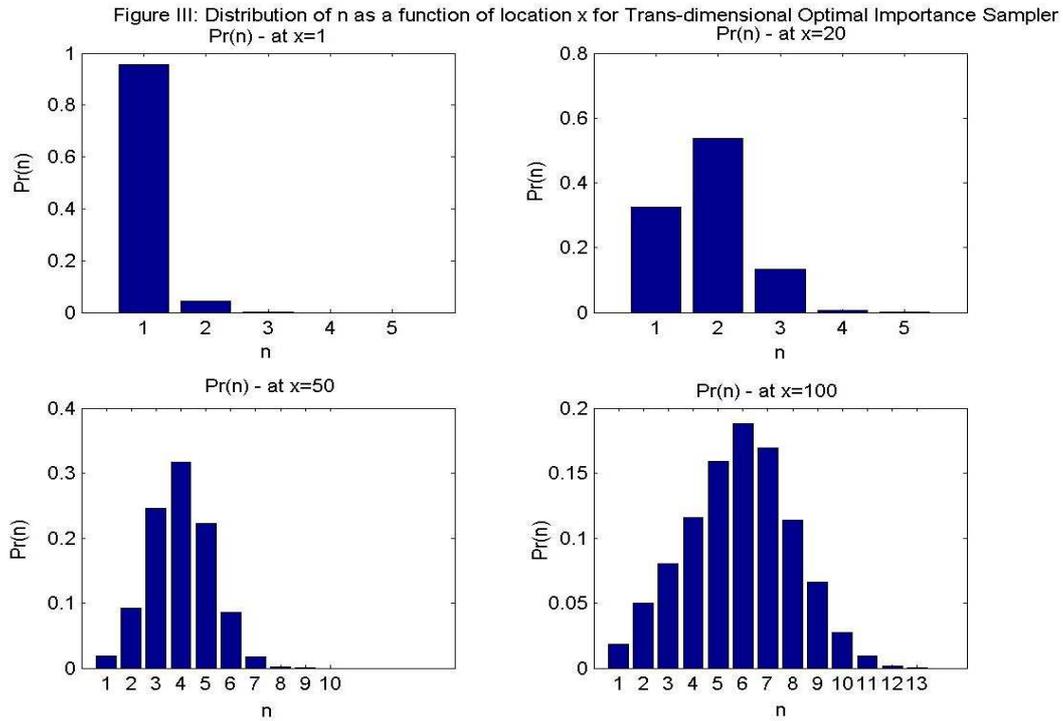


Fig. 10.8.3: Distribution of $p_{opt}(n)$ for selected grid evaluation points, x .

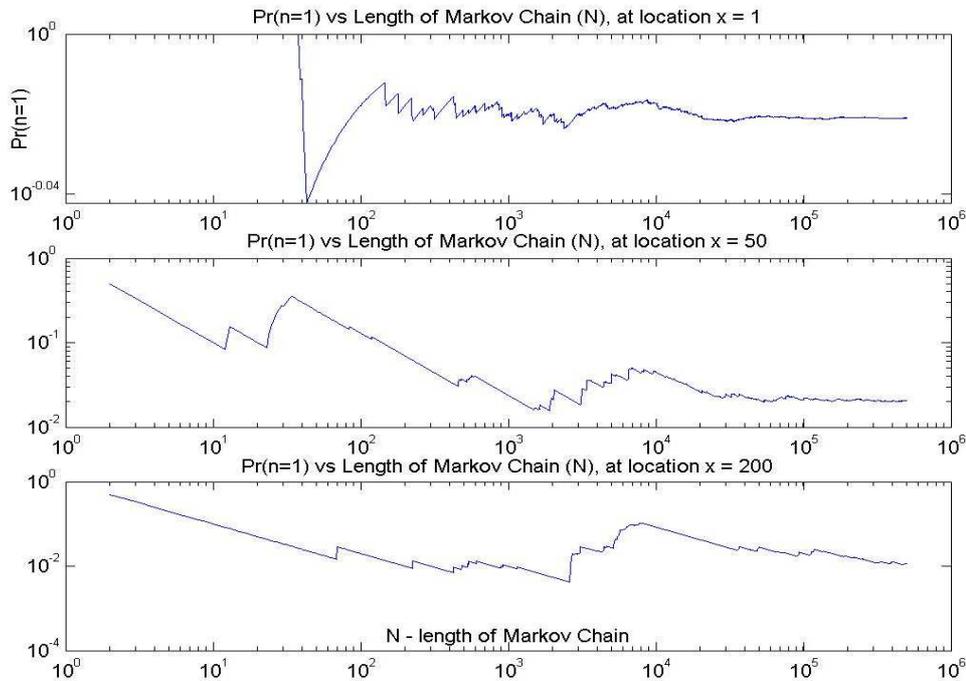


Fig. 10.8.4: Proportion of time $n=1$ versus length of chain N as a function of x .

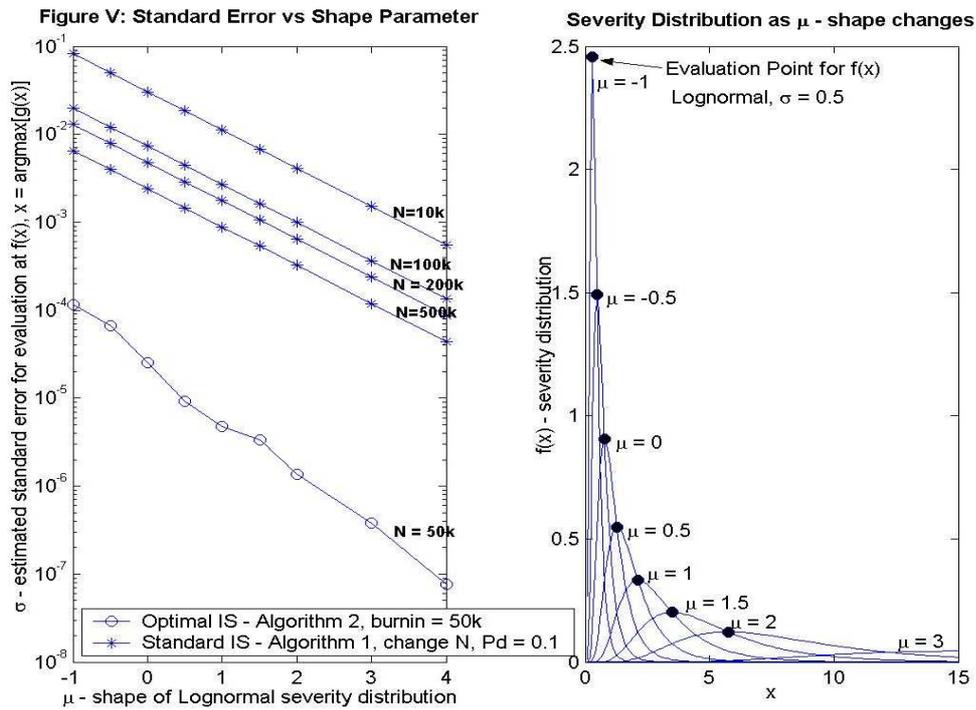


Fig. 10.8.5: Impact of the shape of the severity distribution on the variance estimate obtained when performing evaluation of $f_Y(x_m)$

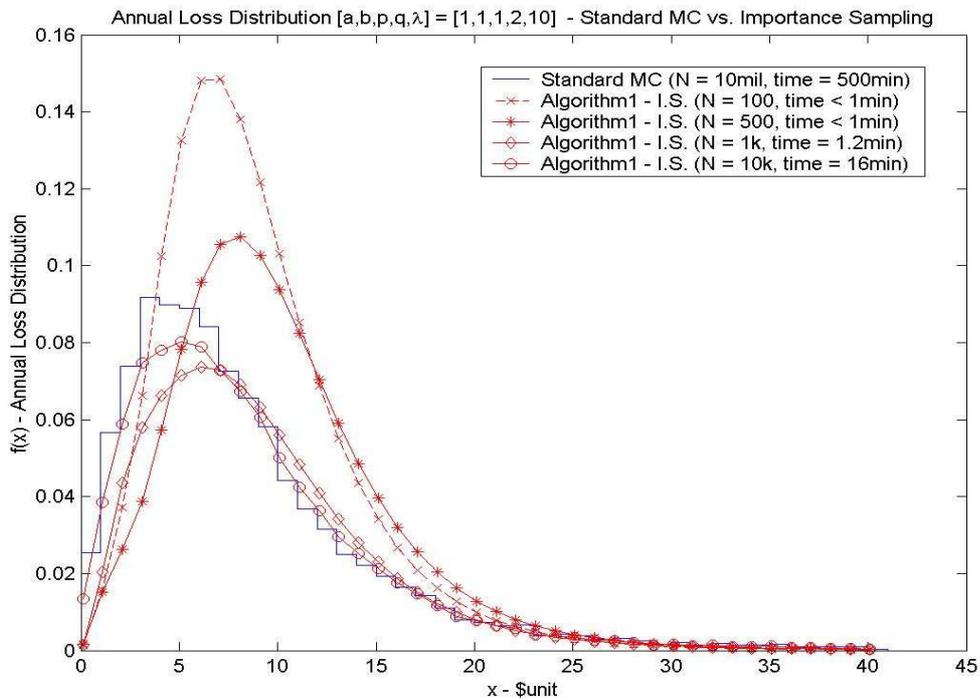


Fig. 10.8.6: Analysis of Annual Loss Distribution Estimate vs N (number of I.S. samples).

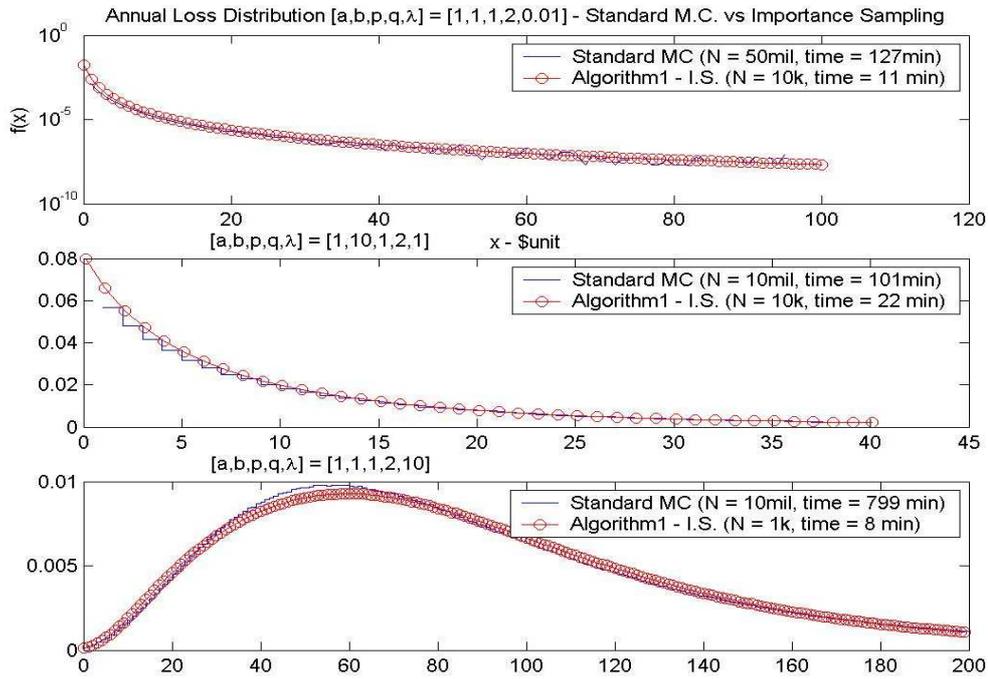


Fig. 10.8.7: Analysis of Annual Loss Distribution Estimate vs λ (mean of frequency distribution).

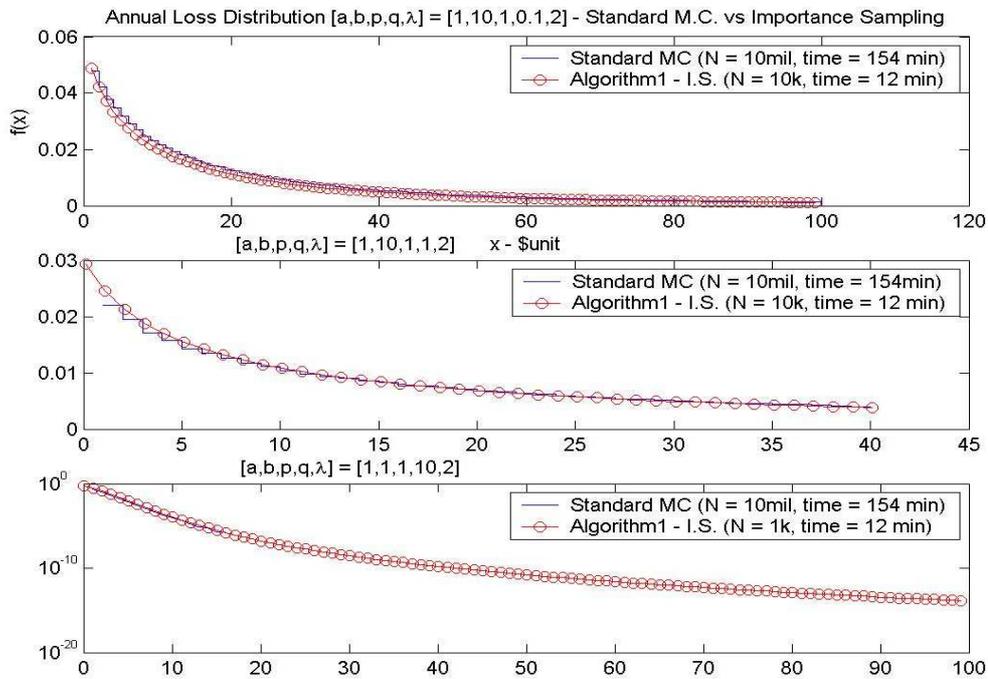


Fig. 10.8.8: Analysis of Annual Loss Distribution Estimate vs q (shape of the severity distribution).

References

- [1] Baker C. (2000). *Volterra Integral Equations*, Numerical Analysis Report No. 366, Manchester Center for Computational Mathematics Numerical Analysis Reports.
- [2] Baker C. (1977). *The Numerical Treatment of Integral Equations*, Oxford Clarendon Press.
- [3] Bartolucci F., Scaccia L. and Mira A. (2006) *Efficient Bayes factor estimation from the reversible jump output*, *Biometrika*, **93**(1), 41-52.
- [4] Bookstaber R. and J. McDonald (1987). *A general distribution for describing security price returns*. *The Journal of Business*, **60** (3), 401-424.
- [5] Brooks S., Giudici P. and Roberts G. (2003). *Efficient Construction of Reversible Jump MCMC Proposal Distributions (with discussion)*, *Journal of the Royal Statistical Society, Series B*, **65**, 3-55.
- [6] Chavez-Demoulin V., Embrechts P. and Neslehova J. (2006). *Quantitative Models for Operational Risk: Extremes, Dependence and Aggregation*. *Journal of Banking and Finance*, **30**(10), 2635-2658.
- [7] Cruz M. (2002). *Modelling, Measuring and Hedging Operational Risk*. John Wiley & Sons, Chapter 4.
- [8] Devroye L. (1986). *Non-Uniform Random Variate Generation*. Springer-Verlag, New York.
- [9] Doucet A. and Tadic V. (2007). *On Solving Integral Equations using Markov Chain Monte Carlo*, Technical Report CUED-F-INFENG Cambridge University no. 444.
- [10] Doucet A., de Freitas N. and Gordon, N. (2001). *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, New York.
- [11] Dutta K. and J. Perry (2006). *A tale of tails: An empirical analysis of loss distribution models for estimating operational risk capital*. Federal Reserve Bank of Boston, Working Papers No. 06-13.
- [12] Embrechts P., H. Furrer and R. Kaufmann (2003). *Quantifying regulatory capital for operational risk*. *Derivatives Use, Trading & Regulation*, **9** (3), 217—223.
- [13] Gelman A., J. B. Carlin, H.S. Stern and D.B. Rubin (1995). *Bayesian Data Analysis*. Chapman and Hall.

- [14] Gilks W., S. Richardson and D. Spiegelhalter (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall.
- [15] Glasserman, P. (2003). *Monte Carlo Methods in Financial Engineering*. Springer.
- [16] Green P. (1995). *Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination*, *Biometrika*, **82**, 711-732.
- [17] Green P. (2003). *Trans-dimensional Markov Chain Monte Carlo*, chapter from *Highly Structured Stochastic Systems*, Oxford University Press.
- [18] Kipnis, C. and Varadhan, S. R. S. (1986). *Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions*. *Communications in Mathematical Physics*, **104**, 1-19.
- [19] Klugman, S. Panjer, H. and Willmot, G. (2004). *Loss Models: From Data to Decisions*. John Wiley, New York.
- [20] Linz P. (1987). *Analytical and Numerical Methods for Volterra Equations*, *Mathematics of Computation*, **48**, No. 178, 841-842
- [21] Menn C. (2006). *Calibrated FFT-based Density Approximation for alpha-stable Distributions*. *Computational Statistics and Data Analysis*, **50**(8), 1891-1904.
- [22] Meyn S. and R. Tweedie (1993). *Markov Chains and Stochastic Stability*, Springer.
- [23] Neslehova J, Embrechts P. and Chavez-Demoulin V. (2006). *Infinite Mean Models and the LDA for Operational Risk*. *Journal of Operational Risk*, **1**(1), 3-25.
- [24] Orsi A. (1996). *Product Integration for Volterra Integral Equations of the Second Kind with Weakly Singular Kernels*, *Mathematics of Computation*, **65**, No. 215, 1201-1212
- [25] Panjer H. (2006). *Operational Risk: Modeling Analytics*, Wiley.
- [26] Panjer H. and Wilmott G. (1992), *Insurance Risk Models*, Chicago Society of Actuaries.
- [27] Panjer H. (1981). *Recursive Evaluation of a Family of Compound Distributions*, *Astin Bulletin* **12**, 22-26.
- [28] Peters G. and Teruads V. (2007). *Low Probability Large Consequence Events*, Australian Centre of Excellence for Risk Analysis, Project No. 06/02.
- [29] Peters G. and Sisson S. (2006). *Bayesian Inference Monte Carlo Sampling and Operational Risk*. *Journal Of Operational Risk*, **1**, No. 3.
- [30] Robert C.P. and Casella G. (2004). *Monte Carlo Statistical Methods, 2nd Edition*. Springer Texts in Statistics.
- [31] Rotar, V. (2007). *Actuarial Models – The Mathematics of Insurance*. Chapman & Hall CRC.
- [32] Shevchenko, P. and M. Wuthrich (2006). *The structural modelling of operational risk via Bayesian inference: Combining loss data with expert opinions*. CSIRO Technical Report Series, CMIS Call Number 2371.
- [33] Sisson, S. (2005). *Trans-dimensional Markov Chains: A Decade of Progress and Future Perspectives*. *Journal Of American Statistical Association*, **100**, 1077-1089.

-
- [34] Stroter, B. (1984). *The Numerical Evaluation of the Aggregate Claim Density Function via Integral Equations*. *Blatter der Deutschen Gesellschaft für Versicherungs-mathematik*, **17**, 1-14.
- [35] Sundt, B and Jewell, W. (1981). *Further Results on Recursive Evaluation of Compound Distributions*. *Astin Bulletin*, **12**, 27-39.
- [36] Willmot, G. and Panjer, H. (1985). *Difference Equation Approaches in Evaluation of Compound Distributions*. *Insurance: Mathematics and Economics*, **6**, 195-202.
- [37] Willmot, G. (1986). *Mixed Compound Poisson Distributions*. *Astin Bulletin*, **16**, S, S59-S79.

Journal Paper 8

"Our imagination is stretched to the utmost, not, as in fiction, to imagine things which are not really there, but just to comprehend those things which are there."

Richard Feynman

Peters G.W., Shevchenko P. and Wuthrich M. (2009) "Dynamic Operational Risk: modeling dependence and combining different sources of information". *Journal of Operational Risk*, 4(2), 69-104.

This work was instigated by Pavel Shevchenko at CSIRO. The first author can claim around 70% of the credit for the contents. This work was presented at an international Statistical conference in Queensland (ISBA) and received positive feedback from academics after the poster session. The first authors work included developing the methodology contained, developing the applications, the implementation and comparison to alternative approaches, writing the drafts of the paper and undertaking revisions. The paper has been accepted and will appear in 2009, *Journal of Operational Risk*. Permission from all the co-authors is granted for inclusion in the PhD.

This paper appears in the thesis in a modified format to that which finally appeared in the *Journal of Operational Risk*, where it was published.

Final print version available at: <http://www.journalofoperationalrisk.com/>

Dynamic operational risk: modelling dependence and combining different sources of information

Gareth W. Peters

CSIRO Mathematical and Information Sciences, Locked Bag 17, North Ryde, NSW, 1670, Australia;

UNSW Mathematics and Statistics Department, Sydney, 2052, Australia;

email: peterga@maths.unsw.edu.au

Pavel V. Shevchenko (*corresponding author*)

CSIRO Mathematical and Information Sciences, Sydney, Locked Bag 17, North Ryde, NSW, 1670,

Australia; e-mail: Pavel.Shevchenko@csiro.au

Mario V. Wüthrich

ETH Zurich, Department of Mathematics, CH-8092 Zurich, Switzerland;

email: wueth@math.ethz.ch

Submitted: 11 April 2009

11.1 abstract

In this paper, we model dependence between operational risks by allowing risk profiles to evolve stochastically in time and to be dependent. This allows for a flexible correlation structure where the dependence between frequencies of different risk categories and between severities of different risk categories as well as within risk categories can be modelled. The model is estimated using the Bayesian inference methodology, allowing for combination of internal data, external data and expert opinion in the estimation procedure. We use a specialized Markov chain Monte Carlo simulation methodology known as Slice sampling to obtain samples from the resulting posterior distribution and estimate the model parameters.

Keywords: dependence modelling, copula, compound process, operational risk, Bayesian inference, Markov chain Monte Carlo, Slice sampling.

11.2 Introduction

Modelling dependence between different risk cells and factors is an important challenge in operational risk (OpRisk) management. The difficulties of correlation modelling are well known and, hence, regulators typically take a conservative approach when considering correlation in risk models. For example, the Basel II OpRisk regulatory requirements for the Advanced Measurement Approach, BIS (2006) p.152, states *“Risk measures for different operational risk estimates must be added for purposes of calculating the regulatory minimum capital requirement. However, the bank may be permitted to use internally determined correlations in operational risk losses across individual operational risk estimates, provided it can demonstrate to the satisfaction of the national supervisor that its systems for determining correlations are sound, implemented with integrity, and take into account the uncertainty surrounding any such correlation estimates (particularly in periods of stress). The bank must validate its correlation assumptions using appropriate quantitative and qualitative techniques.”*

The current risk measure specified by regulatory authorities is Value-at-Risk (VaR) at the 0.999 level for a one year holding period. In this case simple summation over VaRs corresponds to an assumption of perfect dependence between risks. This can be very conservative as it ignores any diversification effects. If the latter are allowed in the model, capital reduction can be significant providing a strong incentive to model dependence in the banking industry. At the same time, limited data does not allow for reliable estimates of correlations and there are attempts to estimate these using expert opinions. In such a setting a transparent dependence model is very important from the perspective of model interpretation, understanding of model sensitivity and with the aim of minimizing possible model risk. However, we would also like to mention that VaR is not a coherent risk measure, see Artzner, Delbaen, Eber and Heath (1999). This means that in principal dependence modelling could also increase VaR, see Embrechts, Nešlehová and Wüthrich (2009) and Embrechts, Lambrigger and Wüthrich (2009).

Under Basel II requirements, the financial institution intending to use the Advanced Measurement Approach (AMA) for quantification of OpRisk should demonstrate accuracy of the internal model within 56 risk cells (eight business lines times seven event types). To meet regulatory requirements, the model should make use of internal data, relevant external data, scenario analysis and factors reflecting the business environment and internal control systems. The definition of OpRisk, Basel II requirements and the possible Loss Distribution Approach for AMA were discussed widely in the literature, see e.g. Cruz (2004), Chavez-Demoulin, Embrechts and Nešlehová (2006), Frachot, Moudoulaud and Roncalli (2004), Shevchenko (2009). It is more or less widely accepted that under the Loss Distribution Approach of AMA Basel II requirements, the banks should quantify distributions for frequency and severity of OpRisk for each business line and event type over a one year time horizon. These are combined into an annual loss distribution for the bank top level (as well as business lines and event types if required) and the bank capital (unexpected loss) is estimated using the 0.999 quantile of the annual loss distribution. If the severity and frequency distribution parameters are known, then the capital estimation can be accomplished using different techniques. In the case of single risks there are: hybrid

Monte Carlo approaches, see Peters, Johansen and Doucet (2007); Panjer Recursions, see Panjer (1981); integration of the characteristic functions, see Luo and Shevchenko (2009); Fast Fourier Transform techniques, see e.g. Embrechts and Frei (2008), Temnov and Warnung (2008). To account for parameter uncertainty, see Shevchenko (2008), and in multivariate settings Monte Carlo methods are typically used.

The commonly used model for an annual loss in a risk cell (business line/event type) is a compound random variable,

$$Z_t^{(j)} = \sum_{s=1}^{N_t^{(j)}} X_s^{(j)}(t). \quad (11.2.1)$$

Here $t = 1, 2, \dots, T, T + 1$ in our framework is discrete time (in annual units) with $T + 1$ corresponding to the next year. The upper script j is used to identify the risk cell. The annual number of events $N_t^{(j)}$ is a random variable distributed according to a frequency counting distribution $P^{(j)}(\cdot | \lambda_t^{(j)})$, typically Poisson, which also depends on time dependent parameter(s) $\lambda_t^{(j)}$. The severities in year t are represented by random variables $X_s^{(j)}(t)$, $s \geq 1$, distributed according to a severity distribution $F^{(j)}(\cdot | \psi_t^{(j)})$, typically lognormal, Weibull or generalized Pareto distributions with parameter(s) $\psi_t^{(j)}$. Note, the index j on the distributions $P^{(j)}$ and $F^{(j)}$ reflects that distribution type can be different for different risks, for simplicity of notation we shall omit this j , using $P(\cdot | \lambda_t^{(j)})$ and $F(\cdot | \psi_t^{(j)})$, hereafter. The variables $\lambda_t^{(j)}$ and $\psi_t^{(j)}$ generically represent distribution (model) parameters of the j^{th} risk that we refer hereafter to as the risk profiles. Typically, it is assumed that given $\lambda_t^{(j)}$ and $\psi_t^{(j)}$, the frequency and severities of the j^{th} risk are independent and the severities within the j^{th} risk are independent too. The total bank's loss in year t is calculated as

$$Z_t = \sum_{j=1}^J Z_t^{(j)}, \quad (11.2.2)$$

where formally for OpRisk under the Basel II requirements $J = 56$ (seven event types times eight business lines). However, this may differ depending on the financial institution and type of problem.

Conceptually under model (11.2.1), the dependence between the annual losses $Z_t^{(j)}$ and $Z_t^{(i)}$, $i \neq j$, can be introduced in several ways. For example via:

- Modelling dependence between frequencies $N_t^{(j)}$ and $N_t^{(i)}$ directly through e.g. copula methods, see e.g. Frachot, Roncalli and Salomon (2004), Bee (2005) and Aue and Klakbrener (2006) or common shocks, see e.g. Lindskog and McNeil (2003), Powojowski, Reynolds and Tuenter (2002). We note that the use of copula methods, in the case of discrete random variables, needs to be done with care. The approach of common shocks is proposed as a method to model events affecting many cells at the same time. Formally, this leads to dependence between frequencies of the risks if superimposed with cell internal events. One can introduce the dependence between event times of different risks, e.g. the 1^{st} event time of the j^{th} risk correlated to the 1^{st} event time of the i^{th} risk, etc., but it can be problematic to interpret such a model.

- Considering dependence between severities (e.g. the first loss amount of the j^{th} risk is correlated to the first loss of the i^{th} risk, second loss in the j^{th} risk is correlated to second loss in the i^{th} risk, etc), see e.g. Chavez-Demoulin, Embrechts and Nešlehová (2006). This can be difficult to interpret especially when one considers high frequency versus low frequency risks.
- Modelling dependence between annual losses directly via copula methods, see Giacometti, Rachev, Chernobai and Bertocchi (2008), Böcker and Klüppelberg (2008) and Embrechts and Puccetti (2008). However, this may create irreconcilable problems with modelling insurance for OpRisk that directly involves event times. Additionally, it will be problematic to quantify these correlations using historical data, and the LDA model (11.2.1) will lose its structure. Though one can consider dependence between losses aggregated over shorter periods.

In this paper, we assume that all risk profiles are stochastically evolving in time. That is we model risk profiles $\lambda_t = (\lambda_t^{(1)}, \dots, \lambda_t^{(J)})$ and $\psi_t = (\psi_t^{(1)}, \dots, \psi_t^{(J)})$ by random variables $\Lambda_t = (\Lambda_t^{(1)}, \dots, \Lambda_t^{(J)})$ and $\Psi_t = (\Psi_t^{(1)}, \dots, \Psi_t^{(J)})$, respectively. We introduce dependence between risks by allowing dependence between their risk profiles Λ_t and Ψ_t . Note that, independence between frequencies and severities in (11.2.1) is conditional on risk profiles (Λ_t, Ψ_t) only. Additionally we assume, for the sake of simplicity, that all risks are independent conditional on risk profiles.

Stochastic modelling of risk profiles may appeal to intuition. For example consider the annual number of events for the j^{th} risk modelled as random variables from Poisson distribution $Poi(\Lambda_t^{(j)} = \lambda_t^{(j)})$. Conditional on $\Lambda_t^{(j)}$, the expected number of events per year is $\Lambda_t^{(j)}$. The latter is not only different for different banks and different risks but also changes from year to year for a risk in the same bank. In general, the evolution of $\Lambda_t^{(j)}$, can be modelled as having deterministic (trend, seasonality) and stochastic components. In actuarial mathematics this is called a mixed Poisson model. For simplicity, in this paper, we assume that $\Lambda_t^{(j)}$ is purely stochastic and distributed according to a Gamma distribution.

Now consider a sequence $(\Lambda_1, \Psi_1), \dots, (\Lambda_{T+1}, \Psi_{T+1})$. It is naive to assume that risk profiles of all risks are independent. Intuitively these are dependent, for example, due to changes in politics, regulations, law, economy, technology (sometimes called drivers or external risk factors) that jointly impact on many risk cells at each time instant. In this paper we focus on dependence between risk profiles.

We begin by presenting the general model and then we perform analysis of relevant properties of this model in a bivariate risk setting. Next, we demonstrate how to perform inference under our model by adopting a Bayesian approach that allows one to combine internal data with expert opinions and external data. We consider both the single risk and multiple risk settings for the example of modelling claims frequencies. Then we present an advanced simulation procedure utilizing a class of Markov chain Monte Carlo (MCMC) algorithms which allow us to sample from the posterior distributions developed. Finally, we demonstrate the performance

of both the model and the simulation procedure in several examples, before finishing with a discussion and conclusions.

The main objective of the paper is to preset the framework we develop for the multivariate problem and to demonstrate estimation in this setting. Application of real data is the subject of further research. To clarify notation, we shall use upper case symbols to represent random variables, lower case symbols for their realizations and bold for vectors.

11.3 Model

Model Assumptions 11.3.1. Consider J risks each with a general model (11.2.1) for the annual loss in year t , $Z_t^{(j)}$, and each modelled by severity $X_s^{(j)}(t)$ and frequency $N_t^{(j)}$. The frequency and severity risk profiles are modelled by random vectors $\mathbf{\Lambda}_t = (\Lambda_t^{(1)}, \dots, \Lambda_t^{(J)})$ and $\mathbf{\Psi}_t = (\Psi_t^{(1)}, \dots, \Psi_t^{(J)})$ respectively and parameterized by risk characteristics $\boldsymbol{\theta}_\Lambda = (\theta_\Lambda^{(1)}, \dots, \theta_\Lambda^{(J)})$ and $\boldsymbol{\theta}_\Psi = (\theta_\Psi^{(1)}, \dots, \theta_\Psi^{(J)})$ correspondingly. Additionally, the dependence between risk profiles is parameterized by $\boldsymbol{\theta}_\rho$. Assume that, given $\boldsymbol{\theta} = (\boldsymbol{\theta}_\Lambda, \boldsymbol{\theta}_\Psi, \boldsymbol{\theta}_\rho)$:

1. The random vectors,

$$\begin{aligned} & \left(\mathbf{\Psi}_1, \mathbf{\Lambda}_1, N_1^{(j)}, X_s^{(j)}(1); j = 1, \dots, J, s \geq 1 \right) \\ & \vdots \\ & \left(\mathbf{\Psi}_{T+1}, \mathbf{\Lambda}_{T+1}, N_{T+1}^{(j)}, X_s^{(j)}(T+1); j = 1, \dots, J, s \geq 1 \right) \end{aligned}$$

are independent. That is, given $\boldsymbol{\theta}$, between different years, the risk profiles for frequencies and severities as well as the number of losses and actual losses are independent.

2. The vectors $(\mathbf{\Psi}_1, \mathbf{\Lambda}_1), \dots, (\mathbf{\Psi}_{T+1}, \mathbf{\Lambda}_{T+1})$ are i.i.d. from a joint distribution with marginal distributions $\Lambda_t^{(j)} \sim G(\cdot | \theta_\Lambda^{(j)})$, $\Psi_t^{(j)} \sim H(\cdot | \theta_\Psi^{(j)})$ and $2J$ -dimensional copula $C(\cdot | \boldsymbol{\theta}_\rho)$.

3. Given $\mathbf{\Lambda}_t = \boldsymbol{\lambda}_t$ and $\mathbf{\Psi}_t = \boldsymbol{\psi}_t$: the compound random variables $Z_t^{(1)}, \dots, Z_t^{(J)}$ are independent with $N_t^{(j)}$ and $X_1^{(j)}(t), X_2^{(j)}(t), \dots$ independent; frequencies $N_t^{(j)} \sim P(\cdot | \lambda_t^{(j)})$; and severities $X_s^{(j)}(t) \stackrel{i.i.d.}{\sim} F(\cdot | \psi_t^{(j)})$, $s \geq 1$.

Calibration of the above model requires estimation of $\boldsymbol{\theta}$. A thorough discussion about the interpretation and role of $\boldsymbol{\theta}$ is provided in Section 11.5, where it will be treated within a Bayesian framework as a random variable $\boldsymbol{\Theta}$ to incorporate expert opinions and external data into the estimation procedure. Also note that for simplicity of notation, we assumed one severity risk profile $\Psi_t^{(j)}$ and one frequency risk profile $\Lambda_t^{(j)}$ per risk - extension is trivial if more risk profiles are required to model risk.

Copula models. To define the above model, a copula function $C(\cdot)$ should be specified to model dependence between the risk profiles. For a description of copulas in the context of financial risk modelling see McNeil, Frey and Embrechts (2005). In general, a copula is a d -dimensional multivariate distribution on $[0, 1]^d$ with uniform marginal distributions. Given a copula function $C(u_1, \dots, u_d)$, the joint distribution of rvs Y_1, \dots, Y_d with marginal distributions $F_1(y_1), \dots, F_d(y_d)$ can be constructed as

$$F(y_1, \dots, y_d) = C(F_1(y_1), \dots, F_d(y_d)). \quad (11.3.1)$$

A well known theorem due to Sklar, published in 1959, says that one can always find a unique copula $C(\cdot)$ for a joint distribution with given continuous marginals. Note that in the case of discrete distributions this copula may not be unique. Given (11.3.1), the joint density can be written as

$$f(y_1, \dots, y_d) = c(F_1(y_1), \dots, F_d(y_d)) \prod_{i=1}^d f_i(y_i). \quad (11.3.2)$$

where $c(\cdot)$ is a copula density and $f_1(y_1), \dots, f_d(y_d)$ are marginal densities. In this paper, for illustration purposes we consider the Gaussian, Clayton and Gumbel copulas (Clayton and Gumbel copulas belong to a so-called family of the Archimedean copulas):

- **Gaussian copula:**

$$c(u_1, \dots, u_d | \Sigma) = \frac{f_N^\Sigma(F_N^{-1}(u_1), \dots, F_N^{-1}(u_d))}{\prod_{i=1}^d f_N(F_N^{-1}(u_i))}, \quad (11.3.3)$$

where $F_N(\cdot)$ and $f_N(\cdot)$ are the standard Normal distribution and its density respectively and $f_N^\Sigma(\cdot)$ is a multivariate Normal density with zero means, unit variances and correlation matrix Σ .

- **Clayton copula:**

$$c(u_1, \dots, u_d | \rho) = \left(1 - d + \sum_{i=1}^d (u_i)^{-\rho} \right)^{-d - \frac{1}{\rho}} \prod_{i=1}^d \left((u_i)^{-\rho - 1} \{(i-1)\rho + 1\} \right), \quad (11.3.4)$$

where $\rho > 0$ is a dependence parameter.

- **Gumbel copula:**

$$c(u_1, \dots, u_d | \rho) = \frac{\partial^d}{\partial u_1 \dots \partial u_d} C(u_1, \dots, u_d | \rho), \quad (11.3.5)$$

$$C(u_1, \dots, u_d | \rho) = \exp \left\{ - \left(\sum_{i=1}^d (-\log(u_i))^\rho \right)^{\frac{1}{\rho}} \right\}, \quad (11.3.6)$$

where $\rho \geq 1$ is a dependence parameter.

In the bivariate case the explicit expression for Gumbel copula is given by

$$\begin{aligned}
 c(u_1, u_2|\rho) &= \frac{\partial^2}{\partial u_1 \partial u_2} C(u_1, u_2|\rho) \\
 &= C(u_1, u_2|\rho) u_1^{-1} u_2^{-1} \left[\sum_{i=1}^2 (-\log(u_i))^\rho \right]^{2\left(\frac{1}{\rho}-1\right)} [\log(u_1) \log(u_2)]^{\rho-1} \\
 &\quad \times \left[1 + (\rho - 1) \left[\sum_{i=1}^2 (-\log(u_i))^\rho \right]^{-\frac{1}{\rho}} \right].
 \end{aligned}$$

An important difference between these three copulas is that they each display different tail dependence properties. The Gaussian copula has no upper or lower tail dependence, the Clayton copula produces lower tail dependence, whereas the Gumbel copula produces upper tail dependence, see McNeil, Frey and Embrechts (2005).

Common factor models. The use of common (systematic) factors is useful to identify dependent risks and to reduce the number of required correlation coefficients that must be estimated. For example, assuming a Gaussian copula between risk profiles, consider one common factor Ω_t affecting all risk profiles as follows

$$\begin{aligned}
 Y_t^{(i)} &= \rho_i \Omega_t + \sqrt{1 - \rho_i^2} W_t^{(i)}, i = 1, \dots, 2J; \\
 \Lambda_t^{(j)} &= G^{-1}(F_N(Y_t^{(j)}|\theta_\Lambda^{(j)}), \Psi_t^{(j)} = H^{-1}(F_N(Y_t^{(j+J)}|\theta_\Psi^{(j)}), j = 1, \dots, J, \quad (11.3.7)
 \end{aligned}$$

where $W_t^{(1)}, \dots, W_t^{(2J)}$ and Ω_t are iid from the standard Normal distribution and all rvs are independent between different time steps t . Given Ω_t , all risk profiles are independent but unconditionally the risk profiles are dependent if the corresponding ρ_i are nonzero. In this example, one should identify $2J$ correlation parameters ρ_i only instead of $J(J-1)/2$ parameters of the full correlation matrix. Often, common factors are unobservable and practitioners use generic intuitive definitions such as: changes in political, legal and regulatory environments, economy, technology, system security, system automation, etc. Several external and internal factors are typically considered. The factors may affect the frequency risk profiles (e.g. system automation), the severity risk profiles (e.g. changes in legal environment) or both the frequency and severity risk profiles (e.g. system security). For more details on the use and identification of the factor models, see Section 3.4 in McNeil, Frey and Embrechts (2005); also, see Sections 5.3 and 7.4 in Marshall (2001) for the use in the operational risk context.

In general, a copula can be introduced between all risk profiles. Though, for simplicity, in the simulation examples below, presented for two risks, we consider dependence between severities and frequencies separately. Also, in this paper, the estimation procedure is presented for frequencies only. The actual procedure can be extended in the same manner as presented to severities but it is the subject of further work.

11.4 Simulation Study - Bivariate Case

We start with a bivariate model, where we study the strength of dependence at the annual loss level obtained through dependence in risk profiles, as discussed above. We consider two scenarios. The first involves independent severity risk profiles and dependent frequency risk profiles. The second involves dependence between the severity risk profiles and independence between the frequency profiles. In both scenarios, we consider three bivariate copulas (Gaussian, Clayton and Gumbel copulas (11.3.3)-(11.3.6)) denoted as $C(u_1, u_2|\rho)$ and parameterized by one parameter ρ which controls the degree of dependence. In the case of Gaussian copula, ρ is a non-diagonal element of correlation matrix Σ in (11.3.3).

Bivariate model for risk profiles. We assume that Model Assumptions 11.3.1 are fulfilled for the aggregated losses

$$Z_t^{(1)} = \sum_{s=1}^{N_t^{(1)}} X_s^{(1)}(t) \quad \text{and} \quad Z_t^{(2)} = \sum_{s=1}^{N_t^{(2)}} X_s^{(2)}(t).$$

As marginals, for $j = 1, 2$ we choose:

- $N_t^{(j)} \sim Poi(\lambda_t^{(j)})$ and $X_s^{(j)}(t) \stackrel{i.i.d.}{\sim} LN(\psi_t^{(j)}, \sigma^{(j)})$, $s \geq 1$.
- $\Lambda_t^{(j)} \sim \Gamma(\alpha_\Lambda^{(j)}, \beta_\Lambda^{(j)})$, $\Psi_t^{(j)} \sim N(\mu_\Psi^{(j)}, \omega_\Psi^{(j)})$.

Here, $\Gamma(\alpha, \beta)$ is a Gamma distribution with mean α/β and variance α/β^2 , $N(\mu, \sigma)$ is a Gaussian distribution with mean μ and standard deviation σ , and $LN(\mu, \sigma)$ is a lognormal distribution.

In analyzing the induced dependence between annual losses, we consider two scenarios:

- Scenario 1: $\Lambda_t^{(1)}$ and $\Lambda_t^{(2)}$ are dependent via copula $C(u_1, u_2|\rho)$ while $\Psi_t^{(1)}$ and $\Psi_t^{(2)}$ are independent.
- Scenario 2: $\Psi_t^{(1)}$ and $\Psi_t^{(2)}$ are dependent via copula $C(u_1, u_2|\rho)$ while $\Lambda_t^{(1)}$ and $\Lambda_t^{(2)}$ are independent.

Here, parameter ρ corresponds to θ_ρ in Model Assumptions 11.3.1. The simulation of the annual losses when risk profiles are dependent via a copula can be accomplished as shown in Appendix 11.11. Utilizing this procedure, we examine the strength of dependence between the annual losses if there is a dependence between the risk profiles. In the next sections we will demonstrate the Bayesian inference model and associated methodology to perform estimation of the model parameters. Here, we assume the parameters are known *a priori* with the following values used in our specific example:

- $\alpha_\Lambda^{(j)} = 5, \beta_\Lambda^{(j)} = 0.1, \mu_\Psi^{(j)} = 2, \omega_\Psi^{(j)} = 0.4, \sigma^{(j)} = 1; j = 1, 2$

These parameters correspond to θ_Λ and θ_Ψ in Model Assumptions 11.3.1. In Figure 11.13.1, we present three cases where $C(\cdot|\rho)$ is a Gaussian, Clayton or Gumbel copula under both scenario 1 and scenario 2. In each of these examples we vary the parameter of the copula model ρ from weak to strong dependence. The annual losses are not Gaussian distributed and to measure the dependence between the annual losses we use a non-linear rank correlation measure, Spearman's rank correlation, denoted by $\rho_{SR}(Z_t^{(1)}, Z_t^{(2)})$. The Spearman's rank correlation between the annual losses was estimated using 10,000 simulated years for each value of ρ . In these and other numerical experiments we conducted, the range of possible dependence between the annual losses of different risks induced by the dependence between risk profiles is very wide and should be flexible enough to model dependence in practice. Note, the degree of induced correlation can be further extended by working with more flexible copula models at the expense of estimation of a larger number of model parameters.

11.5 Bayesian Inference: combining different data sources

In this section we estimate the model introduced in Section 2 using a Bayesian inference method. To achieve this we must consider that the requirements of Basel II AMA (see BIS, p.152) clearly state that: *"Any operational risk measurement system must have certain key features to meet the supervisory soundness standard set out in this section. These elements must include the use of internal data, relevant external data, scenario analysis and factors reflecting the business environment and internal control systems"*. Hence, Basel II requires that OpRisk models include use of several different sources of information. We will demonstrate that to satisfy such requirements it is important that methodology such as the one we develop in this paper be considered in practice to ensure one can soundly combine these different data sources.

It is widely recognized that estimation of OpRisk frequency and severity distributions cannot be done solely using historical data. The reason is the limited ability to predict future losses in a banking environment which is constantly changing. Assume that a new policy was introduced in a financial institution with the intention of reducing an OpRisk loss. This cannot be captured in a model based solely on historical loss data.

For the above reasons, it is very important to include Scenario Analysis (SA) in OpRisk modelling. SA is a process undertaken by banks to identify risks; analyze past events experienced internally and jointly with other financial institutions including near miss losses; consider current and planned controls in the banks, etc. Usually, it involves surveying of experts through workshops. A template questionnaire is developed to identify weaknesses, strengths and other factors. As a result an imprecise, value driven quantitative assessment of risk frequency and severity distributions is obtained. On its own, SA is very subjective and we argue it should be combined (supported) by actual loss data analysis. It is not unusual that correlations between risks are attempted to be specified by experts in the financial institution, typically via SA.

External loss data is also an important source of information which should be incorporated into modelling. There are several sources available to obtain external loss data, for a discussion on

some of the data related issues associated with external data see Peters and Teruads (2007).

Additionally, the combination of expert opinions with internal and external data is a difficult problem and complicated ad-hoc procedures are used in practice. Some prominent risk professionals in industry have argued that statistically consistent combining of these different data sources is one of the most pertinent and challenging aspects of OpRisk modelling. It was quoted in Davis (2006) *"Another big challenge for us is how to mix the internal data with external data; this is something that is still a big problem because I don't think anybody has a solution for that at the moment"* and *"What can we do when we don't have enough data [...] How do I use a small amount of data when I can have external data with scenario generation? [...] I think it is one of the big challenges for operational risk managers at the moment."* Using the methodology that we develop in this paper, one may combine these data sources in a statistically sound approach, addressing these important practical questions that practitioners are facing under Basel II AMA.

Bayesian inference methodology is well suited to combine different data sources in OpRisk, for example see Shevchenko and Wüthrich (2006). A closely related credibility theory toy example was considered in Bühlmann, Shevchenko and Wüthrich (2007). We also note that in general questions of Bayesian model choice must be addressed, adding to this there is the additional complexity that estimation of the required posterior distributions will typically require MCMC, see Peters and Sisson (2006).

A Bayesian model to combine three data sources (internal data, external data and expert opinion) for the case of a single risk cell was presented in Lambrigger, Shevchenko and Wüthrich (2007). In this paper we extend this approach to the case of many risk cells with the dependence between risks introduced as in Section 11.3. Hereafter, for illustrative purposes we restrict to modelling frequencies only.

Hence, our objective will be to utilise Bayesian inference to estimate the parameters of the model through the combination of expert opinions and observed loss data (internal and external).

We note that as part of this Bayesian model formulation an information flow can be incorporated into the model. This could be introduced in many forms. The most obvious example involves incorporation of new data from actual observed losses. However, we stress that more general ideas are possible. For example, if new information becomes available (new policy introduced, etc) then experts can update their prior distributions to incorporate this information into the model.

Additionally, under a Bayesian model we note that SA could naturally form part of a subjective Bayesian prior elicitation procedure, see O'Hagan (2006).

11.5.1 Modelling frequencies for a single risk cell

Here we follow the Lambrigger, Shevchenko and Wüthrich (2007) approach to combine different data sources for one risk cell in the case of the Model Assumptions 11.3.1.

Define a model in which every risk cell of a financial company $j \in \{1, \dots, J\}$ is characterized by a risk characteristic $\Theta_\Lambda^{(j)}$ that describes the frequency risk profile $\Lambda_t^{(j)}$ in risk cell j . This $\Theta_\Lambda^{(j)}$ represents a vector of unknown distribution parameters of risk profile $\Lambda_t^{(j)}$. The true value of $\Theta_\Lambda^{(j)}$ is not known and modelled as a random variable. *A priori*, before having any company specific information, the prior distribution of $\Theta_\Lambda^{(j)}$ is based on external data only. Our aim then is to specify the distribution of $\Theta_\Lambda^{(j)}$ when we have company specific information about risk cell j such as observed losses and expert opinions. This is achieved by developing a Bayesian model and numerical estimation of relevant quantities is performed via MCMC methods. For simplicity, in this section, we drop the risk cell specific superscript j since we concentrate on modelling frequencies for single risk cell j , where $\Theta_\Lambda^{(j)}$ is a scalar Θ_Λ and all other parameters are assumed known.

Model Assumptions 11.5.1. *Assume that risk cell j has a fixed, deterministic volume V (i.e. number of transactions, etc.).*

1. *The risk characteristics Θ_Λ of risk cell j has prior distribution: $\Theta_\Lambda \sim \Gamma(a, b)$ for given parameters $a > 0$ and $b > 0$.*
2. *Given $\Theta_\Lambda = \theta_\Lambda$, $(\Lambda_1, N_1), \dots, (\Lambda_{T+1}, N_{T+1})$ are i.i.d. and the intensity of events of year $t \in \{1, \dots, T + 1\}$ has conditional marginal distribution $\Lambda_t \sim \Gamma(\alpha, \alpha/\theta_\Lambda)$ for a given parameter $\alpha > 0$.*
3. *Given $\Theta = \theta_\Lambda$ and $\Lambda_t = \lambda_t$, the frequencies $N_t \sim Poi(V\lambda_t)$.*
4. *The financial company has K expert opinions Δ_k , $k = 1, \dots, K$ about Θ_Λ . Given $\Theta_\Lambda = \theta_\Lambda$, Δ_k and (Λ_t, N_t) are independent for all k and t , and $\Delta_1, \dots, \Delta_K$ are i.i.d. with $\Delta_k \sim \Gamma(\xi, \xi/\theta_\Lambda)$.*

Remarks 11.5.2.

- In items 1) and 2) we choose a gamma distribution for the underlying parameters. Often, the available data is not sufficient to support such a choice. In such cases, in actuarial practice, one often chooses a gamma distribution. A gamma distribution is neither conservative nor aggressive and it has the advantage that it allows for transparent model interpretations. If other distributions are more appropriate then, of course, one should replace the gamma assumption. This can easily be done in our simulation methodology.
- Given that $\Theta_\Lambda \sim \Gamma(a, b)$, $E[\Theta_\Lambda] = a/b$ and $Var(\Theta_\Lambda) = a/b^2$. These are the prior two moments of the underlying risk characteristics Θ_Λ . The prior can be determined by external data (or regulator). In general, parameters a and b can be estimated by the maximum likelihood method using the data from all banks.
- Note that we have for the first moments

$$\begin{aligned} E[\Lambda_t | \Theta_\Lambda] &= \Theta_\Lambda, \quad E[\Lambda_t] = \frac{a}{b}, \quad E[N_t | \Theta_\Lambda, \Lambda_t] = V \Lambda_t, \\ E[N_t | \Theta_\Lambda] &= V \Theta_\Lambda, \quad E[N_t] = V \frac{a}{b}. \end{aligned}$$

The second moments are given by

$$\begin{aligned} \text{Var}(\Lambda_t | \Theta_\Lambda) &= \alpha^{-1} \Theta_\Lambda^2, & \text{Var}(\Lambda_t) &= \alpha^{-1} \frac{a^2}{b^2} + (\alpha^{-1} + 1) \frac{a}{b^2}, \\ \text{Var}(N_t | \Theta_\Lambda, \Lambda_t) &= V \Lambda_t, & \text{Var}(N_t | \Theta_\Lambda) &= V \Theta_\Lambda + V^2 \alpha^{-1} \Theta_\Lambda^2, \\ \text{Var}(N_t) &= V \frac{a}{b} + V^2 \alpha^{-1} \frac{a^2}{b^2} + V^2 (\alpha^{-1} + 1) \frac{a}{b^2}. \end{aligned} \quad (11.5.1)$$

For model interpretation purposes, consider the results for the coefficient of variation (CV), a convenient dimensionless measure of uncertainty commonly used in the insurance industry:

$$\lim_{V \rightarrow \infty} CV^2(N_t | \Theta_\Lambda) = \lim_{V \rightarrow \infty} \frac{\text{Var}(N_t | \Theta_\Lambda)}{E^2[N_t | \Theta_\Lambda]} = \alpha^{-1} > 0, \quad (11.5.2)$$

and

$$\lim_{V \rightarrow \infty} CV^2(N_t) = \lim_{V \rightarrow \infty} \frac{\text{Var}(N_t)}{E^2[N_t]} = \alpha^{-1} + (\alpha^{-1} + 1) a^{-1} > 0. \quad (11.5.3)$$

That is, our model makes perfect sense from a practical perspective. Namely, as volume increases, $V \rightarrow \infty$, there always remains a non-diversifiable element, see 11.5.2 and 11.5.3. This is exactly what has been observed in practice and what regulators require from internal models. Note, if we model Λ_t as constant and known then $\lim_{V \rightarrow \infty} CV^2(N_t | \Lambda_t) \rightarrow 0$.

- Contrary to the developments in Lambrigger, Shevchenko and Wüthrich (2007), where the intensity Λ_t was constant overtime, now Λ_t is a stochastic process. From a practical point of view, it is not plausible that the intensity of the annual counts is constant over time. In such a setting parameter risks completely vanish if we have infinitely many observed years or infinitely many expert opinions, respectively (see Theorem 3.6 (a) and (c) in Lambrigger, Shevchenko and Wüthrich (2007)). This is because Λ_t can then be perfectly forecasted. In the present model, parameter risks will also decrease with increasing information. As we gain information the posterior standard deviation of Θ_Λ will converge to 0. However, since Λ_{T+1} viewed from time T is always random, the posterior standard deviation for Λ_{T+1} will be finite.
- Note that, conditionally given $\Theta_\Lambda = \theta_\Lambda$, N_t has a negative binomial distribution with probability weights for $n \geq 0$,

$$P[N_t = n | \theta_\Lambda] = \binom{\alpha + n - 1}{n} \left(\frac{\alpha}{\alpha + \theta_\Lambda V} \right)^\alpha \left(\frac{\theta_\Lambda V}{\alpha + \theta_\Lambda V} \right)^n. \quad (11.5.4)$$

That is, at this stage we could directly work with a negative binomial distribution. As we will see below, only in the marginal case can we work with (11.5.4). In the multidimensional model we require Λ_t .

- Δ_k denotes the expert opinion of expert k which predicts the true risk characteristics Θ_Λ

of his company. We have

$$\begin{aligned} E [\Delta_k | \Theta_\Lambda] &= E [\Lambda_j | \Theta_\Lambda] = E [N_j/V | \Theta_\Lambda] = \Theta_\Lambda, \\ Var (\Delta_k | \Theta_\Lambda) &= \Theta_\Lambda^2/\xi, \quad CV (\Delta_k | \Theta_\Lambda) = \xi^{-1/2}. \end{aligned} \tag{11.5.5}$$

That is, the relative uncertainty CV in the expert opinion does not depend on the value of Θ_Λ . That means that ξ can be given externally, e.g. by the regulator, who is able to give a lower bound to the uncertainty. Moreover, we see that the expert predicts the average frequency for his company. Alternatively, ξ can be estimated using method of moments as presented in Lambrigger, Shevchenko and Wüthrich (2007).

Denote $\Lambda_{1:T} = (\Lambda_1, \dots, \Lambda_T)$, $N_{1:T} = (N_1, \dots, N_T)$ and $\Delta_{1:K} = (\Delta_1, \dots, \Delta_K)$. Then the joint posterior density of the random vector $(\Theta_\Lambda, \Lambda_{1:T})$ given observations $N_1 = n_1, \dots, N_T = n_T$, $\Delta_1 = \delta_1, \dots, \Delta_K = \delta_K$ is by Bayes' Theorem

$$\pi(\theta_\Lambda, \lambda_{1:T} | n_{1:T}, \delta_{1:K}) \propto \pi(n_{1:T} | \theta_\Lambda, \lambda_{1:T}) \pi(\lambda_{1:T} | \theta_\Lambda) \pi(\delta_{1:K} | \theta_\Lambda) \pi(\theta_\Lambda). \tag{11.5.6}$$

Here, the likelihood terms and the prior are made explicit,

$$\pi(n_{1:T} | \theta_\Lambda, \lambda_{1:T}) \pi(\lambda_{1:T} | \theta_\Lambda) = \prod_{t=1}^T \frac{(V\lambda_t)^{n_t}}{n_t!} \frac{(\alpha/\theta_\Lambda)^\alpha}{\Gamma(\alpha)} \lambda_t^{\alpha-1} \exp\{-\lambda_t(V + \alpha/\theta_\Lambda)\}, \tag{11.5.7}$$

$$\pi(\delta_{1:K} | \theta_\Lambda) = \prod_{k=1}^K \frac{(\xi/\theta_\Lambda)^\xi}{\Gamma(\xi)} \delta_k^{\xi-1} \exp\{-\delta_k \xi/\theta_\Lambda\}, \tag{11.5.8}$$

$$\pi(\theta_\Lambda) = \frac{b^a}{\Gamma(a)} \theta_\Lambda^{a-1} \exp\{-\theta_\Lambda b\}. \tag{11.5.9}$$

Note that the intensities $\Lambda_1, \dots, \Lambda_T$ are non-observable. Therefore we take the integral over their densities to obtain the posterior distribution of the random variable Θ_Λ given $(N_{1:T}, \Delta_{1:K})$

$$\begin{aligned} \pi(\theta_\Lambda | n_{1:T}, \delta_{1:K}) &\propto \prod_{t=1}^T \binom{\alpha + n_t - 1}{n_t} \left(\frac{\alpha}{\alpha + \theta_\Lambda V}\right)^\alpha \left(\frac{\theta_\Lambda V}{\alpha + \theta_\Lambda V}\right)^{n_t} \\ &\quad \times \prod_{k=1}^K \frac{(\xi/\theta_\Lambda)^\xi}{\Gamma(\xi)} \delta_k^{\xi-1} \exp\{-\delta_k \xi/\theta_\Lambda\} \frac{b^a}{\Gamma(a)} \theta_\Lambda^{a-1} \exp\{-\theta_\Lambda b\} \\ &\propto \left(\frac{1}{\alpha + \theta_\Lambda V}\right)^{T\alpha + \sum_{t=1}^T n_t} \theta_\Lambda^{a-K\xi + \sum_{t=1}^T n_t - 1} \exp\left\{-\theta_\Lambda b - \frac{\xi}{\theta_\Lambda} \sum_{k=1}^K \delta_k\right\}. \end{aligned} \tag{11.5.10}$$

Given Θ_Λ , the distribution of the number of losses N_t is negative binomial. Hence, one could start with a negative binomial model for N_t . The reason for the introduction of the random intensities Λ_t is that we will utilize them to model dependence between different risk cells, by introducing dependence between $\Lambda_t^{(1)}, \dots, \Lambda_t^{(J)}$.

Typically, a closed form expression for the marginal posterior function of Θ_Λ , given $(N_{1:T}, \Delta_{1:K})$ can not be obtained, except in this single risk cell setting. In general, we will integrate out the latent variables $\Lambda_1, \dots, \Lambda_T$ numerically through a MCMC approach to obtain an empirical distribution for the posterior of $\pi(\theta_\Lambda | n_{1:T}, \delta_{1:K})$. This empirical posterior distribution then allows for the simulation of Λ_{T+1} and N_{T+1} , respectively, conditional on the observations $(N_{1:T}, \Delta_{1:K})$.

11.5.2 Modelling frequencies for multiple risk cells

As in the previous section we will illustrate our methodology by presenting the frequency model construction. In this section we will extend the single risk cell frequency model to the general multiple risk cell setting. This will involve formulation of the multivariate posterior distribution.

Model Assumptions 11.5.3 (multiple risk cell frequency model). *Consider J risk cells. Assume that every risk cell j has a fixed, deterministic volume $V^{(j)}$.*

1. The risk characteristic $\Theta_\Lambda = (\Theta_\Lambda^{(1)}, \dots, \Theta_\Lambda^{(J)})$ has a J -dimensional prior density $\pi(\theta_\Lambda)$. The copula parameters θ_ρ are modelled by a random vector Θ_ρ with the prior density $\pi(\theta_\rho)$; Θ_Λ and Θ_ρ are independent.
2. Given $\Theta_\Lambda = \theta_\Lambda$ and $\Theta_\rho = \theta_\rho$: $(\Lambda_1, \mathbf{N}_1), \dots, (\Lambda_{T+1}, \mathbf{N}_{T+1})$ are i.i.d. and the intensities $\Lambda_t = (\Lambda_t^{(1)}, \dots, \Lambda_t^{(J)})$ have a J -dimensional conditional density with marginal distributions $\Lambda_t^{(j)} \sim G(\cdot | \theta_\Lambda^{(j)}) = \Gamma(\alpha^{(j)}, \alpha^{(j)} / \theta_\Lambda^{(j)})$ and the copula $c(\cdot | \theta_\rho)$. Thus the joint density of Λ_t is given by

$$\pi(\lambda_t | \theta_\Lambda, \theta_\rho) = c\left(G(\lambda_t^{(1)} | \theta_\Lambda^{(1)}), \dots, G(\lambda_t^{(J)} | \theta_\Lambda^{(J)}) | \theta_\rho\right) \prod_{j=1}^J \pi(\lambda_t^{(j)} | \theta_\Lambda^{(j)}), \quad (11.5.11)$$

where $\pi(\cdot | \theta_\Lambda^{(j)})$ denotes the marginal density.

3. Given $\Theta_\Lambda = \theta_\Lambda$ and $\Lambda_t = \lambda_t$, the number of claims are independent with $N_t^{(j)} \sim \text{Poi}(V^{(j)} \lambda_t^{(j)})$, $j = 1, \dots, J$.
4. There are expert opinions $\Delta_k = (\Delta_k^{(1)}, \dots, \Delta_k^{(J)})$, $k = 1, \dots, K$. Given $\Theta_\Lambda = \theta_\Lambda$: Δ_k and $(\Lambda_t, \mathbf{N}_t)$ are independent for all k and t ; and $\Delta_k^{(j)}$ are all independent with $\Delta_k^{(j)} \sim \Gamma(\xi^{(j)}, \xi^{(j)} / \theta_\Lambda^{(j)})$.

For convenience of notation, define:

- $\Lambda_{1:T} = \left[\left(\Lambda_1^{(1)}, \dots, \Lambda_1^{(J)} \right), \left(\Lambda_2^{(1)}, \dots, \Lambda_2^{(J)} \right), \dots, \left(\Lambda_T^{(1)}, \dots, \Lambda_T^{(J)} \right) \right]$ - Frequency intensities for all risk profiles and years;
- $N_{1:T} = \left[\left(N_1^{(1)}, \dots, N_1^{(J)} \right), \left(N_2^{(1)}, \dots, N_2^{(J)} \right), \dots, \left(N_T^{(1)}, \dots, N_T^{(J)} \right) \right]$ - Annual number of losses for all risk profiles and years;

- $\Delta_{1:K} = \left[\left(\Delta_1^{(1)}, \dots, \Delta_1^{(J)} \right), \left(\Delta_2^{(1)}, \dots, \Delta_2^{(J)} \right), \dots, \left(\Delta_K^{(1)}, \dots, \Delta_K^{(J)} \right) \right]$ - Expert opinions on mean frequency intensities for all experts and risk profiles.

Prior Structure $\pi(\theta_\Lambda)$ and $\pi(\theta_\rho)$. In the following examples, *a priori*, the risk characteristics $\Theta_\Lambda^{(j)}$ are independent Gamma distributed: $\Theta_\Lambda^{(j)} \sim \Gamma(a^{(j)}, b^{(j)})$ with hyper-parameters $a^{(j)} > 0$ and $b^{(j)} > 0$. This means that *a priori* the risk characteristics for the different risk classes are independent. That is, if the company has a bad risk profile in risk class j then the risk profile in risk class i need not necessarily also be bad. Dependence is then modelled through the dependence between the intensities $\Lambda_t^{(1)}, \dots, \Lambda_t^{(J)}$. If this is not appropriate then, of course, this can easily be changed by assuming dependence within Θ_Λ .

In the simulation experiments below we consider cases when the copula is parameterized by a scalar θ_ρ . Additionally, we are interested in obtaining inferences on θ_ρ implied by the data only so we use uninformative constant prior on the ranges $[-1,1]$, $(0,30]$ and $[1,30]$ in the case of Gaussian, Clayton and Gumbel copulas respectively.

Posterior density. The marginal posterior density of random vector $(\Theta_\Lambda, \Theta_\rho)$ given data of counts $N_1 = n_1, \dots, N_T = n_T$ and expert opinions $\Delta_1 = \delta_1, \dots, \Delta_K = \delta_K$ is

$$\begin{aligned} \pi(\theta_\Lambda, \theta_\rho | n_{1:T}, \delta_{1:K}) &= \prod_{t=1}^T \int \pi(\theta_\Lambda, \theta_\rho, \lambda_t | n_{1:T}, \delta_{1:K}) d\lambda_t \\ &\propto \prod_{t=1}^T \left(\int \prod_{j=1}^J \exp\{-V^{(j)} \lambda_t^{(j)}\} \frac{(V^{(j)} \lambda_t^{(j)})^{n_t^{(j)}}}{n_t^{(j)!}} \pi(\lambda_t | \theta_\Lambda, \theta_\rho) d\lambda_t \right) \\ &\quad \times \prod_{k=1}^K \prod_{j=1}^J \left(\frac{(\xi^{(j)} / \theta_\Lambda^{(j)})^{\xi^{(j)}}}{\Gamma(\xi^{(j)})} (\delta_k^{(j)})^{\xi^{(j)}-1} \exp\{-\delta_k^{(j)} \xi^{(j)} / \theta_\Lambda^{(j)}\} \right) \\ &\quad \times \prod_{j=1}^J \frac{(b^{(j)})^{a^{(j)}}}{\Gamma(a^{(j)})} (\theta_\Lambda^{(j)})^{a^{(j)}-1} \exp\{-b^{(j)} \theta_\Lambda^{(j)}\} \pi(\theta_\rho). \end{aligned} \tag{11.5.12}$$

11.6 Simulation Methodology - Slice sampler

Posterior (11.5.12) involves integration and sampling from this distribution is difficult. Here we present a specialized MCMC simulation methodology known as a Slice sampler to sample from the desired target posterior distribution $\pi(\theta_\Lambda, \theta_\rho, \lambda_{1:T} | n_{1:T}, \delta_{1:K})$. Marginally taken samples of Θ_Λ and Θ_ρ are samples from $\pi(\theta_\Lambda, \theta_\rho | n_{1:T}, \delta_{1:K})$ which can be used to make inference for required quantities.

It will be convenient to define the exclusion operators, $\Lambda_{1:T}^{(-i,-j)}$, $\Lambda_{1:T \setminus k}$ and $\Theta_\Lambda^{(-j)}$. For example:

- $\Lambda_{1:T}^{(-2,-1)} = \left[\left(\Lambda_1^{(1)}, \dots, \Lambda_1^{(J)} \right), \left(\Lambda_2^{(2)}, \dots, \Lambda_2^{(J)} \right), \dots, \left(\Lambda_T^{(1)}, \dots, \Lambda_T^{(J)} \right) \right]$ - Frequency intensities for all risk profiles and years, excluding risk profile 1 from year 2;
- $\Lambda_{1:T \setminus 2} = \left[\left(\Lambda_1^{(1)}, \dots, \Lambda_1^{(J)} \right), \left(\Lambda_3^{(1)}, \dots, \Lambda_3^{(J)} \right), \dots, \left(\Lambda_T^{(1)}, \dots, \Lambda_T^{(J)} \right) \right]$ - Frequency intensities

for all risk profiles and years, excluding all profiles for year 2;

- $\Theta_{\Lambda}^{(-j)} = \left[\Theta_{\Lambda}^{(1)}, \dots, \Theta_{\Lambda}^{(j-1)}, \Theta_{\Lambda}^{(j+1)}, \dots, \Theta_{\Lambda}^{(J)} \right]$.

Sampling from $\pi(\theta_{\Lambda}, \theta_{\rho}, \lambda_{1:T} | \mathbf{n}_{1:T}, \delta_{1:K})$ or $\pi(\theta_{\Lambda}, \theta_{\rho} | \mathbf{n}_{1:T}, \delta_{1:K})$ via closed form inversion sampling or via rejection sampling is not typically an option. There are many reasons for this. Firstly, only for specific copula models will closed form tractable expression for the marginal $\pi(\theta_{\Lambda}, \theta_{\rho} | \mathbf{n}_{1:T}, \delta_{1:K})$ be attainable, certainly not for the models we consider in this paper. Secondly, even for the expression of the joint posterior $\pi(\theta_{\Lambda}, \theta_{\rho}, \lambda_{1:T} | \mathbf{n}_{1:T}, \delta_{1:K})$ it is typically only possible to sample from the conditional distributions sequentially via numerical inversion sampling techniques which is highly computational and inefficient in high dimensions. Additionally, we would like a technique which is independent of the potentially arbitrary choice in specifying a copula function for the dependence between $\Lambda_t^{(1)}, \dots, \Lambda_t^{(J)}$. Hence, we utilize an MCMC framework which we make general enough to work for any choice of copula model, developed next.

11.7 Bayesian Parameter Estimation

We separate the analysis into two parts. Firstly, we condition on knowledge of the copula parameters $\theta_{\rho} = \rho$, where ρ denotes the true copula parameters used to generate the data. This allows us to demonstrate that if the copula parameters θ_{ρ} are known, we can perform estimation of other parameters accurately under joint inference. The second part involves joint estimation of θ_{ρ} and θ_{Λ} to demonstrate the accuracy of the joint inference procedure developed. Note that, the model for this second part has not been formally introduced but is a simple extension of Model Assumptions 11.5.3.

11.7.1 Conditional on a priori knowledge of copula parameter

Here we assume the copula parameter has been estimated *a priori* and so estimation only involves model parameters. Such a setting may arise for example if the copula parameter is already estimated via a ML estimator. The proposed sampling procedure we develop is a particular class of algorithms in the toolbox of MCMC methods. It is an alternative to a Gibbs sampler known as a univariate Slice sampler. We note that to implement a Gibbs sampler or a univariate Slice sampler one needs to know the form of the full conditional distributions. However, unlike the basic Gibbs sampler the Slice sampler does not require sampling from these full conditional distributions. Derivations of the posterior full conditionals,

$$\begin{aligned} \pi(\theta_{\Lambda}^{(j)} | \theta_{\Lambda}^{(-j)}, \lambda_{1:T}, \mathbf{n}_{1:T}, \delta_{1:K}, \theta_{\rho}) &\propto \pi(\lambda_{1:T} | \theta_{\Lambda}^{(-j)}, \theta_{\Lambda}^{(j)}, \theta_{\rho}) \pi(\delta_{1:K} | \theta_{\Lambda}^{(-j)}, \theta_{\Lambda}^{(j)}) \\ &\quad \times \pi(\theta_{\Lambda}^{(-j)} | \theta_{\Lambda}^{(j)}) \pi(\theta_{\Lambda}^{(j)}), \end{aligned} \quad (11.7.1)$$

$$\pi(\lambda_t^{(j)} | \theta_{\Lambda}, \lambda_{1:T}^{(-t, -j)}, \mathbf{n}_{1:T}, \delta_{1:K}, \theta_{\rho}) \propto \pi(\mathbf{n}_{1:T} | \lambda_{1:T}^{(-t, -j)}, \lambda_t^{(j)}) \pi(\lambda_t^{(-j)}, \lambda_t^{(j)} | \theta_{\Lambda}, \theta_{\rho}) \quad (11.7.2)$$

are presented in Appendix 11.12.

11.7.2 Joint inference of marginal and copula parameters.

To include the estimation of the copula parameter θ_ρ jointly with the parameters Θ_Λ and latent intensities $\Lambda_{1:T}$ in our Bayesian framework, we assume that it is constant in time and model it by a random variable Θ_ρ with some prior density $\pi(\theta_\rho)$. The full conditional posterior of the copula parameter, denoted $\pi(\theta_\rho|\theta_\Lambda, \lambda_{1:T}, \mathbf{n}_{1:T}, \delta_{1:K})$, is given by

$$\pi(\theta_\rho|\theta_\Lambda, \lambda_{1:T}, \mathbf{n}_{1:T}, \delta_{1:K}) \propto \pi(\lambda_{1:T}|\theta_\Lambda, \theta_\rho)\pi(\theta_\rho), \quad (11.7.3)$$

For the full derivation of the scalar case see, Appendix 11.12.

In the following section we provide intuition for our choice of univariate Slice sampler as compared to alternative Markov chain algorithms. In particular we describe the advantages that the Slice sampler has compared to more standard Markov chain samplers, though we also point out the additional complexity involved. We verify the validity of the Slice sampling algorithm for those not familiar with this specialized algorithm and we then describe some intricacies associated with implementation of the algorithm. This is followed by a discussion of some extensions we developed when analyzing the OpRisk model. The technical details of the actual algorithm are provided in Appendix 11.13.

11.8 Slice sampling

The full conditionals given in equations (11.7.1), (11.7.2), (11.7.3) do not take standard explicit closed forms and typically the normalizing constants are not known in closed form. Therefore this will exclude straightforward inversion or basic rejection sampling being used to sample from these distributions. Therefore one may adopt a Metropolis Hastings (MH) within Gibbs sampler to obtain samples, see for example Gilks, Richardson and Spiegelhalter (1996) and Robert and Casella (2004) for detailed expositions of such approaches. To utilize such algorithms it is important to select a suitable proposal distribution. Quite often in high dimensional problems such as ours, this requires tuning of the proposal for a given target distribution. Hence, one incurs a significant additional computational expense in tuning the proposal distribution parameters off-line so that mixing of the resulting Markov chain is sufficient. An alternative not discussed here would include an Adaptive Metropolis Hastings within Gibbs sampling algorithm, see Atchade and Rosenthal (2005) and Rosenthal (2008). Here we take a different approach which utilizes the full conditional distributions, known as a univariate Slice sampler, see Neal (2003). We demonstrate how effective a univariate Slice sampler is for our model.

Slice sampling was developed with the intention of providing a "black box" approach for sampling from a target distribution which may not have a simple form. The Slice sampling method-

ology we develop will be automatically tailored to the desired target posterior. As such it does not require pre-tuning and in many cases will be more efficient than a MH within Gibbs sampler. The reason for this, pointed out by Neal (2003), is that a MH within Gibbs has two potential problems. The first arises when a MH approach attempts moves which are not well adapted to local properties of the density, resulting in slow mixing of the Markov chain. Secondly, the small moves arising from the slow mixing typically lead to traversal of a region of posterior support in the form of a Random Walk. Therefore, L^2 steps are required to traverse a distance that could be traversed in only L steps if moving consistently in the same direction. A univariate Slice sampler can adaptively change the scale of the moves proposed avoiding problems that can arise with the MH sampler when the appropriate scale of proposed moves varies over the support of the distribution.

A single iteration of the Slice sampling distribution for a toy example is presented in Figure 11.13.2. The intuition behind Slice sampling arises from the fact that sampling from a univariate distribution $p(\theta)$ can always be achieved by sampling uniformly from the region under the distribution $p(\theta)$. Obtaining a Slice sample follows two steps: sample a value $u_l \sim U[0, p(\theta_{l-1})]$ and then sample a value uniformly from A_l , $\theta_l \sim U[A_l]$. This procedure is repeated and by discarding the auxiliary variable sample u_l one obtains correlated samples θ'_l s from $p(\theta_{l-1})$. Neal (2003), demonstrates that a Markov chain (U, Θ) constructed in this way will have stationary distribution defined by a uniform distribution under $p(\theta)$ and the marginal of Θ has desired stationary distribution $p(\theta)$. Additionally, Mira and Tierney (2002) proved that the Slice sampler algorithm, assuming a bounded target distribution $p(\theta)$ with bounded support, is uniformly ergodic.

Similar to a deterministic scan Gibbs sampler, the simplest way to extend the Slice sampler to a multivariate distribution is by considering each full conditional distribution in turn. Note, discussion relating to the benefits provided by Random Walk behavior suppression, as achieved by the Slice sampler, are presented in the context of non-reversible Markov chains, see Diaconis, Holmes and Neal (2000).

Additionally, we only need to know the target full conditional posterior up to normalization, see Neal (2003) p. 710. This is important in this example since solving the normalizing constant in this model is not possible analytically. To make more precise the intuitive description of the Slice sampler presented above, we briefly detail the argument made by Neal on this point. Suppose we wish to sample from a distribution for a random vector $\Theta \in \mathbb{R}^n$ whose density $p(\theta)$ is proportional to some function $f(\theta)$. This can be achieved by sampling uniformly from the $(n+1)$ -dimensional region that lies under the plot of $f(\theta)$. This is formalized by introducing the auxiliary random variable U and defining a joint distribution over Θ and U which is uniform over the region $\{(\Theta, U) : 0 < u < f(\theta)\}$ below the surface defined by $f(\theta)$, given by

$$p(\theta, u) = \begin{cases} 1/Z, & \text{if } 0 < u < f(\theta), \\ 0, & \text{otherwise,} \end{cases}$$

where $Z = \int f(\boldsymbol{\theta}) d\boldsymbol{\theta}$. Then the target marginal density for Θ is given by

$$p(\boldsymbol{\theta}) = \int_0^{f(\boldsymbol{\theta})} \frac{1}{Z} du = \frac{f(\boldsymbol{\theta})}{Z},$$

as required. There are many possible procedures to obtain samples of (Θ, U) . The details of the implemented algorithm undertaken in this paper are provided in Appendix 11.13.

11.8.1 Extensions

We note that in the Bayesian model we develop, in some cases strong correlation between the parameters of the model will be present in the posterior, see Figure 11.13.3. In more extreme cases, this can cause slow rates of convergence of a univariate sampler to reach the ergodic regime, translating into longer Markov Chain simulations. In such a situation several approaches can be tried to overcome this problem. The first involves the use of a mixture transition kernel combining local and global moves. For example, we suggest local moves via a univariate Slice sampler and global moves via an Independent Metropolis Hastings (IMH) sampler with adaptive learning of its covariance structure, such an approach is known as a hybrid sampler, see comparisons in Brewer, Aitken and Talbot (1994). Alternatively, for the global move if determination of level sets in multiple dimensions is not problematic, for the model under consideration, then some of the multivariate Slice sampler approaches designed to account for correlation between parameters can be incorporated, see Neal (2003) for details. This is beyond the scope of this paper.

Another approach to break correlation between parameters in the posterior is via transformation of the parameter space. If the transformation is effective this will reduce correlation between parameters of the transformed target posterior. Sampling can then proceed in the transformed space, and then samples can be transformed back to the original space. It is not always straightforward to find such transformations.

A third alternative is based on Simulated Tempering, introduced by Marinari and Parisi (1992) and discussed extensively in Geyer and Thompson (1995). In particular a special version of Simulated Tempering, first introduced by Neal (1996) can be utilized in which one considers a sequence of target distributions $\{\pi_l\}$ constructed such that they correspond to the objective posterior in the following way,

$$\pi_l = [\pi(\boldsymbol{\theta}_\Lambda, \boldsymbol{\lambda}_{1:T}, \theta_\rho | \mathbf{n}_{1:T}, \boldsymbol{\delta}_{1:K})]^\gamma$$

with sequence $\{\gamma_l\}$. Then one uses the Slice sampling algorithm presented and replaces π with π_l .

Running a Markov chain such that at each iteration l we target posterior π_l and then only keeping samples from the Markov chain corresponding to situations in which $\gamma_l = 1$ can result in a significant improvement in exploration around the posterior support. This can overcome

slow mixing arising from a univariate sampling regime. The intuition for this is that for values of $\gamma_l \ll 1$ the target posterior is almost uniform over the space, resulting in large moves being possible around the support of the posterior, then as γ_l returns to a value of 1, several iterations later, it will be in potentially new unexplored regions of posterior support.

As an extension we developed a Simulated Tempering Slice sampler to obtain samples from the posterior $p(\theta_\Lambda, \lambda_{1:T}, \theta_\rho | \mathbf{n}_{1:T}, \delta_{1:K})$. In our development we utilize a sine function, $\gamma_l = \min(\sin(\frac{2\pi}{1000}l) + 1, 1)$, for γ_l which has its amplitude truncated to ensure it ranges between $(0, 1]$. That is the function is truncated at $\gamma_l = 1$ for extended iteration periods for our simulation index l to ensure the sampler spends significant time sampling from the actual posterior distribution.

We note that application of the Tempering proved useful and improved mixing of the Markov chain. However, for simulation examples presented in the remainder of this paper it was sufficient to use the basic univariate Slice sampler presented previously, which is more computationally efficient than the Tempered version.

Note, in the application of Tempering one must discard many simulated states of the Markov chain, whenever $\gamma_l \neq 1$. There is however a computational way to avoid discarding these samples, see Gramacy, Samworth and King (2007).

Finally, we note that there are several alternatives to a MH within Gibbs sampler such as a basic Gibbs sampler combined with Adaptive Rejection sampling (ARS), Gilks and Wild (1992). Note ARS requires distributions to be log concave. Alternatively an adaptive version of this known as the Adaptive Metropolis Rejection sampler could be used, see Gilks, Best and Tan (1995).

11.9 Results

In this section we demonstrate and compare the performance of our sampling methodology on several different copula models. We intend to demonstrate the appropriate behavior of our Bayesian models as a function of the number of annual years, in the presence of highly biased expert opinions. This will be achieved through simulation studies using the sampling techniques detailed above to perform inference on model parameters. The intention will be to demonstrate the appropriate convergence and accuracy as a function of data sample size. Hereafter, we study the case of dependence between intensities of two risks and set risk cell volumes $V^{(1)} = V^{(2)} = 1$.

11.9.1 Estimation of model if copula parameter is known.

Here, we study the estimation of model parameters in two cases. The first case involves two low frequency risks. In the second case, one risk has low frequency while another risk has high frequency. In these two cases we present results for the univariate Slice sampler under scenarios involving: data generated independently for each risk profile and data generated

using a Gaussian, Clayton and Gumbel copulas.

Only one expert opinion is assumed for each risk. We present the parameter estimates as a function of data size for each of the specified correlation levels. That is, we study the accuracy of the parameter estimates as the number of observations increases. Simulation results are obtained by creating independently 20 data sets each of length 20 years, then for each data set simulations are performed for subsets of the data going for 1, 2, 5, 10, 15 and 20 years. We then average the performance of posterior estimates over these independent simulations. The Markov chains are run for 50,000 iterations with 10,000 iterations discarded as burnin. The simulation time depends on the number of risk profiles, the number of observations and expert opinions and the length of the Markov chain¹. In performing the analysis we studied three cases and in each case we performed the following steps,

1. Simulate a data set of appropriate number of years according to the procedure specified in Appendix 11.11.
 2. Obtain correlated MCMC samples from the target posterior distribution after discarding burnin samples, $\{\theta_{\Lambda,l}, \lambda_{1:T,l}\}, l = 1001, \dots, 50000$.
 3. Estimate desired posterior quantities such as posterior mean of parameters of interest and posterior standard deviations.
 4. Repeat stages 1 - 3 for 20 independent data realizations and then average the results.
- **Joint:** The results are obtained by MCMC samples taken from $\pi(\theta_{\Lambda}, \lambda_{1:T} | \mathbf{n}_{1:T}, \delta_{1:K}, \theta_{\rho})$ with the correct copula model and copula parameter used in the sampler. This is the procedure that should be performed in a real application.
 - **Marginal:** Results are obtained by MCMC samples taken from

$$\begin{aligned} \pi(\theta_{\Lambda}, \lambda_{1:T} | \mathbf{n}_{1:T}, \delta_{1:K}, \theta_{\rho}) &= \prod_{j=1}^J \pi(\theta_{\Lambda}^{(j)}, \lambda_{1:T}^{(j)} | \mathbf{n}_{1:T}, \delta_{1:K}, \theta_{\rho}) \\ &= \prod_{j=1}^J \pi(\theta_{\Lambda}^{(j)} | \mathbf{n}_{1:T}, \delta_{1:K}, \theta_{\rho}) \end{aligned}$$

which is the posterior in the case of independence. This is equivalent to marginal estimation where single risk cell data is analyzed separately, see Section 4.1.

- **Benchmark:** To verify the results we also consider the case where we assume perfect knowledge of the realized random process for random vector $\Lambda_{1:T}$. We then perform inference on Θ_{Λ} without the additional uncertainty arising from estimating $\Lambda_{1:T}$. In this regard this represents a benchmark for which we may compare the performance of our

¹ A typical run with 5 years of data and 1 expert in the bivariate case for 50,000 simulations took approximately 50sec and approximately 43min for the case of ten risk profiles when coded in Fortran and run on 2.40GHz Intel Core2.

simulations. In particular, it is obtained by samples taken from $\pi(\boldsymbol{\theta}_\Lambda | \boldsymbol{\lambda}_{1:T}, \mathbf{n}_{1:T}, \boldsymbol{\delta}_{1:K}, \theta_\rho)$ conditional on the true simulated realizations of random variables $\Lambda_{1:T}$.

Example 1: low frequency risk profiles. We set the true parameter values of $\Theta_\Lambda^{(1)}$ and $\Theta_\Lambda^{(2)}$ to be $\theta_{true}^{(1)} = 5$ and $\theta_{true}^{(2)} = 5$ respectively. Also we choose the expert's opinion on the true parameters to be an underestimate in risk profile 1 with $\Delta_1^{(1)} = 2$ and an overestimate for risk profile 2 with $\Delta_1^{(2)} = 8$. The model parameters were set to $\xi^{(1)} = \xi^{(2)} = 2$, $\alpha^{(1)} = \alpha^{(2)} = 2$, and prior distribution parameters $a^{(1)} = a^{(2)} = 2$, $b^{(1)} = b^{(2)} = 2.5$. The results for this simulation study, presented in Tables 11.1 and 11.2, show the appropriate convergence of the estimates of parameters $\Theta_\Lambda^{(1)}$ and $\Theta_\Lambda^{(2)}$ as a function of the data size, demonstrating how well this simulation procedure works under these models. In addition we note that as expected from credibility theory we observe that joint estimation is better than the marginal, i.e. the posterior standard deviations for $\Theta_\Lambda^{(1)}$ and $\Theta_\Lambda^{(2)}$ are less when joint estimation is used. In addition the rate of convergence of the posterior mean for Θ_Λ to the true value is faster under the joint estimation. Note, the standard errors in the posterior mean and standard deviation were calculated and found to be strictly in the range of 1-5% for the simulations presented.

In Figure 11.13.3, corresponding to Gaussian, Clayton and Gumbel copula models respectively, we demonstrate the estimated density $\pi(\boldsymbol{\theta}_\Lambda | \boldsymbol{\lambda}_{1:T}, \mathbf{n}_{1:T}, \boldsymbol{\delta}_{1:K}, \theta_\rho)$ if we had perfect knowledge of the latent process parameters $\Lambda_{1:T}$. In this way we compare the exact posterior with perfect knowledge of the correlation structure as captured by the copula model which here we assume is known. Obtaining these plots involved a particular realized data set of length 20 years. For all copulas two values of θ_ρ were considered: $\theta_\rho = 0.9$ and $\theta_\rho = 0.1$ for Gaussian copula; $\theta_\rho = 10$ and $\theta_\rho = 1$ for Clayton copula; and $\theta_\rho = 3$ and $\theta_\rho = 1.1$ for the Gumbel copula. These plots of the joint marginal posterior distribution of Θ_Λ demonstrate clearly that the standard practice in the industry of performing marginal estimation of risk profiles will lead to incorrect results when estimating quantities based on the distribution of Θ_Λ .

Example 2: one low frequency and one high frequency risk profile. We set the true values of $\Theta_\Lambda^{(1)}$ and $\Theta_\Lambda^{(2)}$ to be $\theta_{true}^{(1)} = 5$ and $\theta_{true}^{(2)} = 10$ respectively. Also we choose the expert's opinion on the true parameters to be an under estimate in risk profile 1 with $\Delta_1^{(1)} = 2$ and an over estimate for risk profile 2 with $\Delta_1^{(2)} = 13$. The model parameters were set to $\xi^{(1)} = \xi^{(2)} = 2$, $\alpha^{(1)} = 2$, $\alpha^{(2)} = 2$, $a^{(1)} = a^{(2)} = 2$, $b^{(1)} = 2.5$, $b^{(2)} = 5$. The simulation results and comparisons are developed in the same approach as Example 1 and again the standard errors in the estimates were in the range 1-5%. The results can be found in Tables 11.3 and 11.4.

11.9.2 Joint estimation of marginal and copula parameters

Here we estimate $\Theta_\Lambda^{(1)}$, $\Theta_\Lambda^{(2)}$ and Θ_ρ jointly. For this example, the model settings from Example 1 were used and one data set of length 20 years randomly generated was utilized. The simulation was performed by taking 150,000 iterations of the sampler and discarding the first 20,000 as burnin. Results for these simulations are contained in Table 11.5.

These results demonstrate that our model and estimation methodology is successfully able to estimate jointly the risk profiles and the correlation parameter. This is seen to be the case for all the models we consider in this paper. It is also clear that with few observations, e.g. $T \leq 5$, and a vague prior for the copula parameter, it will be difficult to accurately estimate the copula parameter. This is largely due to the fact that the posterior distribution in this case is diffuse. Additionally, with a small amount of data it appears that accurately estimating the copula parameter is most difficult in the Gumbel model. However, as the number of observations increases the accuracy of the estimate improves in all models and the estimates are reasonable in the case of 15 or 20 years of data. Additionally, we could further improve the accuracy of this prediction if we incorporated expert opinions into the prior specification of the copula parameter, instead of using a vague prior.

Overall, we have demonstrated that combination of all the relevant sources of data can be achieved under our model. Then with this study we show that our sampling methodology has the ability to estimate jointly all the model parameters including the copula parameter. This is a key step forward in model development and estimation for OpRisk models. We further envisage that one can extend this methodology to more sophisticated and flexible copula based models with more than one parameter. This should be relatively trivial since the methodology we developed applies directly. However, the challenge in the case of a more sophisticated copula model relates to finding a relevant choice of prior distribution on the correlation structure.

Full predictive distribution. As a final comment in this section we point out an important additional outcome of obtaining samples from the joint posterior distribution of the model parameters and the correlation. This relates to construction of the full predictive annual loss distribution, accounting for parameter uncertainty.

Typically practitioners will take point estimates of all parameters and then condition on these point estimates to empirically construct the predictive distribution and then calculate risk measures to be reported such as VaR. Here we comment that a more robust approach to prediction can now be performed. Using our methodology, we can construct the full predictive distribution after removing the parameter uncertainty from the model considered, including the uncertainty arising from the correlation parameter. To achieve this we would consider the full predictive distribution:

$$\pi(Z_{T+1}|n_{1:T}, \delta_{1:T}) = \int \pi(Z_{T+1}|\boldsymbol{\theta}_\Lambda, \theta_\rho) \pi(\boldsymbol{\theta}_\Lambda, \theta_\rho|n_{1:T}, \delta_{1:T}) d\boldsymbol{\theta}_\Lambda d\theta_\rho. \quad (11.9.1)$$

Here, we used the model assumptions that given $\boldsymbol{\Theta}_\Lambda$ and Θ_ρ we have that Z_{T+1} is independent from the observations $(N_{1:T}, \Delta_{1:K})$. In practice to obtain samples from this full predictive distribution involves taking the steps demonstrated in Appendix a minor modification. If one wanted to simulate L annual losses from the full predictive distribution, this would involve first running the Slice sampler for L iterations after burnin. Then for each iteration l one would use the state of the Markov chain $(\boldsymbol{\theta}_{\Lambda,l}, \theta_{\rho,l})$ in the simulation procedure detailed in Appendix

11.11. We also note that it is trivial under our methodology to extend this full predictive distribution sampling to the case of frequency and severities.

11.10 Discussion

This paper introduced a dynamic OpRisk model which allows for significant flexibility in correlation structures introduced between risk profiles. Next a Bayesian framework was established to allow inference and estimation under this model to be performed, whilst at the same time allowing incorporation of alternative data sources into the inferential procedure. Then a novel simulation procedure was developed for the Bayesian model presented, in the case of dependence between frequency risk profiles. Simulations were performed to demonstrate the accuracy of this procedure in multiple bivariate examples. Comparisons were made between marginal estimation and a benchmark estimation procedure. In all simulations, the estimation of the model parameters was accurate and behavior of the estimates of posterior mean and standard deviation presented, smoothed over multiple data realizations, was as expected. Initially the influence of the biased expert opinion observation influenced the results and as the size of the data set for actual annual loss counts grew, the estimations improved in accuracy. Clearly, the joint estimation will outperform marginal estimation when forming predictions of future counts and rates in year $T + 1$, given estimates based on data up to year T . Additionally, we demonstrated highly accurate estimation of the copula parameter, jointly with the model parameters.

Additionally, simulations were performed in the models $J = 5$ and $J = 10$ for the Clayton copula model in which the copula parameter is also unknown. Though the simulation time was increased as a factor of the number of risk cells, the results and performance were as presented for the bivariate models, making this approach suitable for practical purposes.

Finally, the main objective of the paper is to preset the framework for the multivariate problem and to demonstrate estimation in this setting. Application of the framework to real data is the subject of further research. In this paper, the estimation procedure is presented for frequencies only but it can be extended in the same manner as presented to severities.

11.11 Appendix A: Simulation of annual losses

In general, given marginal and copula parameters $(\theta_\Lambda, \theta_\Psi, \theta_\rho)$, the simulation of the annual losses for year $t = T + 1$, when risk profiles are dependent, can be done as described below.

1. Simulate $2J$ -variate $u_1, \dots, u_J, v_1, \dots, v_J$ from a $2J$ dimensional copula $C(\cdot|\theta_\rho)$.
2. Calculate $\lambda_t^{(j)} = G^{-1}(u_j|\theta_\Lambda^{(j)})$ and $\psi_t^{(j)} = H^{-1}(v_j|\theta_\Psi^{(j)})$, $j = 1, \dots, J$
3. Sample $n_t^{(j)}$ from $P(\cdot|\lambda_t^{(j)})$, $j = 1, \dots, J$.
4. Sample iid $x_s^{(j)}(t)$, $s = 1, \dots, n_t^{(j)}$, $j = 1, \dots, J$ from $F(\cdot|\psi_t^{(j)})$.
5. Calculate annual losses $z_t^{(j)} = \sum_{s=1}^{n_t^{(j)}} x_s^{(j)}(t)$, $j = 1, \dots, J$.
6. Repeat Steps 1-5, K times to get K random samples of the annual losses $z_t^{(j)}$.

Note, to simulate from the full predictive distribution of annual losses, add simulation of $(\theta_\Lambda, \theta_\Psi, \theta_\rho)$ from the posterior distribution (e.g. using Slice sampler methodology) as an extra step before Step 1. Simulation of the random variates from a copula in Step 1 in the case of Gaussian, Clayton and Gumbel copulas can be done as described below.

Gaussian copula:

1. Simulate d -variate x_1, \dots, x_d from $\Phi_N(\mathbf{0}, \Sigma)$, where $\Phi_N(\mathbf{0}, \Sigma)$ is a Normal distribution with zero means, unit variances and correlation matrix Σ .
2. Calculate $u_1 = F_N(x_1), \dots, u_d = F_N(x_d)$. Obtained (u_1, \dots, u_d) is a d -variate from a Gaussian copula.

Archimedean copulas: The Clayton and Gumbel copulas are members of the Archimedean family of copulas. The d -dimensional Archimedean copulas can be written as

$$C(u_1, \dots, u_d|\rho) = \phi^{-1}(\phi(u_1) + \dots + \phi(u_d)) \quad (11.11.1)$$

with ϕ a decreasing function known as the generator for the given copula, see Frees and Valdez (1998). The generator and inverse generator for the Clayton (ϕ_C) and Gumbel (ϕ_G) copulas are given by

$$\begin{aligned} \phi_C(t) &= (t^{-\rho} - 1); & \phi_C^{-1}(s) &= (1 + s)^{-\frac{1}{\rho}}; \\ \phi_G(t) &= (-\ln t)^\rho; & \phi_G^{-1}(s) &= \exp\left(-s^{\frac{1}{\rho}}\right), \end{aligned} \quad (11.11.2)$$

where ρ is a copula parameter. Simulation from such a copula can be achieved following the algorithm provided in Melchiori (2006):

1. Sample d independent random variates v_1, \dots, v_d from a uniform distribution $U[0, 1]$.
2. Simulate y from $D(\cdot)$ such that Laplace transform of D satisfies $\mathcal{L}(D) = \phi^{-1}$ and $D(0) = 0$.
3. Find $s_i = -(\ln v_i) / y$ for $i = 1, \dots, d$
4. Calculate $u_i = \phi^{-1}(s_i)$ for $i = 1, \dots, d$.

The obtained (u_1, \dots, u_d) is a d -variate from d -dimensional Archimedean copula. What remains is to define the relevant distribution $D(\cdot)$ for the Clayton and Gumbel Copulas. For the Clayton copula, $D(\cdot)$ is a Gamma distribution with shape parameter given by ρ^{-1} and unit scale. For the Gumbel copula, $D(\cdot)$ is from the α -stable family $S_\alpha(\beta, \gamma, \delta)$ with the following parameters shape $\alpha = \rho^{-1}$, skewness $\beta = 1$, scale $\gamma = (\cos(\frac{1}{2}\pi/\rho))^\rho$ and location $\delta = 0$. In the Gumbel case, the density for D has no analytic form and the simulation from this distribution can be achieved using the algorithm from Nolan (2007) to efficiently generate the required samples from the univariate stable distribution.

11.12 Appendix B: Full conditional posterior distributions

Note, in Part 1 and Part 2 we are conditioning on the copula parameter θ_ρ , this notation is dropped for simplicity. It is only explicitly introduced in Part 3.

Part 1: Using Bayes' theorem

$$\begin{aligned} \pi\left(\theta_\Lambda^{(j)} | \theta_\Lambda^{(-j)}, \boldsymbol{\lambda}_{1:T}, \mathbf{n}_{1:T}, \boldsymbol{\delta}_{1:K}\right) &\propto \pi\left(\theta_\Lambda^{(-j)}, \boldsymbol{\lambda}_{1:T}, \mathbf{n}_{1:T}, \boldsymbol{\delta}_{1:K} | \theta_\Lambda^{(j)}\right) \pi\left(\theta_\Lambda^{(j)}\right) \\ &= \pi\left(\boldsymbol{\lambda}_{1:T}, \mathbf{n}_{1:T}, \boldsymbol{\delta}_{1:K} | \theta_\Lambda^{(-j)}, \theta_\Lambda^{(j)}\right) \pi\left(\theta_\Lambda^{(-j)} | \theta_\Lambda^{(j)}\right) \pi\left(\theta_\Lambda^{(j)}\right). \end{aligned} \quad (11.12.1)$$

Using the model structure to exploit conditional independence properties we note that

$$\begin{aligned} \pi\left(\boldsymbol{\lambda}_{1:T}, \mathbf{n}_{1:T}, \boldsymbol{\delta}_{1:K} | \theta_\Lambda^{(-j)}, \theta_\Lambda^{(j)}\right) &= \pi\left(\mathbf{n}_{1:T}, \boldsymbol{\lambda}_{1:T} | \theta_\Lambda^{(-j)}, \theta_\Lambda^{(j)}\right) \pi\left(\boldsymbol{\delta}_{1:K} | \theta_\Lambda^{(-j)}, \theta_\Lambda^{(j)}\right) \\ &= \pi\left(\mathbf{n}_{1:T} | \boldsymbol{\lambda}_{1:T}\right) \pi\left(\boldsymbol{\lambda}_{1:T} | \theta_\Lambda^{(-j)}, \theta_\Lambda^{(j)}\right) \pi\left(\boldsymbol{\delta}_{1:K} | \theta_\Lambda^{(-j)}, \theta_\Lambda^{(j)}\right) \end{aligned} \quad (11.12.2)$$

which specifies the full conditional distributions for the j^{th} component $\theta_\Lambda^{(j)}$ as

$$\begin{aligned} \pi\left(\theta_\Lambda^{(j)} | \theta_\Lambda^{(-j)}, \boldsymbol{\lambda}_{1:T}, \mathbf{n}_{1:T}, \boldsymbol{\delta}_{1:K}\right) &\propto \pi\left(\mathbf{n}_{1:T} | \boldsymbol{\lambda}_{1:T}\right) \\ &\quad \times \pi\left(\boldsymbol{\lambda}_{1:T} | \theta_\Lambda^{(-j)}, \theta_\Lambda^{(j)}\right) \pi\left(\boldsymbol{\delta}_{1:K} | \theta_\Lambda^{(-j)}, \theta_\Lambda^{(j)}\right) \pi\left(\theta_\Lambda^{(-j)} | \theta_\Lambda^{(j)}\right) \pi\left(\theta_\Lambda^{(j)}\right) \\ &\propto \pi\left(\boldsymbol{\lambda}_{1:T} | \theta_\Lambda^{(-j)}, \theta_\Lambda^{(j)}\right) \pi\left(\boldsymbol{\delta}_{1:K} | \theta_\Lambda^{(-j)}, \theta_\Lambda^{(j)}\right) \pi\left(\theta_\Lambda^{(-j)} | \theta_\Lambda^{(j)}\right) \pi\left(\theta_\Lambda^{(j)}\right). \end{aligned} \quad (11.12.3)$$

Part 2: The next full conditional distribution we must specify is given by

$$\begin{aligned} \pi\left(\lambda_t^{(j)}|\boldsymbol{\theta}_\Lambda, \boldsymbol{\lambda}_{1:T}^{(-t,-j)}, \mathbf{n}_{1:T}, \boldsymbol{\delta}_{1:K}\right) &\propto \pi\left(\boldsymbol{\theta}_\Lambda, \boldsymbol{\lambda}_{1:T}^{(-t,-j)}, \mathbf{n}_{1:T}, \boldsymbol{\delta}_{1:K}|\lambda_t^{(j)}\right) \pi\left(\lambda_t^{(j)}\right) \\ &= \pi\left(\boldsymbol{\lambda}_{1:T}^{(-t,-j)}, \mathbf{n}_{1:T}, \boldsymbol{\delta}_{1:K}|\boldsymbol{\theta}_\Lambda, \lambda_t^{(j)}\right) \pi\left(\boldsymbol{\theta}_\Lambda|\lambda_t^{(j)}\right) \pi\left(\lambda_t^{(j)}\right). \end{aligned} \quad (11.12.4)$$

We then use conditional independence properties of the model to get

$$\begin{aligned} \pi\left(\boldsymbol{\lambda}_{1:T}^{(-t,-j)}, \mathbf{n}_{1:T}, \boldsymbol{\delta}_{1:K}|\boldsymbol{\theta}_\Lambda, \lambda_t^{(j)}\right) \\ = \pi\left(\mathbf{n}_{1:T}|\boldsymbol{\lambda}_{1:T}^{(-t,-j)}, \lambda_t^{(j)}\right) \pi\left(\boldsymbol{\lambda}_{1:T}^{(-t,-j)}|\boldsymbol{\theta}_\Lambda, \lambda_t^{(j)}\right) \pi\left(\boldsymbol{\delta}_{1:K}|\boldsymbol{\theta}_\Lambda\right) \end{aligned} \quad (11.12.5)$$

giving the full conditional we are interested in, up to proportionality,

$$\begin{aligned} \pi\left(\lambda_t^{(j)}|\boldsymbol{\theta}_\Lambda, \boldsymbol{\lambda}_{1:T}^{(-t,-j)}, \mathbf{n}_{1:T}, \boldsymbol{\delta}_{1:K}\right) &\propto \pi\left(\mathbf{n}_{1:T}|\boldsymbol{\lambda}_{1:T}^{(-t,-j)}, \lambda_t^{(j)}\right) \\ &\times \pi\left(\boldsymbol{\delta}_{1:K}|\boldsymbol{\theta}_\Lambda\right) \pi\left(\boldsymbol{\lambda}_{1:T}^{(-t,-j)}|\boldsymbol{\theta}_\Lambda, \lambda_t^{(j)}\right) \pi\left(\boldsymbol{\theta}_\Lambda|\lambda_t^{(j)}\right) \pi\left(\lambda_t^{(j)}\right). \end{aligned} \quad (11.12.6)$$

We now demonstrate that this expression simplifies significantly. We can show that the terms $\pi\left(\boldsymbol{\lambda}_{1:T}^{(-t,-j)}|\boldsymbol{\theta}_\Lambda, \lambda_t^{(j)}\right) \pi\left(\boldsymbol{\theta}_\Lambda|\lambda_t^{(j)}\right) \pi\left(\lambda_t^{(j)}\right)$ simplify as follows:

$$\begin{aligned} \pi\left(\boldsymbol{\lambda}_{1:T}^{(-t,-j)}|\boldsymbol{\theta}_\Lambda, \lambda_t^{(j)}\right) \pi\left(\boldsymbol{\theta}_\Lambda|\lambda_t^{(j)}\right) \pi\left(\lambda_t^{(j)}\right) &= \frac{\pi\left(\boldsymbol{\theta}_\Lambda, \boldsymbol{\lambda}_{1:T}^{(-t,-j)}, \lambda_t^{(j)}\right)}{\pi\left(\boldsymbol{\theta}_\Lambda, \lambda_t^{(j)}\right)} \pi\left(\boldsymbol{\theta}_\Lambda|\lambda_t^{(j)}\right) \pi\left(\lambda_t^{(j)}\right) \\ &= \frac{\pi\left(\boldsymbol{\lambda}_{1:T}^{(-t,-j)}, \lambda_t^{(j)}|\boldsymbol{\theta}_\Lambda\right) \pi\left(\boldsymbol{\theta}_\Lambda\right)}{\pi\left(\boldsymbol{\theta}_\Lambda|\lambda_t^{(j)}\right) \pi\left(\lambda_t^{(j)}\right)} \pi\left(\boldsymbol{\theta}_\Lambda|\lambda_t^{(j)}\right) \pi\left(\lambda_t^{(j)}\right) \\ &= \pi\left(\boldsymbol{\lambda}_{1:T}^{(-t,-j)}, \lambda_t^{(j)}|\boldsymbol{\theta}_\Lambda\right) \pi\left(\boldsymbol{\theta}_\Lambda\right). \end{aligned} \quad (11.12.7)$$

Finally, we are left with the full conditional distribution

$$\begin{aligned} \pi\left(\lambda_t^{(j)}|\boldsymbol{\theta}_\Lambda, \boldsymbol{\lambda}_{1:T}^{(-t,-j)}, \mathbf{n}_{1:T}, \boldsymbol{\delta}_{1:K}\right) &\propto \pi\left(\mathbf{n}_{1:T}|\boldsymbol{\lambda}_{1:T}^{(-t,-j)}, \lambda_t^{(j)}\right) \\ &\times \pi\left(\boldsymbol{\delta}_{1:K}|\boldsymbol{\theta}_\Lambda\right) \pi\left(\boldsymbol{\lambda}_{1:T}^{(-t,-j)}|\boldsymbol{\theta}_\Lambda, \lambda_t^{(j)}\right) \pi\left(\boldsymbol{\theta}_\Lambda|\lambda_t^{(j)}\right) \pi\left(\lambda_t^{(j)}\right) \\ &\propto \pi\left(\mathbf{n}_{1:T}|\boldsymbol{\lambda}_{1:T}^{(-t,-j)}, \lambda_t^{(j)}\right) \pi\left(\boldsymbol{\lambda}_{1:T}^{(-t,-j)}, \lambda_t^{(j)}|\boldsymbol{\theta}_\Lambda\right) \\ &\propto \pi\left(\mathbf{n}_{1:T}|\boldsymbol{\lambda}_{1:T}^{(-t,-j)}, \lambda_t^{(j)}\right) \pi\left(\boldsymbol{\lambda}_t^{(-j)}, \lambda_t^{(j)}|\boldsymbol{\theta}_\Lambda\right). \end{aligned} \quad (11.12.8)$$

Part 3: The full conditional distribution for the copula parameter is given by

$$\begin{aligned} \pi\left(\theta_\rho|\boldsymbol{\theta}_\Lambda, \boldsymbol{\lambda}_{1:T}, \mathbf{n}_{1:T}, \boldsymbol{\delta}_{1:K}\right) &\propto \pi\left(\boldsymbol{\theta}_\Lambda, \boldsymbol{\lambda}_{1:T}, \mathbf{n}_{1:T}, \boldsymbol{\delta}_{1:K}|\theta_\rho\right) \pi\left(\theta_\rho\right) \\ &\propto \pi\left(\mathbf{n}_{1:T}|\boldsymbol{\lambda}_{1:T}\right) \pi\left(\boldsymbol{\delta}_{1:K}|\boldsymbol{\theta}_\Lambda\right) \pi\left(\boldsymbol{\lambda}_{1:T}|\boldsymbol{\theta}_\Lambda, \theta_\rho\right) \pi\left(\boldsymbol{\theta}_\Lambda\right) \pi\left(\theta_\rho\right) \\ &\propto \pi\left(\boldsymbol{\lambda}_{1:T}|\boldsymbol{\theta}_\Lambda, \theta_\rho\right) \pi\left(\theta_\rho\right). \end{aligned} \quad (11.12.9)$$

11.13 Appendix C: Slice sampler algorithm.

Here, we provide the explicit details involved into implementation of a Slice sampler algorithm within a Gibbs sampler framework discussed in Section 5. The iterations of the Slice sampler are denoted by simulation index $l \in \mathbb{N}$.

Slice sampling:

1. Initialize $l = 0$ the parameter vector $[\boldsymbol{\theta}_{\Lambda,0}, \boldsymbol{\lambda}_{1:T,0}, \theta_{\rho,0}]$ randomly or deterministically.
2. Repeat while $l \leq L$
 - (a) Set $[\boldsymbol{\theta}_{\Lambda,l}, \boldsymbol{\lambda}_{1:T,l}, \theta_{\rho,l}] = [\boldsymbol{\theta}_{\Lambda,l-1}, \boldsymbol{\lambda}_{1:T,l-1}, \theta_{\rho,l-1}]$
 - (b) Sample j uniformly from set $\{1, 2, \dots, J\}$
 Sample new parameter value $\tilde{\theta}_{\Lambda}^{(j)}$ from the full conditional posterior distribution $\pi\left(\theta_{\Lambda}^{(j)} | \boldsymbol{\theta}_{\Lambda,l}^{(-j)}, \boldsymbol{\lambda}_{1:T,l}, \mathbf{n}_{1:T}, \boldsymbol{\delta}_{1:K}, \theta_{\rho,l}\right)$.
 Set $\theta_{\Lambda,l}^{(j)} = \tilde{\theta}_{\Lambda}^{(j)}$.
 - (c) Sample j uniformly from set $\{1, 2, \dots, J\}$ and t uniformly from set $\{1, \dots, T\}$
 Sample new parameter value $\tilde{\lambda}_t^{(j)}$ from the full conditional posterior distribution $\pi\left(\lambda_t^{(j)} | \boldsymbol{\theta}_{\Lambda,l}, \boldsymbol{\lambda}_{1:T,l}^{(-t,-j)}, \mathbf{n}_{1:T}, \boldsymbol{\delta}_{1:K}, \theta_{\rho,l}\right)$.
 Set $\lambda_{t,l}^{(j)} = \tilde{\lambda}_t^{(j)}$.
 - (d) Sample new parameter value $\tilde{\theta}_{\rho}$ from the full conditional posterior distribution $\pi\left(\theta_{\rho} | \boldsymbol{\theta}_{\Lambda,l}, \boldsymbol{\lambda}_{1:T,l}, \mathbf{n}_{1:T}, \boldsymbol{\delta}_{1:K}\right)$.
 Set $\theta_{\rho,l} = \tilde{\theta}_{\rho}$.
3. $l = l + 1$ and return to 2.

The sampling from the full conditional posteriors in stage 2 uses a univariate Slice sampler, see Figure 11.13.2. We present the case where we wish to sample the next iteration of the Markov chain from $\pi\left(\theta_{\Lambda}^{(j)} | \boldsymbol{\theta}_{\Lambda,l}^{(-j)}, \boldsymbol{\lambda}_{1:T,l}, \mathbf{n}_{1:T}, \boldsymbol{\delta}_{1:K}, \theta_{\rho}\right)$.

Obtaining a sample using a univariate Slice sampler:

1. Sample u from a uniform distribution

$$U \left[0, \pi \left(\theta_{\Lambda}^{(j)} | \boldsymbol{\theta}_{\Lambda,l}^{(-j)}, \boldsymbol{\lambda}_{1:T,l}, \mathbf{n}_{1:T}, \boldsymbol{\delta}_{1:K}, \theta_{\rho} \right) \right].$$

2. Sample $\tilde{\theta}_{\Lambda}^{(j)}$ uniformly from the intervals (level set)

$$A = \left\{ \theta_{\Lambda}^{(j)} : \pi \left(\theta_{\Lambda}^{(j)} | \boldsymbol{\theta}_{\Lambda,l}^{(-j)}, \boldsymbol{\lambda}_{1:T,l}, \mathbf{n}_{1:T}, \boldsymbol{\delta}_{1:K}, \theta_{\rho} \right) > u \right\}.$$

There are many approaches that could be used in determination of the level sets A of our density

$$\pi \left(\theta_{\Lambda}^{(j)} | \theta_{\Lambda,l}^{(-j)}, \lambda_{1:T,l}, \mathbf{n}_{1:T}, \delta_{1:K}, \theta_{\rho} \right),$$

see Neal (2003) [p.712, Section 4]. For simplicity in our proceeding examples we assume that we can restrict our parameter space to the finite ranges and we argue that this is reasonable since we can consider the finite bounds for example set according to machine precision for the smallest and largest number we can represent on our computing platform. This is not strictly required, but simplifies the coding of the algorithm. We then perform what Neal (2003) terms a stepping out and a shrinkage procedure, the details of which are contained in Neal (2003) [p.713, Figure 1]. The basic idea is that given a sampled vertical level u then the level sets A can be found by positioning an interval of width w randomly around $\theta_{\Lambda,l}^{(j)}$. This interval is expanded in step sizes of width w until both ends are outside the slice. Then a new state is obtained by sampling uniformly from the interval until a point in the slice A is obtained. Points that fail can be used to shrink the interval.

| Year | 1 | 2 | 5 | 10 | 15 | 20 |
|---------------------------------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Independent | | | | | | |
| Marginal | 3.72 (2.04) | 4.10 (1.98) | 4.08 (1.62) | 4.64 (1.42) | 5.13 (1.31) | 5.24 (1.27) |
| Gaussian copula with ($\rho = 0.9$) | | | | | | |
| Benchmark | 4.32 (1.88) | 4.50 (1.67) | 4.84 (1.46) | 5.17 (1.26) | 5.19 (1.12) | 5.21 (1.02) |
| Joint | 3.91 (2.01) | 4.41 (1.72) | 4.37 (1.56) | 4.76 (1.33) | 5.10 (1.21) | 4.95 (1.05) |
| Marginal | 3.72 (2.05) | 4.09 (1.97) | 4.06 (1.61) | 4.48 (1.37) | 5.07 (1.29) | 5.04 (1.13) |
| Clayton copula with ($\rho = 10$) | | | | | | |
| Benchmark | 4.81 (1.82) | 5.17 (1.72) | 5.13 (1.42) | 4.96 (1.13) | 5.10 (0.98) | 5.00 (0.84) |
| Joint | 4.19 (2.03) | 4.92 (1.87) | 5.05 (1.56) | 4.87 (1.26) | 4.96 (1.08) | 4.90 (0.93) |
| Marginal | 3.91 (2.12) | 4.43 (2.10) | 4.54 (1.74) | 4.47 (1.36) | 4.75 (1.22) | 4.72 (1.08) |
| Gumbel copula with ($\rho = 3$) | | | | | | |
| Benchmark | 4.32 (1.98) | 4.46 (1.70) | 4.86 (1.41) | 5.08 (1.16) | 5.16 (1.01) | 5.11 (0.88) |
| Joint | 4.33 (2.06) | 4.21 (1.80) | 4.54 (1.56) | 4.96 (1.23) | 5.01 (1.05) | 4.98 (0.93) |
| Marginal | 3.84 (2.08) | 3.76 (1.87) | 4.17 (1.62) | 4.63 (1.41) | 4.74 (1.22) | 4.72 (1.07) |

Tab. 11.1: Average estimates of posterior mean and standard deviation of $\Theta_{\Lambda}^{(1)}$ for 20 data sets. Data are generated using different copula models as specified. The true values are $\theta_{true}^{(1)} = \theta_{true}^{(2)} = 5$.

| Year | 1 | 2 | 5 | 10 | 15 | 20 |
|---------------------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Independent | | | | | | |
| Marginal | 6.74 (2.74) | 6.84 (2.59) | 6.46 (2.16) | 5.91 (1.67) | 5.74 (1.40) | 5.47 (1.31) |
| Gaussian ($\rho = 0.9$) | | | | | | |
| Benchmark | 5.98 (2.29) | 5.84 (2.04) | 5.46 (1.60) | 5.47 (1.31) | 5.43 (1.14) | 5.41 (1.04) |
| Joint | 6.37 (2.55) | 6.01 (2.23) | 5.63 (1.80) | 5.40 (1.43) | 5.43 (1.25) | 5.36 (1.12) |
| Marginal | 6.59 (2.72) | 6.49 (2.54) | 6.01 (2.07) | 5.75 (1.64) | 5.62 (1.43) | 5.57 (1.26) |
| Clayton ($\rho = 10$) | | | | | | |
| Benchmark | 5.57 (1.91) | 5.41 (1.69) | 5.20 (1.40) | 4.90 (1.10) | 5.09 (0.96) | 5.07 (0.85) |
| Joint | 6.39 (2.48) | 5.92 (1.92) | 5.36 (1.64) | 5.06 (1.22) | 5.13 (1.17) | 5.00 (1.02) |
| Marginal | 6.69 (2.74) | 6.56 (2.55) | 5.92 (2.04) | 5.40 (1.56) | 5.37 (1.36) | 5.24 (1.17) |
| Gumbel ($\rho = 3$) | | | | | | |
| Benchmark | 5.83 (2.35) | 5.51 (2.02) | 5.38 (1.57) | 5.15 (1.18) | 5.20 (1.02) | 5.12 (0.89) |
| Joint | 6.05 (2.47) | 5.96 (2.17) | 5.47 (1.76) | 5.21 (1.27) | 5.12 (1.07) | 5.12 (0.94) |
| Marginal | 6.42 (2.67) | 6.26 (2.50) | 5.92 (2.04) | 5.67 (1.62) | 5.52 (1.37) | 5.36 (1.18) |

Tab. 11.2: Average estimates of posterior mean and standard deviation of $\Theta_{\Lambda}^{(2)}$ for 20 data sets. The data are generated using different copula models as specified. The true values are $\theta_{true}^{(1)} = \theta_{true}^{(2)} = 5$.

| Year | 1 | 2 | 5 | 10 | 15 | 20 |
|---------------------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Independent | | | | | | |
| Marginal | 3.72 (2.04) | 4.07 (1.97) | 4.05 (1.61) | 4.48 (1.37) | 4.94 (1.26) | 5.13 (1.13) |
| Gaussian ($\rho = 0.9$) | | | | | | |
| Benchmark | 4.07 (1.76) | 4.22 (1.53) | 4.61 (1.36) | 5.10 (1.20) | 5.10 (1.08) | 5.20 (0.99) |
| Joint | 3.86 (1.96) | 4.39 (1.86) | 4.46 (1.51) | 4.84 (1.33) | 5.10 (1.22) | 5.24 (1.10) |
| Marginal | 3.72 (2.04) | 4.08 (1.97) | 4.05 (1.61) | 4.48 (1.37) | 4.94 (1.26) | 5.13 (1.13) |
| Clayton ($\rho = 10$) | | | | | | |
| Benchmark | 4.45 (1.65) | 4.86 (1.55) | 4.89 (1.32) | 4.82 (1.07) | 5.00 (0.95) | 4.92 (0.83) |
| Joint | 4.15 (2.01) | 4.84 (1.95) | 4.92 (1.59) | 4.69 (1.33) | 4.97 (1.20) | 4.96 (1.04) |
| Marginal | 3.98 (2.10) | 4.53 (2.09) | 4.54 (1.74) | 4.47 (1.36) | 4.74 (1.21) | 4.72 (1.08) |
| Gumbel ($\rho = 3$) | | | | | | |
| Benchmark | 4.14 (1.90) | 4.20 (1.58) | 4.65 (1.32) | 4.95 (1.11) | 5.06 (0.97) | 5.04 (0.87) |
| Joint | 4.36 (2.16) | 4.17 (1.85) | 4.68 (1.57) | 5.10 (1.34) | 5.21 (1.21) | 5.24 (1.01) |
| Marginal | 3.84 (2.17) | 3.75 (1.87) | 4.17 (1.62) | 4.64 (1.41) | 4.75 (1.22) | 4.79 (1.09) |

Tab. 11.3: Average estimates of posterior mean and standard deviation of $\Theta_{\Lambda}^{(1)}$ for 20 data sets. Data are generated using different copula models as specified. The true values are $\theta_{true}^{(1)} = 5$ and $\theta_{true}^{(2)} = 10$.

| Year | 1 | 2 | 5 | 10 | 15 | 20 |
|---------------------------|--------------|--------------|--------------|--------------|--------------|--------------|
| Independent | | | | | | |
| Marginal | 10.89 (3.74) | 10.78 (3.60) | 10.18 (3.19) | 9.70 (2.67) | 9.64 (2.31) | 9.48 (2.14) |
| Gaussian ($\rho = 0.9$) | | | | | | |
| Benchmark | 10.44 (3.48) | 10.50 (3.24) | 10.25 (2.81) | 10.51 (2.39) | 10.78 (2.05) | 10.04 (1.88) |
| Joint | 10.68 (3.63) | 10.13 (3.35) | 9.57 (2.87) | 9.60 (2.36) | 9.31 (2.03) | 9.23 (1.82) |
| Marginal | 10.89 (3.74) | 10.78 (3.60) | 10.18 (3.19) | 9.70 (2.67) | 9.64 (2.31) | 9.48 (2.15) |
| Clayton ($\rho = 10$) | | | | | | |
| Benchmark | 10.04 (3.21) | 10.07 (2.99) | 9.88 (2.58) | 9.59 (2.08) | 9.97 (1.87) | 9.97 (1.66) |
| Joint | 10.58 (3.54) | 10.03 (3.19) | 9.29 (2.69) | 9.82 (2.22) | 9.93 (1.97) | 9.78 (1.73) |
| Marginal | 10.94 (3.75) | 10.92 (3.61) | 10.13 (3.17) | 9.32 (2.60) | 9.30 (2.25) | 9.45 (1.96) |
| Gumbel ($\rho = 3$) | | | | | | |
| Benchmark | 10.10 (3.55) | 9.88 (3.23) | 10.08 (2.74) | 9.98 (2.22) | 10.17 (1.95) | 10.06 (1.74) |
| Joint | 10.10 (3.61) | 10.18 (3.39) | 9.44 (2.87) | 9.98 (2.33) | 9.85 (1.99) | 9.63 (1.75) |
| Marginal | 10.61 (3.76) | 10.51 (3.59) | 10.11 (3.17) | 9.70 (2.66) | 9.45 (2.27) | 9.39 (1.98) |

Tab. 11.4: Average estimates of posterior mean and standard deviation of $\Theta_{\Lambda}^{(2)}$ for 20 data sets. Data are generated using different copula models as specified. The true values are $\theta_{true}^{(1)} = 5$ and $\theta_{true}^{(2)} = 10$.

| Year | 1 | 2 | 5 | 10 | 15 | 20 |
|--|--------------|--------------|--------------|--------------|--------------|--------------|
| Posterior mean and standard deviation for $\Theta_{\Lambda}^{(1)}$ | | | | | | |
| Independent | 2.81 (1.75) | 4.71 (2.21) | 3.05 (1.29) | 4.90 (1.48) | 4.38 (1.14) | 5.13 (1.15) |
| Gaussian ($\rho = 0.9$) | 2.83 (1.74) | 4.49 (2.02) | 3.31 (1.38) | 4.88 (1.29) | 4.36 (1.10) | 5.07 (1.09) |
| Clayton ($\rho = 10$) | 4.27 (2.04) | 3.80 (1.76) | 4.14 (1.46) | 5.62 (1.45) | 4.53 (1.04) | 4.90 (0.99) |
| Gumbel ($\rho = 3$) | 2.81 (1.40) | 2.59 (1.26) | 3.08 (1.17) | 4.28 (1.26) | 4.51 (1.12) | 4.91 (0.95) |
| Posterior mean and standard deviation for $\Theta_{\Lambda}^{(2)}$ | | | | | | |
| Independent | 10.24 (3.97) | 9.56 (3.56) | 9.48 (3.17) | 9.10 (2.58) | 9.33 (2.24) | 9.55 (1.98) |
| Gaussian ($\rho = 0.9$) | 10.23 (3.92) | 10.85 (3.52) | 8.72 (2.95) | 8.91 (2.12) | 8.58 (2.04) | 9.94 (1.85) |
| Clayton ($\rho = 10$) | 11.40 (3.58) | 10.76 (3.47) | 10.85 (3.10) | 11.39 (2.60) | 10.76 (2.27) | 10.17 (2.03) |
| Gumbel ($\rho = 3$) | 12.27 (3.63) | 11.26 (3.59) | 9.11 (2.93) | 9.54 (2.26) | 9.91 (1.72) | 9.88 (1.17) |
| Posterior mean and standard deviation for Θ_{ρ} | | | | | | |
| Independent | 0.20 (0.53) | 0.10 (0.44) | -0.10 (0.38) | -0.02 (0.30) | -0.17 (0.28) | -0.12 (0.16) |
| Gaussian ($\rho = 0.9$) | 0.21 (0.54) | 0.47 (0.39) | 0.61 (0.30) | 0.66 (0.24) | 0.70 (0.19) | 0.74 (0.15) |
| Clayton ($\rho = 10$) | 5.37 (2.81) | 5.83 (2.66) | 6.20 (2.52) | 6.48 (2.25) | 6.70 (2.11) | 8.24 (1.88) |
| Gumbel ($\rho = 3$) | 16.41 (8.33) | 16.59 (8.19) | 16.39 (8.09) | 9.42 (8.76) | 5.12 (6.21) | 3.90 (4.80) |

Tab. 11.5: Posterior estimates for $\Theta_{\Lambda}^{(1)}$, $\Theta_{\Lambda}^{(2)}$ and copula parameter Θ_{ρ} . In this case a single data set is generated using different copula models as specified. Posterior standard deviations are given in brackets next to estimate. Joint estimation was used.

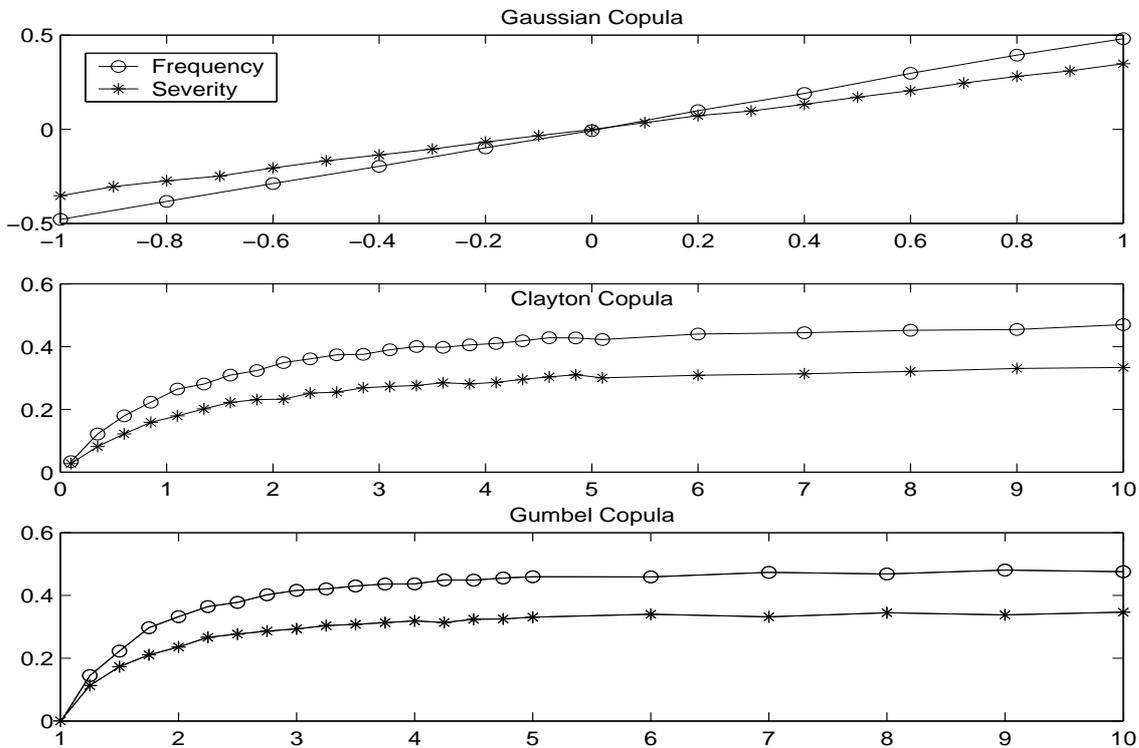


Fig. 11.13.1: Spearman's rank correlation, $\rho_{SR}(Z^{(1)}, Z^{(2)})$, between the annual losses vs copula parameter ρ , also see Section 3. (o) $\rho_{SR}(Z^{(1)}, Z^{(2)})$ vs copula parameter ρ between frequency risk profiles $\Lambda_t^{(1)}$ and $\Lambda_t^{(2)}$; (*) $\rho_{SR}(Z^{(1)}, Z^{(2)})$ vs copula parameter ρ between severity risk profiles $\Psi_t^{(1)}$ and $\Psi_t^{(2)}$.

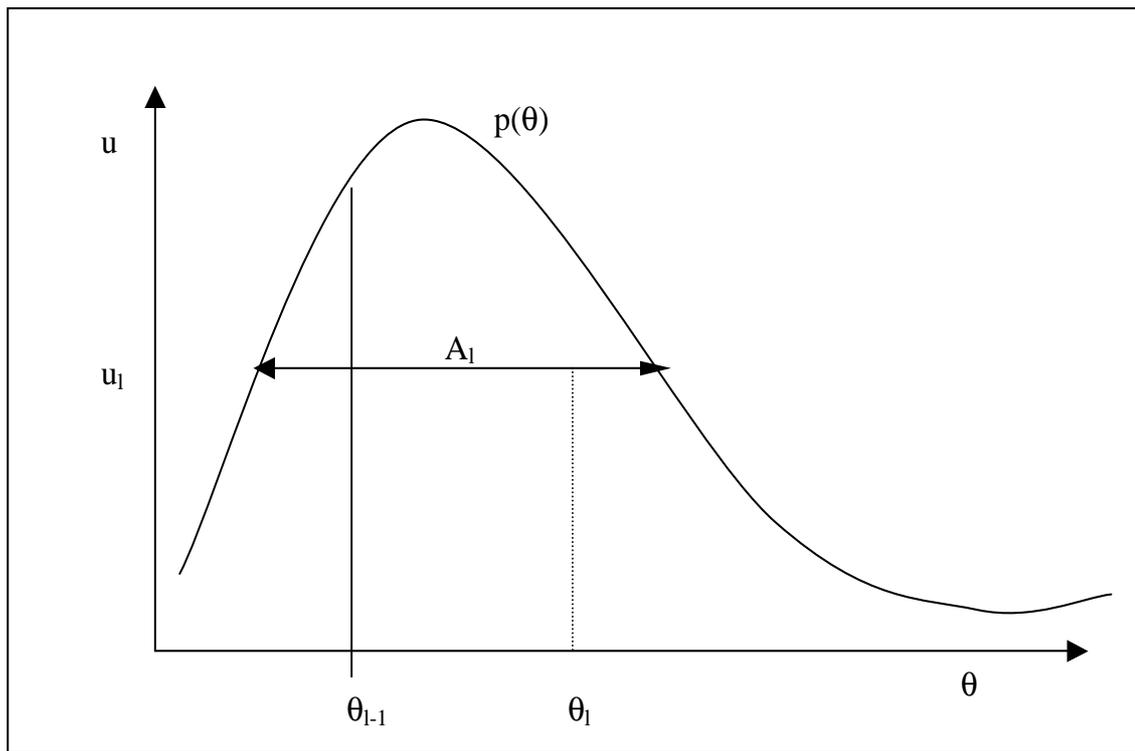


Fig. 11.13.2: Markov chain created for Θ_Λ and auxiliary random variable U , $(u_1, \theta_{\Lambda,1}), \dots, (u_{l-1}, \theta_{\Lambda,l-1}), (u_l, \theta_{\Lambda,l}), \dots$ has stationary distribution with the desired marginal distribution $p(\theta_\Lambda)$.

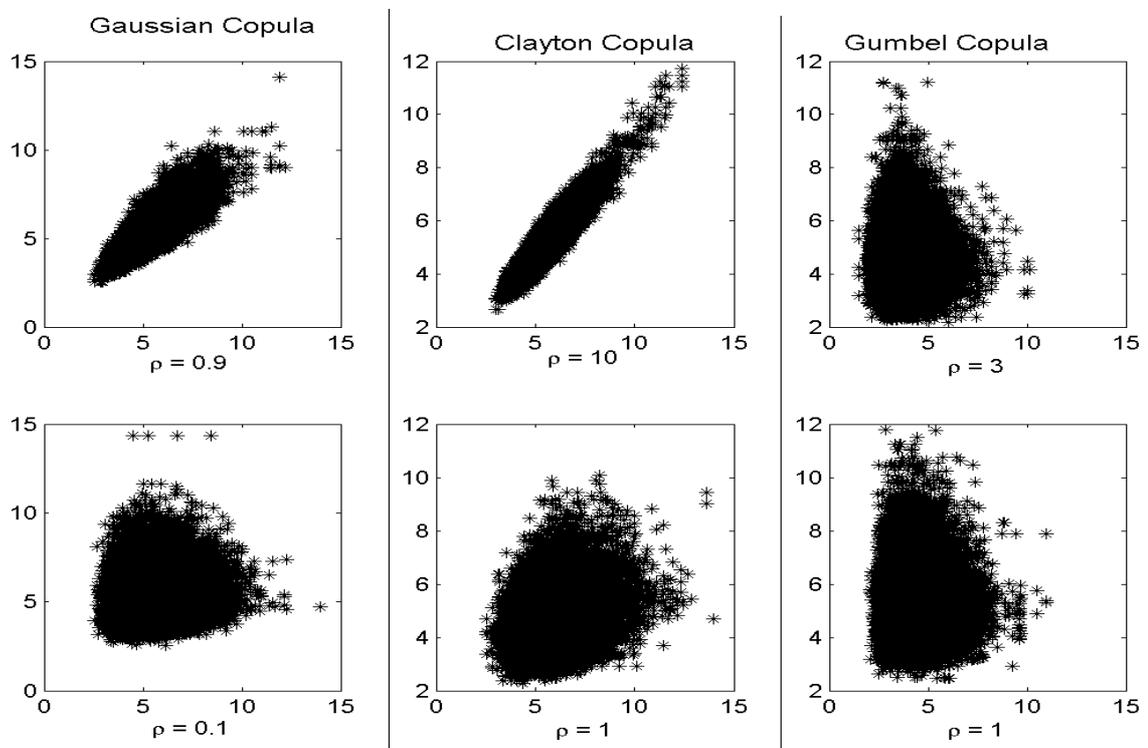


Fig. 11.13.3: Scatter plot of $(\Theta_\Lambda^{(1)}, \Theta_\Lambda^{(2)})$ from $\pi(\theta_\Lambda | n_{1:20}, \delta_{1:1}, \lambda_{1:20}, \theta_\rho)$ with Gaussian, Clayton and Gumbel copulas $C(\cdot | \theta_\rho = \rho)$ between frequency risk profiles. Top row: strong correlation. Bottom row: weak correlation

References

- [1] Artzner P., Delbaen F., Eber J.M. and Heath D. (1999). Coherent measures of risk. *Mathematical Finance* **9**(3), 203-228.
- [2] Atchade Y. and Rosenthal, J. (2005). On adaptive Markov chain Monte Carlo algorithms. *Bernoulli* **11**(5), 815-828.
- [3] Aue F. and Klakbrener, M. (2006). LDA at work: Deutsche Bank's approach to quantifying operational risk. *The Journal of Operational Risk* **1**(4), 49-95.
- [4] Bee M. (2005). Copula-based multivariate models with applications to risk management and insurance. *Preprint, University of Trento*, Available at www.gloriamundi.org.
- [5] BIS (June, 2006). *International Convergence of Capital Measurement and Capital Standards*. Basel Committee on Banking Supervision, Bank for International Settlements, Basel. www.bis.org.
- [6] Böcker K. and Klüppelberg C. (2008). Modelling and measuring multivariate operational risk with Levy copulas. *The Journal of Operational Risk* **3**(2), 3-27.
- [7] Brewer M.J., Aitken C.G.G. and Talbot M. (1996). A comparison of hybrid strategies for Gibbs sampling in mixed graphical models. *Computational Statistics and Data Analysis* **21**(3), 343-365.
- [8] Bühlmann H., Shevchenko P.V. and Wüthrich M.V. (2007). A "Toy" model for operational risk quantification using Credibility theory. *The Journal of Operational Risk* **2**(1), 3-19.
- [9] Chavez-Demoulin V., Embrechts P. and Nešlehová J. (2006). Quantitative models for Operational Risk: Extremes, dependence and aggregation. *Journal of Banking and Finance* **30**(10), 2635-2658.
- [10] Cruz M. (ed.) (2004). *Operational Risk Modelling and Analysis: Theory and Practice*. Risk Books, London.
- [11] Davis E. (2006). Theory vs reality. *OpRisk and Compliance*
www.opriskandcompliance.com/public/showPage.html?page=345305.
- [12] Diaconis P., Holmes S. and Neal R. M. (2000). Analysis of a non-reversible Markov chain sampler. *Annals of Applied Probability* **10**, 726-752.
- [13] Embrechts P. and Frei M. (2008). Panjer recursion versus FFT for compound distributions. *Preprint ETH Zurich*.

- [14] Embrechts P., Lambrigger D.D. and Wüthrich M. (2009). Multivariate extremes and the aggregation of dependent risks: examples and counter-examples. To appear in *Extremes*.
- [15] Embrechts P., Nešlehová J. and Wüthrich M. (2009). Additivity properties for Value-at-Risk under Archimedean dependence and heavy-tailedness. *Insurance: Mathematics and Economics* **44**, 164-169.
- [16] Embrechts P. and Puccetti G. (2008). Aggregation operational risk across matrix structured loss data. *The Journal of Operational Risk*. **3**(2), 29-44.
- [17] Frachot A., Moudoulaud O. and Roncalli T. (2004). Loss distribution approach in practice. *The Basel Handbook: A Guide for Financial Practitioners*. Ong M. (ed), Risk Books.
- [18] Frachot A., Roncalli T. and Salomon E. (2004). The correlation problem in operational risk. *Group de Recherche Opérationnelle, France*. Working paper, www.gloriamundi.org.
- [19] Frees E.W. and Valdez E.A. (1998). Understanding relationships using copulas. *North American Actuarial Journal* **2**, 1-25.
- [20] Geyer C.J. and Thompson E.A. (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. *Journal of the American Statistical Association* **90**(431), 909-920.
- [21] Giacometti R., Rachev S.T., Chernobai A. and Bertocchi M. (2008). Aggregation issues in operational risk. *The Journal of Operational Risk* **3**(3).
- [22] Gilks W.R., Richardson S. and Spiegelhalter D.J. (1996) *Markov Chain Monte Carlo in Practice*. Chapman and Hall, Florida.
- [23] Gilks W.R. and Wild P. (1992). Adaptive Rejection sampling for Gibbs sampling. *Applied Statistics* **41**(2), 337-348.
- [24] Gilks W.R., Best N.G. and Tan K.C. (1995). Adaptive Rejection Metropolis sampling within Gibbs sampling. *Applied Statistics* **44**, 455-472.
- [25] Gramacy R.B., Samworth R. and King R. (2007). Importance tempering. *Cambridge University Statistical Laboratory Technical Report Series*, Preprint on arXiv:0707.4242.
- [26] Lambrigger D.D., Shevchenko P.V. and Wüthrich, M.V. (2007). The quantification of operational risk using internal data, relevant external data and expert opinions. *The Journal of Operational Risk* **2**(3), 3-27.
- [27] Lindskog F. and McNeil A. (2003). Common Poisson shock models: Applications to insurance and credit risk modelling. *ASTIN Bulletin* **33**, 209-238.
- [28] Lou X. and Shevchenko P.V. (2009). Computing Tails of Compound Loss Distributions Using Direct Numerical Integration. To appear in *The Journal of Computational Finance*. Preprint arXiv:0904.0830v2 available on <http://arxiv.org>.
- [29] Marinari E. and Parisi G. (1992). Simulated tempering: a new Monte Carlo scheme. *Europhysics Letters* **19**(6), 451-458.
- [30] Marshall C.L. (2001). *Measuring and Managing Operational Risks in Financial Institutions*, John Wiley & Sons, Singapore.

- [31] McNeil A.J., Frey R. and Embrechts P. (2005). *Quantitative Risk Management: Concepts, Techniques and Tools*, Princeton University Press: Princeton, NJ.
- [32] Melchiori M. (2006). Tools for sampling multivariate Archimedean copulas. *YieldCurve.com*.
- [33] Mira A. and Tierney L. (2002). Efficiency and convergence properties of Slice samplers. *Scandinavian Journal of Statistics* **29**(1), 1-12.
- [34] Neal R.M. (1996). Sampling from multimodal distributions using tempered transitions. *Statistics and Computing* **6**, 353-366.
- [35] Neal R.M. (2003). Slice sampling (with discussions). *Annals of Statistics* **31**, 705-767.
- [36] Nolan J. (2007). *Stable Distributions - Models for Heavy Tailed Data*, Birkhauser.
- [37] O'Hagan A. (2006). *Uncertain Judgements: Eliciting Expert's Probabilities*, Wiley, Statistics in Practice.
- [38] Panjer H. (1981). Recursive evaluation of a family of compound distributions. *ASTIN Bulletin* **12**, 22-26.
- [39] Peters G.W., Johansen A.M. and Doucet A. (2007). Simulation of the annual loss distribution in operational risk via Panjer recursions and Volterra integral equations for value-at-risk and expected shortfall estimation. *The Journal of Operational Risk* **2**(3), 29-58.
- [40] Peters G. W. and Teruads V. (2007). Low probability large consequence events. *Australian Centre for Excellence in Risk Analysis project no. 06/02*.
- [41] Peters G.W. and Sisson S.A. (2006). Monte Carlo sampling and operational risk. *The Journal of Operational Risk* **1**(3), 27-50.
- [42] Powojowski M.R., Reynolds D. and Tuenter H.J.H. (2002). Dependent events and operational risk. *ALGO Research Quarterly* **5**(2), 65-73.
- [43] Robert C.P. and Casella G. (2004). *Monte Carlo Statistical Methods*, 2nd Edition Springer Texts in Statistics.
- [44] Rosenthal J.S. (2008). *MCMC Handbook*, In preparation, Editors: Brooks S., Gelman A., Jones G. and Meng X.L.
- [45] Shevchenko P.V. and Wüthrich M.V. (2006). Structural modelling of operational risk using Bayesian inference: combining loss data with expert opinions. *The Journal of Operational Risk* **1**(3), 3-26.
- [46] Shevchenko P.V. (2008). Estimation of Operational Risk Capital Charge under Parameter Uncertainty. *The Journal of Operational Risk* **3**(1), 51-63.
- [47] Shevchenko P.V. (2009). Implementing Basel II Loss Distribution Approach for Operational Risk. Preprint arXiv:0904.1805 available on <http://arxiv.org>.
- [48] Temnov G. and Warnung R. (2008). A comparison of loss aggregation methods for operational risk. *The Journal of Operational Risk* **3**(1), 3-23.

Journal Paper 9

"Science does not know its debt to imagination."

Ralph Waldo Emerson

Peters G.W., Shevchenko P. and Wuthrich M. (2009) "Model Risk in Claims Reserving within Tweedies Compound Poisson Models" *ASTIN Bulletin*, 39(1).

This work was instigated jointly by the first author and his co-authors. The first author on this major paper can claim at least 50% of the credit for the contents. This work was presented at an international Actuarial (peer reviewed conference) in the UK and has already been cited several times. The first authors work included developing the methodology contained, developing the applications, the implementation and comparison to alternative approaches, writing the drafts of the paper and undertaking revisions. This paper has been accepted for publication in the *ASTIN Bulletin*, one of the top Actuarial Journals and has appeared. Permission from all the co-authors is granted for inclusion in the PhD.

This paper appears in the thesis in a modified format to that which finally appeared in the Actuarial journal *ASTIN Bulletin*, where it was published.

Final print version available at: <http://poj.peeters-leuven.be>

Model uncertainty in claims reserving within Tweedie's compound Poisson models

Gareth W. Peters

CSIRO Mathematical and Information Sciences, Locked Bag 17, North Ryde, NSW, 1670, Australia;

UNSW Mathematics and Statistics Department, Sydney, 2052, Australia;

email: peterga@maths.unsw.edu.au

Pavel V. Shevchenko (*corresponding author*)

CSIRO Mathematical and Information Sciences, Sydney, Locked Bag 17, North Ryde, NSW, 1670,

Australia; e-mail: Pavel.Shevchenko@csiro.au

Mario V. Wüthrich

ETH Zurich, Department of Mathematics, CH-8092 Zurich, Switzerland;

email: wueth@math.ethz.ch

13 October 2008

12.1 abstract

In this paper we examine the claims reserving problem using Tweedie's compound Poisson model. We develop the maximum likelihood and Bayesian Markov chain Monte Carlo simulation approaches to fit the model and then compare the estimated models under different scenarios. The key point we demonstrate relates to the comparison of reserving quantities with and without model uncertainty incorporated into the prediction. We consider both the model selection problem and the model averaging solutions for the predicted reserves. As a part of this process we also consider the sub problem of variable selection to obtain a parsimonious representation of the model being fitted.

Keywords: Claims reserving, model uncertainty, Tweedie's compound Poisson model, Bayesian analysis, model selection, model averaging, Markov chain Monte Carlo.

12.2 Claims reserving

Setting appropriate claims reserves to meet future claims payment cash flows is one of the main tasks of non-life insurance actuaries. There is a wide range of models, methods and algorithms used to set appropriate claims reserves. Among the most popular methods there is the chain-ladder method, the Bornhuetter-Ferguson method and the generalized linear model methods. For an overview, see Wüthrich and Merz (2008) and England and Verrall (2002).

Setting claims reserves includes two tasks: estimate the mean of future payments and quantify the uncertainty in this prediction for future payments. Typically, quantifying the uncertainty includes two terms, namely the so-called process variance and the (parameter) estimation error. The process variance reflects that we predict random variables, i.e. it describes the pure process uncertainty. The estimation error reflects that the true model parameters need to be estimated and hence there is an uncertainty in the reliability of these estimates. In this paper, in addition to these two terms, we consider a third source of error/uncertainty, namely, we analyze the fact that we could have chosen the wrong model. That is, we select a family of claims reserving models and quantify the uncertainty coming from a possibly wrong model choice within this family of models.

Such an analysis is especially important when answering solvency questions. A poor model choice may result in a severe shortfall in the balance sheet of an insurance company, which requires under a risk-adjusted solvency regime an adequate risk capital charge. We analyze typical sizes of such risk capital charges within the family of Tweedie's compound Poisson models, see Tweedie (1984), Smyth and Jørgensen (2002) and Wüthrich (2003).

Assume that $Y_{i,j}$ are incremental claims payments with indices $i, j \in \{0, \dots, I\}$, where i denotes the accident year and j denotes the development year. At time I , we have observations

$$\mathcal{D}_I = \{Y_{i,j}; i + j \leq I\} \quad (12.2.1)$$

and for claims reserving at time I we need to predict the future payments

$$\mathcal{D}_I^c = \{Y_{i,j}; i + j > I, i \leq I\}, \quad (12.2.2)$$

see Table 12.1. Hence, the outstanding claims payment at time I is given by

$$R = \sum_{i=1}^I R_i = \sum_{i+j>I} Y_{i,j}. \quad (12.2.3)$$

Its conditional expectation at time I is given by

$$E[R | \mathcal{D}_I] = \sum_{i=1}^I E[R_i | \mathcal{D}_I] = \sum_{i+j>I} E[Y_{i,j} | \mathcal{D}_I]. \quad (12.2.4)$$

Hereafter, the summation $i + j > I$ is for $i \leq I$. Therefore, we need to predict R and to estimate

$E[R|\mathcal{D}_I]$. Assume that \widehat{R} is an appropriate \mathcal{D}_I -measurable predictor for R and \mathcal{D}_I -measurable estimator for $E[R|\mathcal{D}_I]$. Then, \widehat{R} is used to predict the future payments and is the amount that is put aside in the balance sheet of the insurance company for these payments.

Prediction uncertainty is then often studied with the help of the (conditional) mean square error of prediction (MSEP) which is defined by

$$\text{mse}_{R|\mathcal{D}_I}(\widehat{R}) = E \left[\left(R - \widehat{R} \right)^2 \middle| \mathcal{D}_I \right]. \quad (12.2.5)$$

If \widehat{R} is \mathcal{D}_I -measurable, the conditional MSEP can easily be decoupled as follows, see Wüthrich and Merz (2008), section 3.1:

$$\begin{aligned} \text{mse}_{R|\mathcal{D}_I}(\widehat{R}) &= \text{Var}(R|\mathcal{D}_I) + \left(E[R|\mathcal{D}_I] - \widehat{R} \right)^2 \\ &= \text{process variance} + \text{estimation error}. \end{aligned} \quad (12.2.6)$$

It is clear that the consistent estimator \widehat{R} which minimizes the conditional MSEP is given by $\widehat{R} = E[R|\mathcal{D}_I]$ and is used, hereafter, as the "best estimate" for reserves. Assuming the model is parameterized by the parameter vector θ , $\text{Var}(R|\mathcal{D}_I)$ can be decomposed as

$$\begin{aligned} \text{Var}(R|\mathcal{D}_I) &= E[\text{Var}(R|\theta, \mathcal{D}_I)|\mathcal{D}_I] + \text{Var}(E[R|\theta, \mathcal{D}_I]|\mathcal{D}_I) \\ &= \text{average process variance} + \text{parameter estimation error}. \end{aligned} \quad (12.2.7)$$

These are the two terms that are usually studied when quantifying prediction uncertainties in a Bayesian context, where the unknown parameters θ are modelled stochastically. That is, we obtain in the Bayesian context a similar decomposition as in the frequentist estimation (12.2.6). In the frequentist approach, the second term in (12.2.6) is often estimated by $\text{Var}(\widehat{R})$, see for example section 6.4.3 in Wüthrich and Merz (2008).

As discussed in Cairns (2000), in full generality one could consider several sources of model uncertainty, however unlike Cairns (2000) we focus on a specific class of models. We consider the setting discussed in Bernardo and Smith (1994) termed M Complete modelling. In such a setting the premise is that one considers a set of models in which the "truth" exists but is unknown *a priori*. In this setting we demonstrate the risk associated with the model uncertainty which we analyze jointly as a decomposition into two main parts. The first involves the uncertainty in the parameterization of the model, this is a variable selection problem within a nested model structure in the same vein as discussed in Cairns (2000). It relates to finding a trade-off between parsimony and accuracy in the estimation. The second source of model uncertainty that we study involves the choice of a parameter which determines membership from a spectrum of possible models within the Tweedie's compound Poisson family of models. We restrict the analysis to Tweedie's compound Poisson models and justify this by assuming we are working in the M Complete setting. If we relaxed this assumption and therefore consider competing models not in this family, then the analysis would be difficult to interpret and analyze in the manner we develop in this paper. The second source of model uncertainty will be considered

under both a model selection and a model averaging setting, given the first "variable selection" uncertainty is resolved. As mentioned in Cairns (2000) achieving such an analysis requires advanced simulation methodology. Note, in future work we would also consider the M Open modeling framework of Bernardo and Smith (1994) which relaxes the belief that the truth lies in the set of models considered and hence introduces additional uncertainty associated with the family of models considered. The advanced sampling methodology required to study the M Open model setting will be briefly discussed.

The paper is organised as follows. In section 12.3, we present Tweedie's compound Poisson model and section 12.4 considers parameter estimation in the model, using the maximum Likelihood and Bayesian Markov chain Monte Carlo approaches for a real data set. Having addressed the variable selection question in section 12.5, we then analyze claims reserve estimation and model uncertainty in both a frequentist and Bayesian setting in section 12.6. We finish with conclusions from our findings.

12.3 Tweedie's compound Poisson model

We assume that $Y_{i,j}$ belongs to the family of Tweedie's compound Poisson models. Below we provide three different parameterizations for Tweedie's compound Poisson models, for rigorous derivations we refer to Jørgensen and de Souza (1994), Smyth and Jørgensen (2002) and Wüthrich (2003).

Model Assumptions 12.3.1 (1st Representation). *We assume that $Y_{i,j}$ are independent for $i, j \in \{0, \dots, I\}$ and have a compound Poisson distribution*

$$Y_{i,j} = 1_{\{N_{i,j} > 0\}} \sum_{k=1}^{N_{i,j}} X_{i,j}^{(k)}, \quad (12.3.1)$$

in which (a) $N_{i,j}$ and $X_{i,j}^{(k)}$ are independent for all k , (b) $N_{i,j}$ is Poisson distributed with parameter $\lambda_{i,j}$; (c) $X_{i,j}^{(k)}$ are independent gamma severities with the mean $\tau_{i,j} > 0$ and the shape parameter $\gamma > 0$. Hereafter, we denote 1_{Ω} as an indicator function.

2nd Representation. The random variable $Y_{i,j}$ given in 12.3.1 belongs to the family of Tweedie's compound Poisson models, see Tweedie (1984). The distribution of $Y_{i,j}$ can be reparameterized in such a way that it takes a form of the exponential dispersion family, see e.g. formula (3.5) and Appendix A in Wüthrich (2003):

$Y_{i,j}$ has a probability weight at 0 given by

$$P[Y_{i,j} = 0] = P[N_{i,j} = 0] = \exp \left\{ -\phi_{i,j}^{-1} \kappa_p(\theta_{i,j}) \right\} \quad (12.3.2)$$

and for $y > 0$ the random variable $Y_{i,j}$ has continuous density

$$f_{\theta_{i,j}}(y; \phi_{i,j}, p) = c(y; \phi_{i,j}, p) \exp \left\{ \frac{y \theta_{i,j} - \kappa_p(\theta_{i,j})}{\phi_{i,j}} \right\}. \quad (12.3.3)$$

Here $\theta_{i,j} < 0$, $\phi_{i,j} > 0$, the normalizing constant is given by

$$c(y; \phi, p) = \sum_{r \geq 1} \left(\frac{(1/\phi)^{\gamma+1} y^\gamma}{(p-1)^\gamma (2-p)} \right)^r \frac{1}{r! \Gamma(r\gamma) y} \quad (12.3.4)$$

and the cumulant generating function $\kappa_p(\cdot)$ is given by

$$\kappa_p(\theta) \stackrel{def.}{=} \frac{1}{2-p} [(1-p)\theta]^\gamma, \quad (12.3.5)$$

where $p \in (1, 2)$ and $\gamma = (2-p)/(1-p)$.

The parameters, in terms of the 1st representation quantities, are:

$$p = p(\gamma) = \frac{\gamma + 2}{\gamma + 1} \in (1, 2), \quad (12.3.6)$$

$$\phi_{i,j} = \frac{\lambda_{i,j}^{1-p} \tau_{i,j}^{2-p}}{2-p} > 0, \quad (12.3.7)$$

$$\theta_{i,j} = \left(\frac{1}{1-p} \right) (\mu_{i,j})^{(1-p)} < 0, \quad (12.3.8)$$

$$\mu_{i,j} = \lambda_{i,j} \tau_{i,j} > 0. \quad (12.3.9)$$

Then the mean and variance of $Y_{i,j}$ are given by

$$E[Y_{i,j}] = \frac{\partial}{\partial \theta_{i,j}} \kappa_p(\theta_{i,j}) = \kappa_p'(\theta_{i,j}) = [(1-p)\theta_{i,j}]^{1/(1-p)} = \mu_{i,j}, \quad (12.3.10)$$

$$\text{Var}(Y_{i,j}) = \phi_{i,j} \kappa_p''(\theta_{i,j}) = \phi_{i,j} \mu_{i,j}^p. \quad (12.3.11)$$

That is, $Y_{i,j}$ has the mean $\mu_{i,j}$, dispersion $\phi_{i,j}$ and variance function with the variance parameter p . The extreme cases $p \rightarrow 1$ and $p \rightarrow 2$ correspond to the overdispersed Poisson and the gamma models, respectively. Hence, in this spirit, Tweedie's compound Poisson model with $p \in (1, 2)$ closes the gap between the Poisson and the gamma models. Often in practice, p is assumed to be known and fixed by the modeller. The aim of this paper is to study *Model Uncertainty*, that is, we would like to study the sensitivity of the claims reserves within this subfamily, i.e. Tweedie's compound Poisson models (which are now parameterized through p). This answers model uncertainty questions within the family of Tweedie's compound Poisson models. In this paper the restriction on $p \in (1, 2)$ is taken in the context of practical application of these models to claims reserving, Wüthrich (2003) comments that the majority of claims reserving problems will be captured under this assumption. However, in general, in the exponential dispersion family p can be outside of the $(1, 2)$ range, e.g. $p = 0$ produces a Gaussian density and $p = 3$ leads to an inverse Gaussian model.

3rd Representation. Utilizing the above definitions, the distribution of $Y_{i,j}$ can be rewritten in terms of $\mu_{i,j}$, p and $\phi_{i,j}$ as

$$P[Y_{i,j} = 0] = P[N_{i,j} = 0] = \exp \left\{ -\phi_{i,j}^{-1} \frac{\mu_{i,j}^{2-p}}{2-p} \right\} \quad (12.3.12)$$

and for $y > 0$

$$f_{\mu_{i,j}}(y; \phi_{i,j}, p) = c(y; \phi_{i,j}, p) \exp \left\{ \phi_{i,j}^{-1} \left[y \frac{\mu_{i,j}^{1-p}}{1-p} - \frac{\mu_{i,j}^{2-p}}{2-p} \right] \right\}. \quad (12.3.13)$$

12.4 Parameter estimation

Our goal is to estimate the parameters $\mu_{i,j}$, p and $\phi_{i,j}$ based on the observations \mathcal{D}_I . In order to estimate these parameters we need to introduce additional structure in the form of a multiplicative model.

Model Assumptions 12.4.1. Assume that there exist exposures $\alpha = (\alpha_0, \dots, \alpha_I)$ and a development pattern $\beta = (\beta_0, \dots, \beta_I)$ such that we have for all $i, j \in \{0, \dots, I\}$

$$\mu_{i,j} = \alpha_i \beta_j. \quad (12.4.1)$$

Moreover, assume that $\phi_{i,j} = \phi$ and $\alpha_i > 0$, $\beta_j > 0$.

In addition, we impose the normalizing condition $\alpha_0 = 1$, so that the estimation problem is well-defined. That is we have $(2I + 3)$ unknown parameters p, ϕ, α, β that have to be estimated from the data \mathcal{D}_I . Next we present the likelihood function for this model and then develop the methodology for parameter estimation using the maximum likelihood and Bayesian inference methods.

12.4.1 Likelihood function

Define the parameter vector $\theta = (p, \phi, \alpha, \beta)$. Then the likelihood function for $Y_{i,j}$, $i + j \leq I$, is given by

$$L_{\mathcal{D}_I}(\theta) = \prod_{i+j \leq I} c(Y_{i,j}; \phi, p) \exp \left\{ \phi^{-1} \left[Y_{i,j} \frac{(\alpha_i \beta_j)^{1-p}}{1-p} - \frac{(\alpha_i \beta_j)^{2-p}}{2-p} \right] \right\}, \quad (12.4.2)$$

where we set $c(0; \phi, p) = 1$ for $Y_{i,j} = 0$. The difficulty in the evaluation of the likelihood function is the calculation of $c(y; \phi, p)$ which contains an infinite sum

$$c(y; \phi, p) = \sum_{r \geq 1} \left(\frac{(1/\phi)^{\gamma+1} y^\gamma}{(p-1)^\gamma (2-p)} \right)^r \frac{1}{r! \Gamma(r\gamma) y} = \frac{1}{y} \sum_{r \geq 1} W_r, \quad (12.4.3)$$

where $\gamma = \gamma(p) = (2 - p) / (1 - p)$. Tweedie (1984) identified this summation as Wright's (1935) generalized Bessel function, which can not be expressed in terms of more common Bessel functions. To evaluate this summation we follow the approach of Dunn and Smyth (2005) which directly sums the infinite series, including only terms which significantly contribute to the summation. Consider the term

$$\log W_r = r \log z - \log \Gamma(1 + r) - \log \Gamma(\gamma r),$$

where

$$z = \frac{(1/\phi)^{\gamma+1} y^\gamma}{(p-1)^\gamma (2-p)}.$$

Replacing the gamma functions using Stirling's approximation and approximating γr by $\gamma r + 1$ we get

$$\log W_r \approx r \{ \log z + (1 + \gamma) - \gamma \log \gamma - (1 + \gamma) \log r \} - \log(2\pi) - \frac{1}{2} \log \gamma - \log r,$$

which is also a reasonable approximation for small r . Treating r as continuous and taking the partial derivative w.r.t. r gives

$$\frac{\partial \log W_r}{\partial r} \approx \log z - \log r - \gamma \log(\gamma r).$$

Hence, the sequence W_r is unimodal in r . Solving $\partial W_r / \partial r = 0$, to find (approximately) the maximum of W_r , results in the approximate maximum lying close to

$$R_0 = R_0(\phi, p) = \frac{y^{2-p}}{(2-p)\phi}. \tag{12.4.4}$$

This gives a surprisingly accurate approximation to the true maximum of W_r , $r \in \mathbb{N}$. Finally, the aim is to find $R_L < R_0 < R_U$ such that the following approximation is sufficiently accurate for the use in the evaluation of the likelihood terms,

$$c(y; \phi, p) \approx \tilde{c}(y; \phi, p) = \frac{1}{y} \sum_{r=R_L}^{R_U} W_r. \tag{12.4.5}$$

The fact that $\partial \log W_r / \partial r$ is monotonic and decreasing implies that $\log W_r$ is strictly convex in r and hence the terms in W_r decay at a faster rate than geometric on either side of R_0 . Dunn and Smyth (2005) derive the following bounds,

$$c(y; \phi, p) - \tilde{c}(y; \phi, p) < W_{R_L-1} \frac{1 - q_L^{R_L-1}}{1 - q_L} + W_{R_U+1} \frac{1}{1 - q_U} \tag{12.4.6}$$

with

$$q_L = \exp\left(\frac{\partial \log W_r}{\partial r}\right)\Big|_{r=R_L-1}, \quad q_U = \exp\left(\frac{\partial \log W_r}{\partial r}\right)\Big|_{r=R_U+1}. \tag{12.4.7}$$

These bounds are typically too conservative since the decay is much faster than geometric. In practice, an adaptive approach balancing accuracy and efficiency is to continue adding terms either side of the maximum until the lower and upper terms satisfy the double precision constraints $W_{R_L} \leq e^{-37}W_{R_0}$ (or $R_L = 1$) and $W_{R_U} \leq e^{-37}W_{R_0}$. When evaluating the summation for $\tilde{c}(y; \phi, p)$, it was important to utilize the following identity to perform the summation in the log scale to avoid numerical overflow problems,

$$\log \tilde{c}(y; \phi, p) = -\log y + \log W_{R_0} + \log \left(\sum_{r=R_L}^{R_U} \exp(\log(W_R) - \log(W_{R_0})) \right).$$

We made an additional observation when analyzing this model. For our data set, as p approaches 1 (i.e. when the distribution approaches the overdispersed Poisson model) the likelihood may become multimodal. Therefore, to avoid numerical complications in actual calculations, we restrict to $p \geq 1.1$. At the other extreme, when $p = 2$ the number of terms required to evaluate $c(y; \phi, p)$ may become very large, hence to manage the computation burden, we restrict $p \leq 1.95$. These limitations are also discussed in Dunn and Smyth (2005). For our data set, we checked that this restriction did not have a material impact on the results.

12.4.2 Maximum likelihood estimation

The maximum likelihood estimator (MLE) for the parameters is given by maximizing $L_{\mathcal{D}_I}(\boldsymbol{\theta})$ in $\boldsymbol{\theta} = (p, \phi, \boldsymbol{\alpha}, \boldsymbol{\beta})$ under the constraints $\alpha_i > 0$, $\beta_j > 0$, $\phi > 0$ and $p \in (1, 2)$. This leads to the MLEs $\hat{\boldsymbol{\theta}}^{\text{MLE}} = (\hat{p}^{\text{MLE}}, \hat{\phi}^{\text{MLE}}, \hat{\boldsymbol{\alpha}}^{\text{MLE}}, \hat{\boldsymbol{\beta}}^{\text{MLE}})$ and to the best estimate reserves for R , given \mathcal{D}_I ,

$$\hat{R}^{\text{MLE}} = \sum_{i+j>I} \hat{\alpha}_i^{\text{MLE}} \hat{\beta}_j^{\text{MLE}}. \quad (12.4.8)$$

A convenient practical approach to obtain the MLEs is to use the fact that at the maximum of the likelihood, $\boldsymbol{\beta}$ are expressed through $\boldsymbol{\alpha}$ and p according to the following set of equations, $p \in (1, 2)$:

$$\beta_k = \frac{\sum_{i=0}^{I-k} Y_{i,k} \alpha_i^{1-p}}{\sum_{i=0}^{I-k} \alpha_i^{2-p}}, \quad k = 0, \dots, I, \quad (12.4.9)$$

obtained by setting partial derivatives

$$\begin{aligned} \frac{\partial \ln L_{\mathcal{D}_I}(\boldsymbol{\theta})}{\partial \beta_k} &= \frac{\partial}{\partial \beta_k} \sum_{j=0}^I \sum_{i=0}^{I-j} \phi^{-1} \left(Y_{i,j} \frac{(\alpha_i \beta_j)^{1-p}}{1-p} - \frac{(\alpha_i \beta_j)^{2-p}}{2-p} \right) \\ &= \sum_{i=0}^{I-k} \phi^{-1} \left(Y_{i,k} \alpha_i^{1-p} \beta_k^{-p} - \alpha_i^{2-p} \beta_k^{1-p} \right) \end{aligned} \quad (12.4.10)$$

equal to zero. Hence, after maximizing the likelihood in α, p, ϕ one then calculates the set of equations (12.4.9) for the remaining parameters utilizing the normalization condition $\alpha_0 = 1$.

Under an asymptotic Gaussian approximation, the distribution of the MLEs is Gaussian with the covariance matrix elements

$$\text{cov} \left(\hat{\theta}_i^{\text{MLE}}, \hat{\theta}_j^{\text{MLE}} \right) \approx (\mathbf{I}^{-1})_{i,j}, \tag{12.4.11}$$

where \mathbf{I} is Fisher's information matrix that can be estimated by the observed information matrix

$$(\mathbf{I})_{i,j} \approx - \left. \frac{\partial^2 \ln L_{\mathcal{D}_I}(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^{\text{MLE}}}. \tag{12.4.12}$$

It is interesting to note that, $\hat{\beta}_I^{\text{MLE}} = Y_{0,I}$. Also, it is easy to show (using (12.4.10) and (12.4.11)) that $\hat{\beta}_I^{\text{MLE}}$ is orthogonal to all other parameters, i.e.

$$\text{cov}(\hat{\beta}_I^{\text{MLE}}, \hat{\theta}_i^{\text{MLE}}) = 0, \quad \hat{\theta}_i^{\text{MLE}} \neq \hat{\beta}_I^{\text{MLE}}. \tag{12.4.13}$$

The next step is to estimate the parameter estimation error in the reserve as a function of the parameter uncertainty. We do this via propagation of error by forming a Taylor expansion around the MLEs, see England and Verrall (2002) formulae (7.6)-(7.8) and Wüthrich (2003) formulae (5.1)-(5.2),

$$\text{stdev} \left(\hat{R}^{\text{MLE}} \right) = \sqrt{\text{Var} \left(\hat{R}^{\text{MLE}} \right)} \tag{12.4.14}$$

$$\begin{aligned} \widehat{\text{Var}} \left(\hat{R}^{\text{MLE}} \right) &= \sum_{i_1+j_1>I} \sum_{i_2+j_2>I} \hat{\alpha}_{i_1}^{\text{MLE}} \hat{\alpha}_{i_2}^{\text{MLE}} \text{cov} \left(\hat{\beta}_{j_1}^{\text{MLE}}, \hat{\beta}_{j_2}^{\text{MLE}} \right) \\ &+ \sum_{i_1+j_1>I} \sum_{i_2+j_2>I} \hat{\beta}_{j_1}^{\text{MLE}} \hat{\beta}_{j_2}^{\text{MLE}} \text{cov} \left(\hat{\alpha}_{i_1}^{\text{MLE}}, \hat{\alpha}_{i_2}^{\text{MLE}} \right) \\ &+ 2 \sum_{i_1+j_1>I} \sum_{i_2+j_2>I} \hat{\alpha}_{i_1}^{\text{MLE}} \hat{\beta}_{j_2}^{\text{MLE}} \text{cov} \left(\hat{\alpha}_{i_2}^{\text{MLE}}, \hat{\beta}_{j_1}^{\text{MLE}} \right). \end{aligned} \tag{12.4.15}$$

Additionally, using the independence assumption on $Y_{i,j}$ and (2.11), the process variance is estimated as

$$\widehat{\text{Var}}(R) = \sum_{i+j>I} \left(\hat{\alpha}_i^{\text{MLE}} \hat{\beta}_j^{\text{MLE}} \right)^{\hat{p}^{\text{MLE}}} \hat{\phi}^{\text{MLE}}. \tag{12.4.16}$$

Then the conditional MSEP (12.2.6) is estimated by

$$\begin{aligned} \widehat{\text{mse}}_{P|\mathcal{D}_I} \left(\hat{R}^{\text{MLE}} \right) &= \widehat{\text{Var}}(R) + \widehat{\text{Var}} \left(\hat{R}^{\text{MLE}} \right) \\ &= \text{MLE process variance} + \text{MLE estimation error}. \end{aligned} \tag{12.4.17}$$

Note that, in practice, typically MLE is done for a fixed p (expert choice) and hence model selection questions are neglected. In our context it means that the expert chooses p and then

estimates $\hat{\alpha}^{\text{MLE}}$, $\hat{\beta}^{\text{MLE}}$ and $\hat{\phi}^{\text{MLE}}$ (see also Wüthrich (2003), section 4.1). The case $p = 1$ corresponds to the overdispersed Poisson model and provides the chain-ladder estimate for the claims reserves (see Wüthrich and Merz (2008), section 2.4). It is important to note that, often the dispersion parameter ϕ is estimated using Pearson's residuals as

$$\hat{\phi}^{\text{P}} = \frac{1}{N - k} \sum_{i+j \leq I} \frac{(Y_{i,j} - \hat{\alpha}_i^{\text{MLE}} \hat{\beta}_j^{\text{MLE}})^2}{(\hat{\alpha}_i^{\text{MLE}} \hat{\beta}_j^{\text{MLE}})^p}, \quad (12.4.18)$$

where N is the number of observations $Y_{i,j}$ in \mathcal{D}_I and k is the number of estimated parameters α_i, β_j (see e.g. Wüthrich and Merz (2008), formula (6.58)). Also note that for a given p , \hat{R}^{MLE} given by (12.4.8) does not depend on ϕ and the estimators for the process variance (12.4.16) and estimation error (12.4.15) are proportional to ϕ . Next we present the Bayesian model which provides the posterior distribution of the parameters given the data. This will be used to analyze the model uncertainty within Tweedie's compound Poisson models.

12.4.3 Bayesian inference

In a Bayesian context all parameters, $p, \phi, \alpha_i > 0$ and $\beta_j > 0$, are treated as random. Using Bayesian inference we adjust our *a priori* beliefs about the parameters of the model utilizing the information from the observations. Through the Bayesian paradigm we are able to learn more about the distribution of p, ϕ, α and β after having observed \mathcal{D}_I .

Our *a priori* beliefs about the parameters of the model are encoded in the form of a prior distribution on the parameters $\pi(\theta)$. Then the joint density of $\mathcal{D}_I = \{Y_{i,j} > 0; i + j \leq I\}$ and $\theta = (p, \phi, \alpha, \beta)$ is given by

$$L_{\mathcal{D}_I}(\theta) \pi(\theta). \quad (12.4.19)$$

Now applying Bayes' law, the posterior distribution of the model parameters, given the data \mathcal{D}_I , is

$$\pi(\theta | \mathcal{D}_I) \propto L_{\mathcal{D}_I}(\theta) \pi(\theta). \quad (12.4.20)$$

Usually, there are two problems that arise in this context, the normalizing constant of this posterior is not known in closed form. Additionally, generating samples from this posterior is typically not possible using simple inversion or rejection sampling approaches. In such cases it is usual to adopt techniques such as Markov chain Monte Carlo (MCMC) methods, see for example Gilks *et al.* (1996) and Robert and Casella (2004) for detailed expositions of such approaches.

The Bayesian estimators typically considered are the Maximum a Posteriori (MAP) estimator and the Minimum Mean Square Estimator (MMSE), that is the mode and mean of the posterior,

defined as follows:

$$MAP : \hat{\boldsymbol{\theta}}^{MAP} = \arg \max_{\boldsymbol{\theta}} [\pi(\boldsymbol{\theta} | \mathcal{D}_I)], \quad (12.4.21)$$

$$MMSE : \hat{\boldsymbol{\theta}}^{MMSE} = E[\boldsymbol{\theta} | \mathcal{D}_I]. \quad (12.4.22)$$

We mention here that if the prior $\pi(\boldsymbol{\theta})$ is constant and the parameter range includes the MLE, then the MAP of the posterior is the same as the MLE. Additionally, one can approximate the posterior using a second order Taylor series expansion around the MAP estimate as

$$\begin{aligned} \ln \pi(\boldsymbol{\theta} | \mathcal{D}_I) &\approx \ln \pi(\hat{\boldsymbol{\theta}}^{MAP} | \mathcal{D}_I) \\ &+ \frac{1}{2} \sum_{i,j} \frac{\partial^2}{\partial \theta_i \partial \theta_j} \ln \pi(\boldsymbol{\theta} | \mathcal{D}_I) \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{MAP}} (\theta_i - \hat{\theta}_i^{MAP}) (\theta_j - \hat{\theta}_j^{MAP}). \end{aligned} \quad (12.4.23)$$

This corresponds to $\pi(\boldsymbol{\theta} | \mathcal{D}_I)$ approximated by the Gaussian distribution with the mean $\hat{\boldsymbol{\theta}}^{MAP}$ and covariance matrix calculated as the inverse of the matrix

$$(\tilde{\mathbf{I}})_{i,j} = - \frac{\partial^2}{\partial \theta_i \partial \theta_j} \ln \pi(\boldsymbol{\theta} | \mathcal{D}_I) \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{MAP}}, \quad (12.4.24)$$

which in the case of diffuse priors (or constant priors defined on a large range) compares with the Gaussian approximation for the MLEs (12.4.11)-(12.4.12).

In the Bayesian context, the conditionally expected future payment, for Model Assumptions 3.1, is given by

$$E[R | \mathcal{D}_I] = \sum_{i+j>I} E[\alpha_i \beta_j | \mathcal{D}_I]. \quad (12.4.25)$$

Denote the expected reserves, given the parameters $\boldsymbol{\theta}$, by

$$\tilde{R} = E[R | \boldsymbol{\theta}] = \sum_{i+j>I} \alpha_i \beta_j. \quad (12.4.26)$$

Then, the best consistent estimate of reserves (ER) is given by

$$\hat{R}^B = E[\tilde{R} | \mathcal{D}_I] = \sum_{i+j>I} E[\alpha_i \beta_j | \mathcal{D}_I] = E[R | \mathcal{D}_I], \quad (12.4.27)$$

which is, of course, a \mathcal{D}_I -measurable predictor. Hence, the conditional MSEP is simply

$$\text{mse}_{P_{R|\mathcal{D}_I}}(\hat{R}^B) = E\left[\left(R - \hat{R}^B\right)^2 \Big| \mathcal{D}_I\right] = \text{Var}(R | \mathcal{D}_I). \quad (12.4.28)$$

This term, in the Bayesian approach for Tweedie's compound Poisson model, is decomposed

as, see also (12.2.7),

$$\text{Var}(R|\mathcal{D}_I) = \text{Var}\left(\sum_{i+j>I} Y_{i,j} \middle| \mathcal{D}_I\right) = \sum_{i+j>I} E[(\alpha_i \beta_j)^p \phi | \mathcal{D}_I] + \text{Var}\left(\tilde{R} \middle| \mathcal{D}_I\right). \quad (12.4.29)$$

Hence, we obtain the familiar decoupling into average process variance and estimation error. However, in addition we incorporate model uncertainty within Tweedie's compound Poisson model, which enters the calculation by the averaging over all possible values of the variance parameter p .

12.4.4 Random walk Metropolis Hastings-algorithm within Gibbs

In this section we describe an MCMC method to be used to sample from the posterior distribution (12.4.20). The following notations are used: $\theta = (p, \phi, \alpha, \beta)$ is the vector of parameters; $U(a, b)$ is the uniform distribution on the interval (a, b) ; $f_N(x; \mu, \sigma)$ and $F_N(x; \mu, \sigma)$ are the Gaussian density and distribution correspondingly with the mean $\mu \in \mathbb{R}$ and standard deviation $\sigma > 0$ at position $x \in \mathbb{R}$.

Prior Structure: We assume that all parameters are independent under the prior distribution $\pi(\theta)$ and all distributed uniformly with $\theta_i \sim U(a_i, b_i)$. The prior domains we used for our analysis were $p \in (1.1, 1.95)$, $\phi \in (0.01, 100)$, $\alpha_i \in (0.01, 100)$ and $\beta_j \in (0.01, 10^4)$. These are reasonable ranges for the priors in view of our data in Table 12.2 and corresponding to the MLEs in Table 12.3. Other priors such as diffuse priors can be applied with no additional difficulty. The choice of very wide prior supports was made with the aim of performing inference in the setting where the posterior is largely implied by the data. Subsequently, we checked that making the ranges wider does not affect the results.

Next we outline a random walk Metropolis-Hastings (RW-MH) within Gibbs algorithm. This creates a reversible Markov chain with the stationary distribution corresponding to our target posterior distribution (12.4.20). That is, we will run the chain until it has sufficiently converged to the stationary distribution (=posterior distribution) and in doing so we obtain samples from that posterior distribution. It should be noted that the Gibbs sampler creates a Markov chain in which each iteration of the chain involves scanning either deterministically or randomly over the variables that comprise the target stationary distribution of the chain. This process involves sampling each proposed parameter update from the corresponding full conditional posterior distribution. The algorithm we present generates a Markov chain that will explore the parameter space of the model in accordance with the posterior mass in that region of the parameter space. The state of the chain at iteration t will be denoted by θ^t and the chain will be run for a length of T iterations. The manner in which MCMC samplers proceed is by proposing to move the i th parameter from state θ_i^{t-1} to a new proposed state θ_i^* . The latter will be sampled from an MCMC proposal transition kernel (12.4.30). Then the proposed move is accepted according to a

rejection rule which is derived from a reversibility condition. This makes the acceptance probability a function of the transition kernel and the posterior distribution as shown in (12.4.31). If under the rejection rule one accepts the move then the new state of the i th parameter at iteration t is given by $\theta_i^t = \theta_i^*$, otherwise the parameter remains in the current state $\theta_i^t = \theta_i^{t-1}$ and an attempt to move that parameter is repeated at the next iteration. In following this procedure, one builds a set of correlated samples from the target posterior distribution which have several asymptotic properties. One of the most useful of these properties is the convergence of ergodic averages constructed using the Markov chain samples to the averages obtained under the posterior distribution.

Next we present the algorithm and then some references that will guide further investigation into this class of simulation methodology. Properties of this algorithm, including convergence results can be found in the following references Casella and George (1992), Robert and Casella (2004), Gelman *et al.* (1995), Gilks *et al.* (1996) and Smith and Roberts (1993).

Random Walk Metropolis Hastings (RW-MH) within Gibbs algorithm.

1. Initialize randomly or deterministically for $t = 0$ the parameter vector $\boldsymbol{\theta}^0$ (e.g. MLEs).
2. For $t = 1, \dots, T$
 - a) Set $\boldsymbol{\theta}^t = \boldsymbol{\theta}^{t-1}$
 - b) For $i = 1, \dots, 2I + 3$

Sample proposal θ_i^* from Gaussian distribution whose density is truncated below a_i and above b_i and given by

$$f_N^T(\theta_i^*; \theta_i^t, \sigma_{RWi}) = \frac{f_N(\theta_i^*; \theta_i^t, \sigma_{RWi})}{F_N(b_i; \theta_i^t, \sigma_{RWi}) - F_N(a_i; \theta_i^t, \sigma_{RWi})} \quad (12.4.30)$$

to obtain $\boldsymbol{\theta}^* = (\theta_1^t, \dots, \theta_{i-1}^t, \theta_i^*, \theta_{i+1}^{t-1}, \dots)$.

Accept proposal with acceptance probability

$$\alpha(\boldsymbol{\theta}^t, \boldsymbol{\theta}^*) = \min \left\{ 1, \frac{\pi(\boldsymbol{\theta}^* | \mathcal{D}_I) f_N^T(\theta_i^*; \theta_i^t, \sigma_{RWi})}{\pi(\boldsymbol{\theta}^t | \mathcal{D}_I) f_N^T(\theta_i^*; \theta_i^t, \sigma_{RWi})} \right\}, \quad (12.4.31)$$

where $\pi(\boldsymbol{\theta}^* | \mathcal{D}_I)$ is given by (12.4.20). That is, simulate $U \sim U(0, 1)$ and set $\theta_i^t = \theta_i^*$ if $U < \alpha(\boldsymbol{\theta}^t, \boldsymbol{\theta}^*)$.

\Rightarrow Note that in (12.4.31) the normalizing constant of the posterior $\pi(\boldsymbol{\theta} | \mathcal{D}_I)$ from (12.4.20) is not needed.

Remark. The RW-MH algorithm is simple in nature and easily implemented. However, if one does not choose the proposal distribution carefully, then the algorithm only gives a very slow convergence to the stationary distribution. There have been several studies regarding the optimal scaling of proposal distributions to ensure optimal convergence rates. Gelman *et al.* (1997), Bedard and Rosenthal (2007) and Roberts and Rosenthal (2001) were the first authors to publish theoretical results for the optimal scaling problem in RW-MH algorithms with Gaussian proposals. For d -dimensional target distributions with i.i.d. components, the asymptotic acceptance rate optimizing the efficiency of the process is 0.234 independent of the target density. In this case we recommend that the selection of σ_{RW_i} are chosen to ensure that the acceptance probability is roughly close to 0.234. This number is the acceptance probability obtained for asymptotically optimal acceptance rates for RW-MH algorithms when applied to multidimensional target distributions with scaling terms possibly depending on the dimension. To obtain this acceptance rate, one is required to perform some tuning of the proposal variance prior to final simulations. An alternative approach is to utilize a new class of Adaptive MCMC algorithms recently proposed in the literature, see Atchade and Rosenthal (2005) and Rosenthal (2007), but these are beyond the scope of this paper.

12.4.5 Markov chain results and analysis

This section presents the results comparing both MLE and Bayesian estimates for the parameters of Tweedie's compound Poisson model. It is also demonstrated how additional information in a Bayesian framework can be obtained through the complete knowledge of the target posterior distribution obtained from the MCMC algorithm described above. In this regard we demonstrate how this additional information can be exploited in the claims reserving setting to provide alternative statistical analysis not obtainable if one just considers point estimators. We also analyze model averaging solutions in section 5. These can be obtained by forming estimates using the information given by the full posterior distribution $\pi(\theta | \mathcal{D}_T)$ that we find empirically from the MCMC samples.

The maximum likelihood and MCMC algorithms were implemented in Fortran. The maximization routine for the MLEs utilizes the direct search algorithm DBCPOL (that requires function evaluation only) from the IMSL numerical library. Note that, gradient based optimization routines such as the BFGS algorithm can be more efficient, but the direct search algorithm we used was sufficient for our problem in terms of computing time (≈ 4 seconds on a typical desktop PC¹).

The algorithm was analyzed on synthetic data and found to provide correct estimates. In particular with uniform priors the MAP estimates of the parameters are the same as the MLEs, up to numerical errors. This was confirmed for different sized claims triangles. The actual data set studied in this paper is presented in Table 12.2. The data we study is the standard data set used in Wüthrich and Merz (2008) scaled by 10,000.

The results presented for the Bayesian approach were obtained after pre-tuning the Markov

chain random walk standard deviations, σ_{RW_i} , to produce average acceptance probabilities of 0.234. Then the final simulation was for 10^5 iterations from a Markov chain ($\approx 17\text{min}^1$) in which the first 10^4 iterations were discarded as burnin when forming the estimates. The pretuned proposal standard deviations σ_{RW_i} are presented in Table 12.3. The first set of results in Table 12.3 demonstrates the MLE versus the Bayesian posterior estimator MMSE for all model parameters. Included are the [5%, 95%] predictive intervals for the Bayesian posterior distribution. The MLE standard deviations are calculated using (12.4.11). The numerical standard errors (due to a finite number of MCMC iterations) in the Bayesian estimates are obtained by blocking the MCMC samples post burnin into blocks of length 5000 and using the estimates on each block to form the standard error (these are given in brackets next to the estimates).

The next set of analysis demonstrates the performance of the MCMC approach in converging to the stationary distribution given by the target posterior $\pi(\boldsymbol{\theta} \mid \mathcal{D}_I)$. To analyze this, in Figure 12.7.1, we present the trace plots for the Markov chain for the parameters, $(p, \phi, \alpha_1, \beta_0)$. Also, in Figure 12.7.2, we demonstrate the marginal posterior distribution histograms and pair-wise posterior scatter plots for $(p, \phi, \alpha_1, \beta_0, \alpha_I, \beta_I)$. The lower panels in Figure 12.7.2 are the scatter plots for the pair-wise marginal posteriors, the diagonal contains the marginal posteriors and the upper panels contains the correlations between parameters. These plots demonstrate strong linear correlations between several parameters. Some of these correlations are similar to MLE correlations calculated using (12.4.11). For example, we found that under the posterior distribution $\rho(p, \phi) \approx -0.82$ and $\rho(\beta_0, \alpha_1) \approx -0.63$, see Figure 12.7.2, are similar to $\rho(\hat{p}^{\text{MLE}}, \hat{\phi}^{\text{MLE}}) \approx -0.94$ and $\rho(\hat{\beta}_0^{\text{MLE}}, \hat{\alpha}_1^{\text{MLE}}) \approx -0.68$ correspondingly. However, we also observed that under the posterior distribution $\rho(p, \beta_I) \approx -0.17$ and $\rho(\phi, \beta_I) \approx 0.23$, see Figure 12.7.2, while corresponding MLE correlations are zero, see (12.4.13).

12.5 Variable selection via posterior model probabilities

In the development so far it has been assumed that variable selection is not being performed, that is we are assuming that the model is known and we require parameter estimates for this model. This is equivalent to specifying that the number of α and β parameters is fixed and known in advance. We now relax this assumption and will demonstrate how the variable selection problem can be incorporated into our framework. The procedure we utilize for the variable selection is based on recent work of Congdon (2006) and specifies the joint support of the posterior distribution for the models and parameters under the product space formulation of Carlin and Chib (1995).

In this section we consider the subset of nested models which create homogeneous blocks in the claims reserving triangle ($I = 9$) for the data set in Table 2.

- $M_0 : \boldsymbol{\theta}_{[0]} = (p, \phi, \tilde{\alpha}_0 = \alpha_0, \dots, \tilde{\alpha}_I = \alpha_I, \tilde{\beta}_0 = \beta_0, \dots, \tilde{\beta}_I = \beta_I)$ - **saturated model**.

¹ Intel® Core™2 Duo, 2.13GHz processor.

- $M_1 : \boldsymbol{\theta}_{[1]} = (p, \phi, \tilde{\beta}_0)$ with $(\tilde{\beta}_0 = \beta_0 = \dots = \beta_I), (\alpha_0 = \dots = \alpha_I = 1)$.
- $M_2 : \boldsymbol{\theta}_{[2]} = (p, \phi, \tilde{\alpha}_1, \tilde{\beta}_0, \tilde{\beta}_1)$ with $(\alpha_0 = \dots = \alpha_4 = 1), (\tilde{\alpha}_1 = \alpha_5 = \dots = \alpha_I),$
 $(\tilde{\beta}_0 = \beta_0 = \dots = \beta_4), (\tilde{\beta}_1 = \beta_5 = \dots = \beta_I)$.
- $M_3 : \boldsymbol{\theta}_{[3]} = (p, \phi, \tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2)$ with $(\alpha_0 = \alpha_1 = 1), (\tilde{\alpha}_1 = \alpha_2 = \dots = \alpha_5),$
 $(\tilde{\alpha}_2 = \alpha_6 = \dots = \alpha_I), (\tilde{\beta}_0 = \beta_0 = \beta_1), (\tilde{\beta}_1 = \beta_2 = \dots = \beta_5), (\tilde{\beta}_2 = \beta_6 = \dots = \beta_I)$.
- $M_4 : \boldsymbol{\theta}_{[4]} = (p, \phi, \tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\alpha}_3, \tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2, \tilde{\beta}_3)$ with $(\alpha_0 = \alpha_1 = 1), (\tilde{\alpha}_1 = \alpha_2 = \alpha_3),$
 $(\tilde{\alpha}_2 = \alpha_4 = \alpha_5 = \alpha_6), (\tilde{\alpha}_3 = \alpha_7 = \alpha_8 = \alpha_I), (\tilde{\beta}_0 = \beta_0 = \beta_1), (\tilde{\beta}_1 = \beta_2 = \beta_3),$
 $(\tilde{\beta}_2 = \beta_4 = \beta_5 = \beta_6), (\tilde{\beta}_3 = \beta_7 = \beta_8 = \beta_I)$.
- $M_5 : \boldsymbol{\theta}_{[5]} = (p, \phi, \tilde{\alpha}_1, \tilde{\alpha}_2, \tilde{\alpha}_3, \tilde{\alpha}_4, \tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2, \tilde{\beta}_3, \tilde{\beta}_4)$ with $(\alpha_0 = \alpha_1 = 1), (\tilde{\alpha}_1 = \alpha_2 = \alpha_3),$
 $(\tilde{\alpha}_2 = \alpha_4 = \alpha_5), (\tilde{\alpha}_3 = \alpha_6 = \alpha_7), (\tilde{\alpha}_4 = \alpha_8 = \alpha_I), (\tilde{\beta}_0 = \beta_0 = \beta_1), (\tilde{\beta}_1 = \beta_2 = \beta_3),$
 $(\tilde{\beta}_2 = \beta_4 = \beta_5), (\tilde{\beta}_3 = \beta_6 = \beta_7), (\tilde{\beta}_4 = \beta_8 = \beta_I)$.
- $M_6 : \boldsymbol{\theta}_{[6]} = (p, \phi, \alpha_0, \tilde{\alpha}_1, \beta_0, \beta_1, \dots, \beta_I)$ with $(\tilde{\alpha}_1 = \alpha_1 = \dots = \alpha_I)$.

Now, to determine the optimal model, we first consider the joint posterior distribution for the model probability and the model parameters denoted $\pi(M_k, \boldsymbol{\theta}_{[k]} | \mathcal{D}_I)$, where $\boldsymbol{\theta}_{[k]} = (\tilde{\theta}_{1,[k]}, \tilde{\theta}_{2,[k]}, \dots, \tilde{\theta}_{N_{[k]},[k]})$ is the parameter vector for model $[k]$. Additionally we denote the prior bounds for $\tilde{\theta}_{i,[k]}$ as $[a_{\tilde{\theta}_{i,[k]}}, b_{\tilde{\theta}_{i,[k]}}]$. We assume a prior distribution $\pi(M_k)$ for the model selection and a prior for the parameters conditional on the model $\pi(\boldsymbol{\theta}_{[k]} | M_k)$. It is no longer possible to run the standard MCMC procedure we described in section 3.4 for this variable selection setting. This is because the posterior is now defined on either a support consisting of disjoint unions of subspaces or a product space of all such subspaces, one for each model considered. A popular approach to run Markov chains in such a situation is to develop a more advanced sampler than that presented above, typically in the disjoint union setting. This involves developing a Reversible Jump RJ-MCMC framework, see Green (1995) and the references therein. This type of Markov chain sampler is complicated to develop and analyze. Hence, we propose as an alternative in this paper to utilize a recent procedure that will allow us to use the above MCMC sampler we have already developed for a model M_k . The process we must follow involves first running the sampler in the simulation technique described in section 3.4 for each model considered. Then the calculation of the posterior model probabilities $\pi(M_k | \mathcal{D}_I)$ is performed using the samples from the Markov chain in each model to estimate (12.5.3).

Furthermore, our approach here removes the assumption on the priors across models, made by Congdon (2006), p.348,

$$\pi(\boldsymbol{\theta}_{[m]} | M_k) = 1, m \neq k \quad (12.5.1)$$

and instead we work with the prior

$$\pi(\boldsymbol{\theta}_{[m]} | M_k) = \prod_{i=1}^{N_{[m]}} \left[b_{\tilde{\theta}_{i,[m]}} - a_{\tilde{\theta}_{i,[m]}} \right]^{-1}, m \neq k. \quad (12.5.2)$$

That is, instead we use a class of priors where specification of priors for a model M_k automatically specifies priors for any other model. This is a sensible set of priors to consider given our product space formulation and it has a clear interpretation in our setting where we specify our models through a series of constraints, relative to each other. In doing this we also achieve our goal of having posterior model selection insensitive to the choice of the prior and being data driven. The modified version of Congdon's (2006), formula A.3, we obtain after relaxing Congdon's assumption, allows the calculation of the posterior model probabilities $\pi(M_k | \mathcal{D}_I)$ using the samples from the Markov chain in each model to estimate

$$\begin{aligned} \pi(M_k | \mathcal{D}_I) &= \int \pi(M_k, \boldsymbol{\theta}_{[k]} | \mathcal{D}_I) d\boldsymbol{\theta}_{[k]} = \int \pi(M_k | \boldsymbol{\theta}_{[k]}, \mathcal{D}_I) \pi(\boldsymbol{\theta}_{[k]} | \mathcal{D}_I) d\boldsymbol{\theta}_{[k]} \\ &\approx \frac{1}{T - T_b} \sum_{j=T_b+1}^T \pi(M_k | \mathcal{D}_I, \boldsymbol{\theta}_{j,[k]}) \\ &= \frac{1}{T - T_b} \sum_{j=T_b+1}^T \frac{L_{\mathcal{D}_I}(M_k, \boldsymbol{\theta}_{j,[k]}) \prod_{k=0}^K \pi(\boldsymbol{\theta}_{j,[k]} | M_k) \pi(M_k)}{\sum_{m=0}^K L_{\mathcal{D}_I}(M_m, \boldsymbol{\theta}_{j,[m]}) \prod_{k=0}^K \pi(\boldsymbol{\theta}_{j,[k]} | M_m) \pi(M_m)} \\ &= \frac{1}{T - T_b} \sum_{j=T_b+1}^T \frac{L_{\mathcal{D}_I}(M_k, \boldsymbol{\theta}_{j,[k]})}{\sum_{m=0}^K L_{\mathcal{D}_I}(M_m, \boldsymbol{\theta}_{j,[m]})}. \end{aligned} \quad (12.5.3)$$

Here $K = 6$, and for a proof, see Congdon (2006), formula A.3. Note that, the prior of parameters (given model) contributes in the above implicitly as $\boldsymbol{\theta}_{j,[k]}$ are MCMC samples from the k^{th} models posterior distribution. In the actual implementation we used $T = 100,000$ and the burnin period $T_b = 10,000$. Note, the prior probabilities for each model are considered diffuse and are set such that all models *a priori* are equiprobable, hence $\pi(M_k) = 1/(K + 1)$ and $\pi(\boldsymbol{\theta}_{j,[k]} | M_k)$ is the prior for model M_k 's parameters evaluated at the j^{th} Markov chain iteration. Once we have the posterior model probabilities we can then take the MAP estimate for the optimal model (variable selection) for the given data set. In this paper we do not consider the notion of model averaging over different parameterized models in the variable selection context. Instead we simply utilize these results for optimal variable selection from a MAP perspective for the marginal posterior $\pi(M_k | \mathcal{D}_I)$.

In addition to this model selection criterion we also consider in the Bayesian framework the Deviance Information Criterion (DIC), see Bernardo and Smith (1994). From a classical maximum likelihood perspective we present the likelihood ratio (LHR) p-values.

Application of this technique to the simulated MCMC samples for each of the considered models produced the posterior model probabilities given in Table 12.4. This suggests that within this subset of models considered, the saturated model M_0 was the optimal model to utilize in

the analysis of the claims reserving problem, $\pi(M_0 | \mathcal{D}_I) \approx 0.7$. It is followed by model M_6 with $\pi(M_0 | \mathcal{D}_I) \approx 0.3$. Additionally, the choice of M_0 was also supported by the other criteria we considered: DIC and LHR.

In future research it would be interesting to extend to the full model space which considers all models in the power set $|\theta_{[0]}|$. This is a large set of models including all combinatorial combinations of model parameters for α 's and β 's. In such cases it is no longer feasible to run standard MCMC algorithms in each model since this will involve an impractical number of simulations. Hence, more sophisticated model exploration techniques will be required such as RJ-MCMC, see Green (1995) or the product space samplers of Carlin and Chib (1995).

We note here that we do not claim M_0 is the optimal model in all possible models, only in the subset we consider in this section. In saying this we acknowledge that we aim to work in the saturated model but consider it important to illustrate how variable selection can be performed in this class of models and also raise awareness that this will impact the model uncertainty analysis subsequently performed.

Hence, using these findings and the analysis of the MCMC results for model M_0 provided above, we may now proceed to analyze the claims reserving problem. Of interest to the aim of this paper is the sensitivity of the model choice parameter p to the parameterization of the claims reserving triangle. This is particularly evident when one considers the MMSE estimate of the model specification parameter p estimated under each model. In the most parsimonious, yet inflexible model M_1 the estimate obtained was $MMSE(p) \approx 1.9$, a very similar estimate was obtained in models M_2, M_3, M_4 and M_5 , however, interestingly in the saturated model the estimate was $MMSE(p) \approx 1.3$ which is almost at the other extreme of the considered range for which the parameter p is defined.

12.6 Calculation of the claims reserves

We now demonstrate the results for several quantities in the claims reserving setting, utilizing the MCMC simulation results we obtained for the Bayesian posterior distribution under the variable selection model M_0 (saturated model). In particular, we start by noting that we use uniform prior distributions with a very wide ranges to perform inference implied by the data only. In this case, theoretically, the Bayesian MAP (the posterior mode) and MLEs for the parameters should be identical up to numerical error due to the finite number of MCMC iterations. A large number of MCMC iterations was performed so that the numerical error is not material. In general, the use of more informative priors will lead to the differences between the MAP and MLE. Some of the MMSE estimates (the posterior mean) were close to the MAP estimates, indicating that the marginal posterior distributions are close to symmetric. When the posterior is not symmetric, MMSE and MAP can be very different. Also, note that the uncertainties in the parameter MLEs are estimated using the asymptotic Gaussian approximation (12.4.11)-(12.4.12). In the case of constant priors, this should lead to the same inferences as corresponding Bayesian estimators if the posterior distributions are close to the Gaussian

approximation, see (12.4.23)-(12.4.24). In addition, the MLEs for the reserves, estimation error and process variance, see section 3.2, are based on a Taylor expansion around parameter MLEs assuming small errors. In many cases the posterior is materially different from the Gaussian distribution, has significant skewness and large standard deviation leading to the differences between the MLEs and corresponding Bayesian estimators. Having mentioned this, we now focus on the main point of this paper which involves analysis of the quantities in Table 12.5 related to the model uncertainty within Tweedie's compound Poisson models (introduced by fixing model parameter p) in a Bayesian setting.

It is worth noting that point estimates of model parameters are either in the frequentists approach MLEs or in a Bayesian approach the MAP or MMSE estimates. These are under the auspice that we wish to perform model selection (i.e. selection of p). The focus of this paper is to demonstrate the difference in results obtained for reserve estimates that can arise by performing model averaging instead of the typical approach of model selection, using *a priori* chosen p . In this regard we perform estimation utilizing the full posterior distribution of the parameters and not just point estimators. This allows us to capture the influence of the model uncertainty (uncertainty in p), since in a Bayesian setting we can account for this uncertainty using the posterior distribution. In particular, the Bayesian analysis specifies the optimal p (either in the MAP or the MMSE context) and it also provides a confidence interval for the choice of p (see Figure 12.7.7), which corresponds to the choice of the optimal model within Tweedie's compound Poisson models. Moreover, we demonstrate the impact on the claims reserve by varying p from 1.1 to 1.9 (i.e. for a fixed model choice).

12.6.1 Results: average over p

Initially it is worth considering the predicted reserve distribution for the estimator \tilde{R} . This is obtained by taking the samples $t = 10,001$ to $100,000$ from the MCMC simulation $\{p^t, \phi^t, \alpha^t, \beta^t\}$ and calculating $\{\tilde{R}^t\}$ via (12.4.26). The histogram estimate is presented in Figure 12.7.3. In the same manner, we also estimate the distributions of $\tilde{R}_{i,j} = \alpha_i \beta_j$ for the individual cells of the $I \times I$ claims matrix, presented as subplots in Figure 12.7.4. Note that the total observed loss in the upper triangle (≈ 9274) is consistent with $E[\sum_{i+j \leq I} \alpha_i \beta_j]$ and $[\text{Var}(\sum_{i+j \leq I} \alpha_i \beta_j)]^{1/2}$ estimated using the MCMC samples as (≈ 9311) and (≈ 190) respectively. The maximum likelihood approach results in $\sum_{i+j \leq I} \hat{\alpha}_i^{MLE} \hat{\beta}_j^{MLE} \approx 9275$ with standard deviation ≈ 124 also conforming with the observed total loss.

Now we focus on quantities associated with the estimated distribution for \tilde{R} to calculate the results, see Table 12.5, which can only be estimated once the entire posterior distribution is considered. These quantities are the key focus of this paper since they allow assessment of the conditional MSE as specified in (12.4.28). In particular, we may now easily use the posterior probability samples obtained from the MCMC algorithm to evaluate the estimated reserve (ER), the process variance (PV) and the estimation error (EE) in the conditional MSE. This provides an understanding and analysis of the behavior of the proposed model in both the model av-

eraging and model selection (i.e. selection of p) contexts whilst considering the issue of model uncertainty, the goal of this paper. The Bayesian estimates for ER, PV, EE and MSEP are presented in Table 12.6. The corresponding MLEs were calculated using (12.4.8), (12.4.16), (12.4.15) and (12.4.17) respectively and presented in Table 12.6 for comparison. The results demonstrate the following:

- Claims reserves MLE, \hat{R}^{MLE} , is less than Bayesian estimate \hat{R}^{B} by approximately 3%, which is the estimation bias of the claims reserve MLE (see also Wüthrich and Merz (2008), Remarks 6.15).
- \sqrt{EE} and \sqrt{PV} are of the same magnitude, approximately 6-7% of the total claims reserves.
- MLEs for \sqrt{EE} and \sqrt{PV} are less than corresponding Bayesian estimates by approximately 37% and 30%, respectively.
- The difference between \hat{R}^{MLE} and \hat{R}^{B} is of the same order of magnitude as \sqrt{EE} and \sqrt{PV} and thus is significant.

Note that we use constant priors with very wide ranges, the MLE uncertainties are calculated using an asymptotic Gaussian approximation and numerical error due to the finite number of MCMC iterations is not material (also see the 1st paragraph, section 12.6). The observed significant differences between the MLEs and corresponding Bayesian estimators suggest that our posterior distributions are skewed and materially different from the Gaussian distribution.

We conclude this section with the distribution of R , the total outstanding claims payment, see Figure 12.7.5. This is obtained from the MCMC samples of the parameters (p, ϕ, α, β) which we then transform to parameters (λ, γ, τ) from model representation 1, section 12.3, and simulate annual losses in $i + j > I$. That is, these samples of R are obtained from the full predictive distribution $f(R | \mathcal{D}_I) = \int g(R | \theta) \pi(\theta | \mathcal{D}_I) d\theta$, where $g(R | \theta)$ is the distribution of R given by (12.2.3) and (12.3.1). It takes into account both process uncertainty and parameter uncertainty. We note that while reserving by some measure of centrality such as \hat{R}^{B} may be robust, it will not take into account the distributional shape of R . A viable alternative may be Value-at-Risk (VaR) or a coherent risk measure such as Expected Shortfall. In Table 12.7 we demonstrate estimates of the VaR for \tilde{R} and R at the 75%, 90% and 95% quantiles.

12.6.2 Results: conditioning on p

As part of the model uncertainty analysis, it is useful to present plots of the relevant quantities in the model selection (selection of p) settings, see Figure 12.7.6, where we present $ER_p = E[\tilde{R} | \mathcal{D}_I, p]$, $PV_p = \sum_{i+j>I} E[\phi(\alpha_i \beta_j)^p | \mathcal{D}_I, p]$ and $EE_p = \text{Var}(\tilde{R} | \mathcal{D}_I, p)$ as a function of p . Figure 12.7.6 shows:

- MLE of ER_p is almost constant, varying approximately from a maximum of 603.96 ($p = 1.1$) to a minimum of 595.78 ($p = 1.9$) while the MLE for ER was 602.63.
- The Bayesian estimates for ER_p change as a function of p . Approximately, it ranged from a maximum of 646.4 ($p = 1.9$) to a minimum of 621.1 ($p = 1.5$) while the Bayesian estimator for ER was 624.1. Hence, the difference (estimation bias) within this possible model range is ≈ 25 which is of a similar order as the process uncertainty and the estimation error.
- Bayesian estimators for $\sqrt{PV_p}$ and $\sqrt{EE_p}$ increase as p increases approximately from 33.1 to 68.5 and from 37.4 to 102.0 respectively, while the Bayesian estimators for \sqrt{PV} and \sqrt{EE} are 37.3 and 44.8 correspondingly. Hence, the resulting risk measure strongly varies in p which has a large influence on quantitative solvency requirements. The MLEs for PV_p and EE_p are significantly less than the corresponding Bayesian estimators. Also, the difference between the MLE and the Bayesian estimators increases as p increases.

For interpretation purposes of the above results it is helpful to use the following relations between model averaging and model selection quantities (easily derived from their definitions in Table 12.5):

$$ER = E[ER_p | \mathcal{D}_I], \quad (12.6.1)$$

$$PV = E[PV_p | \mathcal{D}_I], \quad (12.6.2)$$

$$EE = E[EE_p | \mathcal{D}_I] + \text{Var}(ER_p | \mathcal{D}_I). \quad (12.6.3)$$

Here, the expectations are calculated with respect to the posterior distribution of p . The histogram estimate of the later is presented in Figure 12.7.7 and highlights significant uncertainty in p (model uncertainty within Tweedie's compound Poisson model).

We also provide Figure 12.7.8 demonstrating a Box and Whisker summary of the distributions of $\tilde{R} | p$ for a range of values of p . This plot provides the first, second and third quartiles as the box. The notch represents uncertainty in the median estimate for model comparison, across values of p , and the whiskers demonstrate the smallest and largest data points not considered as outliers. The outliers are included as crosses and the decision rule to determine if a point is an outlier was taken as the default procedure from the statistical software package R.

The conclusion from this section is that if model selection is performed (i.e. p is fixed by the modeller), the conditional MSE_P will increase significantly if a poor choice of the model parameter p is made. In particular, though the median is fairly constant for the entire range of $p \in (1, 2)$ the shape of the distribution of $\tilde{R} | p$ is clearly becoming more diffuse as $p \rightarrow 2$. This will lead to significantly larger variance in the reserve estimate. If risk measures such as Value-at-Risk are used in place of the mean, it will result in reserves which are too conservative (if a poor choice of p is made). Also, using the maximum likelihood approach may significantly underestimate the claims reserves and associated uncertainties.

12.6.3 Overdispersed Poisson and Gamma models

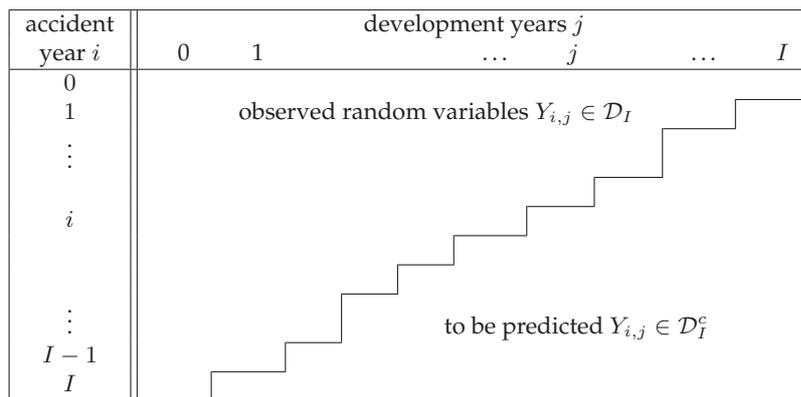
There are several popular claims reserving models, however we restrict our comparison to the overdispersed Poisson and gamma models since they fit into Tweedie's compound Poisson framework when $p = 1$ and $p = 2$ respectively. Note that the overdispersed Poisson model and several other stochastic models lead to the same reserves as the chain ladder method but different in higher moments. The detailed treatment of these models can be found in e.g. England and Verrall (2002) or Wüthrich and Merz (2008), section 3.2.

The MLEs for the reserves and associated uncertainties within the overdispersed Poisson and gamma models are provided in Table 12.8. These results are obtained when the dispersion ϕ is estimated by $\hat{\phi}^P$ using Pearson's residuals (12.4.18) and when ϕ is estimated by $\hat{\phi}^{\text{MLE}}$ obtained from the maximization of the likelihood. The results for the first case are also presented in Wüthrich and Merz (2008), Table 6.4. Firstly note that, the values of $\hat{\phi}^P$ and $\hat{\phi}^{\text{MLE}}$ are significantly different both for the overdispersed Poisson and gamma models. As we mentioned in section 3.2, for a fixed p , the MLE for the reserves does not depend on ϕ while the estimation error, process variance and MSEF are proportional to ϕ . As one can see from Table 12.8, different estimators for the dispersion ϕ lead to the same estimators for the reserves but very different estimators for the uncertainties. Also note that, our MLE calculations for Tweedie's distribution conditional on p , i.e. Figure 12.7.6, are obtained using $\hat{\phi}^{\text{MLE}}$ and are consistent with the corresponding results for the overdispersed Poisson and Gamma models when $p \rightarrow 1$ and $p \rightarrow 2$ respectively. Though, in the case of the overdispersed Poisson we had to use an extended quasi-likelihood to estimate $\hat{\phi}^{\text{MLE}}$. In Figure 12.7.6, we do not show the results based on $\hat{\phi}^P$ but would like to mention that these are always above the MLEs and below the Bayesian estimators for the process variance and estimation error and are consistent with corresponding overdispersed Poisson and gamma model limits. Interestingly, the ratio $\hat{\phi}^P / \hat{\phi}^{\text{MLE}}$ is approximately 1.4 – 1.5 for all considered cases of p within a range[1, 2].

The MLEs obtained using both $\hat{\phi}^{\text{MLE}}$ and $\hat{\phi}^P$ underestimate the uncertainties compared to the Bayesian analysis. Note that, while the MLEs for the uncertainties are proportional to the dispersion estimator, the corresponding Bayesian estimators are averages over all possible values of ϕ according to its posterior distribution. The uncertainty in the estimate for the dispersion is large which is also highlighted by a bootstrap analysis in Wüthrich and Merz (2008), section 7.3. This indicates that ϕ should also depend on the individual cells (i, j) . However, in this case overparameterization needs to be considered with care and Bayesian framework should be preferred.

12.7 Discussion

The results demonstrate the development of a Bayesian model for the claims reserving problem when considering Tweedie's compound Poisson model. The sampling methodology of a Gibbs sampler is applied to the problem to study the model sensitivity for a real data set. The prob-



Tab. 12.1: Claims development triangle.

lem of variable selection is addressed in a manner commensurate with the MCMC sampling procedure developed in this paper and the most probable model under the posterior marginal model probability is then considered in further analysis. Under this model we then consider two aspects, model selection and model averaging with respect to model parameter p . The outcomes from these comparisons demonstrate that the model uncertainty due to fixing p plays a significant role in the evaluation of the claims reserves and its conditional MSEP. It is clear that whilst the frequentist MLE approach is not sensitive to a poor model selection, the Bayesian estimates demonstrate more dependence on poor model choice, with respect to model parameter p . We use constant priors with very wide ranges to perform inference in the setting where the posterior is largely implied by data only. Also, we run a large number of MCMC iterations so that numerical error in the Bayesian estimators is very small. In the case of the data we studied, the MLEs for the claims reserve, process variance and estimation error were all significantly different (less) than corresponding Bayesian estimators. This is due to the fact that the posterior distribution implied by the data and estimated using MCMC is materially different from Gaussian, i.e. more skewed.

Future research will examine variable selection aspects of this model in a Bayesian context considering the entire set of possible parameterizations. This requires development of advanced approaches such as Reversible Jump MCMC and variable selection stochastic optimization methodology to determine if a more parsimonious model can be selected under assumptions of homogeneity in adjacent columns/rows in the claims triangle.

Acknowledgements

The first author is thankful to the Department of Mathematics and Statistics at the University of NSW for support through an Australian Postgraduate Award and to CSIRO for support through a postgraduate research top up scholarship. Thank you also goes to Robert Kohn for discussions.

| Year | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|------|----------|----------|---------|---------|---------|---------|--------|--------|--------|--------|
| 0 | 594.6975 | 372.1236 | 89.5717 | 20.7760 | 20.6704 | 6.2124 | 6.5813 | 1.4850 | 1.1130 | 1.5813 |
| 1 | 634.6756 | 324.6406 | 72.3222 | 15.1797 | 6.7824 | 3.6603 | 5.2752 | 1.1186 | 1.1646 | |
| 2 | 626.9090 | 297.6223 | 84.7053 | 26.2768 | 15.2703 | 6.5444 | 5.3545 | 0.8924 | | |
| 3 | 586.3015 | 268.3224 | 72.2532 | 19.0653 | 13.2976 | 8.8340 | 4.3329 | | | |
| 4 | 577.8885 | 274.5229 | 65.3894 | 27.3395 | 23.0288 | 10.5224 | | | | |
| 5 | 618.4793 | 282.8338 | 57.2765 | 24.4899 | 10.4957 | | | | | |
| 6 | 560.0184 | 289.3207 | 56.3114 | 22.5517 | | | | | | |
| 7 | 528.8066 | 244.0103 | 52.8043 | | | | | | | |
| 8 | 529.0793 | 235.7936 | | | | | | | | |
| 9 | 567.5568 | | | | | | | | | |

Tab. 12.2: Data - annual claims payments $Y_{i,j}$ for each accident year i and development year j , $i + j \leq 9$.

| | MLE | MLE stdev | Bayesian posterior | | | σ_{RW} |
|------------|-------|-----------|--------------------|---------------|------------------------|---------------|
| | | | MMSE | stdev | $[Q_{0.05}; Q_{0.95}]$ | |
| p | 1.259 | 0.149 | 1.332 (0.007) | 0.143 (0.004) | [1.127;1.590] | 1.61 |
| ϕ | 0.351 | 0.201 | 0.533 (0.013) | 0.289 (0.005) | [0.174;1.119] | 1.94 |
| α_1 | 0.918 | 0.056 | 0.901 (0.004) | 0.074 (0.001) | [0.778;1.022] | 0.842 |
| α_2 | 0.946 | 0.051 | 0.946 (0.003) | 0.073 (0.001) | [0.833;1.072] | 0.907 |
| α_3 | 0.861 | 0.048 | 0.861 (0.003) | 0.068 (0.001) | [0.756;0.977] | 0.849 |
| α_4 | 0.891 | 0.049 | 0.902 (0.003) | 0.072 (0.002) | [0.794;1.027] | 0.893 |
| α_5 | 0.879 | 0.051 | 0.876 (0.003) | 0.070 (0.001) | [0.768;0.994] | 0.932 |
| α_6 | 0.842 | 0.048 | 0.843 (0.002) | 0.069 (0.001) | [0.736;0.958] | 0.751 |
| α_7 | 0.762 | 0.046 | 0.762 (0.003) | 0.066 (0.001) | [0.660;0.876] | 0.888 |
| α_8 | 0.763 | 0.047 | 0.765 (0.003) | 0.067 (0.001) | [0.661;0.874] | 0.897 |
| α_9 | 0.848 | 0.059 | 0.856 (0.003) | 0.090 (0.002) | [0.716;1.009] | 1.276 |
| β_0 | 669.1 | 27.7 | 672.7 (2.1) | 39.7 (0.7) | [610.0;740.0] | 296 |
| β_1 | 329.0 | 14.4 | 331.1 (1.0) | 20.6 (0.4) | [298.1;365.9] | 190 |
| β_2 | 77.43 | 4.38 | 78.06 (0.24) | 6.10 (0.06) | [68.58;88.29] | 75.4 |
| β_3 | 24.59 | 1.96 | 24.95 (0.08) | 2.64 (0.03) | [20.89;29.64] | 40.9 |
| β_4 | 16.28 | 1.55 | 16.65 (0.05) | 2.09 (0.03) | [13.44;20.30] | 40.6 |
| β_5 | 7.773 | 1.028 | 8.068 (0.024) | 1.356 (0.020) | [6.064;10.473] | 26.0 |
| β_6 | 5.776 | 0.937 | 6.115 (0.022) | 1.261 (0.016) | [4.246;8.347] | 24.1 |
| β_7 | 1.219 | 0.396 | 1.494 (0.006) | 0.609 (0.013) | [0.739;2.609] | 13.1 |
| β_8 | 1.188 | 0.476 | 1.622 (0.008) | 0.802 (0.016) | [0.674;3.070] | 15.1 |
| β_9 | 1.581 | 0.790 | 2.439 (0.021) | 1.496 (0.026) | [0.829;5.250] | 32.1 |

Tab. 12.3: MLE and Bayesian estimators. σ_{RW} is the proposal standard deviation in the MCMC algorithm and $[Q_{0.05}; Q_{0.95}]$ is the predictive interval, where Q_α is the quantile of the posterior distribution at level α . The numerical standard error, in Bayesian estimators due to finite number of MCMC iterations, is included in brackets next to estimates.

| | M_0 | M_1 | M_2 | M_3 | M_4 | M_5 | M_6 |
|------------------|-------|----------|----------|----------|----------|----------|-------|
| $\pi(M_k D_I)$ | 0.71 | 4.19E-54 | 3.04E-43 | 1.03E-28 | 6.71E-20 | 2.17E-21 | 0.29 |
| DIC | 399 | 649 | 600 | 535 | 498 | 507 | 398 |
| LHR p - value | 1 | 2.76E-50 | 1.67E-40 | 3.53E-28 | 5.78E-21 | 3.03E-23 | 0.043 |

Tab. 12.4: Posterior model probabilities $\pi(M_k | D_I)$, Deviance Information Criterion (DIC) for variable selection models M_0, \dots, M_6 and Likelihood Ratio (LHR) p-values (comparing M_0 to M_1, \dots, M_6).

| | Model Averaging | Model Selection for p |
|--------------------|---|--|
| Estimated Reserves | $ER = \hat{R}^B = E[\hat{R} D_I]$ | $ER_p = E[\hat{R} D_I, p]$ |
| Process Variance | $PV = E[\sum \phi(\alpha_i \beta_j)^p D_I]$ | $PV_p = E[\sum \phi(\alpha_i \beta_j)^p D_I, p]$ |
| Estimation Error | $EE = \text{Var}(\hat{R} D_I)$ | $EE_p = \text{Var}(\hat{R} D_I, p)$ |

Tab. 12.5: Quantities used for analysis of the claims reserving problem under Model Averaging and Model Selection in respect to p .

| Model Averaging | | |
|-----------------|-------------------|--------------|
| Statistic | Bayesian Estimate | MLE Estimate |
| ER | 624.1 (0.7) | 602.630 |
| \sqrt{PV} | 37.3 (0.2) | 25.937 |
| \sqrt{EE} | 44.8 (0.5) | 28.336 |
| \sqrt{MSEP} | 58.3(0.5) | 38.414 |

Tab. 12.6: Model averaged estimates of claim reserve, process variance and estimation error. Numerical error in Bayesian estimates is reported in brackets. See Table 12.5 for definitions of ER, PV, EE and $MSEP=EE+PV$.

| Model Averaging | | |
|-----------------|-------------|-------------|
| VaR_q | R | \tilde{R} |
| $VaR_{75\%}$ | 659.8 (0.9) | 650.6 (1.0) |
| $VaR_{90\%}$ | 698.4 (1.2) | 680.4 (1.3) |
| $VaR_{95\%}$ | 724.0 (1.5) | 701.7 (1.6) |

Tab. 12.7: Bayesian model averaged estimates of Value at Risk for outstanding claims payment R and claim reserves \tilde{R} .

| Statistic | Overdispersed Poisson | | Gamma model | |
|-----------------|------------------------------|----------------------------------|------------------------------|----------------------------------|
| | $\hat{\phi}^P \approx 1.471$ | $\hat{\phi}^{MLE} \approx 0.954$ | $\hat{\phi}^P \approx 0.045$ | $\hat{\phi}^{MLE} \approx 0.031$ |
| ER_p | 604.706 | 604.706 | 594.705 | 594.705 |
| $\sqrt{PV_p}$ | 29.829 | 24.017 | 62.481 | 52.162 |
| $\sqrt{EE_p}$ | 30.956 | 24.925 | 92.826 | 77.496 |
| $\sqrt{MSEP_p}$ | 42.989 | 34.613 | 111.895 | 93.415 |

Tab. 12.8: The MLEs for the overdispersed Poisson ($p = 1$) and Gamma ($p = 2$) models, when the dispersion ϕ is estimated as $\hat{\phi}^P$ using Pearson’s residuals (12.4.18) or $\hat{\phi}^{MLE}$.

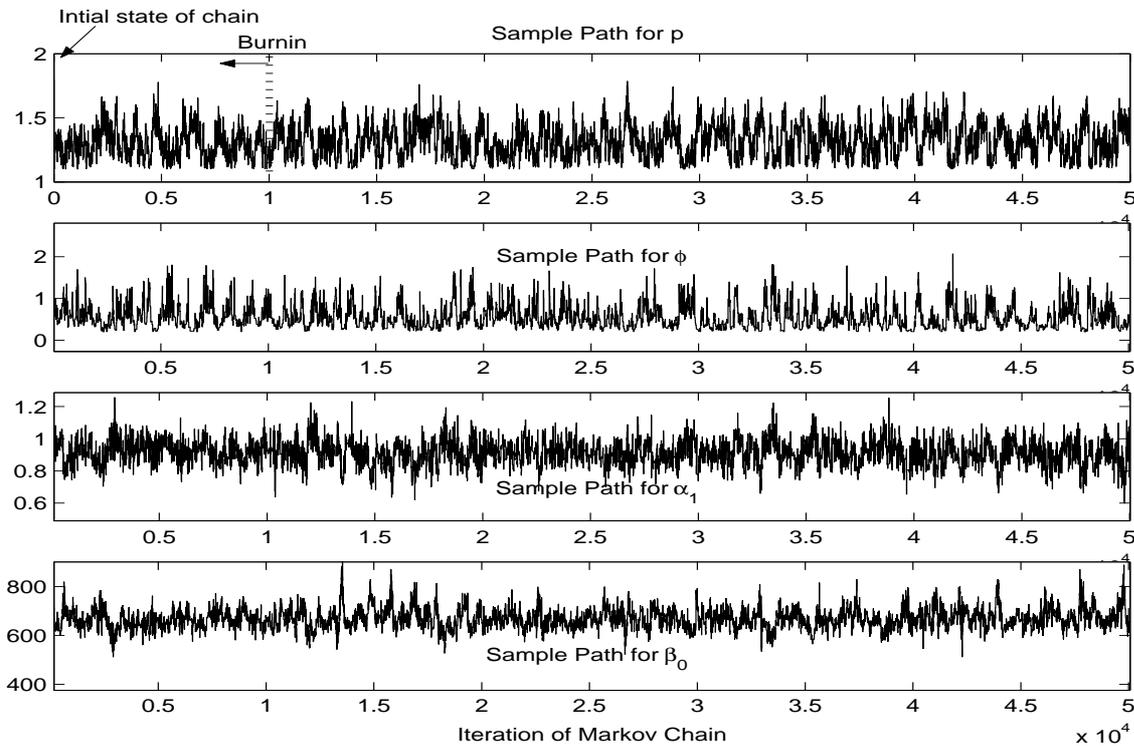


Fig. 12.7.1: Markov chain sample paths $(p, \phi, \alpha_1, \beta_0)$.

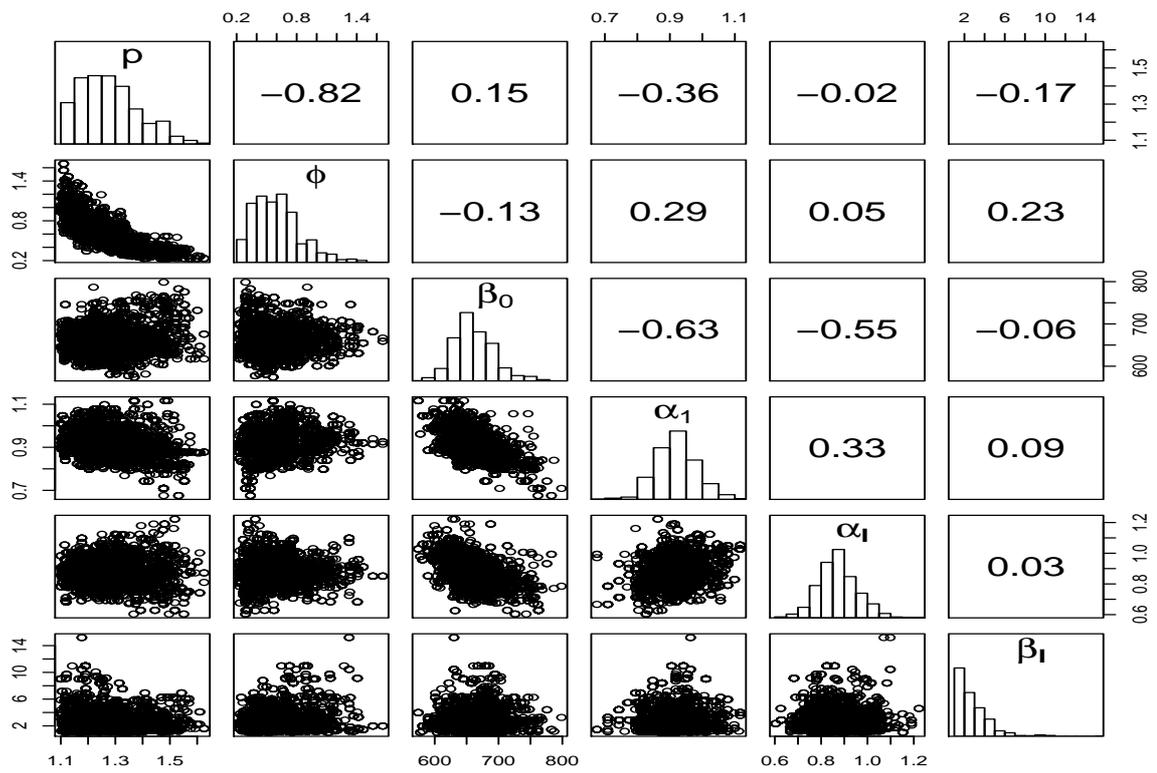


Fig. 12.7.2: Posterior scatter plots, marginal posterior histograms and linear correlations for $(p, \phi, \alpha_1, \beta_0, \alpha_I, \beta_I)$.

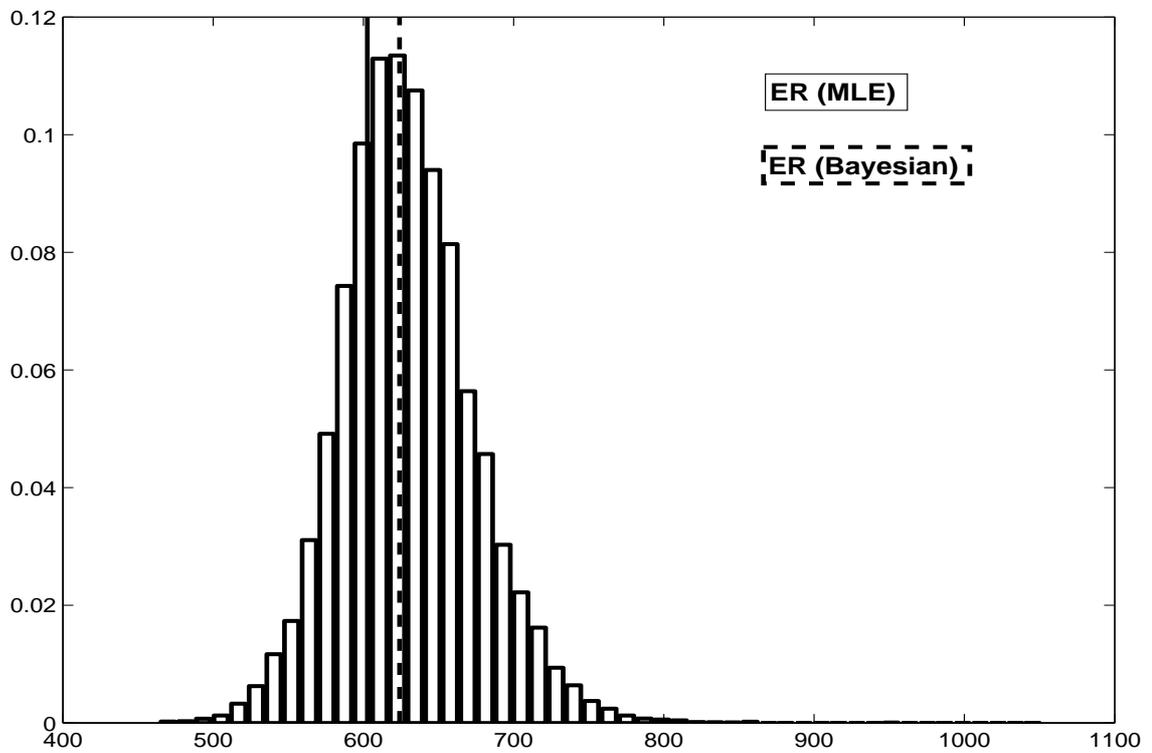


Fig. 12.7.3: Predicted distribution of reserves, $\tilde{R} = \sum_{i+j>I} \alpha_i \beta_j$.

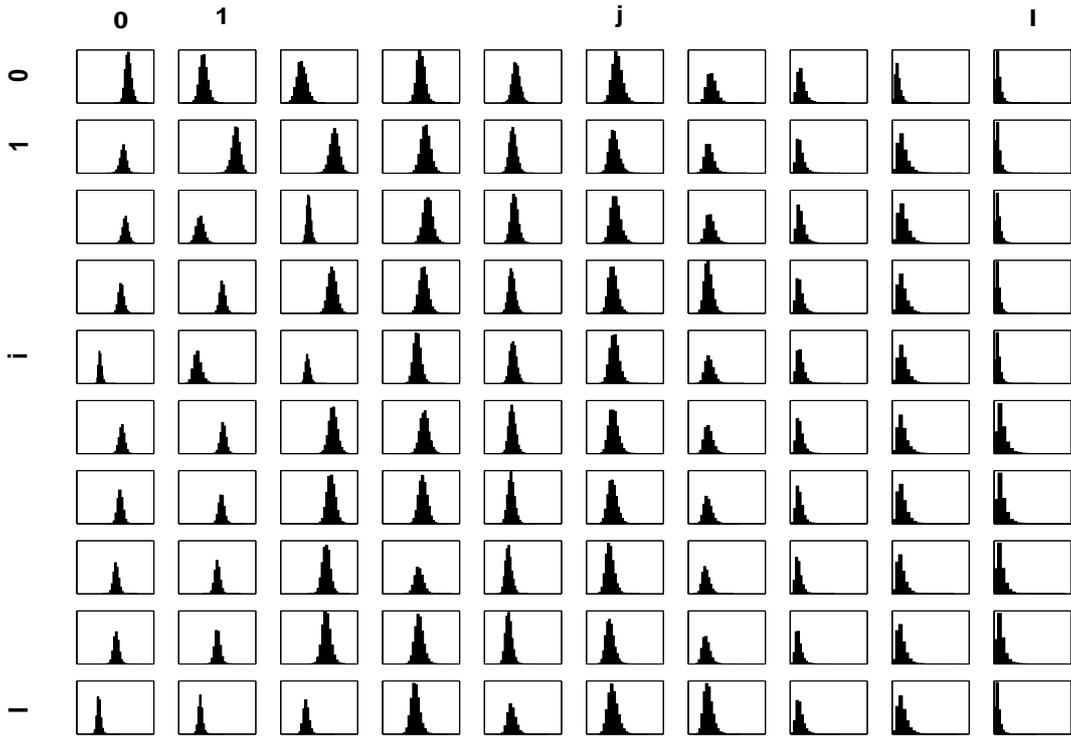


Fig. 12.7.4: Posterior distributions for $\tilde{R}_{i,j} = \alpha_i \beta_j$ estimated using MCMC.

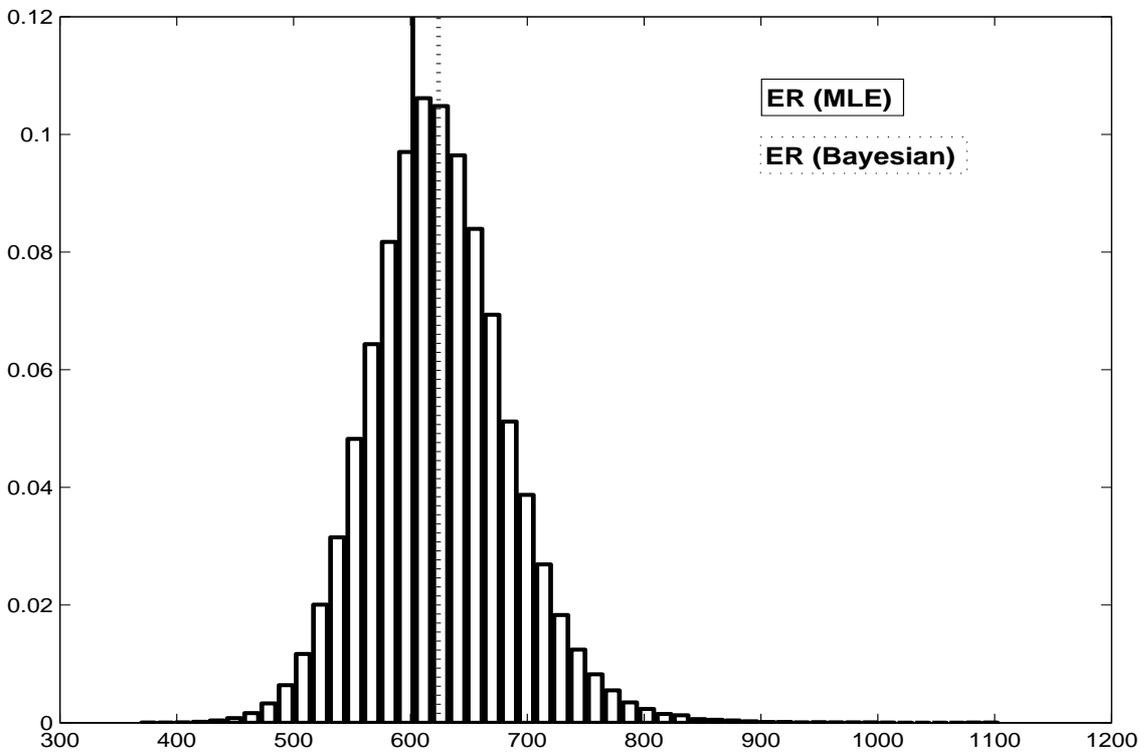


Fig. 12.7.5: Distribution of total outstanding claims payment $R = \sum_{i+j>I} Y_{i,j}$, accounting for all process, estimation and model uncertainties.

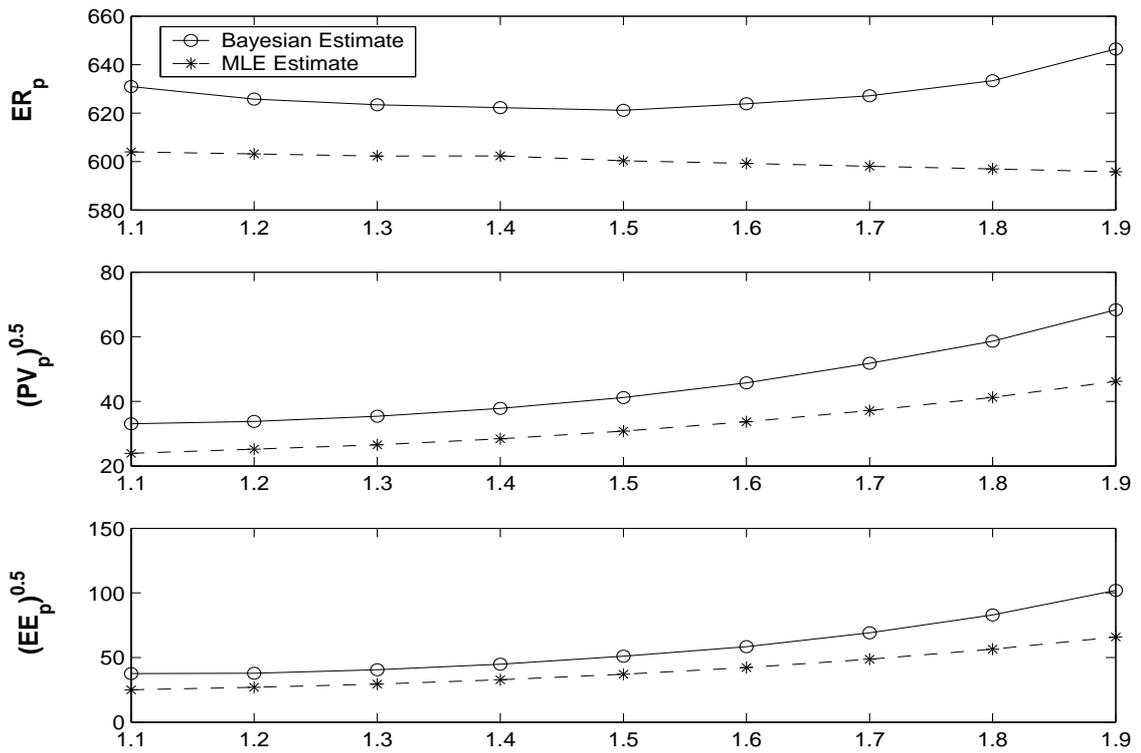


Fig. 12.7.6: Estimates of quantities from Table 12.5 conditional on p . Note, numerical standard errors are not included as they are negligible and are less than the size of the symbols.

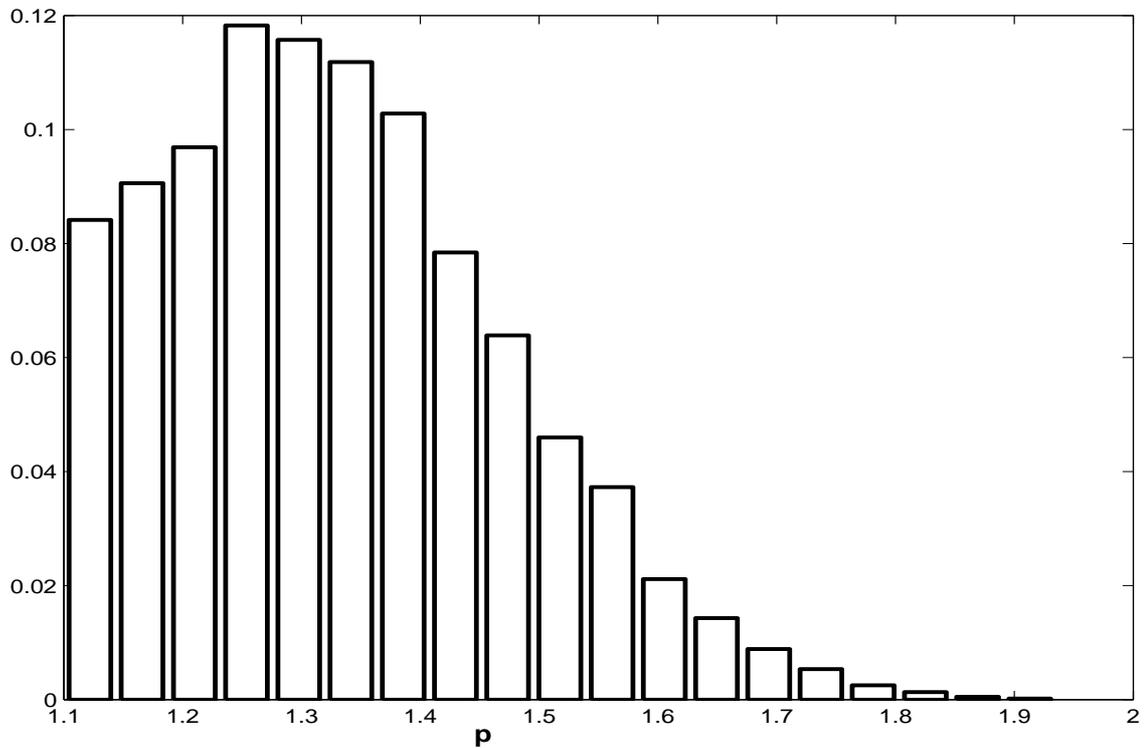


Fig. 12.7.7: Posterior distribution of the model parameter p .

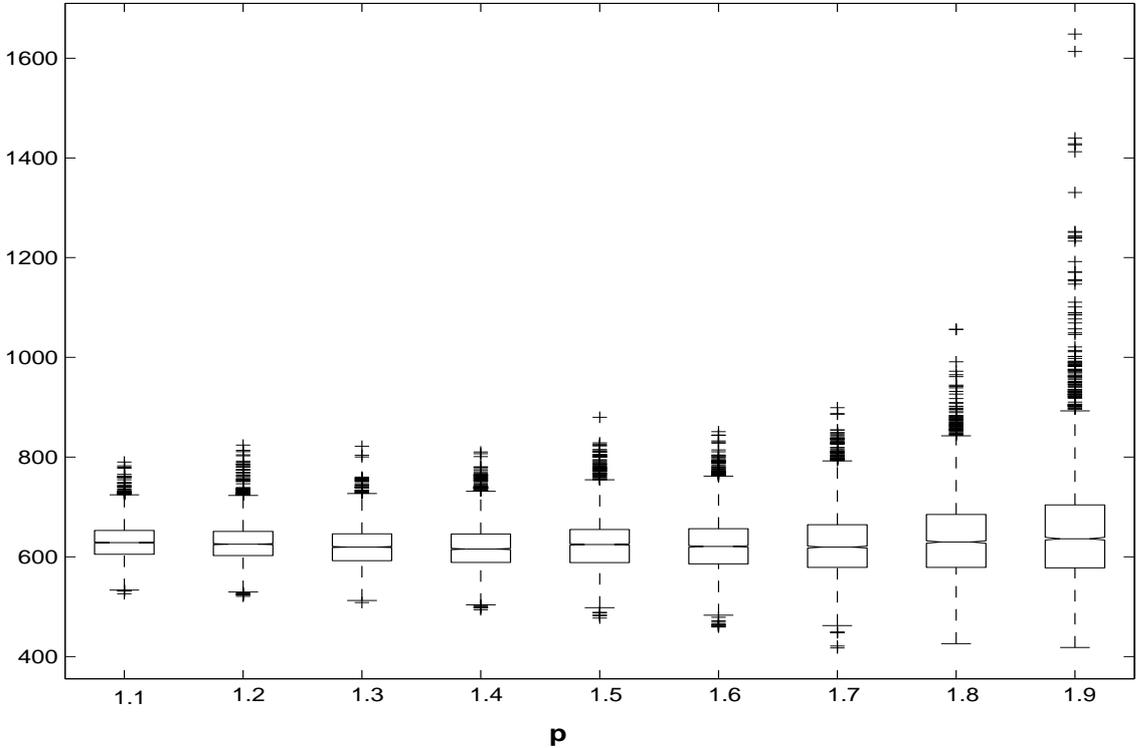


Fig. 12.7.8: Predicted claim reserves \tilde{R} distributional summaries conditional on model parameter p .

References

- [1] Atchade Y. and Rosenthal, J. (2005) On adaptive Markov chain Monte Carlo algorithms. *Bernoulli* **11**(5), 815-828.
- [2] Bedard M. and Rosenthal J.S. (2008) Optimal scaling of Metropolis algorithms: heading towards general target distributions. *The Canadian Journal of Statistics* **36**(4), 483-503.
- [3] Bernardo, J.M. and Smith, A.F.M. (1994) *Bayesian Theory*. John Wiley and Sons, NY.
- [4] Cairns, A.J.G. (2000) A discussion of parameter and model uncertainty in insurance. *Insurance: Mathematics and Economics* **27**, 313-330.
- [5] Carlin, B. and Chib, S. (1995) Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society Series B* **57**, 473-484.
- [6] Casella, G. and George, E.I. (1992) Explaining the Gibbs Sampler. *The American Statistician* **46**(3), 167-174.
- [7] Congdon P. (2006) Bayesian model choice based on Monte Carlo estimates of posterior model probabilities. *Computational Statistics and Data Analysis* **50**(2), 346-357.
- [8] Dunn, P.K. and Smyth, G.K. (2005) Series evaluation of Tweedie exponential dispersion model densities. *Statistics and Computing* **15**, 267-280.
- [9] England P.D. and Verrall R.J. (2002) Stochastic claims reserving in general insurance. *British Actuarial Journal* **8**(3), 443-510.
- [10] Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (1995) *Bayesian Data Analysis*. Chapman and Hall /CRC Texts in Statistical Science Series, **60**.
- [11] Gelman, A., Gilks, W.R. and Roberts, G.O. (1997) Weak convergence and optimal scaling of random walks metropolis algorithm. *Annals of Applied Probability* **7**, 110-120.
- [12] Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1996) *Markov Chain Monte Carlo in Practice*. Chapman and Hall, Florida.
- [13] Green, P. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711-732.

-
- [14] Jørgensen, B. and de Souza, M.C.P. (1994) Fitting Tweedie's compound Poisson model to insurance claims data. *Scandinavian Actuarial Journal*, 69-93.
- [15] Robert, C.P. and Casella, G. (2004) *Monte Carlo Statistical Methods*, 2nd Edition Springer Texts in Statistics.
- [16] Roberts, G.O. and Rosenthal, J.S. (2001) Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science* **16**, 351-367.
- [17] Rosenthal, J.S. (2007) AMCMC: An R interface for adaptive MCMC. *Computational Statistics and Data Analysis* **51**(12), 5467-5470.
- [18] Smith, A.F.M. and Roberts, G.O. (1993) Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of Royal Statistical Society Series B* **55**(1), 3-23.
- [19] Smyth, G.K. and Jørgensen, B. (2002) Fitting Tweedie's compound Poisson model to insurance claims data: dispersion modelling. *Astin Bulletin* **32**, 143-157.
- [20] Tweedie, M.C.K. (1984) An index which distinguishes between some important exponential families. In *Statistics: Applications in new directions. Proceeding of the Indian Statistical Institute Golden Jubilee International Conference*, J.K. Ghosh and J. Roy (eds.), 579-604, Indian Statistical Institute Canada.
- [21] Wright, E.M. (1935) On asymptotic expansions of generalized Bessel functions. *Proceedings of London Mathematical Society* **38**, 257-270.
- [22] Wüthrich, M.V. (2003) Claims reserving using Tweedie's compound Poisson model. *Astin Bulletin* **33**, 331-346.
- [23] Wüthrich, M.V. and Merz, M. (2008) *Stochastic Claims Reserving Methods in Insurance*, Wiley Finance.

Journal Paper 10

"The important thing in science is not so much to obtain new facts as to discover new ways of thinking about them."

William Lawrence Bragg

Peters G.W., Wuthrich M. and Shevchenko P. (2009) "Chain Ladder Method: Bayesian Bootstrap versus Classical Bootstrap". *Insurance: Mathematics and Economics*, (conditionally accepted).

This work was instigated jointly by the first author and his co-authors. The first author on this paper can claim at least 70% of the credit for the contents. The first authors work included developing the methodology contained, developing the applications, the implementation and comparison to alternative approaches, writing the drafts of the paper and undertaking revisions. It is expected that this paper will certainly be accepted for publication and the journal for which it is under review is one of the top Actuarial and Insurance journals. Permission from all the co-authors is granted for inclusion in the PhD.

This paper appears in the thesis in a modified format to that which is in second stage of review at *Insurance: Mathematics and Economics*.

Final print version will appear IME at:

<http://www.elsevier.com>

Chain Ladder Method: Bayesian Bootstrap versus Classical Bootstrap

Gareth W. Peters (*corresponding author*)

CSIRO Mathematical and Information Sciences, Locked Bag 17, North Ryde, NSW, 1670, Australia;

UNSW Mathematics and Statistics Department, Sydney, 2052, Australia;

email: peterga@maths.unsw.edu.au

Mario V. Wüthrich

ETH Zurich, Department of Mathematics, CH-8092 Zurich, Switzerland;

email: wueth@math.ethz.ch

Pavel V. Shevchenko

CSIRO Mathematical and Information Sciences, Sydney, Locked Bag 17, North Ryde, NSW, 1670,

Australia; e-mail: Pavel.Shevchenko@csiro.au

13.1 abstract

The intention of this paper is to estimate a Bayesian Distribution-free chain ladder model using approximate Bayesian computation (ABC) methodology. We demonstrate how to estimate quantities of interest in claims reserving and compare the estimates to those obtained from classical and credibility approaches. In this context, a novel numerical procedure utilizing Markov chain Monte Carlo (MCMC), ABC and a Bayesian bootstrap procedure was developed in a truly distribution-free setting. The ABC methodology arises because we work in a distribution-free setting in which we make no parametric assumptions, meaning we can not evaluate the likelihood point-wise or in this case simulate directly from the likelihood model. The use of a bootstrap procedure allows us to generate samples from the intractable likelihood without the requirement of distributional assumptions, this is critical to the ABC framework. The developed methodology is used to obtain the empirical distribution of the DFCL model parameters and the predictive distribution of the claims conditional on the observed claims. We then estimate predictive Bayesian capital estimates, the Value at Risk (VaR) and the mean square error of prediction (MSEP). The latter is compared with the classical bootstrap and credibility methods.

Keywords: Claims reserving, distribution-free chain ladder, mean square error of prediction, Bayesian chain ladder, approximate Bayesian computation, Markov chain Monte Carlo, annealing, bootstrap

13.2 Motivation

The distribution-free chain ladder model (DFCL) of Mack (13) is a popular model for stochastic claims reserving. In this paper we use a time series formulation of the DFCL model which allows for bootstrapping the claims reserves. An important aspect of this model is that it can provide a justification for the classical deterministic chain ladder (CL) algorithm which is not founded on an underlying stochastic model. Moreover, it allows for the study of prediction uncertainties. Note that there are different stochastic models that lead to the CL reserves (see for example Wüthrich-Merz (24), Section 3.2). In the present paper we use the DFCL formulation to reproduce the CL reserves.

The paper presents a novel methodology for estimating a Bayesian Distribution-free chain ladder model utilizing a framework of approximate Bayesian computation (ABC) in a non-standard manner. A methodology utilizing Markov chain Monte Carlo (MCMC), ABC and a Bayesian bootstrap procedure is developed in a truly distribution-free setting. The ABC framework is required because we work in a distribution-free setting in which we make no parametric assumptions about the form of the likelihood. Effectively, the ABC methodology allows us to overcome the fact that we can not evaluate the likelihood point-wise in the DFCL model. Typically ABC methodology circumvents likelihood evaluations by simulation from the likelihood, however in this case simulation from the likelihood model is not directly available when no parametric assumption is made. The use of ABC methodology combined with bootstrap is how we overcome this additional complexity that the DFCL model presents in the ABC framework. Then by using an MCMC numerical sampling algorithm combined with the novel version of ABC that has the embedded bootstrap procedure we are able to obtain samples from the intractable posterior distribution of the DFCL model parameters.

This allows us to utilize this methodology to obtain the Bayesian posterior distribution of the DFCL model parameters empirically. Then we demonstrate two approaches in which we can utilize the posterior samples for the DFCL model parameters to obtain the Bayesian predictive distribution of the claims. The first approach involves using each posterior sample to numerically estimate the full predictive claims distribution given the observed claims. The alternative approach involves using the posterior samples for the DFCL model parameters to form Bayesian point estimators, then conditional on these point estimators we can obtain the Bayesian conditional predictive distribution for the claims. The second approach will be relevant for comparisons with the classical and credibility approaches and the first approach has the benefit that it integrates out of the Bayesian predictive claims distribution the parameter uncertainty associated with estimation of the DFCL chain ladder parameters.

The paper then analyzes the parameter estimates in the DFCL model, the associated claims reserves and the mean square errors of prediction (MSEP) from both the frequentist perspective and a contrasting Bayesian view. In doing so we analyze CL point estimators for parameters of the DFCL model, the resulting estimated reserves and the associated MSEP from the classical perspective. These include non-parametric bootstrap estimated prediction errors which can be

obtained via one of two possible bootstrap procedures, conditional or unconditional. In this paper we consider the process of conditional back propagation, see (24) for in depth discussion. These classical frequentist estimators are then compared to Bayesian point estimators. The Bayesian estimates considered are the maximum *a posteriori* (MAP) and the minimum mean square error (MMSE) estimators. For comparison with the classical frequentist reserve estimates we also obtain the associated Bayesian estimated reserves conditional upon the Bayesian point estimators.

In addition, since in the Bayesian setting we obtain samples from the posterior for the parameters we use these along with the MSEP obtained by the estimated Bayesian point estimators to obtain associated posterior predictive intervals to be compared with the classical bootstrap procedures. We then robustify the prediction of reserves by Rao-Blackwellization, that is we integrate out the influence of the unknown variance parameters in the DFCL model. Having done this we analyse the resultant MSEP. This is again only achievable since in the Bayesian setting we obtain samples from the joint posterior for the CL factors and the variances.

To summarize our contribution, the novelty within this paper involves the development and comparison of a new estimation methodology to work with the Bayesian CL model for the DFCL model which makes no parametric assumptions on the form of the likelihood function, see also Gisler-Wüthrich (11). This is unlike the works of Yao (25) and Peters-Shevchenko-Wüthrich (19) that assume explicit distributions in order to construct the posterior distributions in the Bayesian context. Instead we demonstrate how to work directly with the intractable likelihood functions and the resulting intractable posterior distribution, using novel Approximate Bayesian Computation methodology. In this regard we demonstrate that we do not need to make any parametric assumptions to perform posterior inference, avoiding potentially poor model assumptions made, as for example in the paper of Yao (25).

Outline of this paper. The paper begins with a presentation of the claims reserving problem and then presents the model we shall consider. This is followed by the description of the classical chain ladder algorithm and the construction of a Bayesian model that can be used to estimate the parameters of the model. The Bayesian model is constructed in a distribution-free setting. Following this is a discussion on classical versus Bayesian parameter estimators along with a bootstrap based procedure for estimation of parameter uncertainty in the classical setting. The next section presents the methodology of ABC coupled with a novel bootstrap based sampling procedure which will allow us to work directly with the distribution-free Bayesian model. We then illustrate the developed algorithm on a synthetic data set and the real data set, comparing performance to the classical results and those obtained via credibility theory.

13.3 *Claims development triangle and DFCL model*

We briefly outline the claims development triangle structure we utilise in the formulation of our models. Assume there is a run-off triangle containing claims development data with the structure given in Table 13.1.

| accident year i | development years j | | | | | |
|----------------------|-----------------------|---|-----|---|-----|-----|
| | 0 | 1 | ... | j | ... | I |
| 0 | | | | | | |
| 1 | | observed random variables $C_{i,j} \in \mathcal{D}_I$ | | | | |
| \vdots | | | | | | |
| i | | | | | | |
| \vdots | | | | | | |
| $I-1$ | | | | | | |
| I | | | | to be predicted $C_{i,j} \in \mathcal{D}_I^c$ | | |

Tab. 13.1: Claims development triangles.

Assume that $C_{i,j}$ are cumulative claims with indices $i \in \{0, \dots, I\}$ and $j \in \{0, \dots, J\}$, where i denotes the accident year and j denotes the development year (cumulative claims can refer to payments, claims incurred, etc). We make the simplifying assumption that the number of accident years is equal to the number of observed development periods i.e. $I = J$. At time I , we have observations

$$\mathcal{D}_I = \{C_{i,j}; i + j \leq I\}, \tag{13.3.1}$$

and for claims reserving at time I we need to predict the future claims

$$\mathcal{D}_I^c = \{C_{i,j}; i + j > I, i \leq I, j \leq J\}. \tag{13.3.2}$$

Moreover, we define for $j \in \{0, \dots, I\}$ the set $\mathcal{B}_j = \{C_{i,k}; i + k \leq I, 0 \leq k \leq j\}$, e.g. \mathcal{B}_0 is the first column in Table 13.1.

13.3.1 Classical chain ladder algorithm

In the classical (deterministic) chain ladder algorithm there is no underlying stochastic model. It is rather a recursive algorithm that was used to estimate the claims reserves and which has proved to give good practical results. It simply involves the following recursive steps to predict unobserved cumulative claims in \mathcal{D}_I^c . Set $\widehat{C}_{i,I-i} = C_{i,I-i}$ and for $j > I - i$

$$\widehat{C}_{i,j} = \widehat{C}_{i,j-1} \widehat{f}_{j-1}^{(CL)} \quad \text{with CL factor estimates } \widehat{f}_{j-1}^{(CL)} = \frac{\sum_{i=0}^{I-j} C_{i,j}}{\sum_{i=0}^{I-j} C_{i,j-1}}. \tag{13.3.3}$$

Since this is a deterministic algorithm it does not allow for quantification of the uncertainty associated with the predicted reserves. To analyse the associated uncertainty there are several stochastic models that reproduce the CL reserves: for example Mack’s distribution-free chain ladder model (13), the over-dispersed Poisson model (see England-Verrall (5)) or the Bayesian chain ladder model (see Gisler-Wüthrich (11)). We use a time series formulation of the Bayesian chain ladder model in order to use bootstrap methods and Bayesian inference.

13.3.2 Bayesian DFCL model

We use an additive time series version of the Bayes chain ladder model (Model Assumptions 3.1, in Gisler-Wüthrich (11)).

Model Assumptions 13.3.1.

1. We define the CL factors by $\mathbf{F} = (F_0, \dots, F_{J-1})$ and the standard deviation parameters by $\Xi = (\Xi_0, \dots, \Xi_{J-1})$. We assume independence between all these parameters, i.e., the prior density of (\mathbf{F}, Ξ) is given by

$$\pi(\mathbf{f}, \boldsymbol{\sigma}) = \prod_{j=0}^{J-1} \pi(f_j) \pi(\sigma_j), \quad (13.3.4)$$

where $\pi(f_j)$ denotes the density of F_j and $\pi(\sigma_j)$ denotes the density of Ξ_j .

2. Conditionally, given $\mathbf{F} = \mathbf{f} = (f_0, \dots, f_{J-1})$ and $\Xi = \boldsymbol{\sigma} = (\sigma_0, \dots, \sigma_{J-1})$, we have:

- Cumulative claims $C_{i,j}$ in different accident years i are independent.
- Cumulative claims satisfy the following time series representation

$$C_{i,j+1} = f_j C_{i,j} + \sigma_j \sqrt{C_{i,j}} \varepsilon_{i,j+1}, \quad (13.3.5)$$

where conditionally, given \mathcal{B}_0 , we have that the residuals $\varepsilon_{i,j}$ are i.i.d. satisfying

$$E[\varepsilon_{i,j} | \mathcal{B}_0, \mathbf{F}, \Xi] = 0 \text{ and } \text{Var}[\varepsilon_{i,j} | \mathcal{B}_0, \mathbf{F}, \Xi] = 1, \quad (13.3.6)$$

and $P[C_{i,j} > 0 | \mathcal{B}_0, \mathbf{F}, \Xi] = 1$ for all i, j .

Remark. Note that the assumptions on the residuals are slightly involved in order to guarantee that cumulative claims $C_{i,j}$ are positive P -a.s.

Corollary 13.3.2. *Under Model Assumptions 13.3.1 we have that conditionally, given \mathcal{D}_I , the random variables $F_0, \dots, F_{J-1}, \Xi_0, \dots, \Xi_{J-1}$ are independent.*

Proof of Corollary 13.3.2. The proof is completely analogous to the proof of Theorem 3.2 in Gisler-Wüthrich (11) and follows from prior independence of the parameters and the fact that $C_{i,j+1}$ only depends on F_j, Ξ_j and $C_{i,j}$ (Markov property).

□

In particular, Corollary 13.3.2 says that we obtain the following posterior distribution for (\mathbf{F}, Ξ) , given \mathcal{D}_I ,

$$\pi(\mathbf{f}, \boldsymbol{\sigma} | \mathcal{D}_I) = \prod_{j=0}^{J-1} \pi(f_j | \mathcal{D}_I) \pi(\sigma_j | \mathcal{D}_I). \quad (13.3.7)$$

This has important implications for the ABC sampling algorithm developed below.

In order to perform the Bayesian analysis we make explicit assumptions on the prior distributions of (\mathbf{F}, Ξ) .

Model Assumptions 13.3.3.

In addition to Model Assumptions 13.3.1 we assume that the prior model for all parameters $j \in \{0, \dots, J-1\}$ is given by:

- $F_j \sim \Gamma(\alpha_j, \beta_j)$, where $\Gamma(\alpha_j, \beta_j)$ is a gamma distribution with mean $E[F_j] = \alpha_j \beta_j = \hat{f}_j^{(CL)}$ (see 13.3.3) and large variance to have diffuse priors.
- The variances $\Xi_j^2 \sim IG(a_j, b_j)$, where $IG(a_j, b_j)$ is an inverse gamma distribution with mean $E[\Xi_j^2] = b_j / (a_j - 1) = \hat{\sigma}_j^{2(CL)}$ (see 13.4.1, below) and large variance.

Remarks

1. The likelihood model is intractable, meaning that no density can be written down analytically in the DFCL model. In formulating the Bayesian model we have only made distributional assumptions on the priors for the parameters (\mathbf{F}, Ξ) but not on the observable cumulative claims $C_{i,j}$. Though we make distributional assumptions for the priors, the model is distribution free because no distributional assumptions on the cumulative claims are made. As a result of only making assumptions on the priors, a standard Bayesian analysis using analytic posterior distributions cannot be performed. One way out of this dilemma would be to re-formulate the Bayesian model by making distributional assumptions (this is, e.g., done in Yao (25)) but then the model is no longer distribution-free. Another approach would be to use credibility methods (see Gisler-Wüthrich (11)) but this only gives statements for second moments. In the present set up we develop ABC methods that allow for a full distributional answer for the posterior without making explicit distributional assumptions for the cumulative claims $C_{i,j}$.
2. Our priors are chosen as diffuse priors with large variances. This again highlights the differences between specification of the prior distributions and making distributional assumptions for the actual likelihood model, these are mutually exclusive ideas.
3. We select the priors to ensure that we maintain several relevant aspects of the DFCL model. In particular, it is important to utilize priors that enforce the strict positivity of the parameters $f_j, \sigma_j > 0$. We note here that the parametric Bayesian model developed in Yao (25) failed in this aspect when it came to prior specification and therefore we develop an alternative prior structure that satisfies these required properties of the DFCL model.

13.4 DFCL model estimators

Classical

In the classical CL method, the CL factors are estimated by $\hat{f}_j^{(CL)}$ given in 13.3.3. The variance parameters are estimated by, see e.g. (3.4) Wüthrich-Merz (24),

$$\hat{\sigma}_j^{2(CL)} = \frac{1}{I-j-1} \sum_{i=0}^{I-j-1} C_{i,j} \left(\frac{C_{i,j+1}}{C_{i,j}} - \hat{f}_j^{(CL)} \right)^2. \tag{13.4.1}$$

Note that this estimator is only well-defined for $j < I - 1$. There is a vast literature and discussion on the estimation of tail parameters. We do not enter this discussion here but we simply choose the estimator given in Mack (13) for the last variance parameter which is defined by

$$\hat{\sigma}_{J-1}^{2(CL)} = \min \left\{ \frac{\hat{\sigma}_{J-2}^{4(CL)}}{\hat{\sigma}_{J-3}^{2(CL)}}, \hat{\sigma}_{J-3}^{2(CL)}, \hat{\sigma}_{J-2}^{2(CL)} \right\}. \tag{13.4.2}$$

Bayesian

In a Bayesian inference context one calculates the posterior distribution of the parameters, given \mathcal{D}_I . As in 13.3.7 we denote this posterior by $\pi(\mathbf{f}, \boldsymbol{\sigma} | \mathcal{D}_I)$. Since the MCMC-ABC bootstrap procedure will allow us to obtain samples from the posterior distribution of the Bayesian DFCL model presented, we can now consider estimating CL point estimators using these samples.

There are two commonly used point estimators in Bayesian analysis that correspond to the posterior mode (MAP) and the posterior mean (MMSE), respectively. Given the posterior independence (see Corollary 13.3.2) they are given by:

$$\hat{f}_j^{(MAP)} = \arg \max_{f_j} \pi(f_j | \mathcal{D}_I), \tag{13.4.3}$$

$$\hat{\sigma}_j^{(MAP)} = \arg \max_{\sigma_j} \pi(\sigma_j | \mathcal{D}_I), \tag{13.4.4}$$

and

$$\hat{f}_j^{(MMSE)} = \int f_j \pi(f_j | \mathcal{D}_I) df_j = E[F_j | \mathcal{D}_I], \tag{13.4.5}$$

$$\hat{\sigma}_j^{(MMSE)} = \int \sigma_j \pi(\sigma_j | \mathcal{D}_I) d\sigma_j = E[\Xi_j | \mathcal{D}_I]. \tag{13.4.6}$$

Note that for diffuse prior we find (see Corollary 5.1 in Gisler-Wüthrich (11))

$$\hat{f}_j^{(MMSE)} \approx \hat{f}_j^{(CL)}. \tag{13.4.7}$$

Hence, using Corollary 13.3.2, we obtain the approximation

$$\begin{aligned}
E[C_{i,J}|\mathcal{D}_I] &= E[E[C_{i,J}|\mathcal{D}_I, \mathbf{F}, \boldsymbol{\Xi}|\mathcal{D}_I]] = C_{i,I-i} E\left[\prod_{j=I-i}^{J-1} F_j \middle| \mathcal{D}_I\right] \\
&= C_{i,I-i} \prod_{j=I-i}^{J-1} E[F_j|\mathcal{D}_I] = C_{i,I-i} \prod_{j=I-i}^{J-1} \hat{f}_j^{(MMSE)} \\
&\approx C_{i,I-i} \prod_{j=I-i}^{J-1} \hat{f}_j^{(CL)} = \hat{C}_{i,J},
\end{aligned} \tag{13.4.8}$$

where on the last line we have an asymptotic equality if the diffusivity of the priors $\pi(f_j)$ tends to infinity. This is exactly the argument why the Bayesian CL model can be used to justify the CL predictors, see Gisler-Wüthrich (11).

Full predictive distribution and VaR

In addition, the posterior samples for the DFCL model parameters, obtained via the MCMC-ABC bootstrap procedure, will allow us to obtain the predictive distribution of the claims in two ways. The first is the full predictive distribution of the claims obtained after integrating out the posterior uncertainty associated with the Bayesian DFCL model parameters to empirically estimate

$$\pi(\mathcal{D}_I^c|\mathcal{D}_I) = \int \int \pi(\mathcal{D}_I^c|\mathbf{f}, \boldsymbol{\sigma}) \pi(\mathbf{f}, \boldsymbol{\sigma}|\mathcal{D}_I) d\mathbf{f} d\boldsymbol{\sigma}. \tag{13.4.9}$$

In practice, this numerical procedure involves taking each posterior sample for the DFCL model parameters and obtaining an estimate of the predicted claims.

The second approach involves using one of the Bayesian point estimators for the parameters such as the MMSE to obtain $\pi(\mathcal{D}_I^c|\mathbf{f}^{MMSE}, \boldsymbol{\sigma}^{MMSE})$ or Rao-Blackwellised version of the Bayesian predictive distribution of claims involving $\pi(\mathcal{D}_I^c|\mathbf{f}^{MMSE})$ having numerically integrated out the Bayesian posterior uncertainty associated with the DFCL variance parameters.

These results can then also be applied to estimate of risk measures. For example, if we fix a security level 95% we can calculate the VaR on that level, which is defined by

$$\text{VaR}_{0.95}\left(C_{i,J} - E[C_{i,J}|\mathcal{D}_I] \middle| \mathcal{D}_I\right) = \min\left\{x; P\left[C_{i,J} - E[C_{i,J}|\mathcal{D}_I] > x \middle| \mathcal{D}_I\right] \leq 0.05\right\}. \tag{13.4.10}$$

13.5 Bootstrap and mean square error of prediction

Assume that we have calculated the Bayesian predictor or the CL predictor given in 13.4.8. Then we would like to determine the prediction uncertainty, i.e., we would like to study the deviation of $C_{i,J}$ around its predictor. If one is only interested in second moments, the so-called conditional mean square error of prediction (MSEP), one can often estimate the error terms analytically. However, other uncertainty measures like Value-at-Risk (VaR) can only be

determined numerically.

A popular numerical method is the bootstrap method. The bootstrap technique was developed by Efron (3) and extended by Efron-Tibshirani (4) and Davison-Hinkley (1). This procedure allows one to obtain information regarding an aggregated distribution given a single realisation of the data. To apply the bootstrap procedure one introduces a minimal amount of model structure such that resampling observations can be achieved using observed samples of the data.

In this section we present a bootstrap algorithm in the classical frequentists approach, i.e., we assume that the CL factors $\mathbf{F} = \mathbf{f}$ and the standard deviation parameters $\mathbf{\Xi} = \boldsymbol{\sigma}$ given in Model Assumptions 13.3.1 are unknown constants. The bootstrap then generates synthetic data denoted by \mathcal{D}_I^* that allow for the study of the fluctuations of $\hat{\mathbf{f}}^{(CL)}$ and $\hat{\boldsymbol{\sigma}}^{2(CL)}$ (for details see Wüthrich-Merz (24), Section 7.4). In the presented text we restrict ourselves to the conditional resampling approach presented in Section 7.4.2 of Wüthrich-Merz (24).

13.5.1 Non-parametric classical bootstrap (conditional version)

1. Calculate estimated residuals $\tilde{\varepsilon}_{i,j}$ for $i + j \leq I, j > 0$, conditional on the estimators $\hat{f}_{0:J-1}^{(CL)}$ and $\hat{\sigma}_{0:J-1}^{2(CL)}$ and the observed data \mathcal{D}_I :

$$\tilde{\varepsilon}_{i,j} = \tilde{\varepsilon}_{i,j}(\hat{f}_{j-1}^{(CL)}, \hat{\sigma}_{j-1}^{(CL)}) = \frac{C_{i,j} - \hat{f}_{j-1}^{(CL)} C_{i,j-1}}{\hat{\sigma}_{j-1}^{(CL)} \sqrt{C_{i,j-1}}}.$$

2. These residuals $(\tilde{\varepsilon}_{i,j})_{i+j \leq I}$ give the empirical bootstrap distribution $\hat{F}_{\mathcal{D}_I}$.
3. Sample i.i.d. residuals $\tilde{\varepsilon}_{i,j}^* \sim \hat{F}_{\mathcal{D}_I}$ for $i + j \leq I, j > 0$.
4. Generate bootstrap observations (conditional resampling)

$$C_{i,j}^* = \hat{f}_{j-1}^{(CL)} C_{i,j-1} + \hat{\sigma}_{j-1}^{(CL)} \sqrt{C_{i,j-1}} \tilde{\varepsilon}_{i,j}^*,$$

which defines $\mathcal{D}_I^* = \mathcal{D}_I^*(\hat{\mathbf{f}}^{(CL)}, \hat{\boldsymbol{\sigma}}^{(CL)})$. Note that for the unconditional version of bootstrap we should generate $C_{i,j}^* = \hat{f}_{j-1}^{(CL)} C_{i,j-1}^* + \hat{\sigma}_{j-1}^{(CL)} \sqrt{C_{i,j-1}^*} \tilde{\varepsilon}_{i,j}^*$. For a discussion on this approach see Section 7.4.1 of (24).

5. Calculate bootstrapped CL parameters \hat{f}_j^* and $\hat{\sigma}_j^{2*}$ by

$$\begin{aligned} \hat{f}_j^* &= \frac{\sum_{i=0}^{I-j-1} C_{i,j+1}^*}{\sum_{i=0}^{I-j-1} C_{i,j}^*}, \\ \hat{\sigma}_j^{2*} &= \frac{1}{I-j-1} \sum_{i=0}^{I-j-1} C_{i,j} \left(\frac{C_{i,j+1}^*}{C_{i,j}^*} - \hat{f}_j^* \right)^2. \end{aligned}$$

6. Repeat steps 3-5 and obtain empirical distributions from the bootstrap samples $\widehat{C}_{i,J}^*$, \widehat{f}_j^* and $\widehat{\sigma}_j^{2*}$. These are then used to quantify the parameter estimation uncertainty.

This non-parametric classical bootstrap method can be seen as a frequentist approach. This means that we do not express our parameter uncertainty by the choice of an appropriate prior distribution. We rather use a point estimator for the unknown parameters and then study the possible fluctuations of this point estimator.

The main difficulty now is that the non-parametric bootstrap method, as described above, underestimates the “true” uncertainty. This comes from the fact that the estimated residuals $\widetilde{\varepsilon}_{i,j}$, in general, have variance smaller than 1 (see formula (7.23) in Wüthrich-Merz (24)). This means that our estimated residuals are not appropriately scaled. Therefore, frequentists use several different scalings to correct this fact (see formula (7.24) in Wüthrich-Merz (24) or England-Verrall (5)). Here, we use a different approach by introducing the novel Bayesian bootstrap method embedded within an MCMC-ABC algorithm to obtain empirically the posterior distribution of the Bayesian DFCL model, see Section 13.6 below. Having obtained this we can then calculate all required Bayesian parameter estimates, capital reserve estimates and associated measures of uncertainty such as VaR. Before presenting the methodology for this novel MCMC-ABC algorithm we will finalize this section with the decompositions of the MSEP under frequentist, Bayesian and credibility approaches.

13.5.2 Frequentist bootstrap estimates

Let us for the time-being concentrate on the conditional MSEP given by

$$\begin{aligned} \text{mse}_{C_{i,J}|\mathcal{D}_I}(\widehat{C}_{i,J}) &= E \left[\left(C_{i,J} - \widehat{C}_{i,J} \right)^2 \middle| \mathcal{D}_I \right] \\ &= \text{Var}(C_{i,J} | \mathcal{D}_I) + \left(E[C_{i,J} | \mathcal{D}_I] - \widehat{C}_{i,J} \right)^2. \end{aligned} \quad (13.5.1)$$

The first term is known as the conditional process variance and the second term as the parameter estimation uncertainty. In the frequentists approach, i.e. for given deterministic $\mathbf{F} = \mathbf{f}$ and $\Xi = \sigma$, this can be calculated, see Wüthrich-Merz (24), Section 3.2. Namely, the terms are given by

$$\text{Var}(C_{i,J} | \mathcal{D}_I) = \left(E[C_{i,J} | C_{i,I-i}] \right)^2 \sum_{j=I-i}^{J-1} \frac{\sigma_j^2 / f_j^2}{E[C_{i,j} | C_{i,I-i}]} \stackrel{\text{def.}}{=} C_{i,I-i} \Gamma_{I-i}, \quad (13.5.2)$$

and

$$\left(E[C_{i,J}|\mathcal{D}_I] - \widehat{C}_{i,J}\right)^2 = C_{i,I-i}^2 \left(\prod_{j=I-i}^{J-1} f_j - \prod_{j=I-i}^{J-1} \widehat{f}_j^{(CL)}\right)^2 \stackrel{def.}{=} C_{i,I-i}^2 \Delta_{I-i}. \quad (13.5.3)$$

The process variance 13.5.4 is estimated by replacing the parameters by its estimators,

$$\widehat{\text{Var}}(C_{i,J}|\mathcal{D}_I) = (\widehat{C}_{i,J})^2 \sum_{j=I-i}^{J-1} \frac{\widehat{\sigma}_j^{2(CL)} / (\widehat{f}_j^{(CL)})^2}{\widehat{C}_{i,j}} \stackrel{def.}{=} C_{i,I-i} \widehat{\Gamma}_{I-i}^{freq}. \quad (13.5.4)$$

The parameter estimation error is more involved and there we need the bootstrap algorithm. Assume that the bootstrap method gives T bootstrap samples $\widehat{f}_j^{*(1)}, \dots, \widehat{f}_j^{*(T)}$. Then the parameter estimation error 13.5.3 is estimated by the sample variance of the product of the bootstrap observation chain ladder parameter estimates $\widehat{f}_j^{*(1)}, \dots, \widehat{f}_j^{*(T)}$, which gives the estimator $C_{i,I-i}^2 \widehat{\Delta}_{I-i}^{freq}$.

13.5.3 Bayesian estimates

In the Bayesian setup, i.e. choosing prior distributions for the unknown parameters \mathbf{F} and $\mathbf{\Xi}$, we obtain a natural decomposition of the conditional MSEP.

$$\begin{aligned} \text{mse}_{C_{i,J}|\mathcal{D}_I}(E[C_{i,J}|\mathcal{D}_I]) &= \text{Var}(C_{i,J}|\mathcal{D}_I) \\ &= E[\text{Var}(C_{i,J}|\mathcal{D}_I, \mathbf{F}, \mathbf{\Xi})|\mathcal{D}_I] + \text{Var}(E[C_{i,J}|\mathcal{D}_I, \mathbf{F}, \mathbf{\Xi}]|\mathcal{D}_I). \end{aligned} \quad (13.5.5)$$

The average process variance is given by (see Wüthrich-Merz (24), Lemma 3.6)

$$\begin{aligned} E[\text{Var}(C_{i,J}|\mathcal{D}_I, \mathbf{F}, \mathbf{\Xi})|\mathcal{D}_I] &= C_{i,I-i} \sum_{j=I-i}^{J-1} E \left[\prod_{m=I-i}^{j-1} F_m \Xi_j^2 \prod_{n=j+1}^{J-1} F_n^2 \middle| \mathcal{D}_I \right] \\ &= C_{i,I-i} \sum_{j=I-i}^{J-1} \prod_{m=I-i}^{j-1} E[F_m|\mathcal{D}_I] E[\Xi_j^2|\mathcal{D}_I] \prod_{n=j+1}^{J-1} E[F_n^2|\mathcal{D}_I] \stackrel{def.}{=} C_{i,I-i} \widehat{\Gamma}_{I-i}^{Bayes}, \end{aligned} \quad (13.5.6)$$

where we have used Corollary 13.3.2. The parameter estimation error is given by

$$\text{Var}(E[C_{i,J}|\mathcal{D}_I, \mathbf{F}, \mathbf{\Xi}]|\mathcal{D}_I) = C_{i,I-i}^2 \text{Var} \left(\prod_{j=I-i}^{J-1} F_j \middle| \mathcal{D}_I \right) \stackrel{def.}{=} C_{i,I-i}^2 \widehat{\Delta}_{I-i}^{Bayes}, \quad (13.5.7)$$

where we have used 13.4.8. Using posterior independence, see Corollary 13.3.2, we obtain for the last term

$$C_{i,I-i}^2 \widehat{\Delta}_{I-i}^{Bayes} = C_{i,I-i}^2 \left[\prod_{j=I-i}^{J-1} E[F_j^2 | \mathcal{D}_I] - \prod_{j=I-i}^{J-1} E[F_j | \mathcal{D}_I]^2 \right]. \quad (13.5.8)$$

In order to calculate these two terms given in 13.5.6 and 13.5.8 we need to calculate the posterior distribution of (\mathbf{F}, Ξ) , given \mathcal{D}_I . Since we do not have a full distributional model, we cannot write down the likelihood function, which would allow for analytical solutions or Markov chain Monte Carlo (MCMC) simulations. Therefore we introduce the ABC framework which allows for distribution-free simulations using appropriate bootstrap samples and a distance metric. This we are going to discuss in the next section.

13.5.4 Credibility Estimates

As mentioned previously, we can also consider the credibility estimates given in Gisler-Wüthrich (11). As long as we are only interested in the second moments, i.e. conditional MSEF, we can also use credibility estimators, which are minimum variance estimators that are linear in the observations. For diffuse priors we obtain the approximation given in Corollary 7.2 of Gisler-Wüthrich (11)

$$\widehat{\text{msef}}_{C_{i,J} | \mathcal{D}_I} (E[C_{i,J} | \mathcal{D}_I]) = C_{i,I-i} \widehat{\Gamma}_{I-i}^{cred} + C_{i,I-i}^2 \widehat{\Delta}_{I-i}^{cred}, \quad (13.5.9)$$

where

$$\widehat{\Gamma}_{I-i}^{cred} = \sum_{j=I-i}^{J-1} \left\{ \prod_{m=I-i}^{j-1} \widehat{f}_m^{(CL)} \widehat{\sigma}_j^{2(CL)} \prod_{n=j+1}^{J-1} \left((\widehat{f}_n^{(CL)})^2 + \frac{\widehat{\sigma}_n^{2(CL)}}{\sum_{i=0}^{I-n-1} C_{i,n}} \right) \right\}, \quad (13.5.10)$$

$$\widehat{\Delta}_{I-i}^{cred} = \prod_{j=I-i}^{J-1} \left((\widehat{f}_j^{(CL)})^2 + \frac{\widehat{\sigma}_j^{2(CL)}}{\sum_{i=0}^{I-j-1} C_{i,j}} \right) - \prod_{j=I-i}^{J-1} (\widehat{f}_j^{(CL)})^2. \quad (13.5.11)$$

In the results section we compare the frequentist bootstrap approach, the credibility approach and the ABC bootstrap approach that is described (see Table 13.5.1).

13.6 ABC for intractable likelihoods and numerical Markov chain sampler

To estimate numerically the parameters, predicted claims and associated uncertainty measures such as the MSEF presented in the previous sections, the Bayesian approach requires the ability to sample from the posterior distribution of the DFCL model parameters. Obtaining samples $\{\mathbf{f}^{(t)}, \boldsymbol{\sigma}^{2(t)}\}_{t=1:T}$ which are realizations of a random vector distributed with a posterior distribution $\pi(\mathbf{f}, \boldsymbol{\sigma} | \mathcal{D}_I)$ in the DFCL model is difficult since the likelihood is intractable. Hence, standard numerical approaches such as Markov chain Monte Carlo (MCMC) algorithms (see, e.g., Gilks et al. (10)) cannot be directly used since they all require explicit repeated evaluation of the likelihood function at each stage of the Markov chain sampling algorithm. It is common

to avoid this difficulty by making distributional assumptions for the form of the likelihood. This then violates a distribution-free CL model assumption but allows for relatively standard sampling procedures to be applied. In this regard, one possible approach involves making a specific Gaussian assumption for the likelihood. One problem with this assumption which is evident immediately is that it precludes skewness in the model. Here, we do not make any such assumptions and instead we work in a truly distribution-free model using approximate Bayesian computation (ABC) to facilitate sampling from an intractable posterior distribution.

There is an additional complexity in the DFCL model not typically encountered when working with ABC methodology. Typically ABC methodology is developed in the case in which the likelihood can not be evaluated point-wise, but conditional on parameter values, it can be simulated from trivially, see examples in Peters and Sisson (15) and Peters et al. (20). This is not the case in the Bayesian DFCL model. Under the DFCL model the likelihood is only expressed by moment conditions, hence we can not evaluate the likelihood point-wise and also the simulation from the likelihood can not be performed directly. This is why we introduce the novel concept of the Bayesian bootstrap which is embedded within the ABC methodological framework.

Hence, to sample from the posterior in our DFCL model we develop a novel formulation of the ABC methodology based on the bootstrap and conditional back transformation procedure, similar to that discussed in Section 13.5.

ABC methods aim to sample from posterior distributions in the presence of computationally intractable likelihood functions. For an application in risk modelling of ABC methodology see Peters-Sisson (15). In this article we present a novel MCMC-ABC algorithm. Before presenting some details of the numerical MCMC procedure we note that alternative numerical algorithms could be considered in the ABC context. For example a sequential Monte Carlo (SMC) based algorithms which can improve simulation efficiency can be found in Del Moral et al. (2), Sisson et al. (23), Peters et al. (16),(17) and Marjoram et al. (14).

13.6.1 ABC methodology

In this section we provide a brief description of ABC methodology, which describes a suite of methods developed specifically for working with models in which the likelihood is computationally intractable. Here we work with a Bayesian model and consider the likelihood intractability to arise in the sense that we may not evaluate the likelihood point-wise.

The ABC method we consider here embeds an intractable target posterior distribution, in our case denoted by $\pi(\mathbf{f}, \boldsymbol{\sigma} | \mathcal{D}_I)$, into a general augmented model

$$\pi(\mathbf{f}, \boldsymbol{\sigma}, \mathcal{D}_I^*, \mathcal{D}_I) = \pi(\mathcal{D}_I | \mathcal{D}_I^*, \mathbf{f}, \boldsymbol{\sigma}) \pi(\mathcal{D}_I^* | \mathbf{f}, \boldsymbol{\sigma}) \pi(\mathbf{f}) \pi(\boldsymbol{\sigma}), \quad (13.6.1)$$

where \mathcal{D}_I^* is an auxiliary vector on the same space as \mathcal{D}_I . In this augmented Bayesian model, the weighting function $\pi(\mathcal{D}_I | \mathcal{D}_I^*, \mathbf{f}, \boldsymbol{\sigma})$ weights the intractable posterior. In this paper we con-

sider the hierarchical model assumption, where we work with $p(\mathcal{D}_I|\mathcal{D}_I^*, \mathbf{f}, \boldsymbol{\sigma}) = g(\mathcal{D}_I|\mathcal{D}_I^*)$, see Reeves and Pettitt (22).

The mechanism in the ABC framework which allows one to avoid the evaluation of the intractable likelihood involves replacing this evaluation with data simulation from the likelihood. That is, given a realisation of the parameters of the model, a synthetic data set, \mathcal{D}_I^* , is generated and compared to the original data set. This is a key aspect of the novel methodology we develop in this paper, since we utilize a bootstrap procedure to perform this simulation in the DFCL model setting.

Then summary statistics, $S(\mathcal{D}_I^*)$, derived from this data are compared to summary statistics of the observed data, $S(\mathcal{D}_I)$, and a distance is calculated, $\rho(S(\mathcal{D}_I^*), S(\mathcal{D}_I))$. Finally, a weight is given to these parameters according to the weighting function $g(\mathcal{D}_I|\mathcal{D}_I^*)$, which may give greater weight when $S(\mathcal{D}_I^*)$ and $S(\mathcal{D}_I)$ are close, (i.e. where $\rho(S(\mathcal{D}_I^*), S(\mathcal{D}_I))$ is small).

For example under the "Hard Decision" (HD) weighting given by

$$g(\mathcal{D}_I|\mathcal{D}_I^*) \propto \begin{cases} 1 & \text{if } \rho(S(\mathcal{D}_I), S(\mathcal{D}_I^*)) \leq \epsilon, \\ 0 & \text{otherwise;} \end{cases} \quad (13.6.2)$$

a reward is given to summary statistics of the augmented auxiliary variables, $S(\mathcal{D}_I^*)$, within an ϵ -tolerance of the summary statistic of the actual observed data, $S(\mathcal{D}_I)$, as measured by distance metric ρ .

Hence, in the ABC context, an approximation to the intractable target posterior marginal distribution, $\pi(\mathbf{f}, \boldsymbol{\sigma}|\mathcal{D}_I)$, for which we are interested in formulating an empirical estimate, is given by

$$\pi_{ABC}(\mathbf{f}, \boldsymbol{\sigma}|\mathcal{D}_I, \epsilon) \propto \int \int \int \pi(\mathcal{D}_I|\mathcal{D}_I^*, \mathbf{f}, \boldsymbol{\sigma}) p(\mathcal{D}_I^*|\mathbf{f}, \boldsymbol{\sigma}) \pi(\mathbf{f}, \boldsymbol{\sigma}) \pi(\mathbf{f}, \boldsymbol{\sigma}) d\mathbf{f} d\boldsymbol{\sigma} d\mathcal{D}_I^*. \quad (13.6.3)$$

As briefly mentioned, obtaining samples from the ABC posterior can be achieved using a number of numerical procedures, in this paper we consider an MCMC approach. The MCMC class of likelihood-free algorithm is justified on a joint space formulation, in which the stationary distribution of the Markov chain is given by $\pi_{ABC}(\mathbf{f}, \boldsymbol{\sigma}, \mathcal{D}_I^*|\mathcal{D}_I, \epsilon)$. The corresponding target distribution for the marginal distribution $\pi_{ABC}(\mathbf{f}, \boldsymbol{\sigma}|\mathcal{D}_I, \epsilon)$ is then obtained via numerical integration. Note that the marginal posterior distribution $\pi_{ABC}(\mathbf{f}, \boldsymbol{\sigma}|\mathcal{D}_I, \epsilon) \rightarrow \pi(\mathbf{f}, \boldsymbol{\sigma}|\mathcal{D}_I)$ as $\epsilon \rightarrow 0$, recovering the "true" (intractable) posterior, assuming that $S(\mathcal{D}_I)$ are sufficient statistics and that the weighting function converges to a point mass on $S(\mathcal{D}_I)$ as $\epsilon \rightarrow 0$, see Peters and Sisson (15) and references therein for detailed discussion. Accordingly, the tolerance ϵ is typically set as low as possible for a given computational budget. In this paper we focus on the class of MCMC-based sampling algorithms.

The ABC methodology is novel both in the statistics literature and in the actuarial literature, as such it is informative to clearly provide the justification for this approach both theoretically and numerically. The simplest understanding of ABC is achieved by considering a rejection

algorithm, therefore we provide a basic argument for how the ABC methodology works in simple rejection sampling in the Appendix. The actuarial DFCL model considered in this paper requires the more sophisticated MCMC-ABC methodology described below.

13.6.2 Technical justification for MCMC-ABC algorithm

For given observations \mathcal{D}_I we want to sample from $\pi_{ABC}(\mathbf{f}, \boldsymbol{\sigma} | \mathcal{D}_I)$ with an intractable likelihood function. We assume that $S(\mathcal{D}_I)$ is either the data itself or a summary of the data such as a sufficient statistic for the model from which we assume data \mathcal{D}_I is a realization. We assume that given a set of parameters values $(\mathbf{f}, \boldsymbol{\sigma})$ we can generate from the DFCL model (via a conditional bootstrap procedure) a synthetic data set denoted \mathcal{D}_I^* . We define a hard decision function $g(\mathcal{D}_I^*, \mathcal{D}_I) = \mathbb{I}\{\rho(S(\mathcal{D}_I^*), S(\mathcal{D}_I)) < \epsilon\}(\mathcal{D}_I^*)$ for a given tolerance level $\epsilon > 0$ and a distance metric $\rho(\cdot, \cdot)$, where $\mathbb{I}\{\cdot\}$ is the indicator function which equals 1 if the event is true and 0 otherwise. As demonstrated in the Appendix we use the approximation, see 13.10.3-13.10.4, which gives us in the Bayesian DFCL model setting,

$$\pi_{ABC}(\mathbf{f}, \boldsymbol{\sigma} | \mathcal{D}_I) \approx \frac{\int g(\mathcal{D}_I^*, \mathcal{D}_I) \pi(\mathbf{f}) \pi(\boldsymbol{\sigma}) \pi(\mathcal{D}_I^* | \mathbf{f}, \boldsymbol{\sigma}) d\mathcal{D}_I^*}{\int \int g(\mathcal{D}_I^*, \mathcal{D}_I) \pi(\mathbf{f}) \pi(\boldsymbol{\sigma}) \pi(\mathcal{D}_I^* | \mathbf{f}, \boldsymbol{\sigma}) d\mathcal{D}_I^* d\mathbf{f} d\boldsymbol{\sigma}} = \frac{\pi(\mathbf{f}, \boldsymbol{\sigma}) E[g(\mathcal{D}_I^*, \mathcal{D}_I) | \mathbf{f}, \boldsymbol{\sigma}]}{E[g(\mathcal{D}_I^*, \mathcal{D}_I)]}, \quad (13.6.4)$$

where $\mathcal{D}_I^* \sim \pi(\mathcal{D}_I^* | \mathbf{f}, \boldsymbol{\sigma})$. In the next step the numerator of 13.6.4 is approximated using the empirical distribution, i.e.

$$\pi_{ABC}(\mathbf{f}, \boldsymbol{\sigma}) E[g(\mathcal{D}_I^*, \mathcal{D}_I) | \mathbf{f}, \boldsymbol{\sigma}] \approx \pi(\mathbf{f}) \pi(\boldsymbol{\sigma}) \frac{1}{L} \sum_{l=1}^L g(\mathcal{D}_I^{*,(l)}(\mathbf{f}, \boldsymbol{\sigma}), \mathcal{D}_I), \quad (13.6.5)$$

where $\mathcal{D}_I^{*,(l)}(\mathbf{f}, \boldsymbol{\sigma}) \stackrel{i.i.d.}{\sim} \pi(\mathcal{D}_I^* | \mathbf{f}, \boldsymbol{\sigma})$. Finally, we need to consider the denominator $E[g(X, y)]$. In general this has a non-trivial form that cannot be calculated analytically. However, since we use an MCMC based method the denominators cancel in the accept-reject stage of the algorithm. Therefore the intractability of the denominator does not impede sampling from the posterior. Therefore, we use

$$\begin{aligned} \pi_{ABC}(\mathbf{f}, \boldsymbol{\sigma} | \mathcal{D}_I) &\approx \frac{\int g(\mathcal{D}_I^*, \mathcal{D}_I) \pi(\mathbf{f}) \pi(\boldsymbol{\sigma}) \pi(\mathcal{D}_I^* | \mathbf{f}, \boldsymbol{\sigma}) d\mathcal{D}_I^*}{\int \int g(\mathcal{D}_I^*, \mathcal{D}_I) \pi(\mathbf{f}) \pi(\boldsymbol{\sigma}) \pi(\mathcal{D}_I^* | \mathbf{f}, \boldsymbol{\sigma}) d\mathcal{D}_I^* d\mathbf{f} d\boldsymbol{\sigma}} \\ &\propto \pi(\mathbf{f}, \boldsymbol{\sigma}) E[g(\mathcal{D}_I^*, \mathcal{D}_I) | \mathbf{f}, \boldsymbol{\sigma}] \\ &\approx \pi(\mathbf{f}) \pi(\boldsymbol{\sigma}) \frac{1}{L} \sum_{l=1}^L g(\mathcal{D}_I^{*,(l)}(\mathbf{f}, \boldsymbol{\sigma}), \mathcal{D}_I) \end{aligned} \quad (13.6.6)$$

in order to obtain samples from $\pi_{ABC}(\mathbf{f}, \boldsymbol{\sigma} | \mathcal{D}_I)$. Almost universally, $L = 1$ is adopted to reduce computation but on the other hand this will slow down the rate of convergence to the stationary distribution.

Note that sometimes one also uses softer decision functions for $g(\cdot, \cdot)$. The role of the distance measure ρ is evaluated by Peters et al. (20), we further extend this analysis to the class of mod-

els considered in this paper. We analyse several choices for the distance measure ρ such as Mahalanobis distance, scaled Euclidean distance and the Manhattan "City Block" distance. Fan et al. (6) demonstrate that it is not efficient to utilise the standard Euclidean distance, especially when summary statistics considered are on different scales.

Additionally, using an MCMC-ABC algorithm, it is important to assess convergence diagnostics. Particularly when using MCMC-ABC where serial correlation in the Markov chain samples can be significant if the sampler is not designed carefully. We assess autocorrelation of the simulated Markov chain, the Geweke (9) time series statistic and the Gelman-Rubin (7) R-statistic convergence diagnostic in an ABC setting.

Concluding: We apply three different techniques in order to treat the intractable likelihood:

1. ABC is used to get a handle on the likelihood and therefore the intractable posterior.
2. As a result of using ABC we need to be able to generate synthetic data samples from the DFCL model given realisations of the parameters, these come from the bootstrap algorithm.
3. We use a well understood MCMC based sampling algorithm that does not require calculation of the non-analytic normalizing constants for the target distribution $\pi_{ABC}(\mathbf{f}, \boldsymbol{\sigma} | \mathcal{D}_I)$. The reason for this is that in the acceptance probability of the MCMC algorithm, the normalizing constant for the target posterior appears both in the numerator and denominator, resulting in cancellation.

The specific details of the MCMC algorithm and ABC choices are provided in the Appendix.

13.7 Example 1: Analysis of MCMC-ABC bootstrap methodology on synthetic data

To test the accuracy of the methodology, first we use synthetic data generated with known parameter values. The tuning of the proposal distribution in this study is done for the simplest "base" distance metric, the weighted Euclidean distance. To study the effect of the distance metric in a comparative fashion we shall keep the proposal distribution unchanged.

The first example we present has a claims triangle of size $I = J = 9$. In this example we fix the true model parameters, denoted by $\mathbf{f} = (f_0, \dots, f_{J-1})$ and $\boldsymbol{\sigma}^2 = (\sigma_0^2, \dots, \sigma_{J-1}^2)$ and given in Table 13.2, used to generate the synthetic data set.

13.7.1 Generation of synthetic data

To generate the synthetic observations for \mathcal{D}_I we generate randomly the first column, denoted \mathcal{B}_0 . Then conditional on this realization of \mathcal{B}_0 we make use of the model given in (13.3.1) to

generate the remaining columns of \mathcal{D}_I , ensuring the model assumptions are satisfied. This requires setting $C_{i,0}$ sufficiently large, for appropriate choices of \mathbf{f} and σ^2 , and then sample i.i.d. realizations of $\varepsilon_{i,j} \sim \mathcal{U}[-\sqrt{3}, \sqrt{3}]$ which are used to obtain \mathcal{D}_I , see the observations in Table 13.2.

13.7.2 Sensitivity analysis and convergence assessment

We perform a sensitivity analysis, studying the impact of the distance metric on the mixing of the Markov chain in the case of joint estimation of the chain ladder factors and the variance parameters.

The pre-tuned coefficient of variation of the Gamma proposal distribution for each parameter of the posterior was performed using the following settings; $T_b = 50,000$, $\tilde{T} = 200,000$, $\epsilon^{min} = 0.1$ and initial values $\gamma_j = 1$ for all $j \in \{1, \dots, 2J\}$. Additionally, the prior parameters for the chain ladder factors F_j were set as $(\alpha, \beta) = (2, 1.2/2)$ and the parameters for the variance parameters Ξ_j^{-2} were set as $(a, b) = (2, 1/2)$.

After tuning the proposal distributions during burn-in and rounding the shape parameters we found that $\gamma_j = 10$ for all $j \in \{1, \dots, 2J\}$ produced average acceptance probabilities for each parameter between 0.3 and 0.5.

Then keeping the proposal distribution constant and using a common data set \mathcal{D}_I we ran three versions of the MCMC-ABC algorithm for 200,000 samples corresponding to;

1. Scaled Euclidean distance and joint estimation of posterior for \mathbf{F}, Ξ^2
2. Mahlanobis distance (modified) and joint estimation of posterior for \mathbf{F}, Ξ^2
3. Manhattan "City Block" distance and joint estimation of posterior for \mathbf{F}, Ξ^2 .

13.7.3 Convergence diagnostics

We estimate the three convergence diagnostics we presented in Section 13.10.3. The results of this analysis are presented as a function of Markov chain iteration t post burin of 50,000 samples.

Autocorrelation Function: In Figure 13.10.1 we demonstrate the estimated autocorrelation functions for the Markov chains of the random variables F_0 and Ξ_0^2 . Since the posterior parameters in this model are independent it is suitable to analyze just the marginal parameters to get a reasonable estimate of the mixing behavior of the MCMC-ABC algorithm on all the posterior parameters. The results demonstrate the degree of serial correlation in the Markov chains generated for these parameters as a function of lag time τ . The higher the decay rate in the tail of the estimated ACF as a function of τ , the better the mixing of the MCMC algorithm. Due to the independence properties of this model there is little difference between results obtained for

Scaled Euclidean and Mahalanobis distances. As shown in the Appendix the estimate of the covariance matrix is diagonal on all but the right lower 2×2 block. Hence, we recommend using the simple Scaled Euclidean distance metric as it provided the best trade-off between simplicity and mixing performance.

Geweke Time Series Diagnostic: Figure 13.10.2 presents results for the Geweke time series diagnostic. Again, we present the results for the random variables F_0 and Ξ_0^2 . Note, we used the posterior mean as the sample function and a set of increasing values for \tilde{T} from $T_b + 5,000$ increasing in steps of 5,000 samples to T . In each case we split the chain in each "window" given by $\{\theta_i^{(t)}\}_{t=1:T_1}$ and $\{\theta_i^{(t)}\}_{t=T^*:\tilde{T}}$ according to recommendations from Geweke et al. (9). We then calculate the convergence diagnostic $Z_{\tilde{T}}$ which is the difference between these 2 means divided by the asymptotic standard error of their difference. As the chain length increases $\tilde{T} \rightarrow \infty$, then the sampling distribution of $Z \rightarrow \mathcal{N}(0,1)$ if the chain has converged. Hence values of $Z_{\tilde{T}}$ in the tails of a standard normal distribution suggest that the chain was not fully converged early on (i.e. during the 1st window). Hence, we plot $Z_{\tilde{T}}$ scores versus increasing \tilde{T} and monitor if they lie within a 95% confidence interval $Z_{\tilde{T}} \in [-1.96, 1.96]$. The results in Figure 13.10.2 demonstrate that clearly the convergence properties of the distance functions differ. Again this is more material in the Markov chain for the variance parameter when compared to the Markov chain results for the chain ladder factor. The main point we note is that again one would advise against use of the "City block" distance metric.

Gelman and Rubin R statistic: In Figure 13.10.3 we present the Gelman and Rubin convergence diagnostic. To calculate this we ran 20 chains in parallel, each of length 10,000 samples and for each chain we discarded 250 samples as burn-in. We then estimated the R statistic as a function of simulation time post burn-in. In Figure 13.10.3 we demonstrate the convergence rate of the R statistic to 1 for each distance metric on increasing blocks of 200 samples. Using this summary statistic all three distance metrics are very similar in terms of convergence rate of the R statistic to 1.

Overall, these three convergence diagnostics demonstrate the simple scaled Euclidean distance metric is the superior choice. Secondly, we see appropriate convergence of the Markov chains under three convergence diagnostics which test different aspects of the mixing of the Markov chains, giving confidence in the performance of the MCMC-ABC algorithm in this model.

13.7.4 Bayesian parameter estimates

In this section we present results for the Scaled Euclidean distance metric, with a Markov chain of length 200,000 samples discarding the first 50,000 samples as burn-in. In Tables 13.3 and 13.4 we present the Chain Ladder parameter estimates for the DFCL model and the associated parameter estimation error. We define the following quantities:

- $\hat{f}_j^{(MAP)} | \sigma_{0:J-1}$, $\hat{f}_j^{(MMSE)} | \sigma_{0:J-1}$, $\hat{\sigma}_{f_j} | \sigma_{0:J-1}$ and $[\hat{q}_{0.05}, \hat{q}_{0.95}] | \sigma_{0:J-1}$ denote respectively the Maximum a-Posteriori, Minimum Mean Square Error, posterior standard deviation of the conditional distribution of chain ladder factor F_j and the posterior coverage probability

estimates at 5% of the conditional distribution of chain ladder factor F_j , each of these estimates is conditional on knowledge of the true $\sigma_{0:J-1}$.

- $\hat{f}_j^{(MAP)}$, $\hat{f}_j^{(MMSE)}$, $\hat{\sigma}_{f_j}$ and $[\hat{q}_{0.05}, \hat{q}_{0.95}]$ denote the same quantities for the unconditional distribution after joint estimation of $F_{0:J-1}, \Xi_{0:J-1}$.
- $Ave.[A(\theta_{1:2J}, f_j)]$ and $Ave.[A(\theta_{1:2J}, \sigma_j)]$ denotes the average acceptance probability of the Markov chain.
- $\hat{\sigma}_j^{2(MAP)}$, $\hat{\sigma}_j^{2(MMSE)}$, $\hat{\sigma}_{\sigma_j^2}$ and $[\hat{q}_{0.05}, \hat{q}_{0.95}]$ denotes the same quantities for the chain ladder variances as those defined above for chain ladder factors.

For the frequentist approach we obtain the standard error in the estimates by using 1,000 bootstrap realizations of $\{\mathcal{D}_I^{(s)}\}_{s=1:1,000}$ to obtain $\{\tilde{f}_j^{(CCL)}, \tilde{\sigma}_j^{2(CCL)}\}_{s=1:1,000}$. We use these bootstrap samples to calculate the standard deviation in the estimates of the parameters in the classical frequentist CL approach, present in brackets (.) next to their corresponding estimators. The standard errors in the Bayesian parameter estimates are obtained by blocking the Markov chain into 100 blocks of length 1,500 samples and estimating the posterior quantities on each block.

13.8 Example 2: Real Claims Reserving data

In this example we consider estimation using real claims reserving data from Wüthrich-Merz (24), see Table 13.3. This yearly loss data is turned into annual cumulative claims and divided by 10,000 for the analysis in this example. We use the analysis from the previous study to justify use of the joint MCMC-ABC simulation algorithm with a Scaled Euclidean distance metric.

We pre-tuned the coefficient of variation of the Gamma proposal distribution for each parameter of the posterior. This was performed using the following settings; $T_b = 50,000$, $\tilde{T} = 200,000$, $\epsilon^{min} = 10^{-5}$ and initial values $\gamma_j = 1$ for all $j \in \{1, \dots, 2J\}$. Here we make a strict requirement of the tolerance level to ensure we have accurate results from our ABC approximation. Additionally, the prior parameters for the chain ladder factors F_j were set as $(\alpha_j, \beta_j) = (1, \hat{f}_j^{(CL)})$ and the parameters for the variance Ξ_j^{-2} priors were set as $(a_j, b_j) = (1, \hat{\sigma}_j^{(CL)})$. The code for this problem was written in Matlab and it took approximately 10 min to simulate 200,000 samples from the MCMC-ABC algorithm, on a Intel Xeon 3.4GHz processor with 2Gb RAM.

After tuning the proposal distributions during burn-in we obtained rounded shape parameters $\gamma_{1:9} = [50, 100, 500, 500, 5,000, 20,000, 100,000, 2,000,000, 3,000,000]$ provided average acceptance probabilities between 0.3 and 0.5.

Estimates of f and σ

In Figures 13.10.4 we present box-whisker plots of estimates of the distributions of the parameters $F_{0:J-1}, \Xi_{0:J-1}$ obtained from the MCMC-ABC algorithm, post burn-in. Figure 13.10.5 presents the Bayesian MCMC-ABC empirical distributions of the ultimate claims, $C_{i,J}$ for $i =$

$1, \dots, I$. In Table 13.5 we present the predicted cumulative claims for each year along with the estimates for the chain ladder factors and chain ladder variances under both the classical approach and the Bayesian model. We see that with this fairly vague prior specified, we do indeed obtain convergence of the MCMC-ABC based Bayesian estimates $\hat{\mathbf{f}}^{(MMSE)}, \hat{\boldsymbol{\sigma}}^{(MMSE)}$ to the classical estimates $\hat{\mathbf{f}}^{(CL)}, \hat{\boldsymbol{\sigma}}^{(CL)}$.

Dependence on tolerance ϵ

In Figure 13.10.6 we present a study of the histogram estimate of the marginal posterior distribution for chain ladder factor $\pi(f_0 | \mathcal{D}_I, \epsilon^{min})$. The plot was obtained by sampling from the full posterior $\pi(\mathbf{f}, \boldsymbol{\sigma} | \mathcal{D}_I, \epsilon^{min})$ for each specified tolerance value, ϵ^{min} . Then the samples for the particular chain ladder parameter in each plot are turned into a smoothed histogram estimate for each epsilon and plotted. The results of this analysis demonstrated that when ϵ is large, in this model greater than around $\epsilon_{min} = 0.1$, the likelihood is not having an influence on the ABC posterior distribution. Hence, under an MCMC-ABC algorithm, this results in acceptance probabilities for the chain being artificially high, resulting in estimates of the posterior which reflect the prior distribution used (in this case a vague prior). As ϵ_{min} is reduced, we notice that the changes in the estimate of the posterior distribution also reduces. The aim of this study is to demonstrate that once ϵ_{min} reaches a small enough level, the effect of reducing it further is minimal on the posterior distribution. We see that changing ϵ_{min} from 10^{-4} to 10^{-5} has not had a material impact on the posterior mean or variance, the change is less than 10%. As a result, reducing ϵ_{min} past this point can not be justified relative to the significant increase in computational effort required to achieve such a further reduction in ϵ_{min} .

Ultimately, we would like an algorithm which could work well for any ϵ^{min} , the smaller the better. However, we note that with a decreasing ϵ^{min} in the sampler we present in this paper, one must take additional care to ensure the Markov chain is still mixing and not "stuck" in a particular state, as is observed to be the case in all MCMC-ABC algorithms. To avoid this acknowledged difficulty with MCMC-ABC requires that one either runs much longer MCMC chains or it requires the use of more sophisticated sampling algorithms such as SMC Samplers PRC-ABC based algorithms, see Sisson, Peters, Fan and Briers (2008).

The conclusion of these findings is that a value of $\epsilon_{min} = 10^{-5}$ which was used for the analysis of the data in this paper is suitable numerically and computationally.

VaR and MSEP.

In Table 13.6 we present the predictive VaR at 95% and 99% levels for the ultimate predicted claims, obtained from the MCMC-ABC algorithm. These are easily obtained under the Bayesian setting, we simply used the MCMC-ABC posterior samples to explicitly obtain samples from the full predictive distribution of the cumulative claims after integrating out the parameter uncertainty numerically. In addition to this we present the analysis of the MSEP under the bootstrap frequentist procedure and the Bayesian MCMC-ABC and credibility estimates for the total predicted cumulative claims for each accident year i . We also present results for the sum of the total cumulative claims for each accident year and the associated parameter uncertainty and process variance (see section 13.5 for details).

We can make the following conclusions from these results:

1. The estimates of process variance for each $C_{i,J}$ demonstrate that the frequentist bootstrap and the credibility estimates are very close for all accident years i . The Bayesian results compare favorably with the credibility results.
2. The results for the parameter estimation error for the predicted cumulative claims $C_{i,J}$ demonstrate for small i the Bayesian approach results in a smaller estimation error compared to the frequentist approach. For large i , the Bayesian approach produces larger estimation error, relative to the credibility approach.
3. The total results for the process variance for $C = \sum_i C_{i,J}$ demonstrate that the frequentist and credibility results are very close. Additionally, Bayesian total results are largest followed by credibility and then frequentist estimates which is in agreement with theoretical bounds.
4. The total results for the parameter estimation error for $C = \sum_i C_{i,J}$ demonstrate that frequentist unconditional bootstrap procedure results in the lowest total error. The Bayesian approach and credibility total parameter errors are close. Additionally, we note that the results in Wüthrich-Merz [(24), Table 7.1] for the total parameter estimation error under an unconditional frequentist bootstrap with unscaled residuals is also very close to the total obtained for the frequentist approach.

13.9 Discussion

This paper has presented a distribution-free claims reserving model under a Bayesian paradigm. A novel advanced MCMC-ABC algorithm was developed to obtain estimates from the resulting intractable posterior distribution of the chain ladder factors and chain ladder variances. We assessed several aspects of this algorithm, including the properties of the convergence of the MCMC algorithm as a function of the distance metric approximation in the ABC component. The methodologies performance was demonstrated on a synthetic data set generated from known parameters. Next it was applied to a real claims reserving data set. The results we obtained for predicted cumulative ultimate claims were compared to those obtained via classical chain ladder methods and via credibility theory. This clearly demonstrated that the algorithm is working accurately and provides us not only with the ability to obtain point estimates for the first and second moments of the ultimate cumulative claims, but an accurate empirical approximation of the entire distribution of the ultimate claims. This is valuable for many reasons, including prediction of reserves which are not based on centrality measures, such as the tail based VaR results we present.

Acknowledgements

The first author thanks ETH FIM and ETH Risk Lab for their generous financial assistance whilst completing aspects of this work at ETH. The first author also thanks the Department of Mathematics and Statistics at the University of NSW for support through an Australian Postgraduate Award and to CSIRO for support through a postgraduate research top up scholarship. Finally, this material was based upon work partially supported by the National Science Foundation under Grant DMS-0635449 to the Statistical and Applied Mathematical Sciences Institute, North Carolina, USA. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation

13.10 Appendix

13.10.1 Section 1

ABC algorithm

The ABC algorithm is typically justified in the simple rejection sampling framework. This then extends in a straightforward manner to other sampling frameworks such as the MCMC algorithm we utilise in this paper. We denote the posterior density from which we wish to draw samples by $\pi(\theta|y) \propto \pi(y|\theta)\pi(\theta)$ with $\theta \in \Omega$, where Ω denotes support of the posterior distribution and \mathcal{Y} is the support for y .

The ABC method aims to draw from this posterior density $\pi(\theta|y)$ without the requirement of evaluating the computationally expensive or in our setting intractable likelihood $\pi(y|\theta)$. The cost of avoiding this calculation is that we obtain an "approximation".

1st case. We assume that the support \mathcal{Y} is discrete. Given an observation $y \in \mathcal{Y}$ we would like to sample from $\pi(\theta|y)$. Then the original rejection sampling algorithm reads as follows:

Rejection Sampling ABC

1. Sample θ' from prior $\pi(\theta)$
2. Simulate synthetic data set of auxiliary variables $x|\theta' \sim \pi(x|\theta')$
3. ABC Rejection Condition: If $x = y$ then accept sample θ' , else reject sample and return to step 1.

Then the chosen $\theta' \sim \pi(\theta|y)$ by simple rejection argument: Denote $\{x = y\}$ if θ' was chosen. Then, the joint density of (θ', x) , conditional on $\{y, x = y\}$, is given by

$$\pi(\theta, x|y, x = y) = \frac{\pi(\theta)\pi(x|\theta)\mathbb{I}\{y\}(x)}{\int \pi(\theta)\pi(y|\theta)d\theta} = \begin{cases} \frac{\pi(\theta,y)}{\pi(y)} = \pi(\theta|y) & \text{if } x = y, \\ 0 & \text{otherwise.} \end{cases} \quad (13.10.1)$$

This implies that

$$\sum_{x \in \mathcal{Y}} \pi(\theta, x|y, x = y) = \pi(\theta|y). \quad (13.10.2)$$

Henceforth, this algorithm generates samples $\theta^{(t)} \sim \pi(\theta|y)$, for $t = 1, \dots, T$.

2nd case. For more general supports \mathcal{Y} one replaces the strict equality $x = y$ with a tolerance $\epsilon > 0$ and a measure of discrepancy or a distance metric, $\rho(x, y) \leq \epsilon$. In this case the posterior distribution is given by $\pi(\theta, x|y, \rho(x, y) < \epsilon)$. Implementing this algorithm in a rejection sampling framework gives the following:

Rejection Sampling ABC

1. Sample θ' from prior $\pi(\theta)$
2. Simulate synthetic data set of auxiliary variables $x|\theta' \sim \pi(x|\theta')$
3. ABC Rejection Condition 2: If $\rho(x, y) < \epsilon$ then accept sample θ' , else reject sample and return to step 1.

In this case the joint density of (θ', x) , conditional on $\{y, \rho(x, y) < \epsilon\}$, is given by

$$\pi(\theta, x|y, \rho(x, y) < \epsilon) = \frac{\pi(\theta) \pi(x|\theta) \mathbb{I}\{\rho(x, y) < \epsilon\}(x)}{\int \pi(\theta) \pi(x|\theta) \mathbb{I}\{\rho(x, y) < \epsilon\}(x) dx d\theta}. \quad (13.10.3)$$

Note that for appropriate choices of the distance metric ρ and assuming the necessary continuity properties for the densities we obtain that

$$\lim_{\epsilon \rightarrow 0} \int_{\mathcal{Y}} \pi(\theta, x|y, \rho(x, y) < \epsilon) dx = \pi(\theta|y). \quad (13.10.4)$$

This concept was taken further with the intention of improving the simulation efficiency by reducing the number of rejected samples. To achieve this sufficient statistics were used to replace the comparison between the auxiliary variables ("synthetic data") x and the observations y . Denoting the sufficient statistics by $S(y)$ and $S(x)$, then this would allow one to decompose the likelihood under the Fisher-Neyman factorization theorem into, $\pi(y|\theta) = f(y)g(S(y)|\theta)$ for appropriate functions f and g . In the ABC context presented above, the consequence of this decomposition is that when $\rho(S(y), S(x)) < \epsilon$ then we obtain samples from the posterior density $\pi(\theta, x|y, \rho(S(y), S(x)) < \epsilon)$ similar to 13.10.3.

13.10.2 Section 2

MCMC-ABC to sample from $\pi_{ABC}(f, \sigma|\mathcal{D}_I)$

We develop an MCMC-ABC algorithm which has an adaptive proposal mechanism and annealing of the tolerance during burn-in of the Markov chain. Having reached the final tolerance post annealing, denoted ϵ^{\min} , we utilize the remaining burn-in samples to tune the proposal distribution to ensure an acceptance probability between the range of 0.3 and 0.5 is achieved. The optimal acceptance probability when posterior parameters are i.i.d. Gaussian was proven to be at 0.234, see Gelman et al. (8). Though our problem does not match the required conditions for this proof, it provides a practical guide. To achieve this, we tune the coefficient of variation of the proposal, in our case the shape parameter of the gamma proposal distribution. We impose an additional constraint that the minimum shape parameter value is set at γ_j^{\min} for $j \in \{1, \dots, 2J\}$.

MCMC-ABC algorithm using bootstrap samples.

1. For $t = 0$ initialize the parameter vector randomly, this gives $\theta_{1:2J}^{(0)} = \left(f_{0:J-1}^{(0)}, \sigma_{0:J-1}^{(0)} \right)$. Initialize the proposal shape parameters, $\gamma_j \geq \gamma_j^{\min}$ for all $j \in \{1, \dots, 2J\}$.
2. For $t = 1, \dots, T$
 - (a) Set $\left(\theta_{1:2J}^{(t)} \right) = \left(\theta_{1:2J}^{(t-1)} \right)$
 - (b) For $j = 1, \dots, 2J$
 - i. Sample proposal θ_j^* from a $\Gamma(\gamma_j, \theta_j^{(t)}/\gamma_j)$ -distribution. We denote the gamma proposal density by $K\left(\theta_j^*; \gamma_j, \theta_j^{(t)}/\gamma_j\right)$. This gives proposed parameter vector $\theta^* = \left(\theta_{1:j-1}^{(t)}, \theta_j^*, \theta_{j+1:2J}^{(t)} \right)$.
 - ii. Conditional on $\theta^* = \left(\theta_{1:j-1}^{(t)}, \theta_j^*, \theta_{j+1:2J}^{(t)} \right)$, generate synthetic bootstrap data set $\mathcal{D}_I^* = \mathcal{D}_I^*(\theta^*)$ using the bootstrap procedure detailed in Section 13.5 where we replace the CL parameter estimates $(\hat{f}^{(CL)}, \hat{\sigma}^{(CL)})$ by the parameters θ^* .
 - iii. Evaluate summary statistics $S(\mathcal{D}_I; 0, 1)$ and $S(\mathcal{D}_I^*; \mu^*; s^*)$ and corresponding decision function $g(\mathcal{D}_I, \mathcal{D}_I^*)$ as described in Section 13.10.3.
 - iv. Accept proposal with ABC acceptance probability

$$A\left(\theta_{1:2J}^{(t)}, \theta^*\right) = \min \left\{ 1, \frac{\pi\left(\theta_j^*\right) K\left(\theta_j^{(t)}; \gamma_j, \theta_j^*/\gamma_j\right)}{\pi\left(\theta_j^{(t)}\right) K\left(\theta_j^*; \gamma_j, \theta_j^{(t)}/\gamma_j\right)} g(\mathcal{D}_I, \mathcal{D}_I^*) \right\}.$$

That is, simulate $U \sim \mathcal{U}(0, 1)$ and set $\theta_j^{(t)} = \theta_j^*$ if $U < A\left(\theta_{1:2J}^{(t)}, \theta^*\right)$.

- v. If $100 \leq t \leq T_b$ and $\epsilon_t = \epsilon^{\min}$ then check to see if require tuning of the proposal. Define the average acceptance probability over the last 100 iterations of updates for parameter i by $\bar{a}_i^{(t-100:t)}$ and consider the adaption,

$$\gamma_j^* = \begin{cases} 0.9\gamma_j & \text{if } \bar{a}_i^{(t-100:t)} < 0.3 \text{ and } \gamma_j > \gamma_j^{\min}, \\ 1.1\gamma_j & \text{if } \bar{a}_i^{(t-100:t)} > 0.5, \\ \gamma_j & \text{otherwise.} \end{cases}$$

Then set the proposal shape parameter as,

$$\gamma_j = \max\{\gamma_j^*, \gamma_j^{\min}\}.$$

The MCMC-ABC algorithm presented can be enhanced by utilizing an idea of Gramacy et al. (12) in an ABC setting. This involves a combination of tempering the tolerance $\{\epsilon_t\}_{t=1:T}$ and importance sampling corrections.

13.10.3 ABC algorithmic choices for the time series DFCL model

We start with the choices of the ABC components.

- **Generation of a synthetic data set:** Note that in this setting not only is the likelihood intractable but additionally the generation of a synthetic data set \mathcal{D}_I^* given the current parameter values $\mathbf{F}, \mathbf{\Xi}$ is not straightforward. The synthetic data set \mathcal{D}_I^* is generated using the bootstrap procedure described in Section 13.5. Note that both the bootstrap residual $\tilde{\varepsilon}_{i,j}$ and the bootstrap samples \mathcal{D}_I^* are functions of the parameter choices. Therefore we generate for given $\mathbf{F} = \mathbf{f}$ and $\mathbf{\Xi} = \boldsymbol{\sigma}$ the bootstrap residuals $\tilde{\varepsilon}_{i,j} = \tilde{\varepsilon}_{i,j}(f_{j-1}, \sigma_{j-1})$ and the bootstrap samples $\mathcal{D}_I^* = \mathcal{D}_I^*(\mathbf{f}, \boldsymbol{\sigma})$ according to the non-parametric bootstrap where we replace the CL parameter estimates $(\hat{\mathbf{f}}^{(CL)}, \hat{\boldsymbol{\sigma}}^{(CL)})$ by the parameters $\theta = (\mathbf{F}, \mathbf{\Xi})$.
- **Summary statistics:** In order to define the decision function g we introduce summary statistics, see Appendix 13.10.1 for details of the role these summary statistics play in the ABC approximation. For the observed data \mathcal{D}_I we define the vector

$$\begin{aligned} S(\mathcal{D}_I; 0, 1) &= (S_1, \dots, S_{n+2}) \\ &= (C_{0,1}, \dots, C_{0,J}, C_{1,1}, \dots, C_{0,J-1}, \dots, C_{I-2,1}, C_{I-2,2}, C_{I-1,1}; 0, 1), \end{aligned}$$

where n denotes the number of residuals $\tilde{\varepsilon}_{i,j}$. For given $\theta = (\mathbf{F}, \mathbf{\Xi})$ we generate the bootstrap sample $\mathcal{D}_I^* = \mathcal{D}_I^*(\mathbf{F}, \mathbf{\Xi})$ as described above. The corresponding residuals $\tilde{\varepsilon}_{i,j} = \tilde{\varepsilon}_{i,j}(F_{j-1}, \Xi_{j-1})$ should also be close to the standardized observations. Therefore we define its empirical mean and standard deviation by

$$\mu^* = \mu^*(\mathbf{F}, \mathbf{\Xi}) = \frac{1}{n} \sum_{i,j} \tilde{\varepsilon}_{i,j}(F_{j-1}, \Xi_{j-1}), \quad (13.10.5)$$

$$s^* = s^*(\mathbf{F}, \mathbf{\Xi}) = \left[\frac{1}{n-1} \sum_{i,j} (\tilde{\varepsilon}_{i,j}(F_{j-1}, \Xi_{j-1}) - \mu^*(\mathbf{F}, \mathbf{\Xi}))^2 \right]^{1/2}. \quad (13.10.6)$$

Hence, the summary statistics for the synthetic data is given by

$$S(\mathcal{D}_I^*; \mu^*, s^*) = (C_{0,1}^*, \dots, C_{0,J}^*, C_{1,1}^*, \dots, C_{0,J-1}^*, \dots, C_{I-2,1}^*, C_{I-2,2}^*, C_{I-1,1}^*; \mu^*, s^*).$$

- **Distance metrics:**

– *Mahlanobis distance and scaled Euclidean distance*

Here we draw on the analysis of Sisson et al. (6) who propose the use of the Mahlanobis distance metric given by,

$$\begin{aligned} \rho(S(\mathcal{D}_I; 0, 1), S(\mathcal{D}_I^*; \mu^*, s^*)) \\ = [S(\mathcal{D}_I; 0, 1) - S(\mathcal{D}_I^*; \mu^*, s^*)]^\top \Sigma_{\mathcal{D}_I}^{-1} [S(\mathcal{D}_I; 0, 1) - S(\mathcal{D}_I^*; \mu^*, s^*)], \end{aligned}$$

where the covariance matrix $\Sigma_{\mathcal{D}_I}$ is an appropriate scaling described in Appendix 13.10.4. The scaled Euclidean distance is obtained when we only consider the diagonal elements of the covariance matrix $\Sigma_{\mathcal{D}_I}$.

Note, the covariance matrix $\Sigma_{\mathcal{D}_I}$ provides a weighting on each element of the vector of summary statistics to ensure they are scaled appropriately according to their influence on the ABC approximation. There are many other such weighting schemes one could conceive.

– *Manhattan "City Block" distance*

We consider the L^1 -distance given by

$$\rho(S(\mathcal{D}_I; 0, 1), S(\mathcal{D}_I^*; \mu^*, s^*)) = \sum_{i=1}^{n+2} |S_i(\mathcal{D}_I; 0, 1) - S_i(\mathcal{D}_I^*; \mu^*, s^*)|.$$

- **Decision function:** We work with a hard decision function given by

$$g(\mathcal{D}_I, \mathcal{D}_I^*) = \mathbb{I}\{\rho(S(\mathcal{D}_I; 0, 1), S(\mathcal{D}_I^*; \mu^*, s^*)) < \epsilon\}.$$

- **Tolerance schedule:** We use the sequence

$$\epsilon_t = \max\{20,000 - 10t, \epsilon^{\min}\}.$$

Note, the use of an MCMC-ABC algorithm can result in "sticking" of the chain for extended periods. Therefore, one should carefully monitor convergence diagnostics of the resulting Markov chain for a given tolerance schedule. There is a trade-off between the length of the Markov chain required for samples approximately from the stationary distribution and the bias introduced by non zero tolerance. In this paper we set ϵ^{\min} via preliminary analysis of the Markov chain sampler mixing rates for a transition kernel with coefficient of variation set to one.

We note that in general practitioners will have a required precision in posterior estimates, this precision can be directly used to determine, for a given computational budget, a suitable tolerance ϵ^{\min} .

- **Convergence diagnostics:** We stress that when using an MCMC-ABC algorithm, it is crucial to carefully monitor the convergence diagnostics of the Markov chain. This is more important in the ABC context than in the general MCMC context due to the possibility of extended rejections where the Markov chain can stick in a given state for long periods. This can be combated in several ways which will be discussed once the algorithm is presented.

The convergence diagnostics we consider are evaluated only on samples post annealing of the tolerance threshold and after an initial burn-in period once tolerance of ϵ^{\min} is reached. If the total chain has length T , the initial burn-in stage will correspond to the first T_b samples and we define $\tilde{T} = T - T_b$. We denote by $\{\theta_i^{(t)}\}_{t=1:\tilde{T}}$ the Markov chain of the i -th parameter after burn-in. The diagnostics we consider are given by,

- *Autocorrelation*. This convergence diagnostic will monitor serial correlation in the Markov chain. For given Markov chain samples for the i -th parameter $\{\theta_i^{(t)}\}_{t=1:\tilde{T}}$ we define the biased autocorrelation estimate at lag τ by

$$\widehat{ACF}(\theta_i, \tau) = \frac{1}{(\tilde{T} - \tau)\hat{\sigma}(\theta_i)} \sum_{t=1}^{\tilde{T}-\tau} [\theta_i^{(t)} - \hat{\mu}(\theta_i)][\theta_i^{(t+\tau)} - \hat{\mu}(\theta_i)], \quad (13.10.7)$$

where $\hat{\mu}(\theta_i)$ and $\hat{\sigma}(\theta_i)$ are the estimated mean and standard deviation of θ_i .

- *Geweke (9) time series diagnostic*. For parameter θ_i we calculate:

1. Split the Markov chain samples into two sequences, $\{\theta_i^{(t)}\}_{t=1:T_1}$ and $\{\theta_i^{(t)}\}_{t=T^*:\tilde{T}}$, such that $T^* = \tilde{T} - T_2 + 1$, and with ratios T_1/\tilde{T} and T_2/\tilde{T} fixed such that $(T_1 + T_2)/\tilde{T} < 1$ for all \tilde{T} .
2. Evaluate $\hat{\mu}(\theta_i^{T_1})$ and $\hat{\mu}(\theta_i^{T_2})$ corresponding to the sample means on each sub sequence.
3. Evaluate consistent spectral density estimates for each sub sequence, at frequency 0, denoted $\widehat{SD}(0; T_1, \theta_i)$ and $\widehat{SD}(0; T_2, \theta_i)$. The spectral density estimator considered in this paper is the classical non-parametric periodogram or power spectral density estimator. We use Welch's method with a Hanning window, for details see Appendix 13.10.5.

4. Evaluate convergence diagnostic given by

$$Z_{\tilde{T}} = \frac{\hat{\mu}(\theta_i^{T_1}) - \hat{\mu}(\theta_i^{T_2})}{T_1^{-1}\widehat{SD}(0; T_1, \theta_i) + T_2^{-1}\widehat{SD}(0; T_2, \theta_i)}.$$

For $\tilde{T} \rightarrow \infty$ one has according to the central limit theorem that $Z_{\tilde{T}} \rightarrow \mathcal{N}(0, 1)$ if the sequence $\{\theta_i^{(t)}\}_{t=1:\tilde{T}}$ is stationary.

- *Gelman-Rubin (7) R-statistic diagnostic*. This approach to convergence analysis requires that one runs multiple parallel independent Markov chains each starting at randomly selected initial starting points, we run 5 chains. For comparison purposes we split the total computational budget of \tilde{T} into $T_1 = T_2 = \dots = T_5 = \tilde{T}/5$. We compute the convergence diagnostic for parameter θ_i with the following steps:

1. Generate 5 independent Markov chain sequences, producing the chains for parameter θ_i , denoted $\{\theta_{i,k}^{(t)}\}_{t=1:T_k}$ for $k \in \{1, \dots, 5\}$.
2. Calculate the sample means $\hat{\mu}(\theta_i^{T_k})$ for each sequence, and the overall mean $\hat{\mu}(\theta_i^{\tilde{T}})$.
3. Calculate the variance of the sequence means, $\frac{1}{4} \sum_{k=1}^5 \left(\hat{\mu}(\theta_i^{T_k}) - \hat{\mu}(\theta_i^{\tilde{T}}) \right)^2 \stackrel{def.}{=} B_i/T_k$.
4. Calculate the within-sequence variances $\hat{s}^2(\theta_i^{T_k})$ for each sequence.
5. Calculate the average within-sequence variance, $\frac{1}{5} \sum_{k=1}^5 \hat{s}^2(\theta_i^{T_k}) \stackrel{def.}{=} W_i$.
6. Estimate the target posterior variance for parameter θ_i by the weighted linear combination, $\hat{\sigma}^2(\theta_i^{\tilde{T}}) = \frac{T_k-1}{T_k} W_i + \frac{1}{T_k} B_i$. This estimate is unbiased for samples which are from the stationary distribution. In the case in which not all sub chains

have reached stationarity, this overestimates the posterior variance for a finite \tilde{T} but it asymptotically, $\tilde{T} \rightarrow \infty$, converges to the posterior variance.

7. Improve on the Gaussian estimate of the target posterior given by $\mathcal{N}(\hat{\mu}(\theta_i^{\tilde{T}}), \hat{\sigma}^2(\theta_i^{\tilde{T}}))$ by accounting for sampling variability in the estimates of the posterior mean and variance. This can be achieved by making a Student-t approximation with location $\hat{\mu}(\theta_i^{\tilde{T}})$, scale $\sqrt{\hat{V}_i}$ and degrees of freedom df_i , each given respectively by;

$$\hat{V}_i = \hat{\sigma}^2(\theta_i^{\tilde{T}}) + \frac{B_i}{\tilde{T}} \text{ and } df_i = \frac{2\hat{V}_i^2}{\widehat{\text{Var}}(\hat{V}_i)}, \text{ where the variance is estimated as,}$$

$$\begin{aligned} \widehat{\text{Var}}(\hat{V}_i) &= \frac{1}{5} \left(\frac{T_1 - 1}{T_1} \right)^2 \widehat{\text{Var}}(\hat{s}^2(\theta_i^{T_k})) + \left(\frac{6}{\sqrt{2\tilde{T}}} \right)^2 B_i^2 \\ &+ \frac{12(T_1 - 1)}{25T_1} \widehat{\text{Cov}}(\hat{s}^2(\theta_i^{T_k}), \hat{\mu}(\theta_i^{\tilde{T}})) \\ &- \frac{24(T_1 - 1)}{25T_1} \hat{\mu}(\theta_i^{\tilde{T}}) \widehat{\text{Cov}}(\hat{s}^2(\theta_i^{T_k}), \hat{\mu}(\theta_i^{\tilde{T}})). \end{aligned} \tag{13.10.8}$$

Note, the covariance terms are estimated empirically using the within sequence estimates of the mean and variance obtained for each sequence.

8. Calculate the convergence diagnostic, $\sqrt{\hat{R}} = \sqrt{\frac{\hat{V}_i df_i}{W_i(df_i - 2)}}$, where as $\tilde{T} \rightarrow \infty$ one can prove that $\hat{R} \rightarrow 1$. This convergence diagnostic monitors the scale factor by which the current distribution for θ_i may be reduced if simulations are continued for $\tilde{T} \rightarrow \infty$.

13.10.4 Section 3

Scaling of statistics in distance metrics

In the Mahlanobis distance metric, estimation of the scaling weights given by the covariance $\Sigma_{\mathcal{D}_I} = \text{Cov}(S(\mathcal{D}_I^*; \tilde{\mu}, \tilde{s}) | \mathcal{D}_I)$, where $\tilde{\mu}$ and \tilde{s} are the sample mean and standard deviation of n i.i.d. residuals $\varepsilon_{i,j}$ (see also 13.10.5-13.10.6). Next we outline the estimation of $\Sigma_{\mathcal{D}_I}$ by a matrix $\hat{\Sigma}_{\mathcal{D}_I}^{CL}$.

- Starting with the elements $\hat{\Sigma}_{\mathcal{D}_I}^{CL}(k, l)$ with $k, l \in \{1, \dots, n\}$, we obtain from the conditional resampling bootstrap

$$\begin{aligned} - \text{Cov}\left(C_{i,j}^*, C_{i',j'}^* \mid \mathcal{D}_I, \hat{\mathbf{f}}^{(CL)}, \hat{\boldsymbol{\sigma}}^{(CL)}\right) &= 0 \text{ if } i \neq i' \text{ or } j \neq j' \\ - \text{Var}\left(C_{i,j}^* \mid \mathcal{D}_I, \hat{\mathbf{f}}^{(CL)}, \hat{\boldsymbol{\sigma}}^{(CL)}\right) &= \hat{\sigma}_{j-1}^{2(CL)} C_{i,j-1}. \end{aligned}$$

- Considering the elements $k \in \{n + 1, n + 2\}, l \in \{1, \dots, n\}$ and also $k \in \{1, \dots, n\}, l \in \{n + 1, n + 2\}$ of the covariance matrix $\Sigma_{\mathcal{D}_I}$, for simplicity we set $\hat{\Sigma}_{\mathcal{D}_I}^{CL}(k, l) = 0$.
- Considering elements $k, l \in \{n + 1, n + 2\}$, we assess now, $\text{Cov}(\tilde{\mu}, \tilde{s})$, either analytically or numerically by simulation of appropriate i.i.d. residuals.

Parametric Approximation

- In approximating $\tilde{\mu}$ and \tilde{s} we assume i.i.d. samples $\varepsilon_{i,j} \sim \mathcal{N}(0, 1)$.
- Using the assumptions we know,

$$\text{Var}(\tilde{\mu}) = \frac{1}{n},$$

$$\text{Var}(\tilde{s}) = \frac{1}{(n-1)^2} \left[\left(1 + \frac{4}{n^2} + \frac{1}{n^2}\right) \sum_{s=1}^n \text{Var}(\tilde{\varepsilon}_s^2) \right] = \frac{1}{(n-1)^2} [2n(1 + \frac{5}{n^2})],$$

$$\text{Cov}(\tilde{\mu}, \tilde{s}) = \frac{1}{2(n-1)^2} [1 - \frac{2}{n}].$$
- Under these assumptions,
 1. If the distribution of $\varepsilon_{i,j}$ is skewed then it is more appropriate to do a numerical approximation with the observed residuals from the bootstrap algorithm.
 2. The precision ϵ_t from the MCMC-ABC algorithm should depend on the size of the claims triangle, i.e. the number of residuals n .

13.10.5 Section 4

Estimating the Spectral Density

This is calculated via a modified technique using Welch's method, see Proakis-Manolakis (21), pages 910-913 . This involves performing the following steps:

- Split each sequence, $\{\theta_i^{(t)}\}_{t=1:T_1}$ and $\{\theta_i^{(t)}\}_{t=T^*:\tilde{T}}$, into $L = 20$ non-overlapping blocks of length N .
- Apply a Hanning window function $w(t) = 0.5 \left(1 - \cos\left(\frac{2\pi t}{N-1}\right)\right)$ to the samples of the Markov chain in each block.
- Take the discrete Fourier transform (DFT) of each windowed block given by, $\tilde{\Theta}_i^l(k) = \sum_{t=0}^{N-1} \theta_i^{(t)} \exp^{-\frac{2\pi i k t}{N}}$.
- Estimate the spectral density (SD) as, $\widehat{SD}(w_k) = \frac{1}{L} \sum_{l=0}^{L-1} \tilde{\Theta}_i^l(k)$.

| Year | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|--------------|--------|--------|--------|--------|--------|--------|--------|--------|--------|----------|
| 0 | 248.97 | 299.47 | 357.00 | 418.61 | 473.63 | 563.35 | 693.22 | 796.84 | 914.95 | 1,084.24 |
| 1 | 186.72 | 201.99 | 227.23 | 271.18 | 305.16 | 379.37 | 466.16 | 554.30 | 660.75 | |
| 2 | 172.58 | 207.48 | 250.37 | 304.44 | 356.92 | 417.60 | 477.99 | 542.25 | | |
| 3 | 195.19 | 229.06 | 290.83 | 320.11 | 367.60 | 469.93 | 543.40 | | | |
| 4 | 131.00 | 168.50 | 198.18 | 219.26 | 270.00 | 344.63 | | | | |
| 5 | 163.58 | 181.16 | 222.10 | 246.78 | 303.00 | | | | | |
| 6 | 294.30 | 373.08 | 477.16 | 566.20 | | | | | | |
| 7 | 529.31 | 577.71 | 805.95 | | | | | | | |
| 8 | 249.00 | 321.83 | | | | | | | | |
| 9 | 140.41 | | | | | | | | | |
| f_j | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 | 1.2 |
| σ_j^2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

Tab. 13.2: Synthetic Data - Cumulative claims $C_{i,j}$ for each accident year i and development year j , $i + j \leq I$.

| Year | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|------|----------|----------|---------|---------|---------|---------|--------|--------|--------|--------|
| 0 | 594.6975 | 372.1236 | 89.5717 | 20.7760 | 20.6704 | 6.2124 | 6.5813 | 1.4850 | 1.1130 | 1.5813 |
| 1 | 634.6756 | 324.6406 | 72.3222 | 15.1797 | 6.7824 | 3.6603 | 5.2752 | 1.1186 | 1.1646 | |
| 2 | 626.9090 | 297.6223 | 84.7053 | 26.2768 | 15.2703 | 6.5444 | 5.3545 | 0.8924 | | |
| 3 | 586.3015 | 268.3224 | 72.2532 | 19.0653 | 13.2976 | 8.8340 | 4.3329 | | | |
| 4 | 577.8885 | 274.5229 | 65.3894 | 27.3395 | 23.0288 | 10.5224 | | | | |
| 5 | 618.4793 | 282.8338 | 57.2765 | 24.4899 | 10.4957 | | | | | |
| 6 | 560.0184 | 289.3207 | 56.3114 | 22.5517 | | | | | | |
| 7 | 528.8066 | 244.0103 | 52.8043 | | | | | | | |
| 8 | 529.0793 | 235.7936 | | | | | | | | |
| 9 | 567.5568 | | | | | | | | | |

Tab. 13.3: Real Data - Incremental claims $Y_{i,j} = C_{i,j} - C_{i,j-1}$ for each accident year i and development year j , $i + j \leq I$.

| DFCL model | $j = 0$ | $j = 1$ | $j = 2$ | $j = 3$ | $j = 4$ | $j = 5$ | $j = 6$ | $j = 7$ | $j = 8$ |
|---|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|
| f_j | 1.20 | 1.20 | 1.20 | 1.20 | 1.20 | 1.20 | 1.20 | 1.20 | 1.20 |
| $\hat{f}_j^{(CL)}$ | 1.20 (2.40E-2) | 1.22 (3.27E-2) | 1.16 (2.46E-2) | 1.17 (2.44E-2) | 1.23 (2.63E-2) | 1.19 (2.78E-2) | 1.16 (2.59E-2) | 1.17 (2.10E-2) | 1.19 (2.51E-2) |
| $\hat{f}_j^{(MAP)} \sigma_{0:J-1}$ | 1.07 (0.02) | 1.19 (0.02) | 1.05 (0.02) | 1.04 (0.02) | 1.10 (0.02) | 1.08 (0.02) | 0.97 (0.02) | 1.19 (0.03) | 1.14 (0.04) |
| $\hat{f}_j^{(MMSE)} \sigma_{0:J-1}$ | 1.19 (1.34E-2) | 1.21 (1.38E-2) | 1.18 (1.27E-2) | 1.19 (1.30E-2) | 1.17 (1.37E-2) | 1.18 (1.53E-2) | 1.20 (1.60E-2) | 1.18 (1.73E-2) | 1.19 (2.35E-2) |
| $\hat{\sigma}_{f_j} \sigma_{0:J-1}$ | 0.23 (4.00E-3) | 0.22 (3.1E-3) | 0.20 (3.1E-3) | 0.21 (3.2E-3) | 0.22 (3.9E-3) | 0.27 (1.01E-2) | 0.35 (1.24E-2) | 0.44 (1.41E-2) | 0.70 (1.60E-2) |
| $[\hat{q}_{0.05}, \hat{q}_{0.95}] \sigma_{0:J-1}$ | [0.75,1.50] | [0.77,1.50] | [0.76,1.41] | [0.75,1.44] | [0.82,1.51] | [0.78,1.52] | [0.65,1.60] | [0.46,1.79] | [0.25,2.50] |
| $\hat{f}_j^{(MAP)}$ | 1.15 (0.02) | 1.13 (0.02) | 1.06 (0.02) | 1.09 (0.02) | 1.15 (0.02) | 1.19 (0.02) | 1.12 (0.03) | 1.08 (0.03) | 1.06 (0.04) |
| $\hat{f}_j^{(MMSE)}$ | 1.19 (0.01) | 1.18 (0.01) | 1.17 (0.01) | 1.18 (0.01) | 1.16 (0.01) | 1.20 (0.02) | 1.18 (0.03) | 1.16 (0.02) | 1.20 (0.02) |
| $\hat{\sigma}_{f_j}$ | 0.24 (5.1E-3) | 0.24 (4.4E-3) | 0.23 (5.0E-3) | 0.26 (5.8E-3) | 0.25 (5.6E-3) | 0.25 (5.7E-3) | 0.40 (0.01) | 0.49 (0.02) | 0.68 (0.02) |
| $[\hat{q}_{0.05}, \hat{q}_{0.95}]$ | [0.66,1.48] | [0.74,1.54] | [0.67,1.42] | [0.65,1.47] | [0.74,1.50] | [0.74,1.50] | [0.22,1.54] | [0.35,1.95] | [0.1,2.50] |
| $Ave.[A(\theta_{1:2J}, f_j)]$ | 0.21 | 0.21 | 0.19 | 0.22 | 0.25 | 0.21 | 0.22 | 0.20 | 0.24 |
| σ_j^2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| $\hat{\sigma}_j^{2(CL)}$ | 1.02 (0.29) | 0.75 (1.44) | 0.51 (1.02) | 0.49 (0.91) | 0.71 (1.18) | 0.72 (1.89) | 0.25 (1.84) | 0.31 (1.40) | 0.25 (0.77) |
| $\hat{\sigma}_j^{2(MAP)}$ | 0.58 (0.06) | 0.96 (0.06) | 0.54 (0.05) | 0.78 (0.05) | 0.78 (0.05) | 0.81 (0.04) | 0.61 (0.04) | 0.79 (0.04) | 0.56 (0.04) |
| $\hat{\sigma}_j^{2(MMSE)}$ | 1.11 (0.03) | 1.18 (0.03) | 1.14 (0.04) | 1.31 (0.03) | 1.29 (0.03) | 1.19 (0.02) | 1.16 (0.03) | 1.14 (0.03) | 1.05 (0.02) |
| $\hat{\sigma}_{\sigma_j}$ | 0.83 (0.02) | 0.79 (0.02) | 0.82 (0.02) | 0.80 (0.02) | 0.79 (0.02) | 0.72 (0.02) | 0.77 (0.02) | 0.78 (0.02) | 0.71 (0.02) |
| $[\hat{q}_{0.05}, \hat{q}_{0.95}]$ | [0.33,2.89] | [0.33,2.79] | [0.25,2.91] | [0.32,2.87] | [0.33,2.82] | [0.27,2.59] | [0.21,2.66] | [0.17,2.62] | [0.22,2.42] |
| $Ave.[A(\theta_{1:2J}, \sigma_j)]$ | 0.23 | 0.24 | 0.24 | 0.23 | 0.24 | 0.24 | 0.24 | 0.24 | 0.25 |

Tab. 13.4: Comparison of Bayesian estimates for the chain ladder factors and variances versus Classical estimates, for synthetic data. Numerical standard errors in estimates are presented in brackets.

| Parameters | Year | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | $\widehat{C}_{i,J}^{(CL)} - C_{i,I-i}$ |
|--------------------------|------|---------|-----------|-----------|-----------|-----------|------------|------------|------------|------------|------------|--|
| $f^{(CL)}$ | 0 | | | | | | | | | | | 0 |
| $f^{(MMSE)}$ | | | | | | | | | | | | 0 |
| $f^{(CL)}$ | 1 | | | | | | | | | | 10,663,318 | 15,126 |
| $f^{(MMSE)}$ | | | | | | | | | | | 10,663,099 | 14,907 |
| $f^{(CL)}$ | 2 | | | | | | | | | 10,646,884 | 10,662,008 | 26,257 |
| $f^{(MMSE)}$ | | | | | | | | | | 10,646,386 | 10,661,291 | 25,541 |
| $f^{(CL)}$ | 3 | | | | | | | | 9,734,574 | 9,744,764 | 9,758,606 | 34,538 |
| $f^{(MMSE)}$ | | | | | | | | | 9,734,765 | 9,744,500 | 9,758,143 | 34,074 |
| $f^{(CL)}$ | 4 | | | | | | | 9,837,277 | 9,847,906 | 9,858,214 | 9,872,218 | 85,302 |
| $f^{(MMSE)}$ | | | | | | | | 9,835,850 | 9,846,669 | 9,856,516 | 9,870,315 | 83,400 |
| $f^{(CL)}$ | 5 | | | | | | 10,005,044 | 10,056,528 | 10,067,393 | 10,077,931 | 10,092,247 | 156,494 |
| $f^{(MMSE)}$ | | | | | | | 10,005,302 | 10,055,329 | 10,066,390 | 10,076,456 | 10,090,563 | 154,811 |
| $f^{(CL)}$ | 6 | | | | | 9,419,776 | 9,485,469 | 9,534,279 | 9,544,580 | 9,554,571 | 9,568,143 | 286,121 |
| $f^{(MMSE)}$ | | | | | | 9,400,832 | 9,466,638 | 9,513,971 | 9,524,436 | 9,533,961 | 9,547,308 | 265,286 |
| $f^{(CL)}$ | 7 | | | | 8,445,057 | 8,570,389 | 8,630,159 | 8,674,568 | 8,683,940 | 8,693,030 | 8,705,378 | 449,167 |
| $f^{(MMSE)}$ | | | | | 8,437,023 | 8,545,017 | 8,604,832 | 8,647,856 | 8,657,369 | 8,666,026 | 8,678,159 | 421,947 |
| $f^{(CL)}$ | 8 | | | 8,243,496 | 8,432,051 | 8,557,190 | 8,616,868 | 8,661,208 | 8,670,566 | 8,679,642 | 8,691,971 | 1,043,242 |
| $f^{(MMSE)}$ | | | | 8,236,916 | 8,417,305 | 8,525,046 | 8,584,722 | 8,627,645 | 8,637,136 | 8,645,773 | 8,657,877 | 1,009,148 |
| $f^{(CL)}$ | 9 | | 8,470,989 | 9,129,696 | 9,338,521 | 9,477,113 | 9,543,206 | 9,592,313 | 9,602,676 | 9,612,728 | 9,626,383 | 3,950,814 |
| $f^{(MMSE)}$ | | | 8,467,380 | 9,118,521 | 9,318,217 | 9,437,490 | 9,503,553 | 9,551,070 | 9,561,577 | 9,571,138 | 9,584,538 | 3,908,970 |
| $\widehat{f}_j^{(CL)}$ | | 1.4925 | 1.0778 | 1.0229 | 1.0148 | 1.0070 | 1.0051 | 1.0011 | 1.0010 | 1.0014 | | 6,047,061 |
| $\sigma_j^{(CL)}$ | | 135.253 | 33.803 | 15.760 | 19.847 | 9.336 | 2.001 | 0.823 | 0.219 | 0.059 | | 5,918,083 |
| $\widehat{f}_j^{(MMSE)}$ | | 1.4919 | 1.0769 | 1.0219 | 1.0128 | 1.0070 | 1.0050 | 1.0011 | 1.0010 | 1.0014 | | |
| $\sigma_j^{(MMSE)}$ | | 154.221 | 33.000 | 16.770 | 22.397 | 8.300 | 2.166 | 0.720 | 0.158 | 0.041 | | |

Tab. 13.5: Predicted cumulative CL claims $\widehat{C}_{i,j}^{(CL)}$ for actual data and estimated CL reserves $\widehat{C}_{i,J}^{(CL)} - C_{i,J-i}$ under the classical and Bayesian DFCL models.

| Accident Year i | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | Total |
|---|-------|-------|--------|--------|--------|---------|---------|---------|-----------|-----------|
| $(C_{i,I-i} \hat{\Gamma}_{I-i}^{freq})^{1/2}$ | 192 | 740 | 2,668 | 6,831 | 30,474 | 68,207 | 80,071 | 126,952 | 389,768 | 424,361 |
| $(C_{i,I-i}^2 \hat{\Delta}_{I-i}^{freq})^{1/2}$ | 503 | 1,560 | 3,059 | 12,639 | 25,761 | 20,776 | 33,771 | 41,554 | 108,547 | 157,680 |
| $(\text{mse}_{C_{i,J} \mathcal{D}_I}^{freq}(\hat{C}_{i,J}))^{1/2}$ | 538 | 1,727 | 4,059 | 14,367 | 39,904 | 71,301 | 86,901 | 133,580 | 404,601 | 452,708 |
| $Vco_i(\%)$ | 3.61% | 6.76% | 11.91% | 17.02% | 25.61% | 25.00% | 19.38% | 12.81% | 9.93% | 7.49% |
| $(C_{i,I-i} \hat{\Gamma}_{I-i}^{Bayes})^{1/2}$ | 134 | 533 | 2,307 | 7,185 | 27,367 | 74,235 | 86,404 | 129,038 | 437,482 | 470,982 |
| $(C_{i,I-i}^2 \hat{\Delta}_{I-i}^{Bayes})^{1/2}$ | 224 | 894 | 1,801 | 4,327 | 15,819 | 29,861 | 32,243 | 49,198 | 152,879 | 211,633 |
| $(\text{mse}_{C_{i,J} \mathcal{D}_I}^{Bayes}(\hat{C}_{i,J}))^{1/2}$ | 261 | 1,040 | 2,927 | 8,387 | 31,610 | 80,016 | 92,224 | 138,099 | 463,425 | 504,934 |
| $Vco_i(\%)$ | 1.75% | 4.07% | 8.59% | 10.06% | 20.42% | 30.16% | 21.86% | 13.68% | 11.86% | 8.53% |
| $\text{VaR}_{0.95}^{Bayes}(C_{i,J} - E[C_{i,J} \mathcal{D}_I] \mathcal{D}_I)$ | 554 | 2,183 | 5,632 | 15,820 | 61,122 | 152,531 | 173,665 | 161,619 | 816,701 | 910,757 |
| $\text{VaR}_{0.99}^{Bayes}(C_{i,J} - E[C_{i,J} \mathcal{D}_I] \mathcal{D}_I)$ | 726 | 2,918 | 7,430 | 22,515 | 79,472 | 201,322 | 228,448 | 211,125 | 1,278,665 | 1,454,966 |
| $(C_{i,I-i} \hat{\Gamma}_{I-i}^{cred})^{1/2}$ | 192 | 740 | 2,668 | 6,831 | 30,474 | 68,207 | 80,071 | 126,952 | 389,769 | 424,362 |
| $(C_{i,I-i}^2 \hat{\Delta}_{I-i}^{cred})^{1/2}$ | 188 | 534 | 1,493 | 3,391 | 13,515 | 27,284 | 29,674 | 43,901 | 129,764 | 185,015 |
| $(\text{mse}_{C_{i,J} \mathcal{D}_I}^{cred}(\hat{C}_{i,J}))^{1/2}$ | 269 | 913 | 3,057 | 7,627 | 33,337 | 73,462 | 85,392 | 134,329 | 410,802 | 462,941 |
| $Vco_i(\%)$ | 1.81% | 3.58% | 8.97% | 9.04% | 21.40% | 25.77% | 19.04% | 12.88% | 10.40% | 7.82% |

Tab. 13.6: Comparison of the frequentist's bootstrap $\text{mse}_{C_{i,J}|\mathcal{D}_I}^{freq}$, the Bayesian MCMC-ABC $\text{mse}_{C_{i,J}|\mathcal{D}_I}^{Bayes}$ and the credibility $\text{mse}_{C_{i,J}|\mathcal{D}_I}^{cred}$. The coefficient of variation is as defined in Wüthrich-Merz (24)

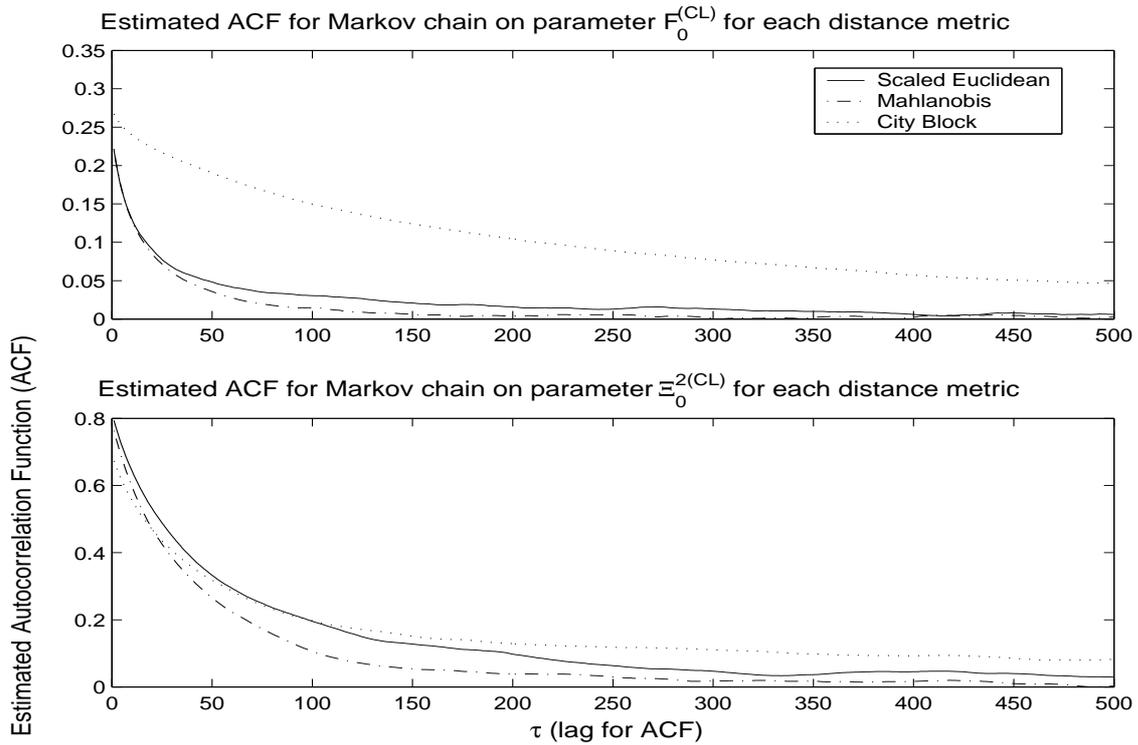


Fig. 13.10.1: Estimated Autocorrelation Function (ACF) for parameters F_0 and Ξ_0^2

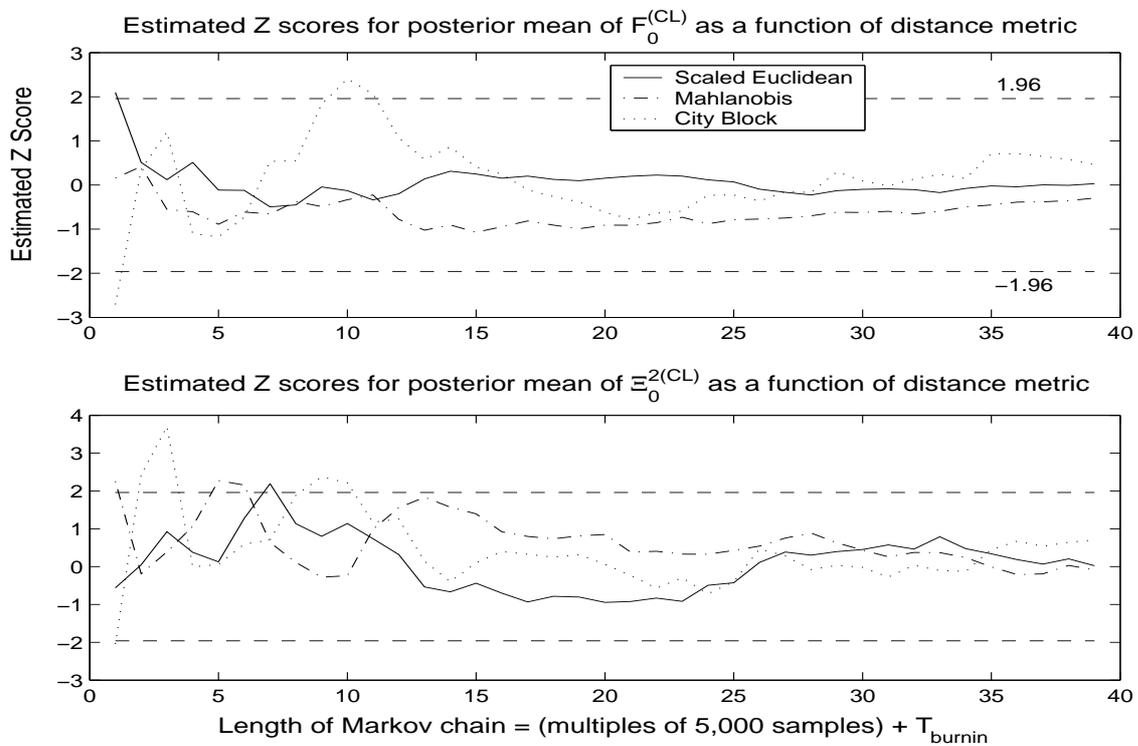


Fig. 13.10.2: Estimated Z scores for the posterior mean of parameters F_0 and Ξ_0^2 as a function of the length of the Markov chain \bar{T} .

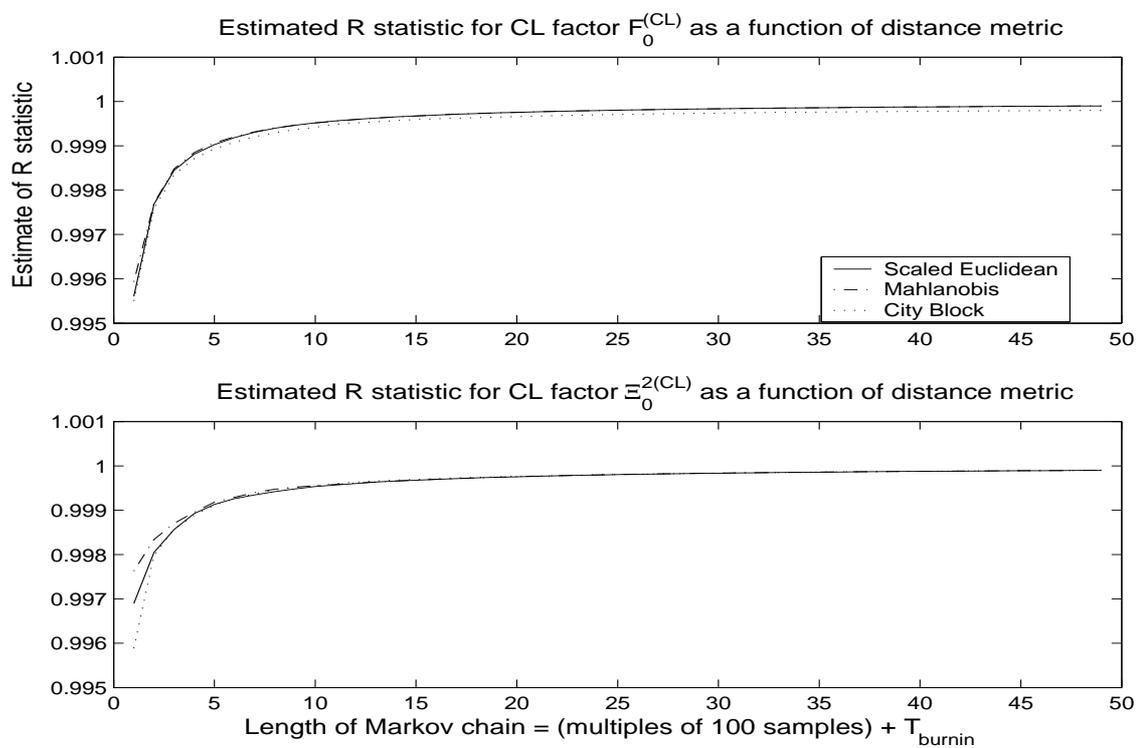


Fig. 13.10.3: Estimated R statistic for parameters F_0 and Ξ_0^2 as a function of the length of the Markov chain \tilde{T} .

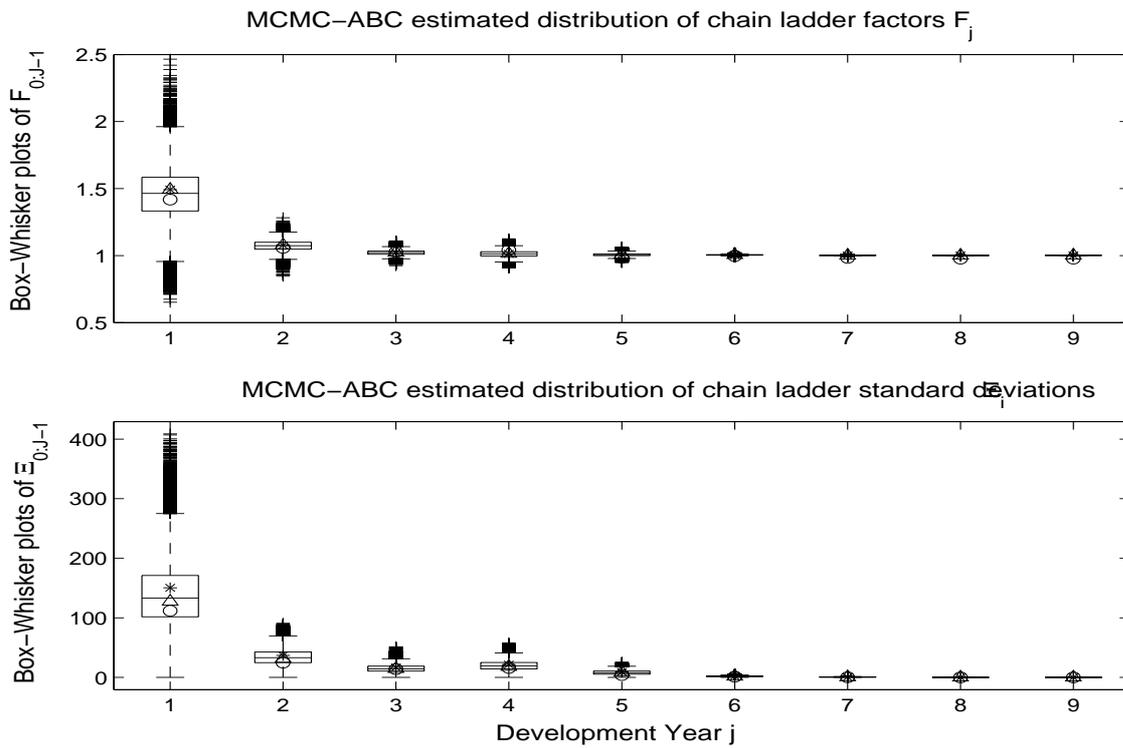


Fig. 13.10.4: Box-Whisker plots of parameters F and E with each box marking the 25th, 50th, 75th percentiles. Top: 200,000 MCMC-ABC samples to estimate posterior for F . The sample mean and mode are denoted by '*' and 'o' respectively. The classical estimators $\hat{f}^{(CL)}$ are denoted by Δ . Bottom: 200,000 MCMC-ABC samples to estimate posterior for E . The sample mean and mode are denoted by '*' and 'o' respectively. The classical estimators $\hat{\sigma}^{(CL)}$ are denoted by Δ .

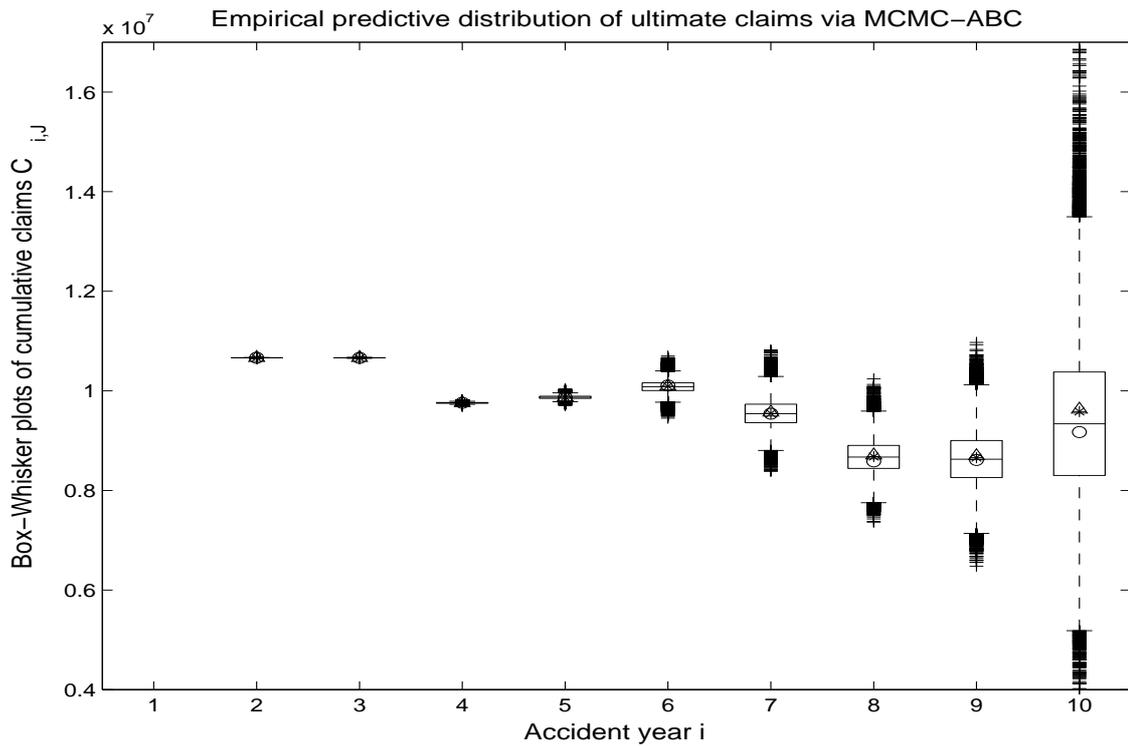


Fig. 13.10.5: Box-Whisker plots of predictive distribution of cumulative ultimate claims $C_{1:J}$ with the box marking the 25th, 50th, 75th percentiles, see also Table 6. The mean predicted ultimate claims under a Bayesian approach (using MMSE point estimates) are marked with *, the predicted mode for the ultimate claims (using MAP point estimates) is marked with o and the mean predicted ultimate claims under the DFCL classical method are marked with Δ .

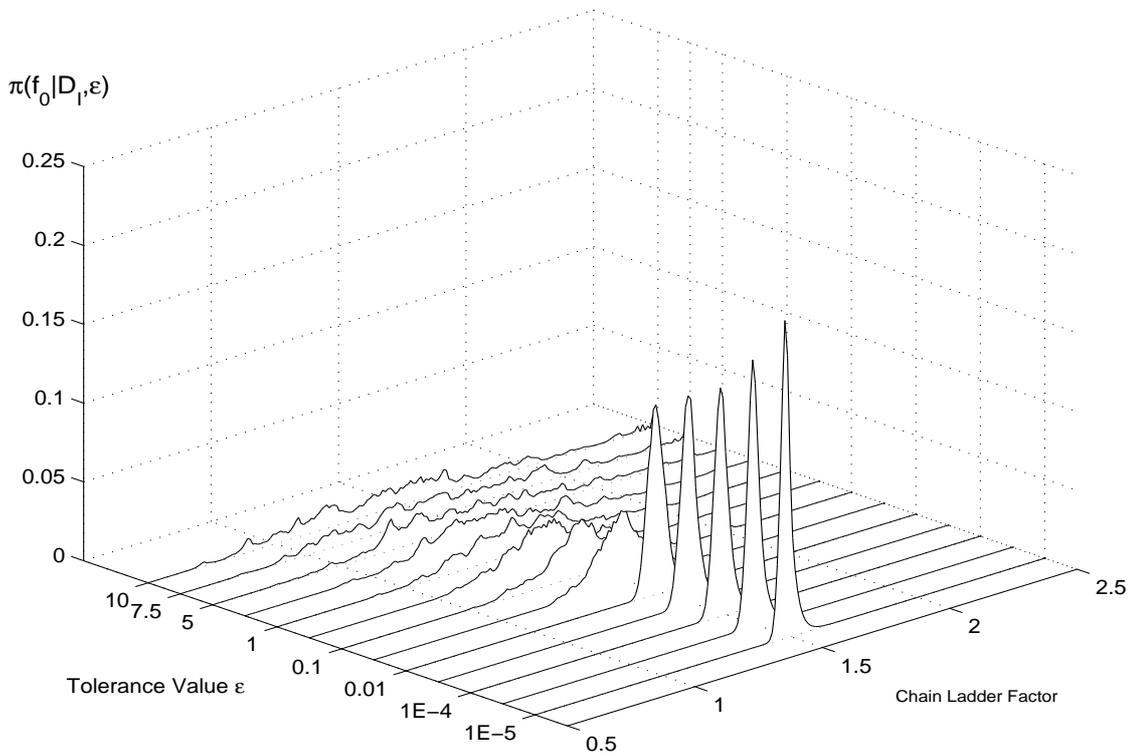


Fig. 13.10.6: Distribution of the chain ladder factor F_0 as a function of tolerance.

References

- [1] Davison, A.C., Hinkley, D.V. (1997). *Bootstrap Methods and Their Application*. Cambridge University Press, Cambridge.
- [2] Del Moral, P., Doucet, A., Jasra, A. (2006). Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society Series. B.* **68**(3), 411-436.
- [3] Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics.* **7**(1), 1-26.
- [4] Efron, B., Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall, NY.
- [5] England, P.D., Verrall, R.J. (2002). Stochastic claims reserving in general insurance. *British Actuarial Journal.* **8**(3), 443-518.
- [6] Fan, Y., Sisson, S.A., Peters, G.W. (2008). Improved efficiency in approximate Bayesian computation. *Technical report, Statistics Department, University of New South Wales*.
- [7] Gelman, A., Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* **4**, 457-472.
- [8] Gelman, A., Gilks, W.R., Roberts, G.O. (1997). Weak convergence and optimal scaling of random walk metropolis algorithm. *Annals of Applied Probability* **7**, 110-120.
- [9] Geweke, J.F. (1991). *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*. In J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith (eds.) *Bayesian Statistics, 4*, Oxford University Press, Oxford.
- [10] Gilks, W.R., Richardson, S., Spiegelhalter, D.J. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.
- [11] Gisler, A., Wüthrich, M.V. (2008). Credibility for the chain ladder reserving method. *ASTIN Bulletin* **38**(2), 483-526.
- [12] Gramacy, R.B., Samworth, R.J., King, R. (2008). Importance tempering. *Preprint, arXiv: 0707.4242v5 [stat.Co]*.
- [13] Mack, T. (1993). Distribution-free calculation of the standard error of chain ladder reserve estimates. *ASTIN Bulletin* **23**, 213-225.

-
- [14] Marjoram, P., Molitor, J., Plagnol, V., Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Science USA* **100**, 15324-15328.
- [15] Peters, G.W., Sisson, S.A. (2006). Bayesian inference, Monte Carlo sampling and operational risk. *Journal of Operational Risk* **1**(3), 27-50.
- [16] Peters, G.W., Fan, Y., Sisson, S.A. (2008). On Sequential Monte Carlo, partial rejection control and approximate Bayesian computation. *Preprint, Statistics Department, University of New South Wales*.
- [17] Peters, G.W., Sisson, S.A., Fan, Y. (2008). Design efficiency for "likelihood free" Sequential Monte Carlo samplers. *Preprint, Statistics Department, University of New South Wales*.
- [18] Sisson, S.A., Peters, G.W., Fan, Y., Briers, M., (2008). Likelihood free samplers. *Preprint, Statistics Department, University of New South Wales*.
- [19] Peters, G.W., Shevchenko, P., Wüthrich, M.V., (2008). Model uncertainty in claims reserving within Tweedie's compound Poisson models. *ASTIN Bulletin* **39**,1-33.
- [20] Peters, G.W., Nevat, I., Sisson, S.A., Fan, Y., Yuan, J. (2009). Bayesian symbol detection in wireless relay networks. *Preprint, Statistics Department, University of New South Wales*.
- [21] Proakis, J.G., Manolakis, D.G., (1996). *Digital Signal Processing*. Upper Saddle River, N.J. Prentice Hall.
- [22] Reeves, R.W., Pettitt, A.N. (2005). A theoretical framework for approximate Bayesian computation. Presented at the International Workshop for Statistical Modelling, Sydney.
- [23] Sisson, S.A., Fan, Y., Tanaka, M. (2007). Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Science USA* **104**, 1760-1765.
- [24] Wüthrich, M.V., Merz, M. (2008). *Stochastic Claims Reserving Methods in Insurance*. Wiley Finance.
- [25] Yao, J. (2008). Bayesian approach for prediction error in chain-ladder claims reserving. *Conference paper presented at the ASTIN Colloquium, Manchester UK*.

Part III

ADVANCES IN BAYESIAN MODELS FOR TELECOMMUNICATIONS ENGINEERING

METHODOLOGY AND APPLICATION

Part III Introduction

“In the world of ideas everything was clear; in life all was obscure, embroiled.”

Aldous Huxley

In this chapter the context and justification for the development of Bayesian models for wireless communications engineering is provided. The technical details relating to likelihood-free methodology and TDMCMC samplers in these contexts is deferred to the journal papers presented in subsequent chapters of Part III. Instead the intention of this chapter is to provide a brief contextual understanding of the significance of the journal papers presented in Part III of this thesis.

To achieve this the introduction chapter of Part III comprises four sections. The first section aims to motivate advanced statistical modelling in the context of wireless communications system design. In particular it provides important contextual information and background on how statistical modelling approaches have contributed significantly to modern telecommunications engineering. The second section then focuses on relevant technical content and background for a class of systems known as Multiple Input Multiple Output communication networks. Section three details a second important aspect of wireless communications relevant to the journal papers contained in subsequent chapters of Part III, covering Orthogonal Frequency Division Multiplexing. Finally, the last section summarises the novelty and contribution introduced in each of the journal papers contained in Part I of the thesis.

14.1 Motivating advanced statistical modelling in wireless communications

The focus of the papers, contained in Part III of the thesis, involve wireless communication systems. Hence, the medium of transmission or channel is an electromagnetic bandwidth. The transmitted signal propagates through the physical medium which contains obstacles and

surfaces of reflection. This causes multiple reflections of the transmitted signal to arrive at the receiver at different times. The reflected signals are distorted and attenuated by the material properties of the surface, that they reflect from or permeate through, depending particularly on aspects such as dielectric constants, permeability, conductivity, thickness, etc..

These propagation effects on the transmitted signal are all encoded in an abstract entity known as the channel. There are many types of channel characteristics, that one can consider. In this thesis consideration is given to multipath channels. These have been classified by communications engineers into popular categories of either large scale fading or small scale fading, see [Haykin, 2001] (63), Chapter 8. The particular form of fading depends on whether the channel characterization is viewed from a frequency domain or time domain perspective, see Figure 14.1.1.

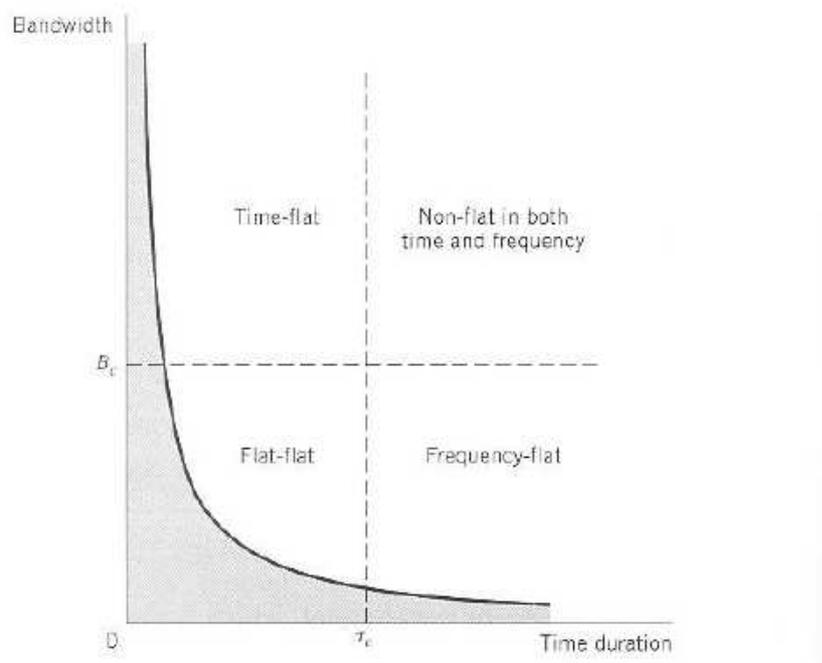


Fig. 14.1.1: Frequency and time domain characterizations

The diagram in Figure 14.1.2 summarises the modelling hierarchy for multipath channels models.

In the class of small scale fading characteristics one can model aspects of signal dispersion or time varying properties of the channel. From this sub-classification a further four popular channel characteristics, typically considered in a wireless communications system are, flat fading; frequency selective; fast fading; and slow fading. To encapsulate or model these characteristics there are a range of proposed channel models to consider such as: Rayleigh fading channels; Jake's model and the associated auto-regressive moving average approximations which include the (all-pole, all-zero and ARMA models), see [Haykin, 2001] (63).

The approach of specifying possible channel characteristics and then developing statistical models for the channel models has arisen as a result of the complexity with modelling the

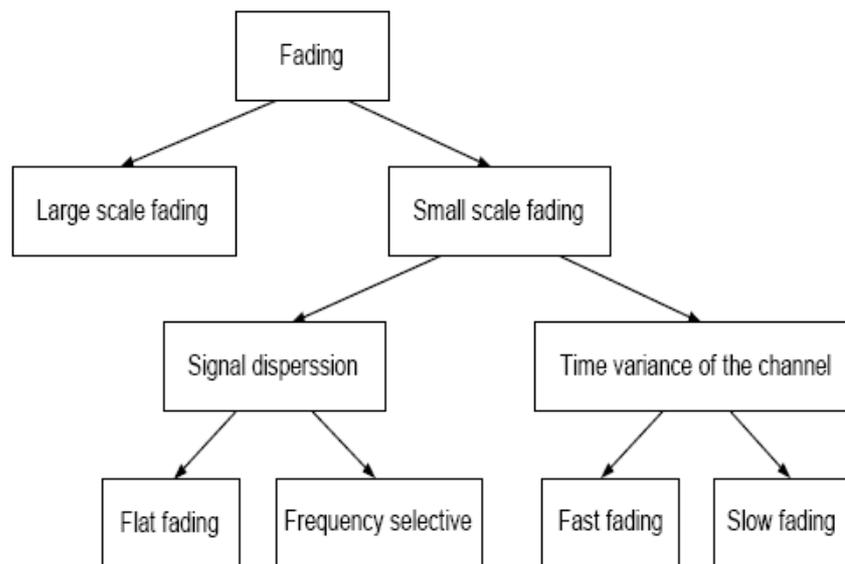


Fig. 14.1.2: Modelling hierarchy in multipath channel models

system using fundamental physical equations. In an ideal world, one would solve directly Maxwell's electromagnetic field equations for the transmitted signal to determine the electromagnetic field at the receiver. However, in practice, this is not only extremely difficult, but impractical, since it would require knowledge of the medium of propagation, such as (building locations, heights, thickness, material on surfaces, locations of trees, speed of moving transmitters etc.). Hence, engineers have developed a usable approximation framework, revolving around classification of properties of the channel and development of approximate statistical models to approximate the true physical processes.

The papers contained in Part III focus on two channel models, the first is based on frequency selective uncorrelated channels and the second is based on flat-fading uncorrelated channels. A flat-fading channel is one, which has properties, which are modelled well by a constant gain and linear phase response to the frequency bandwidth on which transmission will occur. Additionally, it will be assumed, that the channel is slow fading. This assumption states, that the channel properties are such that, the Channel Impulse Response (CIR) changes at a much slower rate than the baseband transmitted signal. Under such an assumption, the channel may be assumed to be static over one or several reciprocal bandwidth intervals.

The design of wireless communications systems must take into account, not only the properties of the channel in the allocated bandwidth of transmission, but also the amount of the bandwidth available as well as the maximum allowed radiated power, which are subject to fundamental physical, regulatory and practical constraints.

At the heart of all wireless communication systems lies Shannon's capacity formula, see [Shannon (1949)] (104) or the review [Shannon (2001)] (105). This statistical model for wireless communications is an application of the noise channel coding theorem to the case of a continuous time analog communications channel, subject to Gaussian noise. In particular, Shannon estab-

lished the channel capacity for such a communications link, which is a bound on the maximum amount of error-free digital data, that can be transmitted, with a specified bandwidth in the presence of noise interference. The assumptions of this result are, that the signal power is bounded and the Gaussian noise process is characterized by a known power spectral density.

For the particular case of a unit-gain band-limited continuous channel, corrupted with additive white Gaussian noise (AWGN), Shannon obtained the capacity as

$$C = W \log \left(\frac{P_T + P_N}{P_N} \right), \quad (14.1.1)$$

where W , P_T and P_N represent the bandwidth, the average transmitted power, and the noise power, respectively.

It is clear that under such a stochastic channel model, increasing the capacity can be achieved in several ways. The most trivial of these is to simply emit more power. This is however not practical, since there is known to be a logarithmic dependence between the spectral efficiency and the transmitted power. The second way to increase capacity would be to utilize a wider electromagnetic band. This is unfortunately also impractical in today's expected system performance criteria for high throughput communications systems. The radio spectrum is a scarce, expensive resource, which is partitioned up by frequency and auctioned for very large sums of money. Consequently, telecommunications engineers are challenged to design wireless systems capable of increased data rates and improved performance in limited frequency bands.

This has led to a range of innovative design ideas, to meet the key design criteria involving high data rates and strong reliability in a wireless communications system. The particular focus of the papers in this thesis each involve the utilization of multiple antennas. Specifically, this thesis, considers Multiple Input Multiple Output (MIMO) antenna systems, see [Foschini *et al.* (1998)] (45) and [Telatar (1999)] (115). MIMO systems are particularly important as they can be shown to increase capacity without the requirement of transmitting additional power or an increase in bandwidth. In principle, the capacity can grow linearly with the minimum number of antennas at the receiver or transmitter. In other words, in addition to diversity gain and array gain, MIMO systems offer multiplexing gain by opening parallel data streams (usually called the spatial mode or the eigen subchannels) within the same frequency band at no additional power expense.

However, in practice, a MIMO system requires behavior which does not display frequency selective characteristics on each channel. This is especially true when the transmission rate is high enough to make the whole channel frequency selective. Therefore, the use of orthogonal frequency division multiplexing (OFDM) which transforms a frequency selective channel into a large set of individual frequency non-selective narrowband channels, is well suited to be used in conjunction with MIMO systems. Put another way, OFDM is a multi-carrier modulation technique with high spectral efficiency and a simple single-tap equaliser structure. It splits the bandwidth into a number of overlapping narrow band sub-channels requiring lower symbol rates. Furthermore, the inter-symbol interference (ISI) and inter-carrier interference (ICI) can

be easily eliminated by inserting a cyclic prefix (CP) in front of each transmitted OFDM block. For more background on OFDM systems see, [Chang *et al.* (1968)] (23) and [Ruiz *et al.* (1992)] (101) and the references contained.

However, though a MIMO-OFDM system can be designed to achieve desired spectral efficiency, the use of multiple antenna elements at the transmitter of a wireless system results in superposition of multiple transmitted signals at the receiver. Each received signal is weighted by the corresponding multipath channels, making the design of the receiver significantly more complex. This thesis considers particular statistical modelling aspects associated with such systems.

In addition to this, as wireless networks continue to expand in geographical size, the distance between the source transmitter and the destination receiver is increasing. This often precludes direct communications between the transmitter and receiver due to attenuation of the signal in the propagation medium. It has therefore become popular to design systems with a repeater between the source and the destination. Implementation complexity and budgetary constraints have prevented the use of relays for mobile communications until recently. The advances in the design and production of relay components, has made it economically feasible to embed into a wireless communications systems, such as: next generation mobile phone networks; ad-hoc mobile networks; and mesh networks for wireless broadband. As a result, relay-based systems have become an important research topic in the wireless research community, see *cooperation diversity* [Laneman *et al.* (2001)] (73), [Sendonaris *et al.* (2003)] (103); *antenna arrays* [Dohler *et al.* (2006)] (33); or *multihop networks* [Pabst *et al.* (2004)] (85).

In the remainder of this chapter, a brief coverage of background models is provided, for the modelling of channels and systems studied in the journal papers contained in Part III. This includes brief overviews of MIMO systems, OFDM systems and wireless relay networks.

14.2 Multiple Input Multiple Output antenna systems

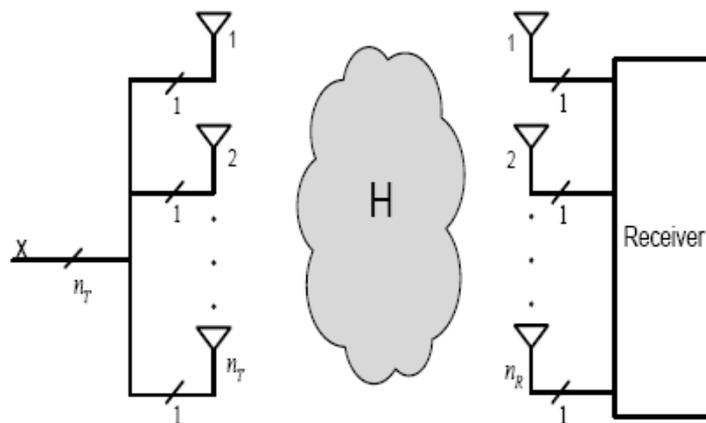


Fig. 14.2.1: MIMO systems

MIMO channels, Figure 14.2.1, are characterized by a transition probability density function, given by $p(\mathbf{y}|\mathbf{x})$, which describes the probability of receiving the vector \mathbf{y} , conditioned on the fact, that the vector \mathbf{x} was actually transmitted. It is typical to assume a linear model approximation in MIMO communications modelling,

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{w}, \quad (14.2.1)$$

where $\mathbf{x} \in \mathbb{A}^{n_T}$ is the MIMO transmitted symbol and $\mathbf{H} \in \mathbb{C}^{n_R \times n_T}$ represents the linear response of the channel, with elements $[H]_{ij}$ denoting the channel path gain between j -th transmitter and i -th receiver. In the papers contained in this thesis, the noise terms \mathbf{w} are modelled as i.i.d. circularly symmetric complex Gaussian random vectors with zero mean and covariance matrix given by $\mathbf{R}_w = \mathbf{I}\sigma_w^2$, ie.

$$\mathbb{E}\{\mathbf{w}\} = \mathbf{0}, \quad (14.2.2)$$

$$\mathbb{E}\{\mathbf{w}\mathbf{w}^H\} = \mathbf{R}_w. \quad (14.2.3)$$

14.2.1 Uncertainty models for the Channel State Information (CSI)

The CSI represents the knowledge of the statistical model of the channel that the receiver is assumed to have during transmission. In practical system designs this knowledge can be categorized into three states of knowledge,

- **No CSI:** the receiver does not have any knowledge of the CSI, or its statistics. Detection of transmitted symbols in this ignorance setting is denoted *non-coherent detection*.
- **Imperfect CSI:** the receiver has inaccurate knowledge about the parameters describing the channel. An example of such a model assumption includes situations in which the receiver has an estimated channel matrix $\hat{\mathbf{H}} \neq \mathbf{H}$, with corresponding covariance error matrix \mathbf{C} . Then the channel model, that can be used, is one in which the channel is a random matrix $\mathbf{H} \sim \mathcal{CN}(\hat{\mathbf{H}}, \mathbf{C})$. Examples of such models are considered in the papers in Part III.
- **Perfect CSI:** the receiver has full knowledge of the instantaneous channel realization. Under these conditions, the detection of symbols is termed *coherent detection*.

Model assumptions such as partial CSI may arise in settings in which one only has access to a noisy channel estimate or there is a quantization error present. It is common to perform channel estimation at the receiver. A training period is established, prior to transmission of the actual information, in which the transmitter sends a pilot or training sequence. Then depending on the complexity of the design of the system and the amount of training data transmitted, this known training sequence can be utilized in some manner to obtain one of the three states of CSI knowledge above.

14.3 Orthogonal Frequency Division Multiplexing (OFDM)

As discussed, several papers in Part III of this thesis consider OFDM systems. Hence, a brief background introduction to OFDM systems is provided. OFDM is a multi-carrier modulation scheme in which data is transmitted by dividing a single wideband stream into several smaller or narrowband parallel bit streams. Each narrowband stream is modulated onto an individual carrier. The narrowband channels are orthogonal to each other, and transmitted simultaneously. In doing so, the symbol duration is increased proportionately, which reduces the effects of inter symbol interference (ISI) induced by multipath Rayleigh-faded environments. The frequency spectra of each of the subcarriers overlaps, making OFDM more spectral efficient as opposed to conventional multicarrier communication schemes. The spectral overlap between subcarriers in OFDM schemes is achieved by using rectangular pulses for data modulation, this results in subchannels having significant spectral overlap with a large number of adjacent subchannels.

In OFDM systems, subchannels (subcarriers) are obtained via an orthogonal transformation on each block of data (OFDM symbol) comprising N subcarriers. Orthogonal transformations are used so that at the receiver, the inverse transformation can be used to demodulate the data without error in the absence of noise. The paper of [Weinstein et al. (1971)] (122) proposed the use of the discrete Fourier transform (DFT) be used for multicarrier modulation. The DFT exhibits the desired orthogonality and can be implemented efficiently through the fast Fourier transform (FFT) algorithm. When the channel distortion is mild relative to the channel bandwidth, data can be demodulated with minimal interference from the other subchannels, due to the orthogonality of the transformation. Subchannel isolation is retained only for channels which introduce virtually no distortion.

To completely remove ISI a cyclic prefix (CP) is inserted in front of every OFDM symbol. The CP is a copy of the OFDM symbol tail, see [Peled *et al.* (1980)] (87). For complete ISI removal the length of the cyclic prefix G must be longer than the essential support of the CIR L . The length of the OFDM symbol after insertion of the cyclic prefix is denoted by $P = N + G$.

Next we describe the process of transmission of information over an OFDM system. The transmission over the channel of an OFDM symbol vector containing N symbols (corresponding to N subcarriers) at time n , denoted by $[n] = [d_1[n], \dots, d_N[n]]^T \in \mathbb{C}^N$, where the subscript i is for the carrier index, according to

$$\mathbf{s}[n] = \mathbf{T}_{CP} \mathbf{W}_N^H [n], \quad (14.3.1)$$

where the CP insertion is described via matrix

$$\mathbf{T}_{CP} \triangleq \begin{bmatrix} \mathbf{I}_{CP} \\ \mathbf{I}_N \end{bmatrix} \in \mathbb{R}^{P \times N}, \quad (14.3.2)$$

where the matrix $\mathbf{I}_{CP} \in \mathbb{R}^{G \times N}$ denotes the last G rows of the identity matrix $\mathbf{I} \in \mathbb{R}^{N \times N}$. The

unitary DFT matrix $\mathbf{W}_N \in \mathbb{C}^{N \times N}$ has elements

$$[\mathbf{W}_N]_{i,k} \triangleq \frac{1}{\sqrt{N}} \exp \left\{ \frac{-j2\pi ik}{N} \right\}, \{i, k\} \in \{0, \dots, N-1\}. \quad (14.3.3)$$

After parallel to serial conversion $\mathbf{s}[n]$ is transmitted over the multipath channel. We express the CIR in vector notation as

$$\mathbf{h} = \begin{bmatrix} h_0 \\ \vdots \\ h_{L-1} \end{bmatrix}, \quad (14.3.4)$$

where we assume that $L < G$. Let

$$\mathbf{H}_{ISI} \triangleq \begin{bmatrix} h_0 & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & & & \vdots \\ h_{L-1} & & \ddots & \ddots & & \vdots \\ 0 & \ddots & & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & & \ddots & 0 \\ 0 & \cdots & 0 & h_{L-1} & \cdots & h_0 \end{bmatrix} \in \mathbb{C}^{P \times P}, \quad (14.3.5)$$

be the lower triangular Toeplitz channel matrix and let

$$\mathbf{H}_{IBI} \triangleq \begin{bmatrix} 0 & \cdots & 0 & h_{L-1} & \cdots & h_1 \\ \vdots & \ddots & & \ddots & \ddots & \vdots \\ \vdots & & \ddots & & \ddots & h_{L-1} \\ \vdots & & & \ddots & & 0 \\ \vdots & & & & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 \end{bmatrix} \in \mathbb{C}^{P \times P}, \quad (14.3.6)$$

be the upper triangular Toeplitz channel matrix. We can express the received signal as

$$\mathbf{r}[m] = \mathbf{H}_{ISI}\mathbf{s}[m] + \mathbf{H}_{IBI}\mathbf{s}[m-1] + \mathbf{w}[m], \quad (14.3.7)$$

where the first term represents the ISI between two consecutive OFDM symbols, the second term corresponds to inter block interference (IBI) between two consecutive OFDM block transmissions at time m and time $(m-1)$ and $\mathbf{w}[m] \in \mathbb{C}^P$ is the additive white Gaussian noise (AWGN), assumed to be i.i.d. complex Gaussian with zero-mean and variance σ_w^2 .

At the receiver the CP of length G is removed, and a DFT is performed on the remaining N samples. The CP removal can be represented by the matrix

$$\mathbf{R}_{CP} \triangleq [\mathbf{0}_{N \times G} \mathbf{I}_N] \in \mathbb{R}^{N \times P}, \quad (14.3.8)$$

which removes the first G entries from the vector $x[m] \in \mathbb{C}^P$ if the product $\mathbf{R}_{CP}\mathbf{x}[m]$ is formed. As long as $G \geq L$

$$\mathbf{R}_{CP}\mathbf{H}_{IBI} = \mathbf{0}_{N \times P}, \quad (14.3.9)$$

which indicates that the ISI between two consecutive OFDM symbols is completely eliminated. Finally, the received signal can be written as:

$$\begin{aligned} \mathbf{y}[m] &= \mathbf{W}_N \mathbf{R}_{CP} (\mathbf{r}[m] + \mathbf{w}[m]) = \mathbf{W}_N \mathbf{R}_{CP} \mathbf{H}_{ISI} \mathbf{s}[m] + \mathbf{W}_N \mathbf{R}_{CP} \mathbf{w}[m] \\ &= \mathbf{W}_N \mathbf{R}_{CP} \mathbf{H}_{ISI} \mathbf{T}_{CP} \mathbf{W}_N^H \mathbf{d}[m] + \mathbf{W}_N \mathbf{R}_{CP} \mathbf{w}[m] \\ &= \mathbf{W}_N \mathbf{H} \mathbf{W}_N^H \mathbf{d}[m] + \mathbf{W}_N \mathbf{R}_{CP} \mathbf{w}[m], \end{aligned} \quad (14.3.10)$$

where

$$\mathbf{H} \triangleq \mathbf{R}_{CP} \mathbf{H}_{ISI} \mathbf{T}_{CP} = \mathbf{W}_N^H \text{diag}(\mathbf{g}) \mathbf{W}_N, \quad (14.3.11)$$

where the CFR $\mathbf{g} \in \mathbb{C}^N$ is defined as the DFT of the CIR

$$\mathbf{g} \triangleq \mathbf{W}_{N \times L} \mathbf{h}, \quad (14.3.12)$$

where \mathbf{W}_L is the partial DFT matrix containing the first L columns of \mathbf{W}_N . Using (14.3.11) we rewrite (14.3.10) as

$$\mathbf{y}[m] = \text{diag}(\mathbf{g}) \mathbf{d}[m] + \mathbf{z}[m], \quad (14.3.13)$$

where the elements of $\mathbf{z}[m] \triangleq \mathbf{W}_N \mathbf{R}_{CP} \mathbf{w}[m]$, are white with variance σ_w^2 . Hence, the covariance matrix of $\mathbf{z}[m]$ has diagonal structure with identical elements

$$\begin{aligned} \mathbf{R}_z[m] &= \mathbb{E} \left\{ \mathbf{W}_N \mathbf{R}_{CP} \mathbf{w}[m] (\mathbf{W}_N \mathbf{R}_{CP} \mathbf{w}[m])^H \right\} \\ &= \sigma_w^2 \mathbf{W}_N \mathbf{R}_{CP} \mathbf{I}_P \mathbf{R}_{CP}^H \mathbf{W}_N^H \\ &= \sigma_w^2 \mathbf{I}_N. \end{aligned} \quad (14.3.14)$$

In an OFDM system, according to (14.3.13), every element of the symbol vector $[m]$ is transmitted over an individual frequency-flat subcarrier. Note that (14.3.13) can also be expressed as

$$\begin{aligned} \mathbf{y}[m] &= \text{diag}(\mathbf{d}[m]) \mathbf{g} + \mathbf{z}[m] \\ &= \text{diag}(\mathbf{d}[m]) \mathbf{W}_{N \times L} \mathbf{h} + \mathbf{z}[m]. \end{aligned} \quad (14.3.15)$$

For the detection of symbols, using (14.3.13) would be more convenient, while for the purpose of channel estimation, the usage of (14.3.15) would be more convenient.

Figure 14.3.1 provides a diagram of an MIMO-OFDM telecommunications system, typical of the ones studied in this thesis.

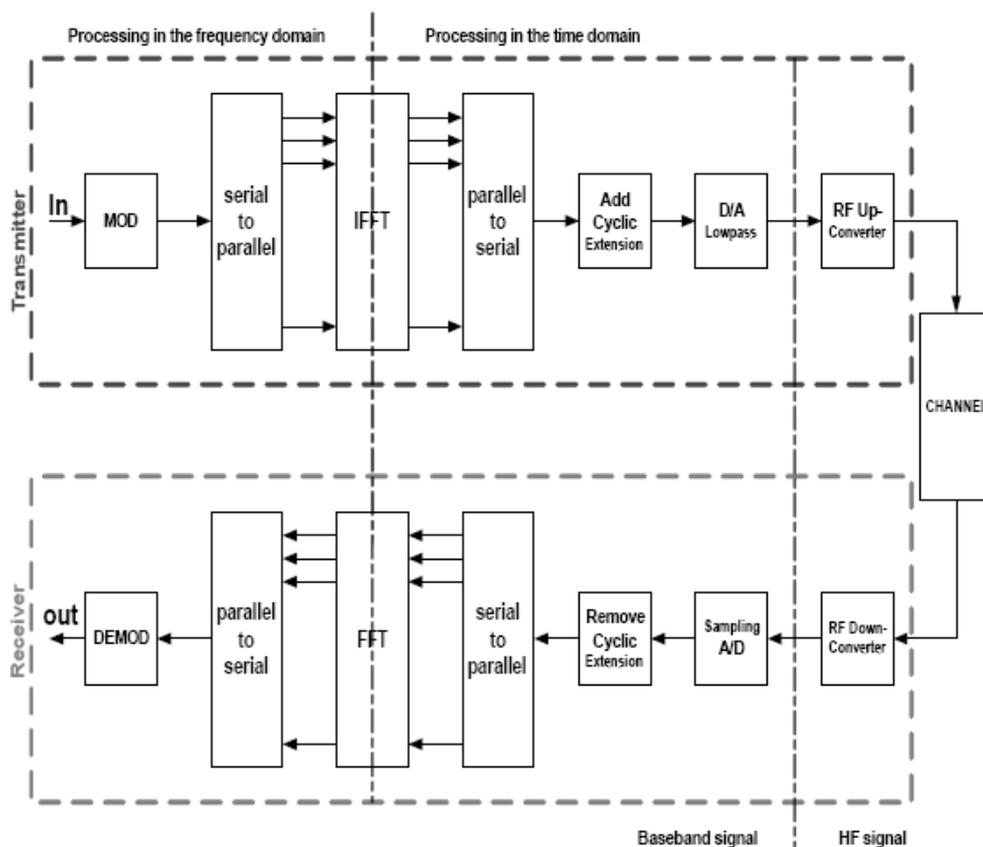


Fig. 14.3.1: MIMO-OFDM telecommunications system

There is an extensive literature on detection schemes at the receiver in MIMO, OFDM and MIMO-OFDM systems that are of relevance to Part III, however they are omitted here since the important aspects are discussed and can be found in suitable detail in the papers contained in Part III.

14.4 Contribution Part III

Journal Papers:

Paper 1: **Peters, G.W., Nevat I. and Yuan J. (2009) "Channel Estimation in OFDM Systems with Unknown Power Delay Profile using Trans-dimensional MCMC". *IEEE Transactions on Signal Processing*, *IEEE Trans. on Signal Processing*, 57(9), 3545-3561.**

This paper considers the problem of channel estimation for OFDM systems, where the number of channel taps and their power delay profile are unknown. Using a Bayesian approach, we construct a model in which we estimate jointly the coefficients of the channel taps, the channel order and decay rate of the power delay profile (PDP). In order to sample from the resulting posterior distribution we develop three novel Trans-dimensional Markov chain Monte Carlo (TDMCMC) algorithms and compare their performance. The

first is the basic *birth and death* TDMCMC algorithm. The second utilizes Stochastic Approximation to develop an adaptively learning algorithm to improve mixing rates of the Markov chain between model subspaces. The third approximates the optimal TDMCMC *proposal* distribution for *between-model moves* using Conditional Path Sampling *proposals*. We assess several aspects of the model in terms of sensitivities to different prior choices. Next we perform a detailed analysis of the performance of each of the TDMCMC algorithms. This allows us to contrast the resulting computational effort required under each approach versus the estimation performance. Finally, using the TDMCMC algorithm which produces the best performance in terms of exploration of the model sub-spaces, we assess its performance in terms of channel estimation mean squared error (MSE) and bit error rate (BER). It is shown that the proposed algorithm can achieve results very close to the case where both the channel length and the PDP decay rate are known.

Paper 2: **Nevat I., Peters, G.W. and Yuan J. (2008) "Detection of Gaussian Constellations in MIMO Systems Under Imperfect Channel State Information", to appear *Journal IEEE Transactions of Communications*.**

This paper considers the problem of symbols detection in MIMO systems at the presence of channel estimation errors. Under this framework we develop a computationally efficient approximation of the MAP detector for non-uniform constellations. First we review the channel estimation error in the setting of known orthogonal training sequences for channel estimation. We analyze the performance degradation due to noisy channel estimation. Next we propose a low complexity detector based on a relaxation of the discrete nature of the digital constellation and the channel estimation error statistics. This leads to a non-convex program that is solved efficiently via a hidden convexity minimization approach. Simulation results in a random MIMO system show that the proposed algorithm outperforms the linear MMSE receiver in terms of BER.

Paper 3: **Nevat I., Peters, G.W. and Yuan J. (2008) "A Low Complexity MAP Estimation in Linear Models with a Random Gaussian Mixing Matrix". *IEEE Transactions on Communications*, to appear**

We consider the formulation of a Bayesian inference approach for a model involving a random Gaussian vector in a linear model containing a random Gaussian matrix for which only the first and second moments are known. We propose an efficient method to finding the MAP estimator for this model and analyze its complexity. The performance in terms of estimation error is evaluated by simulation, which show it's superior to LMMSE.

Paper 4: **Peters, G.W., Nevat I., Yuan J. and Sisson S. (2009) "Bayesian Symbol Detection in Wireless Relay Networks.", in review *IEEE Transactions on Signal Processing*.**

This paper presents a general stochastic model which is developed for a class of cooperative wireless relay networks, in which imperfect knowledge of the channel state information at the destination node is assumed. This general framework incorporates multiple relay nodes operating under arbitrary processing functions. In general for such systems,

due to the intractability of the likelihood function, both the maximum likelihood and the maximum *a posteriori* decision rules do not admit closed form expressions.

We adopt a Bayesian approach, and present three novel computational techniques for maximum *a posteriori* sequence detection in these general networks. These include a Markov chain Monte Carlo approximate Bayesian computation (MCMC-ABC) approach; an auxiliary variable MCMC (MCMC-AV) approach; and a suboptimal exhaustive search zero forcing (SES-ZF) approach. Finally, numerical examples comparing the symbol error rate (SER) performance versus signal to noise ratio (SNR) of the three detection algorithm is studied in simulated examples.

Journal Paper 11

"Innovation is the ability to see change as an opportunity - not a threat"

Steve Jobs

Peters, G.W., Nevat I. and Yuan J. (2009) "Channel Estimation in OFDM Systems with Unknown Power Delay Profile using Trans-dimensional MCMC". *IEEE Transactions on Signal Processing*, IEEE Trans. on Signal Processing, 57(9), 3545-3561.

This work was instigated jointly by the first and second author. The first author can claim at least 50% of the credit for the contents. His work included developing the methodology contained and jointly developing with the second author the applications, the implementation and comparison to alternative approaches, writing the drafts of the paper. This work will be included in a PhD thesis of a co-author, though this thesis is not submitted by sequence of publications. Permission from all the co-authors is granted for inclusion in the PhD.

This paper appears in the thesis in a modified format to that which finally appeared in the journal *IEEE Transactions on Signal Processing*, where it was published.

Final print version available at:

<http://ieeexplore.ieee.org>

Channel Estimation in OFDM Systems with Unknown Power Delay Profile using Trans-dimensional MCMC

Gareth W. Peters (*corresponding author*)

CSIRO Mathematical and Information Sciences, Locked Bag 17, North Ryde, NSW, 1670, Australia;

UNSW Mathematics and Statistics Department, Sydney, 2052, Australia;

email: peterga@maths.unsw.edu.au

Ido Nevat

School of Electrical Engineering and Telecommunications, University of NSW, Sydney, Australia.

Jinhong Yuan

School of Electrical Engineering and Telecommunications, University of NSW, Sydney, Australia.

15.1 Abstract

This paper considers the problem of channel estimation for OFDM systems, where the number of channel taps and their power delay profile are unknown. Using a Bayesian approach, we construct a model in which we estimate jointly the coefficients of the channel taps, the channel order and decay rate of the power delay profile (PDP). In order to sample from the resulting posterior distribution we develop three novel Trans-dimensional Markov chain Monte Carlo (TDMCMC) algorithms and compare their performance. The first is the basic *birth and death* TDMCMC algorithm. The second utilises Stochastic Approximation to develop an adaptively learning algorithm to improve mixing rates of the Markov chain between model subspaces. The third approximates the optimal TDMCMC *proposal* distribution for *between-model moves* using Conditional Path Sampling *proposals*. We assess several aspects of the model in terms of sensitivities to different prior choices. Next we perform a detailed analysis of the performance of each of the TDMCMC algorithms. This allows us to contrast the resulting computational effort required under each approach versus the estimation performance. Finally, using the TDMCMC algorithm which produces the best performance in terms of exploration of the model sub-spaces, we assess its performance in terms of channel estimation mean squared error (MSE) and bit error rate (BER). It is shown that the proposed algorithm can achieve results very close to the case where both the channel length and the PDP decay rate are known.

Keywords: OFDM, channel estimation, Bayesian inference, Trans-dimensional Markov chain Monte Carlo, Stochastic Approximation, Conditional Path Sampling

15.2 Introduction

Orthogonal frequency division multiplexing (OFDM) systems often employ coherent detection that requires accurate information about the *channel impulse response* (CIR) (23). This can be achieved in slow fading channels by using pilot symbols at the beginning of each frame so that the channel can be estimated (12). Based on this channel estimation, the data symbols can be detected during the rest of the frame. When no *a priori* information about the statistics of the channel taps is known, a maximum likelihood (ML) approach is optimal amongst all unbiased estimators (9). A wireless channel is typically modeled as a finite impulse response (FIR) filter with every tap distributed as a complex Gaussian (8). When the channel taps' second order statistics are known, the minimum mean squared error (MMSE) estimate can achieve significant gains compared to the ML estimate (7). In defining these channels, the number of taps and power delay profile (PDP) decay rate need to be *a-priori* known in order for the channel to be estimated using Bayesian inference (14). These parameters, however, are usually unknown *a-priori* and vary as the mobile station changes its location. Existing approaches for estimating the number of channel taps include the most significant taps (MST) idea developed in (24) to estimate the channel order. In (27), an iterative algorithm that re-estimates the channel order using the generalized Akaike information criterion (1) and cancels its contribution is presented. In (26), after the initial channel estimation phase, an auxiliary function is used to distinguish between real taps and the noise contribution in the estimated channel. In (25), both Bayesian model averaging (BMA) and order selection (BMOS) algorithms were proposed for channel estimation based on a finite set of possible channel models. To the best of our knowledge, there has been no work discussing the scenario in which both the length of the channel and its PDP are unknown *a-priori* and need to be estimated jointly. In this paper we explore a Bayesian model which allows for joint estimation of the channel length, the decay rate of the power delay profile and the channel coefficients. To achieve this we develop novel TDMCMC algorithms to sample from the resulting posterior distribution developed in our Bayesian approach. TDMCMC algorithms have become popular in many areas including statistics, machine learning and signal processing since their introduction by (19), (10), (11). Since then they have been successfully developed further in the signal processing literature by (2), (14). We extend aspects of the methodology in these papers and apply our novel algorithms to find solutions to the estimation and sampling problems posed in this paper.

This paper is organized as follows. In Section 15.3 we describe the system model. Then in Section 15.4 we formulate the problem. In Section 15.5 we present a generic TDMCMC algorithm to obtain the quantities of interest. Next, in Section 15.6 we present three TDMCMC algorithms to perform the *between-model moves*. In Section 15.7 we discuss different aspects of the computational complexity of the proposed algorithms, then in Section 15.8 we present several performance bounds. Simulation results are presented in 15.9 and concluding remarks are given in Section 15.10.

The following notation is used throughout, boldface upper case letters denote matrices, boldface lower case letters denote column vectors, and standard lower case letters denote scalars.

The superscripts $(\cdot)^T$ and $(\cdot)^H$ denote Transpose and Hermitian, respectively. By \mathbf{I} and $\mathbb{I}(\cdot)$ we denote the identity matrix and the indicator function, respectively. The functions $Pr(\mathbf{x})$, $p(\mathbf{x})$, $p(\mathbf{x}|\mathbf{y})$ and $\mathbb{E}\{\cdot\}$ denote the probability mass function (PMF) of \mathbf{x} , the probability distribution function (PDF) of \mathbf{x} , the PDF of \mathbf{x} given \mathbf{y} , and the expectation, respectively. By $\mathcal{N}(\cdot; \mu, \sigma)$ we denote a Gaussian random variable with mean μ and standard deviation σ and by $\mathcal{CN}(\cdot; \mu, \sigma)$ a circular complex Gaussian random variable with mean μ and standard deviation σ . The superscripts \mathbb{C} and \mathbb{R}^+ denote the complex plain and positive real plain, respectively. $Tr\{\mathbf{A}\}$ denotes the trace of the matrix \mathbf{A} .

15.3 System Description

We define a sequence of information bits mapped to complex-valued symbols of an M-ary modulation alphabet set by $A = \{a_1, \dots, a_{|A|}\}$. The data symbols are multiplexed to K OFDM subcarriers. After modulation of the OFDM symbols via a K point inverse discrete Fourier transform (IDFT), the signal is transmitted over a frequency selective, time varying channel. We assume that the channel state is fixed (static) over the duration of one frame, but can change significantly between consecutive frames. The use of a proper cyclic prefix (CP) eliminates inter block interference (IBI) between consecutive OFDM symbols. It is assumed that the length of the CP is longer than the maximum path delay and also that inter carrier interference (ICI) caused by Doppler offset is fully compensated for. At the receiver, after CP removal and discrete Fourier transform (DFT), the signal at time index n is given by (23)

$$\mathbf{y}_n = \mathbf{D}_n \mathbf{W}_L \mathbf{h}_n^T + \mathbf{z}_n. \quad (15.3.1)$$

Here \mathbf{y}_n is a $K \times 1$ observation vector, \mathbf{D}_n is a $K \times K$ diagonal matrix and its diagonal entries contain the data symbols, and \mathbf{z}_n is a $K \times 1$ vector of independent identically distributed (i.i.d.) complex zero-mean Gaussian noise with covariance matrix $\mathbb{R} = \mathbf{I}\sigma_z^2$ and it is assumed to be uncorrelated with the channel. The vector $\mathbf{h}_n = [h_1[n], h_2[n], \dots, h_L[n]]$ is the discrete CIR at time instant n of length L , and its components are $h_l[n]$, $l = 1, \dots, L$. For notational simplicity, the block time index is omitted in the remainder of the paper. It is assumed that the channel length L is upper bounded by a predefined maximal possible channel length L_{\max} . The matrix \mathbf{W}_L is a $K \times L$ partial DFT matrix, defined as $\mathbf{W}_L \triangleq \frac{1}{\sqrt{K}} \{e^{-j2\pi kl/K}\}_{k=0, \dots, K-1; l=0, \dots, L-1}$. We consider a fading multipath channel model whose impulse response is represented as $\mathbf{h}_n = \sum_{l=1}^L h_l[n] \delta(n-l)$, where L is the number of paths and $\delta(\bullet)$ is the Kronecker delta function. It is assumed that $\{h_l[n]; l = 1, \dots, L\}$ are mutually independent, wide sense stationary (WSS) circular complex Gaussian random processes, having zero mean and covariance given by

$$\mathbb{E}\{h_k[n] h_j[n]^H\} = \begin{cases} \sigma_{h_k}^2 = \frac{\exp(-\beta k)}{\sum_{l=1}^L \exp(-\beta l)} & , k = j \\ 0 & , \text{otherwise} \end{cases} \quad (15.3.2)$$

The parameter β controls the decay rate of each tap's power, and depends on the physical environment in which the system operates. When $\beta = 0$ the channel taps are uniformly distributed. As β increases, the decay rate increases, and as a result high order channel taps become less significant as their power decreases. The parameters L and β are assumed to be statistically independent. We note that throughout the paper, we assume that \mathbf{D} is known during the channel estimation phase.

15.4 Problem Statement

In this work we concentrate on two Bayesian estimators, namely the MMSE and the maximum *a-posteriori* (MAP) estimators. In the case where L and β are *a-priori* known at the receiver the MAP, MMSE and Linear MMSE estimates of the CIR coincide and expressed as

$$\begin{aligned}\hat{\mathbf{h}} &= \mathbb{E} \{ \mathbf{h} | \mathbf{y}, L, \beta \} = \mathbb{E} \{ \mathbf{h} \mathbf{y}^H \} \mathbb{E}^{-1} \{ \mathbf{y} \mathbf{y}^H \} \mathbf{y} \\ &= \left((\mathbf{D} \mathbf{W}_L)^H (\mathbf{D} \mathbf{W}_L) + \sigma_z^2 \mathbf{C}_h^{-1} \right)^{-1} (\mathbf{D} \mathbf{W}_L)^H \mathbf{y},\end{aligned}\tag{15.4.1}$$

where $\mathbf{C}_h = \text{diag} \{ \sigma_{h_l}^2 \}$, $l = 1, \dots, L$.

Here we build on the model developed in (25). In this model we conditioned on assumed knowledge of the PDP decay rate β . We then modeled the channel length and channel coefficients as random variables in a Bayesian framework. The estimation of the channel length and channel coefficients was performed by considering BMOS estimators for the MMSE and MAP. The procedure in that paper allows an explicit solution to be found for the posterior model probabilities $p(L | \mathbf{y})$ (see (25) equation 12). It was achievable since the required marginalising integrations were analytic. This in turn simplified the model selection and consequent channel coefficient estimation. In this paper, we model the PDP decay rate, channel length and channel coefficients as random variables. In this setting, the addition of the estimation of the power delay profile decay rate β complicates the previous solutions. It is no longer possible to take the approach presented in (25). Instead we must perform joint inference from the posterior of the channel length, PDP decay rate and channel coefficients. That is, in this paper we have the additional complexity that we model the following model parameters: channel model order L , channel coefficients $[h_1, \dots, h_L]$, and channel power decay rate β as random variables. For notational convenience we will explicitly label the elements of \mathbf{h} under model L as $h_{1:L} = [h_1, \dots, h_L]$.

In general, we can consider either performing BMOS or BMA as in (25). However, in this paper we will focus on BMOS to estimate desired properties of the channel. We are interested in a BMOS analysis to obtain the MMSE channel estimator conditional

on the MAP estimate for the channel length, L_{MAP} , which is given by

$$\mathbf{h}_{MMSE} = \mathbb{E} \{ \mathbf{h} | \mathbf{y}, L_{MAP} \} = \int \int \mathbf{h} p(\mathbf{h}, \beta | \mathbf{y}, L_{MAP}) d\beta d\mathbf{h}, \quad (15.4.2)$$

and the MAP channel estimator is given by

$$\mathbf{h}_{MAP} = \arg \max_{\mathbf{h}} \{ p(\mathbf{h} | \mathbf{y}, L_{MAP}) \} = \arg \max_{\mathbf{h}} \left\{ \int p(\mathbf{h}, \beta | \mathbf{y}, L_{MAP}) d\beta \right\}, \quad (15.4.3)$$

where

$$L_{MAP} = \arg \max_L p(L | \mathbf{y}). \quad (15.4.4)$$

In order to estimate (15.4.2) and (15.4.3) we need an analytic expression for $p(\mathbf{h} | \mathbf{y})$. Obtaining an expression for \mathbf{h} involves marginalizing the joint posterior,

$$p(\mathbf{h} | \mathbf{y}) = \sum_{l=1}^{L_{MAX}} \int p(\mathbf{h}, L, \beta | \mathbf{y}) d\beta. \quad (15.4.5)$$

In our system model (15.3.1), typically it will not be possible to obtain an analytical expression for (15.4.5). Hence, we shall sample from the joint posterior via TDMCMC, and estimate this integral numerically.

Given T samples $\{ \mathbf{h}^{(t)}, L^{(t)}, \beta^{(t)} \}_{t=1:T}$ from the distribution $p(\mathbf{h}, L, \beta | \mathbf{y})$ we can then estimate quantities \hat{L}_{MAP} , $\hat{\mathbf{h}}_{MMSE}$ and $\hat{\mathbf{h}}_{MAP}$ and as follows:

$$\hat{L}_{MAP} = \arg \max_L p^E(L | \mathbf{y}), \quad (15.4.6)$$

where we define $p^E(\cdot)$ is the empirical histogram estimate of the density.

$$\hat{\mathbf{h}}_{MMSE} \approx \mathbb{E} \{ \mathbf{h} | \mathbf{y}, \hat{L}_{MAP} \} = \frac{1}{R} \sum_{t=1}^T \mathbf{h}^{(t)} \mathbf{I} \left(L^{(t)} = \hat{L}_{MAP} \right), \quad (15.4.7)$$

where R is the total number of samples corresponding to model \hat{L}_{MAP} .

$$\hat{\mathbf{h}}_{MAP} = \arg \max_{\mathbf{h}} \left\{ p^E \left(\mathbf{h} | \mathbf{y}, \hat{L}_{MAP} \right) \right\}, \quad (15.4.8)$$

where $p^E(\cdot)$ in this case is formed from samples corresponding to model \hat{L}_{MAP} .

To proceed, we define the joint posterior as follows

$$p(\mathbf{h}, L, \beta | \mathbf{y}) \propto p(\mathbf{y} | \mathbf{h}, L, \beta) p(\mathbf{h} | L, \beta) p(L) p(\beta), \quad (15.4.9)$$

where the priors for our model are specified as follows

- $p(L) = Poi(L; \lambda)$, with $L \in \{1, \dots, L_{\max}\}$,
- $p(\beta) = U[0, \beta_{\max}]$,
- $p(\mathbf{h}|L, \beta) = \mathcal{CN}(0, \mathbb{C}_{\mathbf{h}})$,

where $Poi(L; \lambda)$ represents Poisson distribution with mean λ , and $U[a, b]$ represents uniform distribution on the support $[a, b]$. Note that the approaches we develop are general and they allow for any desired prior structure. The choices of the priors made above are chosen to be uninformative.

In the following Sections we shall design novel TDMCMC algorithms to obtain samples from the posterior (15.4.9).

15.5 Generic TDMCMC Algorithm for Channel Estimation

To begin this section we define the following terminology which will be used throughout:

- *proposal*: shall be used to refer to any generic Markov chain transition kernel, which moves the chain from one state to another probabilistically.
- *within-model moves*: shall be used to define generic moves which propose to change the state of the Markov chain within a given model subspace, see (19). In this paper, we work with nested models and this corresponds to changes to parameters $h_{1:L}|L$ and $\beta|L$ with conditioning on the model L remaining fixed, ie. the model subspace is unchanged.
- *between-model moves*: shall be used to define generic moves which propose to change the state of the Markov chain to move it between different model subspaces, see (19). In this paper, this will correspond to increasing or decreasing the channel order, L , and proposing the corresponding model parameters in the new model subspace given by $h_{1:L}$ and β .
- *MH-within-Gibbs*: shall be used to refer to the Metropolis Hastings within Gibbs Markov chain Monte Carlo algorithm, see (17).
- *birth and death*: shall be used to refer to a particular set of *between-model moves*, see (19). In this paper under a *birth move* the model subspace Markov chain state index (L) shall be incremented by one, ie. $L + 1$. This shall involve the first L parameters $h_{1:L}$ remaining unchanged and the $L + 1$ -th parameter sampled from a *proposal*. The corresponding *death move* involves decrementing by one the model subspace Markov chain state index (L), ie. $L - 1$. This shall involve discarding the L -th parameter h_L .

- *mixing*: shall be used generically to refer to the rate at which the generated Markov chain reaches ergodicity, see (17).

The intention of this paper is to estimate $\hat{\mathbf{h}}_{MMSE} = [\hat{h}_{1,MMSE}, \dots, \hat{h}_{L_{MAP},MMSE}|L_{MAP}]$ and $\hat{\mathbf{h}}_{MAP} = [\hat{h}_{1,MAP}, \dots, \hat{h}_{L_{MAP},MAP}|L_{MAP}]$ where we sample realisations of the model parameters $(h_{1:L}, \beta, L)$ jointly from the posterior distribution. To estimate these quantities we must resort to numerical integration and the approach we consider here is based on MCMC algorithms (17). In particular we construct the problem such that it will require use of TDMCMC, see (19), (30) and (16).

We start by specifying the posterior support of each of the parameters of interest, L, β and \mathbf{h} . Here $L \in \{1, 2, \dots, L_{\max}\}$ can be considered as a model index specifying how many channel taps are present in the channel model vector $\mathbf{h} = h_{1:L} = [h_1, \dots, h_L] \in \mathbb{C}^L$. The parameter $\beta \in \mathbb{R}^+$ models the decay rate of the PDP. The joint posterior forms a nested model defined on a disjoint union of subspaces, $\Theta = \uplus \{L\} \times \mathbb{C}^L \times \mathbb{R}^+$.

Having specified the model we develop novel TDMCMC simulation algorithms to sample from the target posterior distribution (15.4.9). The first will be the basic *birth and death* (BD) *MH-within-Gibbs* sampler, called BD-TDMCMC. The second approach develops the ideas of Contour Monte Carlo or Stochastic Approximation (SA) adjusted TDMCMC for model selection and estimation as presented in (22), denoted as SA-TDMCMC. The third is a novel TDMCMC algorithm which automates the construction and sampling of an approximation to the optimal *between-model move* transition distributions, and this will be based on Conditional Path Sampling (CPS *MH-within-Gibbs* approach of (16)) termed CPS-TDMCMC. The BD-TDMCMC algorithm forms a comparison benchmark for the other two approaches. These more sophisticated approaches aim to improve mixing properties of the Markov chain between model subspaces by increasing the probability of acceptance for *between-model move* transitions in our trans-dimensional Markov chain, when compared to the basic BD-TDMCMC sampler.

The SA-TDMCMC utilises a simple *between-model move proposal* based on the BD-TDMCMC *proposal* that we develop. The simplicity of the *birth and death proposal* is not optimised for the target posterior distribution from which we are aiming to obtain samples. As such it will produce proposed transitions to new models which in the majority of cases may be unlikely to be accepted under a standard TDMCMC acceptance probability. This will result in slow between model mixing of the Markov chain.

The SA-TDMCMC algorithm attempts to offset the simplistic choice of *between-model move proposal* distribution. It achieves this by increasing the chance of acceptance of a *between-model move* with adaptively adjusting the acceptance probability of such a move through a stochastic approximation adjustment. This SA adjustment is based on on-line estimation of the Bayes Factors which weight the acceptance probability, with

the intention of improving the between model mixing rate.

The second approach for improving the acceptance probability for *between-model moves* is to approximate the optimal *between-model move proposal* distribution. This is an optimal *proposal* in the sense that it maximises the TDMCMC acceptance probability for a move from one model to another. In this paper we achieve this through use of the approach proposed in (16) which develops a Conditional Path Sampling approximation of the optimal *proposal* distribution.

We note here that we could combine both these approaches, the SA stage and the approximate optimal *between-model move proposal* under a CPS approximation. This would combine the best of both approaches and should be the preferred choice in terms of between model mixing of the Markov chain. However, in this paper our intention is to compare which of these aspects of TDMCMC algorithms will lead to the largest improvement in mixing of the Markov chain between different models, relative to our basic BD-TDMCMC algorithm.

The rest of the paper is organised as follows: we first outline the three different sampling methodologies. Next using the basic BD-TDMCMC sampler we perform sensitivity analysis for the model. We assess sensitivity to quantities such as prior specification (choice of λ) for the mean of the prior on the model order L ; the impact of different values of β on model order estimation; and the impact of SNR on estimation. Next we assess the performance of each of the algorithms in order to decide which provides the best trade-off between computational efficiency and performance. Using the best performing algorithm we then address the question of detection of symbols in the presence of uncertainty about the model L , the channel tap coefficients $h_{1:L}$ and the decay rate for the PDP β .

We introduce the generic notation for the within model (conditional on L) *proposal* to move from state $(h_{1:L}^{(t-1)}, \beta^{(t-1)})$ at iteration $t - 1$ of the Markov chain to state $(h_{1:L}^*, \beta^*)$ at iteration t , denoted by $T \left((h_{1:L}^{(t-1)}, \beta^{(t-1)}) \rightarrow (h_{1:L}^*, \beta^*) \right)$. Such a move will be accepted according to a standard MH acceptance probability.

The generic notation for *between-model moves*, going from state $(h_{1:L}^{(t-1)}, \beta^{(t-1)}, L^{(t-1)})$ at iteration $t - 1$ of the Markov chain to state $(h_{1:L^*}^*, \beta^*, L^*)$ at iteration t will be denoted by $Q \left((h_{1:L}^{(t-1)}, \beta^{(t-1)}, L^{(t-1)}) \rightarrow (h_{1:L^*}^*, \beta^*, L^*) \right)$.

In formulating our TDMCMC algorithms we shall utilise a common *within-model move* for all three algorithms. That is, they only differ in how they propose to move between models, i.e. in the specification of $Q \left((h_{1:L}^{(t-1)}, \beta^{(t-1)}, L^{(t-1)}) \rightarrow (h_{1:L^*}^*, \beta^*, L^*) \right)$. Before providing details of how we design these Markov chain *proposals*, we present the general algorithm we will use in all three cases.

Algorithm: Generic TDMCMC to obtain samples from (15.4.9)

1. Initialise parameters randomly or deterministically:

e.g. $L^{(0)} = 3, h_{1:3}^{(0)} = [0.1, 0.1, 0.1], \beta^{(0)} = 1.$

2. Repeat for $t = 1$ to T

Within-Model Moves (conditional on model $L^{(t-1)}$):

(a) Sample new proposed states of the Markov chain for

$$(h_{1:L^{(t-1)}}^*, \beta^*) \sim T\left((h_{1:L^{(t-1)}}^{(t-1)}, \beta^{(t-1)}) \rightarrow (h_{1:L^{(t-1)}}^*, \beta^*)\right),$$

(b) Calculate acceptance probability of new proposed state represented as

$$\begin{aligned} & \alpha\left((h_{1:L^{(t-1)}}^{(t-1)}, \beta^{(t-1)}), (h_{1:L^{(t-1)}}^*, \beta^*)\right) \\ &= \min\left(1, \frac{p(h_{1:L^{(t-1)}}^*, \beta^*, L^{(t-1)}|\mathbf{y})}{p(h_{1:L^{(t-1)}}^{(t-1)}, \beta^{(t-1)}, L^{(t-1)}|\mathbf{y})} \frac{T((h_{1:L^*}^*, \beta^*) \rightarrow (h_{1:L^{(t-1)}}^{(t-1)}, \beta^{(t-1)}))}{T((h_{1:L^{(t-1)}}^{(t-1)}, \beta^{(t-1)}) \rightarrow (h_{1:L^{(t-1)}}^*, \beta^*))}\right). \end{aligned} \quad (15.5.1)$$

(c) Sample $u \sim U[0, 1]$.

(d) If $u < \alpha\left((h_{1:L^{(t-1)}}^{(t-1)}, \beta^{(t-1)}), (h_{1:L^{(t-1)}}^*, \beta^*)\right)$

$$(h_{1:L^{(t)}}^{(t)}, \beta^{(t)}, L^{(t)}) = (h_{1:L^{(t-1)}}^*, \beta^*, L^{(t-1)}),$$

else

$$(h_{1:L^{(t)}}^{(t)}, \beta^{(t)}, L^{(t)}) = (h_{1:L^{(t-1)}}^{(t-1)}, \beta^{(t-1)}, L^{(t-1)}).$$

Between-Model Moves:

(e) Sample $(h_{1:L^*}^*, \beta^*, L^*) \sim Q\left((h_{1:L^{(t-1)}}^{(t-1)}, \beta^{(t-1)}, L^{(t-1)}) \rightarrow (h_{1:L^{(t-1)}}^*, \beta^*, L^*)\right).$

(f) Calculate acceptance probability of new proposed state according to

$$\begin{aligned} & \alpha\left((h_{1:L^{(t-1)}}^{(t-1)}, \beta^{(t-1)}, L^{(t-1)}), (h_{1:L^*}^*, \beta^*, L^*)\right) \\ &= \min\left(1, \frac{p(h_{1:L^*}^*, \beta^*, L^*|\mathbf{y})}{p(h_{1:L^{(t-1)}}^{(t-1)}, \beta^{(t-1)}, L^{(t-1)}|\mathbf{y})} \frac{Q((h_{1:L^*}^*, \beta^*, L^*) \rightarrow (h_{1:L^{(t-1)}}^{(t-1)}, \beta^{(t-1)}, L^{(t-1)}))}{Q((h_{1:L^{(t-1)}}^{(t-1)}, \beta^{(t-1)}, L^{(t-1)}) \rightarrow (h_{1:L^*}^*, \beta^*, L^*))}\right). \end{aligned} \quad (15.5.2)$$

(g) Sample $u \sim U[0, 1]$.

(h) If $u < \alpha\left((h_{1:L^{(t-1)}}^{(t-1)}, \beta^{(t-1)}, L^{(t-1)}), (h_{1:L^*}^*, \beta^*, L^*)\right)$

$$(h_{1:L^{(t)}}^{(t)}, \beta^{(t)}, L^{(t)}) = (h_{1:L^*}^*, \beta^*, L^*),$$

else

$$(h_{1:L^{(t)}}^{(t)}, \beta^{(t)}, L^{(t)}) = (h_{1:L^{(t-1)}}^{(t-1)}, \beta^{(t-1)}, L^{(t-1)}).$$

Note that the acceptance probability in (15.5.1) and (15.5.2) is obtained by the fact that we wish to construct a reversible Markov chain satisfying detailed balance. The details of this can be found in (19), (18).

We now present our choice of $T\left((h_{1:L}^{(t-1)}, \beta^{(t-1)}) \rightarrow (h_{1:L}^*, \beta^*)\right)$ for the *within-model moves*, given by a *MH-within-Gibbs* sampler. This is just one of many different approaches which could be utilised, see (17). We choose a *MH-within-Gibbs* sampling framework since we are able to obtain expressions for the full conditional posterior distributions of each of our parameters. We then sample from each full conditional posterior distribution of each parameter via a MH procedure. For further details of the convergence properties of this algorithm in terms of mixing rate and optimal acceptance probabilities, see for example (6) and (29).

15.5.1 Within-Model moves $T\left((h_{1:L}^{(t-1)}, \beta^{(t-1)}) \rightarrow (h_{1:L}^*, \beta^*)\right)$: *MH-within-Gibbs*

In this section we shall utilise a Gibbs sampling framework, and sample from each of the full conditionals via a MH stage. This corresponds to decomposing $T\left((h_{1:L}^{(t-1)}, \beta^{(t-1)}) \rightarrow (h_{1:L}^*, \beta^*)\right)$ as follows

$$T\left((h_{1:L}^{(t-1)}, \beta^{(t-1)}) \rightarrow (h_{1:L}^*, \beta^*)\right) = \prod_{i=1}^{L^{(t-1)}} T\left((h_{1:L}^{(t-1)}) \rightarrow (h_i^*)\right) T\left((\beta^{(t-1)}) \rightarrow (\beta^*)\right). \quad (15.5.3)$$

Specification of $T\left((h_{1:L}^{(t-1)}) \rightarrow (h_i^*)\right)$ is as follows

1. Sample $h_i^* \sim \mathcal{CN}\left(h_i^{(t-1)}, \sigma_1(i)\right)$, where $\sigma_1(i)$ is set deterministically, though it could be obtained via a process of pre-tuning, such as in an Adaptive MCMC algorithm, see (5).
2. The proposed new state is then accepted according to an MH rejection stage. Accept h_i^* with probability $a\left(h_i^{(t-1)}, h_i^*\right)$ given by,

$$a\left(h_i^{(t-1)}, h_i^*\right) = \min\left(1, \frac{p\left(h_i^* | \mathbf{h}_{\neq i}^{(t-1)}, L^{(t-1)}, \beta^{(t-1)}, \mathbf{y}\right) \mathcal{CN}\left(h_i^{(t-1)}; h_i^*, \sigma_1(i)\right)}{p\left(h_i^{(t-1)} | \mathbf{h}_{\neq i}^{(t-1)}, L^{(t-1)}, \beta^{(t-1)}, \mathbf{y}\right) \mathcal{CN}\left(h_i^*; h_i^{(t-1)}, \sigma_1(i)\right)}\right) \quad (15.5.4)$$

with $p\left(h_i | \mathbf{h}_{\neq i}, L, \beta, \mathbf{y}\right) \propto p\left(\mathbf{y} | h_i, \mathbf{h}_{\neq i}, L, \beta\right) p\left(h_i | \beta, L\right)$.

Specification of $T\left((\beta^{(t-1)}) \rightarrow (\beta^*)\right)$ is as follows

1. Sample $\beta^* \sim \mathcal{N}\left(\beta^*; \beta^{(t-1)}, \sigma_\beta\right) \mathbb{I}(\beta^* > 0)$, where $\sigma_\beta(i)$ is either set deterministically or obtained via a process of pre-tuning, such as in Adaptive MCMC algorithms.
2. Accept β^* with probability $a\left(\beta^{(t-1)}, \beta^*\right)$ given by

$$a\left(\beta^{(t-1)}, \beta^*\right) = \min\left(1, \frac{p\left(\beta^* | h_{1:L}^{(t-1)}, L^{(t-1)}, \mathbf{y}\right) \mathcal{N}\left(\beta^{(t-1)}; \beta^*, \sigma_\beta\right) \mathbb{I}(\beta^{(t-1)} > 0)}{p\left(\beta^{(t-1)} | h_{1:L}^{(t-1)}, L^{(t-1)}, \mathbf{y}\right) \mathcal{N}\left(\beta^*; \beta^{(t-1)}, \sigma_\beta\right) \mathbb{I}(\beta^* > 0)}\right), \quad (15.5.5)$$

with $p\left(\beta | \mathbf{h}, L, \mathbf{y}\right) \propto p\left(\mathbf{y} | \mathbf{h}, L, \beta\right) p\left(\mathbf{h} | L, \beta\right) p\left(\beta\right)$.

15.5.2 Between-model moves $Q \left((h_{1:L^{(t-1)}}^{(t-1)}, \beta^{(t-1)}, L^{(t-1)}) \rightarrow (h_{1:L^*}^*, \beta^*, L^*) \right)$

In this section we shall specify the choice of between model transition kernel

$Q \left((h_{1:L^{(t-1)}}^{(t-1)}, \beta^{(t-1)}, L^{(t-1)}) \rightarrow (h_{1:L^*}^*, \beta^*, L^*) \right)$ which distinguishes the three algorithms.

Typically, $Q \left((h_{1:L^{(t-1)}}^{(t-1)}, \beta^{(t-1)}, L^{(t-1)}) \rightarrow (h_{1:L^*}^*, \beta^*, L^*) \right)$ would be decomposed as the probability of choosing to perform a move from model $L^{(t-1)}$ to model L^* , denoted $q(L^{(t-1)} \rightarrow L^*)$, followed by a *proposal* distribution for sampling the new parameters in model L^* conditional on current parameters in model $L^{(t-1)}$, denoted $q \left((h_{1:L^{(t-1)}}^{(t-1)}, \beta^{(t-1)}) \rightarrow (h_{1:L^*}^*, \beta^*) \right)$.

We make an assumption that, for $L^* > L^{(t-1)}$, in our nested model structure, we can associate the first $h_{1:L^{(t-1)}}^{(t-1)}$ and $\beta^{(t-1)}$ parameters in model $L^{(t-1)}$ with those in model L^* . This gives the new parameters as, $(h_{1:L^*}^*, \beta^*, L^*) = \left(\mathbf{h}_{1:L^{(t-1)}}^{(t-1)}, \mathbf{h}_{L^{(t-1)}+1:L^*}^*, \beta^{(t-1)}, L^* \right)$, and we can now specify the optimal choice for this generic *between-model move* to sample the new parameters $h_{L^{(t-1)}+1:L^*}^*$ as

$$\begin{aligned} q \left(h_{1:L^{(t-1)}}^{(t-1)}, h_{1:L^*}^* \right) &= p(h_{L^*}^* | \mathbf{h}_{L^{(t-1)}+1:L^*-1}^*, \mathbf{h}_{1:L^{(t-1)}}^{(t-1)}, \beta^{(t-1)}, L^*, \mathbf{y}) \\ &\times p(h_{L^*-1}^* | \mathbf{h}_{L^{(t-1)}+1:L^*-2}^*, \mathbf{h}_{1:L^{(t-1)}}^{(t-1)}, \beta^{(t-1)}, L^*, \mathbf{y}) \dots \\ &\times p(h_{L^{(t-1)}+1}^* | \mathbf{h}_{1:L^{(t-1)}}^{(t-1)}, \beta^{(t-1)}, L^*, \mathbf{y}). \end{aligned} \quad (15.5.6)$$

However, in practice it is not possible to obtain analytic expressions or to sample from any of the marginalized distributions $p(h_{L-i}^* | h_{1:L-i-1}^*, \beta^*, L^*, \mathbf{y})$. In the first instance this typically involves solving marginalizing integrals which don't admit an analytic expression. In the second instance, even when marginalising integrals admit an analytic expression, these marginals will typically not be of a standard parametric form and hence are complicated to sample from.

In this paper we will restrict our proposed moves between models to increasing and decreasing $L^{(t)}$ by one, giving $q(L^{(t-1)} \rightarrow L^*)$ as

- $L^* \sim q(L^{(t-1)} \rightarrow L^*)$, with $q(L^{(t-1)} \rightarrow L^*)$ given by

$$q \left(L^{(t-1)} \rightarrow L^* \right) = \begin{cases} L^{(t-1)} + 1 \text{ w.p. } 0.5 & \text{if } 1 < L^{(t-1)} < L_{\max} \\ L^{(t-1)} - 1 \text{ w.p. } 0.5 & \text{if } 1 < L^{(t-1)} < L_{\max} \\ L^{(t-1)} - 1 \text{ w.p. } 1 & \text{if } L^{(t-1)} = L_{\max} \\ L^{(t-1)} + 1 \text{ w.p. } 1 & \text{if } L^{(t-1)} = 1. \end{cases} \quad (15.5.7)$$

Now we can proceed by specifying $q \left(h_{1:L^{(t-1)}}^{(t-1)} \rightarrow h_{1:L^*}^* \right)$ for the three different algorithms.

15.6 Design of Between Model Birth and Death Proposal

In this section we present three algorithms for the *between-model move proposals*.

15.6.1 Algorithm 1: Between Model Birth and Death Moves (BD-TDMCMC)

At time t , given that the state of the chain at time $t - 1$ is $(h_{1:L^{(t-1)}}^{(t-1)}, \beta^{(t-1)}, L^{(t-1)})$, proposing to move from model $L^{(t-1)}$ to a new model L^* proceeds by first determining if a *birth* or a *death* will be proposed according to (15.5.7). If a *birth move* is proposed, $L^{(t-1)}$ is incremented by one and $h_{1:L^*}^* = [h_{1:L^{(t-1)}}^{(t-1)}, h_{L^*}]$. We obtain h_{L^*} by sampling a new value for h_{L^*} according to a complex Gaussian *proposal* with zero mean and standard deviation σ_1 for real and imaginary components, $\mathcal{N}(0, \sigma_1)$, chosen based on the prior model. In a *death move* case, $L^{(t-1)}$ is decremented by one and $h_{1:L^*}^* = [h_1^{(t-1)}, \dots, h_{L-1}^{(t-1)}]$. These *birth and death* moves form a reversible pair so it is sufficient to specify the probability of acceptance for the *birth move*. The probability of acceptance for a *birth move* is the reciprocal of the death probability. The *birth move* acceptance probability is given in equation (15.5.2) after making the appropriate substitutions for the choice of Q given next

$$\begin{aligned} Q & \left((h_{1:L^{(t-1)}}^{(t-1)}, \beta^{(t-1)}, L^{(t-1)}) \rightarrow (h_{1:L^*}^*, \beta^*, L^*) \right) \\ & = q \left(L^{(t-1)} \rightarrow L^* \right) \mathcal{N}(\mathbb{R} \{h_{L^*}\}; 0, \sigma_2) \mathcal{N}(\mathbb{I} \{h_{L^*}\}; 0, \sigma_2), \end{aligned}$$

where σ_2 is set deterministically.

15.6.2 Algorithm 2: Stochastic Approximation TDMCMC (SA-TDMCMC)

Here we extend the *proposal* of the *birth and death* trans-dimensional sampling methodology detailed in Algorithm 1. We shall draw on the work of (22) and (21). The extension of the TDMCMC involves stochastic approximation and is based on adaptive Monte Carlo strategies and convergence results obtained in (3), (4) and (32). The SA algorithm, also known as Contour Monte Carlo, is itself a generalization of the Wang-Landau algorithm. This algorithm is typically used to calculate the spectral density of a physical system. In this paper we will use this methodology in a similar manner as discussed in (4), however we will be working in a TDMCMC setting. Here we reiterate that we are working with our target posterior distribution, $p(h_{1:L}, L, \beta | \mathbf{y})$, from which we wish to obtain samples. As in (21) we define a mapping $S(h_{1:l}, \beta, l)$ which maps a vector $(h_{1:l}, \beta, l) \in \Theta$ to a model indicator l , distinguishing each model subspace of Θ . The choice of this mapping was made specifically since we are aiming to develop a trans-dimensional sampler. The mapping S effectively partitions the support of the posterior distribution Θ into subspaces corresponding to each model subspace,

$$S(\theta) : E_1 = \{\theta = (h_1, \beta_1, 1) : S(\theta) = 1\}, \dots, E_{L_{\max}} = \{\theta = (h_{1:L_{\max}}, \beta_{L_{\max}}, L_{\max}) : S(\theta) = L_{\max}\}. \quad (15.6.1)$$

Next we define a weight which is attached to each model subspace E_i by

$$g_i = \frac{\int p(h_{1:i}, \beta | \mathbf{y}, L = i) dh_{1:i} d\beta}{\pi_i}, \quad i = 1, \dots, L_{\max}, \quad (15.6.2)$$

where $\pi_i > 0$ are pre-specified weights which satisfy the constraint $\sum_{i=1}^{L_{\max}} \pi_i = 1$. These quantities relate to Bayes Factors between model i and model j via the ratio of $\frac{g_i}{g_j}$. We now re-express the target posterior distribution as

$$p^*(h_{1:L}, L, \beta | \mathbf{y}) \propto \sum_{i=1}^{L_{\max}} \frac{p(\theta | \mathbf{y})}{g_i} \mathbb{I}(\theta \in \mathbf{E}_i), \quad (15.6.3)$$

where $\mathbb{I}(\cdot)$ is an indicator function. Hence, developing an algorithm which samples from $p^*(h_{1:L}, L, \beta | \mathbf{y})$ results in a random walk between the desired model spaces according to sampling frequencies, $\pi = (\pi_1, \dots, \pi_{L_{\max}})$. The SA algorithm effectively provides an automated way to learn the optimal weights $g_1, \dots, g_{L_{\max}}$ simultaneously for given initial $\pi = (\pi_1, \dots, \pi_{L_{\max}})$. In the SA algorithm we denote $\hat{g}_i^{(t)}$ as the estimate of g_i at iteration t . $\psi(\theta)$ is used as generic notation for the unnormalized target posterior. The estimate of the posterior distribution at iteration t is given by

$$\hat{p}^{(t)}(h_{1:L}, L, \beta | \mathbf{y}) \propto \sum_{i=1}^{L_{\max}} \frac{\psi(\theta)}{\hat{g}_i^{(t)}} \pi_i \mathbb{I}(\theta \in \mathbf{E}_i). \quad (15.6.4)$$

Denote by $\left\{ \theta_k^{(t)} \right\}_{k=1:M}$ the samples drawn from $\hat{p}^{(t)}(h_{1:L}, L, \beta | \mathbf{y})$ at iteration t and denote $v^{(t)} = (v_1^{(t)}, \dots, v_{L_{\max}}^{(t)})$ as the realized sampling frequency of each model with

$$v_i^{(t)} = \frac{1}{M} \sum_{k=1}^M \mathbb{I}(\theta_k^{(t)} \in \mathbf{E}_i). \quad (15.6.5)$$

The generic SA algorithm of (21) is presented in the following:

Stochastic Approximation Monte Carlo Algorithm (Generic)

1. **Sampling:** Draw samples $\theta_k^{(t)}, k = 1, \dots, M$, from working density $\hat{p}^{(t)}(h_{1:L}, L, \beta | \mathbf{y})$. This can be achieved via many methods, Importance Sampling, MCMC, Rejection Sampling, TDMCMC etc.
2. **Weight Update:** Update the working estimate of $\hat{g}_i^{(t)}, i = 1, \dots, L_{\max}$, recursively by setting

$$\log \hat{g}_i^{(t+1)} = \log \hat{g}_i^{(t)} + \gamma_t (v_i^{(t)} - \pi_i),$$

where γ_t is prespecified gain factor.

The SA-TDMCMC algorithm is an extension of the BD-TDMCMC algorithm. In this regard the SA-TDMCMC algorithm we develop samples from the target posterior over time by using the BD-TDMCMC algorithm with the addition of a weighting according to the stochastic approximation learning stage. Hence, the idea is to apply a generic TDMCMC algorithm such that the *within-model moves* are unchanged. The *between-model moves* are modified by the SA stage and

it is demonstrated in (22) that under BD-TDMCMC frameworks this corresponds to modifying the *between-model moves* to incorporate the following stages:

SA-TDMCMC: *Between-Model Moves*:

a. Sample $(h_{1:L^*}^*, \beta^*, L^*) \sim Q\left((h_{1:L^{(t-1)}}^{(t-1)}, \beta^{(t-1)}, L^{(t-1)}) \rightarrow (h_{1:L^*}^*, \beta^*, L^*)\right)$ given in equations (15.6.1).

b. Calculate the acceptance probability of proposed state, which is used in a rejection step

$$\alpha\left((h_{1:L^{(t-1)}}^{(t-1)}, \beta^{(t-1)}, L^{(t-1)}), (h_{1:L^*}^*, \beta^*, L^*)\right) = \min\left(1, \frac{\widehat{g}_{L^{(t-1)}}^{(t)}}{\widehat{g}_{L^*}^{(t)}} \frac{p(h_{1:L^*}^*, \beta^*, L^* | \mathbf{y})}{p(h_{1:L^{(t-1)}}^{(t-1)}, \beta^{(t-1)}, L^{(t-1)} | \mathbf{y})} \frac{Q((h_{1:L^*}^*, \beta^*, L^*) \rightarrow (h_{1:L^{(t-1)}}^{(t-1)}, \beta^{(t-1)}, L^{(t-1)}))}{Q((h_{1:L^{(t-1)}}^{(t-1)}, \beta^{(t-1)}, L^{(t-1)}) \rightarrow (h_{1:L^*}^*, \beta^*, L^*))}\right). \quad (15.6.6)$$

c. Weight Update: Update the working estimate of $\widehat{g}_i^{(t)}$, $i = 1, \dots, L_{\max}$, recursively by setting

$$\log \widehat{g}_i^{(t+1)} = \log \widehat{g}_i^{(t)} + \gamma_t \left(v_i^{(t)} - \pi_i\right). \quad (15.6.7)$$

Note, γ_t is the gain factor or learning rate, and it is critical for the convergence of this algorithm that it satisfies the following two conditions $\sum_{t=0}^{\infty} \gamma_t = \infty$ and $\sum_{t=0}^{\infty} \gamma_t^2 < \infty$, see (22). The first condition ensures convergence from any initial starting point and the second condition is an asymptotic damping of errors introduced from the use of $v_i^{(t)}$. In this paper we shall use the suggested choice of (22) $\gamma_t = \frac{\rho \kappa}{\max(\kappa, t)}$, $\kappa > 0$. The recommendation of (22) suggests that an appropriate choice for our problem would be to use $\rho = 1$ and $\kappa = 2 L_{\max}$.

In the model selection context, (22) produces a generalised TDMCMC algorithm. The resulting algorithm adaptively weights the probability of transitioning between different model spaces according to their posterior model probabilities which are learnt on-line. Additionally, the guaranteed convergence results of this algorithm ensure the validity of the approach in eventually generating correlated Markov chain samples from the correct stationary distribution given by our target posterior. Hence, the addition of the stochastic approximation adjustment should help to offset a poorly designed between model transition kernel, such as a simple *birth and death* kernel. That is our Markov chain will still be able to move between different model spaces due to the adjustment to the acceptance probability based on the on-line learning of adjustments based on Bayes Factor estimates.

15.6.3 Algorithm 3: Conditional Path Sampling TDCMC (CPS-TDMCMC)

Algorithm 3 involves an approximation to the optimal *birth proposal*. This requires sampling from the conditional distributions for the real and imaginary components of the new value of h_{L^*} . Henceforth we will work with the complex parameters $h_{1:L}$ on the real domain and by a slight abuse of notation we denote them by $h_{1:2L}$. There are twice as many components to reflect the fact we have L real components and L imaginary components. The *proposal* for a *birth move* is then given by

$$Q\left(\left(h_{1:2L}^{(t-1)}, \beta^{(t-1)}, L^{(t-1)}\right) \rightarrow \left(h_{1:2L^*}^*, \beta^*, L^*\right)\right) = q\left(L^{(t-1)} \rightarrow L^*\right) q\left(h_{1:2L}^{(t-1)} \rightarrow h_{1:2L^*}^*\right) \quad (15.6.8)$$

where

$$q\left(h_{1:2L}^{(t-1)} \rightarrow h_{1:2L^*}^*\right) = p\left(h_{2L}^{(t-1)+2} | L^{(t-1)} + 1, h_{2L}^{(t-1)+1}, h_{1:2L}^{(t-1)}, \beta^{(t-1)}, \mathbf{y}\right) \\ \times p\left(h_{2L}^{(t-1)+1} | L^{(t-1)} + 1, h_{1:2L}^{(t-1)}, \beta^{(t-1)}, \mathbf{y}\right). \quad (15.6.9)$$

We have used here the fact that we can decompose the optimal *proposal* into two distributions. Then by sampling from the first and conditioning on the sample we can sample the second component. Clearly, sampling h_{2L+2} and h_{2L+1} is not trivial since it involves sampling from both the conditional posterior and marginal conditional posterior of new parameters. Additionally sampling of h_{2L+1} is further complicated by the fact that its distribution is only known analytically as the integral $p(h_{2L+1} | h_{1:2L}, \mathbf{y}) = \int p(h_{2L+2}, h_{2L+1} | h_{1:2L}, \mathbf{y}) dh_{2L+2}$. In general this integration cannot be done analytically. Instead we approximate this optimal sampling distribution via a path sampling estimator as described in Section 2.2 of (16) and (15).

Next we develop the automated TDMCMC algorithm of (16) denoted as CPS MCMC. The intention of using this algorithm is twofold. Firstly, we would like to automate the *between-model move proposal* of the MCMC sampler. Secondly, we require that the resulting algorithm has suitably efficient mixing between models. The details of the construction of this *proposal* can be found in section 3.1 of (16).

Generic Construction of the CPS proposal: We utilise the density estimator of (15) as a simple marginal estimator based on path sampling and gradients of the log-posterior. It is straightforward to implement, and flexible for the purpose of estimating densities of the form $p(h_{2L+1} | h_{1:2L}, \mathbf{y})$.

Generically, suppose we have samples $\theta^1, \dots, \theta^t$ from a distribution $\pi(\theta)$ known up to some normalization constant. We are interested in estimating the k -th marginal distribution, $\pi_k(\theta_k)$. If $\varphi(\theta_k) = \log \pi_k(\theta_k)$ denotes the log of the (unnormalized) marginal posterior distribution, (15) shows that

$$\frac{\partial}{\partial \theta_k} \varphi(\theta_k) = \mathbb{E}_{\theta_{-k}}[U_k(\theta)] \quad \text{and} \quad U_k(\theta) = \frac{\partial}{\partial \theta_k} \log \pi(\theta). \quad (15.6.10)$$

An estimator for $\mathbb{E}_{\theta_{-k}}[U_k(\theta)]$ can be obtained by first ordering the samples according to the k -th marginal sample values $\theta_k^{(1)} < \theta_k^{(2)} < \dots < \theta_k^{(m)}$, $m \leq n$, ignoring any duplicates which may arise in the context of the Metropolis algorithm for example. If $n(i)$ is the number of replicates

for each $\theta_k^{(i)}$, then the estimator $\bar{U}_k(i)$ is obtained as

$$\bar{U}_k(i) = n(i)^{-1} \sum_{j=1}^{t(i)} U_k(\theta_j). \quad (15.6.11)$$

The $\bar{U}_k(i)$ approximates the gradient of $\log \pi_k(\theta_k)$ at each of the points in the sample, and may be utilised to derive a density estimate. Of the density estimation methods presented in (15) we implement the simplest approach, since our interest is in a computationally efficient approximate distribution rather than a perfect density estimate. A stepwise linear approximation to $\pi_k(\theta_k)$ is obtained by setting $\hat{\varphi}(\theta_k^{(1)}) = 0$ and defining

$$\hat{\varphi}(\theta_k^{(i)}) = \hat{\varphi}(\theta_k^{(i-1)}) + (\theta_k^{(i)} - \theta_k^{(i-1)}) \times (\bar{U}_k(i) + \bar{U}_k(i-1)) / 2 \quad \text{for } i = 2, \dots, m. \quad (15.6.12)$$

Hence for all points $\theta \in [\theta_k^{(1)}, \theta_k^{(m)}]$, the unnormalized density estimate is given by

$$\hat{\pi}_k(\theta_k) = \exp[\hat{\varphi}(\theta_k^{(i)})] \quad \text{for } \theta \in [\theta_k^{(i)}, \theta_k^{(i+1)}]. \quad (15.6.13)$$

The corresponding distribution function is obtained by integrating over $\hat{\pi}_k(\theta)$, with the normalizing constant given by the value of the distribution function at the point $\theta_k^{(m)}$. Simulation from $\hat{\pi}_k(\theta)$ may proceed via inversion methods.

Hence, our CPS *birth proposal* automates *between-model moves* by using the estimator of (15) to construct a marginal density estimate which is conditioned upon some subset of the remaining parameters θ_{-k} , e.g. $\pi_k(\theta_k \mid \theta_1, \dots, \theta_{k-1})$. This is achieved by fixing the sample margins at their conditioned values and estimating the density as before. Therefore it is feasible to approximate the optimal *between-model move proposal* decomposition in (15.6.9). This may proceed by first estimating and sampling from the density $p(h_{2L+1} \mid h_{1:2L}, \beta, \mathbf{y})$, then estimating and sampling from $p(h_{2L+2} \mid L+1, h_{2L+1}, h_{1:2L}, \beta, \mathbf{y})$ conditional upon the previously sampled point, and so on.

Details on CPS Proposal Construction:

There are several ways one can construct the CPS *proposal*. Choices include: 1) selecting an approximating spline; 2) estimated gradient versus exact gradient calculation; and 3) location of spline points. The choice of splines can include piecewise constant, linear piecewise, cubic piecewise. The spline can be constructed over either a fixed deterministic or a random grid, see (16). We note, the purpose of the approximation is not exact density estimation, but construction of a *proposal* with mass roughly in the region of the target density being approximated. We consider the most computationally efficient choice corresponding to a piecewise constant spline.

Additionally, one must consider the number of grid points in the construction. This will be dependent on: the model; the choice of spline; parameter space (bounded versus unbounded support); and grid placement (deterministic or random). With a bounded support, it is often sensible to utilise a fixed number of grid points placed randomly or deterministically over this

support, see examples in (16). In a continuous support a fixed interval can be used and then continuous tails added to each side of the interval, since the *proposal* must have positive support regions where the target distribution being approximated has positive support.

Specification of the optimal number of grid points is an open question, and we recommend choosing the grid points via a combination of acceptance probability criterion and analysis of *proposal* distributions. If during trial runs the average acceptance probability for *between-model move proposals* is low, and increasing the number of grid points results in a better approximation of the *proposal*, thus increasing the acceptance rate, then increase the number of grid points until the average acceptance rate is no longer improving.

In this paper the number of grid points was obtained by analysis of the constructed *proposal* for several states of the Markov chain under different number of grid points. A compromise between computational cost and accuracy of *proposal* approximation based on analysis of average acceptance rate was used. Details of the construction follow recommendations made in (16). For example estimating $p(h_{2L+1}|h_{1:2L}, \beta, \mathbf{y})$ involved: exact gradient calculations found in Appendix 1; a deterministic grid for the values of h_{2L+1} ; and a piecewise constant density estimator. The grid was centered on the approximate mode that we obtain by solving for the roots of the gradient expression in Appendix 1. Finally, we estimate the average gradient of the log posterior by sampling from the prior for parameters not of interest and not conditioned upon, in our case h_{2L+2} .

The CPS *proposal* clearly requires more computational effort to construct compared to the simple *birth and death proposal*. Generally, this is justified by obtaining much higher acceptance probabilities for *between-model moves* as a result of sampling from an approximation of the optimal *between-model move proposal*.

CPS-TDMCMC: *Between-Model Moves*:

If performing a move from model $L^{(t-1)}$ to $L^* = L^{(t-1)} + 1$:

1. Sample $(h_{1:L^*}^*, \beta^*, L^*) \sim Q\left((h_{1:L^{(t-1)}}^{(t-1)}, \beta^{(t-1)}, L^{(t-1)}) \rightarrow (h_{1:L^*}^*, \beta^*, L^*)\right)$, as follows:
 - (a) Construct approximate optimal *proposal* for parameter h_{2L+1} using CPS approach.
 - i. Sample m points from the prior for parameters h_{2L+2} , h_{2L+1} , denoted by $h_{2L+2}(i)$ and $h_{2L+1}(i)$, which are used to create an array

$$\begin{bmatrix} h_1^{(t-1)}, & \cdots, & h_{2L}^{(t-1)}, & h_{2L+1}(1), & h_{2L+2}(1) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ h_1^{(t-1)}, & \cdots, & h_{2L}^{(t-1)}, & h_{2L+1}(m), & h_{2L+2}(m) \end{bmatrix}. \quad (15.6.14)$$

- ii. Use each row of this array to evaluate the expression for the mode of the posterior (Appendix 1) for parameter h_{2L+1} and average these results to obtain the center point for the grid around which we construct the *proposal*, denoted by $\widehat{Mode}(h_{2L+1})$.

- iii. Construct a grid of n points with $n/2$ points on either side of $\widehat{Mode}(h_{2L+1})$ (linearly spaced or non-linearly spaced) and for each grid point sample m points from the prior for h_{2L+2} and construct the array in (15.6.14). In this paper we used a linear spaced grid with width between adjacent points of length w .

$$\begin{bmatrix} h_1^{(t-1)}, \dots, h_{2L}^{(t-1)}, \widehat{Mode}(h_{2L+1}) - \frac{n}{2}w, & h_{2L+2}(1) \\ \vdots & \vdots \\ h_1^{(t-1)}, \dots, h_{2L}^{(t-1)}, \widehat{Mode}(h_{2L+1}) + \frac{n}{2}w, & h_{2L+2}(n \times (m-1)) \\ h_1^{(t-1)}, \dots, h_{2L}^{(t-1)}, \widehat{Mode}(h_{2L+1}) + \frac{n}{2}w, & h_{2L+2}(n \times m) \end{bmatrix}. \quad (15.6.15)$$

- iv. Use the array in (15.6.15) of n grid points with m samples per grid point to evaluate the gradient (in Appendix 1). Next, for the i -th grid point average the m gradient evaluations to obtain an estimate of $\bar{U}_k(i)$. This is repeated for each of the n grid points.
- v. Construct stepwise constant approximation $\widehat{p}(h_{2L+1}|h_{1:2L}, \beta, \mathbf{y})$ using equations (15.6.12) and (15.6.13). Then add Gaussian tails to this approximate distribution on either side of the left and right end points of the grid, see (16) [p.5, equation 9] for details. Then it is trivial to normalize this stepwise constant approximation and construct an empirical cumulative distribution function (CDF) from which one can easily sample to obtain a new proposed state.

- (b) Sample *proposal* h_{2L+1}^* from normalized approximation $\widehat{p}(h_{2L+1}|h_{1:2L}, \beta, \mathbf{y})$.
- (c) Construct approximate optimal *proposal* for parameter h_{2L+2} using CPS approach, i.e. estimate $p(h_{2L+2}|L+1, h_{2L+1}^*, h_{1:2L}^{(t-1)}, \beta^{(t-1)}, \mathbf{y})$
- Using the array constructed in (15.6.14), replace the elements corresponding to h_{2L+1} with the sampled point h_{2L+1}^* .
 - As above, use each row of this array to evaluate the expression for the mode of the posterior (Appendix 1) for parameter h_{2L+2} . Average these results to obtain the center point for the grid around which we construct the *proposal*, denoted by $\widehat{Mode}(h_{2L+2})$. Then repeat steps iii. and v. above using the newly constructed grid.
- (d) Sample *proposal* h_{2L+2}^* from normalized approximation $\widehat{p}(h_{2L+2}|h_{2L+1}^*, h_{1:2L}^{(t-1)}, \beta^{(t-1)}, \mathbf{y})$.

2. Calculate the acceptance probability of proposed state, which is used in a rejection step

$$\alpha \left((h_{1:L}^{(t-1)}, \beta^{(t-1)}, L^{(t-1)}), (h_{1:L}^*, \beta^*, L^*) \right) = \min \left(1, \frac{p(h_{1:L}^*, \beta^*, L^* | \mathbf{y})}{p(h_{1:L}^{(t-1)}, \beta^{(t-1)}, L^{(t-1)} | \mathbf{y})} \frac{Q((h_{1:L}^*, \beta^*, L^*) \rightarrow (h_{1:L}^{(t-1)}, \beta^{(t-1)}, L^{(t-1)}))}{Q((h_{1:L}^{(t-1)}, \beta^{(t-1)}, L^{(t-1)}) \rightarrow (h_{1:L}^*, \beta^*, L^*))} \right).$$

15.7 Complexity Analysis

The complexity analysis of the class of algorithms presented in the previous Sections can be studied from two perspectives; the most technical of these involves theoretical study of the mixing rate of the TDMCMC algorithm under consideration, the other focus would be on the computational complexity. A theoretical study of the rate at which the generated Markov chain forgets its initial conditions is well beyond the scope of this paper. Instead we focus on a computational complexity comparison between each of the algorithms. The computational cost of each of these algorithms can be split into three parts: the first cost involves constructing and sampling from the *proposal*; the second significant computational cost comes from the evaluation of the acceptance probability for the proposed new Markov chain state; and the third is related to the mixing rate of the overall MCMC algorithm as affected by the length of the Markov chain required to obtain estimators of a desired accuracy.

We define the following building blocks and associated complexity:

1. Sampling a random variable using exact sampling via inversion of CDF has a complexity of $\mathbb{O}(1)$,
2. Evaluation of likelihood density $p(\mathbf{y}|\mathbf{h}, L, \beta)$ has a complexity of $KL(C_m + C_a) + \mathbb{O}(1)$,
3. Evaluation of prior density $p(h_i|\beta, L)$ has a complexity of $\mathbb{O}(1)$,

where C_m and C_a represent the operations of complex multiplication and addition, respectively. The complexity of the TD-MCMC BD (Algorithm 1) is presented in Table 15.1. The additional complexity of the SA-TDMCMC algorithm results from modifying the acceptance probability as shown in (15.6.6) and updating the weights according to (15.6.7). The modification of the acceptance probability is an operation of $\mathbb{O}(1)$. The updating of the weights adds an extra stage after acceptance probability calculation, which has complexity of $L_{\max}\mathbb{O}(1)$.

The complexity of the CPS-TDMCMC algorithm is depicted in Table 15.2. We have now explicitly presented the computational complexity of one iteration of the Markov chain for each of the algorithms. It is clear that although the CPS algorithm is computationally complex than the other algorithms, it's still linear in the dimension of the problem.

15.8 Estimator Efficiency via Bayesian Cramér Rao Type Bounds

In the Section 15.4 we have formulated a Bayesian model which resulted in a posterior distribution (15.4.9). We discussed that we are interested in approximating point estimates (15.4.7) and (15.4.8) from the posterior model. We then established that in order to obtain estimates of these quantities we would need to be able to sample from the posterior distribution over all models, requiring the TDMCMC methodology. In this section we provide some definitions and comparative performance bounds based on MSE of the point estimator of $\hat{\mathbf{h}}_{MMSE}$ obtained from

the posterior $\mathbf{h}|\mathbf{y}$. These bounds will be defined based on the concept of the Bayesian Cramér Rao Lower Bound (BCRLB) adapted for the model selection setting.

To clarify this further, we note that obtaining an analytical expression for (15.4.5) is not feasible, therefore the MMSE estimate for \mathbf{h} given \mathbf{y} and its MSE cannot be calculated analytically. To asses the performance of the estimator of \mathbf{h} we can calculate the MSE for \mathbf{h} via simulation. We can then compare the MSE to the BCRLB.

For deterministic parameters, a commonly used lower bound for the MSE of the parameters is the CRLB, given by the inverse of the Fisher information matrix (13). Since we are assuming the model parameters are random variables and therefore we are working with the posterior distribution under a Bayesian framework we can instead use the BCRLB or posterior CRLB (31).

The BCRLB provides a lower bound on the MSE matrix for random parameters. Let $\hat{\mathbf{h}}$ denote an estimate of \mathbf{h} which is a function of the observations \mathbf{y} . The estimation error is $\hat{\mathbf{h}} - \mathbf{h}$ and the MSE matrix is $\Sigma = \mathbb{E}_{\mathbf{y},\mathbf{h}} \left\{ (\hat{\mathbf{h}} - \mathbf{h}) (\hat{\mathbf{h}} - \mathbf{h})^H \right\}$, where $\mathbb{E}_{\mathbf{y},\mathbf{h}}$ denotes expectation with respect to $p(\mathbf{y}, \mathbf{h})$. The BCRLB \mathbb{C} provides a lower bound on the MSE matrix Σ . It is the inverse of the Bayesian information matrix (BIM) and therefore $\Sigma \geq \mathbb{C} \triangleq^{-1}$, where the matrix inequality indicates that $\Sigma - \mathbb{C}$ is a positive semi-definite matrix. Defining Δ_{β}^{α} to be the $L \times K$ matrix of second-order partial derivatives with respect to the $L \times 1$ parameter vector β and $K \times 1$ parameter vector α , the BIM for \mathbf{h} is defined as

$$\mathbf{J} = \mathbb{E}_{\mathbf{y},\mathbf{h}} \left\{ -\Delta_{\mathbf{h}}^{\mathbf{h}} \ln p(\mathbf{y}, \mathbf{h}) \right\} = \mathbb{E}_{\mathbf{y},\mathbf{h}} \left\{ -\Delta_{\mathbf{h}}^{\mathbf{h}} \left(\ln \sum_{l=1}^{L_{\max}} \int p(\mathbf{y}, \mathbf{h}, L, \beta) d\beta \right) \right\}. \quad (15.8.1)$$

It is clear that the BCRLB for \mathbf{h} can not be obtained analytically since it involves the summation over L and integration over β . This analysis is nonstandard since it involves an unknown model order, as such we will present several settings for which we can define performance bounds, which are based around the BCRLB.

Here we define the following model settings and associated bounds:

1. *BCRLB*: this bound is calculated conditional on knowledge of the true L and true β used to generate the data, as such this would represent the lowest bound that we consider for our comparison.

$$\begin{aligned} Tr \{ \mathbf{J} \} &= Tr \left\{ \mathbb{C}_{\mathbf{h}} - \mathbb{E} \{ \mathbf{h}\mathbf{y}^H \} \mathbb{E}^{-1} \{ \mathbf{y}\mathbf{y}^H \} \mathbb{E} \{ \mathbf{y}\mathbf{h}^H \} \right\} \\ &= Tr \left\{ \mathbb{C}_{\mathbf{h}} - \mathbb{C}_{\mathbf{h}} (\mathbf{D}\mathbf{W}_L)^H \left((\mathbf{D}\mathbf{W}_L) \mathbb{C}_{\mathbf{h}} (\mathbf{D}\mathbf{W}_L)^H + \mathbb{R} \right)^{-1} (\mathbf{D}\mathbf{W}_L) \mathbb{C}_{\mathbf{h}} \right\}. \end{aligned} \quad (15.8.2)$$

2. $B|_{L_{\max},\beta}$: to define this bound we condition on knowledge of β and on a misspecification of L . In particular we consider a saturated model where $L = L_{\max}$ and the true L is less than L_{\max} . In this case we consider based around the BCRLB for the saturated model,

given by

$$\begin{aligned} Tr \{ \mathbf{J}_{|L_{\max}, \beta} \} &= Tr \left\{ \mathbb{C}_{L_{\max}} - \mathbb{C}_{L_{\max}} (\mathbf{D}\mathbf{W}_{L_{\max}})^H \left((\mathbf{D}\mathbf{W}_{L_{\max}}) \mathbb{C}_{L_{\max}} (\mathbf{D}\mathbf{W}_{L_{\max}})^H + \mathbb{R} \right)^{-1} \right. \\ &\quad \left. \times (\mathbf{D}\mathbf{W}_{L_{\max}}) \mathbb{C}_{L_{\max}} \right\}, \end{aligned} \quad (15.8.3)$$

where $[\mathbb{C}_{L_{\max}}]_{l,l} = \frac{\exp(-\beta l)}{\sum_{k=1}^{L_{\max}} \exp(-\beta k)}$, $l = 1, \dots, L_{\max}$.

We note that, under the assumption that we have a nested model structure, the following interesting identity holds

$$Tr \{ \mathbf{J}_{|L_{\max}, \beta} \} > Tr \{ \mathbf{J} \}. \quad (15.8.4)$$

3. $B_{|\beta}$: in this bound we condition on knowledge of β and we consider two settings, namely BMOS and BMA. We define the lower bound for the BMA case as

$$Tr \{ J_{|\beta}^{BMA} \} = \sum_{l=1}^{L_{\max}} Pr(l|\mathbf{y}, \beta) Tr \{ \Sigma_l \}, \quad (15.8.5)$$

where

$$Pr(l|\mathbf{y}, \beta) = \frac{p(\mathbf{y}|l, \beta) Pr(l)}{\sum_{l=1}^{L_{\max}} p(\mathbf{y}|l, \beta) Pr(l)}, \quad (15.8.6)$$

where the marginal likelihood $p(\mathbf{y}|l, \beta)$ can be written as

$$p(\mathbf{y}|l, \beta) = \int p(\mathbf{y}|\mathbf{h}_l, l, \beta) p(\mathbf{h}_l|l, \beta) d\mathbf{h}_l = \mathcal{CN} \left(\mathbf{0}, (\mathbf{D}\mathbf{W}_l) \mathbb{C}_l (\mathbf{D}\mathbf{W}_l)^H + \sigma_w^2 \mathbf{I} \right). \quad (15.8.7)$$

In the case of BMOS we would use the expression given in (15.8.2) after replacing L_{\max} with L_{MAP} .

4. $B_{|L}$: this bound corresponds to the case where we condition on knowledge of the model order. However, we do not condition on knowledge of β . Since we base this bound on the BCRLB, we are interested in

$$\mathbf{J} = \mathbb{E}_{\mathbf{y}, \mathbf{h}} \left\{ -\Delta_{\mathbf{h}}^{\mathbf{h}} \ln p(\mathbf{y}, \mathbf{h}|L) \right\} = \mathbb{E}_{\mathbf{y}, \mathbf{h}} \left\{ -\Delta_{\mathbf{h}}^{\mathbf{h}} \left(\ln \int p(\mathbf{y}, \mathbf{h}, \beta|L) d\beta \right) \right\}. \quad (15.8.8)$$

This can not be written down analytically, hence, we will approximate it as follows

$$\hat{p}(\mathbf{h}|\mathbf{y}, L) = \frac{1}{T} \sum_{t=1}^T p(\mathbf{h}|\mathbf{y}, L, \beta^{(t)}). \quad (15.8.9)$$

We then substitute this Monte Carlo estimate in (15.8.8) into (15.8.8) to obtain an estimate

of the bound

$$\text{Tr} \left\{ \widehat{\mathbf{J}}_{|L} \right\} = \mathbb{E}_{\mathbf{y}, \mathbf{h}} \left\{ -\Delta_{\mathbf{h}}^{\mathbf{h}} \left(\ln \int p(\mathbf{y}, \mathbf{h}, \beta | L) d\beta \right) \right\} \approx \frac{1}{ST} \sum_{s=1}^S \sum_{t=1}^T \text{Tr} \left\{ \Sigma_{\min} \left(\beta^{(t)}, \mathbf{y}^{(s)} \right) \right\} \quad (15.8.10)$$

where S is the number of realisations used and we $\text{Tr} \left\{ \Sigma_{\min} \left(\beta^{(t)}, \mathbf{y}^{(s)} \right) \right\}$ is defined as in (15.8.2) with

$$[\mathbf{C}_{\mathbf{h}}]_{l,l} = \frac{\exp(-\beta^{(t)}l)}{\sum_{k=1}^{L_{\max}} \exp(-\beta^{(t)}k)}, \quad l = 1, \dots, L. \quad (15.8.11)$$

In the next section we shall compare the MSE of the proposed algorithms with the aforementioned bounds.

15.9 Simulation Results

In this section we present the simulation results for the proposed model and sampling algorithms. First we describe the simulated OFDM system.

15.9.1 System Configuration and Algorithms Initialization

Unless stated otherwise, the following specifications were used in the simulations of this paper. The OFDM system setup is $K = 64$ subcarriers employing QPSK symbols and $CP = L_{\max} = K/4$. The channel is modeled as block Rayleigh fading and the *a-priori* channel length distribution follows a truncated Poisson distribution $Poi(l; \lambda, L_{\max})$, with rate $\lambda = 8$.

The prior distribution for the decay parameter $p(\beta) = U[0, \beta_{\max}]$, with $\beta_{\max} = 1$. In all simulations, the realized channel length was $L = 8$ and the decay rate was $\beta = 0.1$.

The initial state of the Markov chain for each of the algorithms was $L^{(1)} = 3$, $h_{1:3}^{(1)} = \mathbf{0.1}$ and $\beta^{(1)} = 1$. The standard deviation for the *within-model move proposal* distribution was $\sigma_1 = 0.25$, the standard deviation for the *between-model move proposal* distribution was $\sigma_2 = 0.1$ and the standard deviation for the *proposal* for β was $\sigma_{\beta} = 0.25$. In **Algorithm 2** the pre-specified weights were set as $\pi_{1:L_{\max}} = \left[\frac{1}{L_{\max}}, \dots, \frac{1}{L_{\max}} \right]$, $\widehat{g}_{1:L_{\max}}^{(1)} = [1, \dots, 1]$ and the gain factor was set as $\gamma_t = \frac{2 L_{\max}}{\max(2 L_{\max}, t)}$. In **Algorithm 3** in constructing the CPS *proposals* the grids specifications were $m = 50$, $n = 50$ and $w = 0.05$. Keeping w constant meant that we used linear spacing and we also attached Gaussian tails at the left and right end points of the grid as in the **CPS-TDMCMC: Between-Model Moves** algorithm.

In the sensitivity studies, data were generated from the true model parameters and Algorithm 1 was run to produce a Markov chain of length 20k after discarding 5k samples as burn-in. These samples were then used to perform analysis of the posterior quantities of interest. In the convergence analysis, 100 independent data sets were generated from the true model using a

seeded random generator. For each realization of the observation we ran each algorithm with Markov chains of length 100k and discarded the first 20k samples as burn-in, ensuring each algorithm was compared on the same realized data sets. Estimates were obtained by averaging over the posterior estimated quantities of interest for each realized data set.

15.9.2 Model Sensitivity Analysis

Here we study the sensitivity of the posterior distribution (15.4.9) to provide insight into the performance of the proposed model. This separates the effects of model sensitivity from the effects of convergence rates of our proposed trans-dimensional sampling algorithms. We set $SNR = 15dB$ and the prior mean of the truncated Poisson model was $\lambda = 8$, the true model order.

We analyse the sensitivity of the MAP model order to the prior choice for model order, and provide recommendations for the prior on the model order. Secondly, we analyse the sensitivity of the MAP model order L , to the value of β used to generate the data set. Finally, we analyse the sensitivity of the marginal posterior distributions for $p(\beta|\mathbf{y})$ and $p(L|\mathbf{y})$ as a function of the SNR level.

Sensitivity of model order to prior choice $p(L)$

A popular choice for the model index prior is a truncated Poisson distribution for L . We study the sensitivity of the MAP estimate under this model order prior as a function of the prior mean. In general if posterior sensitivity to the choice of prior parameters is observed one can remove this sensitivity through the use of a hierarchical Bayesian model structure. This typically involves making the parameters of the prior distribution which directly cause prior sensitivity into random variables to be jointly estimated as part of the posterior inference, and placing an uninformative prior on these parameters. This is beyond the scope of this paper, but is often used in papers in which model selection is found to be sensitive to prior specification of the model order.

In this study we analyse $\lambda = \{2, 4, 6, 8, 10, 12, 14, 16\}$. Fig. 15.11.1 demonstrates the estimated marginal posterior distribution of the model order $p(L|\mathbf{y})$ as a function of the model order prior mean λ . We see that the marginal posterior distribution is only mildly sensitive to the prior mean λ parameterizing the truncated Poisson distribution, hence we did not add additional complexity of hyper priors and hierarchical Bayesian modeling. However, in general we recommend that one should use an uninformative prior choice when performing model selection under our framework. This is typically achieved by either using a uniform discrete prior on the model index or through the use of a truncated Poisson distribution as used in this paper. Generally this allows the data to drive the model selection estimation.

Sensitivity of model order to the true decay rate β used to generate the data

In this analysis we perform model order selection for various values of β . To illustrate this we study the MAP estimate of L as a function of varying the true β used to generate the data,

$\beta = \{0, 0.1, 0.2, 0.6, 0.7, 0.8\}$. The results of this study are depicted in Fig. 15.11.2.

These results demonstrate the estimated posterior model probabilities $p(L|\mathbf{y})$, for a given realization of data. We also present the corresponding PDP decay rate for each value of β .

They confirm that as β increases, the coefficients taps become less uniform. In particular, the coefficients of the last few taps become statistically insignificant from 0. Hence, as β increases, for our fixed SNR, the difference between the power of each tap increases and in particular the first few taps h_1, h_2, \dots will dominate compared to \dots, h_{L-1}, h_L . This results in many of the channel coefficients at higher orders, i.e. \dots, h_{L-1}, h_L , being indistinguishable from noise and ultimately, as demonstrated results in the MAP estimate for the model order L being less than the true model order which we set as $L = 8$.

This behavior is expected in this model. At low SNR levels this implies that generally the true MAP estimate for the posterior model probability, for a given data realisation will be significantly lower than the L used to generate the data. This mismatch in model order disappears asymptotically with the data size. Hence, when we use our trans-dimensional sampling algorithms to estimate the posterior model probabilities and take the MAP estimate these will also be lower than the L used to generate the data.

The impact of this on symbols detection will also be demonstrated in Section 15.9.4. We expect only minor impact on the BER for a give SNR when not detecting taps with very low power relative to the noise level.

Analysis of posterior precision for marginals, $p(\beta|\mathbf{y})$ and $p(L|\mathbf{y})$ as a function of SNR

In Fig. 15.11.3 we assess how the SNR level effects the marginal posterior distribution for β . In particular we demonstrate that at low SNR levels, the distribution of β that results from our model is left skewed and heavy tailed. However, the resulting MAP and MMSE estimates are still reasonably accurate. Then as the SNR increases, the distribution of β becomes more symmetric and centered on the true beta value used to generate the data.

In Fig. 15.11.4 we also present the marginal distributions of L as a function of SNR. Clearly, we also see that as the SNR increases, the precision of the marginal posterior for $p(L|\mathbf{y})$ increases. The mode of this distribution shifts towards the true value of L used to generate the data, demonstrating the convergence rate as a function of SNR.

In summary, these results show that as the SNR increases, the ability to distinguish taps with low power from noise increase. Hence, this corresponds to a more precise distribution for the marginal posterior, $p(\beta|\mathbf{y})$ and also results in the posterior model probabilities $p(L|\mathbf{y})$ producing a MAP estimate for L which corresponds to the true L used to generate the data. Hence these results provide an indication around the rate at which the posterior distributions precision changes as a function of SNR.

Estimated pairwise marginal posterior distributions of $p(h_i, h_j|L_{MAP}, \mathbf{y})$ and $p(h_i, \beta|L_{MAP}, \mathbf{y})$

In Fig. 15.11.5 we see the joint pairwise marginal estimated distributions of $p(h_i, h_j|L_{MAP}, \mathbf{y})$

and $p(h_i, \beta | L_{MAP}, \mathbf{y})$. These demonstrate that no correlation is present between pairs of variables within the posterior distribution. This shows that rate of convergence of our Markov chain sampler, in our case a Gibbs sampler, will be unaffected by correlation between parameters in the posterior.

15.9.3 Comparative Performance of Algorithms

In this section we compare the performance of the proposed algorithms. We generated synthetic data with the true model order $L = 8$. Additionally we also selected SNR=20 and the true $\beta = 0.1$ as we found that this resulted in a very high posterior model probability for $\Pr(L = 8 | \mathbf{y})$, which simplifies the convergence rate analysis. The advantage of this simulation set up is that we can now assess the convergence of the different samplers with respect to time. This was achieved by plotting the convergence of the Mean Squared Error (MSE) in the estimated marginal posterior probability of $\Pr(L^{(t)} = 8 | \mathbf{y})$ as a function of simulation time of each of the algorithms. Averaging these results over the 100 independent realisations allows us to compare the performance of each algorithm in terms of mixing between different models. In obtaining the MSE estimates, we took true posterior model probabilities for $\Pr(L = 8 | \mathbf{y})$ as the posterior model probability after running the BD-TDMCMC algorithm for 10^6 iterations.

In this study, for the *birth and death Gaussian proposal* when sampling two new components (h_L, h_{2L}) we simulated using two different values for the variance, $\sigma_{BD} = 0.05$ and $\sigma_{BD} = 0.2$. These were just selected arbitrarily as the simple BD-TDMCMC approach does not provide a method to determine an optimal parameter for the *proposal* other than via off-line tuning. We did not use this approach for the sake of comparison purposes in this example. In the SA-TDMCMC sampler we used the same *proposal* as the BD-TDMCMC algorithm and included the stochastic approximation stage. Finally, for the CPS-TDMCMC algorithm we selected the adaptive grid centered on the average estimate of the posterior mode.

In Fig. 15.11.6 we present a comparison between the algorithms (BD-TDMCMC, SA-TDMCMC, CPS-TDMCMC) for the average MSE of the marginal posterior model probability, $p(L = 8 | \mathbf{y})$ averaged over each realized observation set. The plots present the average MSE between the estimated posterior model probability of $L = 8$ at simulation time t and the true posterior model probability. Note that the CPS *proposal* performs best in terms of the MSE criterion, however the SA-TDMCMC algorithm also performs well. We also present the distributions of the posterior model probabilities for these simulations. Clearly these results agree with the average MSE results, the CPS *proposal* provides the best performance in terms of convergence of the posterior model probabilities in this analysis. However, the computational cost of this approach is significantly higher than the simple BD-TDMCMC or the SA-TDMCMC. When deciding which algorithm to recommend, we point out that the advantage of the CPS-TDMCMC algorithm is that it is largely automated. One only needs to select the number of grid points to include in the approximation of the optimal *proposal*. In contrast the SA-TDMCMC and BD-TDMCMC require specific choices to be made with respect to properties of the *proposal* distribution. These will impact the performance and should really be tuned off-line. Depending

on the amount of tuning required, the simulation using these approaches could cost as much as the CPS-TDMCMC *proposal* constructions. In our simulations the most computationally efficient algorithm was BD-TDMCMC and was 20 % more efficient in terms of computer time compared to SA-TDMCMC algorithm and 44 % more efficient than CPS-TDMCMC algorithm.

Our analysis suggests that we proceed with use of the CPS algorithm. It had marginally better convergence rate in this study than the SA-TDMCMC algorithm and is an automated TDMCMC algorithm, requiring only simple interpretable user choices in terms of the number of grid points to use in constructing the *proposal*. Having selected the CPS-TDMCMC algorithm we perform detailed studies of MSE of $\hat{\mathbf{h}}_{MMSE}$ in the setting in which we perform joint estimation of \mathbf{h}, L, β . In particular we will compare the performance of this estimator to the bounds presented in Section 15.8. In addition we present the BER vs SNR performance.

15.9.4 Algorithm Performance

In this Section we evaluate both channel estimation MSE and BER Vs. SNR performance of the CPS-TDMCMC algorithm. The frame length was 128 OFDM symbols. At the beginning of each frame one OFDM symbol was composed of known symbols for the purpose of channel estimation using Algorithm 3.

We first evaluate the channel estimation MSE performance of the proposed algorithm in comparison to the bounds presented in Section 15.8. These results are depicted in Fig. 15.11.7. The results demonstrate the following key points:

- As SNR increases the numerical estimate of the MSE of $\hat{\mathbf{h}}_{MMSE}$ obtained using Algorithm 3 converges to the BCRLB. This confirms our previous findings in the sensitivity studies, in that for low SNR the estimated L does not correspond to the true L . As a result we see that the MSE of the estimator of \mathbf{h} is above the BCRLB. However, it is important to point out the following two things:
 1. For low SNR values the MSE for the channel estimate is still close to the BCRLB, and converges as SNR increases.
 2. In the region of SNR values for which the MSE is above the BCRLB this does not adversely affect the BER performance as we demonstrate below.
- For all SNRs the numerical estimate of the MSE of $\hat{\mathbf{h}}_{MMSE}$ obtained using Algorithm 3 is always significantly below the saturated model bound, $B_{|L_{\max}, \beta}$. This is not surprising, since as long as the estimate of β and L is accurate, then we know that the saturated model MSE, given by $B_{|L_{\max}, \beta}$, should upper bound the MSE of $\mathbf{h}_{MMSE} | \beta_{MAP}, L_{MAP}$.
- The bound $B_{|\beta}$ involves BMA. It is interesting to note that for low SNR the posterior model probabilities which we can obtain in this setting exactly, for each realisation of the data, favour underestimation of L as we demonstrated. As a result, when calculating this bound for low SNR values we obtain a non tight bound relative to the BCRLB, since most

of the posterior model probability is given to lower model orders. Hence we find that for low SNR values this bound is not tight.

- When considering $B_{|L}$, we use the estimator $Tr \left\{ \hat{\mathbf{J}}_{|L} \right\}$ given by (15.8.10). When comparing the estimate of this bound to the BCRLB, we see that as expected, it lies between the BCRLB and the MSE of \mathbf{h}_{MMSE} obtained using Algorithm 3.

Next we evaluate the BER performance of the proposed algorithm. As a reference, we compared the BER results of the proposed algorithm with two lower bounds:

- *MMSE perfect* : Here the channel length L and decay parameter β are perfectly known at the receiver, and can be estimated using (15.4.1). This serves as a lower bound for the proposed algorithm.
- *CSI*: Here the CIR \mathbf{h} is perfectly known at the receiver. This will serve as a loose lower bound of the BER of the system.

The detection method for all schemes is based on transforming the estimated channel $\hat{\mathbf{h}}$ to the frequency domain and performing one-tap MMSE equalization. The simulation results are depicted in Fig. 15.11.8. These results show that both MAP and MMSE estimates of the proposed algorithm operate close to *MMSE perfect*. These results demonstrate close to optimal performance of our chosen algorithm in estimating the unknown model parameters which are then used in the detection scheme, resulting in BER performance close to the lower bound.

15.10 Conclusions

In this paper we presented novel algorithms for channel estimation in OFDM systems, where the length of the channel, the channel coefficients and the power decay profile are unknown. We constructed a Bayesian model and then developed novel TDMCMC algorithms to estimate quantities of the posterior distribution such as MMSE and MAP estimates of parameters. We then performed a sensitivity analysis and assessed performance of the algorithms we developed. Finally, after selecting one of the algorithms we performed analysis of the estimation error and BER versus SNR.

15.11 acknowledgement

The authors thank Scott Sisson, Yanan Fan and Pavel Shevchenko. Additionally, thank you goes to UNSW Statistics for APA funding of one of the authors and to CSIRO Mathematical and Information Sciences for support. We would also like to acknowledge the support of the ARC Discovery Project.

Appendix

In this Appendix, we derive the expression for the gradient of the full log posterior with respect to a generic element $h_i \in \mathbf{h}_{1:2L}$. Since we wish to work with only real components, we decompose the model (15.3.1) as follows:

$$\tilde{\mathbf{y}} = \tilde{\mathbf{A}}\tilde{\mathbf{h}}^T + \tilde{\mathbf{z}}, \quad (15.11.1)$$

where

$$\tilde{\mathbf{A}} \triangleq \begin{bmatrix} \text{Re}\{\mathbf{A}\} & -\text{Im}\{\mathbf{A}\} \\ \text{Im}\{\mathbf{A}\} & \text{Re}\{\mathbf{A}\} \end{bmatrix}, \tilde{\mathbf{y}} \triangleq \begin{bmatrix} \text{Re}\{\mathbf{y}\} \\ \text{Im}\{\mathbf{y}\} \end{bmatrix}; \tilde{\mathbf{z}} \triangleq \begin{bmatrix} \text{Re}\{\mathbf{z}\} \\ \text{Im}\{\mathbf{z}\} \end{bmatrix}, \quad (15.11.2)$$

where $\text{Re}(\cdot)$ and $\text{Im}(\cdot)$ are the real and imaginary parts of (\cdot) , respectively, and $\mathbf{A} \triangleq \mathbf{D}\mathbf{W}_L$. The gradient of the full log posterior with respect to a generic element $h_i \in \mathbf{h}_{1:2L}$ can be written as:

$$\frac{\partial}{\partial h_i} \{\log p(h_{1:2L}, \beta, L | \tilde{\mathbf{y}})\} = -\frac{1}{\sigma_z^2} \left(-2\tilde{\mathbf{y}}^T \tilde{\mathbf{A}}_{:,i} + \tilde{\mathbf{A}}_{i,:} \tilde{\mathbf{A}} \tilde{\mathbf{h}}^T + \mathbf{h}^T \tilde{\mathbf{A}} \tilde{\mathbf{A}}_{:,i} \right) - \frac{2}{\sigma_{h_i}^2} h_i, \quad (15.11.3)$$

where $\tilde{\mathbf{A}}_{:,i}$ corresponds to the i -th column of $\tilde{\mathbf{A}}$.

Next we equate (15.11.3) to zero and solve for h_i

$$\frac{-2h_i}{\sigma_{h_i}^2} + \frac{2}{\sigma_z^2} \tilde{\mathbf{y}}^T \tilde{\mathbf{A}}_{:,i} - \frac{1}{\sigma_z^2} \tilde{\mathbf{A}}_{i,:} \tilde{\mathbf{A}} \tilde{\mathbf{h}}^T - \frac{1}{\sigma_z^2} \mathbf{h}^T \tilde{\mathbf{A}} \tilde{\mathbf{A}}_{:,i} = 0. \quad (15.11.4)$$

By defining $\mathbf{h}_{i=0}$ as \mathbf{h} with the element in the i -th location set to zero and \mathbf{e}_i as a column indicator vector, meaning that its elements are all set to zero except for the i -th element which is set to one. We can now rewrite (15.11.4) as

$$\frac{-2h_i}{\sigma_{h_i}^2} + \frac{2}{\sigma_z^2} \tilde{\mathbf{y}}^T \tilde{\mathbf{A}}_{:,i} - \frac{1}{\sigma_z^2} \tilde{\mathbf{A}}_{i,:} \tilde{\mathbf{A}} [\mathbf{h}_{i=0}^T + h_i \mathbf{e}_i] - \frac{1}{\sigma_z^2} [\mathbf{h}_{i=0}^T + h_i \mathbf{e}_i] \tilde{\mathbf{A}}^T \tilde{\mathbf{A}}_{:,i} = 0. \quad (15.11.5)$$

Rearranging (15.11.5) we obtain the expression

$$h_i = \frac{2\tilde{\mathbf{y}}^T \tilde{\mathbf{A}}_{:,i} - \tilde{\mathbf{A}}_{i,:} \tilde{\mathbf{A}} \mathbf{h}_{i=0}^T - \mathbf{h}_{i=0}^T \tilde{\mathbf{A}}^T \tilde{\mathbf{A}}_{:,i}}{\frac{2\sigma_z^2}{\sigma_{h_i}^2} + \tilde{\mathbf{A}}_{i,:} \tilde{\mathbf{A}} \mathbf{e}_i + \mathbf{e}_i \tilde{\mathbf{A}}^T \tilde{\mathbf{A}}_{:,i}}. \quad (15.11.6)$$

15.11.1 Adaptive Grid Placement centred on Estimated Posterior Mode

Constructing the CPS *proposal* computationally efficiently involves concentrating grid points in support regions in which the posterior has most mass, i.e. posterior mode. Estimating the mode of $p(h_i | \mathbf{y}, \mathbf{h}_{\neq i} \beta, L)$ follows from above. The grid can be concentrated around this location.

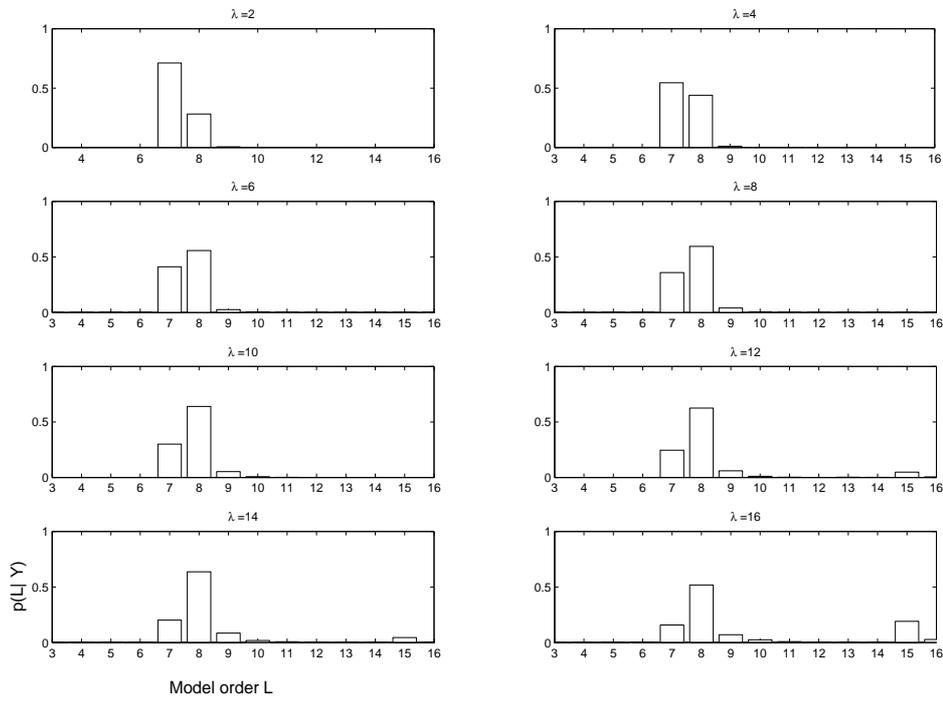


Fig. 15.11.1: Sensitivity of MAP estimate from $P(L|y)$ to prior mean λ

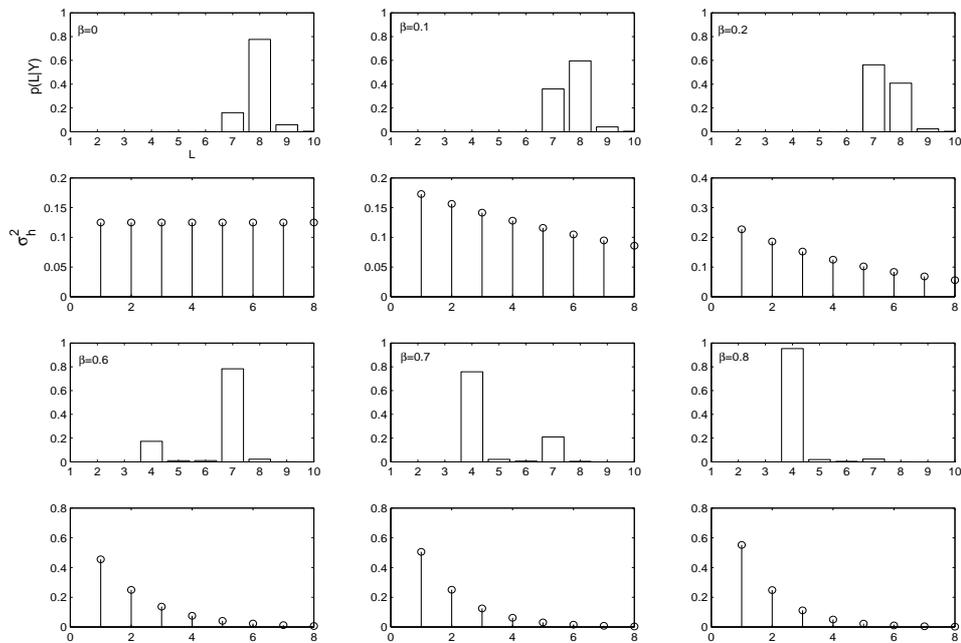


Fig. 15.11.2: Sensitivity of MAP estimate from $P(L|y)$ to β

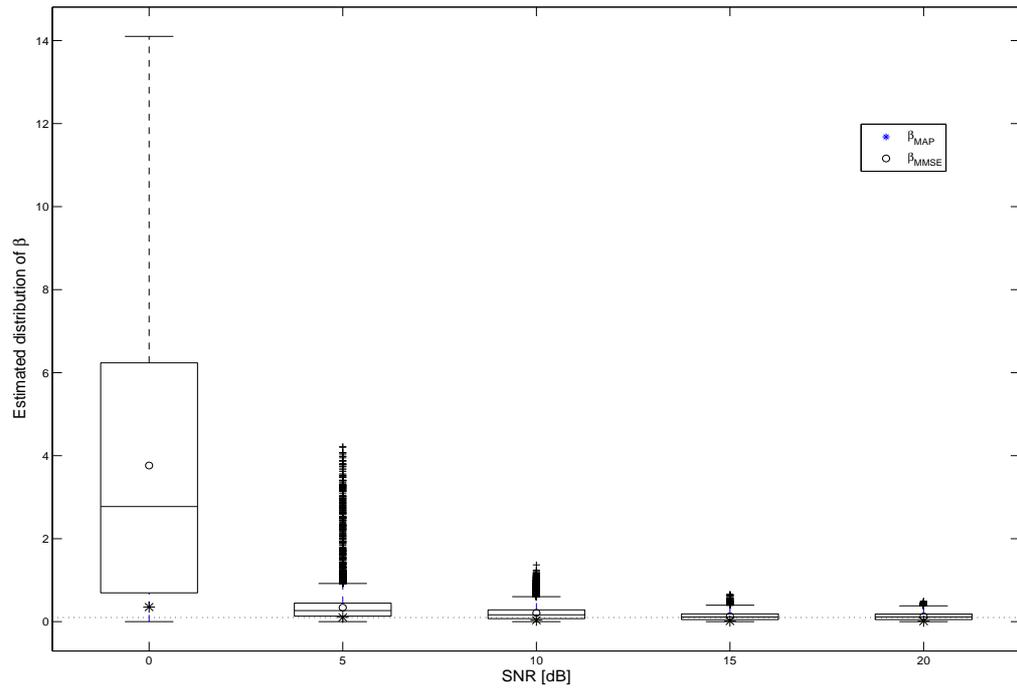


Fig. 15.11.3: Sensitivity of MAP estimate from $P(L|y)$ to β

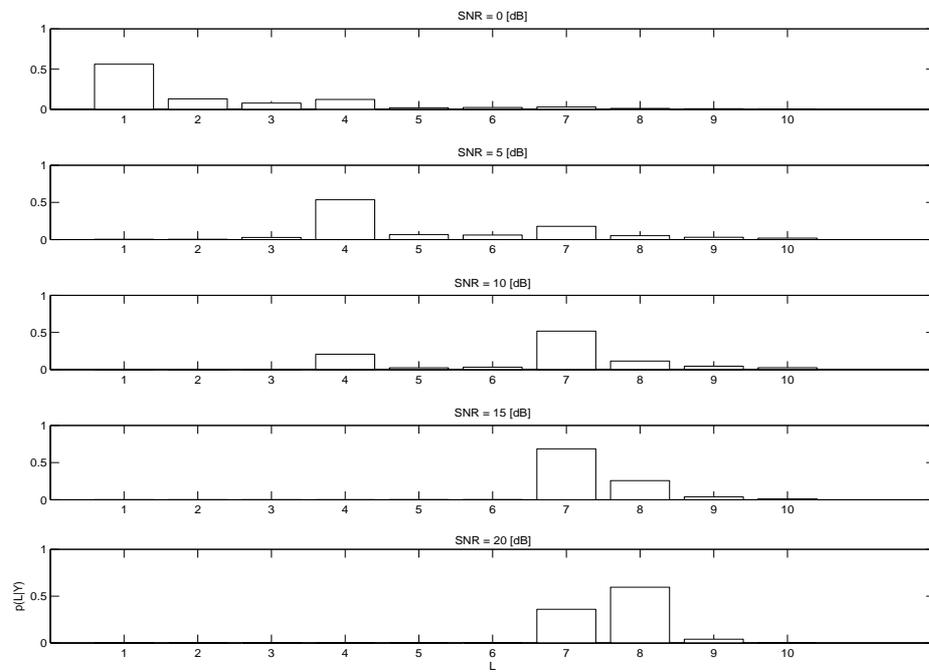


Fig. 15.11.4: Marginal distribution of L , $p(L|y)$ Vs. SNR

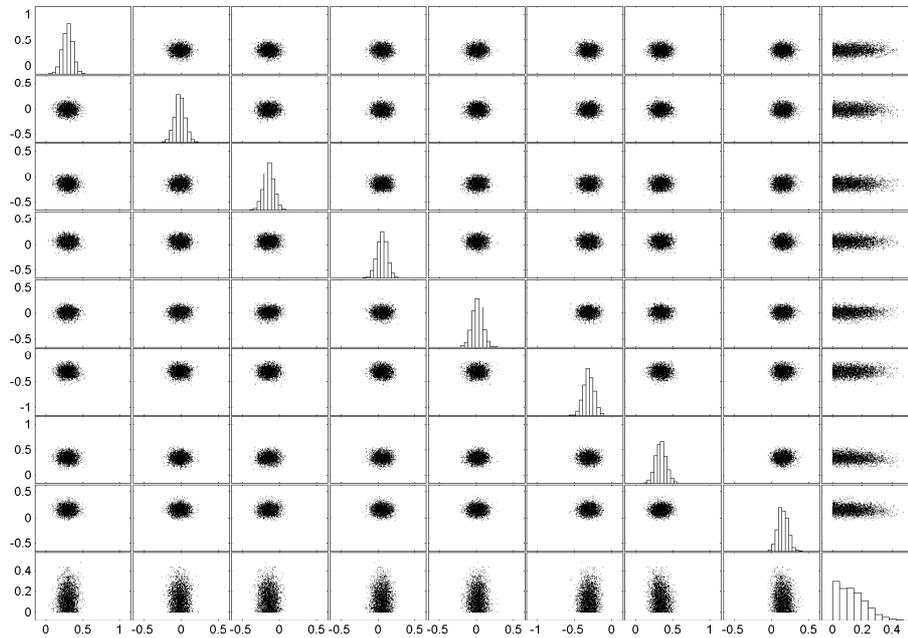


Fig. 15.11.5: Pairwise Marginal Posterior Distributions for $p(h_i, h_j | \mathbf{y})$

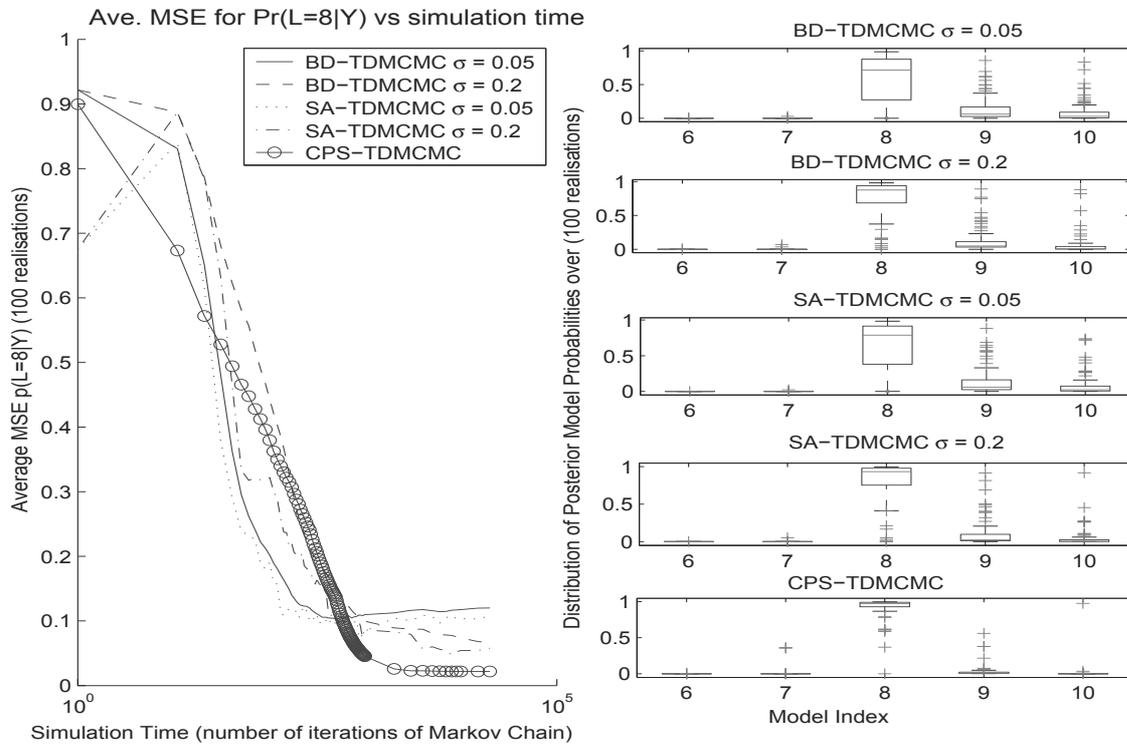


Fig. 15.11.6: Average MSE of the marginal posterior model probability, $p(L = 8 | \mathbf{y})$

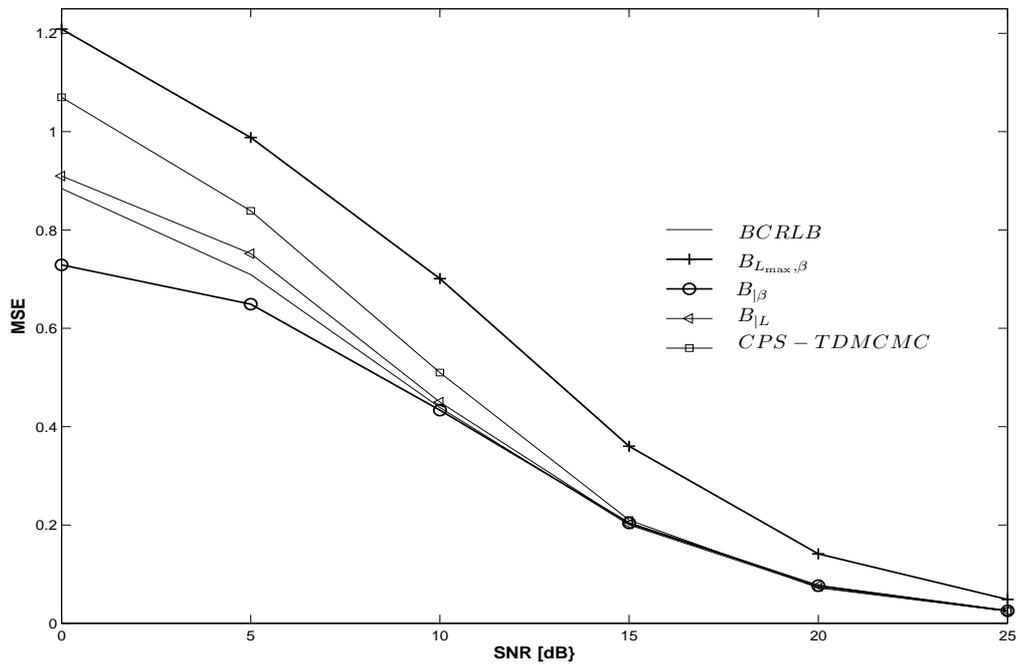


Fig. 15.11.7: MSE performance for the CPS-TDMCMC algorithm with $K = 64, L = 8, \beta = 0.1$ in comparison to several bounds

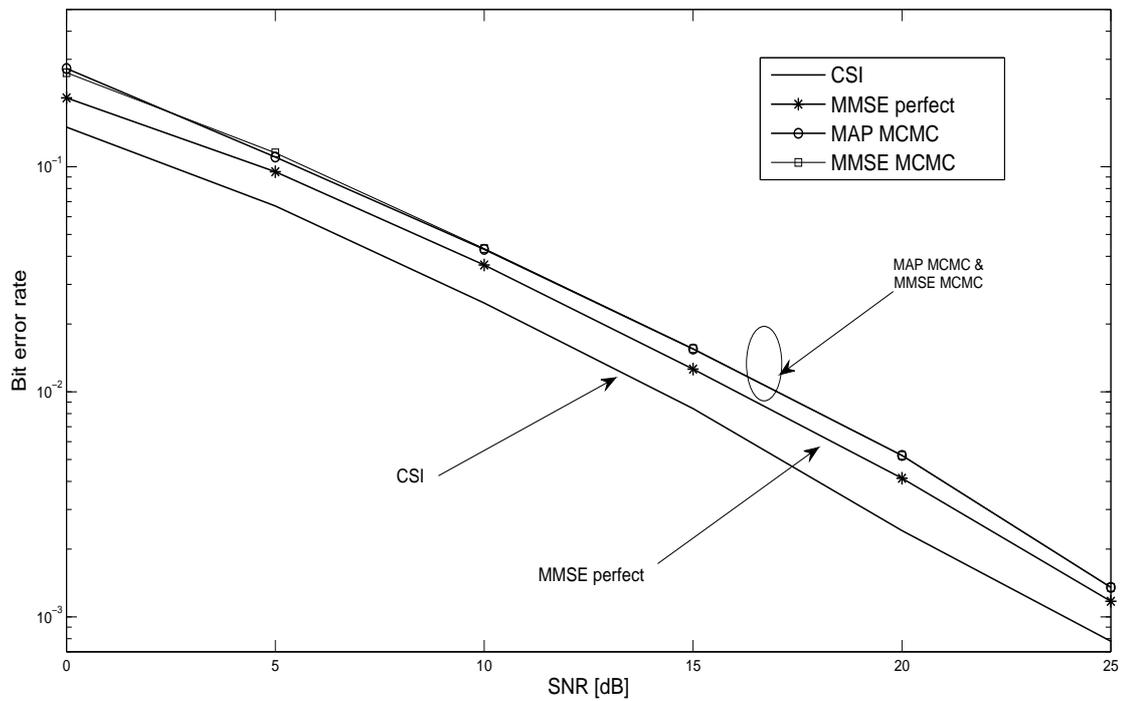


Fig. 15.11.8: BER performance for the CPS-TDMCMC algorithm with $K = 64, L = 8, \beta = 0.1$

| Complexity of <i>within-model moves</i> in (15.5.3) for deterministic scan <i>MH-within-Gibbs</i> | |
|--|-------------------------------------|
| Operation | Number of operations |
| Sampling $\prod_{i=1}^{L^{(t-1)}} T((h_{1:L^{(t-1)}}^{(t-1)}) \rightarrow (h_i^*))$ | $L^{(t-1)} \mathbb{O}(1)$ |
| Evaluating the acceptance probability $a(h_i^{(t-1)}, h_i^*)$ | $KL^{(t-1)} (C_m + C_a)$ |
| $T((\beta^{(t-1)}) \rightarrow (\beta^*))$ | $\mathbb{O}(1)$ |
| Evaluating acceptance probability $a(\beta^{(t-1)}, \beta^*)$ | $KL^{(t-1)} (C_m + C_a)$ |
| Complexity of <i>between-model moves</i> for BD (Algorithm 1) | |
| sampling $Q((h_{1:L^{(t-1)}}^{(t-1)}, \beta^{(t-1)}, L^{(t-1)}) \rightarrow (h_{1:L^*}^*, \beta^*, L^*))$ | $\mathbb{O}(1)$ |
| Evaluating acceptance probability $\alpha((h_{1:L^{(t-1)}}^{(t-1)}, \beta^{(t-1)}, L^{(t-1)}), (h_{1:L^*}^*, \beta^*, L^*))$ | $(L^* + L^{(t-1)}) (K (C_m + C_a))$ |

Tab. 15.1: Computational complexity of Algorithm 1

| Constructing forward <i>proposal</i> $q(h_{1:2L^{(t-1)}}^{(t-1)} \rightarrow h_{1:2L^*}^*)$ | |
|--|--|
| Operation | Number of operations |
| CPS-TDMCMC: <i>Between-Model Moves</i> (1.a.i) | $m \mathbb{O}(1)$ |
| CPS-TDMCMC: <i>Between-Model Moves</i> (1.a.ii), eq. (15.11.6) | $m (C_m + C_a) (5K + 4KL)$ |
| CPS-TDMCMC: <i>Between-Model Moves</i> (1.a.iii) | $(nm + n) \mathbb{O}(1)$ |
| CPS-TDMCMC: <i>Between-Model Moves</i> (1.a.iii), eq. (15.11.3) | $nmK (C_m + C_a) (3 + 2L)$ |
| CPS-TDMCMC: <i>Between-Model Moves</i> (1.a.v) | $n \mathbb{O}(1)$ |
| CPS-TDMCMC: <i>Between-Model Moves</i> (1.c.i) | $\mathbb{O}(1)$ |
| CPS-TDMCMC: <i>Between-Model Moves</i> (1.c.ii) | $nmK (C_m + C_a) (3 + 2L)$ |
| Sampling from forward <i>proposal</i> $q(h_{1:2L^{(t-1)}}^{(t-1)} \rightarrow h_{1:2L^*}^*)$ | |
| CPS-TDMCMC: <i>Between-Model Moves</i> (1.b) and (1.d) | $n \mathbb{O}(1)$ |
| Evaluating the acceptance probability $\alpha((h_{1:L^{(t-1)}}^{(t-1)}, \beta^{(t-1)}, L^{(t-1)}), (h_{1:L^*}^*, \beta^*, L^*))$ | |
| CPS-TDMCMC: <i>Between-Model Moves</i> (2) | $K (C_m + C_a) (L^* + L^{(t-1)} + m(6n + 8L^{(t-1)}))$ |

Tab. 15.2: Computational complexity of CPS (Algorithm 3)

References

- [1] Akaike, H. A new look at the statistical model identification *Automatic Control, IEEE Transactions on*, 1974, 19, 716-723
- [2] Andrieu, C.; de Freitas, N.; Doucet, A. Jordan, M. An Introduction to MCMC for Machine Learning *Machine Learning*, 2003, 50, 5-43
- [3] Andrieu, C. Moulines, E. On the ergodicity properties of some adaptive MCMC algorithms *Annals of Applied Probability of Applied Probability, The Institute of Mathematical Statistics*, 2006, 16, 1462
- [4] Andrieu, C.; Moulines, E. Priouret, P. Stability of Stochastic Approximation under Verifiable Conditions *Decision and Control, 2005 and 2005 European Control Conference. CDC-ECC'05. 44th IEEE Conference on*, 2005, 6656-6661
- [5] Atchade, Y. Rosenthal, J. On adaptive Markov chain Monte Carlo algorithms *Bernoulli-London-, International Statistical Institute*, 2005, 11, 815
- [6] Bedard, M. Rosenthal, J. Optimal scaling of Metropolis algorithms: Heading toward general target distributions
- [7] van de Beek, J.; Edfors, O.; Sandell, M.; Wilson, S. Borjesson, P. On channel estimation in OFDM systems *Vehicular Technology Conference, 1995 IEEE 45th*, 1995, 2
- [8] Bello, P.; Adcom, I. Cambridge, M. Characterization of Randomly Time-Variant Linear Channels *Communications, IEEE Transactions on [legacy, pre-1988]*, 1963, 11, 360-393
- [9] Ben-Haim, Z. Eldar, Y. Minimax Estimators Dominating the Least-Squares Estimator *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, 2005, 4
- [10] Brooks, S.; Giudici, P. Roberts, G. Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Blackwell Synergy, 2003, 65, 3-39
- [11] Carlin, B. Chib, S. Bayesian model choice via Markov chain Monte Carlo methods *Journal of the Royal Statistical Society. Series B. Methodological, Royal Statistical Society*, 1995, 57, 473-484

-
- [12] Choi, J. Lee, Y. Optimum pilot pattern for channel estimation in OFDM systems *Wireless Communications, IEEE Transactions on*, 2005, 4, 2083-2088
- [13] Cramer, H. A contribution to the theory of statistical estimation *Skand. Aktuarietidskr*, 1946, 29, 85-94
- [14] Doucet, A. Wang, X. Monte Carlo methods for signal processing: a review in the statistical signal processing context *Signal Processing Magazine, IEEE*, 2005, 22, 152-170
- [15] Fan, Y.; Brooks, S. Gelman, A. Output Assessment for Monte Carlo Simulations via the Score Statistic *Journal of Computational and Graphical Statistics, American Statistical Association*, 2006, 15, 178
- [16] Fan, Y.; Peters, G. Sisson, S. Automating and Evaluating Reversible Jump MCM proposal distributions. preprint, accepted: *Statistics and Computing*, 2008
- [17] Gilks, W.; Richardson, S. Spiegelhalter, D. *Markov Chain Monte Carlo in Practice* Chapman Hall/CRC, 1996
- [18] Green, P. *Trans-dimensional Markov chain Monte Carlo Highly Structured Stochastic Systems*, Oxford University Press, 2003, 27, 179-98
- [19] Green, P. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination *Biometrika, Biometrika Trust*, 1995, 82, 711-732
- [20] Kay, S. *Fundamentals of Statistical Signal Processing, Volume 2: Detection Theory* Prentice Hall PTR, 1998
- [21] Liang, F. A Generalized WangLandau Algorithm for Monte Carlo Computation *Journal of the American Statistical Association, American Statistical Association*, 2005, 100, 1311-1327
- [22] Liang, F.; Liu, C. Carroll, R. Stochastic Approximation in Monte Carlo Computation *Journal-American Statistical Association, ASA American Statistical Association*, 2007, 102, 305
- [23] Liu, H. Li, G. *OFDM-Based Broadband Wireless Networks: Design and Optimization* Wiley-Interscience, 2005
- [24] Minn, H. Bhargava, V. An investigation into time-domain approach for OFDM channel estimation *Broadcasting, IEEE Transactions on*, 2000, 46, 240-248
- [25] Nevat, I.; Peters, G. Yuan, J. OFDM CIR Estimation with Unknown Length via Bayesian Model Selection and Averaging *IEEE Vehicular Technology Conference, IEEE*; 2008, 2008, 1413-1417
- [26] Nguyen, V.; Kuchenbecker, H. Patzold, M. Estimation of the Channel Impulse Response Length and the Noise Variance for OFDM Systems *IEEE Vehicular Technology Conference, IEEE*; 1999, 2005, 61, 429
- [27] Raghavendra, M. Giridhar, K. Improving channel estimation in OFDM systems for sparse multipath channels *Signal Processing Letters, IEEE*, 2005, 12, 52-55

-
- [28] Rao, C. Information and the Accuracy Attainable in the Estimation of Statistical Parameters Bull. Calcutta Math. Soc., vol. 37, pp. 81 - 91, Springer-Verlag, 1945
- [29] Roberts, G.; Gelman, A. Gilks, W. Weak convergence and optimal scaling of random walk Metropolis algorithms Annals of Applied Probability, The Institute of Mathematical Statistics, 1997, 7, 110-120
- [30] Sisson, S. Transdimensional Markov Chains: A Decade of Progress and Future Perspectives Journal of the American Statistical Association, American Statistical Association, 2005, 100, 1077-1090
- [31] Van Trees, H. Detection, estimation, and modulation theory.. part 1,. detection, estimation, and linear modulation theory Wiley New York, 1968
- [32] Zhang, J. Liang, F. Convergence of stochastic approximation algorithms under irregular conditions Statistica Neerlandica, Blackwell Publishing, 2008, 62, 393-403

16

Journal Paper 12

"When all think alike, then no one is thinking."

Walter Lippman

Nevat I., Peters G.W. and Yuan J. (2009) "Detection of Gaussian Constellations in MIMO Systems Under Imperfect Channel State Information". *Journal IEEE Transactions of Communications, to appear.*

This work was instigated by Ido Nevat. The second author on this paper and can claim 50% of the credit for the contents. His work included jointly developing the methodology contained and the applications, the implementation and comparison to alternative approaches, writing the drafts of the paper. This work will be included in a PhD thesis of a co-author, though this thesis is not submitted by sequence of publications. This paper has been conditionally accepted to appear in the IEEE Transactions on Communications, and several peer reviewed conference papers relating to aspects of this work have been accepted and appeared in conference proceedings. Permission from all the co-authors is granted for inclusion in the PhD.

This paper appears in the thesis in a modified format to that which was submitted to the IEEE Transactions on Communications

Final print version will be available at:

<http://ieeexplore.ieee.org>

Detection of Gaussian Constellations in MIMO Systems under Imperfect CSI

Ido Nevat (*corresponding author*)

School of Electrical Engineering and Telecommunications, University of NSW, Sydney, Australia.

Gareth W. Peters

CSIRO Mathematical and Information Sciences, Locked Bag 17, North Ryde, NSW, 1670, Australia;

UNSW Mathematics and Statistics Department, Sydney, 2052, Australia;

email: peterga@maths.unsw.edu.au

Jinhong Yuan

School of Electrical Engineering and Telecommunications, University of NSW, Sydney, Australia.

16.1 Abstract

This paper considers the problem of symbols detection in MIMO systems at the presence of channel estimation errors. Under this framework we develop a computationally efficient approximation of the MAP detector for non-uniform constellations. First we review the channel estimation error in the setting of known orthogonal training sequences for channel estimation. We analyze the performance degradation due to noisy channel estimation. Next we propose a low complexity detector based on a relaxation of the discrete nature of the digital constellation and the channel estimation error statistics. This leads to a non-convex program that is solved efficiently via a hidden convexity minimization approach. Simulation results in a random MIMO system show that the proposed algorithm outperforms the linear MMSE receiver in terms of BER.

Keywords: MIMO, MAP estimation, Gaussian constellations, Bayesian EM

16.2 Introduction

Multiple-Input Multiple-Output (MIMO) systems can yield vast capacity increases when a rich scattering environment is properly exploited (1). A MIMO system employs multiple antennas at both the transmitter and the receiver, and its capacity increases linearly as the minimum of the number of transmit and receive antennas. However, in practice the channel conditions must be estimated since perfect channel knowledge is never known *a priori*. Typically, this is performed in two distinct stages. The transmitter first sends a known sequence of symbols, which allow channel estimation to be performed at the receiver. Next, the information symbol sequence is transmitted, and sequence detection is performed at the receiver using the estimated channel. In practice, the channel estimation procedure can be aided by transmitting pilot symbols that are known at the receiver. System performance depends on the quality of the channel estimate, and the number of pilot symbols as shown in (2). It is desirable to minimize the number of transmitted pilot symbols in order to maximize spectral efficiency.

We consider the case in which one wishes to transmit symbols designed to achieve maximum throughput of a channel. In the linear Gaussian channel model that we consider in this paper, it is well known that in order to achieve capacity, powerful coding schemes must be combined with shaping methods which result in near-Gaussian distribution of the symbol's amplitude (3; 4). Two practical schemes that obtain shaping gain are "trellis shaping" (5) and "shell mapping" (6). Though theory states that the signal points should be chosen from a continuous Gaussian distribution, in practice, since the constellations are finite this means that optimal gain can not be achieved. In performing constellation shaping, we note that approximating the optimal Gaussian with a discrete distribution can be achieved in many ways. We focus here on the *Maxwell-Boltzman* (M-B) distribution (7).

The optimal detector for digital constellations can be implemented by using a brute-force approach which searches over all symbol possibilities. Typically this is impractical due to the massive computational burden it presents. Several alternative suboptimal, but computationally more tractable, receivers have been considered in the literature e.g. the sphere detector (8) and BLAST (9). The most common class of suboptimal detectors is the class of linear detectors, i.e. the matched filter (MF), the decorrelator or zero forcing (ZF), and the minimum mean-squared error (MMSE) detectors (10).

In this work we focus on the MMSE and the maximum a posteriori (MAP) detectors. A common practice in detection schemes, is to consider a relaxation of the discrete problem. The relaxation leads to a continuous optimization problem, that although suboptimal, in many cases provides a tractable optimization problem, that is simpler to solve. The majority of the literature concentrates on the basic case, in which it is assumed that the channel matrix is completely specified (10). In this setting, the MMSE, Linear MMSE (LMMSE) and MAP estimators coincide and have a simple closed form solution and identical detection performance. In contrast, when noisy channel estimate is considered, the MMSE, LMMSE and MAP approaches lead to different estimators. In fact, we will show that the solution of the MMSE leads to an intractable

integration, whereas the MAP estimator can be efficiently found. The MAP estimator was derived in (1), as an extension of the work in (12)-(13).

The main contributions of this work are firstly to present a complete framework for detection of Gaussian constellations in MIMO systems under imperfect channel state information (CSI). Second, we develop a low complexity receiver which outperforms the LMMSE detector in terms of BER. This receiver will exploit the hidden convexity of the estimation problem. In achieving these goals we also develop and compare this receiver to a receiver based on a Bayesian Expectation Maximization (BEM) algorithm. As part of this framework development we show that the channel estimation error translates into an additive noise term and we proceed to demonstrate the effect on the overall SNR of the system.

This paper is organized as follows. In Section 16.3, we describe the MIMO system model, then in Section 16.4 an overview of pilot aided channel estimation is presented. Section 16.5 presents MAP detectors without taking the channel estimation error into account. Section 16.6 discusses MAP detectors in the framework which includes the channel estimation error. Simulation results are presented in 16.7 and concluding remarks are given in Section 16.8.

The following notation is used throughout, boldface upper case letters denote matrices, boldface lower case letters denote column vectors, and standard lower case letters denote scalars. The superscripts $(\cdot)^T$, $(\cdot)^H$ and $(\cdot)^\dagger$ denote the transpose, Hermitian and the pseudoinverse, respectively. By \mathbf{I} we denote the identity matrix. $\|\cdot\|$ is the standard Euclidean norm and $\|\cdot\|_F^2$ is the Frobenius norm. $\lambda_{\min}(\mathbf{x})$ is the smallest eigenvalue of \mathbf{x} and $\mathbf{x} \succeq 0$ means that the matrix is a symmetric positive semidefinite matrix. The functions $p(\mathbf{x})$, $p(\mathbf{x}|\mathbf{y})$ and $E\{\cdot\}$ denote the probability distribution function (PDF) of \mathbf{x} , the PDF of \mathbf{x} given \mathbf{y} , and the expectation, respectively. The operation $\text{vec}(\mathbf{H})$ denotes the vector obtained from stacking the columns of \mathbf{H} .

16.3 System Description

Consider a flat fading MIMO communication system of M transmit and N receive antennas. The data stream is multiplexed to M data substreams and transmitted by M transmit antennas simultaneously. The baseband equivalent model of the received signal vector at the instant of sampling can be represented by

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{w}, \quad (16.3.1)$$

where $\mathbf{y} = [y_1 \ y_2 \ \dots \ y_N]^T$ is an $N \times 1$ received vector, $\mathbf{x} = [x_1 \ x_2 \ \dots \ x_M]^T$ is an $M \times 1$ transmitted vector with $E\{|x_i|^2\} = \sigma_x^2$, $\mathbf{w} = [w_1 \ w_2 \ \dots \ w_N]^T$ is an $N \times 1$ additive noise vector. We assume that the noise is white complex Gaussian, whose elements have zero mean and covariance matrix $\sigma_w^2 \mathbf{I}$. \mathbf{H} is an $N \times M$ channel matrix, its elements $h_{i,j}$ are independent random complex variables which model the fading gain from antenna j to antenna i . The assumptions made

about the channel are that it is quasi-static and constant over the length of a frame. However, it may change independently between consecutive frames. We assume that the antenna elements are spaced sufficiently apart and there are enough scatterers present, so that the channel paths are suitably modeled as independent and uncorrelated. The elements $h_{i,j}$ follow a complex Gaussian distribution with zero-mean and unit variance. The input symbol vector \mathbf{x} is taken from a (discrete) finite Gaussian distributed signal set \mathbf{D} . This choice of signal distribution has the property that it reduces the average transmitted power.

16.4 Pilot Aided Maximum Likelihood Channel Estimation

For completeness, we now provide an overview of channel estimation using pilot symbols, and the consequently channel estimation error. In the model considered in this paper the receiver has no knowledge of channel state information. Hence, pilot symbols are embedded in the data stream in order to estimate the channel matrix \mathbf{H} . Here we discuss the maximum likelihood (ML) method for channel estimation and build a stochastic model for the estimated channel matrix. We assume that the elements of the channel matrix \mathbf{H} have *deterministic, unknown* values, and therefore the standard ML estimation for the channel matrix \mathbf{H} shall be used. In order to estimate \mathbf{H} , p training vectors $\mathbf{x} = [\mathbf{x}_1, \dots, \mathbf{x}_p]$ are transmitted. The corresponding observation matrix $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_p]$ can be expressed as

$$\mathbf{Y} = \mathbf{H}\mathbf{X} + \mathbf{W}, \quad (16.4.1)$$

where $\mathbf{w} = [\mathbf{w}_1, \dots, \mathbf{w}_p]$ is the $N \times p$ noise matrix. As specified in (14), the ML estimation of the channel matrix $\hat{\mathbf{H}}$ is

$$\begin{aligned} \hat{\mathbf{H}} &= \arg \max_{\mathbf{H}} p(\mathbf{Y}|\mathbf{H}, \mathbf{X}, \sigma_w^2) \\ &= \arg \max_{\mathbf{H}} \exp \left\{ -\frac{1}{2\sigma_w^2} (\mathbf{Y} - \mathbf{H}\mathbf{X})^H (\mathbf{Y} - \mathbf{H}\mathbf{X}) \right\}. \end{aligned} \quad (16.4.2)$$

It can be easily shown that the optimal (in the mean squared error sense) training matrix \mathbf{x} of (16.4.1) should satisfy (2)

$$\mathbf{X}\mathbf{X}^H = \frac{P}{M}\mathbf{I}, \quad (16.4.3)$$

where P is the total transmitted power for training vectors. Therefore, any matrix with orthogonal rows of the same power $\sqrt{P/M}$ is optimal, and can be produced by using the complex Walsh codes generated by the Hadamard matrix. The estimation error under optimal training

is given by (2)

$$\min_{\mathbf{x}} E \left\{ \left\| \mathbf{H} - \hat{\mathbf{H}} \right\|_F^2 \right\} = \frac{\sigma_w^2 N M^2}{P}, \quad (16.4.4)$$

Here we assume the estimated channel matrix can be modeled as

$$\hat{\mathbf{H}} = \mathbf{H} + \Delta, \quad (16.4.5)$$

where Δ is a random matrix of mutually independent, zero mean Gaussian elements, independent of \mathbf{W} . The variance of each element of Δ is

$$\sigma_h^2 = \frac{\sigma_w^2 M}{P}. \quad (16.4.6)$$

The estimated channel matrix $\hat{\mathbf{H}}$ can be interpreted as a Gaussian random matrix with mean equal to the true channel matrix \mathbf{H} , and a diagonal covariance matrix $C_{\mathbf{H}} = \sigma_h^2 \mathbf{I}$. Substituting the estimated channel matrix $\hat{\mathbf{H}}$ in Eq. (16.3.1) and (16.4.5), we obtain the following

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{w} = \hat{\mathbf{H}}\mathbf{x} - \Delta\mathbf{x} + \mathbf{w} \quad (16.4.7)$$

Based on the estimated channel $\hat{\mathbf{H}}$, the overall equivalent additive noise power per receive antenna is

$$\begin{aligned} \sigma_T^2 &= \frac{1}{N} \text{tr} \left\{ E \left\{ (-\Delta\mathbf{x} + \mathbf{w})(-\Delta\mathbf{x} + \mathbf{w})^H \right\} \right\} \\ &= \frac{1}{N} \text{tr} \left\{ E \left\{ \Delta\mathbf{x}\mathbf{x}^H \Delta^H \right\} \right\} + \sigma_w^2 \\ &= M\sigma_x^2\sigma_h^2 + \sigma_w^2. \end{aligned} \quad (16.4.8)$$

Substituting the result of (16.4.6) into (16.4.8), we have

$$\sigma_T^2 = \sigma_w^2 \left(\frac{M^2}{P} \sigma_x^2 + 1 \right). \quad (16.4.9)$$

This shows that the noisy channel estimation results in noise enhancement. As a result, this

causes performance degradation. Using Eq. (16.4.9) we note that one can design the power allocation for the channel estimation phase, based on the target bit error probability. For example, for the case where $P = M^2\sigma_x^2$, one should expect a doubling of the noise power, hence, a degradation of 3 dB compared to the case of perfect CSI.

16.5 Bayesian Detection without Considering Channel Uncertainty

In this section we consider both optimal and suboptimal detection schemes for MIMO symbols ignoring channel uncertainty. In this regard, we are replacing the deterministic model specified by the unknown \mathbf{H} , with the known estimate $\hat{\mathbf{H}}$. We then treat the problem as if it were deterministic, ignoring the fact that the model may be misspecified. An additional assumption that we require in order to obtain a low complexity receiver is that the distribution of the symbol vector \mathbf{x} is Gaussian. As a consequence of these assumptions, a performance degradation occurs when compared with detection utilizing the true distribution for the symbol vector and perfect CSI. In the results section, we will study an important aspect of this degradation by considering the situation with perfect CSI versus an estimated channel matrix.

First, we consider detection of the transmitted signal vector \mathbf{x} based on the received signal \mathbf{y} with the MAP criterion. The MAP detector is optimal in the sense of minimizing the average bit error probability and is given by (14)

$$\hat{\mathbf{x}}_{MAP}(\mathbf{y}) = \arg \max_{\mathbf{x} \in \mathbf{D}^M} p(\mathbf{x}|\mathbf{y}) = \arg \max_{\mathbf{x} \in \mathbf{D}^M} p(\mathbf{y}|\mathbf{x})p(\mathbf{x}), \quad (16.5.1)$$

where \mathbf{D} is the modulation alphabet. Due to the Gaussian assumption, we have that

$$\begin{aligned} p(\mathbf{y}|\mathbf{x}) &= \mathcal{N}(\hat{\mathbf{H}}\mathbf{x}, \sigma_w^2\mathbf{I}), \\ p(\mathbf{x}) &= \mathcal{N}(0, \sigma_x^2\mathbf{I}). \end{aligned} \quad (16.5.2)$$

Therefore, the MAP detector can be expressed as

$$\hat{\mathbf{x}}_{MAP}(\mathbf{y}) = \arg \max_{\mathbf{x} \in \mathbf{D}^M} \frac{1}{(\sigma_w^2)^N} \exp \left\{ -\frac{\|\mathbf{y} - \hat{\mathbf{H}}\mathbf{x}\|^2}{\sigma_w^2} \right\} \frac{1}{(\sigma_x^2)^M} \exp \left\{ -\frac{\|\mathbf{x}\|^2}{\sigma_x^2} \right\}. \quad (16.5.3)$$

Taking the log of the cost function and dropping irrelevant terms, the MAP detector is

$$MAP1 : \begin{cases} \min_{\mathbf{x}} \left\{ \frac{\|\mathbf{y} - \hat{\mathbf{H}}\mathbf{x}\|^2}{\sigma_w^2} + \frac{\|\mathbf{x}\|^2}{\sigma_x^2} \right\} \\ s.t. \mathbf{x} \in \mathbf{D}^M \end{cases}. \quad (16.5.4)$$

To find the solution for (16.5.4), which is a combinatorial problem, a brute-force searching over all of the $|\mathbf{D}|^M$ possibilities can be employed. However, it is impractical as M and $|\mathbf{D}|$ increase.

Instead, a low complexity suboptimal detector based on the MMSE criterion can be derived. This detector minimizes the mean squared error (MSE) between multiple inputs and filtered multiple outputs. Considering that the variables \mathbf{y} and \mathbf{H} are jointly Gaussian, the Bayesian MMSE detector is

$$\hat{\mathbf{x}}_{MMSE}(\mathbf{y}) = E\{\mathbf{x}|\mathbf{y}\} = \hat{\mathbf{H}}^H \left(\hat{\mathbf{H}}\hat{\mathbf{H}}^H + \frac{\sigma_w^2}{\sigma_x^2} \mathbf{I} \right)^{-1} \mathbf{y}. \quad (16.5.5)$$

By using the matrix inversion lemma, a more convenient form can be written as

$$\hat{\mathbf{x}}_{MMSE}(\mathbf{y}) = \left(\hat{\mathbf{H}}^H \hat{\mathbf{H}} + \eta_{MMSE} \mathbf{I} \right)^{-1} \hat{\mathbf{H}}^H \mathbf{y}, \quad (16.5.6)$$

where $\eta_{MMSE} = \frac{\sigma_w^2}{\sigma_x^2}$. The detected symbols are given by

$$\hat{\mathbf{x}}_{D-MMSE}(\mathbf{y}) = \text{quantize}[\hat{\mathbf{x}}_{MMSE}(\mathbf{y})], \quad (16.5.7)$$

where $\text{quantize}[a]$ is defined as the operation of rounding a to the nearest lattice point in the signal constellation.

16.6 Bayesian Detection under Channel Uncertainty

We now discuss the optimal and suboptimal detection schemes for the transmitted symbols, based on the system model in (16.4.7). In this section we consider the knowledge of the uncertainty of the channel matrix as part of our system model, and we incorporate this knowledge into the detection formulation.

First we present the optimal detection scheme and suboptimal linear receivers for this model. Then we present two suboptimal detection schemes with low complexity. The first one finds the global minimum point of the optimization problem using a simple one-dimensional line search. The second method uses the Bayesian EM to find a maximum of the posterior.

16.6.1 Optimal MAP detection

As in Section 16.5, the MAP detector is given by (16.5.1), but after considering the channel estimation error, we now have

$$\begin{aligned} p(\mathbf{y}|\mathbf{x}) &= \mathcal{N}\left(\hat{\mathbf{H}}\mathbf{x}, \left(\sigma_h^2 \|\mathbf{x}\|^2 + \sigma_w^2\right) \mathbf{I}\right) \\ p(\mathbf{x}) &= \mathcal{N}(0, \sigma_x^2 \mathbf{I}), \end{aligned} \quad (16.6.1)$$

where σ_h^2 is the variance of the channel estimation error, given in (16.4.6). Therefore, the MAP detector can be written as

$$\hat{\mathbf{x}}_{MAP}(\mathbf{y}) = \arg \max_{\mathbf{x} \in D^M} \frac{1}{(\sigma_h^2 \|\mathbf{x}\|^2 + \sigma_w^2)^N} \exp \left\{ -\frac{\|\mathbf{y} - \hat{\mathbf{H}}\mathbf{x}\|^2}{\sigma_h^2 \|\mathbf{x}\|^2 + \sigma_w^2} \right\} \frac{1}{(\sigma_x^2)^M} \exp \left\{ -\frac{\|\mathbf{x}\|^2}{\sigma_x^2} \right\} \quad (16.6.2)$$

Taking the log of the cost function and dropping irrelevant terms, the MAP detector is

$$MAP2 : \begin{cases} \min_{\mathbf{x}} & \left\{ N \log \left(\sigma_h^2 \|\mathbf{x}\|^2 + \sigma_w^2 \right) + \frac{\|\mathbf{y} - \hat{\mathbf{H}}\mathbf{x}\|^2}{\sigma_h^2 \|\mathbf{x}\|^2 + \sigma_w^2} + \frac{\|\mathbf{x}\|^2}{\sigma_x^2} \right\} \\ & s.t. \mathbf{x} \in D^M \end{cases} \quad (16.6.3)$$

The program *MAP2*, just as *MAP1*, is computationally intensive. Hence, we investigate two low complexity suboptimal detectors.

16.6.2 Linear MMSE detection

Here we present the MMSE detector of \mathbf{x} for the model in (16.4.7). This can be written as (14)

$$\begin{aligned} \hat{\mathbf{x}}_{MMSE}(\mathbf{y}) &= E \{ \mathbf{x} | \mathbf{y} \} \\ &= E \{ E \{ \mathbf{x} | \mathbf{y}, \mathbf{G} \} | \mathbf{y} \} \\ &= E \left\{ \left(\mathbf{G}^H \mathbf{G} + \frac{\sigma_w^2}{\sigma_x^2} \mathbf{I} \right)^{-1} \mathbf{G}^H \mathbf{y} \middle| \mathbf{y} \right\}. \end{aligned} \quad (16.6.4)$$

Unfortunately, it is easy to see that the computational complexity involved in solving (16.6.4) is too high for practical applications. Instead, a common approach is to consider the LMMSE estimator. The LMMSE minimizes the expected value of the squared error, using a linear estimator

$$\hat{\mathbf{x}}_{LMMSE}(\mathbf{y}) = \mathbf{A}\mathbf{y} \quad (16.6.5)$$

The LMMSE is given by (14)

$$\begin{aligned} \hat{\mathbf{x}}_{LMMSE}(\mathbf{y}) &= \mathbf{COV}(\mathbf{x}, \mathbf{y}) \mathbf{COV}(\mathbf{y}, \mathbf{y})^{-1} \mathbf{y} \\ &= \hat{\mathbf{H}}^H \left(\hat{\mathbf{H}} \hat{\mathbf{H}}^H + K \sigma_h^2 \mathbf{I} + \frac{\sigma_w^2}{\sigma_x^2} \mathbf{I} \right)^{-1} \mathbf{y} \\ &= \left(\hat{\mathbf{H}}^H \hat{\mathbf{H}} + \eta_{LMMSE} \mathbf{I} \right)^{-1} \hat{\mathbf{H}}^H \mathbf{y}, \end{aligned} \quad (16.6.6)$$

where $\eta_{LMMSE} = M \sigma_h^2 + \frac{\sigma_w^2}{\sigma_x^2}$, and σ_h^2 is defined in (16.4.6). For the discrete signal constellation, we obtain

$$\hat{\mathbf{x}}_{D-LMMSE}(\mathbf{y}) = \text{quantize} [\hat{\mathbf{x}}_{LMMSE}(\mathbf{y})] \quad (16.6.7)$$

16.6.3 Hidden Convexity Based Near-Optimal MAP Detector

We now suggest a novel near-optimal approach for the MAP detector. We suggest to relax the discrete constraint over \mathbf{x} and instead assume it stems from a continuous Gaussian distribution. Therefore, the detector can be written as

$$\widehat{\mathbf{x}}_{D-MAP}(\mathbf{y}) = \text{quantize}[\widehat{\mathbf{x}}_{C-MAP}(\mathbf{y})], \quad (16.6.8)$$

where $\widehat{\mathbf{x}}_{C-MAP}(\mathbf{y})$ is the solution to the system with a continuous Gaussian distribution input \mathbf{x} , that is

$$\widehat{\mathbf{x}}_{C-MAP}(\mathbf{y}) = \arg \min_{\mathbf{x} \in C^M} \left\{ N \log(\sigma_h^2 \|\mathbf{x}\|^2 + \sigma_w^2) + \frac{\|\mathbf{y} - \widehat{\mathbf{H}}\mathbf{x}\|^2}{\sigma_h^2 \|\mathbf{x}\|^2 + \sigma_w^2} + \frac{\|\mathbf{x}\|^2}{\sigma_x^2} \right\} \quad (16.6.9)$$

Problem (16.6.9) is an M -dimensional, nonlinear and nonconvex optimization program. In (12)-(13) the authors presented a method to transform a similar problem into a tractable form and can be solved efficiently. Under their setting, the vector \mathbf{x} was treated as an *unknown deterministic* vector. In our setting, the vector \mathbf{x} is treated as a *random Gaussian* vector. This difference results in an additional quadratic term in the MAP objective function, namely $\|\mathbf{x}\|^2/\sigma_x^2$, which incorporates the a-priori information about the random vector \mathbf{x} . The following theorem shows that the technique in (12)-(13) can also be applied in the MAP problem.

Theorem 1. For any $t \geq 0$, let

$$f(t) = \min_{\mathbf{x}: \|\mathbf{x}\|^2=t} \|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2 \quad (16.6.10)$$

and denote the optimal argument by $\mathbf{x}(t)$. Then, the MAP estimator of \mathbf{x} in the model (16.6.9) is $\mathbf{x}(t^*)$, where t^* is the solution to the following unimodal optimization problem

$$\arg \min_{t \geq 0} \left\{ \frac{f(t)}{\sigma_h^2 t + \sigma_w^2} + N \log(\sigma_h^2 t + \sigma_w^2) + \frac{t}{\sigma_x^2} \right\} \quad (16.6.11)$$

Proof. By introducing a slack variable $t = \|\mathbf{x}\|^2$, we can rewrite (16.6.9) as (16.6.11) using $f(t)$ defined in (16.6.10). In **Appendix 1**, we prove that the line search in (16.6.11) is unimodal in $t \geq 0$. \square

The change of variables in Theorem 1 allows for an efficient solution of the MAP problem. This is due to:

1. There are standard methods for evaluating $f(t)$ in (16.6.10) for any $t \geq 0$.
2. The unimodality of (16.6.11) ensures that an efficient one dimensional search can find the global optimum.

First we provide a simple method for evaluating $f(t)$ in (16.6.10). This is a quadratically constrained least squares (LS) problem whose solution can be traced back to (3).

Lemma 16.6.1. ((3), (4)): *The solution to*

$$f(t) = \min_{\mathbf{x}: \|\mathbf{x}\|^2=t} \|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2 \quad (16.6.12)$$

is

$$\mathbf{x}(t) = (\mathbf{H}^H \mathbf{H} + \eta \mathbf{I})^\dagger \mathbf{H}^H \mathbf{y}, \quad (16.6.13)$$

where $\eta \geq -\lambda_{\min}(\mathbf{H}^H \mathbf{H})$ is the unique root of the equation

$$\|\mathbf{x}(t)\|^2 = t. \quad (16.6.14)$$

The only issue is the evaluation of η . This can be done by trying different values of η in (16.6.13) until one that satisfies (16.6.14) is found. This is made simpler due to the monotonicity of $\|(\mathbf{H}^H \mathbf{H} + \eta \mathbf{I})^\dagger \mathbf{H}^H \mathbf{y}\|^2$ in η which enables us to find a value of η that satisfies (16.6.14) using a simple line-search, such as bi-section (6). The search range is $-\lambda_{\min}(\mathbf{H}^H \mathbf{H}) \leq \eta \leq \eta_{max}$, where η_{max} is some sufficiently large upper bound. Next, $f(t)$ can be evaluated by plugging the appropriate $\mathbf{x}(t)$ into $\|\mathbf{y} - \mathbf{H}\mathbf{x}(t)\|^2$. This algorithm is presented in **Algorithm 2**.

Now that we have an efficient method for evaluating $f(t)$, it remains to solve (16.6.11). The unimodality property ensures that this line search can be efficiently implemented using the Golden Section search (6). Theoretically, the search region is defined to be over $0 \leq t \leq \infty$. However, in practice, the search can be confined to $0 \leq t \leq t_{max}$ where t_{max} is a sufficiently large upper bound. This algorithm is presented in **Algorithm 3**.

In the limit of an infinite number of amplitudes in the signal constellation, $\hat{\mathbf{x}}_{D-MAP}$ in (16.6.9) is effectively equal to $\hat{\mathbf{x}}_{MAP}$, and it is optimal. In that case, the detection problem, generally considered to be exponentially complex, can be solved with linear complexity.

16.6.4 Near-Optimal MAP Detector using Bayesian EM

In this section we provide an alternative solution to the one suggested in 16.6.3 using a Bayesian EM (BEM) algorithm. We begin with a short overview of BEM methods. The BEM algorithm is a general technique for finding MAP estimates where the model depends on unobserved latent variables. This method is based on the classical EM algorithm (21), and is known to converge to a stationary point of the posterior density corresponding to a mode, though convergence to the global mode is not guaranteed.

The EM algorithm consists of two major steps: an expectation step, followed by a maximization step. The expectation is with respect to the unknown underlying variables, using the current estimate of the parameters and conditioned upon the observations. During the maximization

step one maximizes the complete data likelihood using the expectations of the previous step. The algorithm is numerically stable and the convergence rate is reasonably fast. However, if the likelihood contains several modes it may be trapped in local extrema. In practice one can attempt to avoid this by several approaches such as careful selection of initial values, or by repeating the procedure several times with different initial values, selected randomly or deterministically under a heuristic criterion.

This missing data problem is overcome by considering the hidden variables as being random variables and averaging over their distribution. BEM is best summarized with the following three steps:

1. Initial guess of the parameters
2. Replace the missing values by their expectations given the guessed parameters
3. Estimate parameters

Steps (2) and (3) are repeated until convergence. We now provide a solution for finding the MAP estimator of \mathbf{x} using the BEM method. For this purpose we rewrite (16.4.7) as

$$\mathbf{y} = \mathbf{G}\mathbf{x} + \mathbf{w}, \quad (16.6.15)$$

where \mathbf{G} is a Gaussian random matrix with mean $\hat{\mathbf{H}}$ and variance of each element σ_h^2 . At each BEM iteration, the algorithm maximizes the expected log likelihood (with respect to \mathbf{G} , given \mathbf{y} and parameterized by \mathbf{x})

$$\begin{aligned} \mathbf{x}_{n+1} &= \arg \max_{\mathbf{x}} E_{\mathbf{G}|\mathbf{y};\mathbf{x}_n} \{ \log p(\mathbf{y}, \mathbf{G}, \mathbf{x}) \} \\ &= \arg \min_{\mathbf{x}} E_{\mathbf{G}|\mathbf{y};\mathbf{x}_n} \left\{ \frac{1}{\sigma_w^2} \|\mathbf{y} - \mathbf{G}\mathbf{x}\|^2 + \frac{1}{\sigma_x^2} \|\mathbf{x}\|^2 \right\} \\ &= \arg \min_{\mathbf{x}} \left\{ \frac{1}{\sigma_w^2} (-2\mathbf{y}\mathbf{x}^H E_{\mathbf{G}|\mathbf{y};\mathbf{x}_n} \{ \mathbf{G}^H \}) + \frac{1}{\sigma_w^2} \mathbf{x}^H E_{\mathbf{G}|\mathbf{y};\mathbf{x}_n} \{ \mathbf{G}^H \mathbf{G} \} \mathbf{x} + \frac{1}{\sigma_x^2} \mathbf{x}^H \mathbf{x} \right\} \\ &= \arg \min_{\mathbf{x}} \left\{ \frac{1}{\sigma_w^2} (-2\mathbf{y}\mathbf{x}^H \Phi_1(\mathbf{y}, \mathbf{x}_n) + \mathbf{x}^H \Phi_2(\mathbf{y}, \mathbf{x}_n) \mathbf{x}) + \frac{1}{\sigma_x^2} \mathbf{x}^H \mathbf{x} \right\} \end{aligned} \quad (16.6.16)$$

where

$$\begin{aligned} \Phi_1(\mathbf{y}, \mathbf{x}_n) &= E_{\mathbf{G}|\mathbf{y};\mathbf{x}_n} \{ \mathbf{G} \} \\ \Phi_2(\mathbf{y}, \mathbf{x}_n) &= E_{\mathbf{G}|\mathbf{y};\mathbf{x}_n} \{ \mathbf{G}^H \mathbf{G} \} \end{aligned} \quad (16.6.17)$$

Next, we take the derivative of (16.6.16) with respect to \mathbf{x} , setting it to 0, to obtain the following recursion:

$$\mathbf{x}_{n+1} = \left(\Phi_2(\mathbf{y}, \mathbf{x}_n) + \frac{\sigma_w^2}{\sigma_x^2} I \right)^{-1} \Phi_1(\mathbf{y}, \mathbf{x}_n)^T \mathbf{y} \quad (16.6.18)$$

The expectations in (16.6.17) can be evaluated based on the jointly Gaussian optimal MMSE estimation theory (14). To demonstrate this, first using the Kronecker product, we rewrite (16.6.15)

as

$$\mathbf{y} = (\mathbf{x}^T \otimes \mathbf{I}) \mathbf{g} + \mathbf{w}, \quad (16.6.19)$$

where $\mathbf{g} = \text{vec}(\mathbf{G})$. The Bayesian MMSE of (16.6.19) becomes

$$E_{\mathbf{g}|\mathbf{y};\mathbf{x}_n} \{\mathbf{g}\} = \text{vec}(\mathbf{H}) + \frac{(\mathbf{x}^T \otimes \mathbf{I})(\mathbf{y} - \mathbf{H}\mathbf{x}_n)}{\|\mathbf{x}_n\|^2 + \frac{\sigma_w^2}{\sigma_h^2}}, \quad (16.6.20)$$

$$\text{COV}_{\mathbf{g}|\mathbf{y};\mathbf{x}_n} \{\mathbf{g}\} = \sigma_h^2 \mathbf{I} - \frac{\sigma_g^4 (\mathbf{x}_n \mathbf{x}_n^T \otimes \mathbf{I})}{\sigma_h^2 \|\mathbf{x}_n\|^2 + \sigma_w^2}. \quad (16.6.21)$$

Now using simple algebraic manipulations, we can find Φ_1 and Φ_2 explicitly as

$$\Phi_1(\mathbf{y}, \mathbf{x}_n) = \mathbf{H} + \frac{1}{\|\mathbf{x}_n\|^2 + \frac{\sigma_w^2}{\sigma_h^2}} (\mathbf{y} - \mathbf{H}\mathbf{x}_n) \mathbf{x}_n^T \quad (16.6.22)$$

$$\Phi_2(\mathbf{y}, \mathbf{x}_n) = \Phi_1(\mathbf{y}, \mathbf{x}_n)^T \Phi_1(\mathbf{y}, \mathbf{x}_n) + \sigma_h^2 N \left(I - \frac{\sigma_h^2 \mathbf{x}_n \mathbf{x}_n^T}{\|\mathbf{x}_n\|^2 + \frac{\sigma_w^2}{\sigma_h^2}} \right) \quad (16.6.23)$$

Initial guess of x_0 : Since our problem is non-convex, convergence to the optimal solution is not guaranteed and it is well known that the initial starting point \mathbf{x}_0 in such settings will influence whether the solution converges to the optimal solution. As a starting point we suggest to take $\hat{\mathbf{x}}_{LMMSE}$ (Eq. (16.6.6)) as the initial guess, and therefore $\mathbf{x}_0 = \hat{\mathbf{x}}_{LMMSE}$. The BEM algorithm is summarized in **Algorithm 4**.

16.6.5 Hidden Convexity Vs. BEM approach

In Sections 16.6.3 and 16.6.4 we developed two detection schemes for the underlined system model in (16.4.7). While the Hidden Convexity based approach is optimal in finding the MAP estimate of the continuous Gaussian, the BEM is suboptimal in this sense and depends on its starting point. In Section 16.7 we compare these approaches in terms of BER.

16.7 Simulation Results

In this section, we compare the performance of the proposed detectors via simulation. First, we describe the simulated MIMO system.

16.7.1 System Configuration

In the simulations performed in this paper, the elements of the channel matrix \mathbf{H} are assumed independent and identically distributed complex Gaussian random variables with zero-mean and unit-variance. We also assume that the channel coefficients remain constant within each

frame and change independently from frame to frame. The additive channel noise is spatially and temporally white complex Gaussian with zero-mean, and the variance σ_w^2 . The average power of the constellation is E_s , and the SNR is defined as $\frac{ME_s}{\sigma_w^2}$. Different numbers of transmit and receive antennas have been simulated. We use frame-by-frame transmission, with each frame consisting of $L = 128$ MIMO symbols. At the beginning of each frame 4 equipower orthogonal pilot symbols were dedicated for channel estimation phase, according to Section 16.4. The energy of each pilot vector is $4E_s$.

16.7.2 Constellation Design

In this example we will illustrate the detection performance on a Gaussian like distribution. A popular choice in the literature (15) to approximate a Gaussian in this setting is the Maxwell-Boltzman (M-B) distribution, motivated in (16) and (17). The M-B distribution (7) gives symbol probabilities $P(x_j)$, $j = 1, \dots, |\mathbf{D}|$ which are obtained using

$$P(x_j) = K(\lambda) \exp\{-\lambda |x_j|^2\}, \quad \lambda \geq 0, \quad (16.7.1)$$

with a normalizing constant,

$$K(\lambda) = \left(\sum_{x_j} \exp\{-\lambda |x_j|^2\} \right)^{-1}. \quad (16.7.2)$$

The parameter λ governs the trade-off between average power of the signal points and the entropy rate. For $\lambda = 0$ a uniform distribution results, whereas for $\lambda \rightarrow \infty$ only the two minimum energy signal points are used.

In this example we illustrate the use of an MB distribution in a 16-ary PAM constellation, using the signal set $\mathbf{D} = \{-15, -13, -11, -9, -7, -5, -3, -1, 1, 3, 5, 7, 9, 11, 13, 15\}$ and $\lambda = 1/40$. The 16-PAM distribution is depicted in Fig. 16.8.1. To analyze the goodness of fit between the M-B discrete constellation with the true continuous Gaussian, we plot the cumulative distribution function (CDF) of the M-B distribution in Eqs. (16.7.1-16.7.2) versus the CDF of a continuous Gaussian with zero-mean and variance, $\sigma^2 = \sum_{k=1}^{16} P(x_k) |x_k|^2$. The results are depicted in the bottom plot of Fig. 16.8.1.

16.7.3 Comparison of Detection Techniques

We now compare the BER results for the different detectors. We compare the performance of the MMSE with perfect CSI, referred to as MMSE_CSI with four other techniques. The first one is the MMSE detector without concern for channel uncertainty (Eq. 16.5.7) and referred to as MMSE_NCU. The second is the LMMSE (Eq. 16.6.7), which takes channel uncertainty into account, and referred to as LMMSE_CU. The third one is the proposed MAP detector (Eq. 16.6.8), referred to as MAP_CU and the fourth one is the BEM based detector, referred

to as BEM_CU. The BER results for the above mentioned methods for $(N=4, M=4)$ and $(N=8, M=8)$ are depicted in Figs. 16.8.2 and 16.8.3, respectively. As the figures clearly depict, there is a significant degradation due to channel estimation error. These results are inline with Eq. (16.4.9) and demonstrate the importance of a good channel estimate. In both MIMO settings the LMMSE_CU detector is about 2 dB better than the MMSE_NCU detector. The proposed MAP detector, MAP_CU produces almost 2 dB gain over the LMMSE_CU detection for $(N=4, M=4)$ system and above 2 dB for $(N=8, M=8)$. We also notice that the MAP_CU and the BEM_CU detectors provide comparable results.

16.7.4 Affect of Training

Next we investigate how much training is required in order to make the channel estimation error have a negligible effect on the BER. We compare our MAP detector with the MMSE with perfect CSI for different SNRs and increased levels of power dedicated for the training phase. The different values for the training power were $P_T = \{4E_s, 16E_s, 64E_s\}$ (four orthogonal pilots, each with a power of P_T were used for training). The results for the case of $(N=4, M=4)$ are depicted in Fig. 16.8.4. The results show that for the $(N=4, M=4)$ case, a training power of $P = 64E_s$ is required to produce negligible BER degradation after which there is no need for further improvement of the training. The cost of increasing the power of training symbols is that additional power is not used directly in sending actual information. We define the percentage of power dedicated for training as the ratio between the training power and the overall transmitted power per frame (training + data)

$$R = \frac{P_T}{M * L * E_s + P_T} * 100. \quad (16.7.3)$$

For the case of $N = 4, M = 4, P_T = 4E_s$ we get that $R = 0.77\%$, where as if we try and remove channel estimation influence by using $P_T = 64E_s$ we get that $R = 11.11\%$.

16.8 Conclusions

In this paper we discussed the detection in MIMO systems of near-Gaussian constellations under imperfect channel estimation. First we reviewed the maximum likelihood channel estimation using optimal pilot sequence and quantified the degradation in performance due to noisy channel estimate, we showed that it can be interpreted as noise enhancement.

Next, we discussed a few suboptimal detection schemes for detection under channel uncertainty. We derived the MAP estimator and provided an efficient method for finding it by transforming the multi-dimensional, nonlinear and nonconvex problem into a simple tractable form. We proposed a detection scheme for near-Gaussian-digitally modulated symbols with linear complexity, based on the new MAP estimator. Simulation results show the improved perfor-

mance offered by our new approach in comparison to the standard LMMSE methods in terms of BER.

Appendix

In this Appendix, we show that (16.6.11) is unimodal in $t \geq 0$. First, we will show that $f(t)$ is convex in $t \geq 0$. In (4) and (7), it was shown that strong duality holds in this special case and that is equal to the value of its dual program

$$f(t) = \begin{cases} \max_{\alpha} & \mathbf{y}^H \mathbf{y} - \mathbf{y}^H \mathbf{H} (\mathbf{H}^H \mathbf{H} + \eta \mathbf{I})^\dagger \mathbf{H}^H \mathbf{y} - \eta t \\ \text{s.t.} & \mathbf{H}^H \mathbf{H} + \eta \mathbf{I} \succeq 0 \\ & \mathbf{H}^H \mathbf{y} \in \mathbb{R} (\mathbf{H}^H \mathbf{H} + \eta \mathbf{I}) \end{cases} \quad (16.8.1)$$

Thus, $f(t)$ is the pointwise maximum of a family of affine functions and therefore is convex in $t \geq 0$. Next, we will show that

$$r(t) = \frac{f(t)}{\sigma_h^2 t + \sigma_w^2} + N \log(\sigma_h^2 t + \sigma_w^2) + \frac{t}{\sigma_x^2} \quad (16.8.2)$$

is unimodal in $t \geq 0$. We use the following result from (7): If $r'(t) = 0$ implies $r''(t) > 0$ for any $t \geq 0$, then $r(t)$ is unimodal in $t \geq 0$.

The condition $r'(t) = 0$ states that

$$r'(t) = \frac{f'(t)}{\sigma_h^2 t + \sigma_w^2} - \frac{f(t)\sigma_h^2}{(\sigma_h^2 t + \sigma_w^2)^2} + \frac{N\sigma_h^2}{\sigma_h^2 t + \sigma_w^2} + \frac{1}{\sigma_x^2} = 0. \quad (16.8.3)$$

Multiplying by $\frac{\sigma_h^2}{\sigma_h^2 t + \sigma_w^2}$ and rearranging yields

$$\frac{f(t)\sigma_g^4}{(\sigma_h^2 t + \sigma_w^2)^3} = \frac{N\sigma_g^4}{(\sigma_h^2 t + \sigma_w^2)^2} + \frac{\sigma_h^2 f'(t)}{(\sigma_h^2 t + \sigma_w^2)^2} + \frac{\sigma_h^2}{\sigma_x^2 (\sigma_h^2 t + \sigma_w^2)}. \quad (16.8.4)$$

The second derivative is

$$r''(t) = \frac{f''(t)}{\sigma_h^2 t + \sigma_w^2} - \frac{f'(t)\sigma_h^2}{(\sigma_h^2 t + \sigma_w^2)^2} - \frac{f(t)\sigma_h^2}{(\sigma_h^2 t + \sigma_w^2)^2} + \frac{2f(t)\sigma_g^4}{(\sigma_h^2 t + \sigma_w^2)^3} - \frac{N\sigma_g^4}{(\sigma_h^2 t + \sigma_w^2)^2}. \quad (16.8.5)$$

Plugging in (16.8.3) results in

$$r''(t) = \frac{f''(t)}{\sigma_h^2 t + \sigma_w^2} + \frac{2\sigma_h^2}{\sigma_x^2 (\sigma_h^2 t + \sigma_w^2)} + \frac{N\sigma_g^4}{(\sigma_h^2 t + \sigma_w^2)^2}. \quad (16.8.6)$$

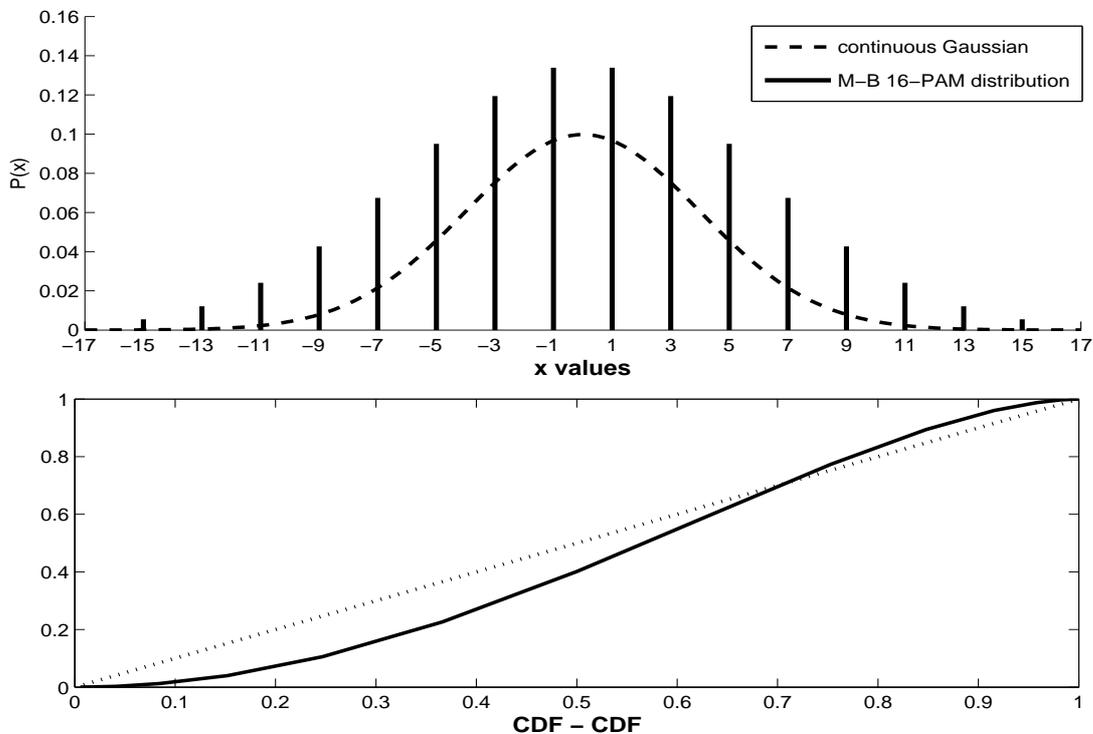
Now, $f(t)$ is convex, which means that $f''(t) \geq 0$. Therefore, the first term of (16.8.6) is non-negative. The second and third terms are positive since $\sigma_x^2 \geq 0$, $\sigma_h^2 \geq 0$ and $\sigma_w^2 > 0$. This concludes the proof.

Algorithm 2 Constrained Least Squares (*Lemma 16.6.1*)**Input:** $t, \mathbf{H}, \mathbf{y}, \lambda_{\min}(\mathbf{H}^H \mathbf{H}), \eta_{\max}$ **Output:** $f(t), \mathbf{x}$

- 1: $\eta_L = -\lambda_{\min}(\mathbf{H}^H \mathbf{H})$
- 2: $\eta_R = \eta_{\max}$
- 3: **repeat**
- 4: $\eta_M = \frac{\eta_L + \eta_R}{2}$
- 5: $\mathbf{x} = (\mathbf{H}^H \mathbf{H} + \eta_M \mathbf{I})^{-1} \mathbf{H}^H \mathbf{y}$
- 6: $\epsilon = \mathbf{x}^H \mathbf{x} - t$
- 7: **if** $\epsilon > 0$ **then**
- 8: $\eta_L = \eta_M$
- 9: **else**
- 10: $\eta_R = \eta_M$
- 11: **end if**
- 12: **until** $|\epsilon| \leq \epsilon_{\min}$
- 13: $f(t) = \|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2$

Algorithm 3 MAP Estimation - Solution of Eq. (16.6.11)**Input:** $\mathbf{y}, \mathbf{H}, \sigma_x^2, \sigma_w^2, \sigma_h^2, N, t_{\max}$ **Output:** \mathbf{x}

- 1: $t_L = 0$
- 2: $t_R = t_{\max}$
- 3: $\rho = \frac{(\sqrt{5}-1)^2}{4}$
- 4: **repeat**
- 5: $\Delta = t_R - t_L$
- 6: $t_A = t_L + \rho\Delta$
- 7: $t_B = t_R - \rho\Delta$
- 8: $r(t_A) = \frac{f(t_A)}{\sigma_h^2 t_A + \sigma_w^2} + N \log(\sigma_h^2 t_A + \sigma_w^2) + \frac{t_A}{\sigma_x^2}$
- 9: $r(t_B) = \frac{f(t_B)}{\sigma_h^2 t_B + \sigma_w^2} + N \log(\sigma_h^2 t_B + \sigma_w^2) + \frac{t_B}{\sigma_x^2}$
- 10: **if** $r(t_A) < r(t_B)$ **then**
- 11: $t_R = t_B$
- 12: **else**
- 13: $t_L = t_A$
- 14: **end if**
- 15: **until** $|r(t_A) - r(t_B)| < \epsilon$

Algorithm 4 BEM detector**Input:** \mathbf{H}, \mathbf{y} **Output:** $\hat{\mathbf{x}}_{BEM}$ 1: $\mathbf{x}_0 = \hat{\mathbf{x}}_{LMMSE}$ (Eq. (16.6.6))2: **repeat**3: $\Phi_1(\mathbf{y}, \mathbf{x}_n) = \mathbf{H} + \frac{1}{\|\mathbf{x}_n\|^2 + \frac{\sigma_w^2}{\sigma_h^2}} (\mathbf{y} - \mathbf{H}\mathbf{x}_n) \mathbf{x}_n^T$ 4: $\Phi_2(\mathbf{y}, \mathbf{x}_n) = \Phi_1(\mathbf{y}, \mathbf{x}_n)^T \Phi_1(\mathbf{y}, \mathbf{x}_n) + \sigma_h^2 N \left(I - \frac{\sigma_h^2 \mathbf{x}_n \mathbf{x}_n^T}{\|\mathbf{x}_n\|^2 + \frac{\sigma_w^2}{\sigma_h^2}} \right)$ 5: $\mathbf{x}_{n+1} = \left(\Phi_2(\mathbf{y}, \mathbf{x}_n) + \frac{\sigma_w^2}{\sigma_x^2} I \right)^{-1} \Phi_1(\mathbf{y}, \mathbf{x}_n)^T \mathbf{y}$ 6: **until** stopping criterion is met7: $\hat{\mathbf{x}}_{BEM} = \text{quantize}[\mathbf{x}_{n+1}]$ Fig. 16.8.1: Near-Gaussian constellation of 16-PAM with $\lambda = 1/40$

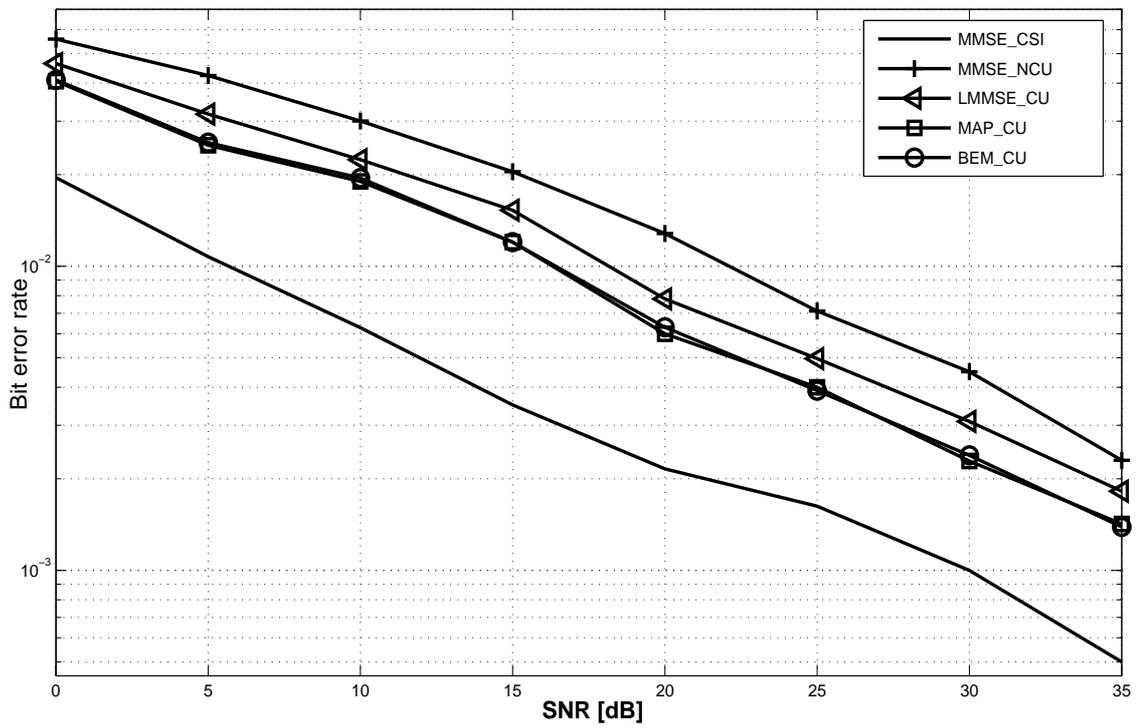


Fig. 16.8.2: Bit error rate of MIMO systems with $M = N = 4$

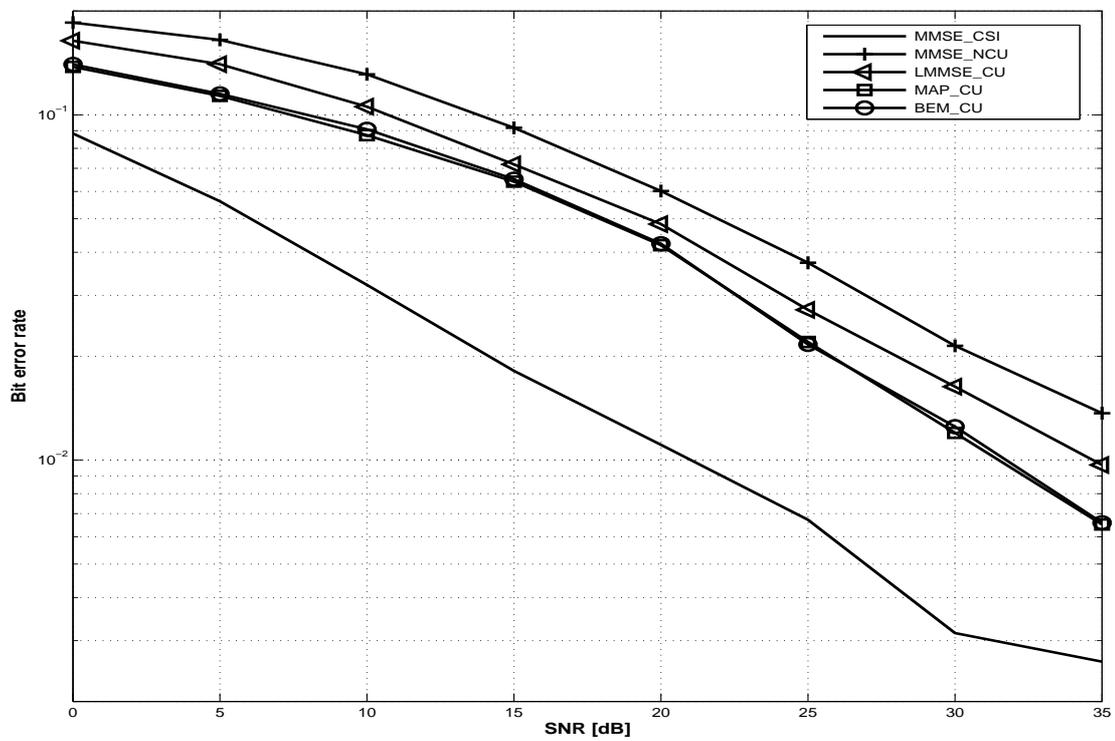


Fig. 16.8.3: Bit error rate of MIMO systems with $M = N = 8$

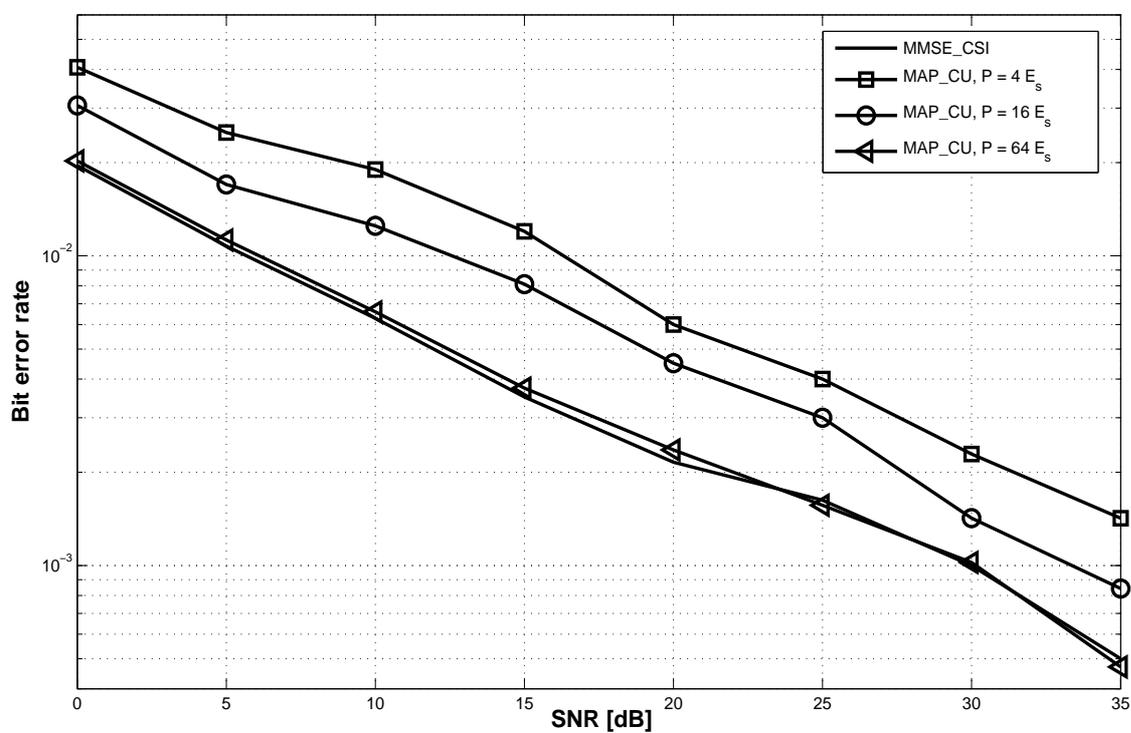


Fig. 16.8.4: Effect of training power for various SNRs for a MIMO system with $M = 4, N = 4$

References

- [1] G. J. Foschini and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Pers. Commun.*, vol. 6, pp. 311-335, Mar. 1998.
- [2] M. Biguesh and A. B. Gershman, "Training-Based MIMO Channel Estimation: A Study of Estimator Tradeoffs and Optimal Training Signals," *IEEE Transactions on Signal Processing*, vol. 54, pp. 884-893, March 2006.
- [3] G. D. Forney and G. Ungerboeck, "Modulation and coding for linear Gaussian channel," *IEEE Transactions on Information Theory*, vol. 44, pp. 2384-2415, Oct. 1998.
- [4] U. Wachsman, R. F. H. Fischer, and J. Huber, "Multilevel codes: Theoretical concepts and practical design rules," *IEEE Transactions on Information Theory*, vol. 45, no. 5, pp. 1361-1391, July 1999.
- [5] G. D. Forney, "Trellis shaping," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 281-300, March 1992.
- [6] P. Fortier, A. Ruiz, and J. M. Cioffi, "Multidimensional signal sets through the shell construction for parallel channels," *IEEE Transactions on Communication*, vol. 40, no. 500-512, pp. 2384-2415, March 1992.
- [7] F. R. Kschischang and S. Pasupathy, "Optimal nonuniform signaling for Gaussian channels," *IEEE Transactions on Information Theory*, vol. 39, pp. 913-929, May 1993.
- [8] E. Agrell, T. Eriksson, A. Vardy, and K. Zeger, "Closest point search in lattices," *IEEE Transactions on Information Theory*, vol. 48, no. 8, pp. 2201-2214, August 2002.
- [9] G. J. Foschini, "Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas," *Bell Labs. Tech. J.*, vol. 1, no. 2, pp. 41-59, 1996.
- [10] A. J. Paulraj, D. A. Gore, R. U. Nabar, and H. Bölcskei, "An Overview of MIMO Communications: A Key to Gigabit Wireless," *Proc. IEEE*, vol. 92, no. 2, pp. 198-218, February 2004.
- [11] I. Neve, A. Wiesel, J. Yuan, and Y. C. Eldar, "Maximum A-Posteriori estimation in linear

models with a gaussian model matrix," *Proc. of conference on Information Sciences and Systems, Baltimore, (CISS-2007)*, March 2007.

- [12] A. Wiesel, Y. C. Eldar, and A. Beck, "Maximum likelihood estimation in linear models with a gaussian model matrix," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 292-295, May 2006.
- [13] A. Wiesel and Y. C. Eldar, "Maximum likelihood estimation in random linear models \mathcal{U} generalizations and performance analysis," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP-2006)*, May 2006.
- [14] S. M. Kay, *Fundamentals of Statistical Signal Processing - Detection Theory*, Prentice Hall, 1998.
- [15] D. Raphaeli and A. Gurevitz, "Constellation shaping for pragmatic turbo-coded modulation with high spectral efficiency," *IEEE Transactions on Communications*, vol. 52, pp. 341-345, 2004
- [16] F. W. Sun and H. C. A. van Tilborg, "Approaching capacity by equiprobable signaling on the gaussian channel," *IEEE Transactions on Information Theory*, vol.39, no. 5, pp. 1714-1716, September 1993.
- [17] A. R. Calderbank and L. H. Ozarow, "Nonequiprobable signaling on the gaussian channel," *IEEE Transactions on Information Theory*, vol. 36, no. 4, pp. 726-740, July 1990.
- [18] G. E. Forsythe and G. H. Golub, "On the stationary values of a second-degree polynomial on the unit sphere," *Journal of the Society for Industrial and Applied Mathematics*, vol. 13, no. 4, pp. 1050-1068, 1965.
- [19] Y. C. Eldar and A. Beck, "Hidden convexity based near maximum-likelihood CDMA detection," in *Proc. of conference on Information Sciences and Systems, Princeton, (CISS-2005)*, March 2005.
- [20] W. H. Press, Saul A. Teukolsky, W. T. Vetterling, and Brian P. Flannery, *Numerical recipes in C*, Cambridge University Press, Cambridge, second edition, 1992, The art of scientific computing.
- [21] A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, in *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1-38, December 1977.
- [22] S. Boyd and L. Vandenberghe, *Introduction to Convex Optimization with Engineering Applications*, Stanford, 2003.

Journal Paper 13

"The best way to have a good idea is to have a lot of ideas."

Linus Pauling

Nevat I., Peters, G.W. and Yuan J. (2008) "A Low Complexity MAP Estimation in Linear Models with a Random Gaussian Mixing Matrix". *IEEE Transactions on Communications, to appear.*

This work was instigated by Ido Nevat. The second author can claim 50% of the credit for the contents. His work included jointly developing the methodology contained and the applications, the implementation and comparison to alternative approaches, writing the drafts of the paper. This work will be included in a PhD thesis of a co-author, though this thesis is not submitted by sequence of publications. This paper is under revision, several peer reviewed conference papers relating to aspects of this work have been accepted and appeared in conference proceedings. Permission from all the co-authors is granted for inclusion in the PhD.

This paper appears in the thesis in a modified format to that which was submitted to the IEEE Wireless Communications Letters

Final print version will be available at:

<http://ieeexplore.ieee.org>

A Low Complexity MAP Estimation in Linear Models With a Random Gaussian Mixing Matrix

Ido Nevat (*corresponding author*)

School of Electrical Engineering and Telecommunications, University of NSW, Sydney, Australia.

Gareth W. Peters

CSIRO Mathematical and Information Sciences, Locked Bag 17, North Ryde, NSW, 1670, Australia;

UNSW Mathematics and Statistics Department, Sydney, 2052, Australia;

email: peterga@maths.unsw.edu.au

Jinhong Yuan

School of Electrical Engineering and Telecommunications, University of NSW, Sydney, Australia.

17.1 Abstract

We consider the formulation of a Bayesian inference approach for a model involving a random Gaussian vector in a linear model containing a random Gaussian matrix for which only the first and second moments are known. We propose an efficient method to finding the MAP estimator for this model and analyze its complexity. The performance in terms of estimation error is evaluated by simulation, which show it's superior to LMMSE.

17.2 Problem formulation

Consider the problem of estimating a random vector \mathbf{x} in the linear model

$$\mathbf{y} = \mathbf{G}\mathbf{x} + \mathbf{w}, \quad (17.2.1)$$

where \mathbf{x} is distributed according to a zero-mean Gaussian vector with independent elements of variance σ_x^2 and \mathbf{w} is a zero-mean Gaussian vector with independent elements of variance σ_w^2 . The matrix \mathbf{G} is $N \times K$ and can be decomposed as

$$\mathbf{G} = \mathbf{H} + \mathbf{\Sigma}, \quad (17.2.2)$$

where \mathbf{H} is a deterministic and known matrix, and $\mathbf{\Sigma}$ is a matrix with i.i.d random Gaussian zero-mean elements and the variance of each element σ_g^2 . In addition, \mathbf{x} , \mathbf{G} and \mathbf{w} are assumed statistically independent. Three standard methods for estimating \mathbf{x} under a Bayesian framework are the minimum mean squared error (MMSE), the linear minimum mean squared error (LMMSE) and the maximum a-posteriori (MAP) estimators. The first two approaches are based on a quadratic cost function whereas the third minimizes a hit-or-miss risk. From a detection point of view, the MAP method is also related to the minimum error probability criterion.

The MMSE estimator of model (17.2.1) is (2)

$$\begin{aligned} \hat{\mathbf{x}}_{MMSE}(\mathbf{y}) &= E\{\mathbf{x}|\mathbf{y}\} \\ &= E\left\{\left(\mathbf{G}^T\mathbf{G} + \frac{\sigma_w^2}{\sigma_x^2}\mathbf{I}\right)^{-1}\mathbf{G}^T\mathbf{y}\middle|\mathbf{y}\right\}. \end{aligned} \quad (17.2.3)$$

Unfortunately, the computational complexity involved in solving (17.2.3) is too high for practical applications. Instead, a common approach is to consider the LMMSE estimator which satisfies the following closed form solution

$$\begin{aligned} \hat{\mathbf{x}}_{LMMSE}(\mathbf{y}) &= E\{\mathbf{x}\mathbf{y}^T\}E^{-1}\{\mathbf{y}\mathbf{y}^T\}\mathbf{y} \\ &= \mathbf{H}^T\left(\mathbf{H}\mathbf{H}^T + K\sigma_g^2\mathbf{I} + \frac{\sigma_w^2}{\sigma_x^2}\mathbf{I}\right)^{-1}\mathbf{y}, \end{aligned} \quad (17.2.4)$$

The MAP estimator of model (17.2.1) is given by

$$\begin{aligned} \hat{\mathbf{x}}_{MAP}(\mathbf{y}) &= \arg\max_{\mathbf{x}}\{\log p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})\} \\ &= \min_{\mathbf{x}}\left\{\frac{\|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2}{\sigma_g^2\|\mathbf{x}\|^2 + \sigma_w^2} + N\log(\sigma_g^2\|\mathbf{x}\|^2 + \sigma_w^2) + \frac{\|\mathbf{x}\|^2}{\sigma_x^2}\right\}. \end{aligned} \quad (17.2.5)$$

Solving (17.2.5) involves a K-dimensional, nonlinear and nonconvex optimization program. The focus of this letter will be to provide an efficient method for solving (17.2.5) and then to analyze its algorithmic complexity.

17.3 MAP Estimation using Hidden Convexity Optimization

In this Section we begin by reviewing the solution of (17.2.5). Next we provide a low complexity method of solving this problem. Using a simple change of variables $\|\mathbf{x}\|^2 = t$, the MAP estimator of \mathbf{x} in the model (17.2.5) is $\mathbf{x}(t^*)$, where t^* is the solution to the following unimodal optimization problem (1):

$$\arg \min_{t \geq 0} \left\{ \frac{f(t)}{\sigma_g^2 t + \sigma_w^2} + N \log(\sigma_g^2 t + \sigma_w^2) + \frac{t}{\sigma_x^2} \right\}, \quad (17.3.1)$$

where

$$f(t) = \begin{cases} \min_{\mathbf{x}} \|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2 \\ \text{s.t. } \|\mathbf{x}\|^2 = t. \end{cases} \quad (17.3.2)$$

The solution to (17.3.1) requires two nested line searches. We have an outer minimization with respect to t and for each fixed t we need to solve (17.3.2). In the rest of this section, we discuss an efficient implementation of (17.3.2). Problem (17.3.2) is a quadratic equality constrained least squares and its solution,

$$\mathbf{x}(t) = (\mathbf{H}^T \mathbf{H} + \eta \mathbf{I})^{-1} \mathbf{H}^T \mathbf{y}, \quad (17.3.3)$$

is given by (3), (4), where $\eta \geq -\lambda_{\min}(\mathbf{H}^T \mathbf{H})$ is the unique root of the equation

$$\|\mathbf{x}(t)\|^2 = t, \quad (17.3.4)$$

and $\lambda_{\min}(\mathbf{H}^T \mathbf{H})$ is the smallest eigenvalue of $\mathbf{H}^T \mathbf{H}$. The only issue is the evaluation of η . This can be done by trying different values of η in (17.3.3) until one that satisfies (17.3.4) is obtained. In practice this is straightforward since $\left\| (\mathbf{H}^T \mathbf{H} + \eta \mathbf{I})^{-1} \mathbf{H}^T \mathbf{y} \right\|^2$ is monotonic decreasing in η which enables us to find a value of η that satisfies (17.3.4) using a simple line-search, such as bi-section (6). The search range is $\eta_{\min} \leq \eta \leq \eta_{\max}$, where $\eta_{\min} = -\lambda_{\min}(\mathbf{H}^T \mathbf{H})$ and η_{\max} is some sufficiently large upper bound. Next, $f(t)$ can be evaluated by plugging the appropriate $\mathbf{x}(t)$ into $\|\mathbf{y} - \mathbf{H}\mathbf{x}(t)\|^2$. However, this procedure may be computationally intensive due to the matrix inversion required for each trial of η . Therefore we suggest an efficient implementation of $(\mathbf{H}^T \mathbf{H} + \eta \mathbf{I})^{-1} \mathbf{H}^T$ in (17.3.3):

Lemma 17.3.1.

$$(\mathbf{H}^T \mathbf{H} + \eta \mathbf{I})^{-1} \mathbf{H}^T = \mathbf{V} \Lambda_\eta \mathbf{U}^T, \quad (17.3.5)$$

where $[\mathbf{U}, \mathbf{S}, \mathbf{V}] = \text{SVD}(\mathbf{H})$, is the Singular Value Decomposition (SVD) of \mathbf{H} , \mathbf{U} and \mathbf{V} are unitary matrices and \mathbf{S} is a diagonal matrix containing the singular values λ_k , $k=1, \dots, K$. The matrix Λ_η of

dimensions $K \times N$ is defined by

$$\Lambda_\eta = \begin{bmatrix} \frac{\lambda_1}{(\lambda_1)^2 + \eta} & 0 & \dots & 0 & \dots & 0 \\ 0 & \frac{\lambda_2}{(\lambda_2)^2 + \eta} & \dots & 0 & \dots & 0 \\ \vdots & & & \vdots & \dots & 0 \\ 0 & 0 & \dots & \frac{\lambda_K}{(\lambda_K)^2 + \eta} & \dots & 0 \end{bmatrix}, \quad (17.3.6)$$

Proof. Consider the L.H.S. expression $(\mathbf{H}^T \mathbf{H} + \eta \mathbf{I})^{-1} \mathbf{H}^T$ and substitute the SVD of \mathbf{H} given by

$$\mathbf{H} = \mathbf{U} \mathbf{S} \mathbf{V}^T. \quad (17.3.7)$$

Next, using the eigenvalue, eigenvector decomposition of $\mathbf{H}^T \mathbf{H} = \mathbf{V} \mathbf{S}^2 \mathbf{V}^T$, we obtain

$$(\mathbf{V} \mathbf{S}^2 \mathbf{V}^T + \eta \mathbf{V} \mathbf{V}^T)^{-1} (\mathbf{U} \mathbf{S} \mathbf{V}^T)^T \quad (17.3.8)$$

Using the fact that the transpose of a diagonal matrix leaves it unchanged gives

$$\begin{aligned} & (\mathbf{V} \mathbf{S}^2 \mathbf{V}^T + \eta \mathbf{V} \mathbf{V}^T)^{-1} \mathbf{V} \mathbf{S} \mathbf{U}^T \\ &= (\mathbf{V} (\mathbf{S}^2 + \eta \mathbf{I}) \mathbf{V}^T)^{-1} \mathbf{V} \mathbf{S} \mathbf{U}^T. \end{aligned} \quad (17.3.9)$$

We now perform some matrix algebra to obtain

$$\begin{aligned} & (\mathbf{V}^T)^{-1} (\mathbf{S}^2 + \eta \mathbf{I})^{-1} \mathbf{V}^{-1} \mathbf{V} \mathbf{S} \mathbf{U}^T \\ &= \mathbf{V} (\mathbf{S}^2 + \eta \mathbf{I})^{-1} \mathbf{S} \mathbf{U}^T \\ &= \mathbf{V} \Lambda_\eta \mathbf{U}^T. \end{aligned} \quad (17.3.10)$$

□

In order to determine η_{min} we need to evaluate the eigenvalues of $\mathbf{H}^T \mathbf{H}$. Noting that: $\mathbf{H}^T \mathbf{H} = \mathbf{V} \mathbf{S}^2 \mathbf{V}^T$, the squares of the non-zero singular values of \mathbf{H} are equal to the non-zero eigenvalues of $\mathbf{H}^T \mathbf{H}$.

Now that we have an efficient method for evaluating $f(t)$, it remains to solve (17.3.1). The property of unimodality ensures that this line search can be efficiently implemented using the Golden Section search (6). Theoretically, the search region is defined to be over $0 \leq t \leq \infty$. However, in practice, the search can be confined to $0 \leq t \leq t_{max}$ where t_{max} is a sufficiently large upper bound. Note, using Lemma 17.3.1, the solution of (17.3.1) does not involve any matrix inversions, but only one SVD operation.

17.3.1 Complexity Analysis

We now assess the complexity of the new algorithm. We define the cost of an algorithm as the total number of floating-point operations (flops) required to carry it out, as a function of problem

dimensions (7). The most intensive part is the SVD operation which needs to be implemented once. Its complexity is $(4N^2K + 8NK^2 + 9K^3)$ [flops] (7). Since the outer line search consists of only scalar operations and does not depend on either \mathbf{x} or \mathbf{y} , it is negligible. Next we assess the complexity of the inner line search. The vector operations are steps (5), (6) and (13). The complexity of step (5) is $2(K^2N + N^2 + KN)$ [flops]. The complexity of step (6) is $(2K)$ [flops] and the complexity of step (13) is $2N(NK + N)$ [flops]. The overall complexity of the inner line search is $C_{INNER} = 2(K^2N + N^2 + KN + K) ITR_1 + 2(N^2K + N^2)$ [flops], where ITR_1 is the number of iterations of the inner line search. The overall complexity is

$$\begin{aligned} C_{HD} &= (SVD \text{ complexity}) + C_{INNER} ITR_2 \\ &= (4N^2K + 8NK^2 + 9K^3) \\ &\quad + 2(K^2N + N^2 + KN + K) ITR_1 ITR_2 \\ &\quad + 2(N^2K + N^2) ITR_2 \end{aligned} \tag{17.3.11}$$

where ITR_2 is the number of iterations in the outer line search.

We now assess the number of iterations needed for the algorithm to converge to a desired ending tolerance ϵ_{min} . The inner loop can be implemented using bi-section, which has a linear convergence rate. The number of iterations required are $ITR_1 = \log_2\left(\frac{\epsilon_1}{\epsilon_{min}}\right)$ (5), where $\epsilon_1 = -\lambda_{min}(\mathbf{H}^T \mathbf{H}) + \eta_{max}$ is the size of the initially bracketing interval. For the outer line search, using the Golden Section search, each iteration of the line search approximately reduces the original interval by a factor of 0.618 and therefore, the number of iterations required are $ITR_2 = \log_{0.618}\left(\frac{\epsilon_2}{\epsilon_{min}}\right) \approx 2 * \log_2\left(\frac{\epsilon_2}{\epsilon_{min}}\right)$ (5), where $\epsilon_2 = t_{max}$ is the original interval size.

To summarize, the overall complexity of the algorithm is *deterministic*, and does not depend on σ_g^2 and σ_w^2 which makes it appealing for real-time systems.

17.4 Simulation Results

In this section numerical results illustrating the behavior of our new estimator in a MIMO system under a range of variance values for the random matrix \mathbf{G} are provided. For this simulation the parameters are $N = 16$, $K = 4$. The matrix \mathbf{H} was chosen as a concatenation of four 4×4 matrices with unit diagonal elements and 0.5 off-diagonal elements. We ran 5000 computer simulation for every σ_w^2 . The SNR is defined as $-10 \log \sigma_w^2$. We first compare the estimation mean squared error (MSE) of the MAP and the LMMSE solutions for various values of σ_g^2 . This is depicted in Fig. 17.5.1. The results show that the MAP estimate provides a smaller MSE than the LMMSE. Also note that the error floor in all the results is due to the uncertainty in the matrix \mathbf{G} , and can't be avoided, even in high SNR values.

17.5 Conclusions

In this work, we discussed the MAP estimator of a random Gaussian vector \mathbf{x} in a linear model with random transformation matrix \mathbf{G} . We proposed a low complexity algorithm for finding the MAP estimate for the underlined model. We also analyzed the complexity of the algorithm.

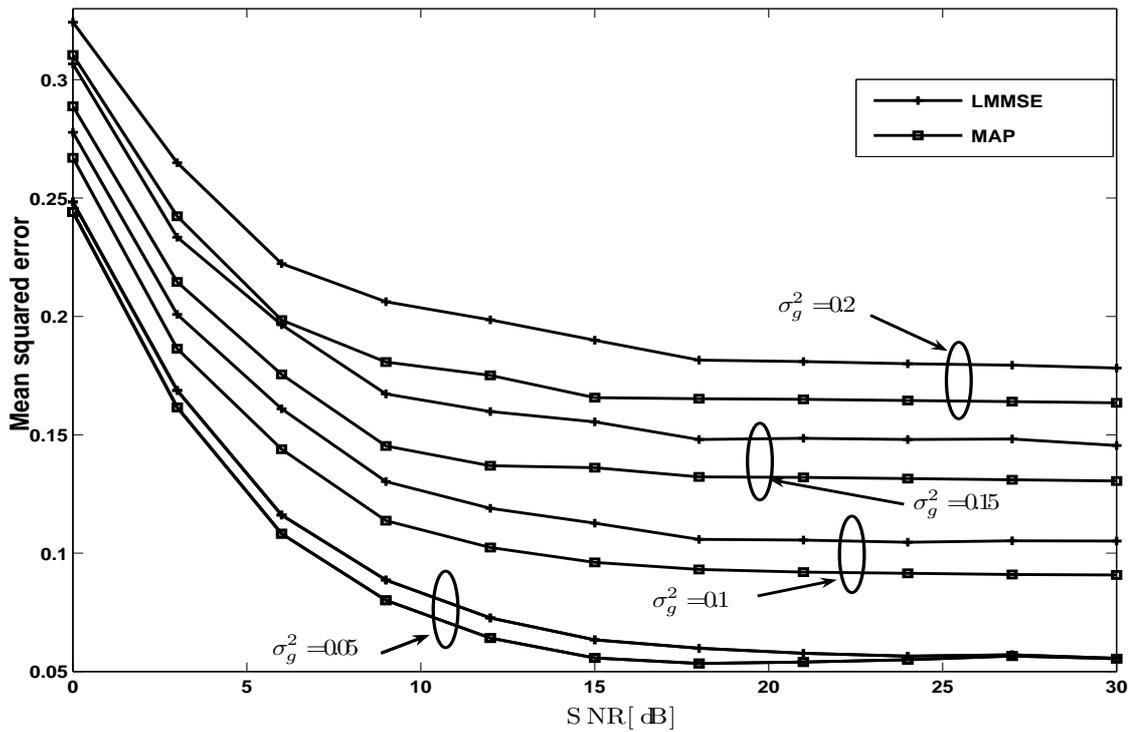


Fig. 17.5.1: Mean squared error for $N = 16, K = 4$

Algorithm 5 Quadratic Equality Constrained Least Squares

Input: $t, \mathbf{H}, \mathbf{y}, \lambda_{\min}(\mathbf{H}^T \mathbf{H}), \Lambda, \mathbf{U}, \mathbf{V}, \eta_{\max}$

Output: $f(t), \mathbf{x}$

1: $\eta_L = -\lambda_{\min}(\mathbf{H}^T \mathbf{H})$

2: $\eta_R = \eta_{\max}$

3: **repeat**

4: $\eta_M = \frac{\eta_L + \eta_R}{2}$

5: $\mathbf{x} = \mathbf{V} \Lambda_{\eta} \mathbf{U}^T \mathbf{y}$

6: $\epsilon = \mathbf{x}^T \mathbf{x} - t$

7: **if** $\epsilon > 0$ **then**

8: $\eta_L = \eta_M$

9: **else**

10: $\eta_R = \eta_M$

11: **end if**

12: **until** $|\epsilon| \leq \epsilon_{\min}$

13: $f(t) = \|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2$

References

- [1] I. Nevat, A. Wiesel, J. Yuan, and Y. C. Eldar, "Maximum A-Posteriori estimation in linear models with a Gaussian model matrix," *Proc. of conference on Information Sciences and Systems, Baltimore*, March 2007.
- [2] S. M. Kay, *Fundamentals of Statistical Signal Processing - Estimation Theory*, Prentice Hall, 1993.
- [3] G. E. Forsythe and G. H. Golub, "On the stationary values of a second-degree polynomial on the unit sphere," *Journal of the Society for Industrial and Applied Mathematics*, vol. 13, no. 4, pp. 1050–1068, 1965.
- [4] Y. C. Eldar and A. Beck, "Hidden convexity based near maximum-likelihood CDMA detection," *Proc. of conference on Information Sciences and Systems*, Princeton, March 2005.
- [5] A. Kharab and R. B. Guenther, *An introduction to numerical methods ũ A MATLAB Approach*, Chapman & Hall/CRC, Cambridge, New York, 2002.
- [6] W. H. Press, Saul A. Teukolsky, W. T. Vetterling, and Brian P. Flannery, *Numerical recipes in C*, Cambridge University Press, Cambridge, second edition, 1992, The art of scientific computing.
- [7] S. Boyd and L. Vandenberghe, *Introduction to Convex Optimization with Engineering Applications*, Stanford, 2003.

Journal Paper 14

"Without the playing with fantasy no creative work has ever yet come to birth. The debt we owe to the play of imagination is incalculable."

Carl Jung

Peters G.W., Nevat I., Yuan J. and Sisson S. (2009) "Bayesian symbol detection in wireless relay networks.", In review at IEEE Transactions on Signal Processing.

This work was instigated jointly by the first and second authors. The first author can claim at least 50% of the credit for the contents. His work included developing the methodology contained and working on the applications, the implementation and comparison to alternative approaches, as well as writing and revising the drafts of the paper. This work will be included in a PhD thesis of a co-author, though this thesis is not submitted by sequence of publications. Permission from all the co-authors is granted for inclusion in the PhD.

This paper appears in the thesis in a modified format to that which was submitted to the journal IEEE Transactions on Signal Processing, where it is in review.

Final print version should be available at:

<http://ieeexplore.ieee.org>

Bayesian symbol detection in wireless relay networks

Gareth W. Peters (*corresponding author*)

CSIRO Mathematical and Information Sciences, Locked Bag 17, North Ryde, NSW, 1670, Australia;

UNSW Mathematics and Statistics Department, Sydney, 2052, Australia;

email: peterga@maths.unsw.edu.au

Ido Nevat

School of Electrical Engineering and Telecommunications, University of NSW, Sydney, Australia.

S. A. Sisson

UNSW Mathematics and Statistics Department, Sydney, 2052, Australia

Y. Fan

UNSW Mathematics and Statistics Department, Sydney, 2052, Australia

Jinhong Yuan

School of Electrical Engineering and Telecommunications, University of NSW, Sydney, Australia.

18.1 Abstract

This paper presents a general stochastic model which is developed for a class of cooperative wireless relay networks, in which imperfect knowledge of the channel state information at the destination node is assumed. This general framework incorporates multiple relay nodes operating under arbitrary processing functions. In general for such systems, due to the intractability of the likelihood function, both the maximum likelihood and the maximum *a posteriori* decision rules do not admit closed form expressions.

We adopt a Bayesian approach, and present three novel computational techniques for maximum *a posteriori* sequence detection in these general networks. These include a Markov chain Monte Carlo approximate Bayesian computation (MCMC-ABC) approach; an auxiliary variable MCMC (MCMC-AV) approach; and a suboptimal exhaustive search zero forcing (SES-ZF) approach. Finally, numerical examples comparing the symbol error rate (SER) performance versus signal to noise ratio (SNR) of the three detection algorithm is studied in simulated examples.

Keywords: approximate Bayesian computation, Markov chain Monte Carlo, likelihood-free computation, auxiliary variable technique, co-operative wireless relay network, partial channel state information.

18.2 Background

Cooperative communications systems, (15) and (9), such as mobile phone networks and the internet, have become a major focus for communications engineers. Particular attention has been paid to the wireless network setting, as it incorporates spatial resources to gain diversity, and enhance connection capability and throughput. More recently, the focus has shifted towards incorporation of relay nodes, (13), which are known to improve energy efficiency, and reduce the interference level of wireless channels, (10).

In simple terms, such a system broadcasts a signal from a transmitter at the source through a wireless channel. The signal is then received by each relay node and a relay strategy is applied before the signal is retransmitted to the destination.

A number of relay strategies have been studied in the literature, (4); (9). We focus on the *amplify-and-forward* (AF) strategy of (3) in which the relay sends a transformed version, via the relay function, of its received signal to the destination. The relay function can be optimized for different design objectives, (8); (7); (5). We demonstrate that the choice of relay function directly affects the tractability of the system model.

In this paper we will focus on a single hop relay design in which the number of relays and the type of relay function are allowed to be general. However, our methodology trivially extends to arbitrary relay topologies and multiple hop networks. Figure 18.9.1 presents the system model considered. We extend the work of (14) to incorporate a stochastic model for the parameters associated with each relay channel link. That is we now consider random variables, associated with the channel model, which must be jointly estimated with the unknown random vector of transmitted symbols.

Since this framework incorporates general relay functions, the resulting likelihood for the model typically cannot be evaluated pointwise, and in some cases cannot even be written down in a closed form. Bayesian "likelihood-free" inference methodologies overcome the problems associated with the intractable likelihood by replacing explicit evaluation of the likelihood with simulation from the model. These methods are also collectively known as approximate Bayesian computation (ABC), see (21), and references therein for a detailed overview of the methodological and theoretical aspects.

Applications of ABC methods are becoming widespread, and include: telecommunications, (14), (21); ecology, (2); extreme value theory, (1); modelling drug-resistance, (11); protein networks, (18), (19); SIR models, (23); species migration (6); pathogen transmission, (22), non-life insurance claims reserving, (17); and operational risk (16).

In this paper we develop a novel sampling methodology based on the Markov chain Monte Carlo approximate Bayesian computation (MCMC-ABC) algorithm of (12). Since the focus of this paper is on detection of the transmitted symbols, we will integrate out the nuisance channel parameters. We also develop two alternative solutions, an auxiliary variable MCMC (MCMC-AV) approach and a suboptimal explicit search zero forcing (SES-ZF) approach. In the

MCMC-AV approach the addition of auxiliary variables results in closed form expressions for the full conditional posterior distributions. We use this fact to develop the MCMC-AV sampler and demonstrate that this works well when small numbers of relay nodes are considered. The SES-ZF approach involves an approximation based on known summary statistics of the channel model and an explicit exhaustive search over the parameter space of code words. This performs well for relatively small numbers of relays and a high signal to noise (SNR) ratio.

In Section 2 we introduce a stochastic system model for the wireless relay system and the associated Bayesian model. We discuss the intractability arising from these models when one considers arbitrary relay functions. Section 3 presents the telecommunications estimation goal, namely maximum *a posteriori* sequence detection. Sections 4 and 5 present the likelihood-free solution and methodology developed for the wireless relay channel Bayesian model. Section 6 presents the alternative auxiliary variable based solution and Section 7 presents results and analysis.

18.3 Bayesian system model and detection

In this section we introduce a new system model and a Bayesian model for inference on the system model parameters. In our system model, the channels in the relay network are modelled stochastically, where we do not know *a priori* the realized channel coefficient values. Instead, we consider partial channel state information (CSI), in which we assume known statistics of the distribution of the channel coefficients.

This significantly extends the work of (14) which demonstrated how to perform sequence detection in arbitrary relay systems in which *perfect* knowledge of the channel model parameters was assumed. Additionally, no knowledge in the form of pilot symbols or training sequences are transmitted as part of the signals.

We consider M-ary Pulse Amplitude Modulated (PAM) sequences of symbols. Transmitted symbols are from an alphabet of M code words, which correspond to a constellation of points non-uniformly distributed on the complex plane, each labeled with a binary sequence of bits. Our methodology extends naturally to other constellation and encoding frameworks.

The following are relevant definitions which are used throughout the wireless telecommunications literature. We define the signal to noise ratio in units of decibels (dB) which is given by $SNR = 10\log_{10} \frac{P_{\text{signal}}}{P_{\text{noise}}}$. We define the symbol error rate (SER) as the proportion of detected symbols that are miss-detected, relative to the number of symbols transmitted. This will vary as a function of the SNR and is a primary focus when considering the performance of wireless communications systems. The term "frame" is defined as the duration over which the channel realization is unchanged. In slow fading channels this can encompass several transmitted sequences. In this paper we only require that it is at least as long as the duration of one transmitted sequence.

18.3.1 Model and assumptions

Here we present the system model and associated assumptions. We will work with notation in which random variables are denoted by upper case letters and their realizations by lower case letters. In addition, bold will be used to denote a vector or matrix quantity, upper subscripts will refer to the relay node index and lower subscripts refer to the element of a vector or matrix.

1. Assume a wireless relay network with one source node, transmitting sequences of K symbols denoted $\mathbf{s} = s_{1:K}$.
2. The sequence of K symbols \mathbf{s} are from an M -ary pulse amplitude modulation (PAM) scheme.
3. The relays cannot transmit and receive on the same time slot on the same frequency band. We thus consider a half duplex system model in which the data transmission is divided into two steps. In the first step, the source node broadcasts a code word $\mathbf{s} \in \Omega$ from the codebook to all the relay nodes. In the second step, the relay nodes then transmit the relay signals to the destination node in orthogonal non-interfering channels. We assume that all channels are independent with a coherence interval larger than the codeword length K .
4. Assume imperfect CSI in which noisy estimates of the channel model coefficients for each relay link are known. This is a standard assumption based on the fact that a training phase has been performed a priori. This involves an assumption regarding the channel coefficients as follows:
 - Source to relay there are L i.i.d. channels parameterized by $\left\{ H^{(l)} \sim F\left(\hat{h}^{(l)}, \sigma_h^2\right) \right\}_{l=1}^L$, where $F(\cdot)$ is the distribution of the channel coefficients, $\hat{h}^{(l)}$ is the estimated channels coefficient and σ_h^2 is the associated estimation error variance.
 - Relay to destination there are L i.i.d. channels parameterized by $\left\{ G^{(l)} \sim F\left(\hat{g}^{(l)}, \sigma_g^2\right) \right\}_{l=1}^L$, where $F(\cdot)$ is the distribution of the channel coefficients, $\hat{g}^{(l)}$ is the estimated channels coefficient and σ_g^2 is the associated estimation error variance.
5. The received signal at the l -th relay is a random vector given by

$$\mathbf{R}^{(l)} = \mathbf{S}H^{(l)} + \mathbf{W}^{(l)}, \quad l \in \{1, \dots, L\}, \quad (18.3.1)$$

where $H^{(l)}$ is the channel coefficient between the transmitter and the l -th relay, $\mathbf{S} \in \Omega_M$ is the transmitted code-word and $\mathbf{W}^{(l)}$ is the noise realization associated with the relay receiver.

6. The received signals at the destination is a random vector given by

$$\mathbf{y}^{(l)} = \mathbf{f}^{(l)}\left(\mathbf{R}^{(l)}\right)G^{(l)} + \mathbf{V}^{(l)}, \quad l \in \{1, \dots, L\}, \quad (18.3.2)$$

where $G^{(l)}$ is the channel coefficient between the l -th relay and the receiver, $\mathbf{f}^{(l)}\left(\mathbf{R}^{(l)}\right) \triangleq \left[f^{(l)}\left(r_1^{(l)}\right), \dots, f^{(l)}\left(r_K^{(l)}\right) \right]^T$ is the memoryless relay processing function (with possibly different functions at each of the relays) and $\mathbf{V}^{(l)}$ is the noise realization associated with the relay receiver.

7. Conditional on $\{h^{(l)}, g^{(l)}\}_{l=1}^L$, we have that all received signals are corrupted by zero-mean additive white complex Gaussian noise. At the l -th relay the noise corresponding to the l -th transmitted symbol is denoted by random variable $W_i^{(l)} \sim \mathcal{CN}(0, \sigma_w^2)$. At the receiver this is denoted by random variable $V_i^{(l)} \sim \mathcal{CN}(0, \sigma_v^2)$.

Additionally, we assume the following properties:

$$\mathbb{E} \left[W_i^{(l)} \overline{W}_j^{(m)} \right] = \mathbb{E} \left[V_i^{(l)} \overline{V}_j^{(m)} \right] = \mathbb{E} \left[W_i^{(l)} \overline{V}_j^{(m)} \right] = 0,$$

$\forall (i, j) \in \{1, \dots, K\}, \forall (l, m) \in \{1, \dots, L\}, i \neq j, l \neq m$, where \overline{W}_j denotes the complex conjugate of W_j .

Prior specification and posterior

Here we present the relevant aspects of the Bayesian model and associated assumptions. We begin by specifying the prior models for the sequence of symbols and the unknown channel coefficients. At this stage we note that in terms of system capacity it is only beneficial to transmit a sequence of symbols if it aids detection. This is achieved by having correlation in the symbol sequence. Hence, typically the transmitted symbols $s_{1:K}$ will be correlated. We assume that since this is part of the system design, this will be known, hence the prior structure for $p(s_{1:K})$ should reflect this information.

1. Under the Bayesian model, the symbol sequence is treated as a random vector $\mathbf{S} = S_{1:K}$. The prior for the random symbols sequence (code word) $S_{1:K}$ is defined on a discrete support denoted Ω_M with $|\Omega_M| = M^K$ elements and pmf denoted by $p(s_{1:K})$.
2. The assumption of imperfect CSI is treated under a Bayesian paradigm by formulating priors for the channel coefficients as follows:

- Source to relay there are L i.i.d. channels parameterized by $\left\{ H^{(l)} \sim \mathcal{CN}(\widehat{h}^{(l)}, \sigma_h^2) \right\}_{l=1}^L$, where $\widehat{h}^{(l)}$ is the estimated channels coefficient and σ_h^2 the associated estimation error variance.
- Relay to destination there are L i.i.d. channels parameterized by $\left\{ G^{(l)} \sim \mathcal{CN}(\widehat{g}^{(l)}, \sigma_g^2) \right\}_{l=1}^L$, where $\widehat{g}^{(l)}$ is the reestimated channels coefficient and σ_g^2 the associated estimation error variance.

The likelihood model $p(\mathbf{y}^{(l)} | \mathbf{s}, h^{(l)}, g^{(l)})$ for this relay system is in general computationally intractable. There are two potential difficulties that arise when dealing with non-linear relay functions. The first relates to finding the distribution of the signal transmitted from each relay to the destination, this involves finding the density of the random vector $\mathbf{f}^{(l)} \left(\mathbf{S} H^{(l)} + \mathbf{W}^{(l)} \right) G^{(l)}$ conditional on realizations $\mathbf{S} = \mathbf{s}, \mathbf{H} = \mathbf{h}, \mathbf{G} = \mathbf{g}$. This is not always possible for a general non-linear multivariate function $\mathbf{f}^{(l)}$. Conditional on $\mathbf{S} = \mathbf{s}, \mathbf{H} = \mathbf{h}, \mathbf{G} = \mathbf{g}$, we know the distribution of $\mathbf{R}^{(l)} | \mathbf{s}, \mathbf{g}, \mathbf{h}$,

$$p_{\mathbf{R}}(\mathbf{R} | \mathbf{s}, \mathbf{g}, \mathbf{h}) = p\left(\mathbf{s}h^{(l)} + \mathbf{w}^{(l)} | \mathbf{s}, h^{(l)}, g^{(l)}\right) = \mathcal{CN}\left(\mathbf{s}h^{(l)}, \sigma_w^2 \mathbf{I}\right). \quad (18.3.3)$$

However, finding the distribution of the random vector after the non-linear function is applied i.e. the distribution of $\tilde{\mathbf{f}}(\mathbf{R}^{(l)}) \triangleq \mathbf{f}(\mathbf{R}^{(l)}) G^{(l)}$ given $\mathbf{s}, h^{(l)}, g^{(l)}$, involves the following change of variable formula

$$p(\tilde{\mathbf{f}}(\mathbf{R}^{(l)}) | \mathbf{s}, h^{(l)}, g^{(l)}) = p_{\mathbf{R}}\left(\left(\tilde{\mathbf{f}}^{(l)}\right)^{-1}\left(\mathbf{R}^{(l)}\right) | \mathbf{s}, h^{(l)}, g^{(l)}\right) \left| \frac{\partial \tilde{\mathbf{f}}^{(l)}}{\partial \mathbf{R}^{(l)}} \right|^{-1}, \quad (18.3.4)$$

which can not always be written down analytically for arbitrary $\tilde{\mathbf{f}}$. The second more serious complication is that even in cases where the density for the transmitted signal is known, one must then solve a K fold convolution to obtain the likelihood:

$$\begin{aligned} p(\mathbf{y}^{(l)} | \mathbf{s}, \mathbf{g}, \mathbf{h}) &= p(\tilde{\mathbf{f}}(\mathbf{R}^{(l)}) | \mathbf{s}, \mathbf{g}, \mathbf{h}) * p_{\mathbf{V}^{(l)}} \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p(\tilde{\mathbf{f}}(\mathbf{z} | \mathbf{s}, \mathbf{g}, \mathbf{h})) p_{\mathbf{V}^{(l)}}(\mathbf{y}^{(l)} - \mathbf{z}) dz_1 \dots dz_K. \end{aligned} \quad (18.3.5)$$

Typically this will be intractable to evaluate pointwise. However, in the most simplistic case of a linear relay function, the likelihood can be obtained analytically as

$$p(\mathbf{y}^{(l)} | \mathbf{s}, h^{(l)}, g^{(l)}) = \mathcal{CN}\left(\mathbf{s}h^{(l)}g^{(l)}, \left(|g^{(l)}|^2 \sigma_w^2 + \sigma_v^2\right) I\right), \quad (18.3.6)$$

where, I is the identity matrix.

Hence, the resulting posterior distribution involves combining the likelihood given in Equation 18.3.5 with the priors for the sequence of symbols and the channel coefficients.

18.3.2 Inference and MAP sequence detection

Since the primary concern in designing a relay network system is on SER versus SNR our goal is oriented towards detection of transmitted symbols and not the associate problem of channel estimation. We will focus on an approach which samples $S_{1:K}$, $\mathbf{G} = G_{1:L}$, $\mathbf{H} = H_{1:L}$, jointly from the target posterior distribution.

In particular we consider the maximum *a posteriori* (MAP) sequence detector at the destination node. Therefore the goal is to design a MAP detection scheme for \mathbf{s} at the destination, based on the received signals $\{\mathbf{y}^{(l)}\}_{l=1}^L$, the noisy channels estimates as given by the partial CSI $\{\hat{h}^{(l)}\}_{l=1}^L$, $\{\hat{g}^{(l)}\}_{l=1}^L$ and the noise variances, σ_n^2 and σ_v^2 .

Since the channels are mutually independent, the received signals $\{\mathbf{R}^{(l)}\}_{l=1}^L$ and $\{\mathbf{y}^{(l)}\}_{l=1}^L$ are conditionally independent given $\mathbf{s}, \mathbf{g}, \mathbf{h}$. Thus, the MAP decision rule, after marginalizing out the unknown channel coefficients, is given by

$$\hat{\mathbf{s}} = \arg \max_{\mathbf{s} \in \Omega} \prod_{l=1}^L \int \int p(\mathbf{s}, \mathbf{g}, \mathbf{h} | \mathbf{y}^{(l)}) d\mathbf{h} d\mathbf{g} = \arg \max_{\mathbf{s} \in \Omega} \prod_{l=1}^L \int \int p(\mathbf{y}^{(l)} | \mathbf{s}, \mathbf{h}, \mathbf{g}) p(\mathbf{s}) p(\mathbf{g}, \mathbf{h}) d\mathbf{h} d\mathbf{g}. \quad (18.3.7)$$

The intractability of the likelihood model in (18.3.5) results in our development of likelihood-free based methodology and associated MCMC-ABC sampler. Two alternative approaches for MAP detection based on auxiliary variables and zero forcing methods will be considered.

18.4 Likelihood-free methodology

Likelihood-free inference is a suite of methods developed specifically for working with models in which the likelihood is computationally intractable. Here we work with a Bayesian model and consider the likelihood intractability to arise in the sense that we may not evaluate the likelihood pointwise, (Section 18.3.2). Additionally, in general we can only obtain an expression for the likelihood in terms of a multivariate convolution integrals and hence we do not have an explicit closed form expression for the likelihood.

It is shown in (21) that the ABC method we consider here embeds an intractable target posterior distribution, in our case denoted by $p(s_{1:K}, h_{1:L}, g_{1:L} | \mathbf{y})$, into a general augmented model,

$$p(s_{1:K}, h_{1:L}, g_{1:L}, \mathbf{x}, \mathbf{y}) = p(\mathbf{y} | \mathbf{x}, s_{1:K}, h_{1:L}, g_{1:L}) p(\mathbf{x} | s_{1:K}, h_{1:L}, g_{1:L}) p(s_{1:K}) p(h_{1:L}) p(g_{1:L}), \quad (18.4.1)$$

where $\mathbf{x} \in \chi$ is an auxiliary vector on the same space as \mathbf{y} . In this augmented Bayesian model, the density $p(\mathbf{y} | \mathbf{x}, s_{1:K}, h_{1:L}, g_{1:L})$ weights the intractable posterior. In this paper we consider the model assumption where we work with, $p(\mathbf{y} | \mathbf{x}, s_{1:K}, h_{1:L}, g_{1:L}) = p(\mathbf{y} | \mathbf{x})$.

The actual ABC mechanism which allows one to avoid the evaluation of the intractable likelihood is via simulation from the likelihood. That is, given a realisation of the parameters of the model, a synthetic data set is generated. Then summary statistics evaluated on this data are compared to evaluation on the observed data, and a distance is calculated. Finally, a weight is given to these parameters according to the weighting function $p(\mathbf{y} | \mathbf{x})$. This is made explicit when the algorithm is presented incorporating the ABC mechanism and also when details of the ABC methodology and algorithm choices are presented in Section .

In this paper we examine the MCMC-ABC under two different weighting functions: the first is a popular "Hard Decision" (HD) weighting given by,

$$p(\mathbf{y} | \mathbf{x}, s_{1:K}, h_{1:L}, g_{1:L}) \propto \begin{cases} 1, & \text{if } \rho(D(\mathbf{y}), D(\mathbf{x})) \leq \epsilon, \\ 0, & \text{otherwise;} \end{cases} \quad (18.4.2)$$

the other utilizes an idea which we term a "Soft Decision" (SD) weighting function given by,

$$p(\mathbf{y} | \mathbf{x}, s_{1:K}, h_{1:L}, g_{1:L}) \propto \exp\left(-\frac{\rho(D(\mathbf{y}), D(\mathbf{x}))}{\epsilon^2}\right). \quad (18.4.3)$$

Hence, the HD weighting function rewards summary statistics of the augmented auxiliary variables, $D(\mathbf{x})$, within an ϵ -tolerance of the summary statistic of the actual observed data, $D(\mathbf{y})$, as measured by distance metric . The SD weighting function clearly penalizes sum-

mary statistics as a non-linear function of the distance between summary statistics. However, a key difference is that, even though the weighing may be small, it will remain non-zero, unlike the HD rule.

Hence, in the ABC context the intractable target posterior marginal distribution, $p(s_{1:K}|\mathbf{y})$, for which we are interested in formulating a MAP detector, is given by:

$$p_{ABC}(s_{1:K}|\mathbf{y}, \epsilon) \propto \int_{\mathcal{X}} \int \int p(\mathbf{y}|\mathbf{x}, s_{1:K}, h_{1:L}, g_{1:L}) p(\mathbf{x}|s_{1:K}, h_{1:L}, g_{1:L}) p(s_{1:K}) p(h_{1:L}) p(g_{1:L}) dh dg dx$$

Here, we make an important note about the MCMC-ABC algorithm. As discussed, (21), this particular class of algorithm is justified on a joint space ABC formulation in which the stationary distribution of the Markov chain is given by, $p_{ABC}(s_{1:K}, \mathbf{x}|\mathbf{y}, \epsilon)$ and then the target distribution corresponding to the marginal distribution $p_{ABC}(s_{1:K}|\mathbf{y}, \epsilon)$ is obtained via numerical integration. Additionally, we note that the marginal posterior distribution

$p_{ABC}(s_{1:K}|\mathbf{y}, \epsilon) \rightarrow p(s_{1:K}|\mathbf{y})$ as $\epsilon \rightarrow 0$ (assuming sufficient statistics for the summary statistics and a weighting function of the form considered in this paper).

Hence, we note that the tolerance ϵ typically should be set as low as possible for a given computational budget. Typically, this will depend in the ABC context on the choice of algorithm used to sample from the ABC posterior distribution. In this paper we focus on the class of sampling algorithms known as MCMC-ABC. In the next section we first present specific details of choices that must be made when constructing a likelihood-free inference model, and then we present the sampling algorithm based on MCMC-ABC that we utilize in this paper.

18.4.1 Approximate Bayesian computation MCMC approach

Here we present a novel algorithm to perform MAP detection of a sequence of transmitted symbols. To achieve this we utilize a MCMC-ABC Sampler based on (12); (1). In particular we utilize a random scan Metropolis-Hastings within Gibbs sampler. The algorithm is depicted in Algorithm 1. Next we present details and discussion around the choices made for the ABC components of the likelihood-free methodology.

Observations and synthetic data

The data $\mathbf{y} = y_{1:K}$ correspond to the observed sequence of symbols at the receiver. The generation of the "synthetic data" in the likelihood-free approximation involves generating auxiliary variables $x_1^{(l)}, \dots, x_K^{(l)}$ from the model, $p(\mathbf{x}^{(l)}|\mathbf{s}^*, \mathbf{g}^*, \mathbf{h}^*)$, for $l = 1, \dots, L$, to obtain a realization of $\mathbf{X} = \mathbf{x} = [\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(L)}]^T$. This is achieved under the following steps:

1. Sample $\mathbf{W}^{(l)*} \sim \mathcal{CN}(\mathbf{0}, \sigma_w^2 I)$.
2. Sample $\mathbf{V}^{(l)*} \sim \mathcal{CN}(\mathbf{0}, \sigma_v^2 I)$
3. Evaluate $\mathbf{X}^{(l)} = \mathbf{f}^{(l)}(\mathbf{S}^* h^{(l)} + \mathbf{W}^{(l)*}) g^{(l)} + \mathbf{V}^{(l)*}, l \in \{1, \dots, L\}$,

Summary statistics

As discussed in the likelihood-free methodology at the beginning of Section 3, summary statistics are used in the comparison between the synthetic data and the actual data via the weighting function on the distance metric. For the observed data \mathbf{y} we define the summary vector $D(\mathbf{y})$ and for the synthetic data \mathbf{x} we define $D(\mathbf{x})$. Note, summary statistics are utilised since they provide a more efficient mechanism of comparison. In particular, under the Neymann-Fisher factorization theorem, when sufficient statistics are utilised no loss of information is incurred when using these summary statistics in place of a comparison of the actual data and synthetic data. However, it is more probable to match summary statistics than to match generations of synthetic data with actual data. However, when non-sufficient summary statistics are used, a bias is introduced at the expense of a more efficient algorithm.

There are many possibilities that can be chosen for the summary statistics, but most critically, we want these as near to sufficient as possible whilst low in dimension, we shall comment on a few choices here. The simplest choice is to use $D(\mathbf{y}) = \mathbf{y}$, hence one works with the actual data, this is optimal in the sense that it does not result in a loss of information from the observations. The reason this choice is rarely used is that it typically it will result in poor performance of the MCMC-ABC algorithm. To understand this, consider the HD rule weighting function in Equation 18.4.2. In this case, even if the true MAP estimate model parameters were utilized to generate the synthetic data \mathbf{x} , it will still be become improbable to realize a non-zero weight as $\epsilon \rightarrow 0$. This is made worse as the number of observations increases, curse of dimensionality. As a result, the acceptance probability in the MCMC-ABC algorithm would be 0 for long periods, resulting in poor sampler efficiency, see (20) for a related discussion on chain mixing. We note that the exception to this rule is when a moderate tolerance ϵ and small number of observations are used.

The next best choice is to utilize a set of summary statistics which are sufficient statistics for the model parameters. In this special case, there is no loss of data through use of the summary statistics versus the whole data set. Unfortunately, generally the sufficient statistics for an arbitrary model are unknown. A practical alternative to utilizing the entire data set, is to use empirical quantile estimates of the distribution of the observations and the generated "synthetic observation". This could correspond to a vector of quantiles for example, $D(\mathbf{y}) = [\hat{q}_{0.1}(\mathbf{y}), \dots, \hat{q}_{0.9}(\mathbf{y})]$, where $\hat{q}_\alpha(\mathbf{y})$ denotes the estimated empirical quantile at $\alpha\%$.

Distance metric

The next requirement for the likelihood-free Bayesian model is the specification of the distance metric. In this paper we consider the Mahlanobis and the scaled Euclidean distances, where the Mahlanobis distance metric is given by,

$$\rho(D(\mathbf{y}), D(\mathbf{x})) = [D(\mathbf{y}) - D(\mathbf{x})]^\top \Sigma_{\mathbf{x}|\mathbf{s},\mathbf{h},\mathbf{g}}^{-1} [D(\mathbf{y}) - D(\mathbf{x})],$$

where the covariance matrix $\Sigma_{\mathbf{x}|\mathbf{s},\mathbf{h},\mathbf{g}}$ is an appropriate scaling. As specified in the weighting function, in this paper we consider the model with $\sigma_{\mathbf{x}}$.

An important question is related to the best way to estimate this covariance matrix. The approach we take in this paper is off-line and can be computed once prior to simulation. The approach used in this paper involved taking the partial CSI estimates of the channel coefficients and the mode of the prior for the symbols. Another approach we also tested involved sampling a realisation from the prior for the symbols. Then given these parameters, we generated 2,000 data sets. The empirical covariance of summary statistics from these data sets are then obtained and used in the Mahalanobis distance metric.

The scaled Euclidean distance is obtained when we only consider the diagonal elements of the covariance matrix Σ_x . The Euclidean distance corresponds to the case in which we set the covariance matrix to the identity matrix.

Note, the covariance matrix Σ_x provides a weighting on each element of the vector of summary statistics according to how much estimated variation they display relative to each other. There are many other such weighting schemes one could conceive for the vector of summary statistics. In general, though any of these weighting schemes will produce accurate results (in the limit as $\epsilon \rightarrow 0$ the exact form of Σ in ρ is immaterial). However, the efficiency of the resulting algorithm is directly affected by the matrix used. We demonstrate this in the results section.

Weighting function

The next requirement is the specification of the decision or weighting function to be used in the likelihood-free model. As presented in (21) We consider both a HD weighting function, see Equation 18.4.2 and a SD weighting function, see Equation 18.4.3.

Tolerance schedule

For the annealed tolerance schedule, during burnin of the Markov chain, we use the sequence

$$\epsilon_t = \max \{20,000 - 10t, \epsilon^{\min}\}.$$

We note that with a HD weighting function the use of an MCMC-ABC algorithm can result in "sticking" of the chain for extended periods, since the acceptance probability can be 0 for long periods when the chain explores the tails of the posterior distribution. Therefore, one should carefully monitor convergence diagnostics of the resulting Markov chain for a given tolerance schedule. There is a trade-off between the length of the Markov chain required for samples approximately from the stationary distribution and the bias introduced by non zero tolerance. In this paper we set ϵ^{\min} via preliminary analysis of the Markov chain sampler mixing rates for a transition kernel with coefficient of variation set to one.

In general practitioners will have a required precision in posterior estimates, this precision can be directly used to determine, for a given computational budget, a suitable tolerance ϵ^{\min} .

Performance diagnostic

The performance diagnostic we consider is the autocorrelation evaluated on samples post annealing ($\tilde{T} = T - T_b$) of the tolerance threshold ie. after the initial burnin period T_b . We denote by $\{\theta_i^{(t)}\}_{t=1:\tilde{T}}$ the Markov chain of the i -th parameter after burnin. Given Markov chain sam-

ples for the i -th parameter $\{\theta_i^{(t)}\}_{t=1:\tilde{T}}$ we define the biased autocorrelation estimate at lag τ by

$$\widehat{ACF}(\theta_i, \tau) = \frac{1}{(\tilde{T} - \tau)\hat{\sigma}(\theta_i)} \sum_{t=1}^{\tilde{T}-\tau} [\theta_i^{(t)} - \hat{\mu}(\theta_i)][\theta_i^{(t+\tau)} - \hat{\mu}(\theta_i)], \quad (18.4.4)$$

where $\hat{\mu}(\theta_i)$ and $\hat{\sigma}(\theta_i)$ are the estimated mean and standard deviation of θ_i .

18.5 Auxiliary variable MCMC approach

In this section we demonstrate that at the expense of increasing the parameter vector to be estimated by the size of the data series, one can develop a standard MCMC algorithm without the requirement of the ABC methodology. To understand this just associate the additional auxiliary variables with the unknown noise realizations at each relay, $\mathbf{w}_{1:K}$. In this case we avoid the intractability of the likelihood model due to the complications described in Remark 2.

If we augment the parameter vector of interest with the auxiliary variables $\mathbf{w}_{1:K}$ to obtain a new parameter vector, $(\mathbf{s}_{1:K}, g_{1:L}, h_{1:L}) \cup (\mathbf{w}_{1:K})$ then we can consider the following posterior distribution $p(\mathbf{s}_{1:K}, g_{1:L}, h_{1:L}, \mathbf{w}_{1:K} | \mathbf{y})$. Then we can decompose this posterior into the following full conditional distributions which form a block Gibbs sampling framework:

$$\begin{aligned} p(\mathbf{s}_{1:K} | g_{1:L}, h_{1:L}, \mathbf{w}_{1:K}, \mathbf{y}) &\propto p(\mathbf{y} | \mathbf{s}_{1:K}, g_{1:L}, h_{1:L}, \mathbf{w}_{1:K}) p(\mathbf{s}_{1:K}) \\ p(g_{1:L} | \mathbf{s}_{1:K}, h_{1:L}, \mathbf{w}_{1:K}, \mathbf{y}) &\propto p(\mathbf{y} | \mathbf{s}_{1:K}, g_{1:L}, h_{1:L}, \mathbf{w}_{1:K}) p(g_{1:L}) \\ p(h_{1:L} | \mathbf{s}_{1:K}, g_{1:L}, \mathbf{w}_{1:K}, \mathbf{y}) &\propto p(\mathbf{y} | \mathbf{s}_{1:K}, g_{1:L}, h_{1:L}, \mathbf{w}_{1:K}) p(h_{1:L}) \\ p(\mathbf{w}_{1:K} | \mathbf{s}_{1:K}, g_{1:L}, h_{1:L}, \mathbf{y}) &\propto p(\mathbf{y} | \mathbf{s}_{1:K}, g_{1:L}, h_{1:L}, \mathbf{w}_{1:K}) p(\mathbf{w}_{1:K}) \end{aligned} \quad (18.5.1)$$

Clearly, if we associate the additional auxiliary parameters as the unknown noise random variables at the relays then we can obtain a simple closed form solution for the likelihood models allowing us to obtain tractable full conditional posterior distributions for equations (18.5.1). In this case the likelihood in each of these full conditional posterior distributions is simply given by,

$$p(\mathbf{y} | \mathbf{s}_{1:K}, g_{1:L}, h_{1:L}, \mathbf{w}_{1:K}) = \prod_{l=1}^L p(\mathbf{y}^{(l)} | \mathbf{s}_{1:K}, g^{(l)}, h^{(l)}, \mathbf{w}^{(l)}), \quad (18.5.2)$$

where

$$p(\mathbf{y}^{(l)} | \mathbf{s}_{1:K}, g^{(l)}, h^{(l)}, \mathbf{w}^{(l)}) = \mathcal{CN}(\mathbf{f}^{(l)} (\mathbf{S}h^{(l)} + \mathbf{W}^{(l)}) g^{(l)}, \sigma_v^2 \mathbf{I}) \quad (18.5.3)$$

We can then proceed by implementing a Metropolis-within Gibbs sampler for this block Gibbs framework to sample from these full conditional distributions. This is outlined in Algorithm 18.5, where we define the joint posterior parameter vector $\Theta = (\mathbf{S}, \mathbf{G}, \mathbf{H}, \mathbf{W})$.

The proposals used for the Metropolis-Hastings algorithm used to sample from each full conditional distribution were given as follows.

- Draw proposal $\mathbf{S}_{1:K}^*$ from distribution $q(\mathbf{s}_{1:K}^{(t-1)} \rightarrow \mathbf{s}_{1:K}^*) = q(i)q(\mathbf{s}_i|\mathbf{s}_i^{(t-1)}) = \frac{1}{K} \frac{1}{\log_2 M} \delta_{\mathbf{s}_{1:i-1}^{(t-1)}} \delta_{\mathbf{s}_{i+1:K}^{(t-1)}}$, where δ_θ denotes a dirac mass on location θ .
- Draw proposal \mathbf{G}_i^* from distribution $q(\mathbf{g}_i^{(t-1)} \rightarrow \mathbf{g}_i^*) = q(l)q(g_i^{(l)}|\mathbf{g}_i^{(t-1)}) = \frac{1}{L} \text{CN}(\mathbf{g}_i^{(t-1)}, \sigma_{g_{rw}}^2)$.
- Draw proposal \mathbf{H}_i^* from distribution $q(\mathbf{h}_i^{(t-1)} \rightarrow \mathbf{h}_i^*) = q(l)q(h_i^{(l)}|\mathbf{h}_i^{(t-1)}) = \frac{1}{L} \text{CN}(\mathbf{h}_i^{(t-1)}, \sigma_{h_{rw}}^2)$.
- Draw proposal \mathbf{W}_i^* from distribution $q(\mathbf{w}_i^{(t-1)} \rightarrow \mathbf{w}_i^*) = q(l)q(i)q(w_i^{(l)}|\mathbf{w}_i^{(t-1)}) = \frac{1}{KL} \text{CN}(\mathbf{w}_i^{(t-1)}, \sigma_{w_{rw}}^2)$.

Hence, this MCMC-AV approach presents an alternative to the likelihood-free Bayesian model sampler which produces exact samples from the posterior once stationarity of the Markov chain is reached. The approach still performs channel estimation and detection jointly, however instead of utilizing ABC methodology, it works with an augmented parameter space. The trade-off is that typically sampling the additional large number of extra parameters will result in requiring overall much longer Markov Chains to achieve the same performance as the ABC algorithm in terms of joint estimation and detection performance. This is especially true in high dimensional problems, such as when the sequence of transmitted symbols is long and the number of relays present in the system is large or when the posterior distribution of the additional auxiliary variables exhibits strong dependence.

18.6 Alternative MAP detectors and lower bound performance

One can define a suboptimal solution to the MAP detector, even with an intractable likelihood, involving a naive, highly computational algorithm based on a Zero Forcing (ZF) solution. The ZF solution is popular in simple system models where it can be efficient and performs well.

Under a ZF solution one conditions on some knowledge of the partial channel state information, and then perform an explicit search over the set of all possible symbol sequences. To our knowledge a ZF solution for MAP sequence detection in arbitrary non-linear relay systems has not been defined. Accordingly we define the ZF solution for MAP sequence detection as the solution which conditions on the mean of the noise at the relay nodes, and also uses the noisy channel estimates given by the partial CSI information, to reduce the detection search space.

18.7 Sub-optimal exhaustive search Zero Forcing approach

In this approach we condition on the mean of the noise $\mathbf{W}^{(l)} = \mathbf{0}$, and use the partial CSI estimates of channels coefficients, $\{\hat{h}^{(l)}, \hat{g}^{(l)}\}_{l=1}^L$, to reduce the dimensionality of the MAP detector search space to just the symbol space Ω . The SES-ZF-MAP sequence detector can be expressed

as

$$\begin{aligned}\hat{\mathbf{s}} &= \arg \max_{\mathbf{s} \in \Omega} \prod_{l=1}^L p\left(\mathbf{s} | \mathbf{y}^{(l)}, G^{(l)} = \hat{g}^{(l)}, H^{(l)} = \hat{h}^{(l)}, \mathbf{W}^{(l)} = \mathbf{0}\right) \\ &= \arg \max_{\mathbf{s} \in \Omega} \prod_{l=1}^L p\left(\mathbf{y}^{(l)} | \mathbf{s}, G^{(l)} = \hat{g}^{(l)}, H^{(l)} = \hat{h}^{(l)}, \mathbf{W}^{(l)} = \mathbf{0}\right) p(\mathbf{s}).\end{aligned}\quad (18.7.1)$$

Thus, the likelihood model results in a complex Gaussian distribution for each relay channel, as follows

$$p\left(\mathbf{y}^{(l)} | \mathbf{s}, G^{(l)} = \hat{g}^{(l)}, H^{(l)} = \hat{h}^{(l)}, \mathbf{W}^{(l)} = \mathbf{0}\right) = \mathcal{CN}\left(\mathbf{f}^{(l)}\left(\mathbf{s}\hat{h}^{(l)}\right)\hat{g}^{(l)}, \sigma_v^2 \mathbf{I}\right). \quad (18.7.2)$$

As a result, the MAP detection can be solved exactly using an explicit search.

Note however that this approach to symbol detection also involves a very high computational cost, as one must evaluate the posterior distribution for all M^K code words in Ω . It is usual for communications systems to utilise M as either 64-ary PAM or 128-ary PAM and the number of symbols can be anything from $K = 1$ to $K = 20$ depending on the channel capacity budget for the designed network and the typical operating SNR level. Typically this explicit search is not feasible to perform. However, the sub-optimal ZF-MAP detector provides a comparison for the MCMC-ABC approach, which at low SNR should be a reasonable upper bound for the SER and for high SNR an approximate optimal solution.

The SES-ZF-MAP sequence detector can be highly sub-optimal for low SNR values. This is trivial to see, since we are explicitly setting the noise realisations to zero when the variance the noise distribution is large. For the same reasoning, in high SNR values, the ZF approach becomes close to optimal.

18.7.1 Lower bound MAP detector performance

We denote the theoretical lower bound for the MAP detector performance as the oracle MAP detector (OMAP). The OMAP detector involves conditioning on perfect oracular knowledge of the channels coefficients $\{h^{(l)}, g^{(l)}\}_{l=1}^L$ and of the realized noise sequence at each relay $\mathbf{W}^{(l)}$. This results in the likelihood model for each relay channel being complex Gaussian, resulting in an explicit solution for the MAP detector. Accordingly, the OMAP detector provides the lower bound for the SER performance. The OMAP detector can be expressed as

$$\begin{aligned}\hat{\mathbf{s}} &= \arg \max_{\mathbf{s} \in \Omega} \prod_{l=1}^L p\left(\mathbf{s} | \mathbf{y}^{(l)}, G^{(l)} = g^{(l)}, H^{(l)} = h^{(l)}, \mathbf{W}^{(l)} = \mathbf{W}^{(l)}\right) \\ &= \arg \max_{\mathbf{s} \in \Omega} \prod_{l=1}^L p\left(\mathbf{y}^{(l)} | \mathbf{s}, G^{(l)} = g^{(l)}, H^{(l)} = h^{(l)}, \mathbf{W}^{(l)} = \mathbf{W}^{(l)}\right) p(\mathbf{s}).\end{aligned}\quad (18.7.3)$$

In this case, the likelihood model results in a complex Gaussian distribution for each relay channel, as follows

$$p(\mathbf{y}^{(l)}|\mathbf{s}, G^{(l)} = g^{(l)}, H^{(l)} = h^{(l)}, \mathbf{W}^{(l)} = \mathbf{W}^{(l)}) = \mathcal{CN}\left(\mathbf{f}^{(l)}\left(\mathbf{s}h^{(l)} + \mathbf{W}^{(l)}\right)g^{(l)}, \sigma_v^2\mathbf{I}\right). \quad (18.7.4)$$

However, clearly this is impossible to evaluate in a real system, since oracular knowledge is not available.

18.8 Results

Here we compare the performance of joint channel estimation and detection under the ABC relay methodology versus the auxiliary MCMC approach for different model configurations. Additionally, we compare the detection performance of the ABC relay methodology, the auxiliary MCMC approach, the Optimal Oracle MAP sequence detector and the SES-ZF MAP sequence detector.

The following specifications are used for all the MCMC algorithms presented: the length of the Markov chain is 20000; the burn in period is 5000; for each MCMC algorithm, the proposals for each of the parameters were tuned off-line to ensure the average acceptance probability post burn-in was in the range of 0.3 to 0.5.

The following specifications for the relay system model are used for all the simulations performed: the symbols are taken from a constellation which is 4-PAM at constellation points in $\{-3, -1, 1, 3\}$; each sequence contains two symbols, $K = 2$; the prior for the sequence of symbols is $Pr((s_1, s_2) = [1, 1]) = Pr((s_1, s_2) = [-1, 1]) = 0.3$ otherwise, all other symbols are equiprobable with probability 0.2; the partial CSI is given by $\sigma_g^2 = \sigma_h^2 = 0.1$; the nonlinear relay function is given by $f(\cdot) = \tanh(\cdot)$. These system parameters were utilised as they allow us to perform the zero forcing solution, with out a prohibitive computational burden.

18.8.1 Analysis of mixing and convergence of MCMC-ABC methodology

In this section analysis of Algorithm 1 (MCMC-ABC) with Euclidean distance metric and Hard Decision weighting function is performed. The impact that the ABC tolerance level has on estimation performance of channel coefficients and the mixing properties of the Markov chain is studied. The study involves joint estimation of channel coefficients and transmitted symbols at an SNR level of 15dB, with $L = 5$ relays present in the system. To empirically monitor the mixing of the Markov chain we consider the Auto-correlation function at different lags.

Note, though results are not presented here, we also performed analysis for all aspects of Algorithm 1 under the setting in which the relay function is linear. We confirmed the MMSE estimates of the channel coefficients and the MAP sequence detector results were accurate for

a range of SNR values. The results presented next are for a much more challenging setting in which the relay is highly non-linear, given by a hyperbolic tangent function.

In Figure 18.9.2 we present a study of the Autocorrelation function of the Markov chains for the channel estimations of G_1 and H_1 as a function of the tolerance ϵ , and the associated estimated marginal posterior distributions $p(g_1|\mathbf{y})$ and $p(h_1|\mathbf{y})$.

For large ϵ the Markov chain mixes over the posterior support efficiently, since when the tolerance is large, the HD weighting function and therefore the acceptance probability, will regularly be non-zero. In addition, with a large tolerance the posterior almost exactly recovers the prior given by the partial CSI. This is expected since, under this setting a large tolerance results in weak contribution from the likelihood weighting. As the tolerance ϵ decreases, the posterior distribution precision increases and there is a translation from the prior partial CSI channel estimates to the posterior distribution over the actual true generated channel coefficients for the given frame of data communications.

It is evident that the mixing properties of the MCMC-ABC algorithm are impacted by the choice of tolerance level. As the tolerance decreases, hence resulting in more accurate posterior distributions in an ABC setting, one should take care to monitor the mixing performance of the MCMC-ABC algorithm. Clearly, the ACF tail decay rate is significantly slower as the tolerance reduces. In this paper we will combat this problem with a soft decision Gaussian weighting function.

18.8.2 Analysis of ABC model specifications

Section two compares the four possible choices we present for the MCMC-ABC algorithm: HD weighting and scaled Euclidean distance; HD weighting and Mahalanobis distance; SD weighting and scaled Euclidean distance; and SD weighting and Mahalanobis distance. The analysis particularly focuses on the mixing properties of the resulting Markov chains under each ABC setting, again as a function of tolerance. In particular it studies the estimated Autocorrelation Function (ACF) of the Markov chains of each of the channel coefficients G and H . The SNR level was 15dB, with $L = 5$ relays present in the system. Note, when presenting the results, since the channels are assumed independent, it is suitable to present the results for one channel, as they will be reflective for the other channels.

To perform this analysis we note that an equivalent posterior accuracy or precision in the ABC posterior estimate should be obtained under each algorithm. Since, the weighting and distance functions are different, this will result in different ϵ values for each choice in the MCMC-ABC algorithm. Therefore, the comparison of the mixing performance and accuracy of each algorithm is established as follows:

1. Take a minimum base epsilon value, $\epsilon_b = 0.2$ and run the MCMC-ABC with soft decision Gaussian weighting and Mahalanobis distance for 100,000 simulations. Ensure the average acceptance rate is between $[0.1, 0.3]$. This will produce an empirical cdf which will act as

the baseline comparison estimate.

2. Consider a set of tolerance values $\epsilon_i = [0.25, 0.5, 0.75, 1]$ and run the MCMC-ABC with soft decision Gaussian weighting and Mahalanobis distance. Tune the random walk proposal standard deviation, σ_i , for each tolerance to obtain average acceptance rates, post burnin, in the interval $[0.1, 0.3]$.
3. Fix the random walk standard deviation $\sigma_{b(i)}$ and run the remaining choices of the MCMC-ABC algorithm, for each tolerance.
4. Record the estimated Kolmogorov-Smirnov test statistic between the ECDF of the baseline ECDF for each algorithm choice.
5. Repeat the above simulation for 20 independently generated data sets.

In Figure 18.9.3 we present a study of the K-S statistic for the first channel, downlink estimated cdf for G_1 , averaged over 20 independent data realisations. The K-S statistic is calculated between the empirical cdf estimated for different tolerances versus the baseline estimated posterior cdf at the baseline tolerance. This plot demonstrates that the algorithm producing the most accurate results involve the use of soft decision and Mahalanobis distance. Clearly, the worst performance involves the hard decision and scaled Euclidean distance, where at low tolerances the average K-S statistic is maximum since the algorithm was not mixing. This demonstrates that such low tolerances under this setting of the MCMC-ABC algorithm will produce poor performance, relative to the soft decision with Mahalanobis distance.

18.8.3 Comparisons of detector performance

In this section we present the symbol error rate (SER) results for the MCMC-ABC with settings selected based on results from the previous two sections, i.e. SD weighting and Mahalanobis distance; the MCMC-AV detector algorithm; the SES-ZF and Oracle detectors. This will involve systematically studying the SER as a function of the following system parameters: number of relays, $L \in \{1, 2, 5, 10\}$; and SNR $\in \{0, 5, 10, 15, 20, 25, 30\}$.

The results of this analysis are presented in Figures 18.9.4 and 18.9.5. The summary of the findings from these results is that they demonstrate clearly that under our proposed system model and detection algorithms, spatial diversity stemming from an increasing number of relays results in measurable improvements in the SER performance. For example Figure 18.9.4 demonstrates that for $L = 1$, there is an insignificant difference between the results obtained for algorithms MCMC-ABC, MCMC-AV and SES-ZF. However, as L increases SES-ZF has the worst performance and degrades relative to the MCMC based approaches. Clearly this demonstrates the utility obtained by developing a more sophisticated detector algorithm. It is clear that the SES-ZF suffers from an error-floor effect: as the SNR increases the SER is almost constant for SNR values above 15dB.

Finally, in Figure 18.9.4 the comparison between the two MCMC based approaches demonstrates that for small L , the performance is comparable. However, as L increases, in the high SNR region, the difference in performance between the MCMC-AV and MCMC-ABC algorithms increases, and as demonstrated by Figure 18.9.5 is statistically significant. This is evident in the separation of the two SER versus SNR curves in figure 18.9.4 as a function of L . This could be due to the significant increase in parameters that must be estimated in the auxiliary based approach as L increases. In particular we note that adding an additional relay introduces K additional nuisance parameters into the auxiliary model posterior. Clearly, this will result in a curse of dimensionality in terms of the number of parameters evident in the model under the MCMC-AV approach.

18.9 Conclusions

In this paper, we proposed a novel cooperative relay system model and then obtained novel detector algorithms. In particular, this involved an approximated-MAP sequence detector for a coherent multiple relay system, where the relay processing function is non-linear. Using the ABC methodology we were able to perform "likelihood free" inference on the parameters of interest. Simulation results validate the effectiveness of the proposed scheme. In addition to the ABC approach, we developed an alternative exact novel algorithm, MCMC-AV, based on auxiliary variables. Finally, we developed a sub-optimal zero forcing solution. We then studied the performance of each algorithm under different settings of our relay system model, including the size of the network and the noise level present. Future research includes the design of detection algorithms for relay systems with partial CSI in which the relay system topology may contain multiple hops on a given channel, or the relay network topology may be unknown. This can include aspects such as an unknown number of relay channels or hops per relay channel.

Acknowledgements

Author Gareth Peters thanks Prof. Arnaud Doucet and Dr. Mark Briers for discussion on aspects of this work whilst at SAMSI. The first author also thanks the Department of Mathematics and Statistics at the University of NSW for support through an Australian Postgraduate Award and to CSIRO for support through a postgraduate research top up scholarship. Finally, this material was based upon work partially supported by the National Science Foundation under Grant DMS-0635449 to the Statistical and Applied Mathematical Sciences Institute, North Carolina, USA. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Authors, Scott Sisson and Yanan Fan were supported by the Australian Research Council through the Discovery Project Scheme (DP0664970).

Algorithm: MAP sequence detection algorithm using MCMC-ABC**1. Initialize Markov chain state:**

2. Initialize $t=0$, $\mathbf{S}^{(0)} \sim p(\mathbf{s})$, $g_{1:L}^{[0]} = \hat{g}_{1:L}$, $h_{1:L}^{[0]} = \hat{h}_{1:L}$

3. **FOR** $t = 1, \dots, T$

(a) **Propose new Markov chain state:** Θ^* **given** $\Theta^{[t-1]}$.

(b) Draw an index $i \sim U[1, \dots, K + 2L]$

(c) Draw proposal $\Theta^* = [\theta_{1:i-1}^{[t-1]}, \theta^*, \theta_{i+1:K+2L}^{[t-1]}]$ from proposal distribution $q(\theta_i^{[t-1]} \rightarrow \theta^*)$.

(Note the proposal will depend on which element of the Θ vector is being sampled.)

(d) **ABC posterior:**

(e) Generate auxiliary variables $x_1^{(l)}, \dots, x_K^{(l)}$ from the model, $p(\mathbf{x}^{(l)} | \theta^*)$, for $l = 1, \dots, L$, to obtain a realization of $\mathbf{X} = \mathbf{x} = [\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(L)}]^\top$ by:

(i) Sample $\mathbf{W}^{(l)*} \sim \mathcal{CN}(\mathbf{0}, \sigma_w^2 \mathbf{I})$, $l \in \{1, \dots, L\}$.

(ii) Sample $\mathbf{V}^{(l)*} \sim \mathcal{CN}(\mathbf{0}, \sigma_v^2 \mathbf{I})$, $l \in \{1, \dots, L\}$.

(iii) Evaluate $\mathbf{X}^{(l)*} = \mathbf{f}^{(l)}(\mathbf{S}^* h^{(l)*} + \mathbf{W}^{(l)*}) g^{(l)*} + \mathbf{V}^{(l)*}$, $l \in \{1, \dots, L\}$.

(f) Calculate a measure of distance $\rho(D(\mathbf{y}), D(\mathbf{x}))$

(g) Evaluate the acceptance probability

$$\alpha(\Theta^{[t-1]}, \Theta^*) = \min \left\{ 1, \frac{p_{ABC}(\Theta^* | \mathbf{y}, \epsilon_t) q(\Theta^* \rightarrow \Theta^{[t-1]})}{p_{ABC}(\Theta^{[t-1]} | \mathbf{y}, \epsilon_{t-1}) q(\Theta^{[t-1]} \rightarrow \Theta^*)} \right\},$$

where $p_{ABC}(\Theta^* | \mathbf{y}, \epsilon_t)$, depending whether HD or SD is used, is given by:

$$\text{HD: } p_{ABC}(\Theta^* | \mathbf{y}, \epsilon_t) \propto \begin{cases} p(s_{1:K}^*) p(h_{1:L}^*) p(g_{1:L}^*), & \text{if } \rho(D(\mathbf{y}), D(\mathbf{x}^*)) \leq \epsilon_t, \\ 0, & \text{otherwise;} \end{cases}$$

$$\text{SD: } p_{ABC}(\Theta^* | \mathbf{y}, \epsilon_t) \propto \exp\left(-\frac{\rho(D(\mathbf{y}), D(\mathbf{x}^*))}{\epsilon_t^2}\right) p(s_{1:K}^*) p(h_{1:L}^*) p(g_{1:L}^*).$$

(h) Sample random variate u , where $U \sim U[0, 1]$.

(i) **IF** $u \leq \alpha(\Theta^{[t-1]}, \Theta^*)$

$\Theta^{[t]} = \Theta^*$

ELSE

$\Theta^{[t]} = \Theta^{[t-1]}$.

ENDIF

4. ENDFOR

Algorithm: MAP sequence detection algorithm using AV-MCMC

1. Initialize Markov chain state:

2. Initialize $t=0$, $\mathbf{S}^{(0)} \sim p(\mathbf{s})$, $g_{1:L}^{[0]} = \hat{g}_{1:L}$, $h_{1:L}^{[0]} = \hat{h}_{1:L}$, $\mathbf{W}^{(0)} \sim p(\mathbf{w})$

3. **FOR** $t = 1, \dots, T$

(a) **Propose new Markov chain state:** Θ^* **given** $\Theta^{[t-1]}$.

(b) Draw an index $i \sim U[1, \dots, K + 2L + KL]$

(c) Draw proposal $\Theta^* = [\theta_{1:i-1}^{[t-1]}, \theta^*, \theta_{i+1:K+2L+KL}^{[t-1]}]$ from proposal distribution $q(\theta_i^{[t-1]} \rightarrow \theta^*)$.

(Note, the proposal will depend on which element of the Θ vector is being sampled.)

(d) Evaluate the acceptance probability

$$\alpha(\Theta^{[t-1]}, \Theta^*) = \min \left\{ 1, \frac{p(\Theta^* | \mathbf{y}) q(\Theta^* \rightarrow \Theta^{[t-1]})}{p(\Theta^{[t-1]} | \mathbf{y}) q(\Theta^{[t-1]} \rightarrow \Theta^*)} \right\}.$$

(e) Sample random variate u , where $U \sim U[0, 1]$.

(f) **IF** $u \leq \alpha(\Theta^{[t-1]}, \Theta^*)$

$\Theta^{[t]} = \Theta^*$

ELSE

$\Theta^{[t]} = \Theta^{[t-1]}$.

ENDIF

4. ENDFOR

Algorithm: Simulation setup

1. Initialize $\text{DataIndex} = 0$

2. Generate each of the true channel coefficients $\{g^{(l)}\}_{l=1}^L, \{h^{(l)}\}_{l=1}^L$

3. Generate transmitted symbols at the source $\mathbf{s}_{1:K}$.

4. Generate observations at the receiver $\{\mathbf{y}^{(l)}\}_{l=1}^L$

5. Estimate covariance matrix for Mahalanobis distance.
6. Run MAP sequence detection algorithm (MCMC-ABC, auxiliary MCMC, ZF) .
7. Count symbol misdetections.
8. $\text{DataIndex} = \text{DataIndex} + 1$.
9. **if** $\text{DataIndex} < \text{NumberSymbols}$ **then** go to 2, **else** quit.

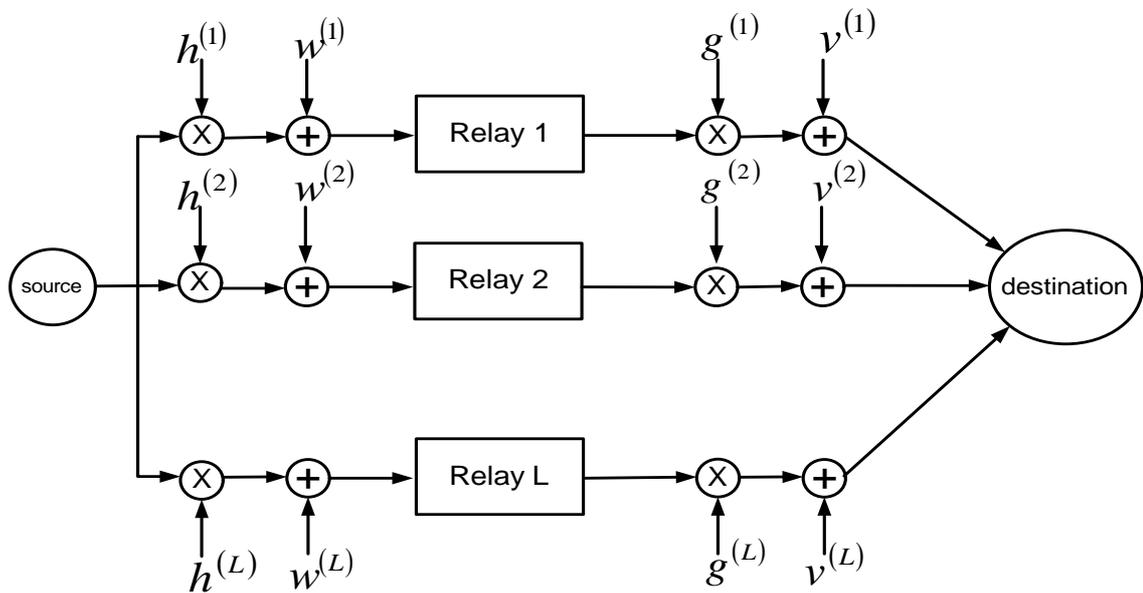


Fig. 18.9.1: The system model studied in this paper involves transmission from the source, through each of the L relay channels on the uplink to the relay, denoted by $h^{(l)}$. Additive complex Gaussian noise is included at the receiver of the relay then the signal is processed and retransmitted by the relay to the destination. In this process the signal is transmitted again through L channels denoted by $g^{(l)}$ and additive complex Gaussian noise is included at the destination.

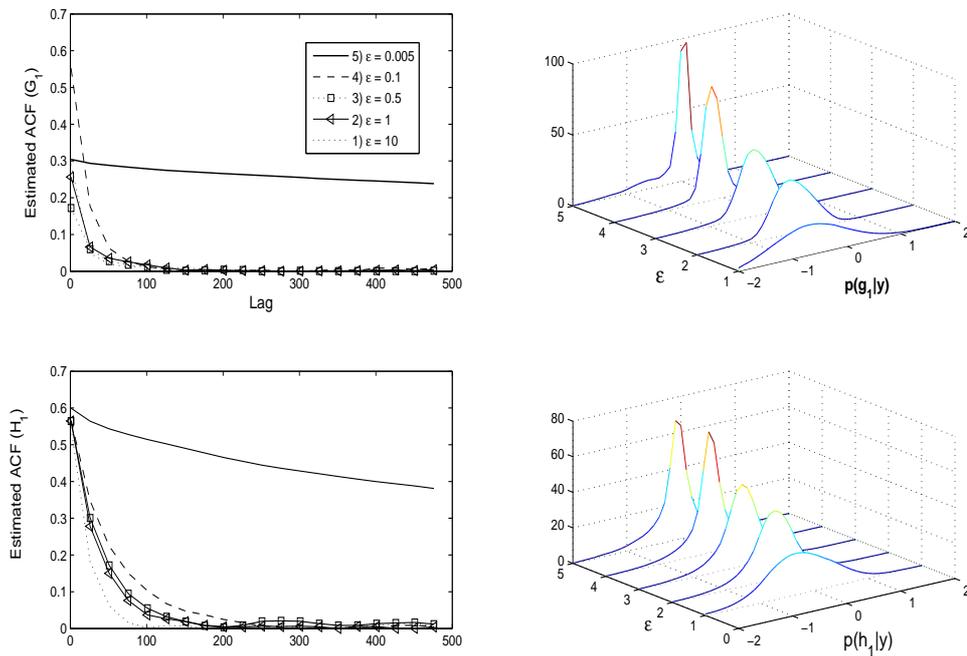


Fig. 18.9.2: Comparison of performance for MCMC-ABC with Hard Decision weighting and Scaled Euclidean distance metric. Subplots on the left of the image display how the estimated ACF changes as a function of tolerance level ϵ for the estimated channels on the up and down links of the relay system. Subplots on the right of the image display a sequence of smoothed marginal posterior distribution estimates for the 1st channel in both the up and down links of the relay, as the tolerance decreases. Note, the indexing of each marginal distribution with labels 1,2,... corresponds to the tolerance given in the legend on the left hand plots.

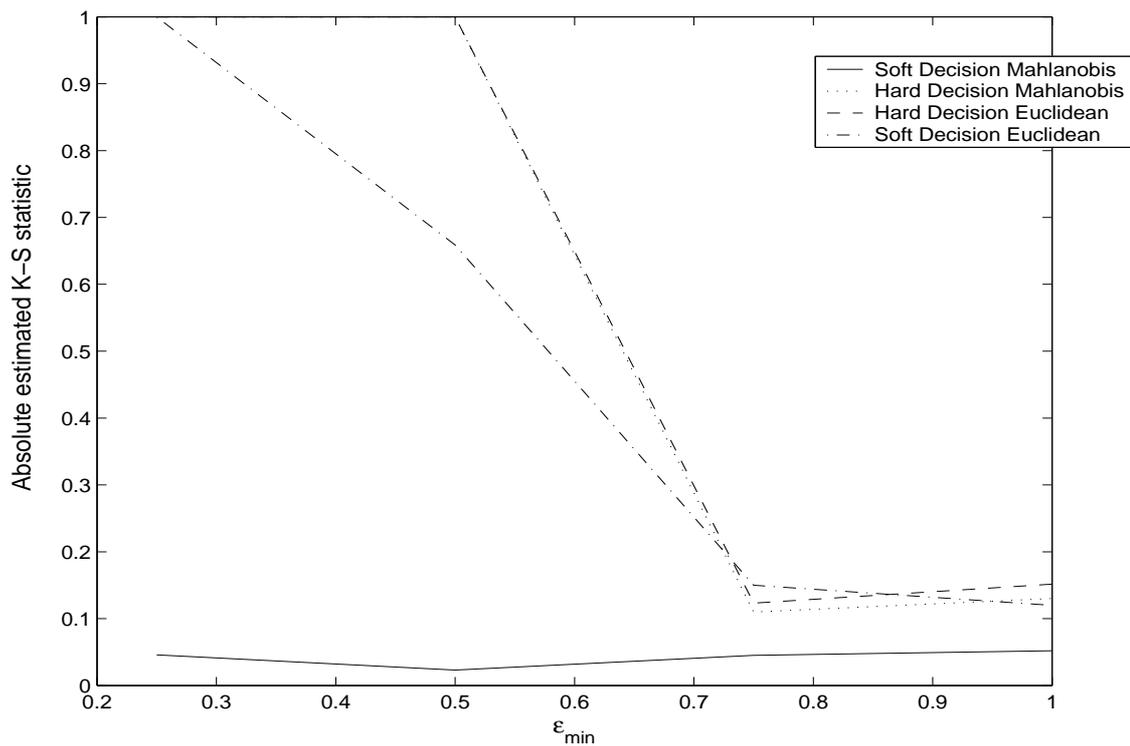


Fig. 18.9.3: These plots present the K-S statistic for the first channel, downlink estimated cdf for G_1 , averaged over 20 independent data realisations. The K-S statistic is calculated between the empirical cdf estimated for different tolerances versus the baseline estimated posterior cdf at the baseline tolerance.

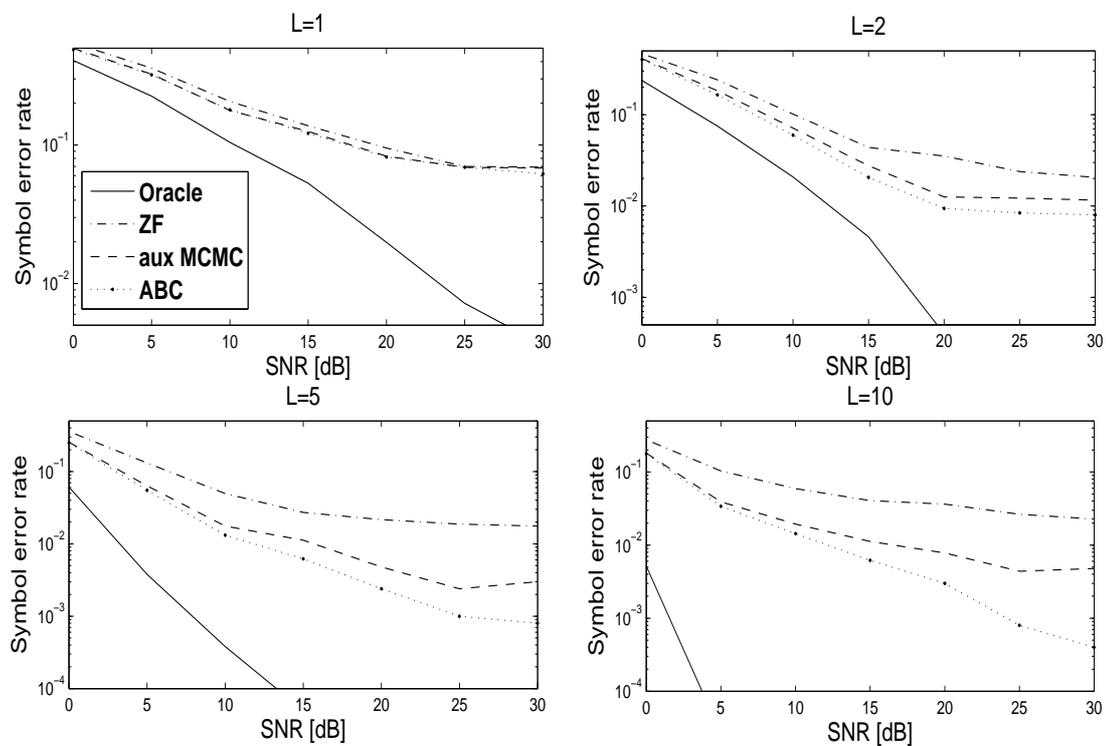


Fig. 18.9.4: These plots demonstrate the performance of each of the proposed detector schemes as a function of the number of relay links, L . For each relay set up the SER is reported as a function of the SNR.

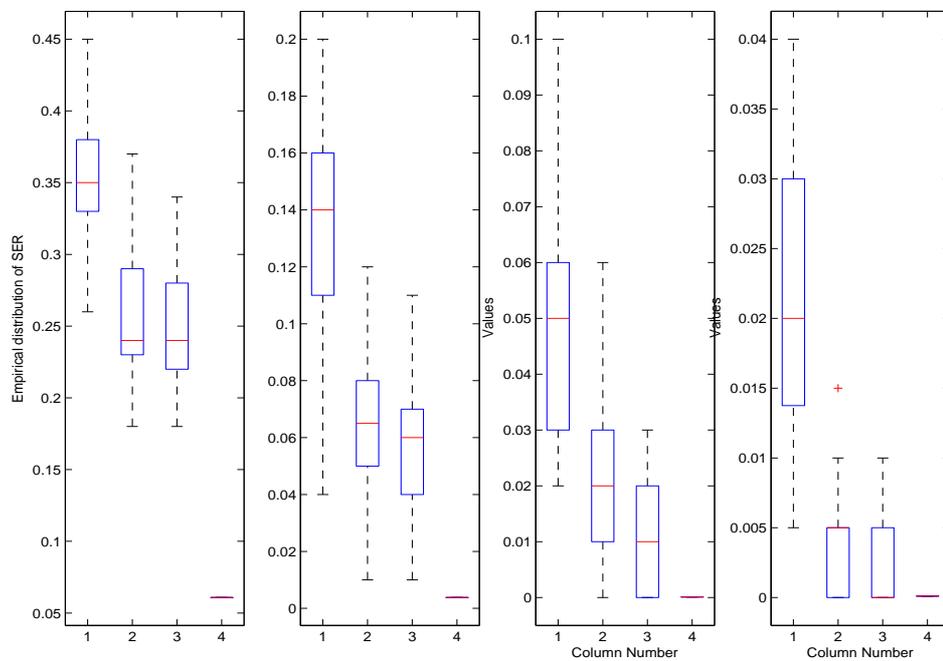


Fig. 18.9.5: Empirical distribution of the SER again as a function of SNR for a relay network with $L =$.

References

- [1] Bortot, P., Coles, S. and Sisson, S. (2007). "Inference for stereological extremes". *Journal of the American Statistical Association*, 102, 84-92.
- [2] Butler, A., Glasbey, C., Allcroft, D. and Wanless, S. (2007). "A latent Gaussian model for compositional data with structural zeroes preparation". *Technical report*, Biomathematics and Statistics, Scotland, Edinburgh.
- [3] Chen, D. and Laneman, J. (2006). "Modulation and Demodulation for Cooperative Diversity in Wireless Systems". *IEEE Transactions on Wireless Communications*, INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, 5, 1785.
- [4] Cover, T. and Gamal, A. (1979). "Capacity theorems for the relay channel". *IEEE Transactions on Information Theory*, 25, 572-584.
- [5] Gomadam, K. and Jafar, S. (2006). "Optimal Relay Functionality for SNR Maximization in Memoryless Relay Networks" Arxiv preprint cs.IT/0510002.
- [6] Hamilton, G., Currat, M., Ray, N., Heckel, G., Beaumont, M. and Excoffier, L. (2005). "Bayesian estimation of recent migration rates after a spatial expansion". *Genetics*, 170, 409-417.
- [7] Khojastepour, M., Sabharwal, A. and Aazhang, B. (2003). "On the capacity of cheap relay networks". *37th Annual Conf. Information Sciences and Systems*
- [8] Kramer, G., Gastpar, M., and Gupta, P. (2005). "Cooperative Strategies and Capacity Theorems for Relay Channels". *IEEE Transactions on Information Theory*, 51, 3037-3063.
- [9] Laneman, J., Tse, D. and Wornell, G. (2004). "Cooperative diversity in wireless networks: Efficient protocols and outage behavior". *IEEE Transactions on Information Theory*, 50, 3062-3080.
- [10] Laneman, J. and Wornell, G. (2003). "Distributed space-time-coded protocols for exploiting cooperative diversity in wireless networks". *IEEE Transactions on Information Theory*, 49, 2415-2425.
- [11] Luciani, F., Sisson, S.A., Jiang, H., Francis, A.R. and Tanaka, M.M. (2009). "The epidemiological fitness cost of drug resistance in mycobacterium tuberculosis: Bayesian analysis of molecular data". *Technical report*, University of New South Wales.

- [12] Marjoram, P., Molitor, J., Plagnol, V. and Tavaré, S. (2003). "Markov chain Monte Carlo without likelihoods". *Proceedings of the National Academy of Science, USA*, 100, 15324-15328.
- [13] van der Meulen, E. (1971). "Three-terminal communication channels". *Advances in Applied Probability*, JSTOR, 3, 120-154.
- [14] Nevat, I., Peters, G.W. and Yuan, J. (2009). "Coherent Detection for Cooperative Networks with Arbitrary Relay Functions using 'Likelihood Free' Inference". *Proc. NEWCOM-ACorn Workshop*, Barcelona, Spain.
- [15] Nosratinia, A., Hunter, T. and Hedayat, A. (2004). "Cooperative communication in wireless networks". *IEEE Communications Magazine*, 42, 74-80.
- [16] Peters, G. and Sisson, S. (2006). "Bayesian inference, Monte Carlo sampling and operational risk". *Journal of Operational Risk*, 1, 27-50.
- [17] Peters, G., Wüthrich, M. and Shevchenko, P. (2009). "Chain Ladder Method: Bayesian Bootstrap versus Classical Bootstrap". *Preprint - UNSW statistics*, UNSW statistics.
- [18] Ratmann, O., Jorgensen, O., Hinkley, T., Stumpf, M., Richardson, S. and Wiuf, C. (2007). "Using likelihood-free inference to compare evolutionary dynamics of the protein networks of *H. pylori* and *P. falciparum*". *PLoS Comput Biol*, 3, e230.
- [19] Ratmann, O., Andrieu, C., Hinkley, T., Wiuf, C. and Richardson, S. (2008). "Model criticism based on likelihood-free inference, with an example in protein network evolution". *Technical report*, Imperial College London.
- [20] Sisson, S., Fan, Y. and Tanaka, M. (2007). "Sequential Monte Carlo without likelihoods". *Proceedings of the National Academy of Sciences*, National Acad Sciences, 104, 1760.
- [21] Sisson, S., Peters, G., Fan, Y. and Briers, M. (2008). "Likelihood-free samplers". *Preprint - UNSW statistics*, UNSW statistics Technical report,
- [22] Tanaka, M.M., Francis, A.R., Luciani, F. and Sisson, S.A. (2006). "Using Approximate Bayesian Computation to estimate Tuberculosis transmission parameters from genotype data". *Genetics*, 173, 1511-1520.
- [23] Toni, T., Welch, D., Strelkowa, N., Ipsen, A. and Stumpf, M.P.H. (2008). "Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems". *Journal of the Royal Society Interface*, In press.
- [24] Wilkinson, R. and Tavaré, S. (2008). "Approximate Bayesian Computation: a simulation based approach to inference".

Summary and future work

“Science never solves a problem without creating ten more.”

Bernard Shaw

19.1 Conclusions

Broadly, this dissertation has addressed three statistical challenges, encapsulated by the questions:

- How can one develop practically useful Bayesian models and corresponding computationally efficient sampling methodology, when the likelihood model is intractable?
- How can one develop methodology in order to automate Markov chain Monte Carlo sampling approaches to efficiently explore the support of a posterior distribution, defined across multiple Bayesian statistical models?
- How can these sophisticated Bayesian modelling frameworks and sampling methodologies be utilized to solve practically relevant and important problems in the research fields of financial risk modelling and telecommunications engineering?

In addressing these questions this thesis was divided into three parts. Part I comprised a primarily statistical development of methodology and theory. Part II then extends and develops the concepts and methodology created in Part I, for advanced statistical modelling of financial risk and non-life insurance. In doing so it clearly demonstrates the impact such methodology can have in extending the models and analysis available to tackle real world financial problems. Part III develops novel statistical models for a range of problems encountered in wireless communications electrical engineering. It achieves this by developing and extending methodology developed in Part I to the setting of communications engineering, again demonstrating

the impact such advanced statistical methodology can have on a real world class of engineering problems.

19.1.1 Outcomes of Part I - Advances in approximate Bayesian computation and trans-dimensional Markov chain Monte Carlo

Part I involved the following specific outcomes:

Likelihood-free Methodology

1. A framework was developed in which one can interpret the posterior distribution of the likelihood-free methodology. This involved making explicit the likelihood-free approximation to the target posterior in both the joint and marginal space representations. As a result a detailed exploration of the properties of these approaches methodologically, numerically and theoretically was undertaken. This clarified and cleared up some ambiguity in the literature on what the target posterior distribution model is in likelihood-free methodology.
2. Several extensions to algorithmic and methodological aspects of the likelihood-free framework including distance metrics, summary statistics, convergence diagnostics, non-parametric bootstrap for simulation from the intractable likelihood and convergence diagnostics were developed.
3. A novel class of SMC Samplers algorithms with PRC stage included in the mutation kernel, termed SMC Samplers PRC was developed.
 - (a) Demonstrated the theoretical properties and validity of this class of SMC Sampler PRC algorithm. This included studying the asymptotic properties of a PRC stage in the SMC Sampler algorithm when compared to the same algorithm without rejection control.
 - (b) Extended the SMC Samplers PRC algorithm to the likelihood-free context and demonstrated properties of this algorithm. Including providing useful guide lines around setting parameters in the ABC methodology and PRC methodology related to the tolerance sequence and rejection control threshold.
 - (c) Demonstrated the numerical properties of this algorithm on several challenging applications such as univariate and multivariate Bayesian models for α -stable settings.
 - (d) Explored and discussed links between existing alternative sampling algorithms in the ABC context, including ABC-SMC, MCMC-ABC and PMC-ABC algorithms.

Trans-dimensional Markov chain Monte Carlo Methodology

1. Developed several versions of a Conditional Path Space proposal mechanism for a Trans-dimensional MCMC algorithm.

- (a) Several possible construction mechanisms for the CPS proposal were explored and demonstrated on actual data examples.
 - (b) Progress was made in the direction of automation of the between model move construction for TD-MCMC algorithms. This is achieved by automation of the process of constructing a proposal that approximates the optimal proposal for between model moves, i.e. the proposal maximizing the between model Markov chain acceptance probability.
 - (c) It was shown that depending on the proposal construction, it could be computational to construct a CPS proposal relative to simpler deterministic construction moves such as RJ-MCMC split and merge type moves. However the proposal under the CPS construction was highly flexible and can be applied generally with good performance as demonstrated in several actual data examples in both Part I and Part III journal papers.
 - (d) The CPS proposal was also demonstrated to be useful as a diagnostic tool for assessment of simpler less computational proposal mechanisms based on for example RJ-MCMC deterministic mappings.
2. The CPS proposal was compared to several other proposal mechanisms. These include the RJ-MCMC proposal constructions based on birth-and-death and split-and-merge type moves constructed via basic moment matching, in an important statistical application of mixture distributions. Additionally, a time series model was analyzed and the CPS proposal was compared to an alternative probabilistic proposal scheme based on Gaussian approximations centered on Maximum-Likelihood estimates. The third comparison involved performance assessment of the CPS proposal relative to a Stochastic Approximation based methodology used in the Contour Monte Carlo framework. In all cases, the CPS proposal was more computational but the efficiency gain measured by the improvement in the acceptance rate of between model moves was significant.
3. Several real world models studied including, Bayesian mixture models in which the number of mixture components and parameters is unknown; Bayesian model for an Autoregressive process in which the lag is unknown in addition to the AR coefficients and the innovation noise variance; Channel Estimation in a Wireless Communications problem, in which the number of channel taps is unknown as well as the channel characteristics.

19.1.2 Outcomes of Part II - Advances in Bayesian financial risk and non-life insurance models

Part II involved the following specific outcomes:

Advanced Statistical Models and Methodology for Operational Risk

1. Developed the first paper introducing Bayesian model selection methodology to the area of Operational Risk modelling. This is particularly relevant in the setting of operational risk as expert opinion is a critical component of all operational risk models, as specified by banking regulation.
2. Demonstrated how to work with Bayesian models for the popular Operational Risk class for severity distribution models based on the g-and-h distribution and the GB2 distribution. These classes of distribution were found to be particularly relevant to Operational Risk modelling when empirical studies were performed. However, these distributions do not admit densities and can only be expressed as transformation expressions of simple Gaussian and exponential random variables. As a result a novel likelihood-free methodology was developed to work with these models. Several methodologies were presented including, for the first time in this literature, an annealing of the tolerance level during burn-in and a Simulated Annealing version of the likelihood-free methodology for MAP parameter estimation.
3. A novel and highly efficient methodology for estimation of the compound process annual loss distribution in Operational Risk models was developed. This included a significant improvement in efficiency relative to the standard Monte Carlo approaches and the Inverse Fourier Transform approaches. This was then demonstrated for the important risk modelling calculations of the risk measures for Value-at-Risk and Expected Shortfall in the setting of Operational Risk.
4. Novel recursions based on those of Panjer were developed for estimation of a general compound process. This was achieved by re-interpreting the Panjer recursion as a Volterra integral equation of the second kind. I then realized that this allowed it to be reformulated as an infinite sum of recursive integrals. This could then be solved by a sophisticated importance sampling methodology. However, the design of the importance sampling distribution is critical to reducing the variance of estimates under this Monte Carlo sampler, so two approaches were developed:
 - Approach 1: Birth-Death path space Importance Sampling
 - Approach 2: Optimal Importance Sampling distribution construction to minimize the variance of the path space importance weights, via a Birth-and-Death Trans-dimensional MCMC algorithm.
5. Compared performance of the algorithms developed to standard Monte Carlo approaches and a basic Edgeworth expansion estimate on several challenging compound process models.

6. Developed a novel dynamic Operational Risk model in the Loss Distributional Approach framework.
 - (a) This involved developing a stochastic model for each individual business and risk type specified in Basel II regulations. Each risk process had latent stochastic risk profiles for both the frequency and severity models.
 - (b) Then correlation structures were introduced between the latent risk profiles over time via a copula model. Gaussian and Archimedian copula models were developed. In particular the Gumbel and Clayton copulas were considered as they provide upper and lower tail dependence.
 - (c) Properties of this model were studied such as the degree of induced correlation between the annual loss compound process random variables when different correlation structures were specified between the latent processes.
 - (d) In introducing the latent processes and copula correlation structures, the risk profile and copula parameters were treated as unknown random variables, termed risk characteristics. A novel Bayesian model was constructed for this model.
 - (e) An estimation framework was developed and demonstrated to accurately estimate this model via slice sampling. Several detailed numerical studies were performed for different models in the framework developed.

Advanced Statistical Models and Methodology for non-life Insurance Claims Reserving

1. Developed novel models and sampling methodology for the problem of non-life insurance claims reserving. This is a family of actuarial problems for predicting the cumulative claims in a run-off triangle structure. The traditional approach is prediction via the chain ladder model, which is a deterministic algorithm used in practice by actuaries. Under such a framework the lack of a stochastic model to describe the estimation and prediction results in an inability to statistically quantify the distribution of the predictive claims, the mean square error of prediction and other risk measures such as Value-at-Risk.
2. Developed two novel stochastic Bayesian models to address the problem of claims reserving that capture the properties of the basic chain ladder factor model:

Model 1: Bayesian formulation of the Poisson-Tweedie model, particularly concentrating on the over dispersed Poisson model. This model captures salient features of the claims reserving problem under the chain ladder model, whilst providing a stochastic model for the estimation of predictive distribution and associated prediction error (MSEP).

- (a) In this class of models, formulation of a Bayesian model is complicated non-trivially by the inability to simply evaluate the likelihood pointwise. We demonstrate a novel, numerically efficient approach to overcome this problem in a Bayesian sampling framework.
- (b) We then consider two important questions in this class of models related to variable selection and model selection. We introduce to the literature the problem of

model selection and parameterization of Poisson-Tweedie model in the claims reserving setting. We then demonstrate how to robustly perform model selection in the Bayesian framework developed for this problem, via numerical sampling methodology developed to tackle the problem of sampling from each of the possible posterior model subspaces.

- (c) As a result, we demonstrate for the claims reserving setting a new framework for consideration of the impact associated with incorporation of model uncertainty into the predictive performance and mean square error of prediction of cumulative claims.
- (d) We study the properties of the proposed model on synthetic and actual data, assessing model parsimony and comparing performance to credibility and Maximum-Likelihood based estimates.

Model 2: Bayesian formulation of the stochastic model for claims reserving via a Distribution Free Chain Ladder, time series model. This model is designed to asymptotically capture the properties of the classical chain ladder model, including the standard predictions for the chain ladder factors.

- (a) We demonstrate how to develop a novel likelihood-free methodology to numerically work with estimation of the empirical predictive claims distribution. As a result we avoid entirely the contentious and currently debated issue of appropriate choice of a parametric distributional assumption typically required to perform estimation of the predictive distribution. We only require the standard conditions on the first and second moments of the DFCL model.
- (b) To achieve this outcome required formulation of a novel likelihood-free methodology. The challenges typically associated with designing a likelihood-free model were compounded by the fact that simulation from the likelihood model was not possible. As a result we formulated a novel non-parametric conditional bootstrap procedure in a Bayesian context, where given parameters we could obtain a synthetic set of data.
- (c) We embedded this into our likelihood-free methodology and also assessed the several novel distance metrics in the likelihood-free framework.
- (d) Simulations were run on both synthetic and actual data sets and results obtained were compared to frequentist bootstrap predictions of the MSE and also credibility theory based estimates.

19.1.3 Outcomes of Part III - Advances in Bayesian models for telecommunications engineering

Part III involved the following specific outcomes:

1. The statistical problems tackled in the journal papers and conference papers of Part III fall into two categories: the design of robust techniques for OFDM receiver design and the

design of data detection in various MIMO systems with different levels of knowledge of CSI. This involved incorporation of both Trans-dimensional MCMC and likelihood-free methodology from the outcomes obtained in Part I. The following areas covered involve:

- (a) Channel estimation for OFDM systems with unknown channel model order and PDP decay rate.
 - (b) Detection of transmitted symbols stemming from non-uniform constellations in MIMO systems with imperfect CSI at the receiver.
 - (c) Wireless relay network design, detection and estimation.
2. Detection of Gaussian constellations in MIMO systems has been devoted to the design of algorithms for detection of Gaussian like constellations in MIMO systems when only partial CSI is given at the receiver. The problem was first formulated as a non convex and non-linear optimization problem. Using the framework of hidden convexity we were able to transform this problem into a tractable convex problem that could be solved efficiently. We have presented a competing approach based on the BEM methodology that achieved comparable performances. We then extended the problem to the case where the noise variance is also unknown. Using the concept of Gibbs annealing we were able to find a solution with close to optimal performance.
 3. We developed a Bayesian inference approach for a model involving a random Gaussian vector in a linear model containing a random Gaussian design matrix for which only the first and second moments are known. We proposed an efficient method to finding the MAP estimator for this model and analyzed its complexity.
 4. Channel estimation in OFDM systems has dealt with the problem of channel estimation in OFDM systems with unknown model order and PDP decay rate. We formulated the problem under the Bayesian framework and using the TDMCMC methodology we were able to construct three different algorithms to sample from the intractable posterior distribution. We analyzed the sensitivity of these algorithms to different choices of prior distributions and various SNR values. Simulation results demonstrate that these algorithms can perform close to the case where the model order and PDP decay rate are known.
 5. We developed a cooperative wireless relay network model in which imperfect knowledge of the channel state information at the destination node is assumed. This general framework incorporated multiple relay nodes operating under arbitrary processing functions.
 6. We demonstrated the complications associated with detection and estimation in such models. In general we show that there will be an intractable likelihood which would render both the maximum likelihood and the maximum *a posteriori* decision rules not analytic.
 7. We developed a Bayesian likelihood-free model and three simulation approaches for maximum *a posteriori* sequence detection in these general networks. These include a Markov

chain Monte Carlo approximate Bayesian computation (MCMC-ABC) approach; an auxiliary variable MCMC (MCMC-AV) approach; and a suboptimal exhaustive search zero forcing (SES-ZF) approach.

19.2 Future work

There exist many possibilities for future work that may extend the results obtained in this dissertation. The extensions and future development directions are again presented according to the three research parts in which the thesis is presented. Starting with the methodology pertaining to the likelihood-free based ideas, the extensions possible for the research in Part I involve first developing further the sampling methodology of the SMC Samplers PRC-ABC algorithm. This would involve automating the design of the schedule of tolerance levels whilst ensuring the rejection rate under the PRC steps is controlled. Secondly, one could explore the development of an auxiliary particle based SMC Sampler PRC algorithm and then determine its utility in the likelihood-free setting in reducing the computational cost associated with the SMC Samplers PRC-ABC algorithm. Additionally, a detailed theoretical and practical study could be performed to understand in greater detail the pros and cons of working in either the marginal or joint likelihood-free model formulations. Another idea of interest in this regard is development of a Particle Markov chain Monte Carlo sampler in the setting of likelihood-free models.

In terms of other advances to the likelihood-free methodology possible, it would be interesting to explore model selection in the context of likelihood-free modelling. This would include Bayes-Factor analysis and TD-MCMC based samplers for the likelihood-free setting. I may also include the development of interacting particle filter formulations for model selection in this context. Additionally, further exploration and development of methods to automate and select appropriate summary statistics is required to be developed. Finally, I would also like to explore the possibilities of the bootstrap type ideas for generating data in the likelihood-free context, for settings in which the model is not tractable (evaluation pointwise) and can not be directly simulated from.

When it comes to developing new models it would be interesting to extend the class of multivariate alpha-stable models to allow for more flexible multivariate distributions such as meta-distributions with stable marginals and dependence defined via a general copula family. In such cases the joint density for the likelihood would still not admit a closed form which could be evaluated pointwise so likelihood-free methodology would be required. The class of multivariate models here would be more flexible as it would not require all marginal distributions to have the same tail index parameter as was the case for the multivariate stable models considered. Other extensions to this model would include developing further a numerical approach to obtaining a sparse grid (defined via a mapping from a sparse grid in Euclidean space to a S^{d-1} sphere in order to perform the projections required for the spectral mass model in the likelihood-free setting).

Extensions to the TD-MCMC methodology would include extending the CPS proposal type ideas to the setting of the Particle Markov chain Monte Carlo algorithm. This would allow for sampling from joint model space for competing state space models. In this regard, one could conceive of running a filter for the proposal mechanism of each of the time series models and moving between these time series model subspaces via an automated proposal construction such as that used in the CPS construction. Another extension possible would involve extending the CPS proposal mechanism into the setting of the likelihood-free model framework. This would automate the between model exploration of likelihood-free models.

When it comes to the extensions possible for the operational risk and insurance models, there are several possibilities. Starting with the Operational Risk setting it is interesting to extend these models to include insurance premiums. This would involve defining relevant insurance policies and to assess their impact on the model formulation, simulation methodology and estimation of risk metrics for differing insurance policies. Note this would likely have important model implications with respect to the introduction of dependence structures in the postulated LDA models. In addition, it would be interesting to explore the extension of the efficient simulation methodology developed for Panjer recursions to the multivariate setting involving Sundt recursions. This would involve development of sophisticated Importance Sampling frameworks as was the case in the univariate recursions. Having this multivariate extension could then also allow for analysis of dependence structures as was the case in the models proposed in the journal papers in Part II.

In terms of some of the extensions possible in the insurance models, the first thing would be to extend the bootstrap based likelihood-free methodology to compare the conditional and unconditional time series approaches. In addition it would be important to further explore the implications such approaches would have with respect to suitable summary statistics for the claims reserving likelihood-free Bayesian model. Further extensions of relevance would be to consider multivariate panel data structures in which several claims triangles are considered with correlation between the loss events in each claims triangle introduced via a copula model. This would provide a challenge to sample and develop such a Bayesian model for the parameterization, estimation and prediction of chain ladder factors and unknown correlation parameters. This could be formulated in the over-dispersed Poisson model or in the likelihood-free Distribution Free Chain Ladder formulations. Finally, it would also be interesting to develop the SMC Samplers PRC methodology for the non-linear state-space formulation of the claims reserving problem. This has been considered for particular claims reserving models and scenarios via a Kalman Filter and an Extended Kalman filter, the SMC Sampler PRC formulation would generalize these approaches.

In terms of the extensions and future work possible for the wireless communications methodology in Part III, one could again consider several aspects relating to modelling and sampling methodology. Concerning the MIMO models, it would be interesting to extend the solutions obtained for the detection methodology to the case in which only partial CSI is given at the receiver. Additionally, it would be beneficial to find a reduced complexity algorithms for very

high modulation schemes, e.g. 64 QAM. That would include finding ways to reduce the number of groups visited in an efficient manner. Concerning the OFDM receiver design as noted the major drawback of the algorithms presented related to the computational complexity. It would be beneficial to add an initial stage in which a coarse estimate of the model order and unknown PDP decay rate value is obtained. Based on these estimates a smaller joint space would need to be explored, thus reducing the length of the Markov chain needed to be run, reducing the overall complexity. From the perspective of extending the model it could be interesting to build a dynamic model to include variations in model order and PDP decay rate. That would correspond to a time varying physical environment.

Finally, when it comes to the wireless relay network models. It would be interesting to extend these models to a state space model formulation under an OFDM framework. One could then perform joint channel tracking and estimation via a Particle Markov chain Monte Carlo algorithm, which could incorporate aspects of the SMC Sampler PRC algorithm. Additionally, one could tackle the model selection and estimation question for wireless relay networks in which it is not known which relay nodes are operational. This is practical in many communications settings such as military settings in which some relay links are destroyed or are transmitting misleading information. It would then be important to detect the appropriate model as well as detecting the transmitted information and possibly performing the channel estimation.

Overall, writing this last chapter I hope it is clear that there are several new and exciting avenues for research involving both methodological and practical model design questions. With each new discovery comes a large number of new and exciting questions which must be addressed and answered, the research in this thesis is no different. I will now be my goal over the next few years to start to answer these questions whilst also posing new and exciting related questions. As is always the case, it is now a matter of time, thought and hard work.

References

- [1] Andrieu C., Doucet A., Berthelsen K.K. and Roberts G.O. (2009). "The expected auxiliary method for Monte Carlo simulation". *Technical report, statistics department UBC*.
- [2] "APRA Prudential Guide APG 115". (October 2006). *Advanced measurement approaches to Operational Risk*.
- [3] "APRA Prudential Standard APS 114". (January 2007). *Capital adequacy: standardised approach to Operational Risk*. Draft Report.
- [4] "APRA Prudential Standard APS 115". (October 2006). *Capital adequacy: advanced measurement approaches to Operational Risk*. Draft Report.
- [5] Artzner P., Delbaen F., Eber J. and Heath D. (2000). "Thinking coherently, pp77 to 82." *Extremes and Integrated Risk Management*, Risk Books, London.
- [6] Arulampalam S., Maskell S., Gordon N. and Clapp T. (2002). "A tutorial on particle filters for on-line nonlinear /non-Gaussian Bayesian tracking". *IEEE Transactions on Signal Process*, **50**(2), 174 to 188.
- [7] "Basel committee on banking supervision, Basel II: international convergence of capital measurement and capital standards: a revised framework". (June 2006). *Comprehensive Version, Bank for International Settlements*. Available from <http://www.bis.org/publ/bcbs128.htm>
- [8] "Basel committee on banking supervision, sound practices for the management and supervision of Operational Risk". (February 2003). *Bank for International Settlements*.
- [9] Bayes T. (1763). "An essay towards solving a problem with the doctrine of Chances". *Philosophical Transactions of the Royal Society London*, **53**, 370 to 418.
- [10] Beaumont M. A., W. Zhang and D. J. Balding. (2002). "Approximate Bayesian computation in population genetics". *Genetics*, **162**, 2025 to 2035.
- [11] Beaumont M. A., Cornuet J.M., Marin J.M., and Robert C.P. (2009). "Adaptive approximate Bayesian computation". *Biometrika*, to appear.
- [12] Bee M. (2006). "Estimating and simulating loss distributions with incomplete data". *Oprisk and Compliance*, **7**(7), 38 to 41.

- [13] Blum M.G.B. and Francois O. (2009). "Nonlinear regression models for approximate Bayesian computation". *Statistics and Computing*, to appear. arXiv:0809.4178.
- [14] Blum M.G.B. (2009) "Approximate Bayesian computation: a non-parameteric perspective". arXiv:0904.0635.
- [15] Bocker K. and Kluppelberg C. (2005). "Multivariate models for Operational Risk". *Hypo Vereinsbank*.
- [16] Bocker K. and Kluppelberg C. (2005). "Operational Var: a closed form approximation". *Hypo Vereinsbank*.
- [17] Bortot P., Coles S.G. and Sisson S.A. (2007). "Inference for stereological extremes". *Journal of the American Statistical Association*. **102**, 84 to 92.
- [18] Brooks S.P., Guidici P. and Roberts G.O. (2003). "Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions". *Journal of the Royal Statistical Society, Series B*, **65**, 3 to 39.
- [19] Butler A., Glasbey C., Allcroft A. and Wanless S. (2006). "A latent Gaussian model for compositional data with structural zeroes". *Technical report*. Biomathematics and Statistics Scotland, Edinburgh.
- [20] Cappe O., Guillin A., Marin J.M. and Robert C.P. (2004). "Population Monte Carlo". *Journal of Computational and Graphical Statistics*, **13**, 907 to 929.
- [21] Cappe O., Godsill S.J. and Moulines, E. (2007). "An overview of existing methods and recent advances in sequential Monte Carlo". *IEEE Proceedings*, **95**(5), 899 to 924.
- [22] Carlin B. and Chib S. (1995). "Bayesian model choice through Markov chain Monte Carlo". *Journal of the Royal Statistical Society, Series B*, **57**, 473 to 484.
- [23] Chang R. and Gibby R. (1968). "A theoretical study of performance of an orthogonal multiplexing data transmission scheme". *IEEE Transactions on Communications, [legacy, pre-1988]*, **16**(4), 529 to 540.
- [24] Cipra B.A. (2000). "The best of the 20th century: editors name top 10 algorithms". *SIAM News*, **33**(4).
- [25] Cornuet J.M., Santos F., Beaumont M.A., Robert C.P., Marin J.M., Balding D.A., Guillemaud T. and Estoup A. (2008). "Inferring population history with DIY ABC: a user-friendly approach to Approximate Bayesian Computation". *Bioinformatics*, **24**(23), 2713 to 2719.
- [26] Cornebise J., Moulines E. and Olsson J. (2008). "Adaptive methods for sequential importance sampling with application to state space models". *Proceedings of 16th European Signal Processing Conference (EUSIPCO)*, August 25 to 29.
- [27] Cornebise J. and Peters G.W. (2009). "Comments on 'Particle Markov Chain Monte Carlo'". *Statistical and Applied Mathematics Sciences Institute (SAMSI), Technical Report 2009 to 7*.
- [28] Cruz M. (2002). *Modelling, Measuring and Hedging Operational Risk*. John Wiley & Sons, Chapter 4.

- [29] Del Moral P. (2004). *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications*. Series: Probability and Applications, Springer-Verlag, New York.
- [30] Del Moral P., Doucet A. and Jasra A. (2006). "Sequential Monte Carlo Samplers". *Journal of Royal Statistical Society, Series B*, **68**, 411 to 436.
- [31] Del Moral P., Doucet A. and Peters G.W. (2007). "Sharp propagation of chaos estimates for Feynman-Kac particle models". *SIAM theory of probability and its applications*, **51**, 459 to 485.
- [32] Del Moral P., Doucet A. and Jasra A. (2008). "An adaptive sequential Monte Carlo method for Approximate Bayesian Computation". *Preprint via Sequential Monte Carlo Methods Homepage*.
- [33] Dohler M. and Aghvami H. (2006). "A step towards MIMO: virtual antenna arrays". In *Cooperation in wireless networks: principles and applications*. [Ed.] Fitzek F.H.P and Katz M.D.K., Springer Netherlands.
- [34] Doucet A., de Freitas N. and Gordon N.J. (2001). *An introduction to sequential Monte Carlo methods in practice*. Springer-Verlag, New York.
- [35] Doucet A. and Johansen A.M. (2007). "A Tutorial on particle filtering and smoothing: fifteen years later". In *Handbook of Nonlinear Filtering (eds. D. Crisan et B. Rozovsky)*, Oxford University Press, to appear.
- [36] Dutta K. and J. Perry (2006). "A tale of tails: an empirical analysis of loss distribution models for estimating operational risk capital". *Federal Reserve Bank of Boston, Working Papers* No. 06 to 13.
- [37] Egan B. (APRA May 2005), *Basel II Changes and Operational Risk*, Speech given at the Enterprise-Wide Risk Management Conference. Available from http://www.apra.gov.au/speeches/05_04.cfm
- [38] Embrechts P., H. Furrer and R. Kaufmann (2003). "Quantifying regulatory capital for operational risk". *Derivatives Use, Trading & Regulation*, **9**(3), 217 to 223.
- [39] Embrechts P., McNeil A. and Rudiger F. (2005). *Quantitative Risk Management, Techniques and Tools*, Princeton Series in Finance.
- [40] Embrechts P., Degen M. and Lambrigger D. (2006). "The quantitative modelling of operational risk: between g-and-h and EVT". *Astin Bulletin* **37**(2), 265 to 291
- [41] Estoup A., Wilson I.J., Sullivan C., Cornuet J.-M., and Moritz C. (2002). "Inferring population history from microsatellite and enzyme data in serially introduced cane toads, *Bufo marinus*". *Genetics*, **159**, 1671 to 168.
- [42] Estoup A., Beaumont M., Sennedot F., Moritz C. and Cornuet J.M. (2004). "Genetic analysis of complex demographic scenarios: spatially expanding populations of the cane toad, *bufo marinus*". *Evolution*, **58**, 2021 to 2036.
- [43] Fearnhead P. (2008). "Computational methods for complex stochastic systems: A review of some alternatives to MCMC". *Statistics and Computing*, **18**, 151 to 171.

- [44] Feldman M.W., Kumm J., and Pritchard J.K. (1999). *Mutation and migration in models of microsatellite evolution*. In "Microsatellites: Evolution and Applications", DG Goldstein and C Schloetterer, eds., pp 98 to 115. Oxford University Press, Oxford.
- [45] Foschini G.J. and Gans M.J. (1998). "On limits of wireless communications in a fading environment when using multiple antennas". *Wireless Personal Communications*, **6**, 311 to 335.
- [46] Franklin J., Sisson S.A., Peters G.W. and Terauds V. (2006). *Assessment of strategies for evaluating extreme risks*. Supplementary report "Quantifying Bank Operational Risk", Australian Centre for Excellence in Risk Analysis (ACERA) Project No. 0602.
- [47] Fu Y.X. and Li W.H. (1997). "Estimating the age of the common ancestor of a sample of DNA sequences". *Molecular Biology and Evolution*, textbf14, 195 to 199.
- [48] Garthwaite P. and O'Hagan A. (2000). "Quantifying expert opinion in the UK water industry: an experimental study". *The Statistician*, **49**(4), 455 to 477.
- [49] Gelfand A. and Dey D. (1994). "Bayesian model choice: asymptotics and exact calculations". *Journal of the Royal Statistical Society, Series B*, **56**, 501-514.
- [50] Gelman A. and Rubin D.B. (1992). "Inference from iterative simulation using multiple sequences". *Statistical Science*, **4**, 457 to 472.
- [51] Geweke J.F. (1991). *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*. In J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith (eds.) *Bayesian Statistics*, 4, Oxford University Press, Oxford.
- [52] Godsill S. J. (2001). "On the relationship between Markov chain Monte Carlo methods for model uncertainty". *Journal of Computational and Graphical Statistics*, **10**, 1 to 19.
- [53] Godsill S.J. (2003). In P. J. Green, N. L. Hjort, and S. Richardson (Eds.), *Highly Structured Stochastic Systems*, Chapter Discussion of Trans-dimensional Markov chain Monte Carlo by P. J. Green, pp. 199 to 203. Oxford University Press.
- [54] Godsill S.J. and Vermaak J. (2004). "Models and algorithms for tracking using trans-dimensional sequential Monte Carlo". In *Proceedings of IEEE ICASSP*.
- [55] Gordon N.J., Salmond D.J. and Smith A.F.M. (1993). "Novel approach to nonlinear/non-Gaussian Bayesian state estimation". *IEE-Proceedings-F*, **140**, 107 to 113.
- [56] Gourieroux C., Monfort A. and Renault E. (1993). "Indirect inference". *Journal of Applied Econometrics*, **8**, 85 to 118.
- [57] Gramacy R.B., Samworth R.J. and King R. (2008). "Importance tempering". *Preprint*, arXiv: 0707.4242v5 [stat.Co].
- [58] Green P. (1995). "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination". *Biometrika*, **82**, 711 to 732.
- [59] Green P. (2003). *Trans-dimensional Markov chain Monte Carlo*. Chapter for the book "Highly structured stochastic systems", with discussion by Simon Godsill and Juha Heikkinen, Oxford University Press.

- [60] Grelaud A., Marin J.M. and Robert C.P. (2009). "ABC methods for model choice in Gibbs random fields". *Notes aux Comptes Rendus de l'Academie des Sciences* **347**(3 to 4), 205 to 210.
- [61] Grenander U. and Miller M.I. (1994). "Representations of knowledge in complex systems". *Journal of the Royal Statistical Society, Series B*, **56**, 549-603.
- [62] Hamilton G., Currat M., Ray N., Heckel G., Beaumont M. and Excoffier L. (2005). "Bayesian estimation of recent migration rates after a spatial expansion". *Genetics*, **170**, 409 to 417.
- [63] Haykin S. (2001). *Communication Systems*. Fourth Edition, Wiley and Sons.
- [64] Jasra A., Stephens D.A. and Holmes C.C. (2007). "Population-based reversible jump MCMC". *Biometrika*, **94**(4), 787 to 807.
- [65] Jasra A., Doucet A., Stephens D.A. and Holmes C.C. (2008). "Interacting SMC samplers for trans-dimensional simulation". *Computational Statistics and Data Analysis*, **52**(4), 1765 to 1791.
- [66] Johansen A.M., Del Moral P. and Doucet A. (2006). "Sequential Monte Carlo samplers for rare events". In *Proceedings of the 6th International Workshop on Rare Event Simulation*, 256 to 267, Bamberg, Germany.
- [67] Joyce P. and Marjoram P. (2008). "Approximately sufficient statistics and Bayesian computation". *Stat. Appl. Genet. Molec. Biol.*, **7**, 23.
- [68] KPMG (2005). "Financial services Basel II: a closer look - managing Operational Risk". *white paper from Advisory*.
- [69] Kitagawa G. (1996). "Monte Carlo filter and smoother for non-Gaussian non-linear state space models". *Journal of Computational and Graphical Statistics*, **5**, 1 to 25.
- [70] Kueck H., de Freitas N. and Doucet A. (2006). "SMC Samplers for Bayesian optimal non-linear design". *Nonlinear Statistical Signal Processing Workshop (NSSPW)*.
- [71] Kunsch H.R. (2001). *State-space and hidden Markov models*. in *Complex Stochastic Systems* (eds. O.E. Barndorff-Nielsen, D.R. Cox and C. Kluppelberg), CRC Press, 109 to 173.
- [72] Laker J. (APRA April 2006), *Basel II - Observations from Down Under*, Speech given at the Second Annual Conference on the Future of Financial Regulation, London School of Economics. Available from <http://www.apra.gov.au/speeches/BASEL-II-OBSERVATIONS-FROM-DOWN-UNDER.cfm>
- [73] Laneman J.N., Wornell G.W. and Tse D.N.C. (2001). "An efficient protocol for realizing cooperative diversity in wireless networks". *Information Theory, Proceedings IEEE International Symposium*
- [74] Leuenberger C., Wegmann D. and Excoffier L. (2009). "Bayesian computation and model selection in population genetics". *preprint*, arXiv:0901.2231v1 [stat.ME]
- [75] Liu J.S. (2001). *Monte Carlo strategies in scientific computing*. Springer-Verlag, New York

- [76] Luciani F., Sisson S.A., Jiang H., Francis A.R. and Tanaka M.M. (2009). "The epidemiological fitness cost of drug resistance in mycobacterium tuberculosis: Bayesian analysis of molecular data". *Technical report*, University of New South Wales.
- [77] Marjoram P., Molitor J., Plagnol V. and Tavare S. (2003). "Markov chain Monte Carlo without likelihoods". *Proceedings of the National Academy of Science, USA*, **100**, 15324 to 15328.
- [78] Marjoram P. and Tavaré S. (2006). "Modern computational approaches for analysing molecular genetic variation data". *Nat. Rev. Genet.*, **7**, 759 to 770.
- [79] Marlo L. and Nyce C. (2006). "Sarbanes-Oxley section 404 internal controls and actuarial process". *Casualty Actuarial Society Forum*, casact.com
- [80] Moller J., Pettitt A.N., Reeves R. and Berthelsen K.K. (2006). "An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants". *Biometrika*, **93**, 451 to 458.
- [81] Moscadelli M. (2004). "The modelling of Operational Risk: experience with the analysis of the data collected by the Basel committee". *Bank of Italy*, available from <http://ssrn.com/abstract=557214>.
- [82] Nevat I., Peters G.W. and Yuan J. (2009). "Coherent detection for cooperative networks with arbitrary relay functions using likelihood-free inference". *NEWCOM-ACorn Workshop*, Barcelona, Spain.
- [83] O'Hagan A. (1998). "Eliciting expert beliefs in substantial practical applications". *The Statistician*, **47**, 21 to 35.
- [84] O'Hagan A. (2006). *Uncertain Judgements: Eliciting Expert's Probabilities*, Wiley, Statistics in Practice.
- [85] Pabst R., Walke B.H., Schultz D.C., Herhold P., Yanikomeroglu H., Mukherjee S., Viswanathan H., Lott M., Zirwas W., Dohler M. and others. (2004). "Relay-based deployment concepts for wireless and mobile broadband radio". *IEEE Communications Magazine*, **42**(9), 80 to 89.
- [86] Panjer H. (2006). *Operational Risk: Modeling Analytics*, Wiley.
- [87] Peled A. and Ruiz A. (1980). "Frequency domain data transmission using reduced computational complexity algorithms". *IEEE International Conference on ICASSP, Acoustics, Speech and Signal Processing*, **5**, 80.
- [88] Peters G.W. (2005). *Topics in Sequential Monte Carlo Samplers*. University of Cambridge, M.Sc. Thesis, Department of Engineering.
- [89] Peters G.W. and Sisson S.A. (2006). "Bayesian inference, Monte Carlo sampling and Operational Risk". *Journal of Operational Risk*, **1**(3).
- [90] Peters G.W., Fan Y. and Sisson S.A. (2008). "On sequential Monte Carlo, partial rejection control and approximate Bayesian computation". *Technical report*, arXiv:0808.3466v1.
- [91] Peters G.W., Nevat I., Sisson S.A., Fan Y. and Yuan J. (2009). "Bayesian symbol detection in wireless communications". *Technical report*, University of NSW.

- [92] Peters G.W., Wüthrich M. and Shevchenko P. (2009). "Chain Ladder Method: Bayesian bootstrap versus classical bootstrap". *Technical report*, Univesity of NSW.
- [93] Peters G. W., Fan Y. and Sisson S.A. (2009). "Likelihood-free Bayesian inference for α -stable models". *Technical report*, Univesity of NSW.
- [94] Plagnol V. and Tavare S. (2004). *Approximate Bayesian Computation and MCMC*. in Monte Carlo and Quasi-Monte Carlo Methods 2002, ed. H. Niederreiter, Heidelberg, Springer-Verlag, pp. 99 to 114.
- [95] Pritchard J.K., Seielstad M.T., Perez-Lezaun A. and Feldman M.W. (1999). "Population growth of human Y chromosomes: a study of Y chromosome microsatellites". *Mol. Biol. Evol.*, **16**, 1791 to 179.
- [96] Ramamurthy S., Arora H. and Ghosh A. (2005). "Operational risk and probabilistic networks - an application to corporate actions processing". *Infosys White Paper*.
- [97] Ratmann O., Jorgensen O., Hinkley T., Stumpf M., Richardson S. and Wiuf C. (2007). "Using likelihood-free inference to compare evolutionary dynamics of the protien networks of h. pylori and p. falciparum". *PLoS Comp. Biol.*, **3**, e230.
- [98] Ratmann O., Andrieu C., Hinkley T., Wiuf C. and Richardson S. (2009). "Model criticism based on likelihood-free inference, with an example in protein network evolution". *Proceedings of the National Academy of Sciences of the United States of America*, **106**(26), 10576 to 10581.
- [99] Reeves R.W. and Pettitt A.N. (2005). "A theoretical framework for approximate Bayesian computation". *Presented at the International Workshop for Statistical Modelling*, Sydney.
- [100] Robert C.P. and Casella G. (2008). "A history of Markov chain Monte Carlo \hat{U} subjective recollections from incomplete data". eprint arXiv: 0808.2902.
- [101] Ruiz A., Cioffi J.M., Kasturia S., Center, I.B.M.T.J.W.R. and Hawthorne N.Y. (1992). "Discrete multiple tone modulation with coset coding for the spectrally shaped channel". *IEEE Transactions on Communications*, **40**(6), 1012 to 1029.
- [102] SarbanesÓxley Act (2002). *Public Laws 107-204, 116 Statute 745, enacted July 30, 2002*.
- [103] Sendonaris A., Erkip E., Aazhang B., Inc Q. and Campbell C.A. (2003). "User cooperation diversity. Part II. Implementation aspects and performance analysis". *IEEE Transactions on Communications*, **51**(11), 1927 to 1948.
- [104] Shannon C.E. (1949). "Communication in the presence of noise". *Proceedings of the Institute of Radio Engineers*, **37**(1), 10 to 21.
- [105] Shannon C.E. (2001). "A mathematical theory of communication". *ACM SIGMOBILE Mobile Computing and Communications Review*, **5**(1), 3 to 55, ACM New York, NY, USA.
- [106] Shevchenko P. and Wuthrich M. (2006). "The structural modelling of operational risk via Bayesian inference: combining loss data with expert opinions". *CSIRO Technical Report Series*, CMIS Call Number 2371.

- [107] Shevchenko P. (2009). Modelling Operational Risk using Bayesian inference. To appear in Springer Finance Series.
- [108] Sisson S.A. (2005). "Trans-dimensional Markov chains: A decade of progress and future perspectives". *Journal of the American Statistical Association*, **100**, 1077 to 108.
- [109] Sisson S.A. (2007). "Genetics and stochastic simulation do mix!" *The American Statistician*, **61**, 112 to 119.
- [110] Sisson S.A., Fan Y. and Tanaka M.M. (2007). "Sequential Monte Carlo without likelihoods". *Proceedings of the National Academy of Science, USA*, **104**, 1760 to 1765. Errata (2009).
- [111] Sisson S.A., Peters G.W., Fan Y. and Briers M. (2008). "Likelihood-free samplers". *Technical report*, University of NSW.
- [112] Tanaka M.M., Francis A.R., Luciani F. and Sisson S.A. (2006). "Using Approximate Bayesian Computation to estimate Tuberculosis transmission parameters from genotype data". *Genetics*, **173**, 1511 to 1520.
- [113] Tavaré S., Balding D.J., Griffiths R.C. and Donnelly P. (1997). "Inferring coalescence times from DNA sequence data". *Genetics*, **145**, 505 to 518.
- [114] Taylor G. and McGuire G. (2004). "Loss reserving with GLMs: a case study". *Spring 2004 Meeting of the Casualty Actuarial Society*, Colorado Springs, Colorado.
- [115] Telatar E. (1999). "Capacity of multi-antenna Gaussian channels". *European transactions on telecommunications*, **10**(6), 585 to 595, Associazione elettrotecnica ed elettronica italiana
- [116] Tishkoff S.A., Varkonyi R., Cahinhinan N., Abbes S. and Argyropoulos G. (2001). "Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance". *Science*, **293**, 455 to 46.
- [117] Toni T., Welch D., Strelkowa N., Ipsen A. and Stumpf M.P.H. (2009). "Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems". *Journal of the Royal Society Interface*, **6**(31), 187.
- [118] Vermaak J., Godsill S.J. and Doucet A. (2003). "Radial basis function regression using trans-dimensional sequential Monte Carlo". In *IEEE Workshop on Statistical Signal Processing*.
- [119] von Neumann J. (1951). "Various techniques used in connection with random digits". *Applied Math Series*, **12**, 36 to 38.
- [120] Weiss G. and von Haeseler A. (1998). "Inference of population history using a likelihood approach". *Genetics*, **149**, 1539 to 154.
- [121] Wilkinson R.D. (2008). "Approximate Bayesian computation (ABC) gives exact results under the assumption of model error". arXiv:0811.3355v1.
- [122] Weinstein S. and Ebert P. (1971). "Data transmission by Frequency-Division Multiplexing using the discrete Fourier transform". *IEEE Transactions on Communications, [legacy, pre-1988]*, **19**(5 part 1), 628 to 634.

-
- [123] Wüthrich M.V. and Merz M. (2008). *Stochastic claims reserving methods in insurance*. Wiley Finance.