Non-centered parameterisations for data augmentation and hierarchical models

with applications to inference for Lévy-based stochastic volatility models

Omiros Papaspiliopoulos, B.Sc.

Submitted for the degree of Doctor of Philosophy
at Lancaster University, June 2003.

I declare that the work presented in this thesis is my own, except where stated otherwise.

Omiros Papaspiliopoulos
June, 2003

Non-centered parameterisations for data augmentation and hierarchical models

Omiros Papaspiliopoulos, B.Sc.

Submitted for the degree of Doctor of Philosophy
at Lancaster University, June 2003.

Abstract

The main purpose of this thesis is to present, in a unified manner, how non-centered parameterisations can be applied in a wide range of hierarchical models. The aim of such parameterisations is to improve the performance of the Gibbs sampler and related componentwise-updating MCMC algorithms which are used to perform Bayesian inference. The main attraction of the centered and the non-centered methodology is that they both provide a general parameterisation strategy. Thus, the performance of the corresponding MCMC algorithms can be studied in a general way abstracting from the technicalities of a particular model. On the other hand, detailed knowledge about a specific model can be useful in finding parameterisations which are preferable to both the non-centered and the centered. Our partially-non-centered parameterisations try to combine the generality of the centered and non-centered methodology together with the specificity of a particular model to produce better parameterisations.

We review the existing theory for comparing different parameterisations in the context of the Gibbs sampler. We apply this theory to investigate when the non-centered is preferable to the centered parameterisation for a wide range of linear Gaussian models. Qualitative comparison of the two schemes, based on the notions of geometric and uniform ergodicity of Markov chains, is provided for a class of linear non-Gaussian models.

We introduce state space expansion techniques which allow us to construct non-centered parameterisations for a wide range of distributions. We also suggest a variety of techniques for non-centering Poisson processes. We link our work with existing methodologies in Bayesian non-parametrics.

We review a family of stochastic volatility models recently introduced in the literature by Barndorff-Nielsen and Shephard (2001). We proceed to conduct Bayesian inference for these models using centered and non-centered MCMC algorithms, which we have introduced for this purpose. The efficiency of these algorithms is assessed using several simulated datasets. We use our methods to provide estimates of the model parameters for a financial series of Deutsch Mark against US Dollar exchange rates.

We introduce a hierarchical model diagnostic tool and apply it to assess identifiability and fit of the stochastic volatility models.

We introduce partially-non-centered methods that attempt to outperform both the centered and non-centered. Some analytic results are given and applications in random-effects

models, hierarchical generalised linear and geostatistical mixed models are considered. We also provide a discussion on how this method relates to other augmentation methods recently introduced in the literature.

Acknowledgements

The biggest gain of my undergraduate studies in Athens was by far that I met Petros Dellaportas. Petros has been a role model, a very good friend and an exciting collaborator during the last years ($\tau o\ \chi o\nu\tau\rho\upsilon\nu\alpha\ \phi\iota\lambda\epsilon$?). I will always be grateful to him for insisting (and convincing me) firstly that I do a PhD, and secondly that the best person to supervise me is Gareth Roberts. He was correct in both.

Working with Gareth has been a great pleasure and lesson for me. His innovative thinking, his energy and enthusiasm for work, his instinct to recognise which research questions are of real interest and his modesty have shaped my own attitude towards research. Gareth has been extremely supportive throughout my PhD. He tried hard and succeeded in finding a PhD studentship for me, he always managed to provide me with funds to travel to conferences and when my studentship expired he found alternative ways to finance me in order to continue our collaboration. He has been very patient and constructive during my un-inspired periods. For all these and many more I would like to thank him.

I feel lucky that I had the chance to study at the department of Mathematics and Statistics in Lancaster University, where I had the chance to exchange ideas and collaborate with many interesting people. In particular I would like to thank my collaborator Martin Sköld, Søren Fiig Jarner for his suggestions and help during the first year of my PhD, Chris Ferro for many interesting discussions, Paulo Ribeiro, Oswaldo Cruz and Theo Kypraios for their invaluable help with many computing issues, and Alexandros Beskos for studying with me (and often teaching me) Markov chain theory during the last year of my PhD.

In addition to the stimulating research environment, the department is characterised by its unique friendly atmosphere. Among many of the good friends I had here, I would particularly like to thank FabriZZio Laurini for the exciting time we had at Lancaster.

An important part of my education as a statistician was the attendance to numerous conference and meetings. Therefore I would like to thank Neil Shephard, Ole Barndorff-Nielsen, Sylvia Fruhwirth-Shnatter and Friedrich Hubalec for inviting me and providing me with financial support to participate to several meetings regarding Lévy processes. I would also like to thank Christian Robert for his invitation and financial support to give two seminars in Paris.

I would like to thank Martin Sköld, Petros Dellaportas, Alexandros Beskos and Pete Neal for their corrections and suggestions on earlier versions of this thesis.

Finally, I would like to thank Lancaster University, Onasis Foundation, TMR network FMRXCT960095 and EPSRC grant GR/M62723 for financial support during my PhD.

To *Patricia*,
who *"for having the music so deep, she has her heart in her hand"*

Contents

List of Figures ii
List of Tables iii

1 Introduction **1**
1.0 Motivation 1
1.1 Outline of the thesis 2
1.2 Notation 7
1.3 Bayesian inference for missing data problems 8
1.4 Graphical models and conditional independence 9
1.5 Markov chain Monte Carlo methods 10
1.5.1 Basic Markov chain theory 10
1.5.2 MCMC algorithms 12
1.6 Hierarchical models 21
1.7 Centering and non-centering 24
1.8 Basics of Lévy processes and infinite divisibility 25

2 Convergence rates and reparameterisations for the Gibbs sampler on normal hierarchical models **30**
2.0 Introduction 30
2.1 Rates of convergence of the Gibbs sampler 30
2.1.1 Gibbs sampler on Gaussian target distributions 34
2.1.2 Measures of efficiency 36
2.2 Parameterisations of hierarchical models 39
2.3 The normal hierarchical model 39
2.3.1 Brownian motion interpretation 42
2.3.2 Effect of prior distribution on the rate of convergence 44
2.4 A general normal hierarchical model 44
2.5 A State-space model 51
2.6 Linear non-Gaussian model 55

3 Convergence of MCMC for linear hierarchical models with heavy-tailed links **58**
3.0 Introduction . . . 58
3.1 Markov chain theory for general state spaces . . . 59
3.1.1 ϕ-irreducibility and small sets . . . 59
3.1.2 Recurrence and Harris chains . . . 60
3.1.3 The ergodic theorem . . . 61
3.1.4 Uniform ergodicity of Markov chains . . . 62
3.1.5 Geometric ergodicity of Markov chains . . . 63
3.1.6 De-initialising chains . . . 66
3.2 Linear non-Gaussian models and robust Bayesian analysis . . . 67
3.2.1 The Dawid/O'Hagan conditions . . . 68
3.2.2 Our approach . . . 71
3.3 Convergence of the CA and the NCA for the Cauchy-Gaussian model . . . 73
3.4 The general result . . . 75
3.5 The double exponential-double exponential model . . . 76

4 General non-centered parameterisations and state space expansion **79**
4.0 Introduction . . . 79
4.1 General non-centered parameterisations . . . 80
4.2 NCPs for gamma random effect models by expanding the state space . . . 83
4.2.1 Non-centering for infinitely divisible and related distributions . . . 86
4.3 Comparison of different non-centering schemes . . . 87
4.3.1 The stable family . . . 88
4.3.2 Gaussian latent distribution . . . 89
4.3.3 Cauchy latent distribution . . . 89

5 Non-centered parameterisations for Poisson processes **94**
5.0 Introduction . . . 94
5.1 Poisson processes: review of basic definitions and properties . . . 95
5.1.1 Restriction, superposition and mapping properties . . . 96
5.1.2 Sums over Poisson processes . . . 97
5.1.3 Marked Poisson processes . . . 98
5.1.4 Likelihood functions for Poisson and Gibbs processes . . . 99
5.1.5 Birth-death-displacement MCMC algorithms for simulating Gibbs processes . . . 106
5.2 NCPs for Poisson processes . . . 108
5.3 NCP by thinning . . . 109

5.3.1 THIN-NCA for homogeneous Poisson processes on a bounded state space . . . 111
5.3.2 THIN-NCA for finite Poisson processes . . . 113
5.4 NCP by the inverse CDF method . . . 114
5.5 NCPs for marked Poisson processes . . . 118
5.5.1 An illustrative example . . . 120
5.6 Completely random measures and subordinators . . . 124
5.7 Completely random measures . . . 125
5.8 Positive independent increments processes, subordinators and representations 127
5.8.1 The Ferguson-Klass representation and approximations . . . 129
5.8.2 Applications to Bayesian non-parametrics . . . 132

6 Inference for Non-Gaussian OU models 133
6.0 Introduction . . . 133
6.1 Financial markets and stylised facts . . . 133
6.2 Stochastic volatility modelling . . . 137
6.3 The Barndorff-Nielsen and Shephard model . . . 140
6.3.1 Construction of the model . . . 140
6.3.2 Integrated volatility . . . 143
6.3.3 Aggregation results . . . 145
6.3.4 Superposition of OU processes . . . 147
6.3.5 Existing estimation methods . . . 148
6.4 OU models with compound Poisson BDLP . . . 151
6.4.1 Superposition of OU models with compound Poisson BDLP . . . 153
6.5 Bayesian inference for the gamma-OU model . . . 154
6.5.1 Superposition of gamma-OU processes . . . 154
6.5.2 Prior specification and posterior inference . . . 155
6.6 Augmentation based on marked Poisson processes . . . 156
6.7 A centered parameterisation . . . 157
6.7.1 MCMC implementation . . . 158
6.8 Alternatives to the centered parameterisation . . . 163
6.9 Non-centering for the the gamma-OU model . . . 164
6.10 MPP-THIN-NCP for the gamma-OU model . . . 165
6.11 Alternative non-centered parameterisations . . . 168
6.11.1 MCMC implementation . . . 169
6.12 Simulation study . . . 170
6.12.1 Comparison of CA vs NCA . . . 170
6.12.2 Comparison of the different NCPs . . . 174

6.13 Augmentation and non-centered parameterisation for the superposition of OU processes . . . 182
6.13.1 Examples using simulated data . . . 184
6.14 Posterior inference and sensitivity analysis . . . 185
6.15 Model diagnostic tools . . . 189
6.16 A real data example . . . 192
6.17 Extensions and further work . . . 196
6.17.1 Drift and risk premium . . . 196
6.17.2 Leverage effect and non-integrable Lévy measures . . . 198

7 Partially Non-centered parameterisations 200
7.0 Introduction . . . 200
7.1 Partial non-centering of hierarchical models . . . 200
7.2 PNCP for the normal hierarchical model . . . 201
7.3 PNCP for the general normal hierarchical model . . . 204
7.3.1 Full conditional distributions . . . 204
7.4 PNCP with proper priors . . . 205
7.5 PNCP outside the Gaussian context . . . 206
7.6 State-space expanded PNCPs . . . 208
7.7 PNCP and correlation analysis . . . 209
7.8 Reparameterisations for GLHM . . . 210
7.8.1 Natural exponential family with quadratic variance function . . . 211
7.8.2 GLHM . . . 211
7.8.3 Reparameterisation based on posterior correlations . . . 212
7.8.4 Simulation results . . . 213
7.8.5 Higher order PNCP . . . 217
7.9 Conditional and marginal augmentation . . . 218
7.9.1 Conditional augmentation . . . 218
7.9.2 Marginal augmentation . . . 219
7.10 Optimising the PNCP . . . 223
7.11 Examples . . . 225
7.11.1 Spatial GLMM . . . 226

List of Figures

1.1 The conditional independence graph of the two-component Gibbs sampler. . 16
1.2 The conditional independence graph of the two-component Hastings-within-Gibbs sampler, when only one (left) and when both (right) conditionals are not sampled from directly. 20
1.3 The graphical model of the centered parameterisation. 22
1.4 Graphical model of an exchangeable model 23
1.5 Graphical model of a partially exchangeable model 23
1.6 The graphical model of the non-centered parameterisation. The dashed arrows correspond to a deterministic link, that is X is a deterministic function of $\tilde{X}$ and Θ. 25
1.7 A path in $[0, 1]$ of a standard Brownian motion. It has been simulated by discretising time in intervals of length 10^{-3} and simulating from the corresponding increments of the process. 27
1.8 A path in $[0, 1]$ of a $\mathrm{Ga}(10, 1)$ Lévy process. It has been simulated by discretising time in intervals of length 10^{-3} and simulating from the corresponding increments of the process. 28
1.9 A path in $[0, 1]$ of a compound Poisson process with finite rate $\lambda = 10$ and $E_j \sim \mathrm{Ex}(1)$. The path has been simulated without any discretisation error by explicitly simulating the jump times and corresponding sizes from the appropriate distributions. 29

2.1 Updating of X given Y and Θ as a Brownian bridge simulation: simulate a Brownian bridge starting at time 0 from Θ and hitting Y at time $\sigma_y^2 + \sigma_x^2$, obtain X as the value of the bridge at time $\kappa(\sigma_y^2 + \sigma_x^2) = \sigma_x^2$. 43
2.2 Convergence rates results for the state-space model. The first two rows plot $-1/\log(\rho_c)$ and $-1/\log(\rho_{nc})$ against the sample size n for various values of ϕ, for $\kappa = 0.8$ and $\kappa = 0.2$ respectively. The last row shows $(1 - \rho_{nc})/(1 - \rho_c)$ against n together with its asymptotic limit (the horizontal line), for $\phi = 0.1$ (left) and $\phi = 0.95$ (right). 54

2.3 Gibbs sampler output for Θ in the Normal-Cauchy model (2.48), where we have taken $m = 1,\ Y_1 = 51.91,\ \sigma_x^2 = 1$.. Top: centered parameterisation started from $\Theta_0 = 50$ (left) and $\Theta_0 = 500$ (right). Bottom: non-centered parameterisation for the same starting values. All chains were run for 10^4 iterations. Notice the different scales in the plots. 56

3.1 A typical example of a drift function V for a unimodal density π on $\mathbb{R}$. The shaded area consists of pairs (z, y) such that $z \in C$, where C is the small set used in the drift condition. C is typically some compact set around the mode of π. 64

3.2 The conditional distribution of X given $Y = 0$ and $\Theta = 10$ in (3.15) where C and $\tilde{X}$ are double exponential random variables. 77

4.1 Steps to update $\tilde{X}_i$ conditionally on Y_i and Θ. The current value of Θ is denoted by θ_0. (a): Simulate the value of the process at time θ_0. Denote this by $x_i := \tilde{X}_i(\theta_0)$. (b): Given $\tilde{X}_i(\theta_0) = x_i$, simulate forwards in time a gamma process started from x_i at time θ_0. (c): Simulate a beta process started at time 0 from 0 and stopped at time θ_0 to x_i. (d): The new configuration for $\tilde{X}_i$. To produce these figures we have assumed the model $Y_i \sim \mathrm{Ex}(X_i)$, initial values $\theta_0 = 3$ and assumed an observed data point $Y_i = y_i = 0.5$. 85

4.2 Updating of Θ conditionally on $\tilde{X}_i, Y_i$. The step function corresponds to the unnormalised density $\pi(Y_i \mid \tilde{X}_1(\Theta))$ as a function of Θ, where $\tilde{X}_1$ has the configuration shown in Figure 4.1.d. The product of this density and $\pi(\Theta)$ is the target density of this step of the algorithm. The proposed value for Θ is $\theta_1 > \theta_0$ in (a) and $\theta_1 < \theta_0$ in (b). Once a new value θ_1 has been proposed, we simulate the value of $\tilde{X}_i(\theta_1)$ from the prior, namely $\tilde{X}_i(\theta_1) \stackrel{d}{=} \tilde{X}_i(\theta_0) + G,\ G \sim \mathrm{Ga}(\theta_1 - \theta_0, 1)$, if $\theta_1 > \theta_0$, and $\tilde{X}_i(\theta_1) \stackrel{d}{=} B\tilde{X}_i(\theta_0),\ B \sim \mathrm{Be}(\theta_1/\theta_0, (\theta_0 - \theta_1)/\theta_0)$, if $\theta_0 > \theta_1$. The paths in red colour in the plot show these simulations as gamma and beta process simulations respectively, as described in the previous step of the algorithm. Once $\tilde{X}_i(\theta_1)$ has been simulated the acceptance probability of the move from θ_0 to θ_1 given in (4.4) can be computed. To produce these figures we have assumed the model $Y_i \sim \mathrm{Ex}(X_i)$, initial values $\theta_0 = 3$, proposed values $\theta_1 = 5$ in (a) and $\theta_1 = 1$ in (b), and assumed an observed data point $Y_i = y_i = 0.5$. 86

4.3 Simulation results from implementation of the sc-NCA and the lp-NCA for the normal hierarchical model with unknown latent variance Θ. First column shows trace plots for the sc-NCA while the second shows the corresponding trace plots for the lp-NCA. The last column superimposes the ACFs of the sample paths from the two implementations where the solid line corresponds to sc-NCA and the dashed to lp-NCA. Both algorithms have been run for 10^4 iterations, the first 10^3 being discarded for estimation of the ACFs as burn-in (clearly larger burn-in should be used for the lp-NCA in the last row). For the simulation we have taken $m = 200$, $\sigma_y^2 = 1$ and $\Theta = 0.2, 1, 5$ (first, second and third row respectively). An $\text{Ig}(1, 1)$ prior was chosen for Θ. 90

4.4 MCMC trace plots of the implementation of the sc-NCA (left) and the lp-NCA (right) for the normal hierarchical model with unknown latent variance Θ. Data were simulated under the specification $m = 200, \sigma_y^2 = \Theta = 1$. Both algorithms are initialised at $\theta_0 = 100$ and the proposal variances are tuned so that the overall acceptance rate is 0.2-0.3, although similar results were obtained for a variety of scaling schemes. The lp-NCA has a random-walk type of behaviour when started in the tails, characteristic of algorithms which fail to be geometrically ergodic. 91

4.5 (Un-normalised) $\log \pi(\Theta \mid \tilde{X}, Y)$ as a function of Θ for sc-NCP (left) and lp-NCP (right) for the normal hierarchical model (4.6), with $m = 1$ (top), $m = 50$ (middle) and $m = 200$ data simulated from the model using $\Theta = 0.2$ and $\sigma_y = 1$. $\tilde{X}$ has been simulated from $\tilde{X} \mid Y, \Theta = 0.2$ for both algorithms. The simulation has been designed in such way that the transformation of $\tilde{X}$ and $\Theta = 0.2$ leads to the same X in both algorithms. For the lp-NCP 20 realisations of $\tilde{X} \mid Y, \Theta = 0.2$ have been drawn, all resulting to the same X and the corresponding functions $\pi(\Theta \mid \tilde{X}, Y)$ are superimposed. 92

4.6 (Un-normalised) $\log \pi(\Theta \mid \tilde{X}, Y)$ as a function of Θ for sc-NCP (left) and lp-NCP (right) for the normal hierarchical model (4.6), with $m = 1$ (top), $m = 50$ (middle) and $m = 200$ data simulated from the model using $\Theta = 5$ and $\sigma_y = 1$. $\tilde{X}$ has been simulated from $\tilde{X} \mid Y, \Theta = 5$ for both algorithms. The simulation has been designed in such way that the transformation of $\tilde{X}$ and $\Theta = 5$ leads to the same X in both algorithms. For the lp-NCP 20 realisations of $\tilde{X} \mid Y, \Theta = 5$ have been drawn, all resulting to the same X and the corresponding functions $\pi(\Theta \mid \tilde{X}, Y)$ are superimposed. 93

5.1 The relationship between the Poisson process Φ with state space the positive half-line, plotted with asterisks, and the jump function $z(\cdot)$. For the simulation we have taken $S = [0, 100]$ and Φ to be a homogeneous Poisson process with intensity 0.1. 96

5.2 The relationship between the marked Poisson process Ψ, produced by marking a homogeneous Poisson process with intensity 0.1, with independent Ex(1) marks, and the compound Poisson process $z(\cdot)$; for the simulation we have taken $S = [0, 100]$. 100

5.3 Simulation from the Strauss model with $r = 0.1,\ \theta_1 = \log(100)$ and for three different values of θ_2, $\log(0.75)$, $\log(0.1)$, $\log(0.001)$ in top, middle and bottom correspondingly. A configuration of the process is shown in the left column, MCMC samples from the stationary distribution of the sufficient statistics are plotted in the middle and right columns. The birth-and-death MCMC algorithm of Section 5.1.5 was used to obtain the samples from the point processes, run for 10^5 iterations but then thinned every 100 to produce the plots. 105

5.4 Transformation from $\tilde{X}$ to X. X is a Poisson process on S with intensity $\lambda(x;\theta)$, $\tilde{X}$ is a unit intensity Poisson process on $S \times (0, \infty)$. Choose all points of $\tilde{X}$ that lie on $\{(x, z) : x \in S,\ y < \lambda(x;\theta)\}$ (white area in the plot) and project them down to S. The resulting points (denoted by the black asterisks) form X. In this example $S = [0, 30],\ \lambda(x;\theta) = \theta,\ \theta = 0.3$. 110

5.5 Updating of Θ given $\tilde{X}$ and Y in the THIN-NCA. The current value of θ is denoted by θ_0 and the proposed by θ_1, the corresponding quantities for X by $X^{(0)}$ and $X^{(1)}$ respectively. If $\theta_1 < \theta_0$, $X^{(1)}$ is found by removing from $X^{(0)}$ all points $\{x \in X^{(0)} : (x, z) \in \tilde{X} \cap S \times [\theta_1, \theta_0)\}$ (lying in the cyan area). If $\theta_1 > \theta_0$, $X^{(1)}$ is derived from $X^{(0)}$ by the addition of all points $\{x \in S : (x, z) \in \tilde{X} \cap S \times [\theta_0, \theta_1)\}$ (lying in the yellow area). Once $X^{(1)}$ has been found the Metropolis-Hastings acceptance probability (5.13) can be calculated. In this example we have taken $S = [0, 10],\ \theta_0 = 2,\ \theta_1 = 1$ and $\theta_1 = 3$. 113

5.6 The increasing (red) and decreasing (blue) transformation h in the CDF-NCP corresponding to the intensity function (5.21). In this plot we have taken $r = 4, \phi = 1$. 118

5.7 The MPP-THIN-NCP of (Θ, X) for the Poisson process with intensity (5.26). Current values of the parameters are assumed to be $r = 0.1, \phi = 1$ and $T = 100$. $\tilde{X}$ is a Poisson process on $[0, T] \times (0, \infty) \times (0, \infty)$ with mean measure $e^{-\tilde{\epsilon}} dc\ dm\ d\tilde{\epsilon}$; choose all $(C_i, M_i, \tilde{E}_i) \in \tilde{X}$ with $M_i \leq r$ (denoted by circles as opposed to the points with $M_i > r$ denoted by asterisks); project them to S; set $E_i = \tilde{E}_i/\phi$. 122

5.8 The MPP-CDF-NCP of (Θ, X) for the Poisson process with intensity (5.26). Current values of the parameters are assumed to be $r = 0.1, \phi = 1$ and $T = 100$. $\tilde{X}$ is a unit rate Poisson process on S; choose all $(C_i, \tilde{E}_i) \in \tilde{X}$ with $\tilde{E}_i < r$; set $E_i = -\log(\tilde{E}_i/r)/\phi$. 122

5.9 The THIN-NCP of (Θ, X) for the Poisson process with intensity (5.26). Current values of the parameters are assumed to be $r = 0.1, \phi = 1$ and $T = 100$. $\tilde{X}$ is a unit rate Poisson process on $S \times (0, \infty)$ and X consists of all (C_i, E_i) such that $(C_i, E_i, M_i) \in \tilde{X}$ and $M_i < \lambda(C_i, E_i)$. 123

6.1 Series of daily prices and log-returns for the exchange rate of the DM (Deutsch Mark) against the US Dollar. 135

6.2 Series of daily prices and log-returns for the exchange rate of the UK Sterling against the US Dollar. 135

6.3 Series of daily prices and log-returns for the Dow Jones index. 135

6.4 The logarithm of the estimate of the unconditional density of the log-returns using kernel density estimation (blue lines), for the three financial datasets introduced in Section 6.1. We superimpose the log-density of the Gaussian distribution (red lines) which has the same first two moments as the data. . 137

6.5 Sample autocorrelations for the series of daily log-returns (y_n), their absolute values $(|y_n|)$ and their squares (y_n^2) for the German DM-US Dollar exchange rate (left), the UK Sterling-US Dollar exchange rate (middle), and the Dow Jones index (right). 138

6.6 Simulation of the BDLP $z(\cdot)$ (left column)and the corresponding OU process (right column), for the gamma-OU model. Here we have taken $v(t) \sim \text{Ga}(3, 8.5)$ and $\mu = 0.01$. Top panel concentrates on a short time horizon, while the bottom panel shows simulation for a much longer period. 144

6.7 The integrated volatility processes corresponding to the OU processes plotted in Figure 6.6. 146

6.8 Graphical model of the first (centered) parameterisation 158

6.9 The local displacement of the point $\{(c_i, \epsilon_i)\}$. The new jump time c is generated uniformly in (c_{i-1}, c_{i+1}) and the new jump size ϵ is set so that the volatility process changes only between time points c and c_i. The new jump time c can lie either side of c_i; both options are shown in the diagram. 162
6.10 The graphical model of the non-centered parameterisation. 165
6.11 Estimates of the autocorrelation function of the marginal chain corresponding to $\lambda = \nu\mu$ for the centered (solid) and the non-centered (dashed) parameterisations for each of the simulated data sets described in Section 6.12.1. dataset 1 is in the top left corner and the rest of the datasets are placed from left to right. $T = 2000$ in the two bottom right plots, $T = 500$ in the rest. The estimates were calculated after thinning each of the chains one every hundredth and discarding the 500 initial points. 172
6.12 MCMC traces for λ for dataset 1, when all parameters are initialised from their prior means. (a): the centered (converging slowly) and the non-centered (converging rapidly) algorithms described in Appendices 1 and 2 are used. (b): a modification of the centered parameterisation is used, where 100 birth-death updates are performed for each update of the parameters. All chains have been thinned one every hundredth. 173
6.13 MCMC output for the point process Ψ for dataset 1, for the non-centered algorithm. The configuration of the jump times $\{C_i\}$ is plotted against every 1000th iteration for the first 60,000 iterations of the algorithms. The degree of darkness of the points within each configuration reflects their relative jump sizes E_i. In the far right the configuration of the jump times used in the simulation of the data is plotted. The picture is similar for the centered algorithm, although for this example convergence is not reached so quickly. . 174
6.14 Estimates of the autocorrelation function of the marginal chain corresponding to $\lambda = \nu\mu$ for the MPP-THIN-NCP (red), the MPP-CDF-NCP (green) and the THIN-NCP (blue) for simulated dataset 1 (left) and 6 (right). The estimates were calculated after thinning each of the chains one every hundredth and discarding the 500 initial points. Runs of length 1.5×10^6 were used for all algorithms. 175

6.15 Estimates of the autocorrelation function of the marginal chain corresponding to $\lambda = \nu\mu$ for the MPP-CDF-NCP and the THIN-NCP for simulated dataset 6. The green and blue lines are the same as those in Figure 6.14. The two algorithms were re-run, performing 100 updates for Ψ, that is 100 birth-death-displacement steps, for every update of the parameters. The light blue line corresponds to the estimated ACF of λ for the THIN-NCP and the purple line for the MPP-CDF-NCP. The estimates were calculated after thinning each of the chains one every hundredth and discarding the 500 initial points. Runs of length 1.5×10^6 were used for all algorithms. 177

6.16 $\log \pi(X \mid \lambda, \theta_T, \mu_T, \tilde{\Psi})$ as a function of λ for dataset 6 for the MPP-THIN-NCP ((a) and (c)) and the MPP-CDF-NCP ((b) and (d)). The function for a single realisation of $\tilde{\Psi}$ is plotted on the top panel ((a) and (b)). The bottom panel superimposes this function calculated for 100 different realisations of $\tilde{\Psi}$. In both algorithms $\tilde{\Psi}$ is transformed to the same Ψ when the parameters take the values λ_T, θ_T. All draws of $\tilde{\Psi}$ have been simulated from its conditional distribution given the data and the true parameter values. 179

6.17 $\log \pi(X \mid \lambda, \theta_T, \mu_T, \tilde{\Psi})$ as a function of λ for dataset 1 for the MPP-THIN-NCP ((a) and (c)) and the MPP-CDF-NCP ((b) and (d)). The function for a single realisation of $\tilde{\Psi}$ is plotted on the top panel ((a) and (b)). The bottom panel superimposes this function calculated for 100 different realisations of $\tilde{\Psi}$. In both algorithms $\tilde{\Psi}$ is transformed to the same Ψ when the parameters take the values λ_T, θ_T. All draws of $\tilde{\Psi}$ have been simulated from its conditional distribution given the data and the true parameter values. 180

6.18 Estimates of the autocorrelation function of the marginal chain corresponding to $\lambda = \nu\mu$ for the MPP-CDF-NCP and the THIN-NCP for simulated dataset 1. The two algorithms were run performing 100 updates for Ψ for every update of the parameters. The blue line corresponds to the estimated ACF of λ for the THIN-NCP and the green line for the MPP-CDF-NCP. The estimates were calculated after thinning each of the chains one every fiftieth and discarding the 100 initial points. Runs of length 10^5 were used for both algorithms. . . 181

6.19 The MPP-THIN-NCP of $(\Psi_1, \Psi_2, \lambda_1, \lambda_2, \theta)$. Current values of the parameters are assumed to be $\lambda_1 = 0.2, \lambda_2 = 0.1, \theta = 1$ and $T = 100$. $\tilde{\Psi}$ is a Poisson process on $[0, T] \times (-\infty, \infty) \times (0, \infty)$ with mean measure $e^{-\tilde{\epsilon}} dc\, dm\, d\tilde{\epsilon}$; choose all $(C_i, M_i, \tilde{E}_i) \in \tilde{\Psi}$ with $-\lambda_2 < M_i < \lambda_1$ (denoted by circles as opposed to the points with $M_i > \lambda_1$ or $M_i < -\lambda_2$ denoted by asterisks). 183

6.20 Estimates of the autocorrelation function of the marginal chain corresponding to $\lambda_1 = \nu_1\mu_1$ and $\lambda_2 = \nu_2\mu_2$ for the dataset 1 (solid) and the dataset 2 (dashed), described in Table 6.2. The estimates were calculated after thinning each of the chains one every hundredth and discarding the 1000 initial points. 185
6.21 Histograms of the posterior distribution for the parameters ν, θ, μ under dataset 1 (see Table 6.1). The black vertical lines indicate the values of the parameters used in the data simulation. The prior density for each of the parameters is plotted in red. 186
6.22 Histograms of the posterior distribution for the parameters ν, θ, μ under dataset 5 (see Table 6.1). The black vertical lines indicate the values of the parameters used in the data simulation. The prior density for each of the parameters is plotted in red. 186
6.23 Histograms of the posterior distribution for the parameters ν, θ, μ under dataset 6 (see Table 6.1). The black vertical lines indicate the values of the parameters used in the data simulation. The prior density for each of the parameters is plotted in red. 187
6.24 Kernel estimates of the posterior density for the parameters ν, θ, μ under dataset 1 (see Table 6.1), for two different prior specifications: $\nu \sim \text{Ga}(1, 0.1),\ \mu \sim \text{Ga}(1, 1)$ and $\theta \sim \text{Ga}(1, 0.1)$ (light blue) and $\theta \sim \text{Ga}(1, 0.01)$ (dark blue). Notice that the light blue lines are density estimates of the same posterior distribution that is represented by histograms in Figure 6.21. The black vertical lines indicate the values of the parameters used in the data simulation. . . . 188
6.25 Kernel estimates of the posterior density for the parameters ν, μ under dataset 1 (see Table 6.1), for two different prior specifications: $\nu \sim \text{Ga}(1, 0.1),\ \mu \sim \text{Ga}(1, 1)$ (light blue) and $\nu \sim \text{Ga}(1, 3/2),\ \mu \sim \text{Ga}(1, 100/3)$ (dark blue). Notice that the light blue lines are density estimates of the same posterior distribution that is represented by histograms in Figure 6.21. The black vertical lines indicate the values of the parameters used in the data simulation. . . . 189
6.26 Posterior density estimates for the parameters $\nu, w_2, \theta, \mu_2, \mu_1 - \mu_2$ under dataset 1 (see Table 6.2). The black vertical lines indicate the values of the parameters used in the data simulation. The prior density for each of the parameters is plotted in red. 190
6.27 Posterior desnity estimates for the parameters $\nu, w_2, \theta, \mu_2, \mu_1 - \mu_2$ under dataset 2 (see Table 6.2). The black vertical lines indicate the values of the parameters used in the data simulation. The prior density for each of the parameters is plotted in red. 191

6.28 Model diagnostic plots for the single-OU model of Section 6.5: For each draw of $\tilde{C}$ and $\tilde{E}$ from the posterior distribution, we estimate empirically $-\log P[\lambda(c_j - c_{j-1}) > t]$ and $-\log P[\theta\epsilon > t]$. Each jagged line in the figures corresponds to such an estimator. If the model is "true" the estimator has to coincide with the straight 45^o line. Dataset 6 (Table 6.1) was used in the left column. The underlying OU process used in the middle and right columns has been generated under the specification $\lambda = 0.2,\ \mu = 0.1, T = 2000$ but the jump sizes have been generated from a Ga(0.1, 1), instead of an Ex(10) used for Dataset6. Daily data were used in the middle column and 100 data per day in the right column. The posterior means for the parameters (ν, θ, μ) are (1.97,9.86,0.09) (left), (0.59,2.43,0.089) (middle) and (0.68,2.7,0.098) (right). 193

6.29 Results from fitting the OU models to the US$/DM data. In all plots dashed and solid lines corresponds to results obtained under the single-OU and the two-OU models respectively. Top: pointwise posterior median v_n^*: (a) short period of time for the single-OU and the two-OU models. (b) long period of time for the two-OU model; the grey dots correspond to y_n^2. Bottom: (c) diagnostic of Section 6.14 applied to the jump sizes from the single-OU model. (d) kernel estimates of the log-predictive density of $\log(v_n^*)$. (e) posterior median of the ACF of the series $\{v_n^*\}$. 195

7.1 The geometry of the normal hierarchical model. 202

7.2 Rate of convergence $\rho_{pnc}(w)$ for the PNCP on the normal hierarchical model. In this example we have taken $\sigma_x^2 = 1,\ \sigma_y^2 = 3$, thus $\kappa = 1/4$. 203

7.3 Rate of convergence $\rho_{pnc}(w)$ different values of $1/\tau^2 = 0, 1, 2, 5, 10, 100$, when $\sigma_x = 1,\ \kappa = 0.4$. Curves that lie above others correspond to smaller values of $1/\tau^2$. 206

7.4 Different ways to simulate $X \sim \text{Ga}(\Theta, 1)$ using gamma processes. We use the general notation $\tilde{X}_{\alpha,\beta}$ for a gamma process with shape parameter α and scale parameter β. $X = \tilde{X}_{1,1}(\Theta)$ (top left), $X = \tilde{X}_{\Theta,1}(1)$ (top right), $X = \tilde{X}_{\Theta^w,1}(\Theta^{1-w})$ (bottom left), $X = \tilde{X}_{1,1}(w\Theta) + X_2,\ X_2 \sim \text{Ga}((1-w)\Theta, 1)$ (bottom right). In this example we have taken $\Theta = 8$ and $w = 1/2$. 209

7.5 MCMC output for the CA for the GLHM (7.16). We have kept $w = 0.8$ fixed and take $n_0 = 1$ (top), $n_0 = 2$ (middle) and $n_0 = 5$ (bottom). Traces of Θ are drawn on the same scale. In the middle column we draw the likelihood function. 215

7.6 MCMC output for the CA (first and third columns) and the PNCA (second and fourth columns) for the GLHM (7.16). We have kept fixed $n_0 = 1$ and varied $w = 1/1.2$ (top), $w = 1/2$ (middle) and $w = 1/6$ (bottom). Notice that traces are drown on the same scale for both the CA and the PNCA. 215
7.7 ACFs for CA (left) and PNCA (middle and right) when $n_0 = 20$ and $c = 10$, $(w = 0.67)$. The plot on the right corresponds to the PNCA algorithm that performs 30 Metropolis-Hastings steps for every update of the missing data. 216
7.8 ACFs for the CA (left) and the PNCA (middle and right) when $n_0 = 3$ and $c = 0.2$ $(w = 0.9375)$. The plot on the right corresponds to the PNCA algorithm that performs 300 Metropolis-Hastings steps for every update of the missing data. 216
7.9 Contour plot of the estimated joint distribution of $\theta, S^w = \sum_i \tilde{X}_i^{(w)}$ when $n_0 = 3$ and $c = 0.2$ $(w = 0.9375)$. 217
7.10 ACFs for the CA (left), the PNCA (middle) and the higher-order PNCA (right) when $n_0 = 3$ and $c = 0.2$ $(w = 0.9375)$. 100 Metropolis-Hastings steps were performed for every update of the missing data in the two PNCAs. . . . 218
7.11 ρ_{pxda} for $D = 2, 1, 0.5$. The asymptote in each of these curves is ρ_{nc}. 223
7.12 The spatial GLMM of Section 7.11.1 and its special case, the random effects model of Section 7.10 (showed in the bottom two plots). ACF for Θ using CA (dotted), NCA (dashed) and data-dependent PNCA (solid) for various parameter values. 228

List of Tables

1.1 Relevance of the review areas of the thesis to the main innovations. 7

6.1 Information about simulated datasets 171
6.2 Information about simulated datasets for the two-OU model 184
6.3 Posterior parameter summaries for the US$/DM series under the single-OU model 193
6.4 Posterior parameter summaries for the US$/DM series under the two-OU model 194

Chapter 1

Introduction

1.0 Motivation

The last 15 years have seen the wide spread of sampling-based methods for performing Bayesian statistical inference. The adoption of these methods is particularly appropriate within this paradigm, see for example Smith and Roberts (1993), where most inferential problems typically involve multidimensional analytically intractable integrations, which however can be easily managed using Monte Carlo methods. Therefore, the computational challenge usually faced within a Bayesian analysis is, given a distribution π on some (arbitrary) state space $\mathcal{Z}$, to obtain samples $Z_0, Z_1, \ldots \sim \pi$. Then, use the samples to approximate for a real function f on $\mathcal{Z}$ with finite mean under π,

$$\int_{\mathcal{Z}} f(z)\pi(dz) \approx \frac{1}{n}\sum_{i=0}^{n} f(Z_i). \tag{1.1}$$

Techniques which attempt to draw values directly from π have been shown to have limited scope and applicability. Instead, a collection of very powerful, general and easy to implement, iterative computational methods, known as Markov chain Monte Carlo (MCMC), have found a huge success within the statistical community since the beginning of the 1990's. The main idea can be traced at least as far back as Metropolis et al. (1953) in the physics context, and Hastings (1970) in the statistical context. MCMC methods had been used in the computer science and operational research (see for example Kirkpatrick et al. (1983)) and image analysis (see for example Geman and Geman (1984)) but it was much later with Gelfand and Smith (1990) that the broad statistical audience became aware of the potential of MCMC for Bayesian inference.

The idea behind MCMC is simple: for a given distribution π on $\mathcal{Z}$ construct a Markov chain with state space $\mathcal{Z}$ whose stationary distribution is π. Then under mild conditions a Markov chain sample path $\{Z_0, Z_1, \ldots\}$ is an approximate and dependent random sample

from π and well known asymptotic results ensure, for example, distributional convergence of the realisations

$$Z_n \xrightarrow{d} Z,\ Z \sim \pi$$

and consistency of the ergodic averages, for any integrable scalar function f,

$$\frac{1}{n}\sum_{i=0}^{n} f(Z_i) \to \int_{\mathcal{Z}} f(z)\pi(dz),\ \text{as } n \to \infty,\ \textit{almost surely}; \tag{1.2}$$

this type of results are more thoroughly presented in Section 1.5.1 and Section 3.1.

The dependence among the simulated values is of serious concern for assessing the efficiency of an MCMC algorithm. The ergodic average estimators (1.2) can have very unstable behaviour and converge (in an appropriate norm) very slowly to their strong law limiting values, in the presence of high serial dependence in the $\{Z_n\}$ series. In worst cases it might be even impossible to assess the error of the estimator for a finite value of n in (1.2) (using for a example a central limit theorem) since the estimator might have infinite asymptotic variance. In such cases it is important to redesign the sampler in order to reduce serial dependence and obtain much more reliable results.

It is often the case that the random variable Z, whose distribution π we want to sample from, admits a natural partition $Z = (Z^{(1)}, \ldots, Z^{(k)})$. A popular MCMC variant samples from π by iteratively sampling (either directly or using an approximate MCMC step) from the full conditional distributions of each $Z^{(i)}$ given the current values of the other components $Z^{(j)},\ j \neq i$. An important special case, which arises in the Bayesian analysis of hierarchical models (see Section 1.3) takes $k = 2$. High dependence (under π) between the updated components $Z^{(1)}$ and $Z^{(2)}$ results in high serial dependence in the MCMC draws $\{Z_n\} = \{(Z_n^{(1)}, Z_n^{(2)})\}$, thus in an inefficient sampler, as it was argued in the previous paragraph. Algorithmic performance can be significantly improved by finding a reparameterisation of the $Z = (Z^{(1)}, Z^{(2)})$ vector, under which the two components are mildly dependent under π. Then, the MCMC successive draws are also mildly serially dependent.

The motivation behind this thesis is to develop a general methodology for constructing such reparameterisations and, thus, improve the performance of componentwise-updating MCMC algorithms.

1.1 Outline of the thesis

The unifying theme of this thesis is the construction and analysis of non-centered parameterisations for data augmentation and hierarchical models. This class of parameterisations is proposed as a general purpose method to improve the speed of convergence of the Gibbs sampler and related component-wise updating MCMC algorithms. We look at the problem

of non-centering from various perspectives:

- Characterising the rate of convergence of the Gibbs sampler on linear hierarchical models under this parameterisation (Chapter 2 and Chapter 3).
- Developing methods which can be used in order to construct non-centered parameterisations for various hierarchical models with latent random variables and stochastic processes, as well as exposing how the corresponding MCMC algorithms are implemented (Chapter 4 and Chapter 5).
- Linking the ideas of reparameterisation to probabilistic representations of stochastic processes (Chapter 5).
- Applying non-centered algorithms to make Bayesian inference about a new class of stochastic models in financial econometrics (Chapter 6).
- Using non-centered techniques as a hierarchical model diagnostic tool (Chapter 6).
- Constructing a continuum of parameterisations which lie in the between of the centered and the non-centered (Chapter 7).

For expositional reasons we have chosen to blend in the new results of the thesis together with review of existing work in the sense that every chapter contains both review and original work. This choice makes each chapter relatively self-contained. Moreover, the introduction of technical details is avoided until they become necessary. The review material contained in Chapter 1 is relevant to all the subsequent chapters. This section serves to clarify which are the main innovations of the thesis and which existing results are reviewed.

The main innovations of this thesis are summarised below, where we also give references to the specific sections where they can be traced.

1. We discuss general issues related with non-centered parameterisations of hierarchical models. In particular:
   (a) We formalise the notions of centered and non-centered parameterisations for three-stage hierarchical models; see Section 1.7, Section 2.2 and Section 4.1.
   (b) We describe the MCMC algorithms under these parameterisations; see Section 4.1.
   (c) Based on analytic results, intuitive arguments and simulation studies we advocate the employment of non-centered methods for certain models and types of data; see Section 1.7, Section 2.3, Section 2.5, Section 3.4, Section 4.1 and Section 6.8.
   (d) We compare empirically different types of non-centered parameterisations and make some conjectures regarding their observed variability; see Section 6.12.2 and Section 4.3.

2. We derive the analytic geometric convergence rate of the Gibbs sampler under the centered and the non-centered parameterisation for a general normal hierarchical model; see Section 2.4. As a special case we study the simple state-space model; see Section 2.5.

3. We express the normal hierarchical model in terms of Brownian motion, which helps us gain intuition about some results concerning the convergence rate of the Gibbs sampler on this model; see Section 2.3.1.

4. We state and prove results concerning the type of convergence of the Gibbs sampler under the centered and the non-centered parameterisation for linear non-Gaussian hierarchical models; see Chapter 3. In particular:

   (a) We prove that when the observation error is Gaussian and the latent error is Cauchy the Gibbs sampler under a non-centered parameterisation is uniformly ergodic, while under a centered it fails to be geometrically ergodic; see Section 3.3.

   (b) We establish a connection between the convergence rates results and a well studied problem in Bayesian robustness; see Section 3.2.1.

   (c) This connection is exploited to provide two conditions on the tails of the observation and latent error distributions which generalise the convergence results for the Cauchy-normal to arbitrary linear hierarchical models; see Section 3.2.1 for a description of the conditions and Section 3.4 for a statement of the general result.

   (d) We prove geometric ergodicity for the Gibbs sampler under both the centered and the non-centered parameterisation for linear hierarchical models with double exponential error distributions for the observation and latent equations; see Section 3.5.

5. We provide an alternative proof of the main result of Dawid (1973), concerning the dominance of the prior over the likelihood in the presence of outliers for models with location structure, where we show that one of his conditions is actually not necessary; see Section 3.2.1.

6. We develop a technique for non-centering random variables with infinitely divisible distributions, which is based on an expansion of the state space; see Section 4.2.1. The implementation of the MCMC algorithm under this parameterisation is illustrated using a gamma random effects model as an example; see Section 4.2.

7. We consider non-centered parameterisations for latent Poisson processes; see Chapter 5. In particular:

   (a) We introduce two different non-centered parameterisations for Poisson processes; see Section 5.3 and Section 5.4.

(b) We describe the implementation of the corresponding MCMC algorithms for a variety of cases; see Section 5.3.1, Section 5.3.2 and Section 5.8.1.

(c) We extend our results to marked Poisson processes; see Section 5.5 and Section 5.5.1.

(d) We establish a connection between these parameterisations and probabilistic representations of positive Lévy processes; see Section 5.8.1.

8. We develop MCMC methodology for performing Bayesian inference for non-Gaussian Ornstein-Uhlenbeck stochastic volatility models; see Chapter 6. In particular:

   (a) We propose a data augmentation scheme based on marked Poisson processes; see Section 6.6.

   (b) We design centered and various non-centered MCMC algorithms for sampling from the posterior distribution of the parameters and the missing data in the model; see Section 6.7, Section 6.7.1, Section 6.9, Section 6.10 and Section 6.11.

   (c) We carry out an extensive simulation study to compare the performance of the MCMC algorithms; see Section 6.12.

   (d) We extend our methods to models constructed by superpositioning Ornstein-Uhlenbeck processes; see Section 6.13. We also discuss possible extensions of our methodology to more general models which include drift, risk premium and leverage terms; see Section 6.17.

   (e) We investigate prior elicitation and sensitivity for the model parameters, as well as identifiability of the latent structure; see Section 6.14.

   (f) We fit the models to a series of US dollar (US$) - Deutsch Mark (DM) daily exchange rates that spans the period from 01/01/1986 to 01/01/1996. We compare our results with those obtained by Barndorff-Nielsen and Shephard (2002a), which used second order methods to fit these models to the same financial series but sampled at higher frequencies; see Section 6.16.

9. We introduce a model diagnostic tool for hierarchical models and apply it to the stochastic volatility models of Chapter 6; see Section 6.15.

10. We introduce the partially non-centered parameterisation, which includes the centered and the non-centered as special cases; see Chapter 7. In particular:

    (a) We describe the parameterisation for the simple normal-hierarchical model and we calculate the rate of convergence of the Gibbs sampler under this parameterisation; see Section 7.2.

(b) We extend our results to the general normal hierarchical model and to models outside the Gaussian family; see Section 7.3, Section 7.5 and Section 7.6.

(c) We introduce a technique which is helpful in finding data-dependent partially non-centered parameterisations which outperform (in terms of the corresponding Gibbs sampler convergence rate) both the centered and the non-centered; see Section 7.10.

(d) We discuss how this methodology relates with the marginal and conditional augmentation but also with parameterisations which are based on posterior correlation analysis; see Section 7.9.1, Section 7.9.2 and Section 7.7.

(e) We apply our methods to improve the convergence rate of MCMC for a non-Gaussian geostatistical model; see Section 7.11.

11. We propose a class of parameterisations for generalised linear hierarchical models. We use simulations to investigate the performance of the Gibbs sampler and related component-wise updating techniques under this parameterisation and compare it with the centered and non-centered; see Section 7.8.

In addition to the original work contained in this thesis, certain key existing results are reviewed in various places. In particular we review

i basic principles of hierarchical modelling and Bayesian analysis of missing data problems; see Section 1.3 and Section 1.6

ii the basic MCMC algorithms in Section 1.5.2 and a specific MCMC algorithm for simulating Gibbs processes in Section 5.1.5

iii theory concerning ergodicity and convergence rates of Markov chains on arbitrary state spaces; see Section 1.5.1, Section 2.1 and Section 3.1

iv convergence rate analysis of the simple normal hierarchical model under the centered and the non-centered parameterisation; see Section 2.3

v marginal and conditional data augmentation methods; see Section 7.9.1 and Section 7.9.2

vi theory regarding the properties and representations of independend increment processes and Lévy processes; see Section 1.8, Section 5.7, Section 5.8 and Section 5.8.1

vii basic properties of, and likelihood-based inference for Poisson and marked Poisson processes; see Section 5.1.

viii stochastic volatility modelling of financial time series; see Section 6.1 and Section 6.2

ix the properties and existing estimation methods of the stochastic volatility model introduced by Barndorff-Nielsen and Shephard (2001); see Section 6.3

x work on Bayesian robustness; see Section 3.2

xi basic results about the natural exponential family; see Section 7.8.1

Table 1.1 associates each review area with the research work for which it is suited.

Innovation	Review area
1	i, ii, iv
2	iii, iv,
3	iv, vi
4	iii, x
5	x
6	ii, vi
7	ii, vi, vii
8	ii, vi, vii, ix
9	i, vii
10	ii, iv, v, vi, xi
11	ii, iv, xi

Table 1.1: Relevance of the review areas of the thesis to the main innovations.

1.2 Notation

This section provides some guidelines about the notation and terminology we have employed in this thesis.

For a random variable X we use both $\pi_X(x)$ and $\pi(X)$ to denote its density function. The former is mathematically more sound, and where such notational precision is important (for example in Chapter 3) we adhere to this form. Otherwise we use the less formal $\pi(X)$. A similar rule holds for conditional densities. With an abuse of notation, we use π to refer to probability measures as well, with the exception of Chapter 3 where the separation between a measure and its density is important and we use Π to refer to the former. "$\Rightarrow$" denotes weak convergence of probability measures, "$\stackrel{d}{=}$" equality in distribution between two random variables and $\stackrel{d}{\rightarrow}$ convergence of random variables in distribution.

In the greatest part of this thesis Y stands for the observed data, X the missing data and Θ the parameters in a three-stage hierarchical model. This generic notation is violated in Chapter 6 to keep consistency with Roberts et al. (2003).

The transition kernel of a Markov chain is denoted by P and its density (if it exists with respect to some reference measure) by p. Proposal kernels of Metropolis-Hastings chains are denoted by Q. Without confusion, the same letter is used to refer to the precision matrix of a multivariate Gaussian distribution.

The indicator function is denoted by $\mathbb{1}$, for example $\mathbb{1}[x \in A]$ is one, if $x \in A$ and zero otherwise. I_p denotes the $p \times p$ identity matrix, $\mathbf{1}$ a vector of 1s, $\mathbf{0}_{r_1 \times r_2}$ an $r_1 \times r_2$ matrix of 0s, where sometimes we omit the subscripts if the dimensions are obvious from the context. The transpose of a matrix A is denoted by A^t.

We have followed Bernardo and Smith (1994) to denote distributions: $\mathrm{Bi}(n, p)$ is the binomial distribution with mean np, $\mathrm{Pn}(\lambda)$ the Poisson with mean λ, $\mathrm{N}(\mu, \sigma^2)$ the Gaussian distribution with mean μ and variance σ^2, $\mathrm{Ga}(\alpha, \beta)$ the gamma with mean α/β and variance α/β^2, $\mathrm{Ig}(\alpha, \beta)$ the inverse gamma with parameters α, β (defined such that the change of variables $x \to 1/x$ leads to a $\mathrm{Ga}(\alpha, \beta)$), $\mathrm{Ex}(\beta)$ the exponential with mean $1/\beta$, $\mathrm{DEx}(\mu, \beta)$ the double exponential with location μ and scale β, $\mathrm{Be}(\alpha, \beta)$ the beta with mean $\alpha/(\alpha + \beta)$, $\mathrm{Un}[\alpha, \beta]$ the uniform in $[\alpha, \beta]$, $\mathrm{Ca}(\mu, \sigma)$ the Cauchy with location μ and scale σ.

We denote the centered parameterisation by CP and the non-centered by NCP, while the corresponding MCMC algorithms are denoted by CA (centered algorithm) and NCA (non-centered algorithm) respectively.

We use directed acyclic graphs to represent hierarchical models and employ the standard notation (see for example Whittaker (1990), where observed variables are included in square nodes, unobserved variables are included in round nodes, and dashed arrows represent deterministic relationship.

1.3 Bayesian inference for missing data problems

The statistical models considered in this thesis share a common structure: the distribution of (Y, X) is specified and depends on the parameters Θ. However, only Y is observed, and therefore X is treated as *missing data*. Section 1.6 discusses a variety of such models and how they appear in statistical modelling; many more examples can be found in the following chapters. The pair (Y, X) is often termed the *augmented* or *complete data*. The term "missing data" should not necessarily be interpreted as data which for some reason we failed to collect, rather as data which are not available to us. In many cases, especially in models with latent variables, random effects or hidden stochastic processes, it is likely that we would never be able to observe X, which could for example represent a collection of unobserved and unknown covariates which are used to explain the variation of the observed

data Y.

Adopting the Bayesian perspective, a prior distribution is assigned to Θ and we are typically interested in deriving or sampling from the posterior distribution of Θ, i.e the conditional distribution of Θ given the observed data Y. Assuming the existence of densities with respect to the Lebesgue measure for simplicity, the latter is given up to proportionality by

$$\pi_{\Theta|Y}(\theta \mid y) \propto \int \pi_{Y,X|\Theta}(y, x \mid \theta) dx \; \pi_{\Theta}(\theta) \tag{1.3}$$

where $\pi_{\Theta}(\theta)$ is the prior on Θ. (1.3) can be easily generalised for arbitrary measures and expresses the fact that in order to perform posterior inference for Θ we need to find the marginal distribution of the observed data given the parameters. In many complex statistical models used now-days (as for example in geostatistics, econometrics, engineering), the integration in (1.3) is neither analytically nor numerically feasible. On the contrary, statistical analysis of the so-called complete data is typically much more straightforward; this for example, is the case in the hierarchical models introduced in Section 1.6.

However, powerful iterative sampling schemes have been developed that sample from the joint posterior distribution of (Θ, X) by exploiting the tractability of the two conditionals $X \mid Y, \Theta$ and $\Theta \mid X, Y$. This methodology, known as the data augmentation (Tanner and Wong (1987)), is described in Section 1.5.2. Once samples have been obtained from this joint distribution, sampling based posterior inference can be easily performed for Θ (or X) using Monte Carlo methods, for example Ripley (1987) and especially Section 2 of Gelfand and Smith (1990) and Section 4 of Smith and Roberts (1993) for accounts of Bayesian inference using sample-based methods. In some cases, even when the integration in (1.3) is manageable, it might still be computationally more efficient to resort to the iterative methods of Section 1.5.

1.4 Graphical models and conditional independence

We say that two variables X and Θ are independent, and we write $X \perp\!\!\!\perp \Theta$, when the distribution of X is the same for all values of Θ, see Dawid (1979). Notice that this definition incorporates the possibility that the distribution of Θ is concentrated at a single value (i.e it is known) or even that this distribution is improper.

The notion of conditional independence is fundamental in this thesis; it is used to construct hierarchical models, for example in Section 6.7; our centered and non-centered parameterisations for a hierarchical model are actually defined in terms of the conditional independence structure they impose between the missing data and the parameters, see for example Section 2.2 and Section 4.1; it is exploited to simplify the implementation of our

state-space expanded MCMC algorithms, as for example in Section 4.2 and Section 5.3.1.

Dawid (1979) builds up (rather heuristically) the theory of conditional independence in a statistical context and we refer to this for the main definitions and properties. In a nutshell, the random variables Y and Θ are said to be conditionally independent given another variable X, and we write

$$Y \perp\!\!\!\perp \Theta | X,$$

when they are independent in their joint distribution given $X = x$, for any value of x. Informally, this implies that X contains all the information which is necessary to predict Y, thus when it is known Θ becomes irrelevant. Marginally though, when X is unknown, Y and Θ could be dependent; we will come across several such examples in Chapter 2 and Chapter 6. The conditional independence is often expressed in terms of a factorisation of the joint density of X, Y, Θ, see for example Section 3.1 of Dawid (1979) and Proposition 2.2.3 of Whittaker (1990). This approach is convenient when working with variables which take values in Euclidean spaces but it becomes less attractive when considering more arbitrary random objects. For example, in Section 5.3.1 we want to express that a Poisson process X is independent of some data Y conditionally on a parameter Θ. Therefore, it is much more natural to work with the general and more abstract principles of conditional independence developed in Dawid (1979).

A compact and illustrative way of expressing conditional independence statements is by means of graphical models and we adopt this approach in this thesis. We refer to Whittaker (1990) for an introduction to graphical modelling and to Section 1.2 for some notational conventions which help the interpretation of the graphs.

1.5 Markov chain Monte Carlo methods

Markov chain Monte Carlo (MCMC) methods are employed to (approximately) draw samples from a specific distribution, π say, which is usually termed the target distribution. π is typically multi-dimensional and often has support on non-Euclidean spaces, for example it can be the distribution of a Gibbs point process as in Section 5.1.4. In our applications π is the joint posterior distribution of the parameters and the missing data in a hierarchical model. Before presenting the main idea and reviewing some well known MCMC algorithms, we give a short introduction to Markov chains on general state-spaces.

1.5.1 Basic Markov chain theory

This section reviews some basic theory for Markov chains on general state spaces; important references for the development of this theory include Meyn and Tweedie (1993) and Roberts and Tweedie (2004); the former is more broadly concerned with Markov chains whereas the

latter is focused on Markov chain theory relevant to MCMC. More specialised results and definitions appear when they become necessary in Section 2.1 and Chapter 3.

Let $\mathcal{Z}$ be a general set and $\mathcal{B}(\mathcal{Z})$ denote a countably generated σ-algebra on $\mathcal{Z}$. Often $\mathcal{Z} \subset \mathbb{R}^d$, for some $d \geq 1$ and $\mathcal{B}(\mathcal{Z})$ is the Borel σ-algebra, although many applications of this thesis involve the construction of Markov chains which live on non-Euclidean spaces, see for example Section 5.1.5. The sample path space is defined as the countable product $\Omega = \prod_{i=0}^{\infty} \mathcal{Z}^{(i)}$, where each $\mathcal{Z}^{(i)}$ is a copy of $\mathcal{Z}$, and $\mathcal{F}$ is the corresponding σ-algebra.

The Markov chain

$$\{Z_0, Z_1, Z_2, \ldots\} =: \{Z_n\}$$

as a stochastic process in discrete time, can be constructed by starting from the transition probabilities (see Chapter 3 of Meyn and Tweedie (1993) for an alternative but equivalent construction).

Definition 1.5.1. *The transition probability kernel P is a function from $\mathcal{Z} \times \mathcal{B}(\mathcal{Z})$ to $[0,1]$, such that*

1 for each $A \in \mathcal{B}(\mathcal{Z})$, $P(\cdot, A)$ is a measurable function on $\mathcal{Z}$

2 for each $z \in \mathcal{Z}$, $P(z, \cdot)$ is a probability measure on $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$.

For any initial measure μ on $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$ and any transition probability kernel P, it can be shown (see Section 3.4 of Meyn and Tweedie (1993)) that there exists a stochastic process $\{Z_n\}$ on Ω, measurable with respect to $\mathcal{F}$, and a probability measure P_μ on $(\Omega, \mathcal{F})$, such that for measurable $A_i \subset \mathcal{Z}^{(i)}$, $i = 0, \ldots, n$ and any n, P_μ exhibits the following conditional independence structure

$$\mathsf{P}_\mu(Z_0 \in A_0, Z_1 \in A_1, \ldots, Z_n \in A_n) = \int_{A_0} \ldots \int_{A_{n-1}} \mu(dy_0) P(y_0, dy_1) \cdots P(y_{n-1}, A_n). \quad (1.4)$$

$\mathsf{P}_\mu(B)$ is the probability of the event that the Markov chain sample path belongs in the set of sample paths B, for $B \in \mathcal{F}$. A stochastic process Z on $(\Omega, \mathcal{F})$ is called a time-homogeneous Markov chain with transition probability kernel $P(z, A)$ and initial distribution μ if its finite dimensional distributions satisfy (1.4) for every n. The conditional independence in (1.4) is known as the Markov property. Notice that μ can be concentrated on a single point $z \in \mathcal{Z}$, in which case we write P_z for the probability measure of the Markov chain. Similarly, expectations under the P_z measure are denoted by E_z. The n-step transition probabilities $P^n(z, A)$ can be defined inductively as

$$P^n(z, A) = \int_{\mathcal{Z}} P(z, dy) P^{n-1}(y, A), \; z \in \mathcal{Z}, A \in \mathcal{B}(\mathcal{Z})$$

and clearly $P^n(z, A) = \mathsf{P}_z(Z_n \in A)$.

A σ-finite measure π on $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$ with the property

$$\pi(A) = \int_{\mathcal{Z}} \pi(dz) P(z, A) \tag{1.5}$$

for all $A \in \mathcal{B}(\mathcal{Z})$ is called invariant. The Markov chain Z with initial distribution π is said to be reversible with respect to π if and only if the detailed balance relation holds

$$\pi(dx)P(x, dy) = \pi(dy)P(y, dx). \tag{1.6}$$

This is understood as an equality of two measures on the product space $(\mathcal{Z}\times\mathcal{Z}, \mathcal{B}(\mathcal{Z})\otimes\mathcal{B}(\mathcal{Z}))$. It follows that π is an invariant measure for Z.

1.5.2 MCMC algorithms

There is already a vast literature about the methodology, theory, implementation and applications of MCMC. Within the statistics community, currently available texts on the subject include, for example, Gilks et al. (1996), Tanner (1996), Gammerman (1997), Robert and Casella (1999), Liu (2001) and Roberts and Tweedie (2004). The following paragraphs describe briefly some of the MCMC algorithms most relevant to our purposes and we refer to the aforementioned books for more details. Nevertheless, much more specialised results about certain algorithms appear later on in this thesis, as for example in Section 2.1 and Chapter 3.

For a given target distribution π, MCMC methods construct a Markov chain $\{Z_n\}$ which has π as its invariant distribution. Rather mild conditions ensure that π is also a limiting distribution of the chain, whatever the initial value Z_0; the main result together with the necessary conditions can be found in Chapter 3. Such Markov chains are called ergodic. Most of the MCMC algorithms used in practice, and certainly all those considered in this thesis, satisfy the conditions which ensure convergence (in an appropriately defined norm) to the invariant distribution π. Thus, the main challenge in designing an MCMC algorithm is to ensure that π is invariant, which is most easily achieved using the idea of reversibility.

From a statistical perspective, the convergence in distribution of the Markov chain to π is exploited to estimate expectations under the invariant measure. Suppose that $f : \mathcal{Z} \to \mathbb{R}$, then we define,

$$\pi f := \int_{\mathcal{Z}} f(x)\pi(dx) \tag{1.7}$$

and

$$\mathcal{L}^1 := \{f : \mathcal{Z} \to \mathbb{R} \text{ such that } \pi|f| < \infty\}.$$

Similarly, we define

$$\mathcal{L}^2 := \{f : \pi f^2 < \infty\}$$

and

$$\mathcal{L}_0^2 := \{f \in \mathcal{L}^2 : \pi f = 0\}.$$

The ergodicity of $\{Z_n\}$ implies that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f(Z_i) = \pi f \text{ for all } f \in \mathcal{L}^1 \tag{1.8}$$

and

$$Z_n \xrightarrow{d} Z, \text{ where } Z \sim \pi$$

for *almost surely* all starting points $Z_0 = z$, where "$\xrightarrow{d}$" denotes convergence in distribution of random variables (see Section 1.2). More details about this sort of convergence results can be found in Chapter 8 of Roberts and Tweedie (2004).

These results facilitate sampling based inference about π, see for example Section 2 of Gelfand and Smith (1990). In Bayesian analysis, π is a posterior distribution and most posterior inference problems come down to calculating posterior expectations, see for example Gelfand and Smith (1990), thus MCMC is a very powerful tool for posterior inference, although it has also found numerous applications outside Bayesian statistics.

Since the conditions which ensure ergodicity are usually met, an important issue is the speed at which an MCMC algorithm "converges to stationarity". This practically determines how much time we should "run" the chain before treating the simulated values as draws from π. A related, but not identical, concern is the dependence among the simulated values. Even if we start in stationarity, by sampling $Z_0 \sim \pi$, the Markov chain generates exact but dependent samples from π. High dependence among the sample can result in very slow convergence of the ergodic average estimates (1.8) to the expectations under π. The former problem is investigated qualitatively in Chapter 3, where the notions of geometric and uniform ergodicity are introduced. The latter is quantitatively answered in Section 2.1, where exact convergence rates are obtained for a particular class of MCMC algorithms, the Gibbs sampler. We now proceed to describe the general forms of the MCMC algorithms used in this thesis.

In the following we assume that we wish to simulate variates from a multi-dimensional distribution π which has support on $\mathcal{Z}$ and we denote a random variable distributed according to π as Z, $Z \sim \pi$.

The Gibbs sampler

The Gibbs sampler decomposes the state space $\mathcal{Z}$ as $\mathcal{Z}_1 \times \mathcal{Z}_2 \cdots \times \mathcal{Z}_k,\ k > 1$ and simplifies a complicated multi-dimensional simulation into a collection of k smaller dimensional and often more manageable ones. Often, $\mathcal{Z} = \mathbb{R}^d,\ \mathcal{Z}_i = \mathbb{R}^{r_i}$ and $\sum_i r_i = d$, but this thesis will be mostly concerned with more general state spaces. The factorisation of the space is usually naturally suggested by the statistical model under consideration, see for example the discussion about the two-component Gibbs sampler for missing data problems later in this section.

We write $z = (z^{(1)}, \ldots, z^{(k)})$ for an element of $\mathcal{Z}$ where $z^{(i)} \in \mathcal{Z}_i$ for all $1 \leq i \leq k$. We also write $z^{(-i)}$ for any vector produced by omitting the ith component,

$$z^{(-i)} = (z^{(1)}, \ldots, z^{(i-1)}, z^{(i+1)}, \ldots, z^{(k)})$$

from the vector z. We also follow the same notational conventions for the random variable $Z \sim \pi$.

Avoiding technical details we assume the existence of the conditional distributions $Z^{(i)} \mid Z^{(-i)} = z^{(-i)}$ for all $i = 1, \ldots, k$, which we denote by

$$\pi_i(dz^{(i)} \mid z^{(-i)}). \tag{1.9}$$

The Gibbs sampler which samples from π is implemented as described below.

The Gibbs sampler

Choose Z_0
Set $n = 0$
Iterate the following steps
{
  Set $i = 1$
  While $i < k + 1$
  {
    Sample $Z_{n+1}^{(i)} \sim \pi_i(\cdot \mid z^{(-i)})$, where
      $z^{(-i)} = (Z_{n+1}^{(1)}, \ldots, Z_{n+1}^{(i-1)}, Z_n^{(i+1)}, \ldots, Z_n^{(k)})$
    $i = i + 1$
  }
  $n = n + 1$
}

The above scheme is also referred to as the deterministic scan Gibbs sampler due to the way the algorithm visits each of the k components. It creates a Markov chain on $\mathcal{Z}$ with transition kernel P which is the composition of k transition kernels $P^{(i)},\ i = 1, \ldots, k$. In particular, if $z, w \in \mathcal{Z}$ we define

$$P^{(i)}(z, dw) = \begin{cases} \pi_i(dw^{(i)} \mid z^{(-i)}), & \text{for } w^{(-i)} = z^{(-i)} \\ 0, & \text{otherwise} \end{cases}$$

and $P = P^{(k)} P^{(k-1)} \cdots P^{(1)}$. There are alternative updating schemes of the Gibbs sampler (see Roberts and Sahu (1997)) of which the most relevant in this thesis is the random scan Gibbs sampler. This algorithm at each iteration picks one of the k components, i say, uniformly at random and updates it according to $P^{(i)}$. The transition kernel of the associated Markov chain has the mixture form $(P^{(1)} + \cdots + P^{(k)})/k$.

It can be checked (see for example Theorem 3.4.2 of Roberts and Tweedie (2004)) that each $P^{(i)}$ is reversible with respect to π, from which easily follows that π is invariant for either the composition, as in the deterministic scan, or the mixture, as in the random scan Gibbs sampler, of the $P^{(i)}$s; see for example Proposition 3.3.3 of Roberts and Tweedie (2004).

Notice that the probabilistic behaviour of the algorithm (for example the convergence rate, see Section 2.1) is not affected by componentwise one-to-one transformations.

Two-component Gibbs sampler (data augmentation)

The data augmentation was originally developed by Tanner and Wong (1987) for finding fixed point solutions to integral equations which appear in statistical inference, and it can be viewed as the stochastic analogue to the well known EM algorithm (see Dempster et al. (1977)). It is most often used to obtain samples from the joint distribution a of random vector, $Z = (Z^{(1)}, Z^{(2)})$ say, by sampling from the two conditional distributions. This scheme shares a lot in common with the Gibbs sampler, but Gelfand and Smith (1990) showed that the latter is at least as efficient as the former. It is a standard practice in the literature (see for example Liu et al. (1994), Liu and Wu (1999), Meng and van Dyk (2001)) to identify the data augmentation with the two-component Gibbs sampler and this thesis conforms to this convention.

Although the data augmentation is considered as a special case of the Gibbs sampler, we treat it separately because of its importance in tackling missing data problems. Moreover, it has some special properties that the more general Gibbs sampler does not possess. The conditional independence structure of the two-component Gibbs sampling Markov chain is depicted in the graphical model in Figure 1.1. From this it can be shown that the marginal

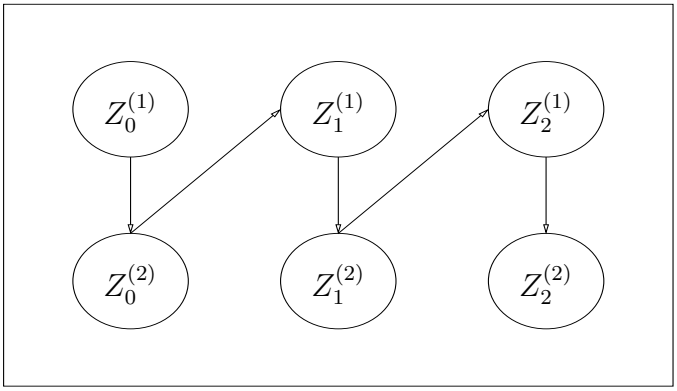


Figure 1.1: The conditional independence graph of the two-component Gibbs sampler.

chains $\{Z_n^{(i)}\},\ i = 1, 2$ are Markov, a feature not shared by the k-component Gibbs sampler. Actually, by directly checking the detailed balance condition, it can be shown that each chain is reversible with respect to the appropriate marginal distribution (see for example Lemma 3.1 of Liu et al. (1994)). These results simplify the theoretical analysis of the algorithm, as for example in Chapter 3. Furthermore, a complete analysis of the covariance structure of the two-component Gibbs sampler is feasible, which leads to a characterisation of its $\mathcal{L}^2$ rate of convergence; see Section 2.1 for a detailed exposition and references.

A rather intuitive property, which however demands considerable effort to be proved in full generality, is that the convergence rate of the Gibbs sampler (either in $\mathcal{L}^2$ norm as in Section 2.1 or in total variation distance as in Chapter 3) is not affected by the order in which components $Z^{(1)}$ and $Z^{(2)}$ are updated. Actually, the result for geometrically converging chains in $\mathcal{L}^2$ norm follows directly from Theorem 3.2 of Liu et al. (1994), see also Section 2.1. The most general statement is much harder to be shown, nevertheless we will assume that this it is true throughout this thesis. This property only becomes relevant in Chapter 3.

Data augmentation is by far the most widely adopted computational method for performing Bayesian analysis of missing data problems. The target distribution is the joint posterior distribution of the missing data X and the parameters Θ (see Section 1.3 for definitions). By construction, simulation from the conditional distributions $\pi_{\Theta|X,Y}$ and $\pi_{X|\Theta,Y}$ is manageable, certainly much more feasible than simulation from the marginal of interest $\pi_{\Theta|Y}$, which in many cases is not even available in closed form due to the integration in (1.3). Therefore, we use the two-component Gibbs sampler (or the more general Hastings-within-Gibbs to be introduced later in this section) which updates X and Θ, to obtain samples from $\pi_{\Theta,X|Y}$ and consequently from $\pi_{\Theta|Y}$.

Adaptive rejection sampling

Adaptive rejection sampling is not an MCMC method, instead it is a rejection sampling algorithm for simulating from log-concave one-dimensional densities; see Ripley (1987) for an introduction to rejection sampling and Wild and Gilks (1993) for a description of the algorithm. The main idea is that when the log-density is concave it can be easily bounded above by its tangents at either side of its (unique) mode. Thus, rejection sampling can be used to sample from the density using piecewise exponential functions as the envelope. Wild and Gilks (1993) propose an adaptive method to build up the envelope, by using proposals from the current version of the envelope which have been rejected as draws from the target density.

This algorithm has been used in numerous applications as a companion to the Gibbs sampler. This is due to the fact that the full conditional densities which are derived from commonly used statistical models are log-concave, see for example Dellaportas and Smith (1993) and Gilks et al. (1994). We will occasionally use this algorithm in this thesis. We use the publicly available (from the web page of W. Gilks) FORTRAN code, which however we have modified in order to correct certain mistakes and numerical instabilities of the original code.

The Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm is a method for constructing a reversible Markov chain $\{Z_n\}$ on a measurable space $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$ with a specified invariant distribution π. A suitable reference for a description of the algorithm for general state spaces is Tierney (1998), which we follow closely.

The algorithm requires a proposal kernel $Q(z, dw)$ and a measurable function $\alpha(z, w) : \mathcal{Z} \times \mathcal{Z} \to [0, 1]$. When the chain is at state z, a candidate value, w say, for the next state is generated by $Q(z, \cdot)$ and it is then accepted with probability $\alpha(z, w)$. If it is rejected, the next state of the chain is z. Therefore, the transition kernel is

$$P(z, dw) = Q(z, dw)\alpha(z, w) + \delta_z(dw) \int (1 - \alpha(z, u))Q(z, du),$$

where $\delta_z(\cdot)$ is the Dirac-delta measure centered at z. The steps of the algorithm are described below.

The Metropolis-Hastings algorithm

Choose Z_0
Set $n = 0$
Iterate the following steps
{

Sample $U_{n+1} \sim \mathtt{Un}[0,1]$
Sample $W_{n+1} \sim Q(Z_n, \cdot)$
If $U_{n+1} \leq \alpha(Z_n, W_{n+1})$ then
set $Z_{n+1} = W_{n+1}$
Else
set $Z_{n+1} = Z_n$
$n = n + 1$

}

For a given proposal kernel Q the aim is to find the acceptance probability α which ensures reversibility of the chain. The first thing to notice is that for reversibility the diagonal component does not matter, that is the Metropolis-Hastings kernel P is reversible with respect to π if and only if

$$\pi(dz)Q(z, dw)\alpha(z, w) = \pi(dw)Q(w, dz)\alpha(w, z) \tag{1.10}$$

which expresses an equality of two measures on the product space $\mathcal{Z} \times \mathcal{Z}$. Let R be the set of all pairs (z, w) for which transitions from z to w and from w to z are both possible for a Markov chain with initial distribution π and transition kernel Q. For the chain to be reversible, we have to ensure that for all $z \in \mathcal{Z}$, we only allow moves to $w \in \mathcal{Z}$ such that $(z, w) \in R$. We also define $\mu(dz, dw) = \pi(dz)Q(z, dw)$ and $\mu^T(dz, dw) = \mu(dw, dz)$. Inside R, μ and μ^T are mutually absolutely continuous and we define their density

$$r(z, w) := \frac{\mu(dz, dw)}{\mu^T(dz, dw)}.$$

The Metropolis-Hastings acceptance probability can then be written as

$$\alpha(z, w) = \begin{cases} \min\{1, r(w, z)\}, & \text{if } (z, w) \in R \\ 0, & \text{otherwise} \end{cases}$$

It has to be said that this α is not the unique valid solution to (1.10), but it is optimal under the Peskun ordering, see Section 2 of Tierney (1998) for details. Notice, that α does not require knowledge of the normalising constant of π, since r does not.

Under further assumptions, the derivation of α can be greatly simplified. For instance if $\pi(dz) = \pi(z)\nu(dz)$ and $Q(z, dw) = q(z, w)\nu(dw)$ for some measure ν on $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$ then $R = \{(z, w) : \pi(z)q(z, w) > 0 \text{ and } \pi(w)q(w, z) > 0\}$, and

$$r(z, w) = \frac{\pi(z)q(z, w)}{\pi(w)q(w, z)}.$$

Once α has been found, rather mild conditions ensure ergodicity of the Metropolis-Hastings algorithm, see for example Chapter 7 of Roberts and Tweedie (2004).

There is flexibility in the choice of the proposal kernel Q, and there is a sense in which all currently employed MCMC algorithms can be thought of as a special case of the Metropolis-Hastings algorithm under a particular choice of Q; see Section 2.5 of Roberts and Tweedie (2004). A very popular algorithm when $\mathcal{Z}$ is a metric space, is the random-walk Metropolis, for which the proposal kernel has a density $q(z, w) = q(|z - w|)$ which is a function only of the distance between z and w. Typically, when $\mathcal{Z} = \mathbb{R}^d$ for some d, q is the multivariate Gaussian density with mean a vector of 0s and covariance matrix $\sigma^2 I_d$. The scaling factor σ^2 can be chosen by the user to optimise algorithmic performance, see Roberts et al. (1997) for a thorough investigation of the optimal scaling of the random-walk Metropolis algorithm.

When $\mathcal{Z}$ is the positive half-line an attractive alternative is the so-called multiplicative random walk algorithm, for which the proposal kernel is described by the following random function of the current value z, $W = z \exp\{U\},\ U \sim \mathrm{N}(0, \sigma^2)$. It can be easily seen that this algorithm is equivalent to the random-walk Metropolis with $\mathrm{N}(0, \sigma^2)$ proposal distribution and target distribution obtained after a logarithmic transformation of the original target.

Hastings-within-Gibbs (component-wise updating algorithm)

The Hastings-within-Gibbs, also known as component-wise updating algorithm, is a hybrid of the Gibbs sampler and the Metropolis-Hastings algorithm, and it is used extensively in this thesis. Suppose that the state space has been factorised as $\mathcal{Z} = \mathcal{Z}_1 \times \mathcal{Z}_2$ and we want to use the Gibbs sampler to obtain samples from the target distribution π. Nevertheless, it is often the case that either or both of the conditional distributions $\pi_i,\ i = 1, 2$ in (1.9) do not admit simple forms that we can easily simulate from. The Hastings-within-Gibbs algorithm replaces the direct simulation by a Metropolis-Hastings updating step which has π_i as the invariant distribution. The conditional independence structure of the resulting Markov chain $\{(Z_n^{(1)}, Z_n^{(2)})\}$ is described by the graphical model in Figure 1.2, and it is interesting to contrast it with Figure 1.1. Notice that the marginal chains are not Markov anymore, although when one of the conditionals can be sampled from directly, one of the marginal chains

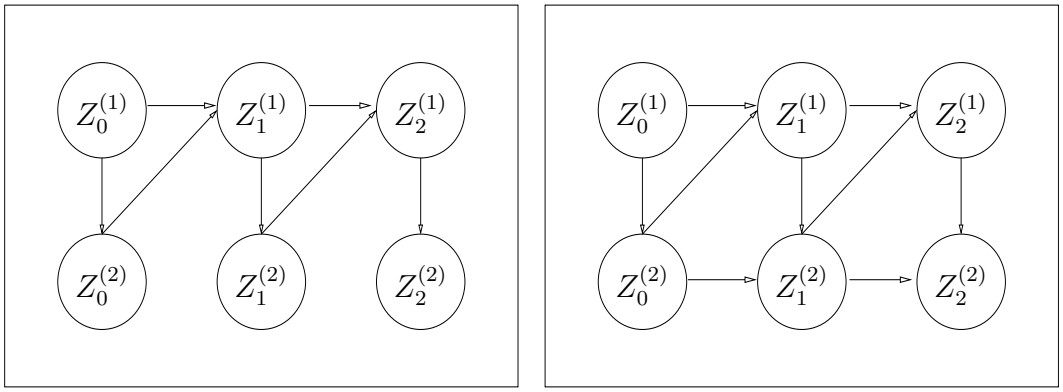


Figure 1.2: The conditional independence graph of the two-component Hastings-within-Gibbs sampler, when only one (left) and when both (right) conditionals are not sampled from directly.

($\{Z_n^{(1)}\}$ in the left panel of Figure 1.2) will still be Markov. In most cases, it is reasonable to assume that the ease in the implementation of the Hastings-within-Gibbs over the Gibbs sampler comes at the expense of speed of convergence. This will be the case in most of the examples in this thesis, where we employ the Hastings-within-Gibbs when direct simulations are not feasible, but in theory we would prefer to use a "pure" Gibbs sampler. Actually, the introduction of Metropolis steps can have serious negative impact on the convergence rate of the algorithm, see for example the discussion in Section 4.3 and Section 6.12.2. Nevertheless, there are Hastings-within-Gibbs samplers with better performance than the "pure" Gibbs, exploting for example antithetic simulation, see examples and references in Section 2.7 of Roberts and Tweedie (2004).

The Hastings-within-Gibbs becomes very relevant when considering missing data problems. The factorisation of the space is natural in terms of the parameters Θ and the missing data X. Moreover, we prefer to work with the two conditional rather than the joint distribution, mainly because X usually lives on a very different space than Θ, thus designing Metropolis-Hastings proposals for the pair (X, Θ) is not straightforward. However, in many complex models the conditionals are not possible to simulate from directly, thus we resort to the Hastings-within-Gibbs sampler.

Clearly, the algorithm described above can be generalised to the case where $\mathcal{Z} = \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_k,\ k > 2$; we describe its steps below.

The Hastings-within-Gibbs sampler

Choose Z_0
Set $n = 0$
Iterate the following steps
{

  Set $i = 1$
  While $i < k + 1$
  {

    Update $Z_{n+1}^{(i)}$ according to $\pi_i(\cdot \mid z^{(-i)})$, where
      $z^{(-i)} = (Z_{n+1}^{(1)}, \ldots, Z_{n+1}^{(i-1)}, Z_n^{(i+1)}, \ldots, Z_n^{(k)})$
    $i = i + 1$
  }
  $n = n + 1$

}

Notice that when the Hastings proposal kernels used at each step i are actually π_i, $i = 1, \ldots, k$, the Hastings-within-Gibbs collapses to the Gibbs sampler.

1.6 Hierarchical models

Essentially all Bayesian models can be viewed as hierarchical models, since we typically assume that the distribution of the observed data Y depends on some unobserved random quantities X, which can live on arbitrary finite or infinite dimensional spaces, whose distribution depends on other random quantities Θ. The distribution of Θ depends on other quantities, which can be assumed either random, thus adding another stage in the hierarchy, or known. We often adopt the second approach and the resulting model is a three-stage hierarchical model. An important aspect of this model, as described above, is the conditional independence between Y and Θ given X and this is illustrated in Figure 1.3.

Some justification of hierarchical modelling as a means to constructing a predictive model for the observables Y (which is the main objective in Bayesian modelling, see for example the discussion in Chapter 4 of Bernardo and Smith (1994)) is provided by the idea of exchangeability and partial exchangeability.

Definition 1.6.1. *The random variables $Y_1, \ldots, Y_n$ are said to be finitely exchangeable if their joint distribution is invariant under permutations of the index set $\{1, \ldots, n\}$. An*

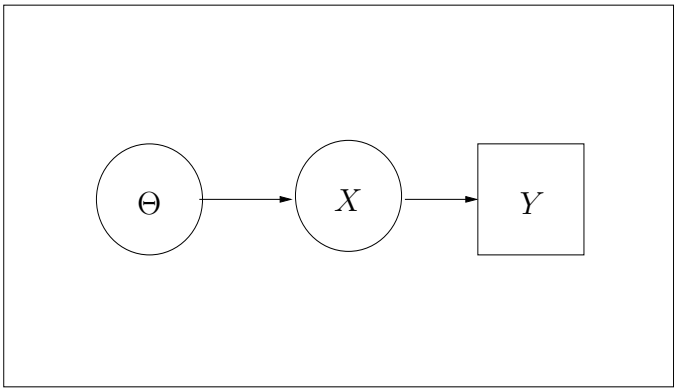


Figure 1.3: The graphical model of the centered parameterisation.

infinite sequence $Y_1, Y_2, \ldots$ *is said to be exchangeable if every finite subsequence is finitely exchangeable.*

Under the assumption of exchangeability for an infinite sequence of random variables $Y_1, Y_2, \ldots \in \{0, 1\}$, de Finetti proved a fundamental representation theorem. He showed that any probability measure P specifying the joint distribution for any subset of $Y_1, Y_2, \ldots$ can be represented in the following hierarchical form

$$\begin{aligned} Y_i \mid X &\sim \text{Bi}(1, X) \; i = 1, 2, \ldots \\ X &\sim Q(\cdot) \end{aligned} \tag{1.11}$$

where $X = \lim_{n\to\infty} \sum_{i=1}^{n} Y_i/n$ and the distribution Q expresses one's prior belief about where X will lie, that is $Q(x) = \lim_{n\to\infty} P[\sum_{i=1}^{n} Y_i/n \leq x]$. Therefore, conditionally on X, which is the unobserved limiting frequency of 1s, the Y_is are independent Bernoulli random variables, while Q is quantifying one's beliefs about this limiting frequency. We often take Q to be of some known parametric form depending on some unknown parameters Θ, which then produces a three-stage hierarchical model as described above.

This theorem has been generalised in various directions. For example when the Y_is live on a Euclidean space, there exists some random distribution function, say F, and a probability measure on the space of all distribution functions, say Q, such that conditional on F the Y_is are independent and identically distributed according to F, while Q expresses our beliefs on how the empirical distribution from a large sample of Y_is would look like (see Proposition 4.3 of Bernardo and Smith (1994)). Bayesian inference for random distribution functions forms the core of the so-called Bayesian non-parametric analysis, see Chapter 5 for more details. By making more assumptions about the probabilistic structure of the Y_is (e.g invariance under some transformations or the existence of fixed dimensional sufficient statistics) we can derive much more specific and easier to handle models. These considerations often lead to models where F is some finite-dimensional distribution with parameters X. Similarly the

prior distribution Q is in some parametric family indexed by the hyperparameters Θ, which in turn are considered unknown and distributed according to some parametric model with typically known parameters. This is an example of a three-stage hierarchical model. For the model described above the corresponding graphical model is given in Figure 1.4.

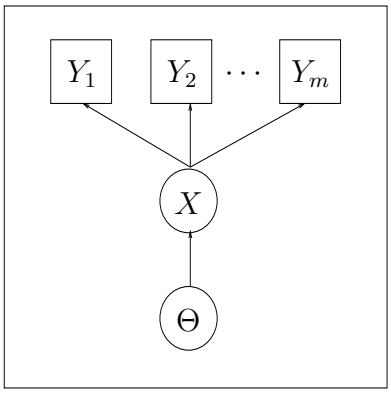


Figure 1.4: Graphical model of an exchangeable model

When modelling a sequence of random variables, exchangeability might be judged too severe an assumption, although we might be prepared to accept it when controlling for some other factors. Suppose for example (as considered in Lindley and Smith (1972)) that Y_{ij} are independent observations on the ith variety in a field trial, of average yield X_i. We might believe that the observations on a given variety are exchangeable although observations on different varieties might have substantial differences. If *a priori* all varieties seem indistinguishable in their performance, it seems reasonable to treat the X_is as an exchangeable sequence with common hyperparameters Θ. Therefore, a reasonable model for the Y_{ij}s is the three-stage hierarchical model described by Figure 1.5, where for simplicity of exposition we take only one observation from each variety. This hierarchical model fits

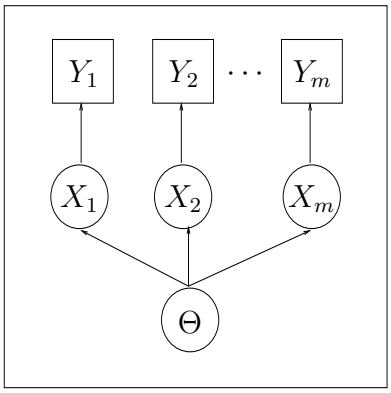


Figure 1.5: Graphical model of a partially exchangeable model

in the general framework describe by Figure 1.3 by setting $Y = (Y_{11}, \ldots, Y_{1n}, Y_{2n}, \ldots)$ and $X = (X_1, \ldots, X_m)$.

A popular example of a partially exchangeable model, which we use for expositional purposes in Chapter 2 is the so-called Gaussian random effects model (see for example Diggle et al. (1994) and references therein), which in its simplest form writes as

$$\begin{aligned} Y_{ij} &= X_i + \sigma_y \epsilon_{ij},\ j = 1, \ldots, n \\ X_i &= \Theta + \sigma_x z_i,\ i = 1, \ldots, m\ . \end{aligned} \tag{1.12}$$

where ϵ_{ij} and z_i are independent standard normal random variables. This model can be derived by the assumption of partial exchangeability and spherical symmetry on the Y_{ij}s (see Section 4.4.1 of Bernardo and Smith (1994)) and exchangeability on the unobserved X_is. In this context X_i is the limiting average of the sequence of $Y_{i1}, Y_{i2}, \ldots$. In many applications, for example in longitudinal studies, i indexes individuals and j measurements on the same individual. Therefore, (1.12) allows sharing of information among different individuals for estimating the individual means X_i, see Section 2.3 for more details. Random effects models can be constructed outside the Gaussian family as well, see for example Diggle et al. (1994) and Lee and Nelder (1996), and are a very a useful tool for modelling a wide variety of features observed in many data sets: clustering (discrete mixture models), complicated marginal distribution (continuous mixture models), heterogeneity among individuals in longitudinal-type studies, over-dispersion with respect to standard sampling distributions.

This thesis is primarily concerned with hierarchical models where X corresponds to a stochastic process. Such models are currently heavily studied in the literature and arise in many areas of science, for example in engineering, geostatistics, stochastic epidemics and econometrics. A slightly more complicated model than (1.12) is the Gaussian state-space model, which is discussed in Section 2.5 and allows for dependence among the X_is conditionally upon Θ. On the other hand, the models considered in Chapter 5 and Chapter 6 are much more complicated, where X is a Poisson process.

We treat hierarchical models, represented by the conditional independence graph Figure 1.3 as missing data problems, and identify X with the missing data.

1.7 Centering and non-centering

Section 1.6 described the general model where $Y \perp\!\!\!\perp \Theta \mid X$, a conditional independence depicted in Figure 1.3. We term the parameterisation in terms of X and Θ the *centered parameterisation* (CP), due to the fact that the missing data are centered between the observed data and the parameters. Suppose, instead, that we can find $\tilde{X}$ and some function (not necessarily invertible) h such that $X = h(\tilde{X}, \Theta)$ and $\tilde{X}$ is *a priori* independent of Θ. We term $(\tilde{X}, \Theta)$ the *non-centered parameterisation* (NCP) and its graphical model is given in Figure 1.6.

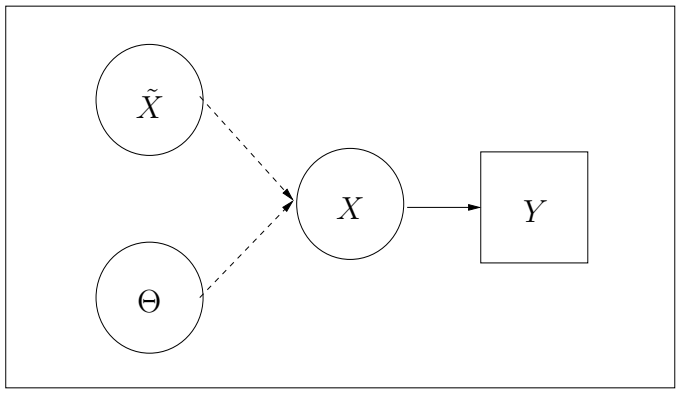


Figure 1.6: The graphical model of the non-centered parameterisation. The dashed arrows correspond to a deterministic link, that is X is a deterministic function of $\tilde{X}$ and Θ.

We intend to use the two-component Hastings-within-Gibbs algorithm to obtain samples from the joint distribution of (X, Θ) and this thesis shows that the parameterisation adopted, either the centered or the non-centered, has a crucial impact on the convergence of the algorithm. This is the motivation behind the NCP: a general purpose reparameterisation to improve the performance of the Hastings-within-Gibbs algorithm when it is slow under a CP; see Section 2.2 and Section 4.1 for an extensive discussion of these issues.

This thesis is concerned with constructing (Chapter 4, Chapter 5) analysing (Chapter 2, Chapter 3), implementing on challenging models (Chapter 6) and improving (Chapter 7) non-centered parameterisations.

1.8 Basics of Lévy processes and infinite divisibility

Lévy processes play an important role in this thesis, either as components of a hierarchical model, as in the Bayesian non-parametric models of Chapter 5 and the stochastic volatility models of Chapter 6, or as tools for constructing non-centered parameterisations, as in Chapter 4. Therefore, it is convenient to introduce, rather informally, some basic concepts and definitions at this early stage. More comprehensive treatment is given in Section 5.7 and Section 5.8.

A stochastic process in time $z(t),\ t \geq 0$ where $z(0) := 0$ *almost surely*, is called a Lévy process if it has independent and stationary increments, that is $z(t+s) - z(t),\ t, s > 0$, is independent of the history of the process up to time t and its distribution depends only on the separation s (see for example Sato (1999), Barndorff-Nielsen and Shephard (2004)). We take a version of the process which has cadlag (continuous from the right, limits from the left) sample paths. Notice however, that the term Lévy process is used occasionaly in the (Bayesian non-parametric) literature (see for example Walker et al. (1999)) to refer to processes with independent but non-stationary increments. In this thesis the

term Lévy process will be used to refer to a stationary increments process, and the term independent increments process to a process with independent but possibly non-stationary increments.

A concept closely linked to the Lévy processes is that of *infinite divisibility*. The following definition is taken from Feller (1971) (p.176):

Definition 1.8.1. *A distribution function F is infinitely divisible if for every n there exists a distribution F_n such that F is the n-fold convolution of F_n. In other words, F is infinitely divisible if and only if for each n it can be represented as the distribution of the sum $S_n = X_{1,n} + \cdots X_{n,n}$ of n independent random variables with common distribution F_n.*

This definition can be extended to higher dimensions but such a generalisation escapes the scope of this thesis. Many of the most commonly used distributions are infinitely divisible, the Gaussian, the Poisson, the gamma, the inverse gamma, the inverse Gaussian, the stable family but also the log-normal are a few examples. It turns out (see Barndorff-Nielsen and Shephard (2004)) that the marginal distributions of a Lévy process are infinitely divisible; see Section 5.7 for a discussion of this property and how it can be exploited to provide representations of Lévy processes. It can be shown that a Lévy process is characterised by its distribution at time 1. For example, we call $z(\cdot)$ a gamma process with parameters α, β when $z(1) \sim \text{Ga}(\alpha, \beta)$. We occasionally use the term standard gamma process to refer to the case where $z(1) \sim \text{Ga}(1, 1)$. A rigorous way to characterise the distribution at time 1 is by means of its cumulant function.

Definition 1.8.2. *The cumulant function of a random variable X is defined as the logarithm of the characteristic function of X,*

$$C(u; X) := \log\{\mathsf{E}[\exp\{iuX\}]\},\ u > 0, i = \sqrt{-1}.$$

For simplicity, especially when dealing with positive random variables, we work with the logarithm of the moment generating function, which we also call cumulant but denote as

$$K(u; X) := \log\{\mathsf{E}[\exp\{-uX\}]\},\ u > 0.$$

These are standard concepts in probability theory, see for example Barndorff-Nielsen and Shephard (2004) for more details.

It is not hard to show, using the independent increments property, that for a Lévy process the following is true

$$C(u; z(t)) = tC(u; z(1)),\ \text{for all } t > 0,$$

which explains why these processes allow great deal of analytic tractability and why it is enough to specify their distribution at time 1 in order to characterise the whole process.

For more independent increments processes we need a family of measures, the so-called Lévy measures for the same purpose, see for example Section 5.8.

We close this introductory section by giving three characteristic examples of Lévy processes. The first is the well known Brownian motion. In its standard form $z(1) \sim \mathrm{N}(0, 1)$, but more generally we can have $z(1) \sim \mathrm{N}(0, \sigma^2)$. The increments of this process are Gaussian

$$z(t+s) - z(t) \sim \mathrm{N}(0, \sigma^2 s),$$

a property which can be used to simulate values from this process; for example Figure 1.7 shows a standard Brownian motion path on $[0, 1]$ which has been simulated by splitting time in small intervals and simulating from the corresponding increments. There is a good reason

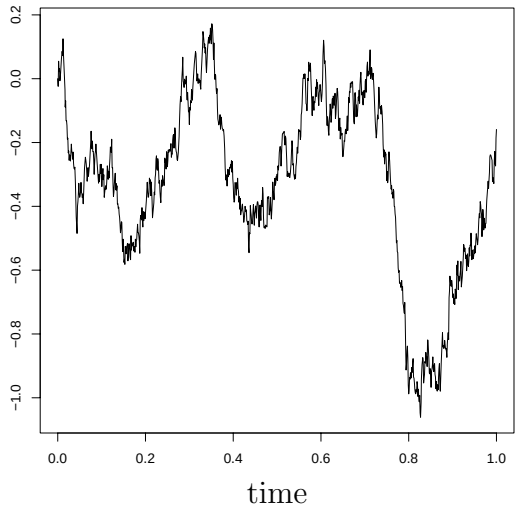


Figure 1.7: A path in $[0, 1]$ of a standard Brownian motion. It has been simulated by discretising time in intervals of length 10^{-3} and simulating from the corresponding increments of the process.

why we chose to plot the path using a continuous line, although we actually simulate only a discrete skeleton of the process. It can be shown (see for example Feller (1971)) that the Brownian motion is the only Lévy process with *almost surely* continuous sample path.

Our second example is the gamma process, specified by asking that $z(1) \sim \mathrm{Ga}(\alpha, \beta)$. The increments of the process are also gamma distributed

$$z(t+s) - z(t) \sim \mathrm{Ga}(\alpha s, \beta)$$

and a simulated path is shown in Figure 1.8. This is a pure jump process, a feature shared by all Lévy processes with positive increments. Actually, the gamma process has an infinite number of jumps in any bounded interval of time, but only a finite number of them are of non-negligible size; see Section 5.8 for more details.

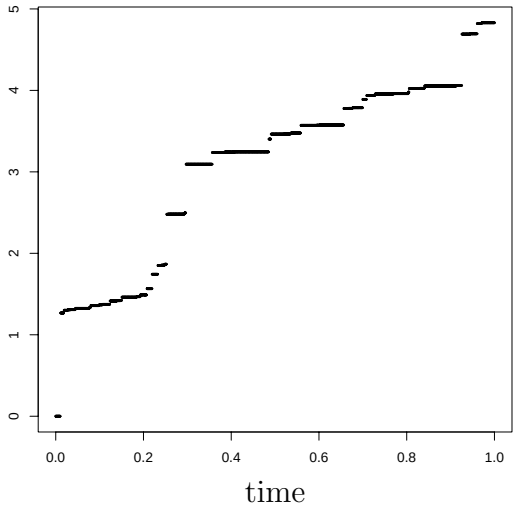


Figure 1.8: A path in $[0, 1]$ of a Ga$(10, 1)$ Lévy process. It has been simulated by discretising time in intervals of length 10^{-3} and simulating from the corresponding increments of the process.

The third example is the compound Poisson process, which turns out to play a fundamental role as a building block of the Lévy processes. It can be represented as

$$z(t) = \sum_{j=1}^{N(t)} E_j, \;\; z(0) := 0 \tag{1.13}$$

where $N(t)$ is the number of arrivals of a Poisson process with finite rate, λ say, in $[0, t]$ and the E_js are IID random variables and also independent from the Poisson process. This representation provides an explicit way to simulate paths from this process without any discretisation error, for example Figure 1.9 shows a realisation of the process when $E_j \sim$ Ex(Θ). The compound Poisson process has only a finite number of jumps at any bounded interval of time, and it is the only Lévy process with this property. This feature is depicted in the example of Figure 1.9.

For the Brownian motion and the gamma processes, we will be interested (in Section 2.3.1 and Section 4.2 respectively) in the stochastic process $z(t), \;\; t \in [0, 1]$ which is constrained to hit a specific value z_1 at time 1, $z(1) = z_1$ *almost surely*. The conditioned Brownian motion is known as the Brownian bridge, we will denote it by $B(t), t \in [0, 1]$, by construction $B(1) = z_1$ and $B(0) = 0$ *almost surely*, and it is a Gaussian process (see for example p.64 of Rogers and Williams (1994)). A useful representation of B in terms of z_1 and an independent Brownian motion w, which can exploited to simulate the process, is

$$B(t) = w(t) - t(w(1) - z_1), \;\; t \in [0, 1].$$

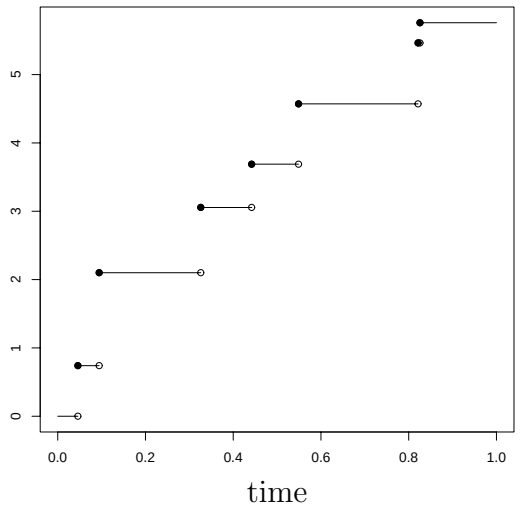


Figure 1.9: A path in $[0, 1]$ of a compound Poisson process with finite rate $\lambda = 10$ and $E_j \sim \text{Ex}(1)$. The path has been simulated without any discretisation error by explicitly simulating the jump times and corresponding sizes from the appropriate distributions.

The conditioned gamma process is known as the beta process. In particular it easy to show that

$$\frac{z(t)}{z(1)} \mid z(1) \sim \text{Be}(t, 1-t), \ t \in [0, 1].$$

It can be shown that, at any bounded interval of time, a sample path of the Brownian motion with scale parameter σ^2 contains infinite information (measured for example by Fisher's information) about σ^2. In particular, σ^2 can be obtained as a specific functional of the sample path by the so-called quadratic variation identity (see Section 6.3.5). Similarly, a sample path at any bounded interval of time of a gamma process with shape parameter α contains infinite information about α. There are important computational implications of these probabilistic results, see for example Section 5.8.2 and Section 7.6.

Chapter 2

Convergence rates and reparameterisations for the Gibbs sampler on normal hierarchical models

2.0 Introduction

This chapter reviews the existing theory for characterising and computing the rate of convergence of the Gibbs sampler. We revisit centered and non-centered parameterisation strategies that have been proposed for hierarchical Gaussian models and discuss the optimal choice between them in terms of their convergence rates. These parameterisation schemes are also considered in the context of some linear non-Gaussian models although this problem is considered in full detail in Chapter 3. This chapter is based and expands on the material contained in sections 1-3 of Papaspiliopoulos et al. (2003).

2.1 Rates of convergence of the Gibbs sampler

This section gives an overview of the existing theory for characterising and computing the rate of convergence of the Markov chain induced by the Gibbs sampler, which was introduced in Section 1.5.2. Based on the corresponding rates, different implementations of the Gibbs sampler for the same target distribution can be compared. In particular, we can decide on the updating and blocking strategies to be employed, but we can also contrast different augmentation schemes. We will be interested in comparing the centered and the non-centered parameterisation schemes on the basis of the convergence rates of the associated Gibbs sampling Markov chains.

There are many techniques in Markov chain theory for obtaining theoretical bounds on rates of convergence, Roberts and Tweedie (1996b), Amit and Grenander (1991), Tierney (1994) are just a few references while Chapter 11 of Roberts and Tweedie (2004) gives a detailed account of the area. However, these bounds are typically very conservative, especially in high dimensions. On the other hand, exact rates for the Gibbs sampler can be obtained when the target distribution is multivariate Gaussian. The methodology for computing such rates, developed by Roberts and Sahu (1997), is briefly outlined in Section 2.1.1.

Due to a very elegant observation by Amit (1991), a particularly insightful, although not very practically useful, characterisation of the rate convergence of the two-component Gibbs sampler exists, which we will now describe.

Section 1.5.1 gave some basic definitions and properties of Markov chains on arbitrary state spaces. There we defined $\mathcal{L}^2$ as the space of all real functions which are measurable and square integrable with respect to π. This is a Hilbert space, where the inner product $< \cdot, \cdot >$ is given by the covariance and the norm $\| \cdot \|$ by the standard deviation with respect to π. The notions of a projection and the angle between subspaces are well understood for Hilbert spaces, and this is why $\mathcal{L}^2$ turns out to be the natural space to describe the Gibbs sampler. We will actually restrict attention to the subspace $\mathcal{L}_0^2$ (see Section 1.5.2); see Liu et al. (1994) for details on this choice.

Recall from Section 1.5.2 that the Markov chain induced by the two-component Gibbs sampler is denoted as $\{Z_n\}$ where each Z_n is partitioned as $Z_n = (Z_n^{(1)}, Z_n^{(2)})$. The marginal chains $\{Z_n^{(i)}\}$, $i = 1, 2$ are also Markov. The invariant distribution of the chain, which is the limiting distribution as well under mild conditions (see Section 1.5.1 and Section 1.5.2), is π. The results of this section are based on the assumption that $Z_0 \sim \pi$.

In the sequel, indexing of expectations with π implies that they are taken with respect to the stationary measure, otherwise with respect to transition kernel $P(\cdot, \cdot)$ of the Markov chain. The latter is defined in Definition 1.5.1. The transition kernels corresponding to the marginal Markov chains $\{Z_n^{(1)}\}$ and $\{Z_n^{(2)}\}$ are denoted by P_1 and P_2 respectively.

The transition kernel P of a Markov chain acts as an operator on $\mathcal{L}_0^2$,

$$Pf(x) := \mathsf{E}[f(Z_1) \mid Z_0 = x], \ f \in \mathcal{L}_0^2.$$

The $\mathcal{L}^2$ rate of convergence is understood as the rate at which expectations $P^n f$ of arbitrary square-integrable functions $f \in \mathcal{L}_0^2$ converge to their stationary values πf (defined in Section 1.5.2) as $n \to \infty$ according to the $\mathcal{L}^2$ norm. This type of convergence is considered in Roberts and Sahu (1997), Amit (1991), Goodman and Sokal (1989) among others.

P is a linear continuous operator (see Section 5 of Rynne and Youngson (2000) for an introduction to operator theory), therefore we can define its norm $\|P\| = \sup \mathrm{Var}_\pi^{1/2}[Pf(X)]$, where the supremum is taken over all $f \in \mathcal{L}_0^2$ with unit variance. We can also define the

spectrum of P, which is the set of all complex numbers λ such that $P - \lambda I$ is not invertible, where I is the identity operator. The spectral radius of P is the maximum modulus λ in its spectrum, and will be denoted here as $\text{spec}(P)$. Since we defined $\|P\|$ as a supremum over unit-variance functions, it is easy to see that $\|P\| \leq 1$. We define the spectral gap for P as $1 - \text{spec}(P)$. The $\mathcal{L}^2$ rate of convergence is given by the spectral radius of P, see for example Roberts (1996) for a discussion on this result for discrete state-spaces. When P is self-adjoint, its spectrum is simply the set of its eigenvalues, the corresponding Markov chain is reversible and $\text{spec}(P) = \|P\|$. For non-self-adjoint kernels the following identity links the two quantities

$$\text{spec}(P) = \lim_{n\to\infty} \|P^n\|^{1/n} \tag{2.1}$$

from which follows (as an effect of the triangle inequality) that $\text{spec}(P) \leq \|P\|$.

In the Gibbs sampler, P is a product of the component-updating kernels, $P = P_k P_{k-1} \cdots P_1$, see Section 1.5.2. Amit (1991) observed that P_i as an operator in $\mathcal{L}_0^2$, is actually an orthogonal projection onto the space

$$V_{-i} = \{f \in \mathcal{L}_0^2 : \ f(X) = f(Y) \text{ if } X^{(-i)} = Y^{(-i)}\}, \ i = 1, \ldots, k \tag{2.2}$$

that is, the space of $\mathcal{L}_0^2$ functions constant with respect to their ith argument.

Assuming that $k = 2$, the angle ϕ between V_{-1} and V_{-2}, which are closed subspaces of a Hilbert space, is defined (see expression (4) of Amit (1991)) in any of the following equivalent ways

$$\begin{aligned} \cos(\phi) &= \sup\{\text{Corr}_\pi\{f(X), g(X)\}, \ f \in V_{-1}, g \in V_{-2}\} && (2.3) \\ &= \sup\{\text{Var}_\pi^{1/2}[P_1 f(X)], \ f \in V_{-1}\} && (2.4) \end{aligned}$$

Lemma 1 of Amit (1991) shows that

$$\begin{aligned} (\text{spec}(P))^{1/2} = \cos(\phi) &= \sup\{\text{Corr}_\pi\{h(Z^{(1)}), g(Z^{(2)})\}, \ h, g \in \mathcal{L}_0^2\} \\ &= \sup_{h:\text{Var}_\pi[h(Z_0^{(1)})]=1} \text{Var}_\pi^{1/2}[\mathsf{E}[h(Z_0^{(1)}) \mid Z_0^{(2)}]]. \end{aligned} \tag{2.5}$$

Therefore, for the two-component Gibbs sampler (2.5) directly links the convergence rate with the correlation structure of the target distribution π. Amit (1991) also provides some bounds on the convergence rate of the Gibbs sampler when $k > 2$ using the angles between the relevant subspaces.

Liu et al. (1995) study the covariance structure of the two-component Gibbs sampler with a view to comparing different estimators and augmentation schemes and link the result

of Amit (1991) with the notion of maximal correlation. The maximal correlation between two random variables $W,\ V$ is defined as

$$\gamma(W,V) = \sup_{h,g\in\mathcal{L}_0^2} \mathrm{Corr}\{h(W), g(V)\}. \tag{2.6}$$

There is also an alternative expression

$$\gamma(W,V) = \sup_{h:\mathrm{Var}[h(W)]=1} \mathrm{Var}^{1/2}[\mathsf{E}[h(W) \mid V]], \tag{2.7}$$

which is much more convenient to handle than (2.6) and it is in this form that the maximal correlation has been found to be useful in the literature, see for example Liu and Wu (1999), Meng and van Dyk (1999) and Meng and van Dyk (2001). The equivalence of (2.6) and (2.7) follows essentially from the corresponding definitions of the angle between two closed subspaces of a Hilbert space, given in (2.3) and (2.4).

The maximal correlation has been used as a general measure of dependence (see for example Breiman and Friedman (1985); Lancaster (1958)) and it has three important properties (see Breiman and Friedman (1985))

1 $0 \le \gamma(W,V) \le 1$.

2 $\gamma(W,V) = 0$ if and only if W and V are independent.

3 If there exist measurable functions $h,\ g$, with $\mathrm{Var}\,[h(W)] > 0$, such that $h(W) = g(V)$ then $\gamma(W,V) = 1$.

In the context of the two-component Gibbs sampler, Theorem 3.2 of Liu et al. (1995) shows that, when the algorithm is started in stationarity, the maximal one-lag autocorrelation of the Markov chain $\gamma(Z_0, Z_1)$ is the same as the maximal correlation between the updated variables under the stationary measure $\gamma(Z_0^{(1)}, Z_0^{(2)})$. They also give a probabilistic proof of the result by Amit (1991) that the rate of convergence is given by $\{\gamma(Z_0^{(1)}, Z_0^{(2)})\}^2$, which is based on the interleaving Markov property. For reasons of completeness, the theorem is stated below.

Theorem 2.1.1. *(Liu et al. (1995)) Assuming that $Z_0 = (Z_0^{(1)}, Z_0^{(2)}) \sim \pi$ then*

$$\gamma(Z_0, Z_1) = \gamma(Z_0^{(1)}, Z_0^{(2)})$$

and

$$\gamma(Z_0^{(1)}, Z_1^{(1)}) = \gamma(Z_0^{(2)}, Z_1^{(2)}) = \{\gamma(Z_0^{(1)}, Z_0^{(2)})\}^2.$$

Hence $\|P\|^2 = \|P_1\| = \|P_2\|$. However, the spectral radii of P, P_1, P_2 are all the same and equal to $\{\gamma(Z_0^{(1)}, Z_0^{(2)})\}^2$.

It has long been recognised that the correlation structure of the target distribution determines the convergence behaviour of the corresponding Gibbs sampler, see Hills and Smith (1992), Gelfand et al. (1995); (2.5) makes this connection precise. However, the characterisation given in (2.5) is of little practical use since in general it will be impossible to evaluate the supremum over all functions $h \in \mathcal{L}_0^2$. We can however restrict the space of functions over which the supremum in (2.7) or in (2.5) is taken by making use of the following important results. When the joint distribution of (W, V) is bivariate Gaussian it has been shown (for example in Lancaster (1958)) that the maximal correlation coincides with the absolute value of the correlation coefficient. When (W, V) has a multivariate Gaussian distribution then the function f leading to maximal correlation must be linear, that is if $W = (W_1, \ldots, W_n)$ then $h(W) = \sum \alpha_i W_i$ for some real coefficients α_i. This result was originally proved by Kolmogorov, see Breiman and Friedman (1985) for some references. Notice also that if $t(V)$ is a sufficient statistic for the conditional distribution of $W \mid V$ then $\gamma(W, V) = \gamma(W, t(V))$. Another important application of the characterisation in (2.5) is in comparing different data augmentation schemes, see for example Meng and van Dyk (1999), Liu and Wu (1999) and Section 7.9.2 of this thesis.

2.1.1 Gibbs sampler on Gaussian target distributions

Explicit convergence rates can be obtained when the target distribution of the Gibbs sampler is multivariate Gaussian. This case is studied thoroughly in Roberts and Sahu (1997), where different updating schemes, blocking strategies and parameterisation issues are investigated. The results for the Gaussian target distributions hold asymptotically for other targets (see Roberts and Sahu (2001)) and therefore provide useful intuition on how the Gibbs sampler might work in a variety of models. The following sections rely on these results to compare different parameterisations of Gaussian hierarchical models.

Assume that the state space $\mathcal{Z} = \mathbb{R}^d$ is decomposed in k components $\mathbb{R}^{r_i}$, for $1 \leq i \leq k$ with $\sum_{i=1}^k r_i = d$. Let π correspond to the multivariate Gaussian density, where without loss of generality, we can take its mean to be $0 \times \mathbf{1}$ ($\mathbf{1}$ is a $k \times 1$ vector of ones here), since as we mentioned in Section 1.5.2 component-wise transformations do not affect the probabilistic behaviour of the Gibbs sampler. Therefore, under this decomposition of $\mathcal{Z}$ the Gibbs sampler on π induces a k-dimensional Markov chain $Z = \{Z_n\}$, where $Z_n = (Z_n^{(1)}, \ldots, Z_n^{(k)})$ and the dimensionality of the random vector $Z_n^{(i)}$ is r_i. We denote the inverse covariance matrix for

π by Q, and partition it accordingly as

$$Q = \begin{pmatrix} Q_{11} & \dots & Q_{1k} \\ \vdots & & \vdots \\ Q_{k1} & \dots & Q_{kk} \end{pmatrix} \tag{2.8}$$

where each Q_{ij} is an $r_i \times r_j$ matrix. Roberts and Sahu (1997) (Lemma 1) notice that the Markov chain induced by the Gibbs sampler on the Gaussian target π, has a Gaussian transition density with mean

$$\mathsf{E}[Z_1 \mid Z_0] = BZ_0$$

and variance matrix $Q^{-1} - BQ^{-1}B^t$, where the matrix B is derived below.

We write $\mathrm{diag}(Q_{11}^{-1}, \dots, Q_{kk}^{-1})$ for the matrix

$$\begin{pmatrix} Q_{11}^{-1} & \mathbf{0}_{r_1 \times r_2} & \dots & \mathbf{0}_{r_1 \times r_k} \\ \vdots & \ddots & & \vdots \\ \mathbf{0}_{r_k \times r_1} & \dots & \dots & Q_{kk}^{-1} \end{pmatrix}$$

and set

$$A = I - \mathrm{diag}(Q_{11}^{-1}, \dots, Q_{kk}^{-1})Q \ . \tag{2.9}$$

Next we need to consider the decomposition of A into its upper and lower-triangular matrices. Partitioning A in k^2 sub-matrices as in the partition of Q, we define the lower triangular matrix L by

$$L = \begin{pmatrix} \mathbf{0} & \dots & \dots & \mathbf{0} \\ A_{21} & \ddots & & \vdots \\ \ddots & \ddots & \ddots & \vdots \\ A_{k1} & \dots & A_{k,k-1} & \mathbf{0} \end{pmatrix} \tag{2.10}$$

and set U to be the upper triangular matrix, $U = A - L$. Then, $B = (I - L)^{-1}U$.

Theorem 1 of Roberts and Sahu (1997) stated below describes how we can use the B matrix constructed above to calculate the rate of convergence of the Gibbs sampler on a Gaussian target distribution:

Theorem 2.1.2. *(Roberts and Sahu (1997)) The Gibbs sampler on π produces a multivariate*

Gaussian AR(1) process and its $\mathcal{L}^2$ rate of convergence ρ, can be characterised as

$$\rho = spec(B), \tag{2.11}$$

therefore ρ is the maximum modulus eigenvalue of B.

2.1.2 Measures of efficiency

When the $\mathcal{L}^2$ rate of convergence ρ of an ergodic Markov chain is available, some informative measures of its efficiency can be calculated. Some of these measures will appear later in this chapter, especially in Section 2.5.

The spectral theory of bounded self-adjoint operators on a Hilbert space provides some powerful representations which can be used to assess the efficiency of reversible Markov chain Monte Carlo methods. Suppose that $\{Z_n,\ n = 1, 2, \ldots\}$ is an ergodic Markov chain (see Section 1.5.1 and Section 3.1.3) with stationary distribution π, and f is a square integrable function with respect to π. The transition operator is denoted by P, as in Section 2.1 and $\rho = \text{spec}(P)$. We also assume that the chain starts in stationarity, $Z_0 \sim \pi$, and that P is a self-adjoint operator, therefore the Markov chain is reversible. Ergodic average estimators of the form

$$S_N = \frac{1}{N}\sum_{i=1}^{N} f(Z_i)$$

were shown to converge to πf in Section 1.5.1 under mild conditions. The error of the estimator can be assessed by its variance $\text{Var}\,(S_N)$. It can be shown (see for example Theorem 3.2 of Roberts (1996) and Theorem 2.1 of Geyer (1992)) that as $N \to \infty$

$$N\text{Var}\,(S_N) \to \sigma_f^2 := \text{Var}_\pi[f(Z_0)] + 2\sum_{i=1}^{\infty} \text{Cov}\{f(Z_0), f(Z_i)\};$$

if $\sigma_f^2 < \infty$ we have a central limit theorem (Theorem 2.1 of Geyer (1992))

$$\sqrt{n}(S_N - \pi f) \xrightarrow{d} S,\ S \sim \text{N}(0, \sigma_f^2).$$

Moreover, the autocovariance function admits the spectral representation (see for example Section 2 of Geyer (1992))

$$\text{Cov}\{f(Z_0), f(Z_i)\} = \int_{-1}^{1} \lambda^i dE_f(\lambda)$$

where E_f is the spectral measure on $[-1, 1]$ associated with f. If $\rho < 1$ (which is true when

the chain is geometrically ergodic; see Section 3.1.5) then, since $(1+\lambda)/(1-\lambda)$ is increasing in λ and $|\lambda| \leq \rho$ for all λ in the spectrum of P by definition, it follows that

$$\sigma_f^2 = \int_{-1}^{1} \frac{1+\lambda}{1-\lambda} dE_f(\lambda) \leq \frac{1+\rho}{1-\rho} \mathrm{Var}_\pi[f(Z_0)]. \tag{2.12}$$

Let

$$\tau_f := \frac{\sigma_f^2}{\mathrm{Var}_\pi[f(Z_0)]} \leq \frac{1+\rho}{1-\rho}, \tag{2.13}$$

in the literature τ_f is called the integrated autocorrelation time of the Markov chain corresponding to the function f. By definition of σ_f and due to (2.13) it follows that $1 \leq \tau_f \leq (1+\rho)/(1-\rho)$. The upper bound of this inequality corresponds to the asymptotic error of the ergodic averages in estimating the "worst" functions, although many other interesting functions might be mixing much faster and having much smaller asymptotic errors.

Actually, when the chain is a Gaussian vector autoregressive process of order one, then it can be shown (see for example Pitt and Shephard (1999)) that $\tau_f = (1+\rho)/(1-\rho)$ and τ_f measures approximately for large N, the sample size that we should require from our Markov chain to estimate πf to the same accuracy as N independent draws from π, where f is some linear function. Similarly a natural way to compare the relative efficiency of two Markov chains with the same stationary distribution π but with different convergence rates, $\rho_c,\ \rho_{nc}$ say, is by calculating $(1+\rho_{nc})(1-\rho_c)/[(1-\rho_{nc})(1+\rho_c)]$.

This thesis is primarily interested in the Gibbs sampler and its variants, therefore the transition operator P of the associated Markov chains is not self-adjoint and the spectral theory results presented above do not hold. However, notice that for the two-component Gibbs sampler each of the marginal chains is reversible, which is proved for example in Lemma 3.1 of Liu et al. (1994). Therefore, these measures can be used to assess efficiency of the sampler in estimating marginal expectations.

Another quantity that contains valuable information about the efficiency of the Markov chain is $-1/\log\{\rho\}$. This is proportional to the number of iterations needed for $P^n f$ to be within a given accuracy close to πf, and it is used in Section 2.5.

A more informal way of assessing the efficiency of a Markov chain $\{Z_n\}$ in estimating expectations under the invariant measure is looking at sample-based estimates $\hat{r}(n)$ of the theoretical autocorrelations

$$r(n) = \mathrm{Corr}\{f(Z_0), f(Z_n)\},\ n = 0, 1, \ldots$$

where f is the function whose expectation we want to estimate. This summary is easy to produce and is informative about the error of ergodic averages in estimating expectations of interest. The $\hat{r}(n)$ are computed using standard time series techniques (for example by dividing appropriate sample means), after discarding a number of initial iterations which are

believed to belong to the transient phase of the chain.

Clearly, determination of this number of iterations is not straightforward and there are various diagnostic tools which have been developed to this end, see for example Brooks and Roberts (1999), Cowles et al. (1996) and references therein as well as the web page

`http://www.statslab.cam.ac.uk/mcmc/pages/links.html`

for some useful related links. There is also the BUGS software package, operating within the S-PLUS and R environment, for performing convergence diagnostics analysis for Gibbs sampler output, see Gilks et al. (1994). Diagnostic tools convey important information about the behaviour of the Markov chain, however they do not prove convergence. In this thesis we don't use any of this technology, instead we shall largely look at trace plots in order to assess the required number of iterations to be discarded. We are primarily interested in the covariance structure of the Gibbs sampler and how this can be improved using non-centered parameterisations, thus convergence of the algorithm from different starting values is somewhat tangential to our considerations. This is why we employ this rather informal method, which however is widely accepted within the MCMC community.

When the autocorrelations remain non-negligible for large number of iterations this is an indication of a slowly mixing chain, which will produce estimates of πf with substantial Monte Carlo error. Actually, if the chain is reversible Geyer (1992) shows that the theoretical autocorrelations under stationarity are non-negative for all lags, therefore it is of interest to plot the estimated autocorrelations for as many lags as they are clearly positive, since negative values are due to the sample variation.

In models as those in Chapter 4, Chapter 6 and Chapter 7 where we cannot calculate analytic convergence rates, we compare different algorithms in terms of these estimated autocorrelations. This is a very rough comparison necessitated by the difficulty in obtaining analytic rates for complex models, whereas the bounds on these rates which some methods provide (see for example Chapter 11 of Roberts and Tweedie (2004)) are typically extremely conservative and very hard to obtain. For the two-component Gibbs sampler the maximal lag-one autocorrelation of the marginal chains coincides with the $\mathcal{L}^2$ rate of convergence, but it is not possible to find which function maximises this autocorrelation unless the target is Gaussian. Nevertheless, this assessment is still useful since it gives an idea of the relative efficiency of different MCMC algorithms with the same target in estimating the expectation of a function of interest. We use such comparison techniques in the following chapters and some more discussion is given in Section 6.12.1, in the specific context of the application considered there.

2.2 Parameterisations of hierarchical models

Section 1.7 introduced the two types of parameterisations of hierarchical models that we will consider in this thesis: the centered (CP), with the conditional independence graph shown in Figure 1.3, and the non-centered (NCP), with corresponding graph shown in Figure 1.6.

The two-component Gibbs sampler (see Section 1.5.2) under the centered parameterisation, simulates iteratively from the conditional distribution of X given Θ and Y, and Θ given X. We call this the centered algorithm (CA). Alternatively, the Gibbs sampler under the non-centered parameterisation simulates iteratively from the conditional distribution of $\tilde{X}$ given Θ and Y, and Θ given $\tilde{X}$ and Y. We call this the non-centered algorithm (NCA). The rest of this chapter is devoted to the calculation and comparison of the $\mathcal{L}^2$ convergence rate of the CA and the NCA for different Gaussian three-stage hierarchical models.

For non-Gaussian models analytic calculation of convergence rates for the Gibbs sampler under any of the suggested parameterisations is typically impossible. Moreover, it might not even be possible to implement a "pure" Gibbs sampler, instead we need to resort to the more general componentwise-updating algorithms like the Hastings-within-Gibbs (see Section 1.5.2). Section 4.1 defines carefully the centered and non-centered parameterisation for arbitrary hierarchical models and exposes how the corresponding Hastings-within-Gibbs algorithms are implemented. Nevertheless, the analytic results of this chapter for the relatively simple Gaussian models will help us gain valuable intuition about non-centering for the much more complicated models that we consider in the following chapters.

2.3 The normal hierarchical model

In this section we calculate analytic results for the rate of convergence of the CA and the NCA for the normal hierarchical model (1.12) which was discussed in Section 1.6; see also (2.14) below. This is a toy example which, however, serves nicely to illustrate and motivate some of the main ideas presented in this thesis. This model has also been used for pedagogical purposes by Liu and Wu (1999).

We initially assume that an improper uniform prior is chosen for Θ, which however can be shown (see Lindley and Smith (1972)) to lead to a proper posterior. We also assume that the variances $\sigma_x^2,\ \sigma_y^2$ are known, which is essential in order to be able to derive analytic results of the convergence rate. Notice that the sample average $\overline{Y_i} = \sum_j Y_{ij}/n$ is sufficient for X_i and has normal distribution with mean X_i and variance σ_y^2/n. Therefore, when σ_y^2 is assumed known multiple observations per X_i is equivalent to rescaling the observation error variance σ_y^2. Hence with no loss of generality, we take $n = 1$ in (1.12), and the model

rewrites as

$$\begin{aligned} Y_i &= X_i + \sigma_y \epsilon_i \\ X_i &= \Theta + \sigma_x z_i,\ i = 1, \ldots, m\ . \end{aligned} \tag{2.14}$$

The CP for this model is (X, Θ), where $X = (X_1, \ldots, X_m)$. It is easy to see that the joint distribution of (X, Θ) is multivariate Gaussian with precision matrix given by

$$Q = \begin{pmatrix} (1/\sigma_x^2 + 1/\sigma_y^2) I_m & -1/\sigma_x^2 \mathbf{1}^t \\ -1/\sigma_x^2 \mathbf{1} & m/\sigma_x^2 \end{pmatrix}\ .$$

We can use the results of Section 2.1.1 but also of Section 2.1 to compute the rate of convergence of the CA, denoted by ρ_c. This was originally done in Roberts and Sahu (1997) (see Section 4.1, but also Section 14.2 of Roberts and Tweedie (2004)), where it was found that

$$\rho_c = 1 - \frac{\sigma_x^2}{\sigma_x^2 + \sigma_y^2} = \frac{\sigma_y^2}{\sigma_x^2 + \sigma_y^2}. \tag{2.15}$$

This expression reveals automatically that the CA works well when the observation error is small compared to the variance of the random effects. There is an interesting connection between the expression (2.15) and a statistical concept crucial in the analysis of missing data problems, that of the *Bayesian fraction of missing information* introduced by Rubin (1987). For a fixed real function $h(\Theta)$, it can be defined as the ratio

$$\kappa = \frac{\mathrm{Var}\left[\mathsf{E}[h(\Theta) \mid Y, X] \mid Y\right]}{\mathrm{Var}\left[h(\Theta) \mid Y\right]} = 1 - \frac{\mathsf{E}[\mathrm{Var}\left[h(\Theta) \mid Y, X\right] \mid Y]}{\mathrm{Var}\left[h(\Theta) \mid Y\right]}. \tag{2.16}$$

It follows that $\rho_c = 1 - \kappa$, where (2.16) is computed for linear functions of Θ. We can interpret κ as the proportion of extra variation caused by not observing X when making inference about $h(\Theta)$. The posterior mean for X_i has a weighted average form

$$\begin{aligned} \mathsf{E}[X_i \mid Y, \Theta] &= \kappa Y_i + (1 - \kappa)\Theta \\ \mathsf{E}[X_i \mid Y] &= \kappa Y_i + (1 - \kappa)\overline{Y} \end{aligned} \tag{2.17}$$

which shows that κ (evaluated for the identity function) is the weight given on the data point Y_i when predicting (under square loss function) the underlying X_i; see Lindley and Smith (1972) for more details on this. Notice that Y_i is the least squares estimator of X_i when we ignore the prior dependence among the X_is. This point estimate is pooled towards the population average $\overline{Y}$ with weight $1 - \kappa$. Actually, we will see in Section 7.8 that posterior

expectations of the canonical parameters in generalised hierarchical models admit similar weighted average forms.

Liu (1994b), using some of the results of Liu et al. (1994), shows that the rate of convergence of the two-component Gibbs sampler equals the maximal Bayesian fraction of missing information, which is obtained as the supremum of (2.16) over all functions h with unit variance. There is a frequentist version of the ratio (2.16) known as the fraction of missing information, which instead of dividing posterior variances it divides the corresponding Fisher informations evaluated either at the true parameter values or at maximum likelihood estimates. It is known that the latter characterises the rate of convergence of the EM and ECM algorithm (see Meng and Rubin (1993), Meng and van Dyk (1997), but also Section 7.10 of this thesis). The two ratios coincide when the joint distribution of (X, Θ) is Gaussian. These connections have been used to find approximations to the Gibbs sampler convergence rate for non-Gaussian target distributions, see for example Sahu and Roberts (1999), Roberts and Sahu (2001), Meng and van Dyk (2001). Moreover, techniques which have been employed to improve the convergence rate of EM algorithm have been found to be successful in improving the Gibbs sampler, see for example Meng and van Dyk (1999), Sahu and Roberts (1999), Liu and Wu (1999) and Section 7.10 of this thesis.

The result in (2.15) can be derived by finding the maximal correlation between X and Θ. As was observed in Section 2.1 this is given by

$$\mathrm{Corr}(\sum X_i, \Theta \mid Y) \;=\; \sqrt{1-\kappa}.$$

An NCP for this model can be constructed by writing (2.14) in an equivalent form as

$$\begin{aligned} Y_i &= \Theta + \tilde{X}_i + \sigma_y \epsilon_i \\ \tilde{X}_i &= \sigma_x z_i,\ i = 1, \ldots, m \end{aligned} \tag{2.18}$$

where $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_m)$ is *a priori* independent of Θ and $X = \tilde{X} + \mathbf{1}\Theta$. Actually, it is in this form that the normal hierarchical model often appears in the literature, see for example Tanner and Wong (1987). $(\tilde{X}, \Theta)$ has multivariate Gaussian distribution with precision matrix

$$Q = \begin{pmatrix} (1/\sigma_x^2 + 1/\sigma_y^2) I_m & \sigma_y^2 \mathbf{1}^t \\ \sigma_y^2 \mathbf{1} & m/\sigma_y^2 \end{pmatrix}$$

and similar calculations as for the CA yield that the rate of convergence of the NCA, denoted

by ρ_{nc}, is

$$\rho_{nc} = \kappa. \tag{2.19}$$

Notice that for the simple normal hierarchical model the rates ρ_c and ρ_{nc} do not depend on the sample size m. This is not true in general for other partially exchangeable models. Moreover, $\tilde{X}$ and X are orthogonal marginally in an $\mathcal{L}^2$ sense, that is $\text{Cov}\{X, \tilde{X} \mid Y\} = 0$ (see Figure 7.1). This explains why $\rho_c = 1 - \rho_{nc}$, bearing in mind Amit's characterisation of the convergence rate as the cosine of the angle between spaces presented in Section 2.1. This property shows that CA and NCA are complementary of each other, in the sense that the one performs best when the other is very poor and provides some justification on why these two parameterisations are natural competitors. Moreover, it inspires the partially non-centered methods which will be developed in Chapter 7. However it is not preserved in more general models, even inside the Gaussian family, as we shall see in Section 2.4.

2.3.1 Brownian motion interpretation

Intuition into how the Gibbs sampler performs under the CP and the NCP can be gained by expressing the model (2.14) inside a simple Brownian motion context.

For reasons of exposition suppose that $m = 1$ (it is straightforward to generalise the ideas for arbitrary m). Then the data Y can be seen as the value of a Brownian motion, $\tilde{X}(\cdot)$ say (the choice of such notation will become clear in the Chapter 4), at time $\sigma_y^2 + \sigma_x^2$, that has been started at time 0 from initial value Θ. The random effect X can be obtained as the value of the Brownian motion at time σ_x^2. Therefore, we want to infer about the values of the Brownian motion at time 0 and σ_x^2 conditional on its value Y at time $\sigma_y^2 + \sigma_x^2$. Under this setting, the total time $\sigma_y^2 + \sigma_x^2$ represents the marginal uncertainty about Θ while the time proportion $\kappa = \sigma_x^2/(\sigma_x^2 + \sigma_y^2)$ represents the relative strengths of the prior and the data in model (2.14).

The Gibbs sampling algorithm based on the CP iterates between the two steps

1 Simulate $X \mid \Theta, Y \sim N(\kappa Y + (1-\kappa)\Theta, \sigma_y^2\kappa)$;

2 Simulate $\Theta \mid X, Y \sim N(X, \sigma_x^2)$

These steps have natural representations in terms of simulations from a Brownian bridge $B(\cdot)$, which is a Brownian motion conditioned to take a prescribed value at some fixed time in the future; see Section 1.8. Figure 2.1 shows the first step of this simulation.

1 Simulate a Brownian bridge $B(t)$, $0 \le t \le \sigma_y^2 + \sigma_x^2$ forwards in time starting from Θ and hitting Y at time $\sigma_y^2 + \sigma_x^2$. Set $X = B(\kappa(\sigma_y^2 + \sigma_x^2))$

2 Simulate a Brownian motion $\tilde{X}(t),\ 0 \leq t \leq \kappa(\sigma_y^2 + \sigma_x^2)$ backwards in time started from X at time $\kappa(\sigma_y^2 + \sigma_x^2)$. Set $\Theta = \tilde{X}(0)$.

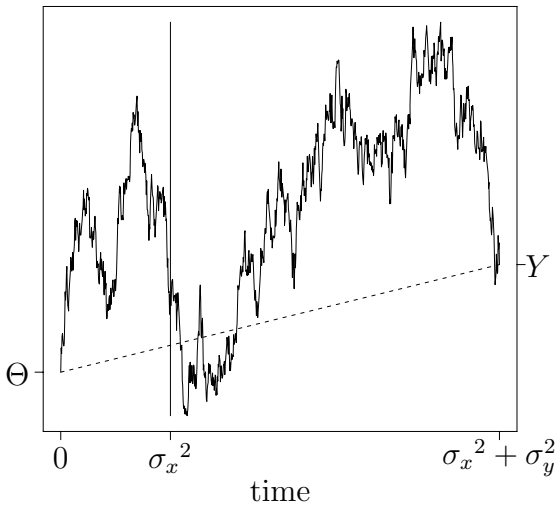


Figure 2.1: Updating of X given Y and Θ as a Brownian bridge simulation: simulate a Brownian bridge starting at time 0 from Θ and hitting Y at time $\sigma_y^2 + \sigma_x^2$, obtain X as the value of the bridge at time $\kappa(\sigma_y^2 + \sigma_x^2) = \sigma_x^2$.

When κ is close to one, the data Y dominate in the sense that the Brownian bridge in Step 1 has to hit Y at time $\sigma_y^2 + \sigma_x^2$ and therefore its value at time $\kappa(\sigma_y^2 + \sigma_x^2)$, will be approximately independent from its initial value Θ. This is exactly the case where the centered parameterisation is preferable. On the other hand, when κ is close to zero, the prior will dominate since the Brownian bridge starts from Θ and the value at time $\kappa(\sigma_y^2 + \sigma_x^2)$ will be very close to it, regardless essentially of the data. This is the case when the non-centered is to be preferred.

This interpretation offers insight into the behaviour of the CA and the NCA, by translating dependence into time. More importantly though, in Chapter 4 we shall show how this interpretation on the expanded state space can be used to construct new classes of non-centered algorithms where variance parameters are not assumed to be known, that is when inference for either σ_x^2 or σ_y^2 is of interest. Similar interpretations can be attempted for other hierarchical models with additive structure.

Finally, this interpretation has close links with the semi-parametric model introduced in Neal (2001) where branching Brownian motions are used to model hierarchical dependence structure explicitly.

2.3.2 Effect of prior distribution on the rate of convergence

Instead of choosing an improper prior for Θ we could choose a proper conjugate Gaussian prior, $\Theta \sim \mathrm{N}(\mu, v^2)$ where μ and v^2 have fixed values. Actually, as $1/v^2 \to 0$ this prior converges to the improper uniform prior. The posterior distribution of (X, Θ) is still multivariate Gaussian and the precision matrix remains unaltered except for the diagonal element of the last row which becomes $m/\sigma_x^2 + 1/v^2$. Similarly, that element of the precision matrix of $(\tilde{X}, \Theta)$ becomes $m/\sigma_y^2 + 1/v^2$ and it is easy to show that

$$\begin{aligned} \rho_c &= \frac{v^2}{\sigma_x^2 + v^2}(1-\kappa) \\ \rho_{nc} &= \frac{v^2}{\sigma_y^2 + v^2}\kappa. \end{aligned}$$

As it can be seen both algorithms improve their performance for any $1/v^2 > 0$, while their relative performance is

$$\begin{aligned} \frac{1-\rho_{nc}}{1-\rho_c} &= \frac{\sigma_y^2 + v^2(1-\kappa)}{\sigma_x^2 + v^2\kappa}\frac{\sigma_x^2 + v^2}{\sigma_y^2 + v^2} \\ &= \frac{1-\kappa}{\kappa}\frac{\kappa/v^2 + 1/(\sigma_y^2 + \sigma_x^2)}{(1-\kappa)/v^2 + 1/(\sigma_y^2 + \sigma_x^2)}. \end{aligned} \tag{2.20}$$

Notice that whether (2.20) is less than one or not, which can be used as a criterion to decide whether to use a centered or a non-centered algorithm, does not depend on the choice of v^2 but only on whether $\kappa < 1/2$. However, the ratio of the convergence rates in (2.20), which also appears in (2.47), does not correspond to some relative measure of efficiency, as for example those described in Section 2.1.2, although it is related to the more informative ratio $\log\{\rho_{nc}\}/\log\{\rho_c\}$ as $\rho_c, \rho_{nc} \to 1$ (see for example Section 2 of Pitt and Shephard (1999)).

Notice that for fixed κ and v^2 if we let $\sigma_y^2 + \sigma_x^2 \to \infty$ then $\rho_c, \rho_{nc} \to 0$ both at the same speed, since the ratio in (2.20) tends to one. This is expected, since if we have very strong prior beliefs about Θ, its posterior dependence with both X and $\tilde{X}$ will be very small.

2.4 A general normal hierarchical model

We will now describe a very general form of the normal hierarchical model (1.12) that encompasses most of the Gaussian models used in practice. We will show how to calculate the rate of convergence for the CA and the NCA and discuss how these results compare with those derived above for the simpler model. Although our results are simple and based on well known properties of the normal hierarchical model as developed by Lindley and Smith (1972), they will prove very useful in the construction of partially non-centered parameterisations

for Gaussian and non-Gaussian models in Chapter 7.

The data are $n_i \times 1$ vectors Y_i , $i = 1, \ldots, m$ and a set of covariates $C_{1i},\ n_i \times p$, is observed for each data point. The random effects are $X_i,\ p \times 1$ and the model is written as

$$\begin{aligned} Y_i &= C_{1i}X_i + (V_i)^{1/2}\epsilon_i \\ X_i &= C_2\Theta + D^{1/2}z_i \ i = 1, \ldots, m \end{aligned} \tag{2.21}$$

where $\epsilon_i,\ z_i$ are vectors of independent standard normal random variables with the appropriate dimensions, V_i is the covariance matrix of the Y_i vector conditional on the X_i, Θ is a $q \times 1$ vector and C_2 is another design $p \times q$ matrix. We will assume for convenience that $C_2^T C_2$ is invertible, as will usually be the case. Actually, in many cases C_2 is a $p \times 1$ vector of ones and then $C_2^T C_2 = p$. Conditional on Θ the X_is are independent. The variance matrices will be assumed known and inference will be made for the missing data $X = (X_1, \ldots, X_m)$ and Θ. The (X, Θ) is the CP for (2.21), while the NCP is given by $(\tilde{X}, \Theta)$ where $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_m)$ and $\tilde{X}_i = X_i - C_2\Theta$.

For $V_i = \sigma_i^2 I_{n_i},\ C_2 = I_p$ and Θ a $p \times 1$ vector, model (2.21) coincides with the model considered in Gelfand et al. (1996). In this paper, the authors compare centered and non-centered parameterisations for this model, however, without using the methodology presented in Section 2.1.1 which can produce exact results. Instead they try to find the parameterisation that minimises the correlation between the X_is and Θ and concluded that the CP is likely to be preferred in most real-data applications. (2.21) is more general than the normal linear models studied by Lindley and Smith (1972), since it allows for unbalanced design, i.e the n_i to vary with i.

We will refer to index $i = 1, \ldots, m$ as *individuals* and to index $j = 1, \ldots, n_i$ for each i as *measurements*, borrowing the terminology from longitudinal data analysis where such models have been used extensively. An interpretation of model (2.21) is as follows: we believe that the measurements for each individual depend linearly on the corresponding covariates C_{1i}, although they are allowed to be correlated (serially correlated when j indexes time, as usually happens in repeated measurements studies). We believe that the regression coefficients differ for each individual, for example due to unmeasured covariates. However, they have an exchangeable structure and we further believe that they are *centered* around some population value $C_2\Theta$ with variance matrix D. Thus we can borrow strength when estimating the individual's coefficient from the information about the other individuals. When the observations can be thought of as a sample from an infinite population then (2.21) is a partially exchangeable model, like those discussed in Section 1.6.

Standard calculations (see Section 2 of Gelfand et al. (1996) and Lindley and Smith

(1972), also Chapter 9 of O'Hagan (1994)) yield:

$$\begin{aligned} Y_i \mid \Theta &\sim N(C_{1i}C_2\Theta, \Sigma_i) \\ \Sigma_i &= V_i + C_{1i}DC_{1i}^t \end{aligned}$$

therefore

$$\begin{aligned} \Theta \mid Y &\sim \mathrm{N}(\hat{\Theta}, T^{-1}) \\ T &= \sum T_i \\ \hat{\Theta} &= T^{-1}C^T\Sigma^{-1}Y \end{aligned} \tag{2.22}$$

where

$$\begin{aligned} T_i &= C_2^tC_{1i}^t\Sigma_i^{-1}C_{1i}C_2 \\ \Sigma &= \mathrm{diag}(\Sigma_1 \ldots \Sigma_m) \\ C^t &= (C_2^tC_{11}^t, \ldots, C_2^tC_{1m}^t) \\ Y^t &= (Y_1^t, \ldots, Y_m^t) \end{aligned} \tag{2.23}$$

If we define the $p \times p$ matrix

$$Q_i = C_{1i}^tV_i^{-1}C_{1i} + D^{-1} \tag{2.24}$$

then conditional on Θ

$$\begin{aligned} \mathsf{E}(X_i \mid Y, \Theta) &= Q_i^{-1}(C_{1i}^tV_i^{-1}Y_i + D^{-1}C_2\Theta) \\ \mathrm{Var}\,(X_i \mid Y, \Theta) &= Q_i^{-1} \end{aligned} \tag{2.25}$$

while marginally

$$\begin{aligned} \mathsf{E}(X_i \mid Y) &= Q_i^{-1}(C_i^tV_i^{-1}Y_i + D^{-1}C_2\hat{\Theta}) \\ \mathrm{Var}\,(X_i \mid Y) &= Q_i^{-1} + Q_i^{-1}D^{-1}T^{-1}D^{-1}Q_i^{-1} \\ \mathrm{Cov}(X_i, X_j \mid Y) &= Q_i^{-1}D^{-1}T^{-1}D^{-1}Q_j^{-1},\ i \neq j \\ \mathrm{Cov}(X_i, \Theta \mid Y) &= Q_i^{-1}D^{-1}C_2T^{-1}. \end{aligned} \tag{2.26}$$

These expressions can be derived using basic properties of the covariance operator, such as bilinearity and that $\mathrm{Cov}(X, Y) = \mathrm{Cov}(X, \mathsf{E}(Y \mid X))$ (see Chapter 5 of Whittaker (1990)) in conjunction with the results in (2.22). Notice that when we write $\mathrm{Cov}(X, Y)$ for two random vectors $X,\ Y$ we mean the matrix containing the covariances between each element of X

and each element of Y.

Our target is to compute the rate of convergence of the CA and the NCA for the normal hierarchical model (2.21). For the simpler (1.12) we have shown that this depends on κ, which is the ratio of the observed to augmented information. We will try to find the corresponding quantity for the more general model, although now this will clearly be a matrix and not a scalar. Considerable insight into this problem can be gained by assuming that $C_{1i}^t V_i^{-1} C_{1i}$ is invertible. This is done just to help gaining some understanding and none of our final results will depend on this assumption. In this case we can rewrite (2.26) as

$$\begin{aligned} \mathsf{E}(X_i \mid Y) &= Q_i^{-1}(C_{1i}^t V_i^{-1} C_{1i} \hat{X}_i + D^{-1} C_2 \hat{\Theta}) \\ \hat{X}_i &= (C_{1i}^t V_i^{-1} C_{1i})^{-1} C_{1i}^t V_i^{-1} Y_i \end{aligned} \tag{2.27}$$

which expresses the posterior expectation of X_i as a weighted average of the least squares estimator $\hat{X}_i$, which ignores the dependence among the X_is, and the estimate of the mean of the X_is, $C_2\hat{\Theta}$. $(C_{1i}^t V_i^{-1} C_{1i})^{-1}$ is the variance of the least squares estimator and therefore (2.24) is the sum of the observation and prior precision. Therefore $Q_i^{-1} D^{-1}$ corresponds to

$$1 - \kappa = \frac{\frac{1}{\sigma_\alpha^2}}{\frac{1}{\sigma_a^2} + \frac{1}{\sigma_e^2}} = \frac{\sigma_e^2}{\sigma_\alpha^2 + \sigma_e^2}, \tag{2.28}$$

$Q_i^{-1} C_{1i}^t V_i^{-1} C_{1i}$ corresponds to κ and clearly by the definition of (2.24)

$$Q_i^{-1} D^{-1} + Q_i^{-1} C_{1i}^t V_i^{-1} C_{1i} = I_p.$$

It is obvious that the weights given to the least squares estimator and the population mean should vary with i, as opposed to the constant weights for the simpler normal model, reflecting the heteroscedasticity among the Y_is introduced by the variance matrices V_i and the design matrices C_{1i}.

The following equality will prove helpful in many calculations,

$$Q_i^{-1} C_{1i}^t V_i^{-1} C_{1i} = D C_{1i}^t \Sigma_i^{-1} C_{1i}. \tag{2.29}$$

It can be proved by first showing that

$$Q_i^{-1} = D - D C_{1i}^t \Sigma_i^{-1} C_{1i} D$$

which is done by using a very important matrix identity, proved for example in Lindley and

Smith (1972) (see p. 5, formula (10)), and then by noticing that

$$\begin{aligned} Q_i^{-1}D^{-1} + DC_{1i}^t\Sigma_i^{-1}C_{1i} &= I_p \\ Q_i^{-1}D^{-1} + Q_i^{-1}C_{1i}^t V_i^{-1}C_{1i} &= I_p. \end{aligned}$$

In order to compute the rate of convergence of the CA for this model, we need to find the B-matrix defined in Section 2.1.1. By writing $\pi(\Theta, X \mid Y) = \pi(\Theta \mid Y)\pi(X \mid \Theta, Y)$ and using the results obtained above we can show that the precision matrix corresponding to the joint distribution of (X, Θ) has the following partitioned form:

$$Q^c = \begin{pmatrix} Q_1 & \mathbf{0} & \dots & \mathbf{0} & -D^{-1}C_2 \\ \mathbf{0} & Q_2 & \dots & \mathbf{0} & -D^{-1}C_2 \\ \vdots & \vdots & \ddots & \dots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & Q_m & -D^{-1}C_2 \\ -C_2^tD^{-1} & -C_2^tD^{-1} & \dots & -C_2^tD^{-1} & mC_2^tD^{-1}C_2 \end{pmatrix}. \tag{2.30}$$

We mention here that there are considerable computational and analytical advantages by partitioning Q^c as in (2.30), i.e with the elements corresponding to Θ in the bottom-right corner, rather than with those elements in the top-left corner, as is done in Roberts and Sahu (1997) for example. The main reason being that in doing so we need in the end to compute the eigenvalues of a $q \times q$ matrix rather than of a $p \times p$ matrix, and typically q will be much less than p.

The A-matrix is then

$$A^c = \begin{pmatrix} \mathbf{0} & \dots & \mathbf{0} & Q_1^{-1}D^{-1}C_2 \\ \mathbf{0} & \dots & \mathbf{0} & Q_2^{-1}D^{-1}C_2 \\ \vdots & \dots & \vdots & \vdots \\ \mathbf{0} & \dots & \mathbf{0} & Q_m^{-1}D^{-1}C_2 \\ \frac{1}{m}(C_2^tD^{-1}C_2)^{-1}C_2^tD^{-1} & \dots & \frac{1}{m}(C_2^tD^{-1}C_2)^{-1}C_2^tD^{-1} & \mathbf{0} \end{pmatrix} \tag{2.31}$$

from which we derive

$$B^c = \begin{pmatrix} \mathbf{0} & \dots & \mathbf{0} & Q_1^{-1}D^{-1}C_2 \\ \mathbf{0} & \dots & \mathbf{0} & Q_2^{-1}D^{-1}C_2 \\ \vdots & \dots & \vdots & \vdots \\ \mathbf{0} & \dots & \mathbf{0} & Q_m^{-1}D^{-1}C_2 \\ \mathbf{0} & \dots & \mathbf{0} & W^c \end{pmatrix} \tag{2.32}$$

where

$$W^c = \frac{1}{m}(C_2^t D^{-1} C_2)^{-1} C_2^t D^{-1} \sum Q_i^{-1} D^{-1} C_2 =: \sum W_i^c \tag{2.33}$$

The rate of convergence of the CA for (2.21) equals the maximum modulus eigenvalue of W given above. Although in a slight simpler setting, Gelfand et al. (1996) (p. 482) noticed the importance of the W_i matrices and remarked that when their determinant is near zero then the CA is efficient. Actually, the matrices they looked at are the corresponding $p \times p$ matrices that we would have obtained had we written Q^c with the Θ elements in the top-left corner, as was discussed before.

We will now derive the rate of convergence of the NCA for (2.21). The precision matrix of $(\tilde{X}, \Theta)$ is partitioned as

$$Q^{nc} = \begin{pmatrix} Q_1 & \mathbf{0} & \dots & \mathbf{0} & (Q_1 - D^{-1})C_2 \\ \mathbf{0} & Q_2 & \dots & \mathbf{0} & (Q_2 - D^{-1})C_2 \\ \vdots & \vdots & \ddots & \dots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & Q_m & (Q_m - D^{-1})C_2 \\ C_2^t(Q_1 - D^{-1}) & C_2^t(Q_2 - D^{-1}) & \dots & C_2^t(Q_m - D^{-1}) & C_2^t \sum (Q_i - D^{-1})C_2 \end{pmatrix}. \tag{2.34}$$

and therefore, if we define $G = (C_2^t \sum (Q_i - D^{-1})C_2)^{-1}$

$$A^{nc} = \begin{pmatrix} \mathbf{0} & \dots & \mathbf{0} & -Q_1^{-1}(Q_1 - D^{-1})C_2 \\ \mathbf{0} & \dots & \mathbf{0} & -Q_2^{-1}(Q_2 - D^{-1}) \\ \vdots & \dots & \vdots & \vdots \\ \mathbf{0} & \dots & \mathbf{0} & -Q_m^{-1}(Q_m - D^{-1}) \\ -GC_2^t(Q_1 - D^{-1}) & \dots & -GC_2^t(Q_m - D^{-1}) & \mathbf{0} \end{pmatrix} \tag{2.35}$$

$$B^{nc} = \begin{pmatrix} \mathbf{0} & \dots & \mathbf{0} & -Q_1^{-1}(Q_1 - D^{-1})C_2 \\ \mathbf{0} & \dots & \mathbf{0} & -Q_2^{-1}(Q_2 - D^{-1})C_2 \\ \vdots & \dots & \vdots & \vdots \\ \mathbf{0} & \dots & \mathbf{0} & -Q_m^{-1}(Q_m - D^{-1})C_2 \\ \mathbf{0} & \dots & \mathbf{0} & W^{nc} \end{pmatrix} \tag{2.36}$$

where

$$W^{nc} = GC_2^t \sum (Q_i - D^{-1})Q_i^{-1}(Q_i - D^{-1})C_2 =: \sum W_i^{nc}. \tag{2.37}$$

In general, it is not true that $W^c + W^{nc} = I_p$, which would imply that the rate of convergence of the NCA is one minus that of the CA, as was shown for the simple normal model (1.12). Consider for example the following slight modification of (1.12),

$$\begin{aligned} Y_i &= X_i + \sigma_{yi}\epsilon_i \\ X_i &= \Theta + \sigma_x z_i, \ i = 1, \dots, m \end{aligned} \tag{2.38}$$

where we have allowed for heteroscedasticity among the observed data. Then, directly from (2.33) and (2.37) we have that

$$\begin{aligned} \rho_c &= \frac{1}{m}\sum (1 - \kappa_i) \\ \rho_{nc} &= \frac{1}{\sum 1/\sigma_{yi}^2} \sum \kappa_i (1/\sigma_{yi}^2) \\ \kappa_i &= \frac{\sigma_x^2}{\sigma_x^2 + \sigma_{yi}^2}. \end{aligned} \tag{2.39}$$

Suppose we take σ_{yi}^2s to be independent and identically distributed and assume for simplicity

that $\mathsf{E}(1/\sigma_{y1}^2) < \infty$. If we define $\rho_c(\sigma_{y1}^2)$ to be the rate of convergence of the CA for (1.12) when the observation variance is σ_{y1}^2 then as $m \to \infty$

$$\begin{aligned} \rho_c &\to \mathsf{E}[\rho_c(\sigma_{y1}^2)] \\ \rho_{nc} &\to \frac{\mathsf{E}[(1-\rho_c(\sigma_{y1}^2))(1/\sigma_{y1}^2)]}{\mathsf{E}[1/\sigma_{y1}^2]} \end{aligned}$$

and since $1/\sigma_y^2$ and $(1-\rho_c(\sigma_y^2))$ have clearly positive covariance,

$$\rho_{nc} \geq 1 - \rho_c.$$

Some simulation results suggest that the same is true when $\mathsf{E}(1/\sigma_{y1}^2) = \infty$, which seems intuitively reasonable since in that case the observation errors σ_{yi}^2s tend to be very small favouring the CA.

2.5 A State-space model

In Section 2.4 we saw that when allowing for dependence within each X_i (through the covariance matrix D) and for heterogeneity among the Y_is (through the covariates C_{i1} and the different covariance matrices V_i), the rates of convergence for the CA and NCA are qualitatively very different than those for the simple normal model (1.12), a major difference being that they don't sum up to one. Here we will try to gain some understanding on these phenomenon by looking at an example from this family of models with a specific structure on D.

Our example is taken from the class of linear Gaussian state-space models (see for example West and Harrison (1990) for an overview of such models). In their simplest form they are expressed as

$$\begin{aligned} Y_i &= X_i + \sigma_y \epsilon_i \\ X_i &= \phi X_{i-1} + \Theta(1-\phi) + \sigma_x (1-\phi^2)^{1/2} z_i, \ i = 1, \ldots, n \end{aligned} \tag{2.40}$$

σ_x^2, σ_y^2 and ϕ will be considered known and we will choose an improper uniform prior distribution for Θ. Notice that the joint distribution of $(X_1, \ldots, X_n, \Theta)$ is multivariate Gaussian therefore an exact analysis of the convergence rates for the CA and the NCA is feasible. This problem was originally considered by Pitt and Shephard (1999), and their results are reviewed later in this section.

(2.40) falls under the general normal hierarchical model in (2.21), if we take $m = 1$, $n_1 =$

$n,\ V_1 = I_n,\ C_2 = \mathbf{1}$ and

$$D^{-1} = \frac{1}{(1-\phi^2)\sigma_x^2}\begin{pmatrix} 1 & -\phi & 0 & \dots & 0 & 0 \\ -\phi & 1+\phi^2 & -\phi & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \dots & \vdots & 0 \\ 0 & 0 & 0 & \dots & -\phi & 1 \end{pmatrix}.$$

Therefore, using directly the results of Section 2.4, the rate of the CA is given by

$$\begin{aligned} \rho_c = W^c &= \frac{1}{\sigma_x^2(1+\phi)(n-(n-2)\phi)} v^t Q^{-1} v \\ v^t &= [1 \ \ 1-\phi \ \ \dots \ \ 1-\phi \ \ 1] \\ Q &= D^{-1} + 1/\sigma_y^2 I_n. \end{aligned} \tag{2.41}$$

Unfortunately, we can't further simplify (2.41) for ρ_c, since Q^{-1} doesn't have a simple tractable form.

The NCP for this model simply parameterises in terms of $\tilde{X} = (\tilde{X}_1, \dots, \tilde{X}_n),\ \tilde{X}_i = X_i - \Theta$ and Θ, just as for the normal hierarchical model in (2.21). Obviously, we can obtain the rate of convergence of the NCA directly from (2.37) as

$$\rho_{nc} = W^{nc} = \frac{1}{m\sigma_y^2}\mathbf{1}^t Q^{-1}\mathbf{1} \tag{2.42}$$

$$\tag{2.43}$$

This expression, although useful for computing ρ_{nc} for specific parameter values, is not very convenient for comparing it with ρ_c. Working from first principles, as described in Section 2.1.1, we can show that

$$\rho_{nc} = \frac{\mathbf{1}^t Q \mathbf{1} - 2(\sigma_x^2(1+\phi))^{-1}\mathbf{1}^t v + (\sigma_x^2(1+\phi))^{-2} v^t Q^{-1} v}{\mathbf{1}^t Q \mathbf{1} - 2(\sigma_x^2(1+\phi))^{-1}\mathbf{1}^t v + (n-(n-2)\phi)/(\sigma_x^2(1+\phi))} \tag{2.44}$$

although derivation of this expression directly from (2.43) is less straightforward. Careful calculations show that

$$\mathbf{1}^t Q \mathbf{1} - 2(\sigma_x^2(1+\phi))^{-1}\mathbf{1}^t v + (n-(n-2)\phi)/(\sigma_x^2(1+\phi)) = \frac{n}{\sigma_y^2}.$$

We now notice that

$$\begin{aligned} 1-\rho_{nc} &= \frac{(n-(n-2)\phi)/(\sigma_x^2(1+\phi)) - (\sigma_x^2(1+\phi))^{-2} v^t Q^{-1} v}{\mathbf{1}^t Q \mathbf{1} - 2(\sigma_x^2(1+\phi))^{-1}\mathbf{1}^t v + (n-(n-2)\phi)/(\sigma_x^2(1+\phi))} \\ &= (1-\rho_c)\frac{1-\kappa}{\kappa}\frac{1}{1-\phi^2}(1-2(n-1)/n\phi + (n-2)/n\phi^2). \end{aligned} \tag{2.45}$$

In the second equality we made use of the definition of κ in (2.16). By letting $n \to \infty$ we can obtain the following result about the asymptotic relative performance of the CA and NCA for the Gaussian state-space model:

$$\frac{1-\rho_{nc}}{1-\rho_c} = \frac{1-\kappa}{\kappa}\frac{1-\phi}{1+\phi}. \tag{2.46}$$

The definition of the integrated autocorrelation time τ_f for a scalar function f of an ergodic Markov chain was given in (2.12). Let τ be the integrated autocorrelation time for the identity function of the stationary Markov chain $\{X_i, i = 1, 2, \ldots\}$ defined in (2.40). Then

$$\frac{1-\rho_{nc}}{1-\rho_c} = \frac{1-\kappa}{\kappa}\frac{1}{\tau}. \tag{2.47}$$

This is a particularly interesting and intuitive expression, which also explains the behaviour of the CA and the NCA when applied in situations where the missing data are stationary stochastic processes (see for example Chapter 6 and particularly Section 6.8). The first term in (2.47) corresponds to $(1-\rho_{nc})/(1-\rho_c)$ when $\phi = 0$, i.e when the state-space model collapses to the normal hierarchical model (2.14). The higher the dependence among the X_is the more preferable the CA becomes over the NCA. That is, the CA is more likely to be preferred for highly autocorrelated hidden processes.

On the one hand, when estimating X_i we can use information not only from the corresponding observed data point Y_i, but also from all the "neighbouring" data, since they all have underlying X value very close to X_i, due to their prior high dependence. Hence, when X is highly autocorrelated it is like having multiple observations $Y_{i1}, Y_{i2}, \ldots$ for every missing data point X_i, which makes the observation error smaller. At the same time, as τ increases X becomes less informative about the stationary mean Θ. Therefore, as the dependence in the X process increases, both the data become more informative about X and the link between parameters and missing data gets weaker, therefore it is not surprising that the CA is increasingly more efficient.

It is important to realise that (2.47) is largely a qualitative expression regarding ρ_c and ρ_{nc}. Namely, it shows how CA becomes preferable over NCA for a fixed κ as we increase the persistence in the hidden process. However, the value of the ratio $(1-\rho_{nc})/(1-\rho_c)$ is not really interpretable (see also the relevant remark at the end of Section 2.3.2). As we

discussed in Section 2.1.2 a more appropriate measure is $\log(\rho_c)/\log(\rho_{nc})$, although this is approximately equal to (2.47) when both rates are close to unity. Another quantity that can be used to this end is $(1+\rho_{nc})(1-\rho_c)/[(1-\rho_{nc})(1+\rho_c)]$. This approach was adopted by Pitt and Shephard (1999) as we will shortly describe.

Figure 2.2 demonstrates the performance of the CA and the NCA for various sample sizes n, values of ϕ and κ.

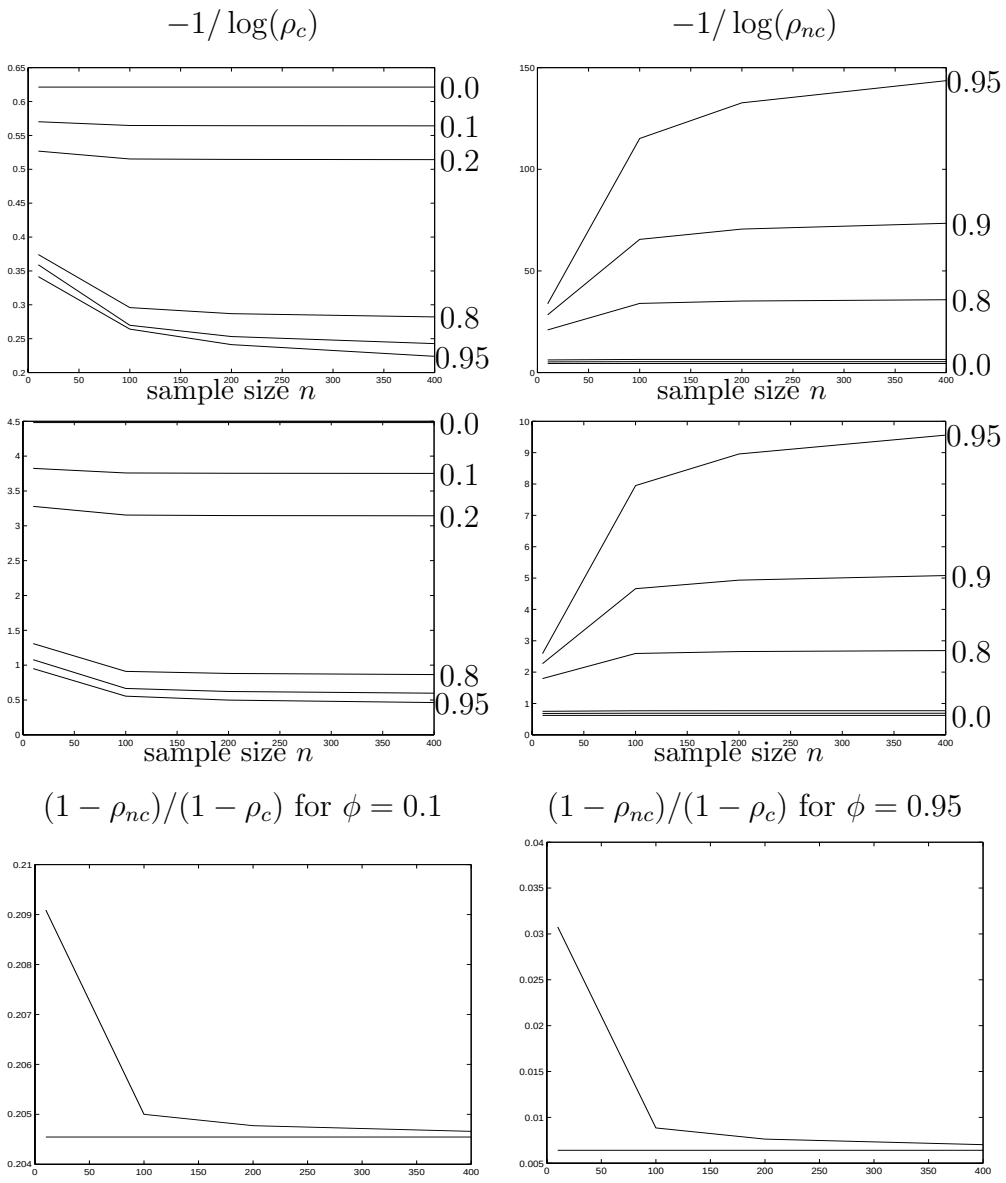


Figure 2.2: Convergence rates results for the state-space model. The first two rows plot $-1/\log(\rho_c)$ and $-1/\log(\rho_{nc})$ against the sample size n for various values of ϕ, for $\kappa = 0.8$ and $\kappa = 0.2$ respectively. The last row shows $(1-\rho_{nc})/(1-\rho_c)$ against n together with its asymptotic limit (the horizontal line), for $\phi = 0.1$ (left) and $\phi = 0.95$ (right).

Pitt and Shephard (1999) were the first to derive analytic convergence rates for the Gibbs

sampler applied to the Gaussian state-space model (2.40), although they used a slightly different formulation where the stationary variance of the hidden process is $\sigma_x^2/(1-\phi^2)$. Using the methodology developed by Roberts and Sahu (1997) for calculating the rate of convergence of the Gibbs sampler on Gaussian target distributions, they considered two issues: when Θ is considered fixed, how efficient is a single site updating Gibbs sampler for simulating from the conditional posterior distribution of $X = (X_1, X_2, \ldots, X_n)$, and when Θ is unknown whether the CP or the NCP should be adopted for simulating from the joint posterior distribution of (Θ, X).

The first problem is outside the scope of this thesis and therefore we will not discuss it in any detail. The main result is that when the hidden process is very persistent (i.e when ϕ is close to 1) and $\kappa \to 0$, the single-site updating scheme is very inefficient and the rate of convergence, for large n, tends to 1. Therefore, either forward-backward updating schemes, that update X as a block, or other blocking schemes that update large chunks of the underlying process should be preferred.

The second problem has been closely examined in this section. Pitt and Shephard (1999) consider the two-component Gibbs sampler that updates X as a block (using for example forward-backward techniques) and study when the CP or the NCP should be preferred. Here, we derived all the results treating (2.40) as a special case of the general normal hierarchical model (2.21) and using the results we obtained for that model in the previous section. Pitt and Shephard (1999) worked from first principles and derived similar expressions. Their equation (4) on page 70, although it appears different, is exactly the same as (2.43). To compare the asymptotic relative efficiency of the NCA over the CA they used $(1+\rho_{nc})(1-\rho_c)/[(1-\rho_{nc})(1+\rho_c)]$, and also provided tight upper and lower bounds for this expression, since $(1+\rho_{nc})/(1+\rho_c)$ is analytically intractable. They found the analytic results of the Gaussian state-space model to be valuable for deciding on the parameterisation to be used for more complex and intractable models. In particular, they considered parameterisation issues for a log-Gaussian discrete time stochastic volatility model, which can be expressed as linear non-Gaussian state-space model.

2.6 Linear non-Gaussian model

The following toy example is very simple, but its results are quite striking. Suppose we modify the simple normal hierarchical model (1.12) such that ϵ_{ij} has a standard Cauchy distribution, while z_i remains standard Gaussian. That is, the model now writes

$$\begin{aligned} Y_i &= X_i + \epsilon_i,\ \epsilon_i \sim \mathrm{Ca}(0,1) \\ X_i &= \Theta + \sigma_x z_i,\ , z_i \sim \mathrm{N}(0,1),\ i = 1, \ldots, m \end{aligned} \tag{2.48}$$

where the latent variance is assumed to be known $\sigma_x^2 = 1$, although inference for that parameter will be considered in Section 4.3.

In this context, the heavy tailed nature of the Cauchy distribution makes the observation equation relatively uninformative for extremal values of X. Following the intuition gained from studying the normal hierarchical model, we might expect the CA to perform poorly in some way in the tail regions in relation to the NCA. Figure 2.3 shows output from the Gibbs sampler for both the CA and the NCA (where $m = 1$) for different starting values for Θ. The CA exhibits unstable heavy-tailed excursions characteristic of algorithms which fail to be geometrically ergodic (see Roberts (2003) and Section 3.1.5 of this thesis), while the NCA appears to return to the distribution mode very rapidly from all starting values. In fact, Chapter 3 proves that the CA fails to be geometrically ergodic whereas the NCA is

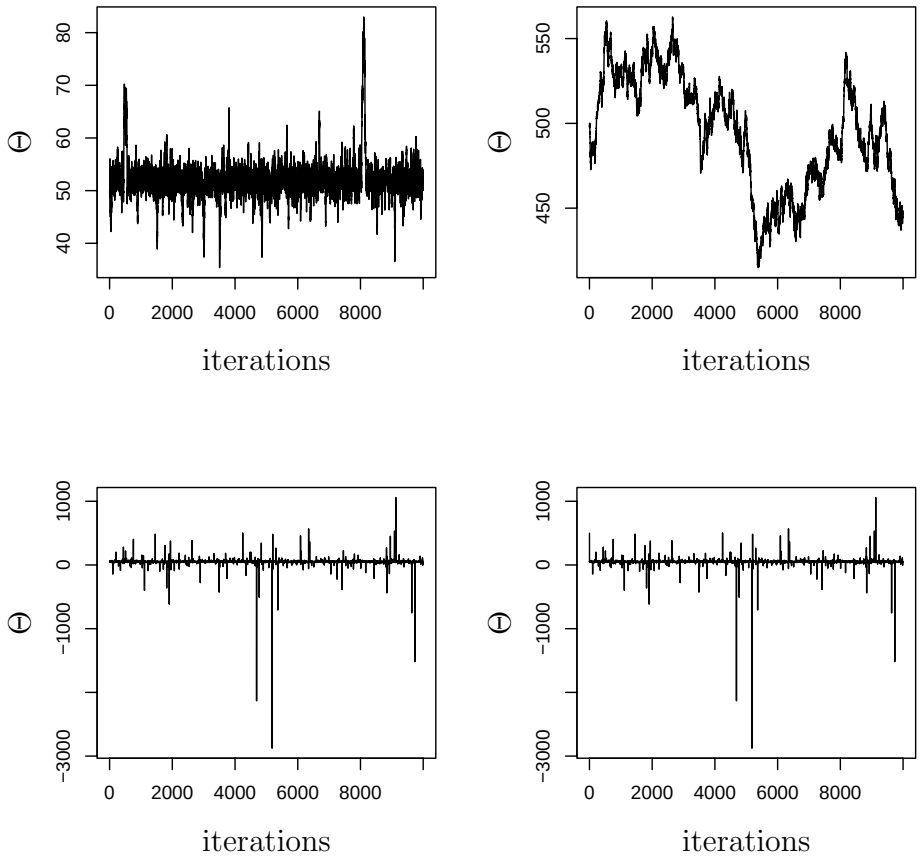


Figure 2.3: Gibbs sampler output for Θ in the Normal-Cauchy model (2.48), where we have taken $m = 1$, $Y_1 = 51.91$, $\sigma_x^2 = 1$.. Top: centered parameterisation started from $\Theta_0 = 50$ (left) and $\Theta_0 = 500$ (right). Bottom: non-centered parameterisation for the same starting values. All chains were run for 10^4 iterations. Notice the different scales in the plots.

uniformly ergodic. We refer to that chapter for definitions, the proofs and considerably more general results about convergence of the Gibbs sampler for linear hierarchical models with

heavy tailed links. For example, we show that for Cauchy latent equation with Gaussian observation (as used for example in Wakefield et al. (1994)), the opposite result holds with the NCA failing to be geometrically ergodic while the CA is uniformly ergodic.

Chapter 3

Convergence of MCMC for linear hierarchical models with heavy-tailed links

3.0 Introduction

This chapter studies the convergence of the two-component Gibbs sampler for linear non-Gaussian models. We show that the CA can have markedly different behaviour from the NCA when the tails of either the observation and/or the latent equation are non-Gaussian. In particular, we establish conditions under which the CA converges uniformly quickly and the NCA fails to be geometrically ergodic. This is for example the case when the prior of the missing data is heavy tailed whereas the observation error has exponential or lighter tails. This result justifies the tremendous success of the early implementation of the Gibbs sampler in simple hierarchical models, see for example Wakefield et al. (1994). The proof of the negative result for the NCA is based on the notion of the capacitance of a Markov chain. Dual conditions imply that the CA is not geometrically ergodic and the NCA is uniformly fast, when for example the prior has light and the observation error heavy tails.

The conditions we use have originally been developed in the Bayesian robustness literature and we make this connection. We also look at a model for which these conditions are not satisfied, where the double exponential distribution is used in both the latent and observation equation. We show geometric ergodicity for both the CA and the NCA by proving the existence of appropriate drift conditions.

The material of this chapter is based on Roberts and Papaspiliopoulos (2003).

In the sequel, we will use Π_X and π_X to denote the probability law and the density with respect to the Lebesgue measure (when it exists) respectively of a random variable X, and $\Pi_{X|Y}$, $\pi_{X|Y}$ for the conditional law and density respectively of X given Y. Capital letters are

used for random variables while lower case for their values. "$\Rightarrow$" denotes weak convergence of probability measures.

3.1 Markov chain theory for general state spaces

Section 1.5.1 introduced the very basic concepts in the theory of Markov chains for general state-spaces, while Section 2.1 went a bit further presenting some results about the spectral analysis of the transition operator of the Gibbs sampling Markov chain. This section goes deeper into Markov chain theory, in order to develop the necessary machinery to prove the main results of this chapter regarding uniform and geometric ergodicity of the Gibbs sampler under the NCP for various linear non-Gaussian hierarchical models. The material below is based heavily on Roberts and Tweedie (2004) and Meyn and Tweedie (1993).

Notice that although most of the definitions and theorems presented in the following subsections hold for quite arbitrary state spaces (with very weak assumptions on their structure), our focus in this chapter is on Euclidean spaces.

3.1.1 ϕ-irreducibility and small sets

Since the Markov chain $\{Z_n\}$ can be thought of as a random variable taking values in the sample-path space Ω (see Section 1.5.1), it is natural to define for any set $A \in \mathcal{B}(\mathcal{Z})$

$$\tau_A := \min\{n \geq 1 : Z_n \in A\} \tag{3.1}$$

$$\eta_A := \sum_{n=1}^{\infty} \mathbb{1}[Z_n \in A] \tag{3.2}$$

the first return and occupation time on A respectively.

We define $L(z, A)$ for $z \in \mathcal{Z}$ and $A \in \mathcal{B}(\mathcal{Z})$ to be the probability that the Markov chain started from $z \in \mathcal{Z}$ ever enters A, that is

$$L(z, A) := \mathsf{P}_z[\tau_A < \infty] = \mathsf{P}_z[Z_n \in A, \text{ for some } n \geq 0].$$

Definition 3.1.1. *The Markov chain $\{Z_n\}$ is called ϕ-irreducible if there exists a non-trivial probability measure ϕ on $\mathcal{B}(\mathcal{Z})$ such that, whenever $\phi(A) > 0$, we have $L(z, A) > 0$ for all $z \in \mathcal{Z}$.*

It is remarkable, that as long as $\{Z_n\}$ has an invariant probability measure π, which will be the case in MCMC by construction, then ϕ-irreducibility for some measure ϕ ensures π-irreducibility and uniqueness of π (see Proposition 4.4.1 of Roberts and Tweedie (2004)).

Definition 3.1.2. *A set $C \in \mathcal{B}(\mathcal{Z})$ is called a small set if there exists an $m > 0$, a non-trivial probability measure ϕ on $\mathcal{B}(\mathcal{Z})$, and an $\epsilon > 0$, such that for all $A \in \mathcal{B}(\mathcal{Z})$ and $z \in C$,*

we have the minorisation condition,

$$P^m(z, A) \geq \epsilon\phi(A). \tag{3.3}$$

Then we say that C is (m, ϵ, ϕ)-small, or simply, m-small.

Trivially, any singleton $\{z\} \subset \mathcal{Z}$ is a 1-small set, since we can choose the minorising measure to be the transition kernel $P(z, \cdot)$. (3.3) expresses that C essentially behaves like a singleton, since with probability at least ϵ, the chain m-steps after it has left C it will have forgotten which point in C it started from. Small sets play a prominent role in the analysis of Markov chains on general state spaces. For example, the technique of coupling, which is used to prove the ergodic theorem (see Section 3.1.3) for irreducible and aperiodic Markov chains, is based on the existence of such small sets. We will see in Section 3.1.4 and Section 3.1.5 that small sets are also related to the concepts of uniform and geometric ergodicity. Therefore, it is important to recognise small sets in the state-space. It turns out that in many cases compact sets are small, see for example Sections 5.1.1 and 5.2 of Roberts and Tweedie (2004) for details. This will be the case in the applications involved in this chapter. We conclude this section with defining another important concept, that of strong aperiodicity:

Definition 3.1.3. *When there exists a small set C satisfying the minorisation condition (3.3) for $m = 1$ (that is, C is 1-small), then the chain is called strongly aperiodic.*

As the name suggests, this is a strong form of aperiodicity. To avoid unnecessary detail we refer to Section 5.3.3 of Roberts and Tweedie (2004) for the definition of aperiodicity for general state spaces. Chapter 5 of Roberts and Tweedie (2004) shows that the MCMC algorithms under mild conditions are aperiodic. This will be the case in all the examples of this chapter.

3.1.2 Recurrence and Harris chains

π-irreducibility for a Markov chain ensures that all "big" (according to π) sets have a chance of being visited. Recurrence relates to whether these sets will be visited in a finite time *almost surely*. This topic is covered in detail in Chapter 6 of Roberts and Tweedie (2004). A set A is called recurrent if $\mathsf{E}_z(\eta_A) = \infty$ for all $z \in A$ and a π-irreducible Markov chain $\{Z_n\}$ is called recurrent if every $A \in \mathcal{B}(\mathcal{Z})$ with $\pi(A) > 0$ is recurrent. Notice that the definition doesn't assert that $\eta_A = \infty$ *almost surely*, which can be shown (see Proposition 6.2.1 of Roberts and Tweedie (2004)) to be equivalent to the demand that

$$L(z, A) = 1, \text{ for all } z \in A. \tag{3.4}$$

A set A for which (3.4) is true for every $z \in A$ is called Harris recurrent and Harris recurrent chains are defined analogously. Certainly, (3.4) is more profoundly expressing the intuitive notion of recurrence than the requirement $\mathsf{E}_z(\eta_A) = \infty$, in the sense that the chain repeatedly visits all "big" sets. Nevertheless, for every recurrent set A there are two options: if A is ever reached by the chain, then it will be revisited infinite number of times. However, for some starting points, there is the chance that the chain never visits A. This is why $\mathsf{E}_z(\eta_A) = \infty$ but not necessarily $\eta_A = \infty$ *almost surely*. Harris chains ensure that such bad starting points do not exist. This dichotomy is manifested in the decomposition of $\mathcal{Z} = H \cup N$ for a recurrent chain, where H is absorbing (that is, if the chain starts in H never leaves it) and non-empty, and every set $A \subset H$ in $\mathcal{B}(\mathcal{Z})$, with $\pi(A) > 0$ is Harris, and N is π-null and transient. Most MCMC chains are Harris recurrent, see Chapter 6 of Roberts and Tweedie (2004).

3.1.3 The ergodic theorem

The total variation norm for a signed measure ν on $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$ is defined as

$$\|\nu\| := \sup_{|f| \leq 1} \left| \int f(y)\nu(dy) \right| = 2 \sup_{A} |\nu(A)|.$$

The ergodic theorem, which is Theorem 7.1.1 of Roberts and Tweedie (2004), for an aperiodic and ϕ-irreducible Markov chain $\{Z_n\}$, is stated as a collection of equivalent conditions which ensure that there exists a unique probability measure π and a π-null set N, such that for every initial condition $z \in \mathcal{Z} - N$,

$$\|P^n(z, \cdot) - \pi(\cdot)\| \to 0, \text{ as } n \to \infty.$$

When $\{Z_n\}$ is Harris, $N = \emptyset$, which shows that the Harris property is important to guarantee convergence from all possible starting points of the sample space. A sufficient condition for the ergodic theorem to hold, for an aperiodic and irreducible Markov chain, is that there exists an invariant probability measure π for the chain, which is always true by construction for MCMC algorithms. Convergence in total variation distance ensures convergence of the type

$$\lim_{n \to \infty} \mathsf{E}_z[f(Z_n)] = \pi f$$

for bounded functions, although typically we are more interested in convergence results for unbounded functions. The f-norm of a signed measure ν is defined as

$$\|\nu\|_f = \sup_{g : |g| \leq f} \left| \int g(y)\nu(dy) \right|$$

where $f \geq 1$. Surprisingly, if $\{Z_n\}$ is ϕ-irreducible and aperiodic, and π is invariant probability measure, the extra assumption that $f \in \mathcal{L}^1$ is enough to generalise the ergodic theorem:

$$\|P^n(z,\cdot) - \pi(\cdot)\|_f \to 0, \text{ as } n \to \infty$$

and, thus, incorporate convergence of moments of unbounded functions g, which increase to ∞ not faster than f.

3.1.4 Uniform ergodicity of Markov chains

Definition 3.1.4. *A Markov chain $\{Z_n\}$ is called uniformly ergodic if there exists an invariant measure π such that*

$$\sup_{z \in \mathcal{Z}} \|P^n(z,\cdot) - \pi\| \to 0, \text{ as } n \to \infty. \tag{3.5}$$

Therefore, uniform ergodicity is a very strong form of ergodicity, since it ensures that there are no starting values which can lead to arbitrarily slow convergence of the chain, minimising somehow the "burn-in" problem. There are many properties which are equivalent to uniform ergodicity and can be used to show that a chain converges or not, uniformly quickly. In the following theorem we state the most relevant to our purposes. A more complete list is given in Theorem 9.1.1 of Roberts and Tweedie (2004).

Theorem 3.1.1. *For any Markov chain $\{Z_n\}$ the following are equivalent:*

1 *$\{Z_n\}$ is uniformly ergodic.*

2 *There exists $r > 1$ and $R < \infty$ such that for all $z \in \mathcal{Z}$*

$$\|P^n(z,\cdot) - \pi\| \leq Rr^{-n},$$

which implies that the convergence takes place at a uniform geometric rate.

3 *The Doeblin condition: the chain is aperiodic and there exists a probability measure ϕ on $\mathcal{B}(\mathcal{Z})$, an $\eta < 1, \delta > 0$ and an integer m, such that, whenever $\phi(A) > \eta$,*

$$\inf_{z \in \mathcal{Z}} P^m(z, A) > \delta. \tag{3.6}$$

4 *The state-space $\mathcal{Z}$ is m-small, for some m.*

Usually, one of the last two conditions are used to establish uniform ergodicity for a given Markov chain.

3.1.5 Geometric ergodicity of Markov chains

Definition 3.1.5. *A Harris Markov chain $\{Z_n\}$ with an invariant probability measure π, is called geometrically ergodic, sometimes also called V-uniformly ergodic, if there exists a function $V \geq 1$ with $\pi V < \infty$, such that*

$$\|P^n(z,\cdot) - \pi\|_V \leq R_V V(z) r_V^{-n} \tag{3.7}$$

for some constants $r_V > 1, R_V < \infty$.

Geometric ergodicity is usually expressed as a collection of equivalent conditions, see for example Theorem 10.1.1 of Roberts and Tweedie (2004). For our purposes the following suffices.

Theorem 3.1.2. *Suppose that the Markov chain $\{Z_n\}$ is ϕ-irreducible and aperiodic. Moreover, suppose that there exists a small set C, constants $b < \infty$, $\lambda < 1$ and a function $V \geq 1$ finite at some $z_0 \in \mathcal{Z}$, satisfying the (Foster-Lyapunov) drift condition*

$$PV(z) \leq \lambda V(z) + b\mathbb{1}[z \in C], \text{ for all } z \in \mathcal{Z}. \tag{3.8}$$

Then the set $S_V := \{z \in \mathcal{Z} : V(z) < \infty\}$ is absorbing, $\pi(S_V) = 1$ and $\{Z_n\}$ is geometrically ergodic.

(3.8) is known as a drift condition and in many examples, especially those occurring in MCMC, it provides a mechanism for showing that a particular chain is geometrically ergodic by finding an appropriate function V. Notice that if (3.8) holds,

$$P^n V(z) \leq \lambda^n V(z) + b\frac{1-\lambda^{n+1}}{1-\lambda}$$

which shows that $\pi V < \infty$.

We will only consider drift conditions in $\mathbb{R}$, since all our applications can be analysed at this level of complexity. Drift conditions can be considered for higher dimensions as well, although they become less intuitive. In $\mathbb{R}$ compact sets are typically small for MCMC chains (see Chapter 5 of Roberts and Tweedie (2004)). This will be the case in our applications and we will assume throughout this chapter that compact sets are small.

Figure 3.1 informally illustrates an example of a drift function V and a small set C for a unimodal density π on $\mathbb{R}$. C is typically taken to be a compact set around the mode of π - an area where the chain keeps returning to. It will typically be necessary that $V(z)$ goes to infinity as $|z| \to \infty$. To see why notice that outside C

$$\frac{PV(z)}{V(z)} \leq \lambda. \tag{3.9}$$

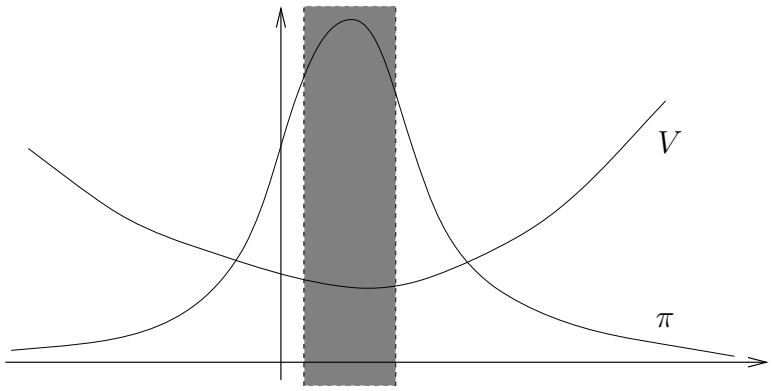


Figure 3.1: A typical example of a drift function V for a unimodal density π on $\mathbb{R}$. The shaded area consists of pairs (z, y) such that $z \in C$, where C is the small set used in the drift condition. C is typically some compact set around the mode of π.

The speed at which it needs to increase as $|z| \to \infty$ depends on how quickly the chain drifts back to C. If the chain returns very quickly to C, PV will be small compared to V even when the latter increases slowly. Notice that πV has to be finite, therefore the tails of π put further restrictions on how quickly V can afford to increase as $|z| \to \infty$.

It is important to have an idea how to derive a drift condition for a given Markov chain with transition kernel P. We sketch two techniques when $\mathcal{Z} \subseteq \mathbb{R}$, which will become valuable in Section 3.5. We assume throughout that V is increasing as $z \to \infty$, decreasing as $z \to -\infty$ and that compact sets are small. Suppose that we can show that

$$\limsup_{|z| \to \infty} \frac{PV(z)}{V(z)} < 1,$$

then by definition there exists some $c > 0$ and a $0 < \lambda < 1$ such that

$$\frac{PV(z)}{V(z)} \leq \lambda \text{ for all } |z| > c.$$

Take the small set to be $C = [-c, c]$. The existence of a drift condition can easily be established if V is continuous and P is weak Feller, which means that $PV(z)$ is continuous in z whenever V is (see Section 5.1.1 of Roberts and Tweedie (2004)). Most MCMC chains are weak Feller, for example both the Metropolis and the Gibbs algorithm under mild smoothness assumptions on the target density π. Under these assumptions, $PV(z)/V(z)$ is continuous and when C is compact there exists some $0 < b < \infty$ such that $PV(z)/V(z) < b$ for all $z \in C$ and (3.8) follows.

Sometimes it is possible to show that

$$PV(z) \leq \lambda' V(z) + b', \text{ for all } z \in \mathcal{Z}. \tag{3.10}$$

This is not in the form required in (3.8), nevertheless once (3.10) has been shown it is not hard to establish the drift condition. From (3.10) follows that

$$PV(z)/V(z) \leq \lambda' + b'/V(z).$$

Take $C = [-c, c]$ for some $c > 0$. Then, assume that $V(z) > V(c)$ for all $|z| > c$, which implies that outside C

$$PV(z)/V(z) \leq \lambda' + b'/V(c) := \lambda.$$

Therefore, we try to find a $c > 0$ such that $V(c)$ is the lower bound of V outside C and $\lambda < 1$ in the inequality above. Typically $\lambda' < \lambda$, since we need to "sacrifice" some of the speed at which we drift towards C in order to introduce the innovation b once in C. Inside C,

$$PV(z) - \lambda V(z) \leq \lambda' V(z) + b' - \lambda V(z) = (\lambda' - \lambda)V(x) + b'.$$

If $\lambda' < \lambda$ then (3.8) follows with $b = b'$. Otherwise, we try to bound V on C by some B and take $b = b' + B$.

Geometric ergodicity can also be studied in the $\mathcal{L}^2$ framework. This allows the introduction of powerful geometrical results about geometric convergence involving the notion of capacitance of a Markov chain.

Definition 3.1.6. *For a given Markov chain $\{Z_n\}$ with an invariant probability measure π and transition kernel P, the conductance of a set $A \in \mathcal{B}(\mathcal{Z})$ with $\pi(A) > 0$ is defined as*

$$c(A) = \frac{\int_A P(x, A^c)\pi(dx)}{\pi(A)} \tag{3.11}$$

and the chain capacitance as

$$\kappa = \inf_{0<\pi(A)\leq 1/2} c(A), \tag{3.12}$$

where the infimum is taken over all $A \in \mathcal{B}(\mathcal{Z})$. The conductance has the interpretation as the probability of leaving A having started according to π restricted to A. The κ defined in (3.12) should not be confused with the signal-to-noise ratio defined in (2.16). To keep consistency with the literature, we use the same notation for both, without creating confusion, since the two quantities never appear in the same context.

Let $g = 1 - \|P\|$, then the following inequality holds only for reversible Markov chains and is known as the Cheeger inequality

$$\kappa^2/2 \leq g \leq 2\kappa; \tag{3.13}$$

see Lawler and Sokal (1988) for example. Notice that for reversible chains, $\text{spec}(P) = \|P\|$ and g is the spectral gap (see Section 2.1), thus (3.13) can be used to derive bounds on the $\mathcal{L}^2$ geometric convergence rate of P. It is typically difficult to compute the capacitance of a chain, due to the infimum operation involved in (3.12). On the other hand, this notion can effectively be exploited to prove negative results, as is done in Theorem 3.3.3 for example. In particular, in many examples we aim at showing that $\kappa = 0$ by finding sets with arbitrarily small capacitance. This then demonstrates that the Markov chain fails to be geometrically ergodic, by Cheeger's inequality (3.13).

In Section 3.3 we are interested in using this result to prove that the two-component Gibbs sampler is not geometrically ergodic for some linear non-Gaussian models. However, the Gibbs sampler is not reversible so it is not feasible to apply the result directly. Instead a two-stage procedure is needed. Firstly, recall from Section 1.5.2 that the two marginal chains of the sampler are reversible. Secondly, we can benefit from the result about de-initialising chains in Section 3.1.6 to show that the convergence rate of the bivariate chain coincides with the rate at which the marginal chains converge. Thus, we can use the Cheeger inequality to bound the rate of convergence of the two-component Gibbs sampler.

3.1.6 De-initialising chains

The notion of de-initialising chains, introduced by Roberts and Rosenthal (2001), proves very convenient in the convergence rate analysis of the two-component Gibbs sampler. Generally, let $\{Z_n\}$ be a Markov chain on a state space $\mathcal{Z}$ and $\{U_n\}$ be another chain (not necessarily Markovian) on $\mathcal{U}$. We call $\{U_n\}$ de-initialising for $\{Z_n\}$ if for each $n > 0$, Z_n is independent of Z_0 given U_n. This situation arises naturally in the two-component Gibbs sampler (see Section 1.5.2). Let π be the target distribution of the sampler and $Z = (Z^{(1)}, Z^{(2)}) \sim \pi$. Then, as shown in the graphical model of the Gibbs sampler in Figure 1.1, the marginal Markov chain $\{Z_n^{(2)}\}$ is de-initialising for $\{Z_n^{(1)}\}$, but also for the joint chain $\{Z_n\}$. Trivially, $\{Z_n\}$ is de-initialising for $\{Z_n^{(2)}\}$. When two chains are de-initialising for each other they are termed co-de-initialising. Then Corollary 1 and Theorem 1 of Roberts and Rosenthal (2001) can be used to show that the convergence rate of $\{Z_n\}$ is the same as that of $\{Z_n^{(2)}\}$. Since the rate of convergence of the two-component Gibbs sampler is invariant to the order of the updated components (see Section 1.5.2, but also Section 2.1), the convergence rate of $\{Z_n^{(1)}\}$ coincides with that of $\{Z_n^{(2)}\}$ and $\{Z_n\}$.

3.2 Linear non-Gaussian models and robust Bayesian analysis

This chapter deals with hierarchical models which have the same linear structure as the normal hierarchical model (1.12),

$$\begin{aligned} Y_{ij} &= X_i + \epsilon_{ij},\ j = 1, \ldots, n \\ X_i &= \Theta + z_i,\ i = 1, \ldots, m \end{aligned} \tag{3.14}$$

but where the distribution of the error terms $\{\epsilon_{ij}, z_i\}$ in either or both of the two equations are not Gaussian, but they are assumed to be symmetric. All random variables above take values in $\mathbb{R}$. For simplicity in the sequel we take $n = 1$. Unlike the Gaussian model, when ϵ_{ij}s are not normally distributed it is not certain that one-dimensional sufficient statistics exist, thus the assumption $n = 1$ can affect the generality of the results. Nevertheless, we believe that our results still hold but the case of arbitrary n will be investigated elsewhere.

When the error distributions are heavy tailed (see Section 3.2.1 for definitions), (3.14) can handle outlying observations in a more satisfactory way than the normal model (1.12); see for example Pericchi and Smith (1992), O'Hagan (1979), Dawid (1973) and references therein. The Bayes estimate (under square loss function) of X_i given Y and Θ in the normal model is shown in (2.17) to be

$$\mathsf{E}[X_i \mid Y, \Theta] = \kappa Y_i + (1 - \kappa)\Theta$$

where $\kappa \in [0, 1]$ is defined in (2.16). It can be shown that whatever the prior we assign to Θ the Bayes estimate of X_i will tend to infinity when $Y_i \to \infty$. On the other hand, it is often desirable that such an outlying observation is ignored. This can be achieved if, for example ϵ_i has a heavy tailed distribution, the Cauchy for instance; Section 3.2.1 gives conditions under which such outlier-proneness can be guarranted.

Similarly, in the three-stage hierarchical model (3.14), although the data Y_{ij} might be believed to be normally distributed, there are situations where the analysis needs to be protected from the effects of a small number of outlying X_is. This situation arises often in hierarchical modelling, especially in random effect models where X_i is the random effect of the ith individual in the population. (It is typical in medical studies, e.g. in pharmacokinetics (see Wakefield et al. (1994)) that there will be a small number of individuals who behave very differently from the rest.) One way of protecting against aberrant individuals is to use a heavy tailed second-stage distribution (see for example Wakefield et al. (1994) and the discussion of Choy and Smith in Lee and Nelder (1996)). This downweighs the influence of the outliers so that the corresponding parameters are shrunk less and their contribution to

the overall population mean is reduced. We also note that there are extensions of the simple model (3.14) in the time series modelling, where the X_is are serially correlated (i indexes time in that context) and z_i is modeled by a heavy tailed distribution (see for example Harvey et al. (1994)).

The purpose of this chapter is to study the convergence of the Gibbs sampler on the joint posterior distribution of $X = (X_1, \ldots, X_m)$ and Θ in (3.14), under the two different parameterisations: the centered (X, Θ), and the non-centered $(\tilde{X}, \Theta)$, where $\tilde{X} = X - \mathbf{1}\Theta$ is *a priori* independent of Θ; see Section 2.2 and Section 4.1 for a description of non-centering for hierarchical models.

For simplicity we will further assume that $m = 1$ (on top of the assumption that $n = 1$) throughout this chapter. Actually, this is not very crucial, since the independence of among the X_is conditionally on Θ still allows us a good deal of analytic tractability. Nevertheless, this extension is beyond the scope of this chapter and will be reported elsewhere. Thus, we rewrite (3.14) in a notationally more convenient form (which is consistent with Roberts and Papaspiliopoulos (2003))

$$\begin{aligned} Y &= X + C \\ X &= \Theta + \tilde{X}. \end{aligned} \tag{3.15}$$

The simple linear structure of (3.15) allows us to derive analytic results concerning the convergence rate of the Gibbs sampler. The results are not quantitative as those for the Gaussian models of Chapter 2 but qualitative. Nevertheless, they are striking since they reveal that the CA and the NCA can have markedly different performance when the tails of the latent and observation equations are not normal-like. The link with the Bayesian robustness literature is fruitful, since Section 3.3 and Section 3.4 show that the conditions which have been developed in that context to ensure dominance of either the prior or the likelihood in the presence of outliers, turn out to characterise the speed at which the CA and the NCA converge to stationarity. These conditions are described in the following section.

3.2.1 The Dawid/O'Hagan conditions

Dawid (1973) assumes a single observation $Y = y$ from the linear model $Y = X + C$ and investigates the asymptotic behaviour of the posterior distribution of X as $y \to \infty$. In particular, he establishes conditions on Π_C and Π_X under which $\Pi_{X|Y=y} \Rightarrow \Pi_X$ as $y \to \infty$, that is the posterior distribution of X converges to its prior as the observation becomes large. He also remarked that due to the symmetry of X and C in the model, if their distributions are interchanged the same conditions ensure that $\Pi_{y-X|Y=y} \Rightarrow \Pi_C$ as $y \to \infty$, therefore the prior is ignored. When Π_C and Π_X are both Gaussian neither of these situations can happen, since the posterior mean of X in (2.17) is always a compromise between the prior and the

data.

Dawid (1973) asks that Π_C and Π_X have densities with respect to the Lebesgue measure π_C, π_X respectively, and expresses his conditions in terms of those, although as we shall see some of them are more naturally understandable using probability measures. These conditions correspond to the case where $y \to \infty$, however they can be obviously modified to cater for the case where $y \to -\infty$. We will also provide a proof of the main result for reasons of completeness, but also in order to motivate the choice of these particular conditions. Actually, our proof shows that the second condition is redundant, since conditions 1 and 3 are enough to prove the result.

The Dawid's conditions for $\Pi_{y-X|Y=y} \Rightarrow \Pi_C$ as $y \to \infty$

- D1. Given $\epsilon > 0$ and $h > 0$, there exists A such that when $x > A$, then

$$|\pi_X(x') - \pi_X(x)| < \epsilon \pi_X(x) \text{ whenever } |x' - x| < h. \tag{3.16}$$

- D2. For some constants B and M, $0 < \pi_X(x') < M\pi_X(x)$ whenever $x' > x > B$.
- D3. Defining

$$k(x) = \sup_z \{\pi_X(x - z)/\pi_X(z)\} \tag{3.17}$$

then

$$\int_{-\infty}^{\infty} k(x)\pi_C(x)dx < \infty.$$

There are various ways to define a heavy tailed distribution with support on the real line, see for example Section 1.4 of Embrechts et al. (1997) for a collection of different definitions. The following coincides with the class $\mathcal{L}$ of heavy tailed distributions defined in Section 1.4 of Embrechts et al. (1997).

Definition 3.2.1. *A random variable X in $\mathbb{R}$ is said to have a right heavy tailed distribution if for all $h > 0$,*

$$\frac{\Pi_X[A < X < A + h]}{\Pi_X[X > A]} \to 0, \text{ as } A \to \infty \tag{3.18}$$

which implies that

$$\frac{\log \Pi_X[X > A]}{A} \to 0 \text{ as } A \to \infty. \tag{3.19}$$

The definition extends obviously to left heavy tailed distributions. (3.19) shows that the tails of a heavy-tailed Π_X decay slower than exponential. For the standard Cauchy distribution for example, the ratio in (3.18) is for large A approximately $h/(A+h)$. Generally, the limit in (3.18) is a constant for exponentially decreasing tails and infinity for those which decay faster than exponential.

Lemma 3.2.1. *D1 implies that Π_X is right heavy tailed.*

PROOF Take any $h > 0$. Then for every $\epsilon > 0$ there exists an $A > 0$ such that for all $x > A$

$$|\pi_X(x+h) - \pi_X(x)| < \epsilon \pi_X(x).$$

Then

$$\begin{aligned} \int_A^\infty |\pi_X(x+h) - \pi_X(x)| dx &\geq \left| \int_A^\infty \{\pi_X(x+h) - \pi_X(x)\} dx \right| \\ &= \Pi_X[A < X < A+h] \end{aligned}$$

thus

$$\frac{\Pi_X[A < X < A+h]}{\Pi_X[X > A]} < \epsilon.$$

Notice that the above inequality holds also for all $L > A$, therefore the result follows since ϵ can be chosen arbitrarily small. □

Lemma 3.2.2. *D1 implies that for every $h \in \mathbb{R}$*

$$\frac{\pi_X(x-h)}{\pi_X(x)} \to 1, \text{ as } x \to \infty.$$

The proof is immediate. For densities with exponential tails this limit is a constant and with tails lighter than exponential is infinity.

Lemma 3.2.3. *We define*

$$f(y) := \int \pi_X(x)\pi_C(y-x)dx = \int \pi_X(y-x)\pi_C(x)dx \tag{3.20}$$

where π_C is any density on $\mathbb{R}$. Then, D1 and D3 imply that

$$\frac{f(y)}{\pi_X(y)} \to 1 \text{ as } y \to \infty.$$

PROOF For any $\delta > 0$ there exists some $h > 0$ such that

$$\int_{-h}^{h} \pi_C(x)dx = 1 - \delta$$

since π_C is a probability density function. For that h, D1 implies that for any $\epsilon > 0$ there exists some $A > 0$, such that for all $y > A$

$$\left| \frac{\pi_X(y-x)}{\pi_X(y)} - 1 \right| < \epsilon.$$

On the other hand, D3 implies that both

$$\int_{h}^{\infty} \frac{\pi_X(y-x)}{\pi_X(y)} \pi_C(x)dx$$

and

$$\int_{-\infty}^{-h} \frac{\pi_X(y-x)}{\pi_X(y)} \pi_C(x)dx$$

converge to 0 as $h \to \infty$. Therefore, it is not difficult to see that $f(y)/\pi_X(y) \to 1$ as $y \to \infty$. □

Lemma 3.2.4. *D1 and D3 imply that* $\Pi_{y-X|Y=y} \Rightarrow \Pi_C$ *as* $y \to \infty$.

PROOF Lemma 3.2.2 and Lemma 3.2.3 immediately imply point-wise convergence of the densities $\pi_{y-X|Y=y}(x) \to \pi_C(x)$ as $y \to \infty$ for every $x \in \mathbb{R}$. By Lemma 3.2.3 follows that there exists some $A > 0$ such that $\pi_X(y)/f(y) < 2$ for all $y > A$, therefore $\pi_{y-X|Y=y}(x) \leq 2k(x)\pi_C(x)$ for all $y > A$ ($k(x)$ is defined in (3.17)), the function on the right side of the inequality being integrable as a consequence of D3. Thus, we can use the dominated convergence theorem to prove the lemma. □

As we remarked earlier, if the distributions of X and C are interchanged, D1 and D3 imply that $\Pi_{X|Y=y} \Rightarrow \Pi_X$. O'Hagan (1979) strengthens D2 and D3 slightly but only to make them easier to verify. Whereas Dawid imposes conditions on both Π_X and Π_C, O'Hagan (1979) studies outlier proneness and resistance for linear models only in terms of Π_C. Although his work is relevant to our purposes, we will not pursue this connection further, since Dawid's conditions are enough for the results of this chapter; see Roberts and Papaspiliopoulos (2003) for extensions.

3.2.2 Our approach

This chapter concentrates on the simple model (3.15) where the linear structure is imposed on both the observation and the latent equations, and where the improper uniform prior is

chosen for Θ. The location structure of the model ensures that the posterior for Θ is proper under this prior elicitation. Moreover, we want to keep the observed $Y = y$ fixed and use the Dawid/O'Hagan conditions to derive results concerning

$$\Pi_{X|Y,\Theta=\theta} \text{ as } \theta \to \infty.$$

However, we can rearrange the equations in (3.15) as

$$\begin{aligned} \Theta &= X - \tilde{X} \\ X &= Y - C \end{aligned}$$

and notice that when $\tilde{X}$ and C have symmetric distributions, model (3.15) can be written equivalently (in distribution) as

$$\begin{aligned} \Theta &= X + \tilde{X} \\ X &= Y + C \end{aligned}$$

This form allows us to use the Dawid/O'Hagan conditions directly to study the limiting form of $\Pi_{X|Y,\Theta=\theta}$ as $\theta \to \infty$.

We use the two-component Gibbs sampler (see Section 1.5.2) to obtain samples from the posterior distribution of (X, Θ) in (3.15) given an observation $Y = y$. The algorithm alternates between updating X and Θ from their conditional distributions, thus it produces a Markov chain $\{(X_n, \Theta_n)\ , n = 0, 1, \dots\}$ with transition density

$$P[(x_0, \theta_0), (x_1, \theta_1)] = \pi_{\Theta|X,Y}(\theta_1 \mid x_0, y)\pi_{X|\Theta,Y}(x_1 \mid \theta_1, y).$$

The transition kernels of the marginal reversible Markov chains $\{X_n,\ n = 0, 1, \dots\}$ and $\{\Theta_n,\ n = 0, 1, \dots\}$ are denoted by P_X and P_Θ respectively and their densities with respect to the Lebesgue measure by p_x and p_Θ respectively; for example

$$p_X(x_0, x_1) = \int \pi_{\Theta|X,Y}(\theta \mid x_0, y)\pi_{X|\Theta,Y}(x_1 \mid \theta, y)d\theta.$$

The above definitions extend naturally when the NCP parameterisation is used and the Gibbs sampler updates $\tilde{X}$ and Θ.

3.3 Convergence of the CA and the NCA for the Cauchy-Gaussian model

This section proves that the NCA for the model in (3.15), where $\tilde{X} \sim \mathrm{N}(0,1)$ and $C \sim \mathrm{Ca}(0,1)$, is uniformly ergodic and the CA fails to be geometrically ergodic.

The joint posterior density of (X, Θ) given an observation $Y = y$ is given by

$$\pi_{X,\Theta|Y}(\theta, x \mid y) \propto \frac{e^{-(x-\theta)^2/2}}{1+(y-x)^2}. \tag{3.21}$$

and the corresponding posterior density of $(\tilde{X}, \Theta)$ is

$$\pi_{\tilde{X},\Theta|Y}(x, \theta \mid y) \propto \frac{1}{1+(y-x+\theta)^2} \exp\{-x^2/2\}. \tag{3.22}$$

Lemma 3.3.1. *When $\Pi_{\tilde{X}}$ is a standard Gaussian and Π_C a standard Cauchy distribution then*

$$\Pi_{\tilde{X}|Y,\Theta=\theta} \;\Rightarrow\; \Pi_{\tilde{X}}$$

Proof This statement follows immediately by Lemma 3.2.4, since the Cauchy and the normal satisfy D1 and D3 given in Section 3.2.1. □

This lemma formalises the notion of asymptotic ($\Theta \geq \theta,\ \theta \to \infty$) posterior independence between $\tilde{X}$ and Θ. On the contrary, this result implies that for large θ, $\Pi_{X|Y,\Theta=\theta}$ is roughly a $\mathrm{N}(\theta, 1)$ distribution, therefore X becomes independent of the observed data $Y = y$.

Notice that $\tilde{X}$ is marginally independent of Y, namely it is not identified by the data, and $X \mid Y = y \sim \mathrm{Ca}(y, 1)$. These statements can actually be shown using the rather general result proved in Lemma 3.4.1.

Lemma 3.3.2. *$P_X(x_0, \cdot) \Rightarrow\ N(0,2)$ as $x_0 \to \infty$.*

Proof Direct calculation shows that

$$p_X(x_0, x_1) \propto \frac{1}{1+(y-x_1)^2} \exp\{-(x_1-x_0)^2/4\}.$$

Essentially, this can be seen as a posterior density arising from a model like (3.15), but where $C \sim N(0,2), \tilde{X} \sim \mathrm{Ca}(0,1)$, $\Theta = y$ and the observation is $Y = x_0$. Therefore Lemma 3.2.4 implies that $P_X(x_0, \cdot) \Rightarrow\ N(0,2)$ when $x_0 \to \infty$, as desired. □

Theorem 3.3.3. *The centered algorithm is not geometrically ergodic for the model (3.15) where $C \sim Ca(0,1)$ and $\tilde{X} \sim N(0,1)$.*

PROOF The proof is based on the notion of the capacitance of the Markov chain $\{X_n\,, n = 0, 1, \dots\}$ defined in Section 3.1.5. We aim to show that $\kappa = 0$, by identifying sets with arbitrarily small capacitance, which then demonstrates the result by Cheeger's inequality (3.13) and the result about de-initialising chains of Section 3.1.6; see Section 3.1.5 for details about the capacitance. Let $\kappa(h) = \kappa([h, \infty))$ then, for any $l > 0$ if the chain is started according to the stationary measure restricted to $[h, \infty)$

$$\begin{aligned}
\kappa(h) &= \mathsf{P}[X_1 < h \mid X_0 > h] \\
&= \mathsf{P}[X_1 < h \mid X_0 > h + l]\Pi_{X|Y}[X_0 > h + l \mid X_0 > h] \\
&+ \mathsf{P}[X_1 < h \mid h < X_0 < h + l]\Pi_{X|Y}[X_0 < h + l \mid X_0 > h] \\
&\leq \mathsf{P}[X_1 - X_0 < -l | X_0 > h] + \Pi_{X|Y}[X_0 < h + l \mid X_0 > h].
\end{aligned}$$

As $h \to \infty$ the second term converges to zero, due to (3.18), while the first term converges to $\Phi(-l/\sqrt{2})$, due to Lemma 3.3.2, which can be chosen to be arbitrarily small for sufficiently large l.

This implies that the algorithm's capacitance κ must be 0 which implies that geometric ergodicity fails by Cheeger's inequality (3.13). □

Lemma 3.3.4. *Let* $p(\theta) = \Pi_{\tilde{X}|\Theta,Y}(\tilde{X} \in [-1, 1]|\theta, y)$. *Then*

1 p *is continuous;*

2 $p(\theta) > 0$ *for all* $\theta \in \mathbb{R}$*;*

3 $\lim_{\theta \to \pm\infty} p(\theta) = 1 - 2\Phi(-1)$*;*

4 $\delta := \inf_{\theta \in \mathbb{R}} p(\theta) > 0$.

PROOF From (3.22) immediately follows that $\pi_{\tilde{X}|\Theta,Y}(x \mid \theta, y)$ is continuous in θ. Therefore, we can use a standard result about continuity of integrals over bounded areas of continuous functions to show 1. 2 follows since p is obtained by integrating an everywhere positive function. 3 follows directly from Lemma 3.3.1. 4 follows from 1, 2 and 3 using standard compactness and continuity arguments. □

Theorem 3.3.5. *The non-centred algorithm is uniformly ergodic for the model (3.15) where* $C \sim Ca(0, 1)$ *and* $\tilde{X} \sim N(0, 1)$.

PROOF

We will prove the theorem by showing that the marginal Markov chain $\{\Theta_n\,, n = 0, 1, \dots\}$ is 1−small (see Definition 3.1.2). This will then prove the theorem, due to condition 4 of Theorem 3.1.1.

First note that

$$\pi_{\Theta|\tilde{X},Y}(\theta \mid x, y) = \frac{1}{\pi}\frac{1}{1+(\theta-(x-y))^2}$$

which implies that Θ given Y and $\tilde{X}$ has a Cauchy distribution. If $x \in [-1, 1]$ then

$$\begin{aligned}\pi_{\Theta|\tilde{X},Y}(\theta \mid x, y) &\geq \frac{1}{\pi}\frac{1}{1+(|\theta|-(x-y))^2} \\ &\geq \frac{1}{\pi}\frac{1}{1+(|\theta|+1+y)^2}.\end{aligned}$$

Therefore,

$$\begin{aligned}p_\Theta(\theta_0, \theta_1) &\geq \int_{-1}^{1} \pi_{\Theta|\tilde{X},Y}(\theta_1 \mid x, y)\pi_{\tilde{X}|\Theta,Y}(x \mid \theta_0, y)dx \\ &\geq \frac{1}{\pi}\frac{1}{1+(|\theta_1|+1+y)^2}\delta\end{aligned}$$

due to property 4 given in Lemma 3.3.4. The function appearing on the right hand side is clearly integrable, therefore we take $\phi(\cdot)$ in (3.3) to be have density the normalised version of this function and ϵ to be the product of δ/π and the normalising constant. □

Figure 2.3 shows output from the implementation of both the CA and the NCA for different starting values for Θ. The CA exhibits unstable heavy-tailed excursions characteristic of algorithms which fail to be geometrically ergodic, while the NCA appears to return to the distribution mode very rapidly.

3.4 The general result

The steps in the proofs of Theorem 3.3.3 and Theorem 3.3.5 can be easily replicated for the more general setting, where we don't assume particular forms for Π_C and Π_X but we only ask that they satisfy the Dawid's conditions of Section 3.2.1 and they are symmetric. We first need to establish the following useful result, where we assume the existence of densities for the corresponding measures for simplicity. Using this, we can prove the general theorem.

Lemma 3.4.1. *If C is symmetric in (3.15) and $\pi_\Theta(\theta) \propto 1$ then $\pi_{\tilde{X}|Y}(x \mid y) = \pi_{\tilde{X}}(x)$. When $\tilde{X}$ is symmetric, $\pi_{X|Y}(x \mid y) = \pi_C(y - x)$.*

Proof

$$\pi_{\tilde{X},\Theta|Y}(x, \theta \mid y) \propto \pi_C((y - x) - \theta)\pi_{\tilde{X}}(x)$$

since $\pi_\Theta(\theta) \propto 1$ and $\tilde{X}$ is apriori independent of Θ. Integrating both sides with respect to θ and exploiting the symmetry of C yields the required result. Similar argumentation shows the second property. □

Theorem 3.4.2. *If $\Pi_{\tilde{X}}$ satisfies D1 and Π_C is such that D3 holds, then the CA is uniformly ergodic, while the NCA fails to be geometrically ergodic. If the roles of Π_C and $\Pi_{\tilde{X}}$ are interchanged then the NCA is uniformly ergodic and the CA converges at a sub-geometric rate.*

For example, if Π_C is a $\text{Ca}(0, c_y)$, then it satisfies conditions 1 and 2. On the other hand it can be shown that

$$\frac{\pi_C(y-x)}{\pi_C(y)} \leq M + (x/c_y)^2$$

for a suitably chosen $M > 0$. Thus, whenever the tails of $\pi_{\tilde{X}}$ are such that polynomial moments exists the NCA is uniformly ergodic and the CA fails to be geometrically ergodic. Examples of $\pi_{\tilde{X}}$ with this tail behaviour include the double exponential, the Gaussian and more generally all densities whose tails decay faster than exponential.

Nevertheless, there are many other interesting combinations of distributions Π_X and Π_C which do not satisfy D1 and D3 of Section 3.2.1. Roberts and Papaspiliopoulos (2003) characterise the rate of convergence for linear hierarchical models where $\Pi_{\tilde{X}}$ and Π_C can be any among the Cauchy, the Double exponential, the Gaussian and distributions with tails lighter than normal. This chapter concludes with a characterisation of the rate of convergence when both Π_X and Π_C are double exponential distributions.

3.5 The double exponential-double exponential model

This section shows geometric ergodicity for both the CA and the NCA when $\tilde{X}$ and C are standard double exponential random variables with density $\exp\{-|x|\},\ x \in \mathbb{R}$. We will do so establishing the existence of a drift condition as described in Section 3.1.5.

Theorem 3.5.1. *Both the non-centered and the centered algorithms are geometrically ergodic for the model (3.15) where $C \sim DEx(0,1)$ and $\tilde{X} \sim DEx(0,1)$.*

Proof We first show the result for the centered. We will do so by establishing the existence of a drift condition (3.8) for the function $V(x) = 1+|x|$. In the following paragraph we will present an argument which shows that

$$\lim_{x\to\infty} \frac{PV(x)}{V(x)} = 1/2. \tag{3.23}$$

The same argument can be applied to prove that the same limit is obtained as $x \to -\infty$. A byproduct of our argument is that $PV(x)$ is continuous in x, thus having established (3.23) and using the results of Section 3.1.5 the existence of a drift condition follows.

Without loss of generality we take $Y = y = 0$. It is easy to derive that

$$\text{when } \theta > 0, \quad \pi_{X|Y,\Theta}(x \mid 0, \theta) = \begin{cases} 1/\{2(1+\theta)\} \exp\{x\} & \text{when } x < 0 \\ 1/(1+\theta) & \text{when } 0 < x < \theta \\ 1/\{2(1+\theta)\} \exp\{\theta - x\} & \text{when } x > \theta \end{cases},$$

which is graphically illustrated in Figure 3.2, and similarly that

$$\text{when } \theta < 0, \quad \pi_{X|Y,\Theta}(x \mid 0, \theta) = \begin{cases} 1/\{2(1-\theta)\} \exp\{x - \theta\} & \text{when } x < \theta \\ 1/(1-\theta) & \text{when } \theta < x < 0 \\ 1/\{2(1-\theta)\} \exp\{x\} & \text{when } x > 0 \end{cases}.$$

These explicit forms allow us to directly compute after some algebra

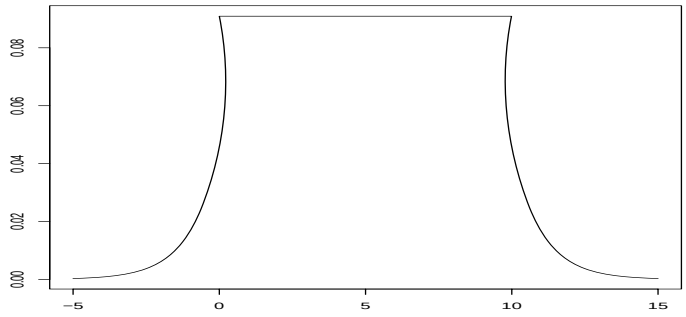


Figure 3.2: The conditional distribution of X given $Y = 0$ and $\Theta = 10$ in (3.15) where C and $\tilde{X}$ are double exponential random variables.

$$2\mathsf{E}[|X_1| \mid \Theta_1 = \theta] = \frac{\theta^2 + |\theta| + 2}{|\theta| + 1}; \tag{3.24}$$

notice that

$$\begin{aligned} \frac{\theta^2 + |\theta| + 2}{|\theta| + 1} &\to |\theta|, \text{ as} |\theta| \to \infty \\ \frac{\theta^2 + |\theta| + 2}{|\theta| + 1} &\leq 1 + \frac{|\theta|}{2}. \end{aligned} \tag{3.25}$$

Then, since $\Theta_1 \mid X_0 = x \sim \mathrm{DEx}(x, 1)$

$$\begin{aligned}
2\mathsf{E}[|X_1| \mid X_1 = x] &= \int \mathsf{E}[|X_1| \mid \Theta_1 = \theta]\pi_{\Theta|X}(\theta \mid x)d\theta \\
&= \int_{-\infty}^{x} \mathsf{E}[|X_1| \mid \Theta_1 = \theta]e^{\theta - x}d\theta + \int_{x}^{\infty} \mathsf{E}[|X_1| \mid \Theta_1 = \theta]e^{x-\theta}d\theta \\
&= e^{-x}\int_{-\infty}^{x} \frac{\theta^2 + |\theta| + 2}{1 + |\theta|}e^{\theta}d\theta + e^{x}\int_{x}^{\infty} \frac{\theta^2 + \theta + 2}{1 + \theta}e^{-\theta}d\theta \\
&\leq e^{-x}\int_{-\infty}^{x} (1 + \frac{|\theta|}{2})e^{\theta}d\theta + e^{x}\int_{x}^{\infty} (1 + \frac{\theta}{2})e^{-\theta}d\theta \\
&= (x + 1) + 1 + e^{-x}
\end{aligned}$$

from which immediately follows (3.23). Therefore, the theorem is proved for he CA.

The same procedure can be repeated for the NCA, since $\pi_{\tilde{X}|Y,\Theta}(x \mid 0, \theta)$ has a similar mixture form as X and $\Theta \mid \tilde{X}, Y \sim \mathrm{DEx}(Y - \tilde{X}, 1)$. □

Chapter 4

General non-centered parameterisations and state space expansion

4.0 Introduction

We have been rather loose in defining the NCP for general hierarchical models so far. This chapter gives the general definition and describes the corresponding Metropolis-Hastings algorithm. It also introduces a technique which can be used to construct easy to implement non-centered parameterisations for a wide range of univariate distributions. The aim is, when X is a one-dimensional random variable with some parameters Θ, to find another random object $\tilde{X}$ which is *a priori* independent of Θ and such that X is a deterministic function of $\tilde{X}$ and Θ. We observe that this function can be non-invertible and therefore $\tilde{X}$ might live on a larger space than X.

This section introduces a specific state space expanded class of NCPs for infinitely divisible and related distributions. The transformed missing data $\tilde{X}$ is taken to be a Lévy process. By means of an example, we illustrate that this technique can be easily implemented. An immediate application is the construction of NCPs for most of the hierarchical generalised linear models, which are presented for example in Lee and Nelder (1996).

We carry out a simulation study in order to assess the performance of the Hastings-within-Gibbs sampler under a state space expanded NCP. We compare it with the performance of the sampler under an alternative NCP, which avoids the state space expansion, for some hierarchical models where Θ is a scale parameter. It is found that the state space expanded NCP has worse performance and some conjectures are made in order to explain the difference in efficiency among the competing NCPs.

Nevertheless, the state space expansion proves to be a very useful tool when constructing

NCPs for models with hidden stochastic process. This problem is investigated in Chapter 5.

4.1 General non-centered parameterisations

The notions of centered and non-centered parameterisations for a hierarchical model with graphical representation as in Figure 1.3 have been introduced in Section 1.7 and Section 2.2. In Chapter 2 we gave some examples of these parameterisations applied to linear models. In this chapter we firstly formalise the notion of a non-centered parameterisation and describe a general MCMC algorithm for its implementation. Secondly we develop some methods that expand greatly the range of models that an NCP can be applied to.

The general setting is as follows. The distribution of the observed data Y depends on the unobserved/latent/missing data X and the distribution of the latter depends on some parameters Θ. X can live on an arbitrary space, in our examples in Chapter 5 and Chapter 6 it is a point process, but Θ typically takes values on some subset of the Euclidean space. We will assume the existence of a joint posterior density

$$\pi(X, \Theta \mid Y)$$

although not necessarily with respect to the Lebesgue measure, from which the conditionals up to proportionality can be derived.

The important feature of the NCP for the simple models we have studied so far that can be extracted to a much more general context, is the orthogonality of the prior structure. Specifically, we need to find some random quantity $\tilde{X}$ which is *a priori* independent of Θ and some function h such that

$$X = h(\tilde{X}, \Theta) \ . \tag{4.1}$$

Notice at this point that *a priori* independence between $\tilde{X}$ and Θ makes sense even if an improper prior is chosen for Θ (see Section 1.4).

We assume that the Hastings-within-Gibbs sampler is used to obtain samples from the joint posterior distribution of parameters and missing data. When the centered parameterisation is employed the target distribution is the distribution of (X, Θ) given Y and the algorithm is termed the centered algorithm (CA). When a non-centered parameterisation is employed then the target distribution is the distribution of $(\tilde{X}, \Theta)$ given Y and the algorithm is called the non-centered algorithm (NCA). Thus, the CA is described below.

A Hastings-within-Gibbs to sample from $(\Theta, X) \mid Y$ *(CA)*

Iterate the following steps:

1. Update Θ according to $\pi(\Theta \mid X)$
2. Update X according to $\pi(X \mid \Theta, Y)$

The difference between the CA and the NCA lies in the step which updates the parameters given the missing data. The NCA is described below.

A Hastings-within-Gibbs to sample from $(\Theta, \tilde{X}) \mid Y$ *(NCA)*

Iterate the following steps:

1. Update Θ according to $\pi(\Theta \mid \tilde{X}, Y)$
2. Transform $(\Theta, \tilde{X}) \to X$
3. Update X according to $\pi(X \mid \Theta, Y)$
4. Transform $(\Theta, X) \to \tilde{X}$.

This is a convenient way to implement the NCA, which shows how to exploit existing computer code made for the corresponding CA. Step 3 of the NCA coincides with Step 2 of the CA, while Steps 2 and 4 are transformations. Notice that when a direct simulation is possible at Step 3, the transformation at Step 2 is unnecessary.

Chapter 2 studied when a non-centered parameterisation is preferable to a centered parameterisation for Gaussian models, based on exact convergence rate analysis. For more general models such exact quantitative results are not available, nevertheless Chapter 3 provided a qualitative comparison of the two schemes for linear non-Gaussian models. Generally speaking, we expect the NCA to perform well when the missing data are weakly identified by the observed data. Of course, by construction if X is not identified (that is its posterior is the same as its prior), neither is Θ and certainly this does not represent an interesting inferential problem. Instead, we are interested in cases where certain aspects of the missing

data are not identified. For example, in the normal linear models of Chapter 2 we found the NCP to be preferable when there was little information about the individual effects (therefore the variance of $X_i - \overline{X}$ where $\overline{X} = \sum_{i=1}^{m} X_i/m$ is very small). For that model we can keep the information about the population mean Θ fixed and vary the information about individual effects by changing κ defined in (2.16). In Chapter 6 we find that for stochastic volatility models some ergodic characteristics are well identified, nevertheless other aspects of the latent structure are weakly identified. This feature is often observed in models with latent stochastic processes.

In principle, an NCP always exists, although it might not be analytically sufficiently tractable to be of any practical use. If for example X is a unidimensional random variable with distribution function $F_\Theta(x)$ then it is well known (see e.g. Ripley (1987)) that if $U \sim \text{Un}(0,1)$, then $X \stackrel{d}{=} F_\Theta^{-1}(U)$, therefore we could take $\tilde{X} = U$ and then $X = h(\tilde{X}, \Theta) = F_\Theta^{-1}(\tilde{X})$. However, in most cases F_Θ^{-1} is analytically intractable and such a construction would be of no practical use. Specifically, we are interested in NCPs for which we can easily perform the transformations Steps 2 and 4 of the algorithm, and where Step 1 can be handled using some Metropolis-Hastings mechanism.

We have already seen cases where it is straightforward to construct NCPs that are very easy to implement. For example, for the Gaussian model with unknown location parameter (2.14) h and $\tilde{X}$ are identified simply as

$$\begin{aligned} X = h(\tilde{X}, \Theta) &= \Theta + \tilde{X} \\ \tilde{X} &\sim N(0, \sigma_x^2). \end{aligned}$$

More generally, whenever Θ is a location parameter for the prior distribution of X, the transformation

$$X = \tilde{X} + \Theta$$

will produce a valid and tractable NCP, while when Θ is a scale parameter we can set

$$X = \Theta \tilde{X}.$$

However, it is less obvious how to devise an NCP for a variety of distributions often used in practice, consider for example the case where $X \sim \text{Ga}(\Theta, 1)$ and $\Theta > 0$.

In the following sections we will describe a technique, which extends the range of models for which an NCP can be constructed, and it is based on the observation that the function h in (4.1) can be non-invertible. Therefore, we find an $\tilde{X}$ which lives on a higher dimensional space than X, and we term this a *state space expanded NCP*. It will then be the case that, the transformation Step 4 of the NCA will be stochastic.

It is important to note that, once a state space expanded NCP has been constructed, it

is most likely that a direct simulation at Step 1 of the Hastings-within-Gibbs algorithm will be impossible and a Metropolis-Hastings step will have to be used instead; see for example Section 4.2 and Section 5.3.1. The state space expanded NCPs we propose typically result in discontinuous conditional densities $\pi(\Theta \mid \tilde{X}, Y)$ (see for example Section 4.2). It turns out that this can be a serious drawback of the method, in the sense that although an NCP might be preferable (for example due to poor identifiability of the latent structure), a state space expanded NCP might be not very efficient due to the slow mixing of the Metropolis step used to update the parameters, as a result of the roughness of the target density. This issue is thoroughly discussed in Section 4.3 and Section 6.12.2.

When X is a unidimensional random variable, we can explore the relationship between infinitely divisible distributions and Lévy processes (see Section 1.8 for definitions) to construct state space expanded NCPs. The following section illustrates this method by means of an example. Section 4.3 compares state space expanded NCPs with other NCPs and makes some general comments about the efficiency of the former. Chapter 5 shows how similar techniques can be employed to construct NCPs for hidden stochastic processes.

4.2 NCPs for gamma random effect models by expanding the state space

This section describes a method for constructing state space expanded NCPs when the missing data are gamma random variables. To ease exposition and allow implementational and computational aspects to be discussed in some detail, we consider gamma random effect models. These models have a partially exchangeable structure as described by the graphical model in Figure 1.5 together with the specification that $X_i \sim \text{Ga}(\Theta, 1)$. At this moment we will not make specific assumptions regarding the distribution of Y_i given X_i.

For this example finding $\tilde{X}$ is far from straightforward, since location-scale transformations are not appropriate. Nevertheless, recall from Section 1.8 that the gamma distribution is infinitely divisible and it is the marginal distribution of a standard gamma process. That is, if $\tilde{X}_i(t),\ t \in [0, \infty)$ is a standard gamma process then $\tilde{X}_i(t) \sim \text{Ga}(t, 1)$. Therefore, we can construct an NCP by taking $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_m)$, a collection of m mutually independent standard gamma processes $\tilde{X}_i$, and the function h to be

$$h(\tilde{X}_i, \Theta) = \tilde{X}_i(\Theta) \text{ for all } i = 1, \ldots, m. \tag{4.2}$$

It is not straightforward to write down the joint posterior distribution of $(\tilde{X}, \Theta)$ but it is not necessary either. In order to implement the NCA (see Section 4.1) we only need to know up to proportionality the the conditional distributions $\tilde{X} \mid \Theta, Y$ and $\Theta \mid \tilde{X}, Y$, but these are very easy to derive. The two steps of the algorithm are described in detail below.

In parallel to demonstrating the general methodology, we provide a numerical and graphical illustration, for which we take $m = 1$ and assume that $Y_1 \sim \text{Ex}(X_1)$.

Step 1: Simulate $\tilde{X}$ conditionally on $\Theta,\ Y$

Due to the partially exchangeable structure of the model

$$\tilde{X}_i(\cdot) \perp\!\!\!\perp \tilde{X}_j(\cdot) \mid (\Theta, Y) \text{ for all } i \neq j.$$

Therefore, simulation of $\tilde{X}$ conditional on Y and Θ is done by simulating independently each of the processes $\tilde{X}_i$ conditionally on Y_i and Θ.

By construction, the distribution of $\tilde{X}_i(\Theta)$ conditionally on Y_i and Θ coincides with that of $X_i \mid Y_i, \Theta$, therefore a sample can be drawn using exactly the same procedure as for the corresponding step of the CA. Moreover, notice that also by construction

$$Y_i \perp\!\!\!\perp \{\tilde{X}_i(t),\ t \neq \Theta\} \mid \tilde{X}_i(\Theta).$$

Therefore, conditionally on $\tilde{X}_i(\Theta)$, the state of the process at any other time can be directly simulated from the prior as described in Section 1.8. An illustration of these simulations is given in Figure 4.1.

We will use a Metropolis-Hastings step to update Θ according to $\pi(\Theta \mid \tilde{X}, Y)$, which is described in the following paragraph. It turns out that only $\tilde{X}_i(\Theta)$ for each $i = 1, \ldots, m$ needs to be stored at the current step of the algorithm, rather than the whole path from each $\tilde{X}_i$ process.

Step 2: Update Θ conditionally on $\tilde{X}$ and Y

At this step of the algorithm we update Θ given Y and a sample path from each of the processes $\tilde{X}_i$. Recalling that $\tilde{X}$ and Θ are *a priori* independent, the conditional density that we wish to simulate from is proportional to

$$\prod_{i=1}^{m} \pi(Y_i \mid \tilde{X}_i(\Theta))\pi(\Theta). \tag{4.3}$$

Typically, some sort of a Metropolis-Hastings step is needed to (approximately) simulate from this conditional distribution. Suppose that the current value of the parameter is θ_0 and θ_1 has been proposed from some density $q(\theta_0, \theta_1)$. The Metropolis-Hastings acceptance ratio is

$$r = \frac{\pi(\theta_1 \mid Y, \tilde{X})q(\theta_1, \theta_0)}{\pi(\theta_0 \mid Y, \tilde{X})q(\theta_0, \theta_1)}, \tag{4.4}$$

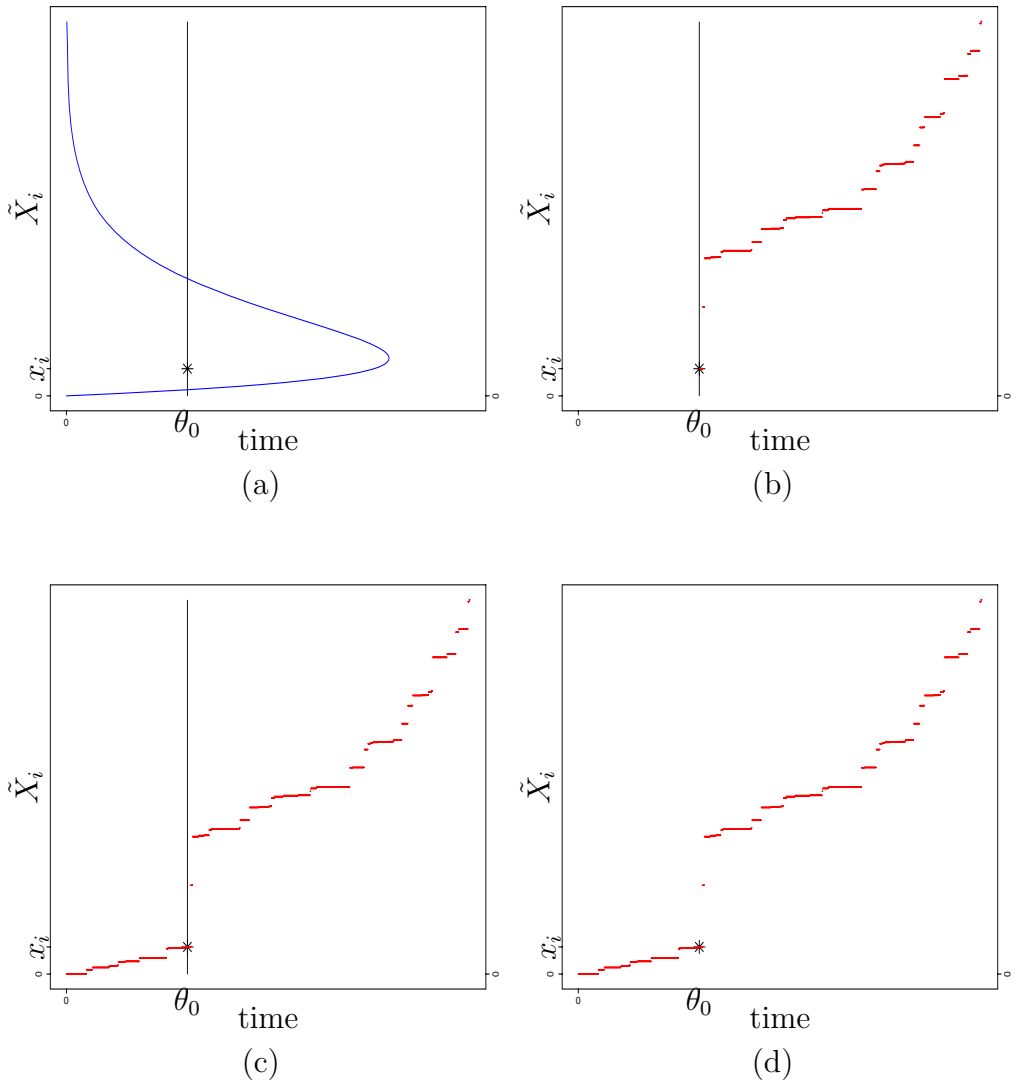


Figure 4.1: Steps to update $\tilde{X}_i$ conditionally on Y_i and Θ. The current value of Θ is denoted by θ_0. (a): Simulate the value of the process at time θ_0. Denote this by $x_i := \tilde{X}_i(\theta_0)$. (b): Given $\tilde{X}_i(\theta_0) = x_i$, simulate forwards in time a gamma process started from x_i at time θ_0. (c): Simulate a beta process started at time 0 from 0 and stopped at time θ_0 to x_i. (d): The new configuration for $\tilde{X}_i$. To produce these figures we have assumed the model $Y_i \sim \text{Ex}(X_i)$, initial values $\theta_0 = 3$ and assumed an observed data point $Y_i = y_i = 0.5$.

where $\pi(\Theta \mid Y, \tilde{X})$ is derived from (4.3), and the move from θ_0 to θ_1 will be accepted with probability $\min\{1, r\}$.

Therefore, in order to perform this updating step we only need to know the value of each stochastic process $\tilde{X}_i$ at times θ_0 and θ_1. $\tilde{X}_i(\theta_0)$ is available from the previous step of the algorithm and we have already shown how to simulate the value of the process at any time t conditionally on this value; an illustration is given in Figure 4.2.

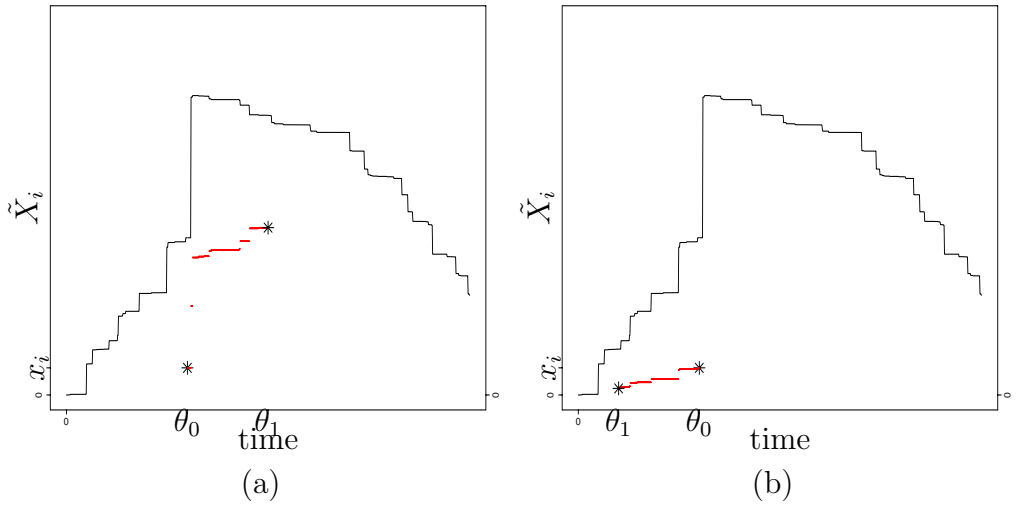


Figure 4.2: Updating of Θ conditionally on $\tilde{X}_i, Y_i$. The step function corresponds to the unnormalised density $\pi(Y_i \mid \tilde{X}_1(\Theta))$ as a function of Θ, where $\tilde{X}_1$ has the configuration shown in Figure 4.1.d. The product of this density and $\pi(\Theta)$ is the target density of this step of the algorithm. The proposed value for Θ is $\theta_1 > \theta_0$ in (a) and $\theta_1 < \theta_0$ in (b). Once a new value θ_1 has been proposed, we simulate the value of $\tilde{X}_i(\theta_1)$ from the prior, namely $\tilde{X}_i(\theta_1) \stackrel{d}{=} \tilde{X}_i(\theta_0) + G$, $G \sim \text{Ga}(\theta_1 - \theta_0, 1)$, if $\theta_1 > \theta_0$, and $\tilde{X}_i(\theta_1) \stackrel{d}{=} B\tilde{X}_i(\theta_0)$, $B \sim \text{Be}(\theta_1/\theta_0, (\theta_0 - \theta_1)/\theta_0)$, if $\theta_0 > \theta_1$. The paths in red colour in the plot show these simulations as gamma and beta process simulations respectively, as described in the previous step of the algorithm. Once $\tilde{X}_i(\theta_1)$ has been simulated the acceptance probability of the move from θ_0 to θ_1 given in (4.4) can be computed. To produce these figures we have assumed the model $Y_i \sim \text{Ex}(X_i)$, initial values $\theta_0 = 3$, proposed values $\theta_1 = 5$ in (a) and $\theta_1 = 1$ in (b), and assumed an observed data point $Y_i = y_i = 0.5$.

4.2.1 Non-centering for infinitely divisible and related distributions

The NCP described in the previous section can be applied whenever we can write $X = \tilde{X}(\Theta)$, where $\tilde{X}(\cdot)$ is a Lévy process, therefore it is applicable whenever the distribution of X is infinitely divisible (see Section 1.8). In the example of Section 4.2 $X \sim \text{Ga}(\Theta, 1)$ and $\tilde{X}(\cdot)$ is a standard gamma process. Another example, which is studied more closely in Section 4.3,

is when $X \sim \mathrm{N}(0, \Theta)$ and $\tilde{X}(\cdot)$ is taken to be a standard Brownian motion. Actually, this scheme is inspired by the Brownian motion interpretation of the normal linear model of Section 2.3.1.

Similar methods can be used for functions of random variables with infinitely divisible distributions. For example if $X \sim \mathrm{Be}(\alpha, \beta)$ and $\Theta = (\alpha, \beta)$ then $X = X_1/(X_1 + X_2),\ X_1 \sim \mathrm{Ga}(\alpha, 1), X_2 \sim \mathrm{Ga}(\beta, 1)$. One option in this example is taking $\tilde{X}(\cdot)$ to be a standard gamma process and writing $X = \tilde{X}(\alpha)/\tilde{X}(\alpha + \beta)$.

4.3 Comparison of different non-centering schemes

We have argued that the decision on whether to employ a CP or not depends on the statistical properties of the model and the type of data available. Nevertheless, even when an NCP is believed to be preferable to a CP, the computational implementation of the former might be much less efficient than of the latter. For example, this could happen when "pure" Gibbs steps, which are feasible under a CP, have to be replaced by Metropolis-Hastings steps when an NCP is employed. There is clearly a trade-off, since in presence of very high dependence between X and Θ the performance of the NCA might still be better than the CA, despite having to do relatively inefficient Metropolis-Hastings steps. Nevertheless, if more than one NCPs exist for the same model it is possible that their performance might be very different, as a result of the varying efficiency of the Hastings steps used in the algorithms. A challenging and open question is whether, if "pure" Gibbs steps were possible, all NCPs for a particular model would have similar performance. Our experience and results (reported here and in Section 6.12.2) suggest that the smoothness of the density $\pi(\Theta \mid \tilde{X}, Y)$ crucially affects the efficiency of the corresponding NCA. A Metropolis-Hastings step is typically used to update Θ and the performance of the algorithm critically depends on the smoothness of the target density; see Section 6.12.2 for more details and relevant references.

The results of this section are complementary to those of Section 6.12.2 and try to shed some light on two issues: firstly, whether the performance of the Hastings-within-Gibbs sampler varies with different NCPs for the same model, and if yes why. And, secondly, why NCAs perform poorly in situations where we would expect them to be very successful, as is observed for example in Chapter 6. The preliminary results presented here and in Section 6.12.2, as well as other simulation studies we have carried out but not included in this thesis, suggest that the fact that Metropolis-Hastings steps are used instead of direct simulations from the conditionals is very crucial and is related with both of the above questions. Whenever one NCA uses more efficiently Hastings steps than another, its performance is considerably better. The efficiency of the Hastings step might be very poor in cases where an NCP is preferable, thus the NCA performs poorly although a different implementation of the algorithm might have been very efficient.

This section compares empirically two different NCPs, one based on a state space expansion and one based on a scale transformation, for a class of latent distributions for which both of these parameterisations can be easily implemented: the stable family.

4.3.1 The stable family

We begin with some definitions. The distribution of a random variable X is said to belong in the stable family if it is infinitely divisible (see Section 1.8) and for any $c > 0$ there is a $b > 0$ such that

$$K(bu; X) = cK(u; X).$$

Stable Lévy processes are defined as those with marginal stable disributions. If $\tilde{X}(\cdot)$ is a stable Lévy process then

$$\tilde{X}(ct) \stackrel{d}{=} c^H \tilde{X}(t), \text{ for some } H \geq 1/2; \tag{4.5}$$

see Theorem 13.11 of Sato (1999) for a proof of this result. Property (4.5) is known as self-similarity and H is called the exponent of the stable process and it is uniquely determined by the process. The (self-similarity) index of the stable process is defined as $k = 1/H$, therefore $0 < k \leq 2$, where the boundary value $k = 2$ characterises the Brownian motion (and the Gaussian distribution correspondingly), while for $k < 2$ the associated stable distribution has infinite variance.

In the following suppose that $X \stackrel{d}{=} \tilde{X}(\Theta)$ where $\tilde{X}(\cdot)$ is a stable Lévy process. In this case, there exist two obvious NCPs for (X, Θ). The first is based on a state space expansion and simply takes $\tilde{X}$ to be the stable Lévy process and sets

$$X = \tilde{X}(\Theta).$$

We will refer to this as the lp-NCP. The second NCP exploits the self-similarity of the distribution of the Lévy process $\tilde{X}(\cdot)$. With a slight abuse of notation, let $\tilde{X} = \tilde{X}(1)$. Then a transformation of $(\tilde{X}, \Theta)$ to X is given by

$$X = \Theta^H \tilde{X},$$

where H is the exponent of the process. The validity of this transformation is ensured by (4.5). We refer to this non-centered scheme as the sc-NCP.

In the rest of this chapter we will try to compare the empirical performance of these two schemes focusing on some examples. In general, stable distributions have intractable densities thus simulating the increments of the corresponding Lévy processes is hard. Therefore, we choose to study the two most tractable stable distributions, the Gaussian and the

Cauchy.

4.3.2 Gaussian latent distribution

Suppose that $X \sim \mathrm{N}(0, \Theta)$, therefore $X \stackrel{d}{=} \tilde{X}(\Theta)$ where $\tilde{X}(\cdot)$ is a Brownian motion, which is a stable Lévy process with exponent $H = 1/2$ (see for example Sato (1999)). Therefore both the lp-NCP and the sc-NCP can be applied easily in this case.

We try both parameterisations for the linear hierarchical models

$$\begin{aligned} Y_i &= X_i + \epsilon_i \\ X_i &\sim N(0, \Theta),\ i = 1, \ldots, m. \end{aligned} \tag{4.6}$$

where $\epsilon_i \sim N(0, \sigma_y^2)$.

We simulate three different data sets from the first model with $m = 200,\ \sigma_y^2 = 1$ and three different values for Θ, $0.2, 1, 5$. An inverse gamma prior is chosen for Θ with both parameters equal to 1. For both NCAs the updating of $\tilde{X}$ given Θ and Y is straightforward. A Metropolis-Hastings step on the log-scale is used to update Θ given $\tilde{X},\ Y$. MCMC traces and autocorrelation plots are shown in Figure 4.3 for both algorithms.

Figure 4.4 shows MCMC traces when both algorithms are started far out in the target distribution tails, for the data corresponding to the middle row in Figure 4.3, i.e when $m = 200,\ \sigma_y^2 = \Theta = 1$. The lp-NCA has a random-walk type of behaviour when started in the tails, characteristic of algorithms which fail to be geometrically ergodic (see Chapter 3).

Figure 4.5 and Figure 4.6 plot $\log \pi(\Theta \mid \tilde{X}, Y)$ as a function of Θ for both the sc-NCP and the lp-NCP for data of various sizes simulated using $\Theta = \theta_T$, where $\theta_T = 0.2$ and $\theta_T = 5$ respectively. $\tilde{X}$ has been simulated from its conditional distribution given Y and the "true" value of Θ. Moreover, the simulation is done in a way where the transformation of $\tilde{X}$ and θ_T yields the same X in both algorithms. For the lp-NCP 20 realisation of $\tilde{X}$ have been drawn in order to show the variability of $\log \pi(\Theta \mid \tilde{X}, Y)$. Nevertheless, all realisations result in the same X when transformed using θ_T. See also Section 6.12.2 for some similar plots in a comparison of different NCPs for Poisson processes.

The relative qualitative behaviour of the two algorithms is very similar to the present case when the Gaussian observation error in (4.6) is replaced by Cauchy, therefore $\epsilon_i \sim \mathrm{Ca}(0, c_y)$. Again the lp-NCP has very slow convergence and very unstable excursions into the tails.

4.3.3 Cauchy latent distribution

Suppose that $X \sim \mathrm{Ca}(0, \Theta)$, then $X \stackrel{d}{=} \tilde{X}(\Theta)$ where $\tilde{X}(\cdot)$ is a standard Cauchy process, which is a stable Lévy process with exponent $H = 1$ (see for example Sato (1999)). For the standard

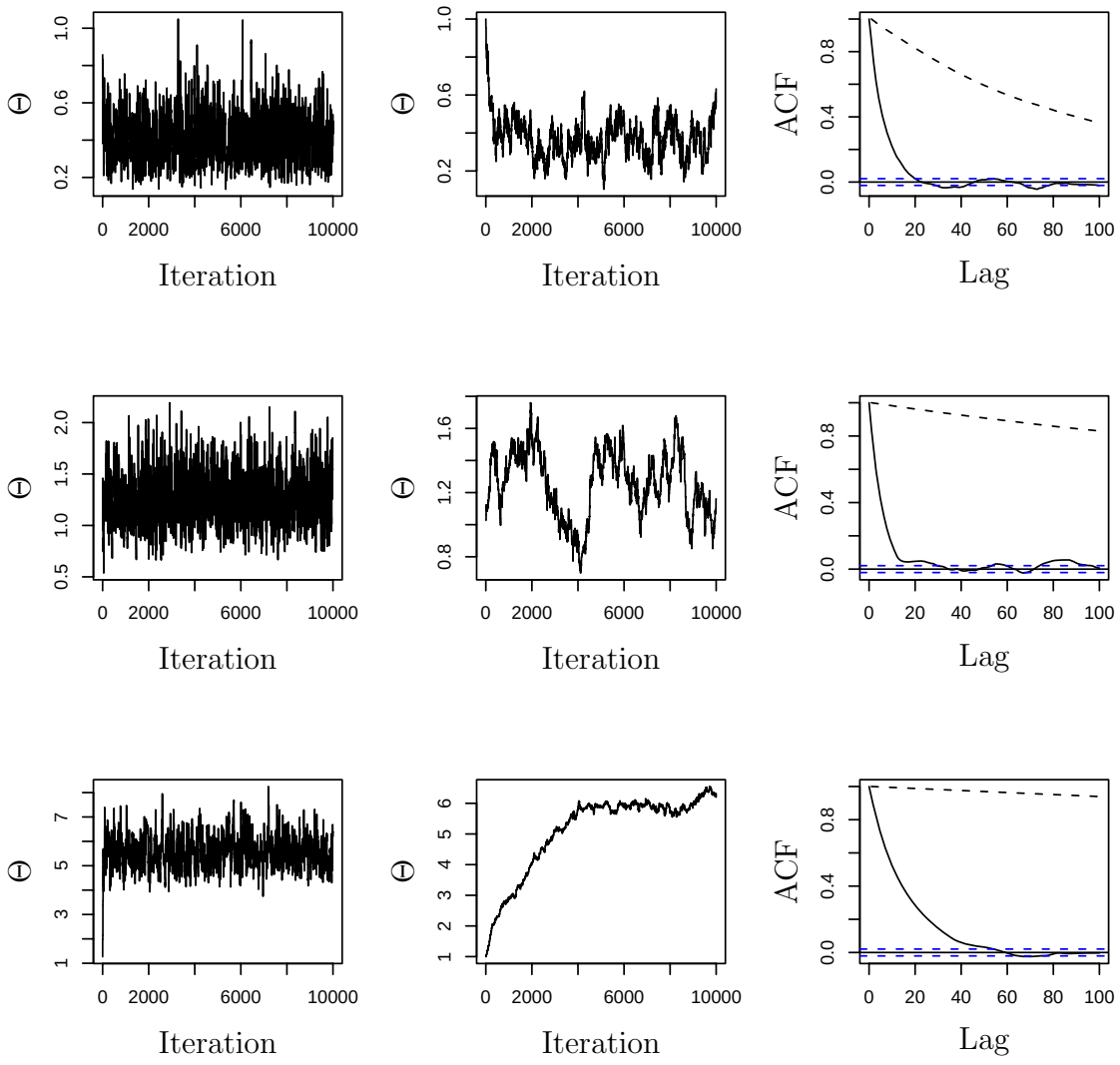


Figure 4.3: Simulation results from implementation of the sc-NCA and the lp-NCA for the normal hierarchical model with unknown latent variance Θ. First column shows trace plots for the sc-NCA while the second shows the corresponding trace plots for the lp-NCA. The last column superimposes the ACFs of the sample paths from the two implementations where the solid line corresponds to sc-NCA and the dashed to lp-NCA. Both algorithms have been run for 10^4 iterations, the first 10^3 being discarded for estimation of the ACFs as burn-in (clearly larger burn-in should be used for the lp-NCA in the last row). For the simulation we have taken $m = 200$, $\sigma_y^2 = 1$ and $\Theta = 0.2, 1, 5$ (first, second and third row respectively). An $\text{Ig}(1, 1)$ prior was chosen for Θ.

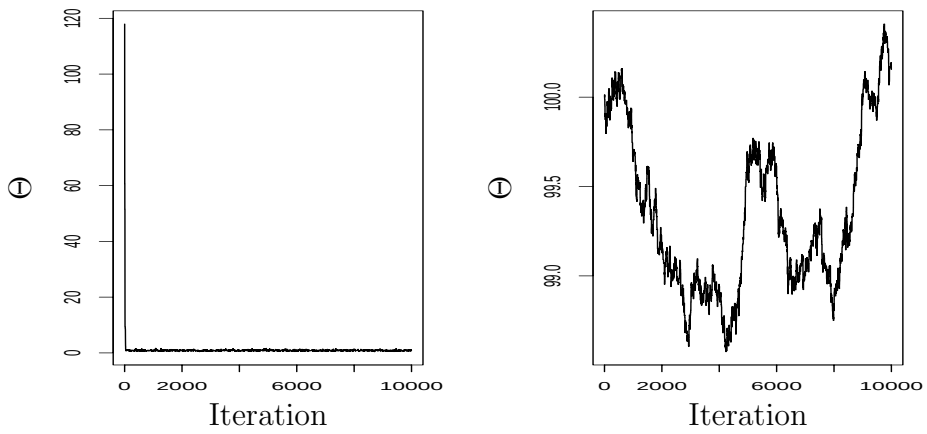


Figure 4.4: MCMC trace plots of the implementation of the sc-NCA (left) and the lp-NCA (right) for the normal hierarchical model with unknown latent variance Θ. Data were simulated under the specification $m = 200, \sigma_y^2 = \Theta = 1$. Both algorithms are initialised at $\theta_0 = 100$ and the proposal variances are tuned so that the overall acceptance rate is 0.2-0.3, although similar results were obtained for a variety of scaling schemes. The lp-NCA has a random-walk type of behaviour when started in the tails, characteristic of algorithms which fail to be geometrically ergodic.

Cauchy process,

$$\tilde{X}(t+s) - \tilde{X}(t) \mid \tilde{X}(t) \sim \text{Ca}(0, s),\ t, s > 0.$$

This setting allows us to use both the lp-NCP and the sc-NCP introduced earlier.

We apply them to the linear model

$$\begin{aligned} Y_i &= X_i + \epsilon_i,\ \epsilon_i \sim \text{Ca}(0, c_y) \\ X_i &\sim \text{Ca}(0, \Theta),\ i = 1, \ldots, m \end{aligned} \tag{4.7}$$

where c_y is considered to be known. Simulation results (not shown here) are in total agreement with those of the previous section and reveal that the lp-NCA is mixing much slower than the sc-NCA and has a random walk type behaviour when started from the tails.

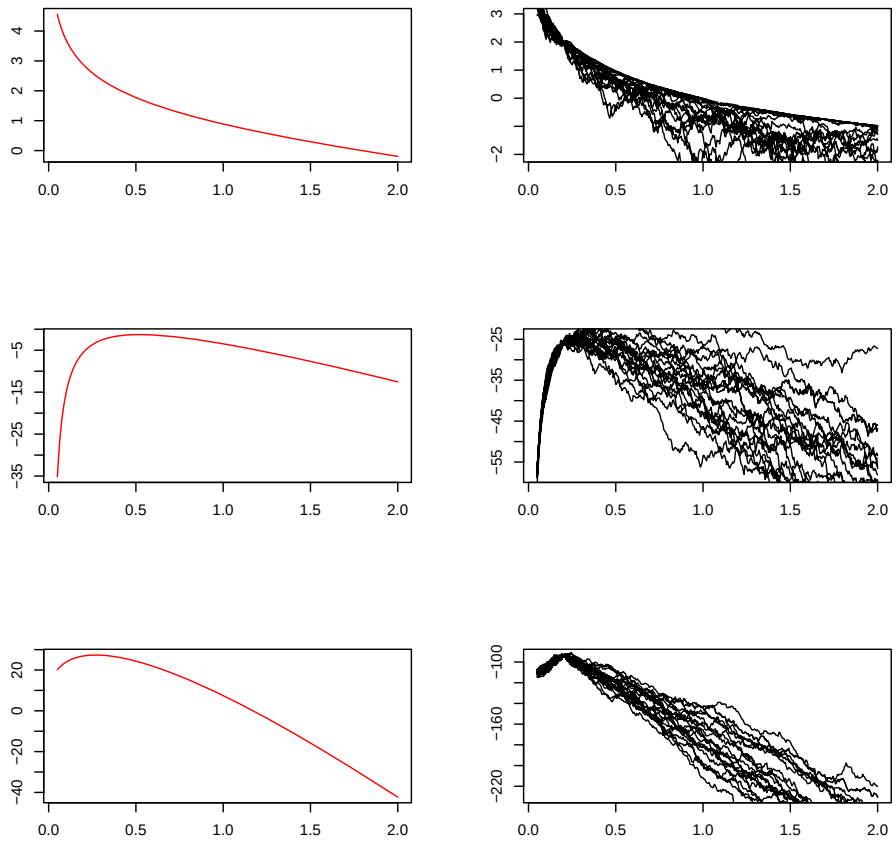


Figure 4.5: (Un-normalised) $\log \pi(\Theta \mid \tilde{X}, Y)$ as a function of Θ for sc-NCP (left) and lp-NCP (right) for the normal hierarchical model (4.6), with $m = 1$ (top), $m = 50$ (middle) and $m = 200$ data simulated from the model using $\Theta = 0.2$ and $\sigma_y = 1$. $\tilde{X}$ has been simulated from $\tilde{X} \mid Y, \Theta = 0.2$ for both algorithms. The simulation has been designed in such way that the transformation of $\tilde{X}$ and $\Theta = 0.2$ leads to the same X in both algorithms. For the lp-NCP 20 realisations of $\tilde{X} \mid Y, \Theta = 0.2$ have been drawn, all resulting to the same X and the corresponding functions $\pi(\Theta \mid \tilde{X}, Y)$ are superimposed.

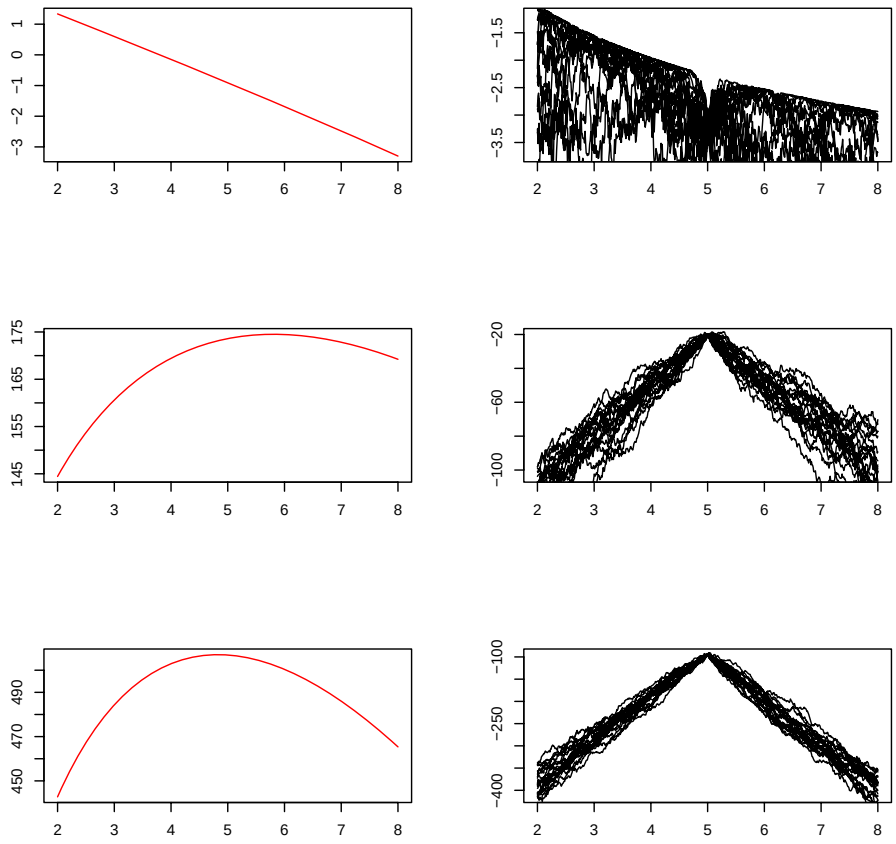


Figure 4.6: (Un-normalised) $\log \pi(\Theta \mid \tilde{X}, Y)$ as a function of Θ for sc-NCP (left) and lp-NCP (right) for the normal hierarchical model (4.6), with $m = 1$ (top), $m = 50$ (middle) and $m = 200$ data simulated from the model using $\Theta = 5$ and $\sigma_y = 1$. $\tilde{X}$ has been simulated from $\tilde{X} \mid Y, \Theta = 5$ for both algorithms. The simulation has been designed in such way that the transformation of $\tilde{X}$ and $\Theta = 5$ leads to the same X in both algorithms. For the lp-NCP 20 realisations of $\tilde{X} \mid Y, \Theta = 5$ have been drawn, all resulting to the same X and the corresponding functions $\pi(\Theta \mid \tilde{X}, Y)$ are superimposed.

Chapter 5

Non-centered parameterisations for Poisson processes

5.0 Introduction

This chapter develops a collection of non-centering techniques for latent Poisson processes, some of which are based on state space expansion. These methods are employed in Chapter 6, where inference for the non-Gaussian stochastic volatility models of Barndorff-Nielsen and Shephard (2001) is considered. The connection between Poisson processes, positive indepepndent increments processes and completely random measures, which is established in Section 5.7 and Section 5.8, renders our methodology potentially useful in fully Bayesian non-parametric inference for random distributions and related functions. For a review of this area see Walker et al. (1999), for applications in survival analysis see Walker and Damien (1998), in spatial modelling see Wolpert and Ickstadt (1998) and in inverse problems see Wolpert et al. (2003).

The chapter starts by reviewing some basic theory for Poisson processes, which will be used in our NCPs. We then formulate the problem of likelihood-based inference for Poisson processes and more generally Gibbs processes, and describe an MCMC algorithm for simulating Gibbs processes. We give two NCPs for Poisson processes, and show how they are implemented. We then establish the connection between completely random measures, positive independent increments and Poisson processes. In particular, we show that the well known Ferguson-Klass representation for positive Lévy processes is essentially an NCP.

Non-centered methodology has been developed for two more important families of stochastic processes; diffusion processes by Roberts and Stramer (2001) and Gaussian processes by Christensen and Waagepetersen (2003) and Christensen et al. (2003).

5.1 Poisson processes: review of basic definitions and properties

We begin by formally defining the Poisson process and presenting some of its properties that are more relevant to our purposes. The material presented below is largely based on the monograph by Kingman (1993), which is a short but thorough study of Poisson processes. Other important references include Stoyan et al. (1995), Karr (1991) and Norris (1997).

Informally, a point process on a set S is a random set of discrete points of S. More formally, it is a measurable function, Φ say, from some set Ω (associated with a probability triple) into the set of all countable subsets of S, which is called the state space of Φ. The careful measure-theoretic construction can be found, for example, in p.7-9 of Kingman (1993) and p. 99-101 of Stoyan et al. (1995). Some more details will be given in Section 5.1.4. The (random) number of points of Φ that lie on the measurable set $A \subset S$ is denoted by $\Phi(A)$. Every configuration of Φ, ϕ say, acts as a counting measure on S, therefore Φ can be thought of as a random counting measure on S (see Section 5.7). There is a dual treatment of point processes, namely as random counting measures and as random closed sets, see Stoyan et al. (1995) for details.

The Poisson process is a point process which is characterised by the following two properties:

1 for any disjoint measurable subsets $A_1, \ldots, A_n$ of S, the random variables $\Phi(A_1), \ldots, \Phi(A_n)$ are independent, and

2 $\Phi(A) \sim \mathrm{Pn}(\Lambda(A))$, where $0 \leq \Lambda(A) \leq \infty$.

It follows that

$$\Lambda(A) = \mathsf{E}[\Phi(A)]$$

and that Λ is a measure on S, hence it is called the mean measure of the Poisson process Φ. The only restriction on Λ is that it has to be non-atomic, that is $\Lambda(\{x\}) = 0$ for all $x \in S$. It is often the case that Λ is σ-finite, which means that there exists a partition $S = \cup_{i=1}^{\infty} S_i$ such that $\Lambda(S_i) < \infty$ for all $i = 1, 2, \ldots$. There are however, interesting examples where this does not hold, which will be encountered in Section 5.8.

When $\Lambda(S) < \infty$, that is when Φ has a finite number of points *almost surely*, conditionally on the event that $\Phi(S) = n$, the random points of $\Phi = \{X_1, \ldots, X_n\}$ are independent random variables, identically distributed according to the normalised mean measure $\Lambda(\cdot)/\Lambda(S)$.

In all the applications we will encounter in this thesis, S is some Borel subset of a Euclidean space. In these cases, if Λ is some multiple of the Lebesgue measure on S, the corresponding Poisson process is called homogeneous. When $S = [0, \infty)$ we will occasionally

refer to the random function

$$z(t) = \Phi((0, t])$$

as a Poisson process (see also Section 5.8); Figure 5.1 depicts the relationship between $z(\cdot)$ and Φ.

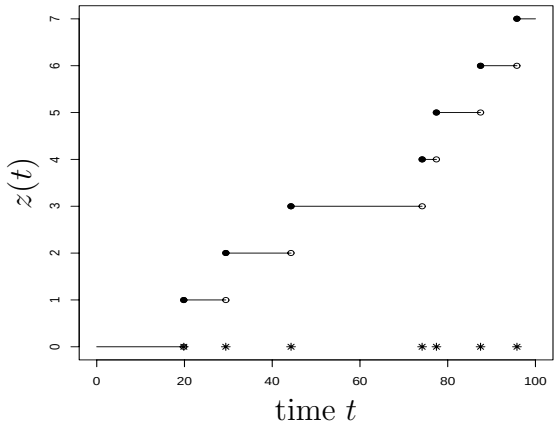


Figure 5.1: The relationship between the Poisson process Φ with state space the positive half-line, plotted with asterisks, and the jump function $z(\cdot)$. For the simulation we have taken $S = [0, 100]$ and Φ to be a homogeneous Poisson process with intensity 0.1.

5.1.1 Restriction, superposition and mapping properties

Suppose that Φ is a Poisson process on a set S with mean measure Λ and $S_1 \subset S$ is a measurable subset. Essentially by definition, the random closed set $\Phi \cap S_1$ is a Poisson process on S_1 with mean measure the restriction of Λ to S_1. It is also easy to show that if $S_1 \cap S_2 = \emptyset$ then $\Phi \cap S_1$ and $\Phi \cap S_2$ are independent Poisson processes. These properties will prove invaluable in the implementation of the NCAs described in Section 5.3.2 and Section 5.4.

Superposition, together with thinning and clustering (see Chapter 5. of Stoyan et al. (1995)), comprise the three basic operations that can be used to produce more complicated point processes from simpler ones. The description of thinning will be deferred to Section 5.2 while clustering will not be considered in this thesis. The superposition of the point processes $\Phi_1, \Phi_2, \ldots$ is the point process

$$\Phi = \bigcup_{n=1}^{\infty} \Phi_n$$

that contains the points of all the constituent processes. If $\Phi_n,\ n = 1, 2, \ldots$ are independent Poisson processes with corresponding mean measures $\Lambda_n,\ n = 1, 2, \ldots$ the superposition

theorem (p.16 in Kingman (1993)) states that Φ is also a Poisson process with mean measure

$$\Lambda = \sum_{n=1}^{\infty} \Lambda_n.$$

Another remarkable property of the Poisson process is that when its state space is mapped into another space, the transformed points form again a Poisson process. Suppose that Φ is a Poisson process on S with mean measure $\Lambda_\Phi(\cdot)$ and that h is a measurable function from S into some space T. The points $h(X)$ for $X \in \Phi$ form a random countable subset of T, Ψ say, and the mapping theorem asserts that under some mild conditions they form a Poisson process with mean measure

$$\Lambda_\Psi(A) = \Lambda_\Phi(h^{-1}(A)),\ A \subset T. \tag{5.1}$$

The conditions are that Λ_Φ is σ-finite and that the induced measure (5.1) has no atoms. It is still possible however that Λ_Ψ is not σ-finite. This theorem has far reaching implications (see the relevant discussion on p.21 of Kingman (1993)) and will explicitly be used in our NCPs for Poisson processes. Since we will come back to it many times in this chapter, we formally state it below and refer to Section 2.3 of Kingman (1993) for a formal proof.

Theorem 5.1.1. *Let Φ be a Poisson process with σ-finite mean measure $\Lambda_\Phi(\cdot)$ on the state space S, and let $h : S \to T$ be a measurable function such that the induced measure (5.1) has no atoms. Then $h(\Phi) =: \Psi$ is a Poisson process on T with mean measure defined in (5.1).*

5.1.2 Sums over Poisson processes

This section briefly reviews some results concerning sums of the form

$$C_f = \sum_{X \in \Phi} f(X) \tag{5.2}$$

where f is some real valued measurable function on S. Moreover, for the purposes of this chapter it is enough to assume that f takes positive values only, although the theory presented in this section does not require this assumption. The material of this section serves to demonstrate the connection between the Poisson process and other positive Lévy processes, which are going to be discussed later in this chapter. Moreover, the theory developed here provides some strong analytic tools to tackle problems raised in Chapter 6.

The main result, which is known as Campbell's theorem, establishes the conditions under which C_f is absolutely convergent with probability 1 and gives an expression for its cumulant function (defined in Definition 1.8.2):

Theorem 5.1.2. *Let Φ be a Poisson process on S with mean measure Λ and C_f be defined as in (5.2), where f is a positive function. Then C_f is absolutely convergent with probability 1 if and only if*

$$\int_S \min\{f(x), 1\}\Lambda(dx) < \infty$$

and then

$$K(u; C_f) := \log \mathsf{E}\{e^{-uC_f}\} = -\int_S (1 - e^{-uf(x)})\Lambda(dx). \tag{5.3}$$

In particular, whenever they exist

$$\begin{aligned} \mathsf{E}\{C_f\} &= \int_S f(x)\Lambda(dx) \\ \mathrm{Var}\{C_f\} &= \int_S f(x)^2\Lambda(dx). \end{aligned}$$

A proof, which is based on the 'standard machine' (in David Williams' terms, see p.59 of Williams (1991)) of integration theory can be found in Chapter 3 of Kingman (1993). The equality expressed in (5.3) can be used as a characterisation of the Poisson process. In particular, if (5.3) is true for $u = 1$, for some measure $\Lambda(\cdot)$ and for a sufficiently rich family of functions f (for example all functions that take a finite number of different values) then Φ is a Poisson process with mean measure $\Lambda(\cdot)$. For more details on this characterisation and its use in proving the celebrated Rényi's theorem see Section 3.3 and 3.4 of Kingman (1993). This is also used to prove Theorem 5.1.3.

5.1.3 Marked Poisson processes

Suppose that Φ is a Poisson process and that, to every $x \in \Phi$ it is assigned a random mark $m_x \in M$, which is generated independently of any other point of Φ and of the rest of the marks. It is yet another remarkable property of the Poisson process that when it is marked as described above it produces a Poisson process on the product space $S \times M$. The space M is subject to some mild measure-theoretic constraints, which are going to be satisfied in all the examples we will encounter. Essentially, we will only be concerned with problems where M is a Euclidean space, although there are many applications in stochastic geometry where M is a much more complicated space (see Stoyan et al. (1995) for such examples). The marks are generated from a probability "transition" kernel $P(x, \cdot)$ (defined in Definition 1.5.1).

Theorem 5.1.3. *The random set $\Psi = \{(X, m_X); X \in \Phi\}$, $m_X \in M$, constructed by marking a Poisson process Φ with mean measure Λ_Φ is a Poisson process on $S \times M$ with mean measure*

given by

$$\Lambda_\Psi(A) = \int_{(x,m)\in A} P(x, dm)\Lambda_\Phi(dx). \tag{5.4}$$

The proof is based on the characterisation of the Poisson process via (5.3) and is given in p.55 of Kingman (1993). Often m_X is independent of X and as a consequence Λ_Ψ is a product measure.

As a result of the marking and mapping theorems, the random set containing the marks $\{m_X; X \in \Phi\}$ forms a Poisson process on M with mean measure given by

$$\int_S P(x, A)\Lambda_\Phi(dx), \ A \subset M.$$

We can therefore immediately use Campbell's Theorem 5.1.2 to find the moment generating function of the sum

$$C_f = \sum_{X\in\Phi} f(m_X)$$

for measurable positive functions f on M. Suppose for example that m_X is positive and independent of X, then

$$K(u; C_f) = -\int_0^\infty (1 - e^{-uf(x)})\Lambda_\Phi(S)P(dx). \tag{5.5}$$

We will see in Section 5.7 and Section 5.8 that such expressions lie at the heart of the Lévy-Khinchine representation theorem for positive Lévy processes.

When S is the positive half-line and m_X is independent of X, we will call

$$z(t) = \sum_{X\in\Phi} m_X \mathbb{1}[X < t]$$

the compound Poisson process; the relationship between the marked Poisson process $\Psi = \{(X, m_X); X \in \Phi\}$ and $z(\cdot)$ is depicted in Figure 5.2. This relationship is explored in Chapter 6 to provide a data augmentation method based on marked Poisson processes for the non-Gaussian OU stochastic volatility models considered there.

5.1.4 Likelihood functions for Poisson and Gibbs processes

The results of this and the following section are particularly relevant to any computational method used to perform likelihood-based inference for point processes. Specifically, the implementation of both the CP and the NCP for the volatility model presented in Chapter 6

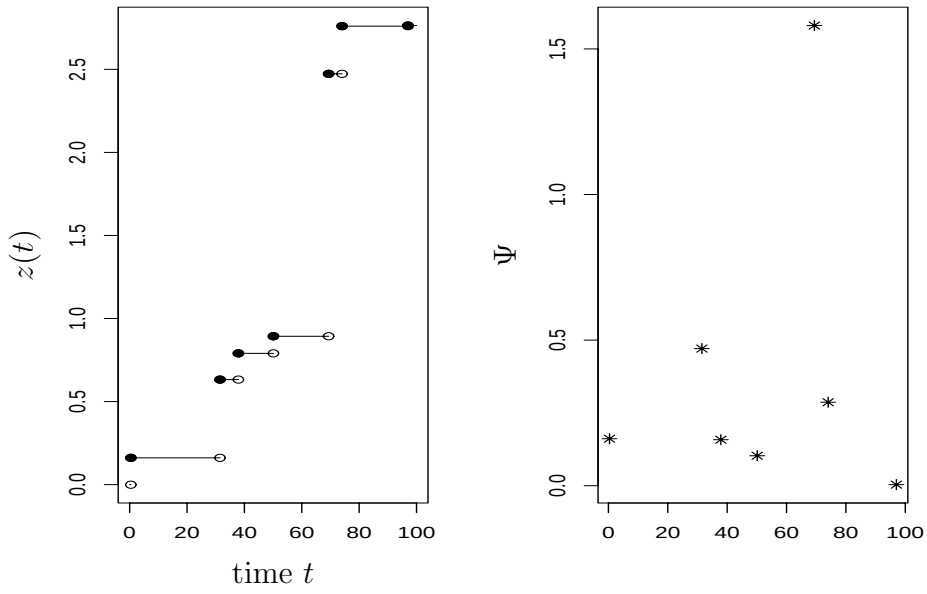


Figure 5.2: The relationship between the marked Poisson process Ψ, produced by marking a homogeneous Poisson process with intensity 0.1, with independent Ex(1) marks, and the compound Poisson process $z(\cdot)$; for the simulation we have taken $S = [0, 100]$.

is based explicitly on the results presented below. Moreover, it will become apparent later in this chapter that the choice of dominating measures and the issue of absolute continuity between stochastic processes, which are discussed in this section in the context of point processes, are very relevant to the construction of non-centered parameterisations. This connection was first observed and exploited in Roberts and Stramer (2001) in the context of inference for partially observed diffusions.

Our target is to derive the likelihood function for an arbitrary Poisson process and consequently to define the Gibbs process in terms of its density with respect to the Poisson process. We first need to add some detail about the measure-theoretic construction sketched in the beginning of Section 5.1. The development below follows Stoyan et al. (1995) and Geyer and Møller (1994) closely.

Let $(S, \mathcal{S}, \Lambda)$ be a measure space, where S is the state space of the Poisson process Φ (typically S is some measurable subset of a Euclidean space), $\mathcal{S}$ is the corresponding σ-algebra that contains all the singletons $\{x\} \subset S$ (typically a Borel σ-algebra), and $\Lambda(\cdot)$ is the mean measure (typically a measure with a density with respect to the Lebesgue measure on $(S, \mathcal{S})$). The theory presented in this section is concerned with processes for which $\Lambda(S) < \infty$. The corresponding exponential space (see Geyer and Møller (1994) and Carter and Prenter (1972) for example) is denoted by $(\Omega, \mathcal{F}, \mathcal{Q})$. Ω is the space containing

all sequences of elements from S, sometimes represented as

$$\Omega = \bigcup_{n=0}^{\infty} \Omega_n, \text{ where } \Omega_n = \{\{x_1, \ldots, x_n\} \subset S\}, \ \Omega_0 = \{\emptyset\}. \tag{5.6}$$

Notice that $\Omega_0 = \{\emptyset\}$ is the empty set configuration, i.e the configuration of Φ that contains no points, which is distinct from the empty subset of Ω. Clearly, realisations of Φ, denoted here with the lower case letter ϕ, take values in Ω. $\mathcal{F}$ is the smallest σ-algebra on Ω that makes measurable all mappings $\phi \to \phi(B)$ for $B \in \mathcal{S}$. The law of Φ on $(\Omega, \mathcal{F})$ is denoted by $\mathcal{Q}$ which is uniquely determined by the system of finite-dimensional distributions

$$\mathcal{Q}[\Phi(B_1) = n_1, \ldots, \Phi(B_k) = n_k], \ k = 1, 2, \ldots, \ n_1, \ldots, n_k \geq 0, \ B_1, \ldots, B_k \in \mathcal{S}$$

where the sets $B_1, \ldots, B_k$ can be taken to be pairwise disjoint. Actually, an important theorem for random closed sets (see for example Chapter 6 of Stoyan et al. (1995)) states that for any point process Φ the system of the so-called void probabilities

$$\mathcal{Q}[\Phi(B) = 0], \ B \in \mathcal{S}$$

is enough to characterise its distribution. Notice that this result holds for general point processes, not just for the Poisson process.

When Φ is a Poisson process with finite mean measure $\Lambda(\cdot)$, $\mathcal{Q}$ can be represented, for any $F \in \mathcal{F}$, as (see p.21 of Geyer and Møller (1994))

$$\begin{aligned} \mathcal{Q}(F) &= \exp\{-\Lambda(S)\} \left[\mathbb{1}[\Omega_0 \in F] + \right. \\ &+ \left. \sum_{n=1}^{\infty} \frac{1}{n!} \int \cdots \int \mathbb{1}[\{x_1, \ldots, x_n\} \in F] \Lambda(dx_1) \cdots \Lambda(dx_n) \right]. \end{aligned} \tag{5.7}$$

In particular, if f is some real measurable function on Ω

$$\begin{aligned} \int_F f(\phi) \mathcal{Q}(d\phi) &= \exp\{-\Lambda(S)\} \left[1(\Omega_0 \in F) f(\{\emptyset\}) \right. \\ &+ \sum_{n=1}^{\infty} \frac{1}{n!} \int \cdots \int \mathbb{1}[\{x_1, \ldots, x_n\} \in F] \qquad (5.8) \\ &\times \left. f(\{x_1, \ldots, x_n\}) \Lambda(dx_1) \ldots \Lambda(dx_n) \right]. \qquad (5.9) \end{aligned}$$

It is easy to show that a Poisson process Φ with finite mean measure $\Lambda(\cdot)$ has the distribution given by (5.7) using a conditioning argument (on the number of points of Φ) and the basic definitions and properties presented in the beginning of Section 5.1. We can also

show the converse, that if a point process Φ has a distribution given by (5.7) it is then a Poisson process with mean measure $\Lambda(\cdot)$. This can be easily achieved using the expression (5.9) together with the characteristic functional idea, discussed briefly at the end of Section 5.1.2 and in more detail in Section 3.3 of Kingman (1993).

Suppose now that a realisation of an arbitrary Poisson process Φ has been observed, $\phi = \{x_1, \ldots, x_n\}$ say. Moreover, some parametric form for the mean measure $\Lambda(\cdot) = \Lambda(\cdot; \Theta)$ has been assumed, where Θ is a vector of unknown parameters, and we are interested in obtaining the likelihood function, i.e the density of the data ϕ given the parameters Θ in order to perform likelihood-based inference for Θ. Since Ω is an infinite-dimensional space we can't expect this likelihood to be given with respect to to some Lebesgue measure, as it is conventionally done in statistics. Instead the dominating measure has to be infinite-dimensional.

When S is a bounded subset of a Euclidean space, a natural choice is the probability measure of a standard Poisson process. Namely, the dominating measure, denoted by $\mathcal{Q}$, is the one derived from (5.7) by setting Λ to be the Lebesgue measure on $(S, \mathcal{S})$. Most of the applications considered in this thesis are like that. An important exception is problems involving marked Poisson processes, which live on spaces of the form $S \times M$ where M can be unbounded. However, an immediate extension of the idea suggested above can be applied as long as the marked Poisson process is finite *almost surely*. At this early stage, we remark that actually the choice of a dominating measure is rather strongly related to the choice of an NCP for a stochastic process. We will come back to this point in Section 5.2, where NCPs for Poisson processes will be constructed.

Abstracting from specific choices, suppose that a dominating measure $\mathcal{Q}$ has been chosen, and that it corresponds to the distribution of a Poisson process on S with mean measure $K(\cdot)$. Suppose that we are interested in expressing the density of the distribution of Φ, $\mathcal{P}$ say, with respect to $\mathcal{Q}$. This can be done as described in the following lemma. For reasons of completeness we provide a proof.

Lemma 5.1.4. *Suppose that $\mathcal{Q}$ and $\mathcal{P}$ are the distributions of two Poisson processes with mean measures $K(\cdot)$ and $\Lambda(\cdot)$ respectively, where $\Lambda(S) + K(S) < \infty$. If Λ is absolutely continuous with respect to K with density $\Lambda(dx)/K(dx) = f(x)$, then $\mathcal{P}$ is absolutely continuous with respect to $\mathcal{Q}$ and the Radon-Nikodym derivative between the two measures evaluated at $\phi = \{x_1, \ldots, x_n\} \in \Omega$ is given by*

$$\frac{d\mathcal{P}}{d\mathcal{Q}} = \exp\{K(S) - \Lambda(S)\} \prod_{i=1}^{\phi(S)} f(x_i)$$

where the product is replaced by 1 if $\phi(S) = 0$.

PROOF For any $F \in \mathcal{F}$,

$$\begin{aligned}
\mathcal{P}(F) &= \int_F \mathcal{P}(d\phi) \\
&= \exp\{-\Lambda(S)\} \left[\mathbb{1}[\Omega_0 \in F] \right. \\
&+ \left. \sum_{n=1}^{\infty} \frac{1}{n!} \int \cdots \int \mathbb{1}[\{x_1, \ldots, x_n\} \in F] \Lambda(dx_1) \cdots \Lambda(dx_n) \right] \\
&= \exp\{-\Lambda(S)\} \left[\mathbb{1}[\Omega_0 \in F] \right. \\
&+ \left. \sum_{n=1}^{\infty} \frac{1}{n!} \int \cdots \int \mathbb{1}[\{x_1, \ldots, x_n\} \in F] f(x_1) \ldots f(x_n) K(dx_1) \ldots K(dx_n) \right] \\
&= \exp\{K(S) - \Lambda(S)\} \int_F g(\phi) \mathcal{Q}(d\phi)
\end{aligned}$$

where the last equality is true due to (5.9), with $g(\phi) = f(x_1) \cdots f(x_n)$, when $\phi = \{x_1, \ldots, x_n\}$, $n > 0$ and $g(\{\emptyset\}) = 1$. This proves the lemma. □

We note that a special case of the previous result is proved in p.167 of Stoyan et al. (1995) using the void probabilities mentioned in the beginning of this section.

Informally, the density derived in the lemma above evaluated at a particular realisation of the Poisson process Φ, ϕ, can be interpreted as a quantification of how more likely it is that ϕ was generated by a Poisson process with mean measure $\Lambda(\cdot)$ relative to have been generated by one with mean measure $K(\cdot)$.

Notice that application of the lemma above can fail when the point process contains infinite number of points on S. It can then be true that $\mathcal{Q}$ and $\mathcal{P}$ are mutually singular, despite absolute continuity of Λ with respect to K. An example is given in p.168 of Stoyan et al. (1995). Consider two homogeneous Poisson processes Φ_1, Φ_2 on $[0, \infty)$, with different intensities, $\lambda_1 \neq \lambda_2$. The mean measures are apparently equivalent, however the distributions of the processes are mutually singular, since with probability 1, $\lim_{n\to\infty} \Phi_i([0, n))/n = \lambda_i$, $i = 1, 2$ (see p.42 of Kingman (1993)). Therefore the measure of Φ_1 puts all the mass on configurations which have zero probability under the distribution of Φ_2, and vice versa. This situation is related with and motivates the non-centered parameterisations of Section 5.2. For a similar problem in the context of diffusion processes, where the diffusion sample path uniquely determines the variance of the process via the quadratic variation identity, see Roberts and Stramer (2001).

Gibbs processes and hierarchical modelling

New point processes can be produced from old by transforming their distribution by means of probability densities. This is the idea behind the theory of Gibbs processes. These

processes first arose in statistical physics (see for example Preston (1976)) and have been used as a modelling devise in many spatial statistics applications (see Section 5.5 of Stoyan et al. (1995)). We will shortly see how these processes arise in the context of hierarchical modelling.

Formally, a Gibbs process Φ with finite number of points *almost surely*, is defined by specifying the density of its distribution, $\mathcal{P}$ say, with respect to the distribution of a Poisson process, $\mathcal{Q}$ say, $g = d\mathcal{P}/d\mathcal{Q}$. This construction is developed briefly in Section 5.5 of Stoyan et al. (1995), where it is shown how hard-core and clustered point processes can be obtained.

As an example consider the (unconditional) Strauss model (Strauss (1975)), which is used to model repulsion between points. Assuming that the state space is the unit square $S = [0,1]\times[0,1]$, the density g with respect to the distribution of the homogeneous Poisson process on S with intensity 1, is in the exponential family

$$g(\phi) \propto \exp\{\phi(S)\theta_1 + s(\phi)\theta_2\}$$

where

$$s(\phi) = \#\{\{x_i, x_j\} \subset \phi, i \neq j : \|x_i - x_j\| < r\}$$

is the number of unordered pairs of points having distance less than $r > 0$. It is necessary that $\theta_2 \leq 0$ for the density to be integrable, where the upper bound corresponds to the homogeneous Poisson process with intensity e^{θ_1}. For fixed $\phi(S)$ and $\theta_2 < 0$ the density is decreasing in $s(\phi)$ therefore the points tend to be repulsing from each other, actually the limit $\theta_2 \to -\infty$ corresponds to a hard-core process, where no points at distance less than r are allowed. The first column of Figure 5.3 demonstrates three different simulated data sets from the Strauss model with $r = 0.1,\ \theta_1 = \log(100)$ and $\theta_2 = \log(0.75),\ \log(0.1),\ \log(0.001)$ in top, middle and bottom correspondingly. Details on the method used to simulate these processes are provided in Section 5.1.5.

Gibbs processes will not be used directly to model data in this thesis. Nevertheless, they arise naturally as described below (see Chapter 6 for a specific application). Assume that in the hierarchical model shown in Figure 1.3 the distribution of the data Y conditionally on X has some Lebesgue density $\pi_{Y|X}(y \mid x)$. Suppose that X is modelled as a Poisson process with finite mean measure $\Lambda(\cdot;\Theta)$, where Θ is a vector of unknown parameters, and that $\pi_{X|\Theta}(x \mid \theta)$ is the density of its distribution with respect to some suitable reference measure, as described by Lemma 5.1.4. It follows that by construction the process X conditioned on Y and Θ is a Gibbs process. Its density with respect to to the Poisson reference measure is

$$\pi_{Y|X}(y \mid x)\pi_{X|\Theta}(x \mid \theta).$$

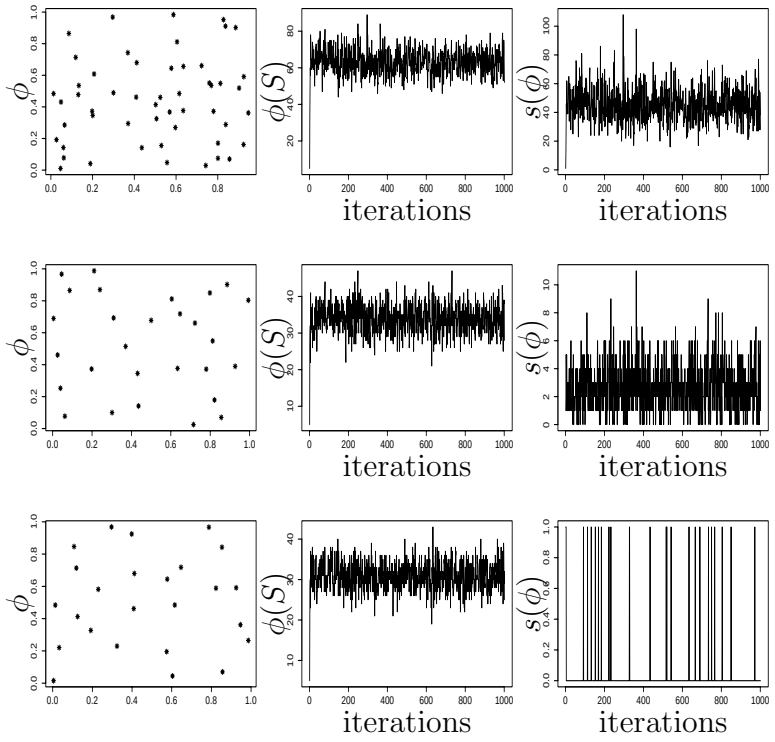


Figure 5.3: Simulation from the Strauss model with $r = 0.1$, $\theta_1 = \log(100)$ and for three different values of θ_2, $\log(0.75)$, $\log(0.1)$, $\log(0.001)$ in top, middle and bottom correspondingly. A configuration of the process is shown in the left column, MCMC samples from the stationary distribution of the sufficient statistics are plotted in the middle and right columns. The birth-and-death MCMC algorithm of Section 5.1.5 was used to obtain the samples from the point processes, run for 10^5 iterations but then thinned every 100 to produce the plots.

Availability of this density is necessary when performing a two-component Hastings-within-Gibbs algorithm to sample from the join posterior distribution of (X, Θ), as for example described in the application of Chapter 6.

MCMC simulation from the distribution of a Gibbs process is described in the following section.

5.1.5 Birth-death-displacement MCMC algorithms for simulating Gibbs processes

Simulating Gibbs point processes in a direct way is usually either impossible or extremely inefficient (see for example the discussion in Section 5.5 of Stoyan et al. (1995)). Instead, Markov chain methods have been proposed for this purpose that are very easy to implement and typically much more efficient than the direct rejection sampling type methods. An algorithm that has been known since at least Preston (1975), is based on running a spatial continuous time birth-death Markov process which converges towards the Gibbs process. Interest for similar algorithms has been recently revived in the statistical community, see for example Stephens (2000) and Cappé et al. (2003) for a recent review. In contrast, we will use the Metropolis-Hastings birth-and-death algorithm proposed by Geyer and Møller (1994). There are some advantages of this algorithm over the continuous time one, most importantly its simplicity in implementation, however a comparison of the two methodologies is far beyond the scope of this thesis and we simply refer to Cappé et al. (2003), Stoyan et al. (1995), Geyer and Møller (1994) and Stephens (2000). Moreover, all these approaches relate to the reversible jump MCMC methods, see Green (1995).

We aim at sampling from a Gibbs process on state space S which has density g with respect to a Poisson process reference distribution. The mean measure of the Poisson process is denoted by $\Lambda(\cdot)$ and it is assumed that $\Lambda(S) < \infty$. Geyer and Møller (1994) describe how to create a reversible Metropolis-Hastings Markov chain, with state space the exponential space Ω defined in Section 5.1.4, which converges to the distribution of the Gibbs process. The transition kernel is a mixture of two reversible kernels,

$$P(\phi, F) = w_d P_d(\phi, F) + (1 - w_d) P_{bd}(\phi, F), \ F \in \mathcal{F}, \phi \in \Omega$$

where $0 \leq w_d \leq 1$ is the probability of choosing kernel P_d. The latter represents a dimension-preserving move, that tries to displace one of the existing points of ϕ. P_{bd} represents a dimension-changing move that attempts to either kill one of the existing or give birth to a new point.

In particular, suppose that $\phi = \{x_1, \ldots, x_n\}$ with $g(\phi) > 0$. The displacement kernel chooses with equal probability one of the existing points, x_i say, and proposes to displace it

according to the measure $q(x_i, dy) = q(x_i, y)\Lambda(dy)$. The proposal distribution $Q(\phi, F)$ can be written as

$$Q(\phi, F) = \frac{1}{n}\sum_{i=1}^{n}\int_S \mathbb{1}[\phi - \{x_i\} \cup \{y\} \in F]q(x_i, y)\Lambda(dy)$$

This can be viewed as a kernel that updates $x \in \phi$ conditionally on $\phi - \{x\}$ and therefore the Metropolis-Hastings acceptance probability is given simply by

$$\alpha_d(\phi, \phi - \{x\} \cup \{y\}) = \min\left\{1, \frac{g(\phi - \{x\} \cup \{y\})q(y, x)}{g(\phi)q(x, y)}\right\} \tag{5.10}$$

whenever $n \geq 1$, $q(x, y)q(y, x) > 0$, $g(\phi - \{x\} \cup \{y\}) > 0$, and 0 otherwise.

The proposal kernel corresponding to P_{bd} is concentrated on $\Omega_{n-1} \cup \Omega_n \cup \Omega_{n+1}$ (or $\Omega_0 \cup \Omega_1$ when $n = 0$) defined in (5.6). With probability $p(\phi)$ we generate a new point $x \in S$ from some distribution $b(\phi, x)\Lambda(dx)$, and with the remaining probability we delete a randomly chosen point $y \in \phi$ with some probability $d(\phi - \{y\}, y)$ or if $n = 0$ we do nothing. Notice that we adopt the slightly unusual notation $d(\phi - \{y\}, y)$ following Geyer and Møller (1994), because it simplifies some of the formulae below. The second argument in $d(\cdot, \cdot)$ denotes the point which is proposed to be deleted, while the first argument is the closed set of points which is common in the point process before and after the proposal of the death. Thus for example, $d(\phi, y)$ is the probability of deleting y from $\phi \cup \{y\}$.

The transition probability for $F_n \subset \Omega_n,\ n = 0, 1, \ldots$ is

$$\begin{aligned}
P(\phi, F_{n+1}) &= p(\phi)\int_S \mathbb{1}[\phi \cup \{x\} \in F_{n+1}]\alpha_{bd}(\phi, \phi \cup \{x\})b(\phi, x)\Lambda(dx), n \geq 0 \\
P(\phi, F_{n-1}) &= (1 - p(\phi))\sum_{x\in\phi} \mathbb{1}[\phi - \{x\} \in F_{n-1}]\alpha_{bd}(\phi, \phi - \{x\})d(\phi - \{x\}, x), n \geq 1 \\
P(\phi, F_n) &= \mathbb{1}[\phi \in F_n]\left\{p(\phi)\int_S (1 - \alpha_{bd}(\phi, \phi \cup \{x\})b(\phi, x)\Lambda(dx)\right. \\
&+ \left.(1 - p(\phi))\sum_{x\in\phi}(1 - \alpha_{bd}(\phi, \phi - \{x\}))d(\phi - \{x\}, x)\right\}, n \geq 0
\end{aligned}$$

and reversibility holds if and only if

$$\begin{aligned}
&\int \mathbb{1}[\phi \in F_n, \phi \cup \{x\}F_{n+1}]g(\phi \cup \{x\})\alpha_{bd}(\phi \cup \{x\}, \phi)d(\phi, x)\Lambda(dx_1)\ldots\Lambda(dx_n)\Lambda(dx) \\
&\quad = \int \mathbb{1}[\phi \in F_n, \phi \cup \{x\}F_{n+1}]g(\phi)\alpha_{bd}(\phi, \phi \cup \{x\})b(\phi, x)\Lambda(dx_1)\ldots\Lambda(dx_n)\Lambda(dx)
\end{aligned}$$

for all measurable $F_n \subset \Omega_n,\ F_{n+1} \subset \Omega_{n+1},\ n \geq 0$. When $n = 0$, $\Lambda(dx_1)\ldots\Lambda(dx_n)$ is replaced

by $\mathbb{1}[\phi = \{\emptyset\}]$. This implies that the Metropolis-Hastings acceptance probability is

$$\alpha_{bd}(\phi, \phi \cup \{x\}) = \min\{1, r(\phi, x)\}$$

if $g(\phi \cup \{x\}) > 0$, and 0 otherwise, and

$$\alpha_{bd}(\phi \cup \{x\}, \phi) = \min\{1, 1/r(\phi, x)\}$$

if $g(\phi \cup \{x\}) > 0$ and 0 otherwise, where

$$r(\phi, x) = \frac{g(\phi \cup \{x\})}{g(\phi)} \frac{1 - p(\phi \cup \{x\})}{p(\phi)} \frac{d(\phi, x)}{b(\phi, x)}. \tag{5.11}$$

Convergence of the birth-death-displacement algorithm is established in Section 4 of Geyer and Møller (1994).

As an illustration of the algorithm, we simulated from the Strauss model described in the previous section and summarise the results in Figure 5.3. In our simulations we chose $p(\phi) = 0.5$, $b(\phi, x) \propto 1$, $d(\phi, x) = 1/n$, when $\phi(S) = n$, and $w_d = 0$, therefore no displacement moves were attempted, following the advice given in Section 5 of Geyer and Møller (1994). The birth-death-displacement algorithm will be used as part of a larger MCMC algorithm in Chapter 6.

5.2 NCPs for Poisson processes

Often we are interested in making inference about a hierarchical model as in Figure 1.3, where X is a Poisson process on a state space S with mean measure Λ depending on some parameters Θ. We will encounter such hierarchical models in Chapter 6, in the context of non-Gaussian stochastic volatility models. Moreover, such hierarchical models arise in Bayesian non-parametric modelling, see for example Wolpert and Ickstadt (1998) in the context of modelling spatial variation. In our applications S is typically some subset of $\mathbb{R}^d$, $d = 1, 2, \ldots$. The distribution of the data Y given X is defined through some Lebesgue density $\pi_{Y|X}(y \mid x)$, but we will not make any assumptions about its form here. Furthermore, we will take for granted that the mean measure Λ admits a Lebesgue density $\lambda(x)$, $x \in S$, which will occasionally be denoted as $\lambda(x; \Theta)$ to stress its functional dependence on the parameters Θ.

We intend to use the two-component Hastings-within-Gibbs sampler as described in Section 4.1 to obtain samples from the joint posterior distribution of (X, Θ). In many applications, due to ergodicity, the augmented information about Θ can be much larger than the marginal information. An extreme example was given in Section 5.1.4, where actually the information contained in X about Θ is infinite, since the latter is obtained as

a strong law limit from the former. In such cases, of very high prior dependence between X and Θ, unless the data are very informative about X, we expect (see Chapter 6) the centered algorithm to have poor convergence properties. We can't be very precise in our statements here without making specific assumptions about the distribution of Y given X.

Therefore, there might be potential benefit from using an NCP for models with hidden Poisson processes. According to the general presentation of Section 4.1 these parameterisations work with $(\tilde{X}, \Theta)$ where $\tilde{X}$ is *a priori* independent of Θ, and X is obtained from them by means of some (not necessarily invertible) function. The MCMC algorithm that samples from the joint posterior of $(\tilde{X}, \Theta)$ is described in Section 4.1, where its relationship with the centered algorithm is shown. Typically, there is a collection of NCPs for a particular pair (X, Θ) which differ in the prior of $\tilde{X}$ and the way $(\tilde{X}, \Theta)$ is transformed to X.

The following sections develop two NCPs, which have quite different properties and applicability. Both have their grounds on methods for simulating complex Poisson processes from simpler ones. NCP by thinning (referred to by THIN-NCP) is motivated by the independent thinning operation on point processes and it can be applied for any $S \subset \mathbb{R}^d,\ d = 1, 2, \ldots$. NCP by the inverse CDF method (referred to by CDF-NCP) is usually employed when S is a subset of the real line, and has its grounds on a method used to simulate Poisson processes in time by generating their independent increments.

5.3 NCP by thinning

Assume for the moment that λ is bounded by one on S, $\lambda(x) < 1$ for all $x \in S$. A Poisson process X on S with intensity function λ can be obtained from another Poisson process on S with intensity one, $\tilde{X}$ say, by thinning, that is by independently deleting each point of the latter, $x \in \tilde{X}$ with probability $1 - \lambda(x)$. It can easily be shown that the process resulting from this rejection sampling procedure is indeed a Poisson process with intensity function λ, since

$$\mathsf{E}[X(B)] = \int_S \lambda(x)dx = \Lambda(B), \text{ for all } B \subset S.$$

This technique is discussed, for example, in Section 6.1 of Devroye (1986), Section 4.2 of Ripley (1987) and Section 5.1 of Stoyan et al. (1995). Thinning can be seen as a random mapping from a Poisson process $\tilde{X}$ on S to another Poisson process $X \subset \tilde{X}$ on S. However, it can also be interpreted as a deterministic mapping from Poisson processes on the expanded state space $S \times (0, 1)$ to Poisson processes on S. Namely, a Poisson process X with intensity function $\lambda(x) < 1,\ x \in S$ can be obtained from a Poisson process $\tilde{X}$ on $S \times (0, 1)$ as

$$X = \{X_j : (X_j, Z_j) \in \tilde{X} \text{ and } Z_j < \lambda(X_j)\}; \tag{5.12}$$

see Figure 5.4 for an illustration. It is an immediate consequence of the restriction property of Poisson processes, described in Section 5.1.1, and the mapping Theorem 5.1.1 that this selection procedure produces a Poisson process on S with intensity function λ. Actually, under the above interpretation there is no need to bound λ. Therefore, an NCP for a Poisson process X on S with intensity function λ can be constructed by taking $\tilde{X}$ to be a unit intensity Poisson process on $S \times (0, \infty)$ and the deterministic but not invertible transformation from $\tilde{X}$ to X to be (5.12). This was originally proposed by Roberts et al. (2003)

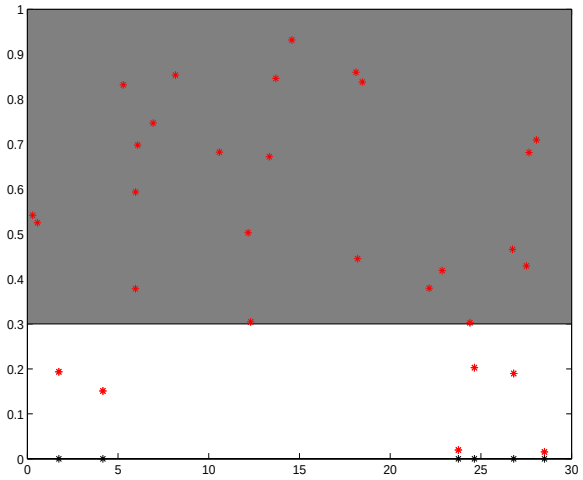


Figure 5.4: Transformation from $\tilde{X}$ to X. X is a Poisson process on S with intensity $\lambda(x;\theta)$, $\tilde{X}$ is a unit intensity Poisson process on $S \times (0, \infty)$. Choose all points of $\tilde{X}$ that lie on $\{(x, z) : x \in S,\ y < \lambda(x;\theta)\}$ (white area in the plot) and project them down to S. The resulting points (denoted by the black asterisks) form X. In this example $S = [0, 30],\ \lambda(x;\theta) = \theta,\ \theta = 0.3$.

and is reviewed in Chapter 6, in the context of non-Gaussian stochastic volatility models. In the terminology of Chapter 4, this is a state space expanded NCP, since $\tilde{X}$ lives on a higher dimensional space than X and the transformation $(\Theta, \tilde{X}) \to X$ is not invertible.

It is possible to construct NCPs by thinning where $\tilde{X}$ is not a homogeneous Poisson process on $S \times (0, \infty)$. This is interesting, for example, when S is a product space. Of course, (5.12) is not the appropriate transformation in this case. We will consider such parameterisations in Section 5.5 and discuss how they compare with the one suggested in this section.

In Section 4.2 we showed that the expanded state space NCAs were just as easy to implement as the corresponding CAs. This was attributed to the independent increments property of $\tilde{X}$ and the fact that the data depended only on some values of $\tilde{X}$ rather than on the whole process. The situation is very similar here, where $\tilde{X}$ is a Poisson process. The points of $\tilde{X}$ on disjoint sets form independent processes while the data Y is independent of the points of $\tilde{X}$ on $\{(x, z) : z > \lambda(x),\ x \in S\}$ (lying on the grey area in Figure 5.4) conditionally on its points on $\{(x, z) : z < \lambda(x),\ x \in S\}$ (lying on the white area).

Assume that a CA as described in Section 4.1 already exists for sampling from the joint posterior distribution of (Θ, X). Typically, the birth-death-displacement algorithm of

Section 5.1.5 is used to update X given Y and Θ, for example this is the strategy we adopt in Chapter 6. In Section 4.1 we gave the two transformation steps necessary to implement a state space expanded NCP. Step 2 transforms $(\Theta, \tilde{X}) \to X$ and it is deterministic and for the THIN-NCA this transformation is given in (5.12). Step 4 is a stochastic transformation, $(\Theta, X) \to \tilde{X}$, therefore we need to 'rebuild' $\tilde{X}$ given Θ, Y and X, which is a projection of a subset of $\tilde{X}$ on S. Section 4.1 suggested that this step can be incorporated into Step 1, which updates Θ given $\tilde{X}$ and Y. Implementation of Steps 1-3 of the THIN-NCA is not complicated by the form of the intensity function λ or the state space S. Nevertheless, the feasibility of the random mapping $(\Theta, X) \to \tilde{X}$, either when explicitly performed at Step 4 or when it is implicitly done at Step 1, depends crucially on the functional form of λ.

The following section describes the transformation $(\Theta, X) \to \tilde{X}$ and how it can be incorporated into Step 1 for the special case where X is a homogeneous Poisson process. Section 5.3.2 generalises to locally finite non-homogeneous Poisson processes, for which the intensity function satisfies the integrability condition

$$\int_C \lambda(z)dz < \infty$$

on any bounded set $C \subset S$. Implementation of the THIN-NCA for processes with non-σ-finite mean measures is briefly considered in Section 5.8.1.

5.3.1 THIN-NCA for homogeneous Poisson processes on a bounded state space

Let X be a homogeneous Poisson process with intensity θ on a bounded state space $S \subset \mathbb{R}^d$ for some $d = 1, 2, \ldots$. We initially describe the transformation $(\Theta, X) \to \tilde{X}$ and then show how to merge this step into the step which updates Θ given $\tilde{X}$ and Y.

Stochastic transformation $(\Theta, X) \to \tilde{X}$

We define

$$\begin{aligned} \tilde{X}_{(0,\Theta)} &= \tilde{X} \cap (S \times [0, \Theta)) \\ \tilde{X}_{[\Theta,\infty)} &= \tilde{X} - \tilde{X}_{(0,\Theta)} \end{aligned}$$

so that $\tilde{X}$ is the disjoint union

$$\tilde{X} = \tilde{X}_{(0,\Theta)} \cup \tilde{X}_{[\Theta,\infty)}$$

and recall from (5.12) that

$$X = \{X_j : (X_j, Z_j) \in \tilde{X}_{(0,\Theta)}\}.$$

$\tilde{X}_{(0,\Theta)}$ and $\tilde{X}_{[\Theta,\infty)}$ (which correspond to the white and grey areas in Figure 5.4 respectively) are independent conditionally on Θ, due to the restriction property of Poisson processes (Section 5.1.1). By construction, the observed data Y depends only on X as we described in Section 5.2. Therefore, conditionally on Θ, $\tilde{X}_{[\Theta,\infty)}$ is independent of Y. As a result, it is also independent of $\tilde{X}_{(0,\Theta)}$ and can, in principle, be simulated from its prior.

We write

$$\tilde{X}_{(0,\Theta)} = \{(X_j, Z_j), j = 1, \ldots, N, X_j \in S, 0 \leq Z_j \leq \Theta\}$$

for some random integer $N > 0$, otherwise $\tilde{X}_{(0,\Theta)} = \{\emptyset\}$. The Z_js are independent of the data Y, since the likelihood depends only on the projection X on S. Therefore, conditionally on N, the Z_js are independent of the X_js and independent of each other, distributed as $\mathrm{Un}[0, \Theta]$. Hence, $\tilde{X}_{(0,\Theta)}$ is obtained from a configuration of $X = \{x_1, \ldots, x_n\}$ simply as

$$\tilde{X}_{(0,\Theta)} = \{(x_j, Z_j) : x_j \in X, Z_j \sim \mathrm{Un}[0, \Theta]\}$$

(or $\tilde{X}_{(0,\Theta)} = \{\emptyset\}$ if $n = 0$).

Summarising, we transform (Θ, X) to $\tilde{X}$, where $X = \{x_1, \ldots, x_n\}$ for some $n \geq 0$, as follows:

Stochastic transformation $(\Theta, X) \to \tilde{X}$

`Simulate` $\tilde{X}_{[\Theta,\infty)}$ `as a Poisson process with unit intensity on` $S \times [\Theta, \infty)$

`Simulate independent variables` $Z_j \sim \mathtt{Un}[0, \Theta], \ \ j = 1, \ldots, n$

`Set` $\tilde{X}_{(0,\Theta)} = \{(x_j, Z_j) : x_j \in X, Z_j \sim \mathtt{Un}[0, \Theta]\}$

`Set` $\tilde{X} = \tilde{X}_{(0,\Theta)} \cup \tilde{X}_{[\Theta,\infty)}$

Update Θ conditionally on $\tilde{X}$ and Y

Assume for the moment that $\tilde{X}$ is available at Step 1 of the NCA given in Section 4.1. Typically, a Metropolis-Hastings step will be used to update Θ given $\tilde{X}$ and Y. Suppose that θ_0 is the current value of Θ and θ_1 has been generated from some proposal kernel

$q(\theta_0, \theta_1)$. The Metropolis-Hastings acceptance ratio is

$$r = \frac{\pi(Y \mid X^{(1)})\pi(\theta_1)q(\theta_1, \theta_0)}{\pi(Y \mid X^{(0)})\pi(\theta_0)q(\theta_0, \theta_1)} \tag{5.13}$$

where $X^{(0)}$, $X^{(1)}$ are defined through (5.12) with $\lambda(x) = \theta_0$ and $\lambda(x) = \theta_1$ for all $x \in S$ respectively, see also Figure 5.5. The transition from θ_0 to θ_1 is accepted with probability $\min\{1, r\}$. Thus, if we had $\tilde{X}$ then we could easily perform Step 1. However, the previous section shows that this is not necessary and we only need to have X at this step. If we propose $\theta_1 < \theta_0$ then $X^{(1)}$ can be obtained as a random thinning of $X^{(0)}$, where its point of the latter is killed with probability $1 - \theta_1/\theta_0$. If $\theta_1 > \theta_0$ then $X^{(1)}$ is the superposition of $X^{(0)}$ and an independent Poisson process on S with intensity $\theta_1 - \theta_0$.

The next section generalises this algorithm for arbitrary locally finite Poisson processes.

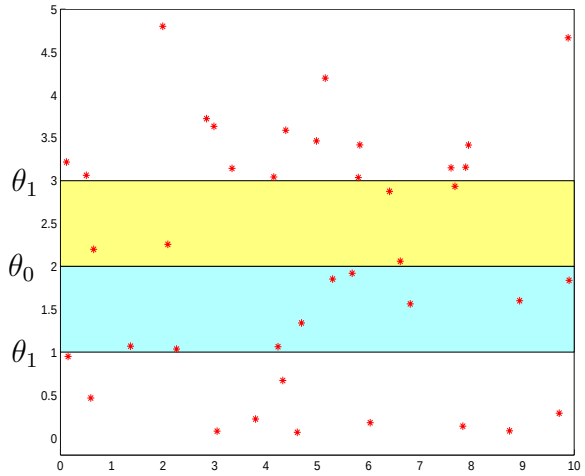


Figure 5.5: Updating of Θ given $\tilde{X}$ and Y in the THIN-NCA. The current value of θ is denoted by θ_0 and the proposed by θ_1, the corresponding quantities for X by $X^{(0)}$ and $X^{(1)}$ respectively. If $\theta_1 < \theta_0$, $X^{(1)}$ is found by removing from $X^{(0)}$ all points $\{x \in X^{(0)} : (x, z) \in \tilde{X} \cap S \times [\theta_1, \theta_0)\}$ (lying in the cyan area). If $\theta_1 > \theta_0$, $X^{(1)}$ is derived from $X^{(0)}$ by the addition of all points $\{x \in S : (x, z) \in \tilde{X} \cap S \times [\theta_0, \theta_1)\}$ (lying in the yellow area). Once $X^{(1)}$ has been found the Metropolis-Hastings acceptance probability (5.13) can be calculated. In this example we have taken $S = [0, 10]$, $\theta_0 = 2$, $\theta_1 = 1$ and $\theta_1 = 3$.

5.3.2 THIN-NCA for finite Poisson processes

The ideas of the previous section extend naturally to arbitrary Poisson processes. The main assumption here is that $\Lambda(S) < \infty$, apart from which the intensity function can be totally arbitrary.

Let θ_0, θ_1, $X^{(0)}$ and $X^{(1)}$ be defined as in the previous section. We also define

$$S_0 = \{x \in S : \lambda(x; \theta_0) < \lambda(x; \theta_1)\}$$

and $S_1 = S - S_0$. If $\lambda(x;\Theta)$ is not monotonic in Θ for all $x \in S$, determination of S_0 might be a tedious task. $X^{(1)}$ is derived from $X^{(0)}$ by the following steps:

Derivation of $X^{(1)}$ from $X^{(0)}$

Kill each point $X_j \in X^{(0)} \cap S_1$ with probability $1 - \lambda(X_j;\theta_1)/\lambda(X_j;\theta_0)$

Calculate $\mu = \int_{S_0} \{\lambda(z;\theta_1) - \lambda(z;\theta_0)\} dz$, or set $\mu = 0$ if $S_0 = \emptyset$

Simulate $N \sim \mathrm{Pn}(\mu)$

Add to $X^{(0)}$ N new points generated independently from the density proportional to $\lambda(x;\theta_1) - \lambda(x;\theta_0)$, $x \in S_0$

The second and fourth step in the above algorithm might be difficult when $\lambda(x;\theta)$ is totally arbitrary. However, notice that THIN-NCA can be relatively easily implemented for a class of intensity functions frequently encountered in applications

$$\lambda(x;\Theta) = \Theta q(x)$$

where $q(x)$ is a density function on S. Then, $S_0 = S$ when $\theta_0 < \theta_1$ otherwise $S_0 = \emptyset$, $\mu = \max\{0, \theta_1 - \theta_0\}$ and the new points are generated from q.

5.4 NCP by the inverse CDF method

This reparameterisation can be employed whenever the state space S is a subset of the real line, although it can be extended when we consider marked Poisson processes. This section assumes that $S = (0, \infty)$. It is based on the inverse CDF method for simulating random variables (see for example Section 3.2 of Ripley (1987)) and suffers from its limitations, in particular the necessity to invert functions and the difficulty to extend to high dimensions. Nevertheless, it can be very useful especially when the intensity function satisfies (5.18), as we shall see in Section 5.8.1.

Let $X = \{X_1, X_2, \ldots\}$ be a Poisson process in time where $0 < X_1 < X_2 < \cdots$, with intensity function λ, for which we will temporarily assume that

$$\int_0^x \lambda(z) dz < \infty$$

for all $x > 0$. This intensity corresponds to locally finite Poisson processes, however extensions will be considered later in this section. We also define $\tilde{X} = \{\tilde{X}_1, \tilde{X}_2, \ldots\}$ to be a homogeneous Poisson process in time with intensity one and $0 < \tilde{X}_1 < \tilde{X}_2 < \cdots$. The increments of this process $\tilde{X}_{j+1} - \tilde{X}_j,\ j = 1, 2, \ldots$ (where $\tilde{X}_0 := 0$) are independent and distributed as Ex(1). Some authors define the homogeneous Poisson process on a subset of the real line as a continuous-time Markov chain with exponentially distributed inter-arrival times, see for example Section 2.4 of Norris (1997). For a proof of this property, when the definition of a Poisson process is as in Section 5.1, see for example Section 4.1 of Kingman (1993).

Let Y_1 be the time until the first arrival of X, i.e $Y_1 := X_1$. Then

$$P[Y_1 > x] = P[X([0, x)) = 0] = \exp\left\{-\int_0^x \lambda(z)dz\right\}$$

therefore Y_1 can be simulated by the inverse CDF method as the solution with respect to x of the following equation

$$-\log(1 - U) = \int_0^x \lambda(z)dz,\ U \sim \mathrm{Un}[0, 1]$$

which is equivalent to

$$\tilde{X}_1 = \int_0^x \lambda(z)dz.$$

Let X_1 be the solution to this equation and therefore the first point of X. Conditional on this value, the time until the next arrival $Y_2 := X_2 - X_1$ can be simulated by solving

$$\tilde{X}_2 - \tilde{X}_1 = \int_{X_1}^{X_1+x} \lambda(z)dz.$$

Recalling that $\tilde{X}_1 = \int_0^{X_1} \lambda(z)dz$, X_2 is obtained as the solution to

$$\tilde{X}_2 = \int_0^x \lambda(z)dz.$$

By a recursive application of the above argument X_n is obtained as the solution to

$$\tilde{X}_n = \int_0^x \lambda(z)dz \tag{5.14}$$

and generally

$$X = \{h(x) : x \in \tilde{X}\} \tag{5.15}$$

where h is the increasing function determined by

$$h^{-1}(x) = \int_0^x \lambda(z)dz. \tag{5.16}$$

If

$$\int_0^\infty \lambda(z)dz < \infty$$

then there exists some finite $n > 0$ *almost surely* such that (5.14) cannot be solved for all x. In that case X_{n-1} is the last point of the Poisson process (under the usual time-ordering) which is finite *almost surely*, as it would have been expected by the integrability condition satisfied by its intensity function. By convention, in (5.15) we map all points x such that $h(x) > \int_0^\infty \lambda(z)dz$ to 0.

The relationship between X and $\tilde{X}$ described by (5.15) and (5.16) can also be deduced using an argument based on the mapping Theorem 5.1.1. In particular, suppose that we want to find the increasing, measurable and differentiable h such that X derived from $\tilde{X}$ by (5.15) has intensity function λ. For $A \subset \mathbb{R}$

$$\begin{aligned}
\int_A \lambda(z)dz &= \mathsf{E}[X(A)] = \mathsf{E}[\tilde{X}(h^{-1}(A))] \\
&= \int_{h^{-1}(A)} dz \\
&= \int_A \frac{dh^{-1}(z)}{dz} dz
\end{aligned}$$

therefore h solves the differential equation

$$\lambda(x) = \frac{dh^{-1}(x)}{dx}. \tag{5.17}$$

Under the initial condition $h^{-1}(0) = 0$, (5.17) is solved by the function defined in (5.16).

We will now show how CDF-NCP works when

$$\begin{aligned}
\int_0^x \lambda(z)dz &= \infty \\
\int_x^\infty \lambda(z)dz &< \infty
\end{aligned} \tag{5.18}$$

for all $x > 0$. In this case, we start by simulating the latest arrival X_1 in X, which will finite *almost surely* by (5.18), and write $X = \{X_1, X_2, \ldots\}$ where now $\infty > X_1 > X_2 > \cdots > 0$.

$$P[X_1 < x] = P[X([x, \infty)) = 0] = \exp\left\{-\int_x^\infty \lambda(z)dz\right\}$$

therefore X_1 can be simulated by the inverse CDF method as the solution with respect to x

of the following equation

$$\tilde{X}_1 = \int_x^\infty \lambda(z)dz.$$

The same arguments which lead to (5.14) yield that X_n is obtained as the solution to

$$\tilde{X}_n = \int_x^\infty \lambda(z)dz. \tag{5.19}$$

Notice that there are infinite points of X in any neighbourhood of 0. It follows that X is obtained by $\tilde{X}$ via (5.15), but where now h is the decreasing function determined by

$$h^{-1}(x) = \int_x^\infty \lambda(z)dz. \tag{5.20}$$

We can arrive at the same result using the mapping theorem, as described earlier.

When $\int_0^\infty \lambda(z)dz < \infty$, h can be given either by (5.16) or by (5.20). If the latter is chosen then *almost surely* there will be a point $\tilde{X}_n > \int_0^\infty \lambda(z)dz$ and (5.19) will be impossible to solve. In this case we follow the convention to map all x in (5.20) such that $x > \int_0^\infty \lambda(z)dz$ to 0.

As an illustration, consider the intensity function

$$\lambda(x) = r\phi \exp\{-\phi x\},\ x > 0. \tag{5.21}$$

This is the intensity of a Poisson process which is obtained as a projection of a Poisson process which we will revisit in Section 5.5.1 and Section 6.5. Since $\int_0^\infty \lambda(z)dz < \infty$ either (5.16) or (5.20) can be used. The solution to (5.14) is given by

$$X_n = -\frac{1}{\phi}\log(1 - \tilde{X}_n/r),\ \tilde{X}_n < r$$

and the solution to (5.19) by

$$X_n = -\frac{1}{\phi}\log(\tilde{X}_n/r),\ \tilde{X}_n < r.$$

The solutions as functions of $\tilde{X}$ are superimposed in Figure 5.6.

If $\lambda(x)$ depends on some parameters Θ, a non-centered reparameterisation of (X, Θ) can be constructed by taking $\tilde{X}$ to be a unit intensity Poisson process and X is derived from $\tilde{X}$ and Θ by (5.15) and (5.16) or (5.20). This is the CDF-NCP, which can be directly implemented only for Poisson processes on subsets of $\mathbb{R}$, since it explicitly uses the time ordering of the Poisson process points, although extensions for marked Poisson processes are possible as in Section 5.5.

CDF-NCP can be applied whenever (5.16) or (5.20) is easily invertible, which, however,

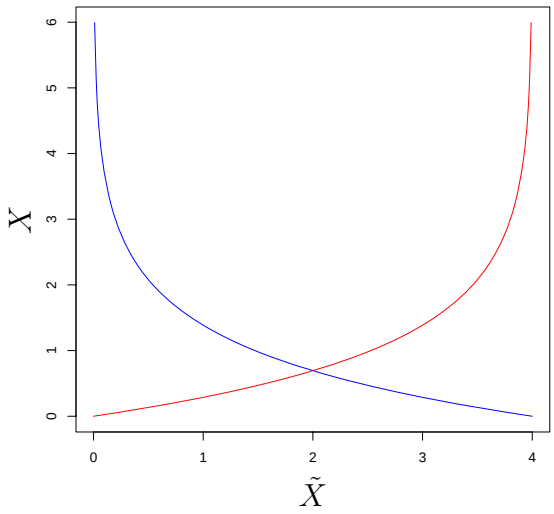


Figure 5.6: The increasing (red) and decreasing (blue) transformation h in the CDF-NCP corresponding to the intensity function (5.21). In this plot we have taken $r = 4, \phi = 1$.

can only be achieved numerically in many applications. Implementation of the CDF-NCA for locally finite processes on bounded intervals $S = [0, T],\ T > 0$ is on the same lines as of the THIN-NCA, therefore we will omit any details and refer back to Section 5.3. For example, when $\lambda(x) = \Theta$ for all $x \in S$, only the finite number of points of $\tilde{X}$ that lie on $[0, T\theta]$ need to be monitored during the iterations of the algorithm.

Section 5.8.1 reveals the connection of this method to the Ferguson and Klass (1972) representation for positive Lévy processes. There it is argued that it might be preferable to the THIN-NCP when the intensity function of the latent Poisson process satisfies (5.18).

5.5 NCPs for marked Poisson processes

Here we discuss how the ideas for non-centering of Poisson processes can be applied when the state space has a product space structure and there exists a representation of the process as marked Poisson. Rather than pursuing the greatest possible generality, we prefer to focus on a class of processes which are very relevant to our applications in Chapter 6. Nevertheless, at the end of this section it should be clear how to extend the ideas to more general marked Poisson processes. Section 5.5.1 illustrates the methods developed here through a specific example.

We take $S = [0, T] \times (0, \infty)$, where $0 < T < \infty$ and assume that the intensity function factorises as

$$\lambda(c, \epsilon) = r(c)q(\epsilon),\ \ (c, \epsilon) \in S \tag{5.22}$$

where r is integrable. In many applications r will be a constant function but there are interesting examples where q will not be integrable, as for example for the point process representation of subordinators in Section 5.8. A Poisson process $X = \{(C_i, E_i),\ i = 1, 2, \ldots\}$ with mean measure (5.22) admits a marked Poisson process representation: it is obtained by marking the Poisson process $\{E_i, i = 1, 2, \ldots\}$ on $(0, \infty)$ with intensity function $q(\epsilon) \int_0^T r(c)dc$, with marks $C_i \sim r$ (up to proportionality). If further q is integrable we can obtain X by marking the Poisson process $\{C_i, i = 1, 2, \ldots\}$ with intensity function $r(c) \int_0^\infty q(\epsilon)d\epsilon$, with marks $E_i \sim q$ (up to proportionality).

Suppose that the parameter vector is decomposed as $\Theta = (\Theta_1, \Theta_2)$, r is a function of Θ_1 only and q a function of Θ_2 only. We describe three different NCPs below. When both r and q are integrable they can all be used (see Section 5.5.1 for an example) but when q is not integrable the second will be much more suitable.

We term the first parameterisation the MPP-THIN-NCP, which as described below presupposes that q is integrable. We non-center the marginal process $\{C_i, i = 1, 2, \ldots\}$ using the THIN-NCP of Section 5.3 and we find a non-centered transformation $(\Theta_2, E_i) \to (\Theta_2, \tilde{E}_i)$ as for example described in Chapter 4. Then the MPP-THIN-NCP takes

$$\tilde{X} = \{(C_i, M_i, \tilde{E}_i), i = 1, 2, \ldots\} \tag{5.23}$$

where the process $\{(C_i, M_i), i = 1, 2, \ldots\}$ is a unit intensity Poisson process on $[0, T] \times (0, \infty)$. The transformation of $(\Theta_1, \Theta_2, \tilde{X}) \to X$ is done by first constructing the process $\{C_i, i = 1, 2, \ldots\}$ from $\{(C_i, M_i), i = 1, 2, \ldots\}$ as described in Section 5.3 (see in particular the transformation in (5.12)) and then for each $i = 1, 2, \ldots$ transforming $(\Theta_2, \tilde{E}_i) \to E_i$, as described in Chapter 4.

A similar NCP, which however can be constructed even when q is not integrable, interchanges the roles of C_i and E_i in the above construction. That is, we non-center the Poisson process $\{E_i, i = 1, 2, \ldots\}$ with mean measure $\int_0^T r(c)dcq(\epsilon)$ using the unit rate Poisson process $\{(E_i, M_i), i = 1, 2, \ldots\}$. Moreover, suppose that $(\Theta_1, C_i) \to (\Theta_2, \tilde{C}_i)$ is a non-centered transformation of the random variable C_i with density proportional to $r(\cdot)$. Then

$$\tilde{X} = \{(E_i, M_i, \tilde{C}_i), i = 1, 2, \ldots\}$$

and the transformation $(\Theta_1, \Theta_2, \tilde{X}) \to X$ is performed by adjusting appropriately the procedure described in the previous paragraph for the case where q is integrable. When $r(\cdot)$ is a constant function the scheme discussed above coincides with the THIN-NCP which is described later in this section.

The MPP-CDF-NCP which we construct below is based on the CDF-NCP proposed in

Section 5.4. We take

$$\tilde{X} = \{(\tilde{C}_i, \tilde{E}_i), i = 1, 2, \ldots\} \tag{5.24}$$

where $\{\tilde{E}_i, i = 1, 2, \ldots\}$ is a unit intensity Poisson process on $(0, \infty)$ and where $(\Theta_1, \tilde{C}_i) \to C_i$ is a non-centered transformation of the random variable C_i with density proportional to $r(\cdot)$. Transformation of $(\Theta_1, \Theta_2, \tilde{X}) \to X$ is done by first constructing the process $\{E_i, i = 1, 2, \ldots\}$, as described in Section 5.4 (using either the increasing or the decreasing transformation), and then for each $i = 1, 2, \ldots$ transforming $(\Theta_1, \tilde{C}_i) \to C_i$, as shown in Chapter 4. When q is integrable we can reverse the roles of C and E.

A third NCP can be considered for a marked Poisson process, one which simply ignores the product space structure of S and takes

$$\tilde{X} = \{(C_i, E_i, M_i), i = 1, 2, \ldots\} \tag{5.25}$$

to be a unit intensity Poisson process on $S \times (0, \infty)$. The transformation $\tilde{X} \to X$ is achieved via (5.12), that is by

$$X = \{(C_i, E_i) : (C_i, E_i, M_i) \in \tilde{X} \text{ and } M_i < \lambda(C_i, E_i)\}\}.$$

In accordance to Section 5.3 we term this the THIN-NCP. In order to pinpoint the different construction between the MPP-THIN-NCP and the THIN-NCP we have used a different notation for the points of the corresponding $\tilde{X}$ processes: a typical point of the $\tilde{X}$ process is denoted by (C_i, E_i, M_i) for the THIN-NCP and by $(C_i, M_i, \tilde{E}_i)$ for the MPP-THIN-NCP. Firstly, this highlights that the latter non-centers X by first non-centering the marginal $\{C_i, i = 1, 2, \ldots\}$ and then the conditional distribution of the E_is given the C_is. Instead, the THIN-NCP non-centers the process as a whole. Secondly, the notation indicates that after projection on S, the points resulting from the MPP-THIN-NCP need a further transformation, that of $\tilde{E}_i \to E_i$, while this is unnecessary for the THIN-NCP.

5.5.1 An illustrative example

In order to illustrate the techniques we have developed so far we consider a specific example. Let X be a Poisson process on $S = [0, T] \times (0, \infty)$ with intensity function

$$\lambda(c, \epsilon; \Theta) = r\phi \exp\{-\phi\epsilon\}, \ (c, \epsilon) \in S, \Theta = (r, \phi), \tag{5.26}$$

We will encounter a latent process with such mean measure in Chapter 6, specifically in Section 6.5, and we will show that it is related with the shot-noise continuous time Markov process. Recognise that the intensity function in (5.26) is of the form given in (5.22), where

$r(c) = r$ and $q(\epsilon) = \phi \exp\{-\phi\epsilon\}$ which integrates to 1.

There are (at least) three interesting NCPs which can be constructed for this process. The first is the MPP-THIN-NCP suggested in Section 5.5 which takes $\tilde{X}$ to be a Poisson process on $[0, T] \times (0, \infty) \times (0, \infty)$ with mean measure

$$e^{-\tilde{\epsilon}} dc\ dm\ d\tilde{\epsilon}. \tag{5.27}$$

and X is retrieved from $\tilde{X} = \{(C_i, M_i, \tilde{E}_i), i = 1, 2, \ldots\}$ and Θ as follows (see also Figure 5.7).

MPP-THIN-NCP when $\lambda(c, \epsilon; \Theta) = r\phi \exp\{-\phi\epsilon\}$

Let $\tilde{X} = \{(C_i, M_i, \tilde{E}_i), i = 1, 2, \ldots\}$ as in (5.23).

Select all points from $\tilde{X}$ for which $M_i < r$.

Project these points to $[0, T] \times (0, \infty)$.

Transform $\{(C_i, \tilde{E}_i)\}$ to $\{(C_i, E_i)\}$ where $E_i = \tilde{E}_i/\phi$.

X consists of the transformed points.

We can also apply the MPP-CDF-NCP proposed in Section 5.5. We take $\tilde{X} = \{(C_i, \tilde{E}_i), i = 1, 2, \ldots\}$ to be a unit rate Poisson process on S and transform $\tilde{X} \to X$ as follows (see also Figure 5.8).

MPP-CDF-NCP when $\lambda(c, \epsilon; \Theta) = r\phi \exp\{-\phi\epsilon\}$

Let $\tilde{X} = \{(C_i, \tilde{E}_i), i = 1, 2, \ldots\}$ as in (5.24).

Select all points $(C_i, \tilde{E}_i) \in \tilde{X}$ for which $\tilde{E}_i < r$.

Set $E_i = -\log\{\tilde{E}_i/r\}/\phi$.

$X = \{(C_i, E_i), i = 1, 2, \ldots\}$.

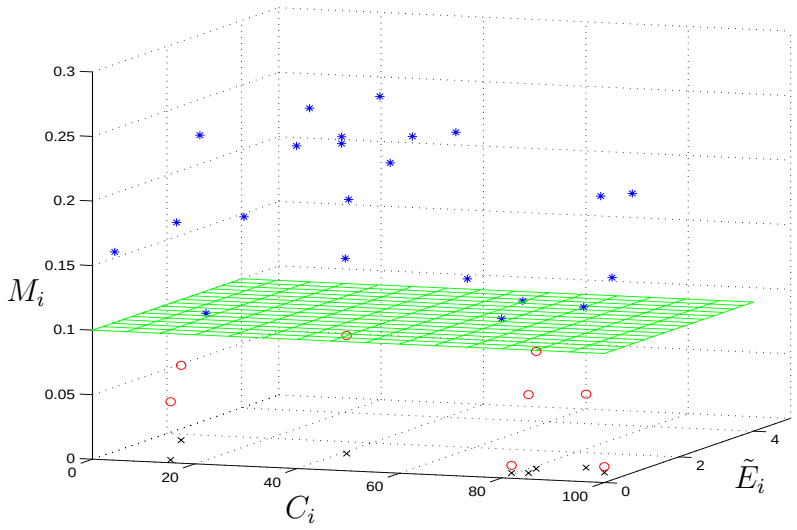


Figure 5.7: The MPP-THIN-NCP of (Θ, X) for the Poisson process with intensity (5.26). Current values of the parameters are assumed to be $r = 0.1, \phi = 1$ and $T = 100$. $\tilde{X}$ is a Poisson process on $[0, T] \times (0, \infty) \times (0, \infty)$ with mean measure $e^{-\tilde{\epsilon}} dc\ dm\ d\tilde{\epsilon}$; choose all $(C_i, M_i, \tilde{E}_i) \in \tilde{X}$ with $M_i \leq r$ (denoted by circles as opposed to the points with $M_i > r$ denoted by asterisks); project them to S; set $E_i = \tilde{E}_i/\phi$.

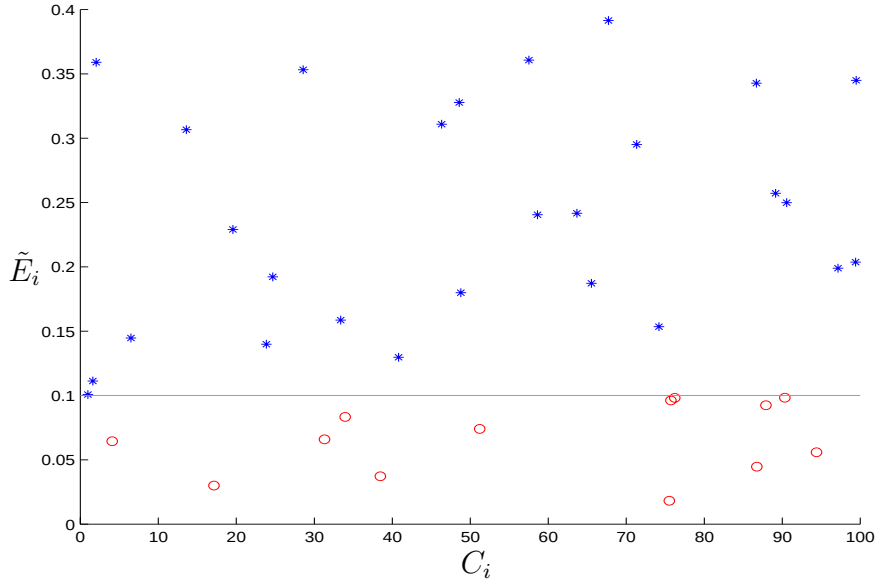


Figure 5.8: The MPP-CDF-NCP of (Θ, X) for the Poisson process with intensity (5.26). Current values of the parameters are assumed to be $r = 0.1, \phi = 1$ and $T = 100$. $\tilde{X}$ is a unit rate Poisson process on S; choose all $(C_i, \tilde{E}_i) \in \tilde{X}$ with $\tilde{E}_i < r$; set $E_i = -\log(\tilde{E}_i/r)/\phi$.

Notice that the decreasing transformation (5.20) is used, therefore $E_1 > E_2 > \cdots$, where $E_1 < \infty$ *almost surely* by the form of the intensity function in (5.26).

The third possible NCP ignores the product space structure of S and is the THIN-NCP suggested in Section 5.5 and Section 5.3. That is, $\tilde{X} = \{(C_i, E_i, M_i), i = 1, 2, \ldots\}$ is a unit rate Poisson process on $S \times (0, \infty)$ and obtain X from $\tilde{X}$ as described below (see also Figure 5.9).

THIN-NCP when $\lambda(c, \epsilon; \Theta) = r\phi \exp\{-\phi\epsilon\}$

Let $\tilde{X} = \{(C_i, E_i, M_i), i = 1, 2, \ldots\}$ as in (5.25).

Select all points from $\tilde{X}$ for which $M_i < r\phi \exp\{-\phi E_i\}$.

Project these points to $[0, T] \times (0, \infty)$.

X consists of the projected points.

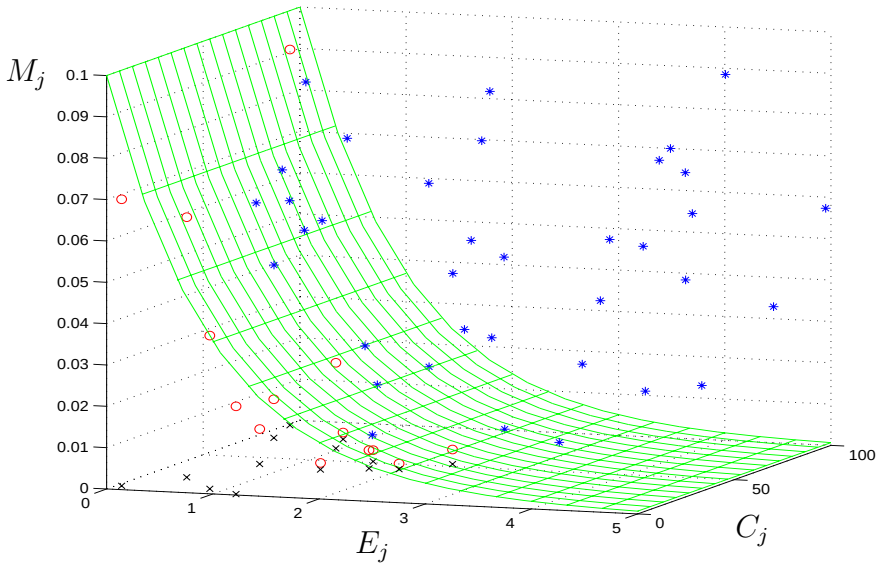


Figure 5.9: The THIN-NCP of (Θ, X) for the Poisson process with intensity (5.26). Current values of the parameters are assumed to be $r = 0.1, \phi = 1$ and $T = 100$. $\tilde{X}$ is a unit rate Poisson process on $S \times (0, \infty)$ and X consists of all (C_i, E_i) such that $(C_i, E_i, M_i) \in \tilde{X}$ and $M_i < \lambda(C_i, E_i)$.

The MCMC implementation of the THIN-NCA is not straightforward, unlike the other

two algorithms, therefore we spend this paragraph to show how to implement the algorithm outlined in Section 5.3.2. Let $\theta_0 = (r_0, \phi_0)$, $\theta_1 = (r_1, \phi_1)$, $X^{(0)}$ and $X^{(1)}$ be defined as in Section 5.3.1, and S_0, S_1, μ and N as defined in Section 5.3.2. It is easy to show that

$$S_0 = \{(c, \epsilon) : \epsilon(\phi_1 - \phi_0) > \log[r_0\phi_0/(r_1\phi_1)]\}$$

which for given current and proposed values of the parameters can be easily found. Let $d = \max\{0, (\phi_1 - \phi_0)^{-1} \log[r_0\phi_0/(r_1\phi_1)]\}$, then

$$\mu = \begin{cases} T[r_1(1 - e^{-\phi_1 d}) - r_0(1 - e^{-\phi_0 d})] & \text{when } \phi_0 > \phi_1 \\ T[r_1 e^{-\phi_1 d} - r_0 e^{-\phi_0 d}] & \text{when } \phi_0 \leq \phi_1 \end{cases}$$

which can be 0. Therefore we can easily draw $N \sim \text{Pn}(\mu)$. Nevertheless, the simulation of any of the $N > 0$ points (c, ϵ) from the density (assuming $\mu > 0$)

$$\frac{1}{\mu}(r_1\phi_1 e^{\phi_1\epsilon} - r_0\phi_0 e^{\phi_0\epsilon}),\ (c, \epsilon) \in S_0 \tag{5.28}$$

is not straightforward. However, we note that the second derivative of the logarithm of this density is

$$-\frac{e^{-(\phi_1+\phi_0)\epsilon} r_0 r_1 \phi_0 \phi_1 (\phi_0 - \phi_1)^2}{(\phi_1 r_1 e^{-\phi_1\epsilon} - \phi_0 r_0 e^{-\phi_0\epsilon})^2}$$

therefore we can use the ARS technique of Wild and Gilks (1993), see Section 1.5.2.

5.6 Completely random measures and subordinators

This section reviews some theory about completely random measures and subordinators, and serves a double purpose. Firstly, it aims at establishing the connection between completely random measures and Poisson processes. This will then allow our methods to be extended to various contexts, ranging from Bayesian non-parametrics to volatility modelling. Secondly, it gives some definitions and properties of positive Lévy processes and subordinators, which provide the blocks which the stochastic volatility models proposed by Barndorff-Nielsen and Shephard (2001) are built upon. Inference for these models is studied in Chapter 6.

The next section follows closely Chapter 8 of Kingman (1993), while the material about subordinators and positive Lévy processes is largely based on Walker and Damien (2000), Ferguson and Klass (1972), Ferguson (1974), Ferguson (1973) and Barndorff-Nielsen and Shephard (2004). Our account will be fairly informal, nevertheless references are given for proofs of the results.

5.7 Completely random measures

The Poisson process, introduced in Section 5.1, is an example of a random measure. If Φ is a Poisson process on a state space S, then each realisation of the process, ϕ say, is a counting measure on S, therefore Φ is a random counting measure. Generally, a random measure on a space S is a stochastic process every realisation of which acts as a measure on S. A research area which is particularly interested in random measures is Bayesian non-parametric modelling, where prior distributions are constructed on spaces of probability measures. This modelling approach, which has its foundations on de Finetti's representation theorem (see Section 1.6), is described in Ferguson (1974) and a more recent review is Walker et al. (1999); see also Section 5.8.2. Bayesian modelling using random measures arises also in the context of spatial statistics, see for example Wolpert and Ickstadt (1998), and more recently in the statistical analysis of inverse problems, see for example Wolpert et al. (2003).

A completely random measure Φ is defined to be a random measure such that for disjoint sets $A_i \subset S$,

$$\Phi\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \Phi(A_i) \tag{5.29}$$

where the random summands on the right are independent random variables. This property suggests that the joint distribution of the random variables $\Phi(A_1), \ldots, \Phi(A_n)$ for arbitrary $n > 0$ and measurable sets A_i is determined by the distribution of $\Phi(A)$ for all measurable $A \subset S$. The distribution of the latter is characterised by its cumulant function

$$K_u(A) := -K(u; \Phi(A)) = -\log \mathsf{E}\{e^{-u\Phi(A)}\},\ u > 0. \tag{5.30}$$

It is easy to show that $K_u(\cdot)$ is a positive measure on S for all $u > 0$. Under the assumption that Φ is non-atomic measure and that $K_u(\cdot)$ is σ-finite (see p.80 of Kingman (1993)) an elegant argument from measure theory (see p.81 of Kingman (1993)) shows that the distribution of the positive random variable $\Phi(A)$ is infinitely divisible (see Section 1.8 for a definition). The Lévy-Khinchine representation for positive infinitely divisible distributions (see Kingman (1993), Feller (1971)) can be used to show that

$$K_u(A) = \beta(A)u + \int_{(0,\infty]} (1 - e^{-uz}) W(A, dz) \tag{5.31}$$

where $\beta(\cdot)$ is a measure on S, $W(A, \cdot)$ is a measure on $(0, \infty]$ for each $A \subset S$, $W(\cdot, B)$ is a measure on S for each $B \subset (0, \infty]$ and W must make the integral in (5.31) converge for all $u > 0$. Both measures β and W are uniquely determined in terms of K_u by (5.31).

Ignoring the deterministic measure $\beta(\cdot)$ for the moment, the similarity between (5.31)

and (5.3) from Campbell's theorem is not coincidental. It can actually be shown (see Section 8.2 of Kingman (1993)) that $\Phi(A)$ has the same distribution as $\beta(A) + C_A$, where C_A is the sum

$$C_A = \sum_i E_i \mathbb{1}[C_i \in A]$$

where $(C_i, E_i) \in \Psi$ and Ψ is a Poisson process on $S \times (0, \infty]$ with mean measure W^*, such that

$$W^*(A \times B) = W(A, B), \ A \subset S, B \subset (0, \infty]. \tag{5.32}$$

It is often possible to represent Ψ as a marked Poisson process. One option is described below. Suppose that the measure defined as

$$\mu(A) = W(A, (0, \infty]) \tag{5.33}$$

is σ-finite. Then, for every $y > 0$, the measure

$$\mu_y(A) = W(A, (0, y]) \leq \mu(A)$$

is absolutely continuous with respect to μ. Since both measures are σ-finite, the Radon-Nikodym theorem (see for example Section 5.14 of Williams (1991)) establishes the existence of a density $F(x, y)$ such that

$$\mu_y(A) = \int_A F(x, y)\mu(dx).$$

Clearly, for every $x \in S$, $0 \leq F(x, y) \leq 1$, $F(x, 0) = 0$ and $\lim_{y\to\infty} F(x, y) = 1$. It is not surprising therefore that $F(x, y)$ is the distribution function of a random variable with values in $(0, \infty]$ for all x, although some measure-theoretic care is necessary to derive this result.

Hence, when μ in (5.33) is σ-finite, the Poisson process Ψ corresponding to the completely random measure, can be represented as a marked Poisson process. The points of the process form a Poisson process on S with mean measure μ. The distribution function of the mark of a point at x is $F(x, y)$. There are interesting examples where $W(A, (0, y]) = \infty$ for all $y > 0$ and all sets A of positive measure, hence the σ-finiteness assumption fails. In such cases the representation is not true anymore, since essentially there does not exist a conditional distribution of the mark y for a given point x. Nevertheless, when S is a subset of the real line, there might still exist an alternative marked Poisson process representation, even when μ in (5.33) is not σ-finite. This construction is shown in the next section.

5.8 Positive independent increments processes, subordinators and representations

Independent increments processes were defined in Section 1.8, and denoted by $z(t), t \geq 0$. The increments of a positive independent increments process are positive random variables, therefore the stochastic process has non-decreasing sample paths. Such processes can be explicitly constructed from completely random measures on the real line. We focus on random measures Φ on the positive half-line and define

$$z(t) = \Phi((0,t]),\ t \geq 0. \tag{5.34}$$

It is easy to verify that the definition in (5.34) ensures that the increments

$$z(t+s) - z(t),\ t, s > 0 \tag{5.35}$$

are positive and independent random variables. When the distribution of the increments in (5.35) is independent of t, z is called a subordinator.

The kinship between completely random measures and Poisson processes uncovered in the previous section, suggests that we can write (see Figure 5.2)

$$z(t) = \sum_i E_i \mathbb{1}[C_i \leq t] \tag{5.36}$$

where $(C_i, E_i) \in \Psi$ and Ψ is a Poisson process on $(0,\infty) \times (0,\infty]$ with mean measure W^* defined by (5.32). This representation guides the choice of the augmentation scheme for latent subordinators in Chapter 6.

We now try to find a marked Poisson process representation of positive independent increments processes. A first step is to write down the measure

$$\mu(A) = W(A, (0,\infty])$$

and check if it is σ-finite. In this case, the previous section suggests a marked Poisson process representation. However, in many interesting examples this measure is not σ-finite. Nevertheless, an alternative representation exists and it is outlined below.

Without loss of generality we take $t \in [0,1]$, that is we focus on $\Psi \cap ([0,1] \times (0,\infty])$, which will be denoted simply as Ψ. With a slight abuse of notation, we define the family of measures on $(0,\infty]$, such that for $B \subset (0,\infty]$,

$$W_t(B) \ = \ W((0,t], B),\ 0 \leq t \leq 1. \tag{5.37}$$

In view of Theorem 5.1.2, the integral in (5.31) converges if and only if $W_t((y,\infty]) < \infty$ for all $y > 0$ and $t \geq 0$, although it may still be possible that $W_t((0,y]) = \infty$. Thus, W_1 is σ-finite, since we can write, for example, $(0,\infty] = \cup_{n=1}^{\infty}(1/n,\infty]$ where $W_1(1/n,\infty] < \infty$ for all $n > 0$. Clearly W_t is absolutely continuous with respect to W_1 and by the Radon-Nikodym theorem there exists a density

$$n_t(y) = \frac{dW_t}{dW_1}(y). \tag{5.38}$$

For every y, $n_t(y)$ is a non-decreasing function of t, $n_0(y) = 0$, $n_1(y) = 1$, therefore it behaves as a distribution function of a random variable on $[0,1]$. Thus, Ψ can be simulated as a marked Poisson process, by first simulating a Poisson process on $(0,\infty]$ with mean measure W_1, and for each point y of this process, we simulate a mark from the distribution function $n_t(y)$ on $[0,1]$.

In the literature, the measures defined in (5.37) are known as Lévy measures. When the Lévy measure $W_t(\cdot)$ possesses a Lebesgue density, usually denoted by w_t, it is termed a Lévy density. When z is a subordinator the mean measure of Ψ is a product measure and

$$W_t(B) = W((0,t],B) = tW_1(B).$$

The stochastic volatility models developed by Barndorff-Nielsen and Shephard (2001) are solely concerned with such Lévy measures. As an example of a subordinator, consider the compound Poisson process introduced in Section 5.1.3 (although using a different notation) which is specified as in (5.36) with the further assumption that the C_is form a Poisson process in time with a finite rate $r > 0$ and that the E_is are positive, IID from some distribution $P(dx)$ and also independent from the C_is. The stationarity of its increments can be easily verified and the Lévy measures are

$$W_t(B) = t\ rP(B); \tag{5.39}$$

see for example Section 2.2.2 of Barndorff-Nielsen and Shephard (2004) for a derivation of this result, which can however be shown easily from first principles. Consequently,

$$n_t(y) = t, \text{ for all } y > 0.$$

On the other hand, the Bayesian non-parametric literature is more interested in non-homogeneous processes, that is where the distibution of the increments in (5.35) varies with t. An important example is the non-homogeneous gamma process (see for example Ferguson

(1973) and Example 1 of Walker and Damien (2000)) for which

$$dW_t(y) = \alpha(t)y^{-1}\exp(-y)dy,\ t \geq 0$$

where $\alpha(\cdot)$ is a bounded non-decreasing function on $(0, \infty)$ and $\alpha(0) = 0$. This process is explicitly involved in the construction of the Dirichlet process, which plays a prominent role in Bayesian non-parametrics (see for example Ferguson (1974)). Notice that for all $t > 0$ and $y > 0$, $W_t((0, y]) = \infty$, while $W_t((y, \infty]) < \infty$. Therefore, the gamma process can be represented as a marked Poisson process, as described in this section.

5.8.1 The Ferguson-Klass representation and approximations

Ferguson and Klass (1972) proposed a representation of processes with independent and positive increments having no Gaussian components and fixed points of discontinuity. The purpose of this representation is both theoretical, to assist the investigation of *almost sure* sample path properties of these prossesses, but also practical in terms of simulating positive independent increments processes. Here we argue that there is another aspect in this representation, it can be viewed as a non-centered transformation of the positive independent increments process, which relates to the CDF-NCP constructed in Section 5.4.

Recall from Section 5.7 that completely random measures can be constructed by means of Poisson processes. Section 5.8 showed how this construction specialises for positive independent increments processes, in particular (5.36) expresses the value $z(t)$ as the sum over all "marks" E_i of the "marked Poisson process" Ψ with points C_i in $[0, t]$ (we will explain the use of quotations in the next lines). We would then typically take the C_is ordered in time, that is $C_1 < C_2 < \ldots$, and then the index i of E_i would just denote which point C_i it corresponds to.

Of course, we noted in Section 5.8 that this interpretation of Ψ as a marked Poisson process is not valid if the measure $W(\cdot, (0, \infty])$ defined in (5.33) is not σ-finite; this explains the use of the quotation marks in the sentence above. Essentially, when this finiteness condition fails we cannot talk about the conditional distribution of E_i given C_i, or equivalently we cannot talk about the marginal distribution of the E_is. However, since the conditional distribution of the C_i given E_i exists and is given in (5.38), Section 5.8 gave an alternative marked Poisson process representation of Ψ, one where the marginal Poisson process $\{E_i, i = 1, 2, \ldots\}$, with mean measure $W_1(\cdot)$, is marked by the C_is. In this representation, the E_is are ordered in $(0, \infty)$, which can be thought of as "time". Moreover, when the σ-finiteness condition fails, then for any $t > 0$ there will be an infinite number of terms in the sum (5.36). This suggests that it is both very convenient and valid to order the E_is by taking E_1 to be the largest value. Section 5.8 (see in particular the discussion after expression (5.37)) argued that $W_1((y, \infty]) < \infty$ for all $y > 0$, which ensures that the largest

E_i in the Poisson process $\Psi = \{(C_i, E_i), i = 1, 2 \ldots\}$ is finite *almost surely*. Actually, the smallest E_i is not well defined in this setting since 0 is a limit point of the marginal process $\{E_i, i = 1, 2, \ldots\}$.

The mean measure of the Poisson process $\{E_i, i = 1, 2, \ldots\}$ satisfies the conditions (5.18) and Section 5.4 gave an algorithm for obtaining the points of this process in a decreasing order, that is where $E_1 > E_2 > \ldots$ via a decreasing transformation (given in (5.20)) of a unit rate Poisson process. This is exactly the representation proposed by Ferguson and Klass (1972) and it is described below, where we mainly follow Walker and Damien (2000). We use the notation set up in the previous section.

Let $M(x) = W_1([x, \infty))$, which is finite for all $x > 0$, although it is possible that $M(0) = \infty$. Then define the non-negative random variables $E_1, E_2, \ldots$ by

$$\begin{aligned} P[E_1 \leq x_1] &= \exp\{-M(x_1)\} \\ P[E_i \leq x_i \mid E_{i-1} = x_{i-1}] &= \exp\{-M(x_i) + M(x_{i-1})\} \quad (x_i < x_{i-1}). \end{aligned}$$

It can be seen that $E_1 > E_2 > \cdots$, actually we can obtain E_i by solving the equation

$$\tilde{E}_i = M(E_i), \tag{5.40}$$

where $\tilde{E}_1, \tilde{E}_2, \ldots$ are the arrival times of a unit rate Poisson process on $(0, \infty)$. If $\tilde{E}_i > M(0)$ then E_i is defined to be 0 (which is the same convention adopted in Section 5.4). The Ferguson and Klass (1972) representation is given by

$$z(t) = \sum_i M^{-1}(\tilde{E}_i) \mathbb{1}[U_i \leq n_t(M^{-1}(\tilde{E}_i))], \ t \in [0, 1] \tag{5.41}$$

where M^{-1} denotes the inverse function of M, the U_is are independent and identically distributed random variables distributed as Un$[0, 1]$. We showed earlier and in Section 5.8 that C_i conditionally on E_i has a distribution function $n_t(E_i), \ t \in [0, 1]$ defined in (5.38). Therefore, a draw from this conditional can be obtained by the inverse CDF method (see Ripley (1987)) using a Un$[0, 1]$ random variable U_i. It is then immediate that the condition $\mathbb{1}[U_i \leq n_t(E_i)]$ is the same in distribution as $\mathbb{1}[C_i \leq t]$ where C_i is a random variable with distribution function $n_t(E_i), \ t \in [0, 1]$. Thus, (5.41) could be written equivalently as

$$z(t) = \sum_i E_i \mathbb{1}[C_i \leq t], \ t \in [0, 1]. \tag{5.42}$$

As an example consider the compound Poisson process

$$z(t) = \sum_i E_i \mathbb{1}[C_i \leq t] \tag{5.43}$$

where $E_i \sim \text{Ex}(\phi)$ and the points C_i form a Poisson process with rate r on $[0,1]$. The Poisson process $\Psi = \{(C_i, E_i), i = 1, 2, \ldots\}$ has intensity function given in (5.26). Specialising the general result (5.39), the Lévy density of z is given by

$$w_t = t\ r\phi \exp\{-\phi x\},\ x > 0 \tag{5.44}$$

and

$$\begin{aligned} M(x) &= r \exp\{-\phi x\} \\ M^{-1}(x) &= -\frac{1}{\phi} \log\{x/r\},\ x \leq r. \end{aligned}$$

Then

$$z(t) = \sum_i -\frac{1}{\phi} \log\{\tilde{E}_i/r\} \mathbb{1}[C_i \leq t],\ t \in [0,1] \tag{5.45}$$

where $C_i \sim \text{Un}[0,1]$ and are independent of the $\tilde{E}_i$s and where we take the sum over all $\tilde{E}_i < r$.

It should be recognisable the similarity of the Ferguson-Klass representation with the CDF-NCP of Section 5.4. In fact, the Ferguson-Klass representation can be thought of as a non-centered transformation. Let Θ denote any parameters of the Lévy measures W_t. Then the collection of points $\{(U_i, \tilde{E}_i), i = 1, 2, \ldots\}$ is independent of Θ and it can be transformed to yield the Poisson process Ψ, which the independent increments process is constructed from, by first finding the E_is from the $\tilde{E}_i$s and then the C_is from the U_is by inverting their distribution function. When, for example, z is a subordinator, this transformation coincides with the MPP-CDF-NCP constructed in Section 5.5 and illustrated through an example in Section 5.5.1. In particular, the Poisson process $\{(C_i, \tilde{E}_i), i = 1, 2, \ldots\}$ is the $\tilde{X}$ defined for the MPP-THIN-NCP in Section 5.5.1 and its transformation to yield Ψ (and consecutively z) is given in the algorithm proposed there.

When the Lévy density w_1 is not integrable, the point process Ψ corresponding to the process $z(\cdot)$ is not locally finite and has an infinite number of points on the sets $[0,1] \times (0, y]$ for all $y > 0$. In turn, this means that there are infinite number of terms in the Ferguson-Klass sum (5.41) for all $t \in (0,1]$. If we are interested in using this expansion to simulate the Lévy process, for example its value at time one $z(1)$, we need to do some truncation and use only a finite number of terms. This problem is investigated in several papers, for example Barndorff-Nielsen and Shephard (2001), Rosinski (2002), Bondesson (1982), Walker and Damien (2000), Damien et al. (1995). For many processes the series of decreasing jumps $\{E_i, i = 1, 2, \ldots\}$ is quite quickly converging to 0 allowing for easy truncation, however, care generally must be taken. Bondesson (1982) discusses several methods of truncation and also

possible approximations to the terms not included in the expansion. We will not go into any more detail here, since this thesis only considers integrable Lévy measures.

5.8.2 Applications to Bayesian non-parametrics

Typically, Bayesian non-parametric modelling is concerned with constructing priors over spaces of distributions and functions; for a review of this area see Walker et al. (1999). The models often have a hierarchical structure, where the latent process is either a completely random measure or a positive independent increments process. For example, in the context of survival analysis Ferguson and Phadia (1979) introduced the neutral to the right random distribution function $F(t), t \geq 0$, which can be written in the form

$$F(t) = 1 - e^{-z(t)}$$

where z is a positive independent increments process. Then, conditionally on F, the data $Y_1, \ldots, Y_m$ are independent and identically distributed according to F.

The underlying independent increments process or more generally the random measure depend on certain parameters Θ, which often control important characteristics of the random measure, and it is sensible to try to be least informative about them. Therefore we would like to assign a prior on Θ, thus giving rise to a three-stage hierarchical model, and make inferences about it based on its posterior distribution. It is of course of interest whether the data (or more appropriately, what kind of data) contain sufficient information about Θ and there are suggestions (see for example Walker and Damien (1998)) that in some cases these parameters might be weakly identified. In any case, it is necessary to be able to sample from the posterior distribution of Θ to address such questions appropriately.

However, sampling from the joint distribution of Θ and the random measure is far from straightforward. Current ongoing work with G. Roberts and M. Sköld reveals that for many common problems componentwise-updating algorithms are essentially reducible, when a centered parameterisation is used. This is consequence of the fact that the random measure might contain infinite information about its parameters (see for example Section 1.8). When the random measure is discretised (see for example Walker et al. (1999) for an illustration), which is common practice when working with infinite activity (i.e non-compound Poisson) Lévy processes, reducibility is avoided, nevertheless the finer the discretisation (and hence the approximation to the true random measure) the worse the performance of the centered algorithm. This bears a strong similarity to the situation encountered in the inference for partially observed diffusions, see Roberts and Stramer (2001). Non-centered techniques are particularly relevant in this context since they avoid the problem of reducibility. When the measure is expressed via a positive Lévy process, the representation of the latter as a Poisson process can be exploited to find an NCP as described in Section 5.3 and in Section 5.4.

Chapter 6

Inference for Non-Gaussian OU models

6.0 Introduction

This chapter begins with a short introduction to financial modelling and stochastic volatility (SV). We present a model introduced by Barndorff-Nielsen and Shephard (2001), where the latent volatility process is modelled by a non-Gaussian Ornstein-Uhlenbeck (OU) process. Such processes are described by stochastic differential equations driven by Lévy processes. We consider Bayesian inference for these models, we introduce a data augmentation method based on marked Poisson processes and a corresponding centered parameterisation. We then employ the methods developed in Chapter 5 to develop various non-centered parameterisations for this augmentation scheme. We compare all methods using simulated data. We also describe algorithms which can be used to infer about the more general SV model based on superpositions of OU processes. We propose a new graphical model diagnostic tool, which is used to investigate whether certain aspects of the latent structure can be identified from the data. We apply our methods to a series of DM/US$ exchange rates. We finish by discussing further extensions of our work. The work in this chapter is based on Roberts et al. (2003).

The notation used in this chapter differs slightly from that used in the rest of the thesis; see Section 1.2 for more details. This is done to keep consistency with Roberts et al. (2003) and generally the chapter is written in a self-contained manner, so that no confusion due to notation is caused.

6.1 Financial markets and stylised facts

Financial data consist, among others, of currency exchange rates, share prices and stock market index values. The data typically take the form of a discrete time series of asset values

$\{P_n, n = 1, 2, \ldots\}$ obtained at times $\{t_n, n = 1, 2, \ldots\}$. It is usually assumed that the time points are equally spaced and then $\Delta = t_n - t_{n-1}$ is the data frequency. Most applications, including those in this thesis, take $\Delta = 1$ to be one day, but there is an increasing interest in analysing high-frequency data, for which $\Delta = 1$ can be as small as five minutes. Such analyses can be found for example in Barndorff-Nielsen and Shephard (2002a) and Andersen et al. (2001). On the other hand, market micro-structure characteristics complicate the analysis of finer than five-minute transaction data, see for example Rydberg and Shephard (1998). When we deal with daily data, the year consists of approximately 261 trading days, after the removal of weekends and bank holidays, when no transactions happen. Thus, Monday and Friday are treated as consecutive days, see Section 2.5 of Taylor (1986) for details on this issue and its effect on modelling.

It is widely accepted that the price series $\{P_n\}$ is non-stationary, see for example Taylor (1986), Mikosch (2002), Mills (1999). Instead, once Δ has been fixed, it is more interesting to study the series of log-returns $\{y_n, n = 1, 2, \ldots\}$ defined as

$$x_n = \log\{P_n\} \tag{6.1}$$

$$y_n = x_n - x_{n-1}. \tag{6.2}$$

A Taylor-series argument shows that this series is close to the series of relative returns, $(P_n - P_{n-1})/P_{n-1},\ n = 1, 2, \ldots$, which describes the relative change over time of the price process. Both series are free of scale and thus comparable among different financial assets. However, it is mathematically more convenient to work with the $\{y_n\}$ series for several reasons (see p.13 of Taylor (1986) and p.9 of Campbell et al. (1997)). From a modelling perspective, the log-returns have an additive structure and are therefore easier to model. On top of that, continuous time generalisations of the discrete time series, as for example in Section 6.2, are easier when working with log-returns. It is generally believed that $\{y_n\}$ can be modelled by a stationary stochastic process. For example the stochastic volatility (SV) models discussed in Section 6.2 all satisfy this assumption.

Figure 6.1, Figure 6.2 and Figure 6.3 show three different data sets, which are going to be used in this chapter. The first two are currency exchange rates and the third is stock index values. All series take $\Delta = 1$ to be one day and are closing prices. We plot both the original series of values and the log-returns series, for each of the assets.

Extensive empirical work has revealed characteristics that the returns series $\{y_n\}$ corresponding to different financial assets have in common, the so-called 'stylised facts'. As Mikosch (2002) points out, these similar properties depend on the time scale chosen, that is on the size of $\Delta = 1$. Depending on whether the latter is a week, a day or five minutes qualitative differences in the returns series are expected. A review of the stylised facts of daily returns series can be found in Chapter 2 of Taylor (1986). More recent reviews include

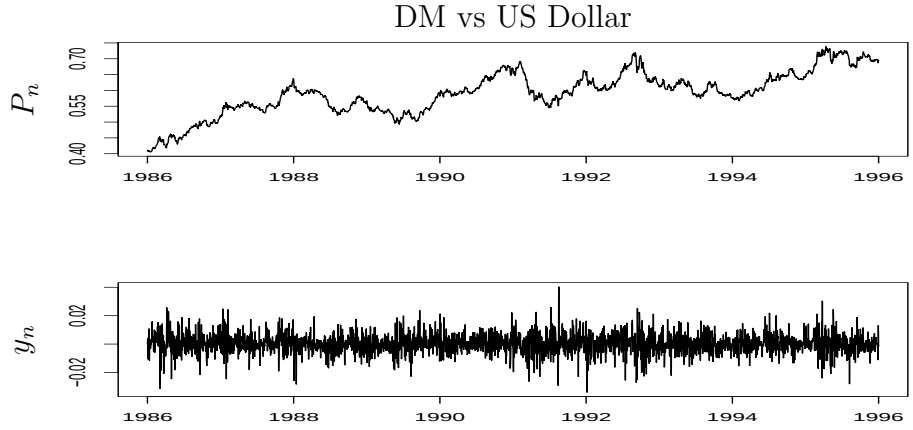


Figure 6.1: Series of daily prices and log-returns for the exchange rate of the DM (Deutsch Mark) against the US Dollar.

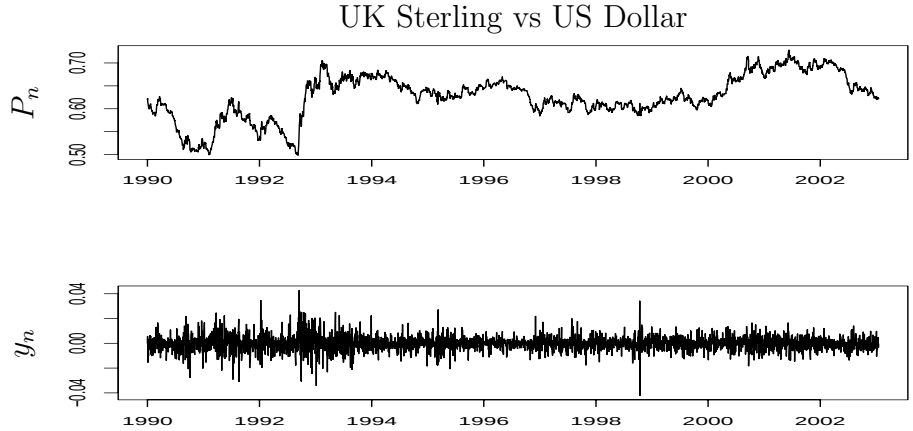


Figure 6.2: Series of daily prices and log-returns for the exchange rate of the UK Sterling against the US Dollar.

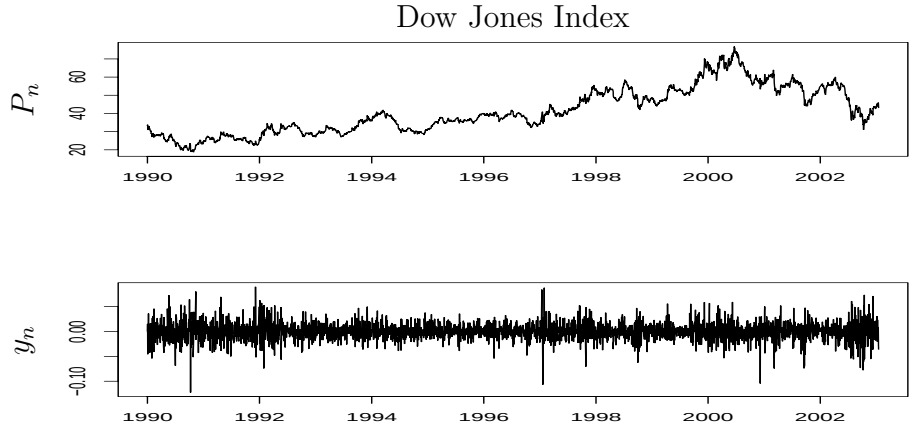


Figure 6.3: Series of daily prices and log-returns for the Dow Jones index.

the Chapter 5 of Mills (1999), Mikosch (2002), Chapter 1 of Campbell et al. (1997) and Chapter 1 of Barndorff-Nielsen and Shephard (2004). The latter also discusses stylised facts of high-frequency data, see also Andersen et al. (2001). We now briefly review the stylised facts of daily series and relate them to the data sets we have introduced in this section.

There are certain characteristics of the marginal distribution of the daily log-returns that seem to be shared by most financial series. The sample mean is close to zero, and it is much larger for shares and indices than currencies. The sample variance is very small, of the order 10^{-4} or even smaller. There is some evidence for skewness, but it is not very large. Typically skewness is negative (although Taylor (1986) finds the opposite in many stock-price series he studies) and its magnitude is much larger for equities than currencies. The negative skewness is a peculiar characteristic from an economic theory viewpoint, since investors should have preference for positively skewed returns, as a compensation for the risk they take (see also Section 6.17). The most profound stylised fact is the heavy tailed nature of the returns distribution, as for example measured by kurtosis. One thing all studies agree with, is that returns over short periods, for example daily, are much more heavy tailed than Gaussian. This is illustrated in Figure 6.4, where a non-parametric estimate of the log-density of the log-return distribution is plotted for each of the series, with the Gaussian log-density fitted to the data superimposed. Mikosch (2002) points out that the tails can be successfully modelled by distributions with power law tails. It has been suggested that the stable distributions (see Section 4.3 for some definitions), which do have power tails but also imply an infinite variance for the log-returns, could be used for the marginal distribution. This has been advocated by Mandelbrot and Fama, see Mills (1999). However, these models are not consistent with another observed feature: "aggregation to Gaussianity". That is, returns over long time horizons, for example weekly or monthly, "look" much more Gaussian than daily data and can be satisfactorily modelled by the normal distribution; see for example Figure 1.7 of Barndorff-Nielsen and Shephard (2004). Stable distributions do not obey such a central limit type result.

Some facts about the dependence structure of the log-returns series also seem to be agreed upon in the literature. There is very little autocorrelation in the $\{y_n\}$ series, usually all but the first lags are negligible. On the other hand, the log-returns are not independent in time. Mandelbrot (see Barndorff-Nielsen and Shephard (2004) for references) was the first to observe that large changes in the returns tend to be followed by large changes of either sign. In terms of the observed series, there is positive and significant autocorrelation over many lags in the $\{|y_n|\}$ and $\{y_n^2\}$ series. This is known as "volatility clustering". This feature is apparent in Figure 6.5, where estimated autocorrelations are plotted for each of our series. Mikosch (2002) also reviews findings about the dependence in the tails of the series and thus clustering of extreme absolute or squared returns. Black (1976) was the first to observe that the returns of equities are negatively correlated with future volatilities. This is known as

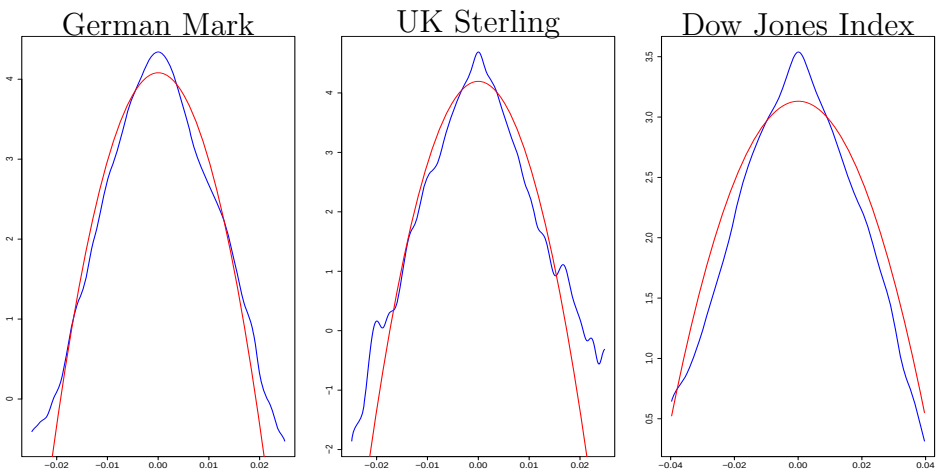


Figure 6.4: The logarithm of the estimate of the unconditional density of the log-returns using kernel density estimation (blue lines), for the three financial datasets introduced in Section 6.1. We superimpose the log-density of the Gaussian distribution (red lines) which has the same first two moments as the data.

the leverage effect and it is supported by some economic arguments, which also explain why similar pattern is not observed in the exchange rates market. The leverage effect suggests that low returns tend to be followed by high volatility.

6.2 Stochastic volatility modelling

There is a strong interest in developing statistical models which can capture the observed characteristics of financial data. However, the non-linearity and non-Gaussianity which characterises such data and was exposed in Section 6.1, complicates the modelling task. For instance, the classical ARMA-type time series models, as for example developed in Brockwell and Davis (1991), cannot capture any of the features of financial returns. On the other hand, it is desirable to construct stochastic models which describe the evolution of prices in order to solve problems in theoretical finance, as for example the pricing of financial derivatives. For a thorough presentation of this theory see Duffie (1992). Typically the two tasks are conflicting since models which are satisfactory from an econometric point of view might be too complicated to allow for analytic solutions to mathematical finance problems.

An empirically and mathematically attractive class of models for financial returns is the stochastic volatility (SV) model. These are essentially hierarchical models which construct the distribution of the observed price process conditionally on an unobserved (latent) stochastic process of conditional variances (volatilities). SV models are constructed either in continuous or discrete time, although often there exist equivalent representations of the same model in either framework.

For example, the log-Gaussian SV model is widely studied and extensively used in prac-

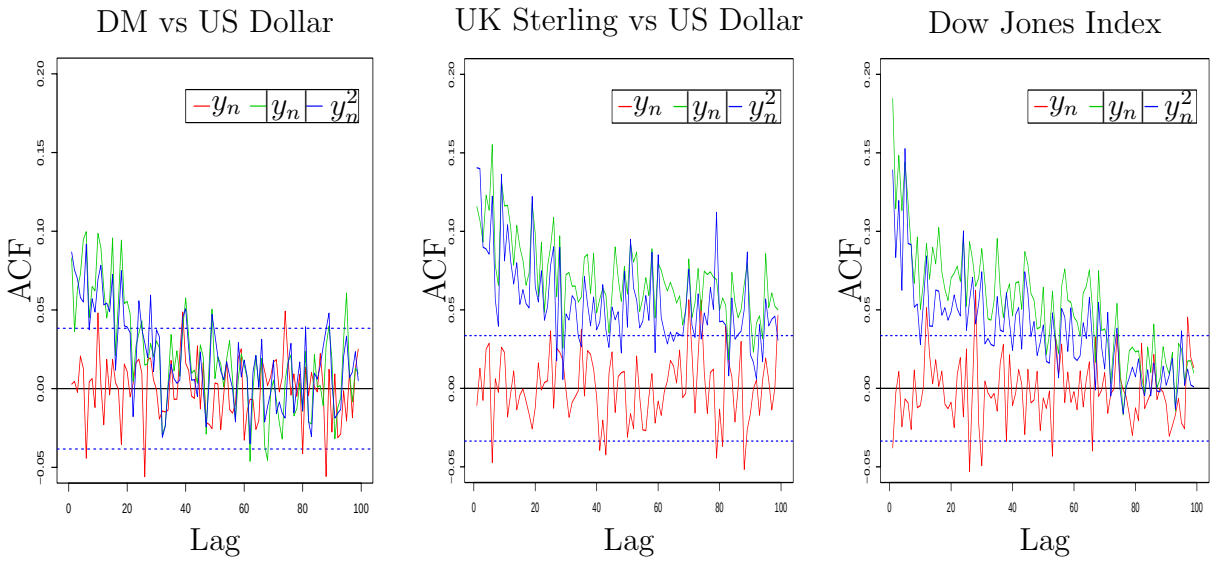


Figure 6.5: Sample autocorrelations for the series of daily log-returns (y_n), their absolute values ($|y_n|$) and their squares (y_n^2) for the German DM-US Dollar exchange rate (left), the UK Sterling-US Dollar exchange rate (middle), and the Dow Jones index (right).

tice, and it is usually encountered in a discrete time form as

$$\begin{aligned} y_n &= (v_n^*)^{1/2}\epsilon_n,\ \epsilon_n \sim \mathrm{N}(0,1) \\ \log\{v_{n+1}^*\} &= \mu + \phi(\log\{v_n^*\} - \mu) + \sigma z_n,\ z_n \sim \mathrm{N}(0,1) \\ \log\{v_1^*\} &\sim \mathrm{N}\left(\mu, \frac{\sigma^2}{1-\phi^2}\right). \end{aligned}$$

The series $\{v_n^*, n = 1, 2, \ldots\}$ contains the conditional variances of the log-returns and it is independent of the white-noise series $\{\epsilon_n, n = 1, 2, \ldots\}$. The stationarity assumption $|\phi| < 1$ is usually made and then the marginal distribution of the log-volatility is given by the last expression in the above model. There are many review papers discussing the econometric properties and the statistical estimation of such SV models, see for example Taylor (1986), Kim et al. (1998), Shephard (1996), Ghysels et al. (1996), Jacquier et al. (1994).

In continuous time, SV models are most often described by stochastic differential equations (SDEs). In particular, the log-price of an asset is the solution to an SDE of the form

$$dx(t) = (\alpha + \beta v(t))dt + v(t)^{1/2}dB(t),\ t \in [0, T] \tag{6.3}$$

where $B(\cdot)$ is a standard Brownian motion, α is the drift and β is the risk premium. These two parameters will be ignored for the moment and will be reconsidered in Section 6.17. The

returns series $\{y_n\}$ is obtained through aggregation,

$$y_n = \int_{(n-1)\Delta}^{n\Delta} dx(t) = x(n\Delta) - x((n-1)\Delta). \tag{6.4}$$

When $v(t)$ is constant in time, (6.3) collapses to the so-called Black-Scholes model (which was also proposed by Samuelson and Merton though). A great deal of the mathematical finance theory has been based on this model to develop pricing formulae for contingent claims. However, the Black-Scholes model is inappropriate to model financial data, since it implies that returns are Gaussian and independent in time, two assumptions which contradict the empirical findings of Section 6.1.

Instead, $v(\cdot)$ is modelled as a stationary stochastic process, latent and independent of $B(\cdot)$. It is also usually assumed that $v(\cdot)$ is cadlag (right-continuous with limits from the left). This is enough to ensure that $x(\cdot)$ has continuous sample paths (see Section 6.2 of Barndorff-Nielsen and Shephard (2004)). The model is completed by deciding on a specific form for the stochastic volatility $v(\cdot)$, which typically depends on some parameters.

The integrated volatility process is defined through

$$v^*(0,t) = \int_0^t v(s)ds, \ t \geq 0 \tag{6.5}$$

from which the so-called actual volatilities are obtained as

$$v_n^* = v^*(0, n\Delta) - v^*(0, (n-1)\Delta). \tag{6.6}$$

It follows that

$$y_n \mid v_n^* \sim \mathrm{N}(\alpha\Delta + \beta v_n^*, v_n^*)$$

so the marginal distribution of the returns is a scaled mixture of normals, and therefore it can exhibit heavy tails and skewness. The dependence in the returns series is implicitly induced by the dependence in the volatility process. Moreover if $v(\cdot)$ is ergodic then it can be shown that the log-returns over long lags tend to normality, but the rate will depend on the memory of the volatility process. See Section 4 of Barndorff-Nielsen and Shephard (2001) for general aggregation results and proofs of the above statements, and Section 6.1 of the same paper for the connection between stochastic volatility and subordination. In Barndorff-Nielsen and Shephard (2001) and Barndorff-Nielsen and Shephard (2002a) it is shown that the second order properties of the actual volatility series (6.6) solely depend on the second order properties of the volatility process. Barndorff-Nielsen and Shephard (2001) recently introduced a continuous-time SV model based on Lévy processes, which is reviewed in Section 6.3.

In applications of SV models to financial data, smoothing, filtering and prediction of the

volatilities, as well as parameter estimation are major goals of the statistical analysis. Of particular interest are the parameters which control the memory of the volatility process and thus the volatility clustering. Likelihood-based inference for SV models is complicated, since likelihood functions are not available in closed form, and computer intensive methods such MCMC and the EM algorithm have to be employed. Considerable progress has been made though in the inference for the log-Gaussian discrete time model. A lively discussion on the available estimation methods can be found in Jacquier et al. (1994). The most effective computational algorithm seems to be the one reviewed by Kim et al. (1998), where lots of other issues are resolved, such as filtering and smoothing using particle filters and model selection. However, this technology is not applicable to the class of models described in Section 6.3.

6.3 The Barndorff-Nielsen and Shephard model

A new class of SV models was introduced by Barndorff-Nielsen and Shephard (2001), where volatility is modelled as a linear but non-Gaussian Ornstein-Uhlenbeck (OU) process. This paper, together with a series of other papers by the same authors, develops the relevant theory, derives the econometric properties and deals with lots of other issues regarding these SV models; some important references include Barndorff-Nielsen and Shephard (2002a), Barndorff-Nielsen and Shephard (2003), Barndorff-Nielsen et al. (2002), Barndorff-Nielsen and Shephard (2002b), and Barndorff-Nielsen and Shephard (2004) for a book-length review of this area. This section reviews the main results in order to prepare the ground for the main contribution of this chapter, which is Bayesian inference for non-Gaussian OU SV models using MCMC.

6.3.1 Construction of the model

A stationary stochastic process $v(t),\ t \geq 0$ is of OU type if it can be represented as

$$v(t) = e^{-\mu t}\left\{v(0) + \int_0^t e^{\mu s} dz(s)\right\}$$

where $z(\cdot)$ is a Lévy process, that is a process with stationary and independent increments and such that $z(0) = 0$ *almost surely*; see Section 1.8 for the basic definitions of Lévy processes and Sato (1999) for a detailed exposition of the area. The initial volatility $v(0)$ is a random variable assumed to be distributed according to the stationary distribution of $v(t)$. That is, we assume that the process $v(\cdot)$ is started in stationarity.

The OU process is often expressed as an SDE

$$dv(t) = -\mu v(t)dt + dz(t). \tag{6.7}$$

Some conditions on $z(\cdot)$ need to be imposed to ensure the existence of a stationary solution of (6.7). This will be considered after the statement of Theorem 6.3.1. The process $z(\cdot)$ is termed the background driving Lévy process (BDLP), due to its role in the above SDE. The OU process is a continuous time generalisation of the well known discrete time autoregressive (AR) process. It is not surprising therefore that when second moments exist,

$$r(t) = \text{Corr}(v(0), v(t)) = \exp(-\mu t),\ t > 0. \tag{6.8}$$

The specification in (6.7) is such that the stationary distribution of the OU process, when it exists, it depends on μ. It is mathematically and statistically desirable to parameterise $v(\cdot)$ in terms of its stationary and transient characteristics separately, therefore it is preferable to rewrite the model in a way that the stationary distribution of $v(\cdot)$ is independent of μ. This can be achieved using the time-change suggested in p.2 of Barndorff-Nielsen and Shephard (2001),

$$dv(t) = -\mu v(t)dt + dz(\mu t) \tag{6.9}$$

which implies that

$$v(t) = e^{-\mu t}\left\{v(0) + \int_0^{\mu t} e^s dz(s)\right\}. \tag{6.10}$$

This solution has marginal distribution independent of μ, a result which follows from equation (11) of Barndorff-Nielsen and Shephard (2001).

When $z(\cdot)$ is a subordinator (see Section 5.8) $v(t)$ is positive for all $t \geq 0$ and Barndorff-Nielsen and Shephard (2001) suggested using OU processes driven by subordinators as models for the volatility.

Section 5.8 established the representation of subordinators as sums over Poisson processes. This representation is mathematically and computationally convenient for the treatment of OU processes. For example, properties of the stochastic integrals

$$\int_0^t f(s)dz(s)$$

which play a prominent role in the theory of OU processes, can be derived using Campbell's Theorem 5.1.2. On the other hand, simulations from these random variables can be performed by using a direct extension of the Ferguson-Klass representation (see Section 5.8.1). Moreover, our data augmentation methods in Section 6.6 are based on this representation.

The marginal distribution of $v(\cdot)$ defined as the solution to (6.9) cannot be arbitrary. In

fact, the following theorem, which is Theorem 1 of Barndorff-Nielsen and Shephard (2001), describes exactly the family of possible distributions. We begin with a definition.

Definition 6.3.1. *Let ϕ be the characteristic function of a random variable V. V is self-decomposable if for all $c \in (0,1)$*

$$\phi(u) = \phi(cu)\phi_c(u) \tag{6.11}$$

where ϕ_c is a valid characteristic function for all $c \in (0,1)$.

Theorem 6.3.1. *If V is self-decomposable, then there exists a stationary stochastic process $v(t)$ and a Lévy process $z(t)$, linked by (6.10), such that $V \stackrel{d}{=} v(t)$ for all $\mu > 0$ in (6.10). Conversely, if $v(t)$ is a stationary stochastic process and $z(t)$ is a Lévy process such that $v(t) \stackrel{d}{=} V$ and $v(t)$ and $z(t)$ satisfy (6.10) for all $\mu > 0$, then V is self-decomposable.*

Before exploring the modelling aspects of this theorem, we explain briefly self-decomposability. Self-decomposable distributions are also infinitely divisible, see Section 1.8 for a definition of the latter. For a proof of this property we refer to Barndorff-Nielsen and Shephard (2002b) (in particular their Theorem 4.1). Therefore, self-decomposable distributions are characterised by a Lévy measure (see Section 5.8), and we will use the generic notation $U(dx)$ for the Lévy measure of the distribution of $v(t)$. Notice that Definition 6.2.1 implies that for all $c \in (0,1)$

$$V \stackrel{d}{=} cY + \epsilon$$

where Y has the same distribution as V, the characteristic function of the random variable ϵ is ϕ_c (and thus depends on c), and ϵ is independent of Y. It is therefore not surprising that such distributions appear as the only choice for marginals of stationary OU processes, recall, for example, that a stationary discrete-time AR process $\{V_n, n = 1, 2, \ldots\}$ is represented as

$$V_n = cV_{n-1} + \epsilon. \tag{6.12}$$

Important examples of self-decomposable distributions on the positive half-line are the log-normal and the generalised inverse Gaussian family (see for example Barndorff-Nielsen and Shephard (2003)). The latter contains the gamma, the inverse Gaussian, the inverse gamma and the positive hyperbolic distributions as special cases.

Theorem 6.3.1 suggests two alternative modelling approaches when working with non-Gaussian OU processes. The first is to specify the marginal distribution D of $v(t)$, which has to be self-decomposable, and then find the Lévy process $z(t)$ for which the solution to (6.10) is distributed as D. Barndorff-Nielsen and Shephard (2001) term these D-OU processes. Alternatively, we can choose the Lévy process $z(t)$, by specifying its Lévy measure $W(\cdot)$ (see Section 5.8 for a definition of W) or equivalently its (infinitely divisible) distribution D at

time 1, and then find the corresponding marginal for $v(t)$. In fact,

$$\int_1^\infty \log(x) W(dx) < \infty$$

has to be satisfied for (6.10) to have a stationary solution. The resulting process is termed OU-D. (D is a generic notation, it is not assumed that the same distribution D is used in the D-OU and the OU-D construction.) In discrete-time these approaches correspond to choosing either the marginal or the error distribution in (6.12).

Assuming the existence of densities w, u for the measures W and U respectively,

$$w(x) = -u(x) - xu'(x) \tag{6.13}$$

(see formula (15) of Barndorff-Nielsen and Shephard (2001)) assuming differentiability of $u(x)$. This simple formula makes both modelling approaches equally tractable. However, empirical studies (see for example Andersen et al. (2001)) provide some information about possible forms for the stationary distribution of the variance, therefore the D-OU approach is very attractive. Nevertheless, most of the empirical evidence refers to the actual volatilities, not to the volatility process directly, an issue that will be discussed in Section 6.3.2. For details on these constructions and explicit forms for the cumulants of $v(t)$ in various situations see Section 2 of Barndorff-Nielsen and Shephard (2001) and Chapter 4 of Barndorff-Nielsen and Shephard (2004).

A simulation from the gamma-OU model is shown in Figure 6.6. It can be easily shown using (6.13) that the Lévy density of $z(1)$ is integrable for this model, thus the Lévy process is a compound Poisson process (see Section 5.8). Therefore simulation from this process proceeds in a straightforward way, by first simulating the compound Poisson process and then using (6.25); see Section 6.4 for details.

6.3.2 Integrated volatility

The integrated volatility process was defined in (6.5). For the OU SV processes it has a very simple form

$$v^*(0,t) = \frac{1}{\mu}\{z(\mu t) - v(t) + v(0)\} \tag{6.14}$$

which can be derived directly from (6.9). Although both $z(t)$ and $v(t)$ are jump processes, $v^*(0,t),\ t \geq 0$, has continuous sample paths, see Barndorff-Nielsen and Shephard (2003). Chapter 5 of Barndorff-Nielsen and Shephard (2004) and Barndorff-Nielsen and Shephard (2003) contain many results about closed forms of conditional (on $v(0)$) and unconditional cumulants of the integrated volatility process. These quantities are very important in financial mathematics, for example in pricing of derivatives, and in the study of the marginal

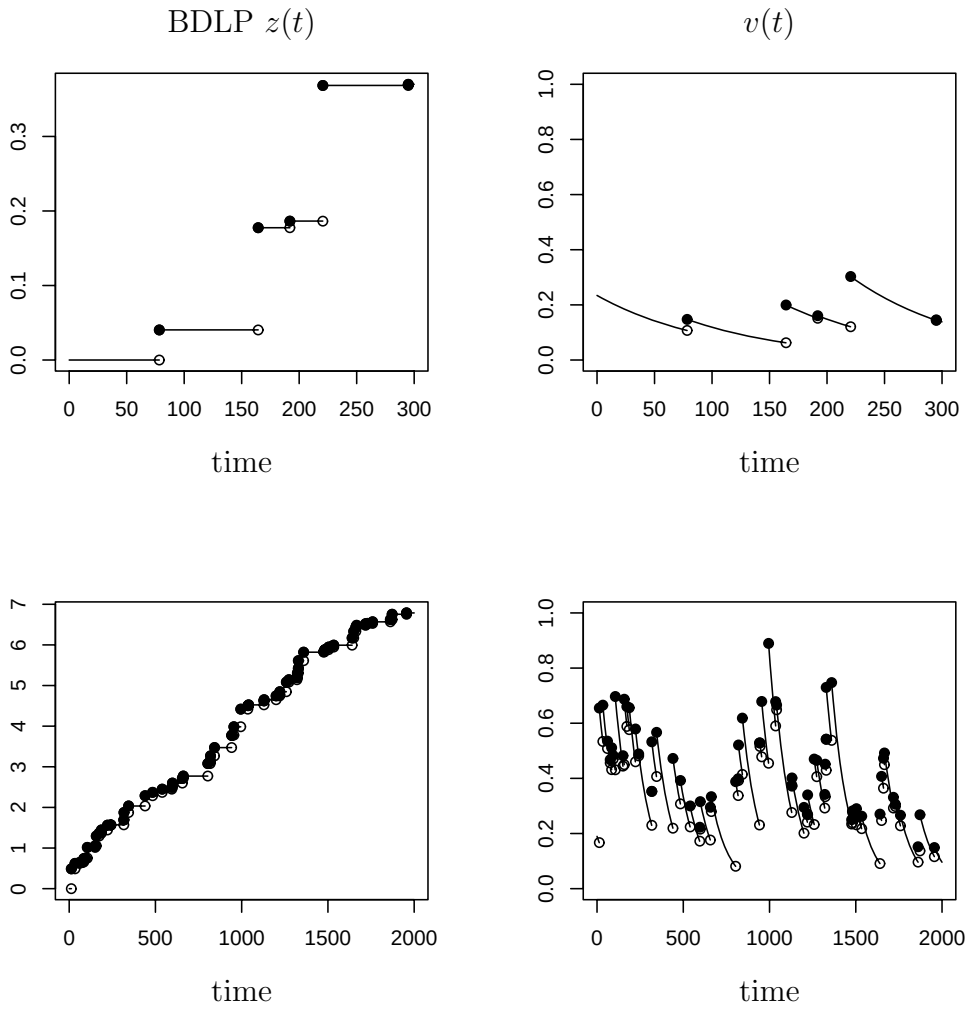


Figure 6.6: Simulation of the BDLP $z(\cdot)$ (left column)and the corresponding OU process (right column), for the gamma-OU model. Here we have taken $v(t) \sim \text{Ga}(3, 8.5)$ and $\mu = 0.01$. Top panel concentrates on a short time horizon, while the bottom panel shows simulation for a much longer period.

distributions of the log-returns.

Section 6.1 of Barndorff-Nielsen and Shephard (2001) and Chapter 6 of Barndorff-Nielsen and Shephard (2004) link SV with subordination. In particular, they introduce the notion of a chronometer, defined to be a positive, non-decreasing process starting from 0 at time 0. Such processes can be used as a random time-change in the Brownian motion, a procedure known as subordination. The subordinators introduced in Section 5.8 are an example of chronometers with independent increments. On the other hand, the integrated volatility process (6.14) is an example of a chronometer with continuous sample paths and dependent increments. It can be shown, that when $\alpha = \beta = 0$ in (6.3) the SV model can be written equivalently as a subordinated Brownian motion, that is

$$x(t) = B(v^*(0,t)).$$

The interpretation of SV as a random time-change is empirically appealing, since financial markets sometimes seem to speed up, an observation that goes at least as back as Taylor (1986) (but see Chapter 2 of that book for more and even earlier references). There are also mathematical conveniences associated with this interpretation, since for example it can be used to show that the sample paths of the log-price $x(t)$ in an SV model are continuous. For an exposition of subordination in the context of SV see Chapter 6 of Barndorff-Nielsen and Shephard (2004), for a brief discussion and references of its role in the problem of local time for Brownian motion see Section 8.4 of Kingman (1993).

An interesting question (which was raised in the discussion of Barndorff-Nielsen and Shephard (2001)) is, how 'close' is the distribution of integrals of the volatility, for example over a single day, to the distribution of the instantaneous volatility. This point was highlighted earlier, when we noted that there is some empirical evidence concerning such integrated quantities, see for example Andersen et al. (2001) for a study of the distribution of daily exchange rate actual volatility. This problem is investigated in Section 3 of Barndorff-Nielsen and Shephard (2003), where it is found that inverse Gaussian OU processes yield actual volatilities that have distributions which can be well approximated by the inverse Gaussian. On the contrary, actual volatilities of log-normal OU processes do not behave like log-normal variables.

The integrated volatility processes corresponding to the OU processes plotted in Figure 6.6, are shown in Figure 6.7. It can be seen that they have continuous sample paths, although their gradient changes in a discontinuous manner.

6.3.3 Aggregation results

Section 4 of Barndorff-Nielsen and Shephard (2001) and Chapter 5 of Barndorff-Nielsen and Shephard (2004) derive analytic results concerning the marginal distribution and the

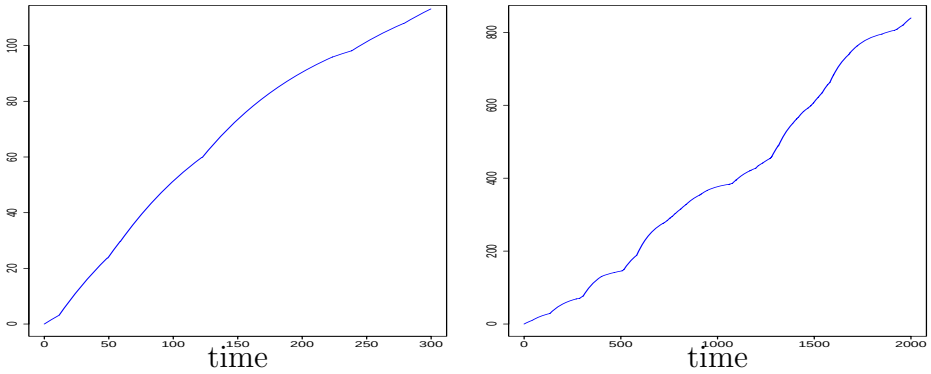


Figure 6.7: The integrated volatility processes corresponding to the OU processes plotted in Figure 6.6.

dependence structure of the log-returns series (6.4). They treat general SV models, but particularly simple forms arise when considering non-Gaussian OU models. Of great importance for analytic calculations is the derivation of the cumulant of the stochastic integral

$$\int_0^\infty f(s)dx(s)$$

for arbitrary functions, which is achieved in formulae (36)-(38) of Barndorff-Nielsen and Shephard (2001).

They also show that if ξ, ω^2 are the mean and the variance of $v(t)$ respectively, and the volatility process has an exponentially decaying autocorrelation function as in (6.8), then

$$\begin{aligned} \text{Corr}(v_n^*, v_{n+s}^*) &= d\exp\{-\mu\Delta(s-1)\} \\ \text{Corr}(y_n^2, y_{n+s}^2) &= c\exp\{-\mu\Delta(s-1)\},\ s>0 \end{aligned} \tag{6.15}$$

where

$$\begin{aligned} d &= \frac{(1-\exp\{-\mu\Delta\})^2}{2(\exp(-\mu\Delta)-1+\mu\Delta)} \\ c &= \frac{(1-\exp\{-\mu\Delta\})^2}{6(\exp(-\mu\Delta)-1+\mu\Delta)+2(\mu\Delta)^2(\xi/\omega^2)} \\ 1 \geq d &\geq c \geq 0. \end{aligned}$$

These expressions are useful when estimating these models (see Section 6.3.5) but they also imply that, if we believe in the SV model, the observed autocorrelation in the squared returns is smaller than the autocorrelation of the actual volatility series.

6.3.4 Superposition of OU processes

Barndorff-Nielsen and Shephard have shown (see for example Barndorff-Nielsen and Shephard (2001) and in particular Barndorff-Nielsen and Shephard (2002a)) that the OU SV models described in Section 6.3.1 are able to capture many of the marginal stylised characteristics of financial returns, while retaining mathematical tractability. However, the dependence structure observed in financial time series is more complicated than that implied by the model and is given in (6.15). It is typically observed a fast initial decay of the autocorrelation in the squared or absolute returns series, followed by a slower decay for many lags; see also Figure 6.5. Instead, the OU model implies an exponential decay for the autocorrelation of squared returns.

A way to construct models which capture this dependence structure without sacrificing analytic tractability, is by superimposing OU processes as suggested in Section 3 of Barndorff-Nielsen and Shephard (2001). In particular, we write for some $m > 0$

$$v(t) = \sum_{i=1}^{m} v_i(t) \tag{6.16}$$

with

$$dv_i(t) = -\mu_i v_i(t)dt + dz_i(t) \tag{6.17}$$

where $z_i(\cdot)$ are independent Lévy processes. The autocorrelation function of the volatility is now

$$r(t) = \sum_{j=1}^{m} w_j e^{-\mu_j t} \text{ where } w_j = \frac{\text{Var}(v_j(t))}{\sum_{i=1}^{m} \text{Var}(v_i(t))}. \tag{6.18}$$

The integrated volatility process is simply

$$v^*(0,t) = \sum_{j=1}^{m} v_j^*(0,t) \tag{6.19}$$

where each v_j^* is obtained from (6.14) in an obvious way.

For example, with $m = 2$, we can capture both short-term variation, represented by an OU process with high decaying rate μ_1 and also long-term movements in the volatility modelled as an OU processes with smaller decaying rate. Therefore we typically take $\mu_2 < \mu_1$. The empirical findings of Barndorff-Nielsen and Shephard (2002a) suggest that a superposition of two OU processes seems to be enough to model financial data, thus we will restrict attention to models where $m = 2$ in (6.16).

The distribution of the z_is might not be the same among $i = 1, \ldots, m$, i.e the OU processes might have different stationary distributions. However, it is convenient and possible

to construct models where v and the v_is have stationary distributions in the same family. For example, we can construct a gamma-OU model which is the superposition of m processes $v_i(t) \sim \text{Ga}(\nu_i, \theta),\ i = 1, \ldots, m$, where $\sum_{i=1}^{m} \nu_i = \nu$. The same modelling approach can be adopted for processes with inverse Gaussian marginal laws.

6.3.5 Existing estimation methods

Here we review some of the existing methods which can be used for inference about the non-Gaussian OU SV models. We argue that our preferred likelihood-based inference is a very challenging problem which demands involved computer intensive methodology. We will address this problem in the following sections where we show how the reparameterisation strategies developed in Chapter 5 can be used in this context. We begin by reviewing the second-order estimation sketched in Barndorff-Nielsen and Shephard (2001) and described in much greater detail in Barndorff-Nielsen and Shephard (2002a). We then discuss the problems encountered when considering likelihood-based inference.

Second-order estimation

Barndorff-Nielsen and Shephard have developed second-order methods for estimating the underlying volatility and its parameters in rather general SV models. The most relevant reference is Barndorff-Nielsen and Shephard (2002a).

The starting point is that the integrated volatility process $v^*(0,t)$ can, in theory, be recovered entirely using continuous observations from the price process $x(t)$, using the well-known quadratic variation identity from stochastic analysis:

$$\text{p-lim}_{q\to\infty} \sum \{x(t_{i+1}^q) - x(t_i^q)\}^2 = v^*(0,t). \tag{6.20}$$

The limit in (6.20) is in probability and holds for any sequence of partitions $t_0^q = 0 < t_1^q < \cdots < t_{m_r}^q = t$ with $\sup_i(t_{i+1}^q - t_i^q) \to 0$ as $q \to \infty$. This is a powerful "non-parametric" identity, which, however, cannot be used in practice. Even when extremely finely-spaced data exist for a financial asset, the general SV model would then be a poor approximation to its dynamic structure, due to market micro-structure effects. For instance, the assumed continuity of sample paths for $x(t)$ would easily collapse, when looking at prices obtained in very short intervals of time even for thickly traded assets.

Instead, Barndorff-Nielsen and Shephard (2002a) recognise that (6.20) can be used as an indication that the actual, daily say, volatility can be estimated using sums of squared returns over short periods of time, for example where $\Delta = 1$ is in the range of five minutes to a few hours. Nevertheless, there is error in this estimation and they study the exact second-order properties of this error and its asymptotic distribution.

More precisely, suppose that M intra-day observations exist for each day, for example $M = 288$ corresponds to five-minute transaction data. The realised volatility series is defined as

$$\{y\}_n = \sum_{j=1}^{M} \{x[(n-1)\Delta + \Delta j/M] - x[(n-1)\Delta + \Delta(j-1)/M]\}^2 .$$

For $M = 1$ this coincides with the squared daily log-returns series. Then we write

$$\{y\}_n = v_n^* + u_n, \text{ where } u_n = \{y\}_n - v_n^*. \tag{6.21}$$

The series $\{u_n\}$ contains the errors in the estimation of v_n^* by $\{y\}_n$ and it is not hard to show that it is a weak white noise uncorrelated with (but not independent of) the actual volatility series $\{v_n^*\}$. It is also easy to show that $\{y\}_n$ is an unbiased and consistent estimator of v_n^*. Barndorff-Nielsen and Shephard (2002a) (Section 2) derive the exact second order structure of the error series $\{u_n\}$ for an arbitrary SV model when $\alpha = \beta = 0$ in (6.3). Moreover, they find the asymptotic distribution of this error, which is a mixed normal and is independent of α, β.

They have also derived (in Barndorff-Nielsen and Shephard (2001) and Barndorff-Nielsen and Shephard (2002a)) the second-order structure of volatility processes with autocorrelation function as in (6.8). In particular, it corresponds to a (constrained) ARMA model. We note here that other SV models share this correlation structure, for example the constant elasticity of variance model (see for example Section 2.2 of Barndorff-Nielsen and Shephard (2002a) for a description and references).

Knowledge of the second-order properties of the "signal" process $\{v_n^*\}$ and the noise process $\{u_n\}$ is enough to allow the adoption of the Kalman filter to provide unbiased and efficient estimations of the actual volatilities by prediction and smoothing. As a by-product quasi-likelihood estimation of the parameters of the SV model is feasible. These methods can easily be extended to cover the superposition of OU models. Details can be found in Section 3 of Barndorff-Nielsen and Shephard (2002a). Section 5 of the same paper discusses some ideas about the estimation of α, β when they are assumed to be unknown.

Likelihood-based inference

We will consider here, and in the remainder of this chapter until Section 6.17, that $\alpha = \beta = 0$ in (6.3). The main difficulty is in estimating the parameters of the volatility process and any techniques developed for this purpose can be easily extended to infer for α, β. We will discuss likelihood inference only for the single-OU SV models here. The main message is that this is already very challenging and many of the existing computational techniques are inappropriate to handle this problem. Inference for superposition of processes is much more complicated. The parameters of interest are the stationary parameters of the volatility

and its memory parameter μ. To simplify exposition and notation we will assume that the marginal distribution of the volatility is described by two parameters only, ξ, ω^2 (defined in Section 6.3.3).

The data are discrete observations from the logarithmic price process $X = \{x(t_1), \ldots x(t_n)\}$ at (possibly irregularly spaced) time points $0 = t_1 < \ldots < t_n = T$, where we take $x(0) = 0$. According to the general SV model,

$$x(t_i) - x(t_{i-1}) \mid v^*(t_{i-1}, t_i) \sim \mathrm{N}(0, v^*(t_{i-1}, t_i)) \tag{6.22}$$

and therefore the conditional density of the data given the integrated volatility process is

$$\pi(X \mid v^*) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi v^*(t_{i-1}, t_i)}} \exp\left\{ -\frac{(x(t_i) - x(t_{i-1}))^2}{2v^*(t_{i-1}, t_i)} \right\}. \tag{6.23}$$

Likelihood-based inference requires the marginal likelihood $\pi(X \mid \mu, \xi, \omega)$ which is obtained by integrating (6.23) with respect to the prior measure of the integrated volatilities. In the OU SV model this prior distribution is defined by (6.7), (6.14) and the specification of the BDLP $z(\cdot)$. This integration is neither analytically nor numerically feasible, since it is over a highly dimensional space.

The solution to these problems, already outlined in Chapter 1 and Chapter 2 is the data augmentation and related Gibbs sampling methods. In particular, suppose that we add to the SV hierarchical model one more level, by assigning a prior distribution on the parameters (μ, ξ, ω). Then the well-rehearsed two-component Gibbs sampler alternates by updating the collection of integrated volatilities $\{v^*(t_{i-1}, t_i),\ i = 1, \ldots, n\}$ conditional on the data and the parameters, and the parameters given the integrated volatilities. According to the terminology of Section 2.2 this is a centered parameterisation, since the data and the parameters are independent conditionally on the augmented data, which are the integrated volatilities.

This augmentation scheme has proved very successful in the context of log-Gaussian SV models, especially when used in conjunction with efficient methods for updating the volatilities conditionally on the parameters all in one block (rather than component-wise). For a review of this methodology see Kim et al. (1998), but also Section 6.2 of this chapter for more references.

Chapter 2 and Chapter 4 argued extensively about potential convergence problems of the Gibbs sampler under a centered parameterisation. The augmentation scheme described above for OU SV models is one more example where centered methods can have very poor convergence properties due to the strong dependence between the missing data and the parameters, in particular between μ and the integrated volatilities. This was first observed in Barndorff-Nielsen and Shephard (2001), where it was argued that knowledge of the integrated

volatilities essentially determines uniquely μ in the single-OU model. Some theoretical justification is given, based on the related work of Nielsen and Shephard (2003) on the properties of the MLE of the memory parameter in auto-regressions with exponential innovations. They show that the MLE estimator is consistent with its standard deviation going to zero as the sample size $n \to \infty$ as n^{-1} (whereas the rate is $n^{1/2}$ for Gaussian auto-regressions). This is known as super-consistency. Moreover, the likelihood is zero for all values of the memory parameter bigger than the MLE. These results are relevant to the OU case, since when it is discretised it has a similar autoregressive structure (especially when the time intervals $t_i - t_{i-1}$ are small). Thus the information contained in the imputed data about μ greatly exceeds that contained in the observed data and consequently the data augmentation has been found (see p.189 of Barndorff-Nielsen and Shephard (2001)) to have extremely poor performance in exploring the posterior distribution.

The following sections develop centered and non-centered parameterisations which are based on a different augmentation scheme from the one suggested above, and which avoid the problem of super-consistency. Our augmentation is based on the Poisson process representation of the BDLP; see Section 5.8 for a discussion on these representations.

6.4 OU models with compound Poisson BDLP

This chapter focuses on OU models for which the BDLP is a compound Poisson process. Compound Poisson processes are characterised by integrable Lévy measures and admit representations by means of locally finite Poisson processes (see Section 5.8). This chapter considers inference for this class of models only. This is done mainly due to simplicity, since it is easier to handle and to transform locally finite Poisson processes using the techniques developed in Chapter 5. Nevertheless, our methodology could be extended to cover BDLPs with non-integrable Lévy measures, which is briefly discussed in Section 6.17.

Compound Poisson processes were introduced in Section 5.1.3. In Section 5.8 they were classified as a special case of subordinators which can be represented as

$$z(t) = \sum_{j=1}^{N(t)} E_j, \;\; z(0) := 0 \tag{6.24}$$

where $N(t)$ is the number of arrivals of a Poisson process with finite rate, λ say, in $[0, t]$ and the E_js are IID random variables and also independent from the Poisson process. Since we are restricting our attention to positive processes we will assume that the jumps of the process E_j are positive random variables and denote their common distribution by $F(\cdot)$. We denote the Poisson process arrivals in $[0, t]$ as $C_1, C_2, \ldots, C_{N(t)}$. An example of a compound Poisson process is shown in the left column of Figure 5.2.

We have already encountered an example of an OU process driven by a compound Poisson process. Section 6.3.1 introduced the gamma-OU process and Figure 6.6 shows simulations from this process. The marginal distribution of the volatility in this model is $\text{Ga}(\nu, \theta)$ with mean $\xi = \nu/\theta$ and variance $\omega^2 = \nu/\theta^2$. It is not hard to show (using (6.13)) that the BDLP is a compound Poisson process, details can be found in Section 6.5. The gamma is the only distribution in the generalised inverse Gaussian family for which the BDLP of the D-OU process is compound Poisson. Nevertheless it is straightforward to construct OU-D processes by specifying the distribution of the jumps E_j.

When the BDLP is compound Poisson, the instantaneous and integrated volatility processes have simple expressions in terms of μ, the (unknown) initial volatility $v(0)$, the jump times and corresponding sizes of $z(\cdot)$ in (6.24). Specifically

$$v(t) = e^{-\mu t} v(0) + \sum_{j=1}^{N(t)} e^{-\mu(t-C_j)} E_j \tag{6.25}$$

and the integrated volatility becomes

$$\begin{aligned} v^*(0,t) &= \frac{1}{\mu} \sum_{j=1}^{N(t)} E_j - \frac{1}{\mu}(v(t) - v(0)) \\ &= \frac{1}{\mu} \sum_{j=1}^{N(t)} E_j - \frac{1}{\mu} \left(v(0)(e^{-\mu t} - 1) + \sum_{j=1}^{N(t)} e^{-\mu(t-C_j)} E_j \right). \end{aligned} \tag{6.26}$$

Notice that this specification forces the volatility to move up entirely by jumps and then to tail off exponentially. This feature is illustrated in Figure 6.6 and in terms of financial modelling can be thought of as new information arriving in packets increasing the transactions variability.

Section 5.8 established the connection between Lévy and Poisson processes, in particular, we showed that each subordinator $z(t),\ t \in [0,T]$ can be constructed in terms of a Poisson process on $S = [0,T] \times (0,\infty)$, whose mean measure $\Lambda(\cdot)$ is given by the product measure

$$\Lambda(dc \times d\epsilon) = W(d\epsilon)dc,\ c \in [0,T], \epsilon > 0$$

where $W(\cdot)$ is the Lévy measure. When the Lévy process is a compound Poisson its Lévy measure is given by

$$W(d\epsilon) = \lambda F(d\epsilon); \tag{6.27}$$

see Section 5.8 and specifically expression (5.39). Let Ψ be the Poisson process on S corre-

sponding to the subordinator (6.24), thus

$$\Psi = \{(C_j, E_j), j = 1, 2, \ldots\}$$

where the C_js and E_js are defined by (6.24), and its mean measure is

$$\Lambda(dc \times d\epsilon) = \lambda F(d\epsilon)dc, \ (c, \epsilon) \in S. \tag{6.28}$$

The relationship between Ψ and the BDLP is depicted in Figure 5.2.

6.4.1 Superposition of OU models with compound Poisson BDLP

Section 6.3.4 showed that more complicated dependence structures in the data can be captured by modelling the volatility process as the superposition of a number of independent OU processes as described by (6.16) and (6.17).

In this section we assume that the BDLP in each of the components of the superposition is a compound Poisson process. Therefore for $i = 1, \ldots, m$

$$v_i(t) = e^{-\mu_i t} v_i(0) + \sum_{j=1}^{N_i(t)} e^{-\mu_i (t - C_{ij})} E_{ij} \tag{6.29}$$

where $C_{i1} < \ldots < C_{iN_i(t)}$ are the arrival times of a Poisson process in $[0, t]$ with rate λ_i, $N_i(t)$ is the corresponding number of arrivals, E_{ij} are IID positive random variables with distribution $F_i(\cdot)$ and $v_i(0)$ are assumed to be random and distributed according to the stationary distribution of the $v_i(t)$ process. The integrated volatility process was shown in (6.19) to satisfy

$$v^*(0, t) = \sum_{i=1}^{m} v_i^*(0, t)$$

where the $v_i^*(0, t)$ can be derived from (6.26) in an obvious way.

To each of the background driving compound Poisson processes there corresponds a Poisson process

$$\Psi_i = \{(C_{ij}, E_{ij}), j = 1, 2, \ldots\}$$

with mean measure

$$\Lambda_i(dc \times d\epsilon) = \lambda_i F_i(d\epsilon)dc, \ (c, \epsilon) \in S. \tag{6.30}$$

The empirical findings of Barndorff-Nielsen and Shephard (2002a) suggest that a superposition of two OU processes seems to be enough to model financial data, therefore we will not consider more than two components in this thesis.

6.5 Bayesian inference for the gamma-OU model

We now formalise the problem of Bayesian inference for the gamma-OU SV model. Actually, the methodology developed in the following sections can be immediately extended to other models with background driving compound Poisson processes. Nevertheless, for reasons of clarity and exposition, as well as because it is by far the most widely used model with integrable Lévy measure, we concentrate on the gamma-OU process. The single-OU model is initially considered while Section 6.5.1 formalises the inference problem for the superposition of gamma-OU processes.

The gamma-OU process has gamma marginal distribution $\mathrm{Ga}(\nu, \theta)$ with mean $\xi = \nu/\theta$ and variance $\omega^2 = \nu/\theta^2$. We remarked earlier that the BDLP of this process has integrable Lévy measure. In fact, it is easy to show using (6.13) that

$$W(d\epsilon) = \lambda\theta e^{-\theta\epsilon} d\epsilon$$

where $\lambda = \nu\mu$. This Lévy measure was shown in Section 5.8.1 (see (5.44)) to correspond to a compound Poisson process for which $E_j \sim \mathrm{Ex}(\theta)$ and the jump times C_j form a Poisson process in time with intensity λ. The volatility process (6.25) driven by this compound Poisson process is known as the shot-noise continuous time Markov process; see for example Cox and Isham (1980). From (6.30) follows that the mean measure of Ψ is

$$\Lambda(dc \times d\epsilon) = \lambda\theta e^{-\theta\epsilon} dc\, d\epsilon. \tag{6.31}$$

The Poisson process with this mean measure was the subject of study in Section 5.5.1, where, different marked Poisson process representations where suggested for Ψ. Following the suggestion of Section 6.3.1 we assume that $v(0)$ has a $\mathrm{Ga}(\nu, \theta)$ prior distribution.

The parameters of interest are ν, θ and μ. In the sequel, for notational convenience we will work with both the (ν, θ, μ) and the $(\lambda = \nu\mu, \theta, \mu)$ parameterisation. When discussing missing data transformations (in Section 6.10 and Section 6.11) the latter is more natural, but when referring to posterior inference the former will be of interest.

6.5.1 Superposition of gamma-OU processes

We will also consider the model where the volatility is the superposition of two gamma-OU processes. The main purpose of superpositioning OU processes is to model the dependence structure and not the stationary distribution of the variance. Therefore, as suggested in Section 6.3.4 we exploit the infinite divisibility of the gamma distribution, and assume that $v_i(\cdot) \sim \mathrm{Ga}(\nu_i, \theta)$, thus

$$v(t) \sim \mathrm{Ga}(\nu, \theta),\ \nu = \sum_{i=1}^{2} \nu_i.$$

In the representation given in (6.29) for each $v_i(t)$, $\lambda_i = \nu_i \mu_i$, the jump sizes all are IID from an $\text{Ex}(\theta)$ and $v_i(0) \sim \text{Ga}(\nu_i, \theta)$.

The parameters of interest are $\nu_1, \nu_2, \mu_1, \mu_2, \theta$.

6.5.2 Prior specification and posterior inference

For the single-OU model, we choose a $\text{Ga}(\alpha_\theta, \beta_\theta)$ prior for θ, a $\text{Ga}(\alpha_\mu, \beta_\mu)$ prior for μ and also a $\text{Ga}(\alpha_\nu, \beta_\nu)$ prior for ν. These priors are adopted mainly for simplicity and computational convenience (as we will see in the sequel). However, a discussion on the choice of the hyperparameters and more general prior sensitivity issues will be made in Section 6.14.

The prior elicitation for the superposition of OU processes needs more careful consideration, see also Section 6.14. As a general strategy, we choose to order the memory parameters *apriori*, that is we assume that $\mu_1 > \mu_2$ *almost surely*. When both the single-OU and the two-OU models are applied to a dataset a prior specification for the latter which is consistent with the one suggested above for the former, is the following. We parameterise in terms of $(\nu, w_2, \theta, \mu_1, \mu_2)$, where $\nu = \nu_1 + \nu_2$ and $w_2 = \nu_2/(\nu_1 + \nu_2)$ (see (6.18)). Under this setting (ν, θ) are the parameters of the stationary distribution of the volatility in both models, and we specify a common prior. It is also reasonable to assume that $w_2 \sim \text{Un}[0, 1]$. Under this scenario, the joint prior density for (ν_1, ν_2) becomes

$$\pi(\nu_1, \nu_2) \propto (2\nu_2 + \nu_1)(\nu_1 + \nu_2)^{\alpha_\nu - 3} \exp\{-\beta_\nu(\nu_1 + \nu_2)\}. \tag{6.32}$$

We construct the joint prior of (μ_1, μ_2) by imposing a $\text{Ga}(\alpha_{\mu_2}, \beta_{\mu_2})$ to μ_2 and assuming that $\mu_1 - \mu_2$ given μ_2 is a $\text{Ga}(\alpha_{\mu_1}, \beta_{\mu_1})$ random variable. Therefore prior elicitation on (μ_1, μ_2) is done by specifying the marginal prior distribution of μ_2 and by quantifying our beliefs about how much bigger μ_1 is expected to be than μ_2. This prior choice is adopted in the simulated examples in Section 6.13.1 but also in analysis of the exchange rates data in Section 6.16. We also note that the gamma prior chosen for θ is computationally convenient, since it leads to conditional conjugacy (see Section 6.13).

We are interested in obtaining either in closed form or samples from the posterior distribution of the parameters, that is from

$$\pi(\nu, \theta, \mu \mid X) \propto \pi(X \mid \nu, \theta, \mu)\pi(\nu, \theta, \mu)$$

in the single-OU model and from

$$\pi(\nu_1, \nu_2, \mu_1, \mu_2, \theta \mid X) \propto \pi(X \mid \nu_1, \nu_2, \mu_1, \mu_2, \theta)\pi(\nu_1, \nu_2, \mu_1, \mu_2, \theta)\mathbb{1}[\mu_1 > \mu_2]$$

in the two-OU model. This presupposes the existence of the marginal likelihoods $\pi(X \mid$

$\nu, \theta, \mu)$ and $\pi(X \mid \nu_1, \nu_2, \mu_1, \mu_2, \theta)$ in closed form, but Section 6.3.5 showed that these quantities are not available. Therefore we resort to MCMC methods and in particular data augmentation type techniques, in order to obtain posterior samples. These augmentation methods are described in the next section for the single-OU model, while Section 6.13 develops MCMC methods for posterior inference for the two-OU model.

6.6 Augmentation based on marked Poisson processes

This section develops a missing data methodology to tackle the inference problem for the gamma-OU model of Section 6.5.

It was observed in Section 6.4 that the integrated volatility process is a simple function of the decaying rate of the OU process μ, the (random) initial volatility $v(0)$ and the latent Poisson process Ψ. Thus by (6.26)

$$\pi(X \mid v^*) = \pi(X \mid \mu, v(0), \Psi).$$

Specialising the argument made in Section 6.3.5 to the case where we have a background driving compound Poisson process

$$\pi(X \mid \nu, \theta, \mu) \quad = \quad \int \pi(X \mid \mu, v(0), \Psi)\pi(v(0) \mid \nu, \theta)\pi(d\Psi \mid \nu, \theta, \mu)dv(0), \qquad (6.33)$$

where $\pi(d\Psi \mid \nu, \theta, \mu)$ denotes the measure of the Poisson process Ψ; see Section 5.1.4 and later in this section.

Section 6.3.5 argued that it is infeasible to obtain the likelihood function $\pi(X \mid \nu, \theta, \mu)$ either explicitly or numerically, due to the integration required in (6.33). Therefore we are dealing with the typical situation in the so-called missing data problems, which were introduced in Section 1.3: the likelihood is obtained through an integration over the distribution of some random objects, Ψ and $v(0)$ in (6.33). Therefore, the data augmentation methodology suggests treating Ψ and $v(0)$ as missing data, and thus making use of Gibbs sampling techniques to sample from the joint posterior distribution of the parameters (ν, θ, μ) and missing data; see for example Tanner and Wong (1987), Smith and Roberts (1993) and Section 1.5.2 of this thesis. This is the data augmentation scheme proposed by Roberts et al. (2003) in the context of the gamma-OU models.

Before discussing different parameterisations for this augmentation scheme, we resolve one technical issue raised in the previous paragraph concerning the measure of the Poisson process Ψ. Section 5.1.4 gave the measure-theoretic background for constructing measures on point process spaces, therefore we refer to that section for any details. We recall that Ψ is a Poisson process on S with mean measure given in (6.31) where $\lambda = \nu\mu$. We choose as a

basic reference measure $\mathcal{Q}$ that of a Poisson process on S with mean measure

$$e^{-\epsilon} dc \, d\epsilon. \tag{6.34}$$

By means of Lemma 5.1.4

$$\pi(d\Psi \mid \lambda, \theta) = e^{-(\lambda-1)T} (\lambda\theta)^{\Psi(S)} \exp\left\{ -(\theta - 1) \sum_{j=1}^{\Psi(S)} E_j \right\} \mathcal{Q}(d\Psi) \tag{6.35}$$

where (see Section 5.1.4) $\Psi(S)$ is the number of points of Ψ on S, and the sum is replaced by 0 if $\Psi(S) = 0$. It is now clear that the density of Ψ is

$$\pi(\Psi \mid \lambda, \theta) = e^{-(\lambda-1)T} (\lambda\theta)^{\Psi(S)} \exp\left\{ -(\theta - 1) \sum_{j=1}^{\Psi(S)} E_j \right\} \tag{6.36}$$

or $\pi(\Psi \mid \lambda, \theta) = \exp\{-(\lambda - 1)T\}$ if $\Psi(S) = 0$.

6.7 A centered parameterisation

A straightforward parameterisation can be constructed by writing the posterior distribution of the parameters and the missing data as

$$\begin{aligned} \pi(\nu, \theta, \mu, \Psi, v(0) \mid X) \propto & \pi(X \mid \mu, v(0), \Psi) \pi(v(0) \mid \nu, \theta) \pi(\Psi \mid \nu, \theta, \mu) \pi(\nu, \theta, \mu) \\ \propto & \pi(X \mid \mu, v(0), \Psi) \frac{\theta^\nu}{\Gamma(\nu)} v(0)^{\nu-1} e^{-\theta v(0) - (\nu\mu - 1)T} \\ & \times (\nu\mu\theta)^{\Psi(S)} \exp\left\{ -(\theta - 1) \sum_{j=1}^{\Psi(S)} E_j \right\} \pi(\nu, \theta, \mu) \end{aligned} \tag{6.37}$$

where $\pi(\nu, \theta, \mu)$ is the joint prior density and $\Gamma(\cdot)$ is the gamma function. The dependence structure between the parameters and the missing data is presented as a graphical model in Figure 6.8.

There are two key features of this parameterisation. Firstly, it is a non-centered parameterisation in terms of the missing data and μ and therefore it circumvents the main drawback of the augmentation scheme described in Section 6.3.5. There, the missing data were taken to be the integrated volatilities conditionally on which μ was independent of the observed data. This was shown to be very inefficient because the augmented information about μ greatly exceeded the observed one. In fact, all the parameterisations we will consider are non-centered as far as μ and the missing data are concerned.

On the other hand, the second important feature of this parameterisation is that the

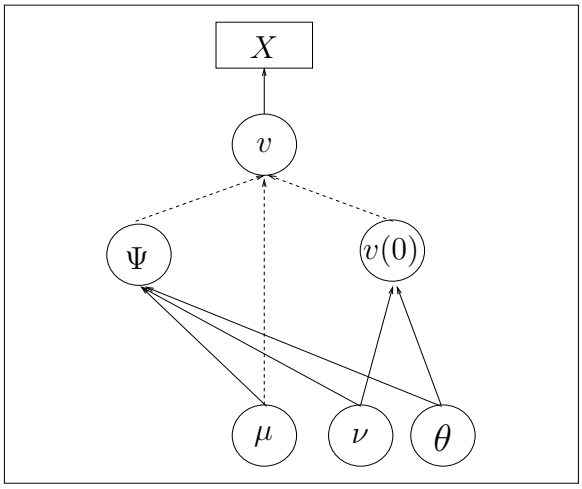


Figure 6.8: Graphical model of the first (centered) parameterisation

parameters ν and θ are conditionally independent of the data given μ and the missing data Ψ, $v(0)$. Put in another way, λ is independent of the observed given the missing data. It is therefore centered in that respect, and we will call it the centered parameterisation. This terminology is used to distinguish it from alternative parameterisations that we will consider in the following sections for which the transformed missing data and λ lie at the same level of the hierarchy.

6.7.1 MCMC implementation

We use a Hastings-within-Gibbs algorithm (see Section 1.5.2) to obtain samples from the joint posterior distribution of the parameters and the missing data (6.37):

A Hastings algorithm to sample from $(\nu, \theta, \mu, \Psi, v(0)) \mid X$

Iterate the following steps:

1. Update (ν, θ) according to $\pi(\nu, \theta \mid \Psi, v(0), \mu)$
2. Update μ according to $\pi(\mu \mid X, \Psi, v(0), \nu, \theta)$
3. Update $v(0)$ according to $\pi(v(0) \mid X, \Psi, \nu, \mu, \theta)$
4. Update Ψ according to $\pi(\Psi \mid X, v(0), \nu, \mu, \theta)$.

The densities in steps $1-4$ are derived up to proportionality from (6.37). Direct simulation is not feasible in Steps $2-3$ but we can easily perform a Metropolis-Hastings updating scheme on the logarithmic scale for the parameters involved.

In Step 1 we exploit the fact the full conditional distribution of θ is known,

$$\theta \mid \cdot \sim \mathrm{Ga}\left(\Psi(S)+\nu+\alpha_\theta, \sum_{j=1}^{\Psi(S)} E_j + v(0) + \beta_\theta\right) \tag{6.38}$$

due to the conditional conjugacy of the gamma prior for θ. This is one of the computational advantages behind the choice of this prior, that we hinted at in Section 6.5.2. Consequently we can derive

$$\begin{aligned}\pi(\nu \mid v(0), \Psi, \mu) &\propto \frac{\Gamma(\nu+\alpha_\theta+\Psi(S))}{\Gamma(\nu)}\left(\frac{v(0)}{\beta_\theta + v(0) + \sum_{j=1}^{\Psi(S)} E_j}\right)^\nu \\ &\times \nu^{\Psi(S)+\alpha_\nu-1}\exp\{-(\beta_\nu+\mu T)\nu\}. \end{aligned}\tag{6.39}$$

We update (ν, θ) by first using a Metropolis-Hastings step (on the logarithmic scale) with target (6.39) and then conditionally on ν we simulate θ directly from (6.38). A less efficient scheme in the Peskun ordering (see Section 3 of Tierney (1998)) is to propose new values for ν and θ, first for ν from some proposal kernel and then for θ from (6.38) substituting ν with its proposed value, and then jointly accepting the move. The acceptance ratio is the same as the one for updating ν in the previous scheme.

Updating the conditional distribution of Ψ in Step 4 is more involved since it requires a Metropolis-Hastings step which operates on a point process space. In particular, we want to construct a Metropolis-Hastings Markov chain which has the conditional distribution of Ψ as a stationary measure. This distribution is characterised by its density

$$\pi(\Psi \mid \nu, \theta, \mu, v(0), X) \propto \pi(X \mid \mu, v(0), \Psi)\pi(\Psi \mid \nu, \theta, \mu) \tag{6.40}$$

with respect to the reference Poisson measure (see Section 6.6), where the first term is given in (6.23) and the second in (6.36).

We use the general methodology presented in Section 5.1.5 together with some specific ideas relevant in the gamma-OU context. At each iteration, we randomly choose between two types of moves. The first move type is a dimension-changing move that proposes either to add or to remove a point from the current configuration of the point process. The second move type selects one of the existing points at random and proposes to displace it. The probability of choosing a displacement move is fixed throughout the simulation and we choose it by pilot tuning.

Formally, let the current configuration of points be $\psi = \{(c_1, \epsilon_1), \ldots, (c_m, \epsilon_m)\}$ where $(c_i, \epsilon_i) \in S$, $i = 1, \ldots, m$.

Birth-or-death move

The details of this step can be found in Section 5.1.5. We choose the probability of a death move to be a half. If we choose a birth move, we propose to move to $\psi \cup \{(c, \epsilon)\}$, where $\{(c, \epsilon)\}$ is the new-born point. We generate it from the prior, therefore the proposal distribution has density with respect to the reference mean measure (6.34)

$$b(c, \epsilon) = \frac{1}{T}\theta \exp\{-(\theta - 1)\epsilon\}. \tag{6.41}$$

When we choose a death move, we propose to move to the configuration $\psi - \{(c, \epsilon)\}$ by removing $\{(c, \epsilon)\}$ uniformly among the existing points of ψ, that is

$$d(\psi - \{(c, \epsilon)\}, (c, \epsilon)) = 1/m \tag{6.42}$$

Therefore, by (6.41), (6.42), (6.40) and (5.11) the Metropolis-Hastings acceptance probability is

$$\alpha_{bd}(\psi, \psi \cup \{(c, \epsilon)\}) = \min\{1, r(\psi, c, \epsilon)\}$$

if $\pi(\psi \cup \{(c, \epsilon)\} \mid \nu, \theta, \mu, v(0), X) > 0$, and 0 otherwise, and

$$\alpha_{bd}(\psi \cup \{(c, \epsilon)\}, \psi) = \min\{1, 1/r(\psi, c, \epsilon)\}$$

where

$$\begin{aligned} r(\psi, c, \epsilon) &= \frac{\pi(\psi \cup \{(c, \epsilon)\} \mid \nu, \theta, \mu, v(0), X))}{\pi(\psi \mid \nu, \theta, \mu, v(0), X)} \frac{T}{(m+1)\theta \exp\{-(\theta - 1)\epsilon\}} \\ &= \frac{\pi(X \mid \mu, v(0), \psi \cup \{(c, \epsilon)\})}{\pi(X \mid \mu, v(0), \psi)} \frac{\nu\mu T}{m+1}. \end{aligned} \tag{6.43}$$

Thus the birth-or-death move proceeds as follows.

Birth-or-death move

Sample $U \sim \mathtt{Un}[0,1]$

If $U < 1/2$ then

    (birth) simulate $(c, \epsilon) \sim b(\cdot, \cdot)$

    move from ψ to $\psi \cup \{(c, \epsilon)\}$ with probability $\min\{1, r(\psi, (c, \epsilon))\}$

Else

    (death) choose $(c, \epsilon) \in \psi$ uniformly

    move from ψ to $\psi - \{(c, \epsilon)\}$ with probability $\min\{1, 1/r(\psi, (c, \epsilon))\}$.

Displacement move

We construct the displacement transition kernel as a mixture of m Metropolis-Hastings transition kernels; the ith kernel is reversible with respect to the conditional posterior distribution of the ith point $\{(c_i, \epsilon_i)\}$ given $\psi - \{(c_i, \epsilon_i)\}$. This can be seen as a distribution on S with Lebesgue density

$$\pi(c, \epsilon \mid X, \lambda, \theta, \mu, v(0), \psi - \{(c_i, \epsilon_i)\}) \propto$$
$$\pi(X \mid \mu, v(0), \psi - \{(c_i, \epsilon_i)\} \cup \{(c, \epsilon)\})\theta \exp\{-\theta\epsilon\}.$$

Each kernel is chosen with equal probability and it is adequate to describe how the ith kernel is constructed. We propose to move from ψ to $\psi - \{(c_i, \epsilon_i)\} \cup \{(c, \epsilon)\}$ and we suggest two strategies to generate $\{(c, \epsilon)\}$.

The first strategy uses independence sampling from the proposal density

$$q(c, \epsilon) = T^{-1}\theta \exp\{-\theta\epsilon\},\ \epsilon > 0,\ 0 < c < T. \tag{6.44}$$

The calculation of the acceptance ratio is straightforward combining (6.44) and (5.10). Nevertheless, we will not use this strategy in any of our examples.

The second is a strategy that achieves local change in the volatility process (see Figure 6.9) and is used in our MCMC programs throughout this chapter. Assume, without loss of generality, that $c_1 < c_2 < \ldots < c_m$. We generate c uniformly in $[c_{i-1}, c_{i+1}]$ with $c_0 := 0$ and

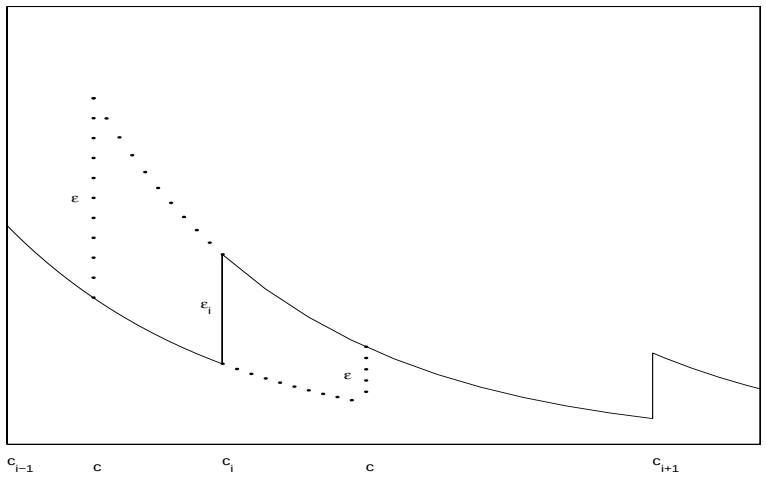


Figure 6.9: The local displacement of the point $\{(c_i, \epsilon_i)\}$. The new jump time c is generated uniformly in (c_{i-1}, c_{i+1}) and the new jump size ϵ is set so that the volatility process changes only between time points c and c_i. The new jump time c can lie either side of c_i; both options are shown in the diagram.

$c_{m+1} := T$, and we define the transformation

$$(\epsilon_i, c_i, c) \to (\epsilon, c, c_i) = (\epsilon_i \exp\{-\mu(c - c_i)\}, c, c_i).$$

This transformation is invertible and its Jacobian is $\exp\{\mu(c - c_i)\}$. Intuitively, the invertibility of the transformation can be seen by inspection of Figure 6.9. Thus, this updating scheme can be made reversible with respect to the joint posterior density of (c, ϵ). However, there is a slight technical obstacle in deriving the resulting Metropolis-Hastings acceptance ratio $\alpha_i[(c_i, \epsilon_i), (c, \epsilon)]$, since we are updating a two-dimensional distribution but there is only one degree of freedom. In other words, the proposal distribution is singular with respect to the target measure, in fact it lives on a slice of the parameter space S, defined by the function $c \to \epsilon_i \exp\{-\mu(c - c_i)\}$. This setting is usually encountered in reversible-jump MCMC algorithms and $\alpha_i[(c_i, \epsilon_i), (c, \epsilon)]$ can be calculated using the results of Green (1995). We favour the more general approach of Tierney (1998) (see also Section 1.5) for deriving such results, see Section 2 of his paper and specifically his Theorem 2 and Examples 2 and 3. It turns out that $\alpha_i[(c_i, \epsilon_i), (c, \epsilon)]$ is given by

$$\min\left\{1, \frac{\pi(X \mid \mu, v(0), \psi - \{(c_i, \epsilon_i)\} \cup \{(c, \epsilon)\})}{\pi(X \mid \mu, v(0), \psi)} \exp\{\theta(\epsilon_i - \epsilon) - \mu(c - c_i)\}\right\}.$$

Therefore, the displacement move proceeds as follows.

Displacement move

Sample i uniformly from $\{1, \ldots, m\}$

Simulate $c \sim \mathtt{Un}[c_{i-1}, c_{i+1}]$

Set $\epsilon = \epsilon_i \exp\{-\mu(c - c_i)\}$

Move from ψ to $\psi - \{(c_i, \epsilon_i)\} \cup \{(c, \epsilon)\}$ with probability $\alpha_i[(c_i, \epsilon_i), (c, \epsilon)]$.

Apart from the well-documented advantages of local moves in reversible-jump type algorithms (see for example Section 4 in Dellaportas et al. (2002)), this move results in increased computational efficiency since it requires evaluation of only a small part of the likelihood function (corresponding to the time between the current and proposed jump times).

To improve the mixing, at each iteration we simulate from the distribution of the jump sizes conditional on the jump times using a random walk Metropolis step. Since the dimension of this vector changes with iterations we set the variance of the proposal distribution to be inversely proportional to the the current number of jumps (see Roberts et al. (1997)).

6.8 Alternatives to the centered parameterisation

The parameterisation shown in Figure 6.8 is statistically natural, representing the hierarchy in which the parameters (ν, θ, μ) are used to construct the latent process Ψ, which in turn (together with μ and $v(0)$) determines the distribution of the observed data. In terms of $\lambda = \nu\mu$ (which is not explicitly shown in the graphical model) and Ψ it is a centered parameterisation.

Chapter 2 and Chapter 4 argued about the potential convergence problems of data augmentation under a centered parameterisation. In particular, we showed that these algorithms converge extremely slowly when the statistical information about the unknown parameters contained in the latent (imputed) data is considerably greater than that actually contained in the observed data.

In the hidden Markov process case (of which the model in Section 6.5 is a special case) problems caused by dependence between the parameters and the hidden process itself can be acute; see the relevant discussion in Section 5.2 in the context of latent Poisson processes. This is particularly problematic for long time series where the prior structure will be bounded by "ergodicity constraints" which link long term empirical properties of the hidden process

with their stationary expectations (which are just functions of the parameters). For our model, an example is the following

$$\frac{1}{T}\int_0^T v(s)\ ds \approx \frac{\lambda}{\theta\mu}.$$

Thus, unless the data are sufficiently informative to override this relationship, extremely high posterior correlation will exist between $\int_0^T v(s)\ ds$ and $\lambda/(\theta\mu)$, or equivalently between Ψ and λ, leading to very poor convergence of the centered algorithm proposed in Section 6.7.

This thesis has advocated the use of non-centered parameterisations to combat convergence problems as those described above. Section 6.9 applies the general methodology for non-centering Poisson processes developed in Chapter 5 in the context of the OU SV models.

Notice that the information of the observed data about the latent process increases

(i) the higher the number of data points per Poisson jump (small λ compared to data frequency)

(ii) the higher the persistence in the volatility (small μ)

(iii) the smaller the variance of the volatility, when the mean is kept fixed.

(ii) and (iii) are empirically supported by the simulation study in Barndorff-Nielsen and Shephard (2002a). Moreover, they are consistent with analytic convergence rate results available for Gaussian models, see Chapter 2 of this thesis, Roberts and Sahu (1997), Pitt and Shephard (1999) and Papaspiliopoulos et al. (2003) for a recent review. (ii) is intuitive, since the posterior for each "location" of the latent process will be influenced not only by the corresponding observed data point but also by the "neighbouring" ones. This phenomenon was highlighted in Section 2.5 in the context of Gaussian state-space models. On the other hand, the information about λ contained in the latent process Ψ, increases with T and decreases with $\Psi(S)$.

Section 6.12.1 investigates whether these characteristics are reflected in the behaviour of the centered and non-centered algorithms when applied to different simulated datasets.

6.9 Non-centering for the the gamma-OU model

We first recognise that the missing data Ψ is a Poisson process with mean measure given in (6.31), which depends on the parameters (ν, θ, μ). Moreover, it can be represented as a marked Poisson process in a variety of ways, we refer back to Section 5.5 for a summary of different available options for Poisson processes with product mean measures.

We wish to construct a non-centered parameterisation $(\lambda, \theta, \Psi) \rightarrow (\lambda, \theta, \tilde{\Psi})$. A further transformation, that turns out to be computationally very useful, as we will shortly see in

Section 6.10, and makes θ *a priori* independent of $v(0)$, is

$$v(0) \rightarrow \tilde{v}(0) = \theta v(0). \tag{6.45}$$

The conditional independence structure under this parameterisation can now be written as

$$\begin{aligned} \pi(\nu, \theta, \mu, \tilde{\Psi}, \tilde{v}(0) \mid X) &\propto \pi(X \mid \tilde{\Psi}, \tilde{v}(0), \nu, \mu, \theta)\pi(\tilde{v}(0) \mid \nu)\pi(\tilde{\Psi})\pi(\nu, \theta, \mu) && (6.46) \\ &\propto \pi(X \mid \tilde{\Psi}, \tilde{v}(0), \nu, \theta, \mu)\frac{1}{\Gamma(\nu)}\tilde{v}(0)^{\nu-1}e^{-\tilde{v}(0)} \\ &\times \pi(\tilde{\Psi})\pi(\nu, \theta, \mu). && (6.47) \end{aligned}$$

and is depicted as a graphical model in Figure 6.10, where we choose the same prior $\pi(\nu, \theta, \mu)$ as in the first parameterisation.

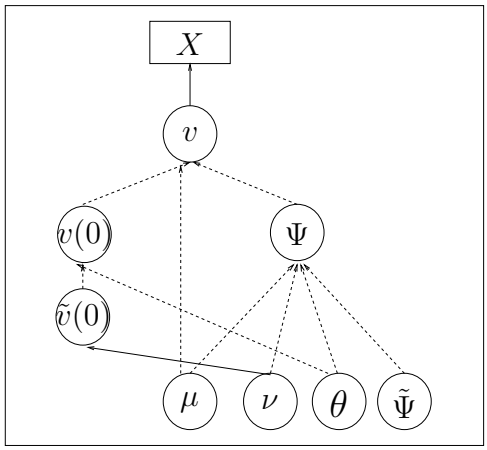


Figure 6.10: The graphical model of the non-centered parameterisation.

Generally speaking, there is a collection of NCPs for the gamma-OU model with the conditional independence structure represented by the graphical model in Figure 6.10. They all result in the same expression for the posterior distribution of missing data and parameters as given in (6.46), however they differ in the prior measure of $\tilde{\Psi}$ and the way $(\tilde{\Psi}, \nu, \theta, \mu)$ is transformed to Ψ. The problem of finding non-centered transformations for Poisson and marked Poisson processes was thoroughly investigated in Chapter 5. Actually, Section 5.5.1 describes three different constructions for latent Poisson processes with mean measure (6.31), although a slightly different notation is used there (contrast for example (5.26) with (6.31)). The next sections implement these ideas in the context of the gamma-OU model.

6.10 MPP-THIN-NCP for the gamma-OU model

Roberts et al. (2003) proposed a non-centered parameterisation for the data augmentation designed in Section 6.6, which is exactly the MPP-THIN-NCP described in Section 5.5.

Section 5.5.1 gave the details of this construction when the mean measure of Ψ is as defined in (6.31). Therefore, in order to make this section self-contained and recast things using the notation of this chapter, we just sketch the construction and refer to Section 5.5.1 for more details. We take $\tilde{\Psi}$ to be a marked Poisson process with points $\{(C_j, M_j)\}$ on $[0,T] \times (0,\infty)$ and marks $\tilde{E}_j$ on $(0,\infty)$ independent of each other and independent of the points. Its mean measure is given by (see (5.27))

$$\tilde{\Lambda}(dc \times dm \times d\tilde{\epsilon}) = e^{-\tilde{\epsilon}} dc\, dm\, d\tilde{\epsilon} \tag{6.48}$$

and we retrieve Ψ from $(\tilde{\Psi}, \lambda, \theta)$ as follows (see also Figure 5.7).

MPP-THIN-NCP transformation $(\tilde{\Psi}, \lambda, \theta) \to \Psi$

Select all points from $\tilde{\Psi}$ for which $M_j < \lambda$.

Project these points to $[0,T] \times (0,\infty)$.

Transform $\{(C_j, \tilde{E}_j)\}$ to $\{(C_j, E_j)\}$ where $E_j = \tilde{E}_j/\theta$.

Ψ consists of the transformed points.

The corresponding MCMC algorithm is as follows.

A Hastings algorithm to sample from $(\nu, \theta, \mu, \tilde{\Psi}, \tilde{v}(0)) \mid X$

Iterate the following steps:

1. Update (ν, θ, μ) according to $\pi(\nu, \theta, \mu \mid X, \tilde{\Psi}, \tilde{v}(0))$

2. Transform $\tilde{v}(0) \to \tilde{v}(0)/\theta = v(0)$ and $(\lambda, \theta, \tilde{\Psi}) \to \Psi$

3. Update $v(0)$ according to $\pi(v(0) \mid X, \Psi, \nu, \mu, \theta)$

4. Update Ψ according to $\pi(\Psi \mid X, v(0), \nu, \mu, \theta)$

5. Transform $\tilde{v}(0) = \theta v(0)$ and stochastically transform $(\lambda, \theta, \Psi) \to \tilde{\Psi}$.

Steps 3-4 are exactly the same as the corresponding steps of the centered algorithm given in Section 6.7.1. The two algorithms differ in Step 1, where the parameters are updated conditionally on the missing data, and in the two transformation Steps 2 and 5. Step 2 is totally deterministic, $\tilde{v}(0) \to v(0)$ is given in (6.45) and $(\lambda, \theta, \tilde{\Psi}) \to \Psi$ is described above and is illustrated in Figure 5.7. The transformation $(\lambda, \theta, \Psi) \to \tilde{\Psi}$ in Step 5 is actually stochastic, however it is not necessary. Section 5.3.1 and Section 5.3.2 showed how to implement this step incorporating in Step 1 of the algorithm; we skip details to avoid repetition. However, we use a rather efficient scheme tailored to this problem in order to update the parameters.

A blocking scheme to update the parameters

We first establish the conditional joint density of ν and μ given the observed and imputed data, but marginalised with respect to θ. The integrated volatility in $[t_{i-1}, t_i)$, as given in (6.26), can be written as

$$v^*(t_{i-1}, t_i) = \frac{g_i}{\theta}$$

where

$$g_i = \frac{\sum_{t_{i-1}}^{t_i} \tilde{E}_j}{\mu} - \frac{1}{\mu}\left((e^{-\mu t_i} - e^{-\mu t_{i-1}})(\tilde{v}(0) + \sum_{0}^{t_{i-1}} e^{\mu C_j} \tilde{E}_j) + \sum_{t_{i-1}}^{t_i} e^{-\mu(t_i - C_j)} \tilde{E}_j \right)$$

where $\{(C_j, \tilde{E}_j)\}$ are the resulting points from steps 1 and 2 of the transformation $(\lambda, \theta, \tilde{\Psi}) \to \Psi$ and where the sums $\sum_s^t$, with $s \le t$, are interpreted as sums over all (if any) $s \le C_j < t$. Hence, g_i is known given $\tilde{v}(0)$, $\tilde{\Psi}$, ν and μ. Therefore

$$\pi(X \mid \nu, \mu, \theta, \tilde{\Psi}, \tilde{v}(0)) = d(\nu, \mu, \tilde{\Psi}, \tilde{v}(0)) \theta^{n/2} \exp\{-\theta\ k(\nu, \mu, \tilde{\Psi}, \tilde{v}(0), X)\}$$

where

$$\begin{aligned} k(\nu, \mu, \tilde{\Psi}, \tilde{v}(0), X) &= \sum_{i=1}^{n} \frac{(x(t_i) - x(t_{i-1}))^2}{2 g_i} \\ d(\nu, \mu, \tilde{\Psi}, \tilde{v}(0)) &= \prod_{i=1}^{n} (g_i)^{-1/2}. \end{aligned} \tag{6.49}$$

This implies that by choosing a $\mathrm{Ga}(\alpha_\theta, \beta_\theta)$ prior on θ its full conditional posterior distribution is

$$\theta \mid \cdot \sim \mathrm{Ga}(n/2 + \alpha_\theta, k(\nu, \mu, \tilde{\Psi}, \tilde{v}(0), X) + \beta_\theta).$$

Furthermore, by integrating out θ we deduce that

$$\pi(\nu, \mu \mid X, \tilde{\Psi}, \tilde{v}(0)) \propto \frac{\tilde{v}(0)^{\nu}}{\Gamma(\nu)} d(\nu, \mu, \tilde{\Psi}, \tilde{v}(0)) \left(k(\nu, \mu, \tilde{\Psi}, \tilde{v}(0), X) + \beta_{\theta} \right)^{-(n/2+\alpha_{\theta})} \pi(\nu)\pi(\mu).$$

We will update (ν, θ, μ) by first updating (ν, μ) using a Metropolis-Hastings step on the logarithmic scale with target density $\pi(\nu, \mu \mid X, \tilde{\Psi}, \tilde{v}(0))$, and then simulating θ directly from its full conditional. Therefore, as we noted in Section 6.5.2 there are computational conveniences when using a gamma prior for θ. Moreover, it can be seen that the transformation $v(0) \to \tilde{v}(0)$ is necessary in order to be able to derive in closed form the full conditional distribution of θ.

6.11 Alternative non-centered parameterisations

Section 6.9 remarked that it is possible to construct a variety of non-centered parameterisations corresponding to the same graphical model shown in Figure 6.10. Recall from Section 6.6 that we augment a marked Poisson process Ψ with mean measure Λ given in (6.31), which depends on some parameters (λ, θ in particular). Chapter 5 discussed a variety of non-centered parameterisations for marked Poisson processes, which can all be used to construct an NCP for the OU model. The parameterisation proposed in Section 6.10 is the MPP-THIN-NCP initially discussed in Section 5.5, where two more NCPs were suggested, the MPP-CDF-NCP and the THIN-NCP. In particular, Section 5.5.1 constructed the latter two NCPs for a Poisson process Ψ with mean measure as in (6.31). In the sequel we sketch the constructions using the notation of this chapter, but we refer back to Section 5.5.1 where more details can be found.

The MPP-CDF-NCP is related with the Ferguson-Klass representation. This construction takes $\tilde{\Psi} = \{(C_i, \tilde{E}_i), i = 1, 2, \ldots\}$ to be a unit rate Poisson process on S and transforms $(\lambda, \theta, \tilde{\Psi}) \to \Psi$ as follows (see also Figure 5.8).

MPP-CDF-NCP transformation $(\tilde{\Psi}, \lambda, \theta) \to \Psi$

`Select all points` $(C_i, \tilde{E}_i) \in \tilde{\Psi}$ `for which` $\tilde{E}_i < \lambda$.

`Set` $E_i = -\log\{\tilde{E}_i/\lambda\}/\theta$.

Ψ `consists of the transformed points.`

We highlight that the decreasing transformation (5.20) is used, therefore $E_1 > E_2 > \cdots$ in the above transformation where $E_1 < \infty$ *almost surely*.

The THIN-NCP is based on random thinning and takes $\tilde{\Psi} = \{(C_i, E_i, M_i), i = 1, 2, \ldots\}$ to be a unit rate Poisson process on $S \times (0, \infty)$ and the transformation of $(\lambda, \theta, \tilde{\Psi})$ to Ψ is described below (see also Figure 5.9).

THIN-NCP transformation $(\tilde{\Psi}, \lambda, \theta) \rightarrow \Psi$

`Select all points from` $\tilde{X}$ `for which` $M_i < \lambda\theta \exp\{-\theta E_i\}$.

`Project these points to` $[0, T] \times (0, \infty)$.

Ψ `consists of the projected points.`

6.11.1 MCMC implementation

Both the MPP-CDF-NCA and the THIN-NCA iterate exactly the same steps as the MPP-THIN-NCA, which are given in Section 6.10. However, the way the transformation Steps 2 and 5 (in the Hastings-within-Gibbs algorithm of Section 6.10) are carried out is different for each of the proposed NCAs; specific details can be found in Section 5.3.1, Section 5.3.2 and Section 5.5. Moreover, Step 1 of the algorithm has a different implementation. Both the MPP-CDF-NCA and the THIN-NCA lead to an intractable full conditional for θ, therefore the blocking scheme proposed in Section 6.10 is not feasible anymore. Instead, we use an alternative generic blocking scheme for both algorithms. Let (ν_0, θ_0, μ_0) denote the current values of the parameters when entering Step 1 of the algorithm. We compute $(\xi_0, \omega_0^2, \mu_0)$ where ξ, ω^2 have been defined in Section 6.3.3 as the stationary mean and the variance of the volatility process. For the gamma-OU model $\xi = \nu/\theta$ and $\omega^2 = \nu/\theta^2$. We then use a random-walk proposal on the logarithmic scale to generate new values $(\xi_1, \omega_1^2, \mu_1)$. The covariance matrix of the proposal distribution is diagonal of the form

$$\Sigma = c \begin{pmatrix} \sigma_\xi^2 & 0 & 0 \\ 0 & \sigma_{\omega^2}^2 & 0 \\ 0 & 0 & \sigma_\mu^2 \end{pmatrix} \tag{6.50}$$

where the diagonal elements are pilot-run estimates of the variance of the corresponding parameters (after a log-transformation), and c is a scaling constant which is tuned to yield acceptance rates for this updating step around 0.3. The proposed values are then transformed back to (ν_1, θ_1, μ_1) and the acceptance ratio is calculated, which however takes into account the Jacobian of the above transformation. We choose this strategy because of the *a priori* belief, which is however supported by simulation results, that the dependence among (ξ, ω^2, μ) is less than among (ν, θ, μ), hence the spherical proposal distribution we use is more likely to be successful in exploring the space.

6.12 Simulation study

This section attempts to assess the relative performance of the various algorithms proposed in the previous sections under different kinds of datasets. Ideally, we would like to be able to compute exact $\mathcal{L}^2$ rates of convergence, as we did in Chapter 2. However, this is infeasible for any Gibbs sampler on a highly structured non-Gaussian target distribution. Instead, we take a simulation-based approach. We apply our MCMC algorithms to a variety of simulated datasets and we look at autocorrelation plots (under "stationarity") of the marginal chains of the parameters. In particular, we plot the estimated autocorrelation function for the MCMC time series (remaining after removing the burn-in samples and thinning) corresponding to $\lambda = \nu\mu$ for each algorithm, for each of the different datasets. The parameter λ was considered since its dependence with Ψ is expected to have an impact on the convergence of the centered algorithm. However, summaries for the other parameters (not included here) convey the same message. See Section 2.1.2 for an argument for using estimated autocorrelations as a means of assessing the speed of convergence of an MCMC chain.

We first examine how the MPP-THIN-NCA compares with its centered counterpart and then we investigate the extent to which the efficiency of the different non-centering schemes varies.

6.12.1 Comparison of CA vs NCA

In order to assess the performance of the two MCMC algorithms proposed in Section 6.7 and Section 6.9, we applied them to a varied collection of data simulated from the OU model of Section 6.5. Six experiments were conducted representing different types of dynamic structure and stationary moments of the stochastic volatility process, and different time series lengths. The aim of the study is to obtain an understanding of the kind of data for which each algorithm is more suitable and consequently to provide guidelines on when each should be preferred.

All experiments are done on the assumption that the data are daily. Having finer data

would be equivalent to making μ smaller while fixing ν, and so variation in data frequency is redundant in this comparison. Furthermore, variation in θ is also not necessary since its impact can be removed by scaling the observed data. A different simulation design is considered in Section 3 of Barndorff-Nielsen and Shephard (2002a), where the mean volatility ξ $(= \nu/\theta)$ is kept fixed, and ξ/ω^2 $(=\theta)$ varies, in order to assess how different estimators (non-parametric and model-based) of the integrated volatility perform, using high frequency data.

Dataset	θ	ν	μ	Length of series	Volatility stationary law
1	10	2/3	0.03	500	$\mathrm{Ga}(2/3, 10)$
2	10	2	0.03	500	$\mathrm{Ga}(2, 10)$
3	10	2/3	0.1	500	$\mathrm{Ga}(2/3, 10)$
4	10	2	0.1	500	$\mathrm{Ga}(2, 10)$
5	10	2/3	0.03	2000	$\mathrm{Ga}(2/3, 10)$
6	10	2	0.1	2000	$\mathrm{Ga}(2, 10)$

Table 6.1: Information about simulated datasets

Table 6.1 summarises the parameter values used in simulating the six datasets. The first four use shorter (length 500) time series. The first two of these have the same memory decay ($\mu = 0.03$) but different stationary distribution for the volatility process. The situation is similar in the Series 3 and 4, but where now $\mu = 0.1$, therefore the volatility processes are less persistent. Series 5 and 6 consist of longer sequences corresponding to the parameter values in Series 1 and 4 respectively.

The purpose of this study is not to analyse real data but to compare the two proposed algorithms. The choice of the hyperparameters for the priors on the parameters (see Section 6.5.2) is considered in some detail in Section 6.14. Here we use values which lead to relatively flat densities in the posterior modal area of the parameters, in order to test the efficiency of our MCMC methods. Extremely informative priors could potentially mask convergence problems in any of the algorithms. In particular, in all experiments, identical prior distributions were assigned to the parameters, a $\mathrm{Ga}(1, 0.1)$ was chosen for θ and ν and a $\mathrm{Ga}(1, 1)$ for μ.

For each of the 6 series each algorithm was run for 6 million iterations, where the first 50,000 iterations were removed as a burn-in period and subsequently parameter values were stored every 100th iteration. For reasons which will be described later in this section, we considered the following asymmetric initialisation of the two algorithms: the parameter values were initiated at their known values for the centered algorithm and at their prior means

for the non-centered algorithm. The computing time for the two algorithms is comparable, 1000 iterations (while chains were in stationarity) took 2.75 seconds of CPU time for the centered algorithm for dataset 1 while 4.47 seconds for the non-centered on a Pentium III 450MHz processor (all programs are coded in Fortran).

The results are summarised in Figure 6.11. The estimated autocorrelation function for the MCMC time series (remaining after removing the burn-in samples and thinning) corresponding to $\lambda = \nu\mu$ is plotted for the centered and non-centered algorithm, for each of the datasets.

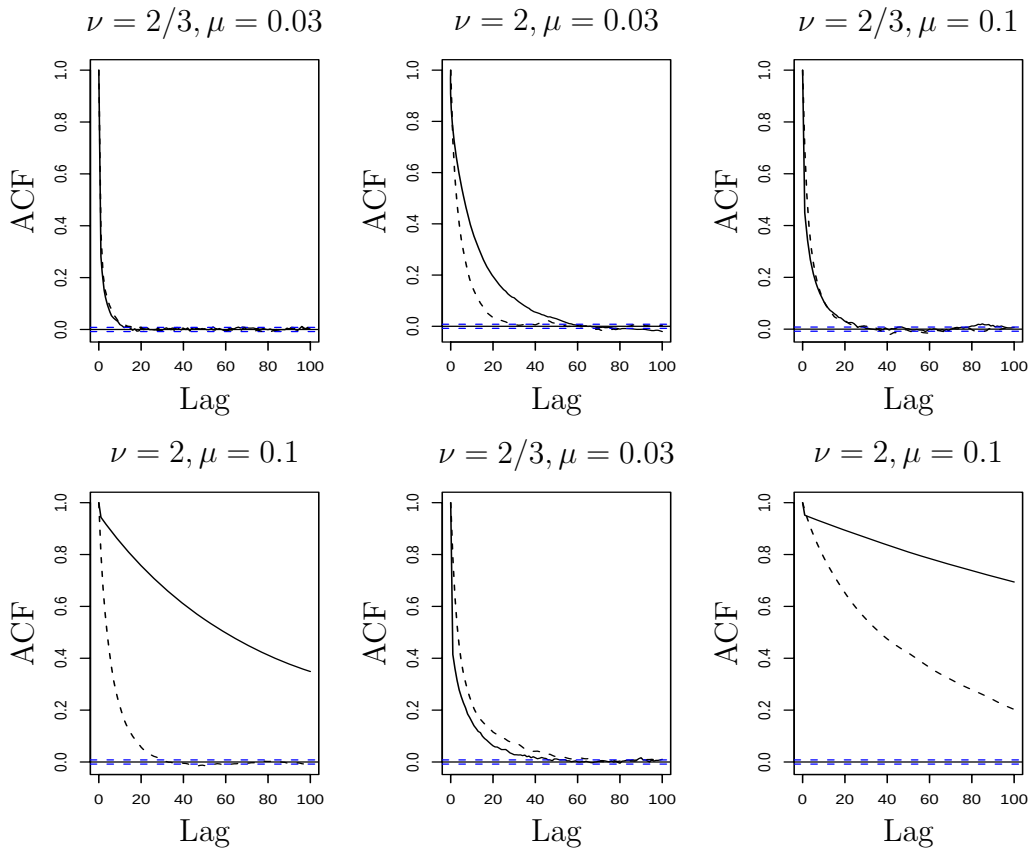
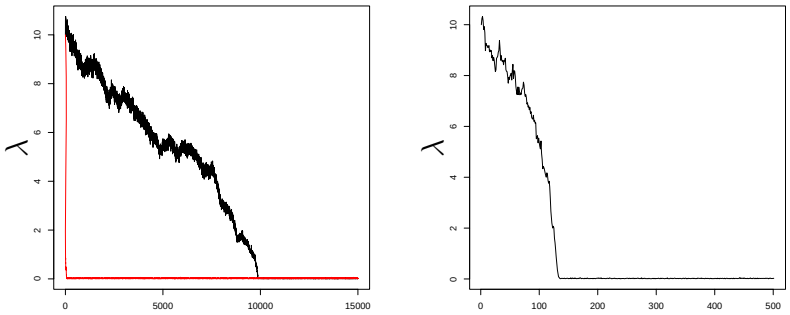


Figure 6.11: Estimates of the autocorrelation function of the marginal chain corresponding to $\lambda = \nu\mu$ for the centered (solid) and the non-centered (dashed) parameterisations for each of the simulated data sets described in Section 6.12.1. dataset 1 is in the top left corner and the rest of the datasets are placed from left to right. $T = 2000$ in the two bottom right plots, $T = 500$ in the rest. The estimates were calculated after thinning each of the chains one every hundredth and discarding the 500 initial points.

Initial inspection of the results reveals considerable robustness of the non-centered algorithm to a variety of different datasets, as opposed to the centered algorithm which seems to be doing badly in some cases. Due to the complexity of the problem, it is a daunting task to provide detailed explanation of the observed behaviour of the algorithms. Moreover, since we cannot quantify the relative strength of the marginal and augmented information (see Meng and van Dyk (1997)), it is difficult to know apriori which algorithm is to be preferred

for a given dataset. For example, if the underlying Poisson rate is small, then although the data are very informative about Ψ, as argued in Section 6.8, the prior dependence between Ψ and λ increases and it is not clear whether the centered or the non-centered parameterisation should be used. Our experience suggests that the centered should be favoured for high frequency data, otherwise the non-centered appears to be a more efficient and robust algorithm. Notice however, that the computer algorithms for both are quite similar since they differ only in the step of updating the parameters given the missing data, and it is not much harder to code both than just one of them. This point was also made in Section 4.1.

Another advantage of the non-centered algorithm is that, compared to its centered counterpart, it is considerably more robust to initial values. Figure 6.12a shows a run of both algorithms on dataset 1 with all parameters started from their prior means. The difference



(a) (b)

Figure 6.12: MCMC traces for λ for dataset 1, when all parameters are initialised from their prior means. (a): the centered (converging slowly) and the non-centered (converging rapidly) algorithms described in Appendices 1 and 2 are used. (b): a modification of the centered parameterisation is used, where 100 birth-death updates are performed for each update of the parameters. All chains have been thinned one every hundredth.

is striking; λ, when started from the tails, it zooms immediately into the modal area for the non-centered while it takes more than 10^6 iteration to reach near the mode for the centered. The behaviour of the centered can be improved at the expense of large computational effort, for example by doing multiple updates of Ψ for each update of the parameters (see Figure 6.12b). It appears that the posterior of the Poisson process depends less on λ as $\lambda \to \infty$, allowing the non-centered algorithm to return from the tails to the mode very quickly. We are currently investigating how this behaviour relates to the findings of Chapter 3 about uniform ergodicity of some NCAs. The instability to initial values of the centered-algorithm is a serious problem since choosing reasonable starting values is particularly difficult in this case where maximum likelihood estimates (for instance) are not available. Moreover, such unstable excursions indicate problems of the sampler in exploring tail areas and therefore

underestimating uncertainty, see for example Roberts (2003).

Figure 6.13 provides a visual summary to monitor the convergence of the high dimensional Ψ for the non-centered algorithm for dataset 1. The current configuration of the jump times

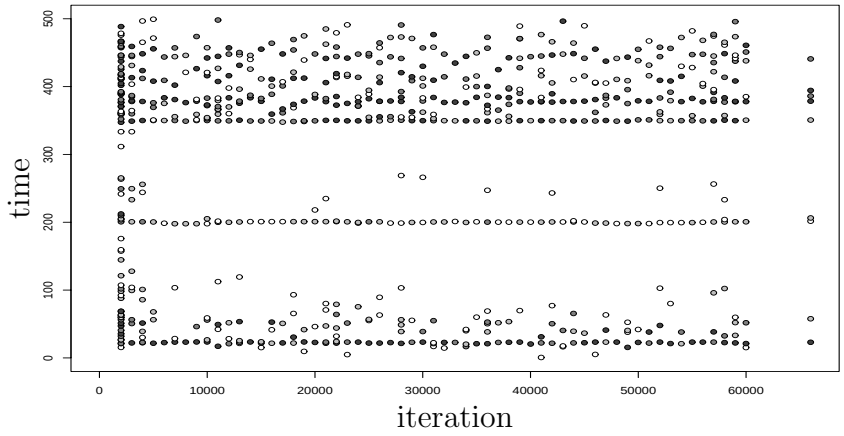


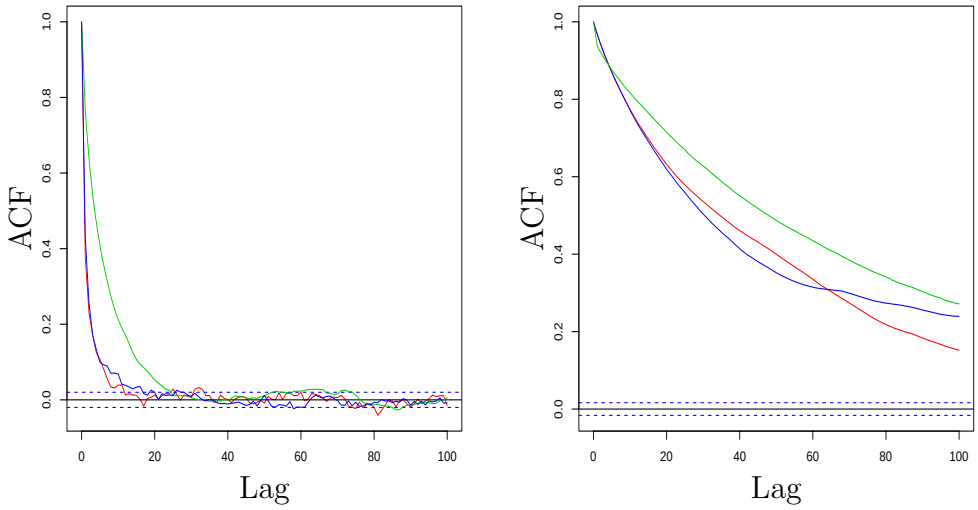
Figure 6.13: MCMC output for the point process Ψ for dataset 1, for the non-centered algorithm. The configuration of the jump times $\{C_i\}$ is plotted against every 1000th iteration for the first 60,000 iterations of the algorithms. The degree of darkness of the points within each configuration reflects their relative jump sizes E_i. In the far right the configuration of the jump times used in the simulation of the data is plotted. The picture is similar for the centered algorithm, although for this example convergence is not reached so quickly.

is being plotted against every 1000th iteration for the first 60,000 iterations. The degree of darkness of the points within each configuration reflects their relative jump sizes on a four colour grey scale (with black corresponding to the largest jumps). In the far right part of the figure the true configuration of the jump times produced from the simulation, coloured as described, is given. For this dataset the data are quite strong in identifying the hidden jump times, as is evident in Figure 6.13.

6.12.2 Comparison of the different NCPs

This section compares the three different non-centered parameterisations we proposed in Section 6.10 and Section 6.11. We look at estimated autocorrelations for λ, as in the simulation study of Section 6.12.1. Figure 6.14 plots these estimates for each of the NCPs applied to the simulated datasets 1 and 6. The results indicate that the MPP-THIN-NCP is slightly better. We note that the computational cost associated with implementing the THIN-NCA is greater than for the other two algorithms.

This simulation study revealed a very interesting scaling property of the MPP-CDF-NCA. Recall that the blocking scheme used in this algorithm for updating the parameters is the same as for the THIN-NCA and it is outlined in Section 6.11.1, and the same estimates



(a) (b)

Figure 6.14: Estimates of the autocorrelation function of the marginal chain corresponding to $\lambda = \nu\mu$ for the MPP-THIN-NCP (red), the MPP-CDF-NCP (green) and the THIN-NCP (blue) for simulated dataset 1 (left) and 6 (right). The estimates were calculated after thinning each of the chains one every hundredth and discarding the 500 initial points. Runs of length 1.5×10^6 were used for all algorithms.

$\sigma_\xi^2, \sigma_{\omega^2}^2, \sigma_\mu^2$ were used in both algorithms. The constant c was found to be very similar for both in order to achieve similar acceptance rates when they were applied to the dataset 1. Nevertheless, when they were applied to the dataset 6, the c chosen for the THIN-NCP was about 10 times smaller than for dataset 1, while the one for the MPP-CDF-NCP remained essentially unchanged. This means that the latter was attempting much larger steps on the parameter space than the former with the same acceptance rates, thus we would expect to see very different ACF plots from those presented in Figure 6.14. Intuitively, we would anticipate that the ACFs for the MPP-CDF-NCP decay faster than for the THIN-NCP, however this is not the case.

On the other hand, from the general results of Chapter 2, we know that a two component Gibbs sampler converges not faster than any of its components. This suggests that a potential explanation for this counter-intuitive algorithmic behaviour is that the mixing of Ψ is much slower in the MPP-CDF-NCP. Ψ is updated twice in any of the NCPs: once explicitly in Step 4 and once implicitly in Step 1 of the algorithm given in Section 6.10. By construction, all our proposed NCPs share the updating Step 4. Clearly, the changes that both the MPP-THIN-NCP and the THIN-NCP are attempting to make on Ψ in Step 1 are drastic, since points are randomly deleted or added, generated from the prior distribution. Therefore, when such moves are accepted, relatively large steps in the point process space are made. In this respect, especially when λ is very large, Step 4 adds little to the mixing of the algorithm.

On the contrary, the MPP-CDF-NCP attempts local changes to Ψ while updating the parameters in Step 1, since points corresponding to the smallest jump sizes are either removed or added. Therefore more ambitious steps in the parameter space can be achieved without changing the likelihood drastically, thus with high probability of acceptance. Nevertheless, Ψ is not moving fast around the parameter space and the updating at Step 4 is not enough for it to mix appropriately, since Ψ can change at most by one point, which is really not enough when λ is large.

To test empirically the above considerations we re-run the THIN-NCP and the MPP-CDF-NCP for dataset 6, but we performed 100 updates of Ψ at Step 4, for every update of the parameters. The results shown in Figure 6.15 seem to support our claims, since now the MPP-CDF-NCP is clearly converging much faster than the THIN-NCP.

Multiple updates of Ψ are computationally extremely expensive in our problem, and should be avoided. We have experimented extensively with more sophisticated approaches for updating Ψ, without significant success. However, if more efficient methods for this updating step could be found, then the MPP-CDF-NCP would become a very attractive option.

The observed behaviour of the different NCPs also relates with the success of the Metropolis-Hastings step we use to update the parameters given the missing data in Step 1 of the MCMC algorithm outlined in Section 6.10. We wish to get an impression of what $\log \pi(X \mid \lambda, \theta, \mu, \tilde{\Psi})$

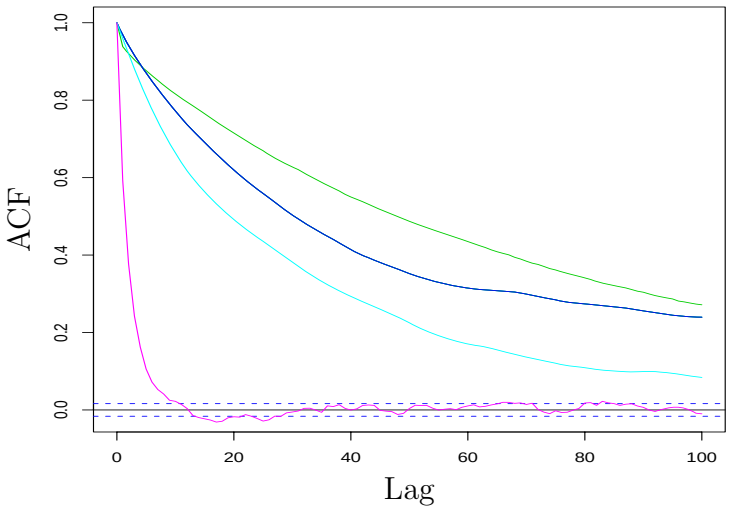


Figure 6.15: Estimates of the autocorrelation function of the marginal chain corresponding to $\lambda = \nu\mu$ for the MPP-CDF-NCP and the THIN-NCP for simulated dataset 6. The green and blue lines are the same as those in Figure 6.14. The two algorithms were re-run, performing 100 updates for Ψ, that is 100 birth-death-displacement steps, for every update of the parameters. The light blue line corresponds to the estimated ACF of λ for the THIN-NCP and the purple line for the MPP-CDF-NCP. The estimates were calculated after thinning each of the chains one every hundredth and discarding the 500 initial points. Runs of length 1.5×10^6 were used for all algorithms.

"looks like" as a function of the parameters when $\tilde{\Psi}$ is kept fixed, for each of our proposed NCPs. In particular, we want to investigate the smoothness of these functions. This issue was investigated also in Section 4.3 for some state space expanded NCPs. For simplicity all parameters are kept fixed except for λ, which is allowed to vary in the range $[\lambda_{\min}, \lambda_{\max}]$, containing the greatest part of its posterior mass. Actually, when θ is kept fixed the MPP-THIN-NCP coincides with the THIN-NCP therefore the latter is not considered in the sequel. To facilitate comparison among the different non-centering schemes, we simulate $\tilde{\Psi}$ for each NCP, not from its marginal posterior, but from the conditional posterior given specific parameter values. Thus, when these values are used, all NCPs yield the same realisation of the missing data Ψ. The experiment, which was conducted on the simulated datasets 1 and 6, is described in detail in the following paragraph.

To simplify exposition we introduce the notation $(\lambda_T, \theta_T, \mu_T)$ to denote the true parameter values used to simulate the datasets. Having fixed the parameters at their true values, we simulated from the conditional posterior distribution of the missing data by running our MCMC algorithm until "convergence" has been reached. Thus, a draw was obtained from the posterior distribution of Ψ and $v(0)$ conditional on the true parameter values, ψ and v respectively say. We then produced 100 draws from the posterior distribution of $\tilde{\Psi}$ given $\Psi = \psi, v(0) = v$ and $(\lambda, \theta, \mu) = (\lambda_T, \theta_T, \mu_T)$, for both the MPP-THIN-NCP and the MPP-

CDF-NCP. This is simply achieved by performing the stochastic transformation Step 5 of the MCMC algorithm in Section 6.10. Specifically, for the MPP-THIN-NCP each draw is obtained by simulating a collection of independent $\text{Un}[0, \lambda_T]$ marks, one for each of the points in ψ, and by simulating the Poisson process $\{(C_i, M_i, \tilde{E}_i), \lambda_T < M_i < \lambda_{\max}\}$ from the prior. For the MPP-CDF-NCP we first need to transform $E_i \to \tilde{E}_i$, so that $\tilde{E}_i \leq \lambda_T$, and subsequently to simulate $\{(C_i, \tilde{E}_i), \lambda_T < \tilde{E}_i < \lambda_{\max}\}$ from the prior. Notice that for each draw $\tilde{\Psi}$ produced for each of the algorithms, $(\tilde{\Psi}, \lambda_T, \theta_T) \to \psi$, although this is not the case for other parameter values. The aim of the experiment is to compute $\log \pi(X \mid \lambda, \theta_T, \mu_T, \tilde{\Psi})$ for each of the posterior draws $\tilde{\Psi}$ for each of the two algorithms.

Figure 6.16 and Figure 6.17 exhibit the results of the experiment for datasets 6 and 1 respectively. Top rows plot the conditional log-likelihood for a single realisation of $\tilde{\Psi}$, emphasizing the smoothness of the target density. The bottom rows superimpose the log-likelihoods for each of the draws. We have split the range of λ in three parts: the left tail (left), the area around the mode (middle) and the right tail (right).

Initial inspection of the figures reveals some fundamental differences in the two algorithms. All realisations of $\tilde{\Psi}$ for the MPP-CDF-NCP are transformed to the same Ψ for all $\lambda < \lambda_T$. On the contrary, the MPP-THIN-NCP induces randomness when λ moves in either direction of the current parameter value (λ_T in our example). The difference in smoothness of $\log \pi$ in the two algorithms is remarkable. The MPP-CDF-NCP, especially when there are many points in ψ, leads to very smooth target densities as opposed to the MPP-THIN-NCP.

All non-centered algorithms require a Metropolis-Hastings step for updating the parameters given the missing data. The experiment reveals that the target density of this step is much smoother for the MPP-CDF-NCP than the MPP-THIN-NCP (and similarly the THIN-NCP). It is known (Roberts and Yuen (2003)) that the performance of the Metropolis-Hastings algorithm critically depends on the smoothness of the target density. In fact, although it is known (Roberts et al. (1997)) that for high dimensional target densities on $\mathbb{R}^d$ the random walk Metropolis has a mixing time of order $O(d)$, Roberts and Yuen (2003) recently showed that for sufficiently well behaved discontinuous densities the mixing time of the algorithm is $O(d^2)$. Actually, when the density is as rough as those in the left panels of Figure 6.16 and Figure 6.16, the mixing time can be even worse.

It remains to be explored whether the difference in the behaviour of the algorithms is due to the efficiency of the Metropolis-Hastings step. This speculation seems to be supported by another simulation result. The THIN-NCP and the MPP-CDF-NCP were re-run for dataset 1, performing 100 updates of Ψ at Step 4 for every update of the parameters. The results in Figure 6.18 show that the difference in the performance in the two algorithms is not as distinct as in Figure 6.15 and actually the THIN-NCP appears to be more efficient than the the MPP-CDF-NCP. Thus, both algorithms behave in a similar way when there is not much difference in the smoothness of the posterior density of the parameters given the missing

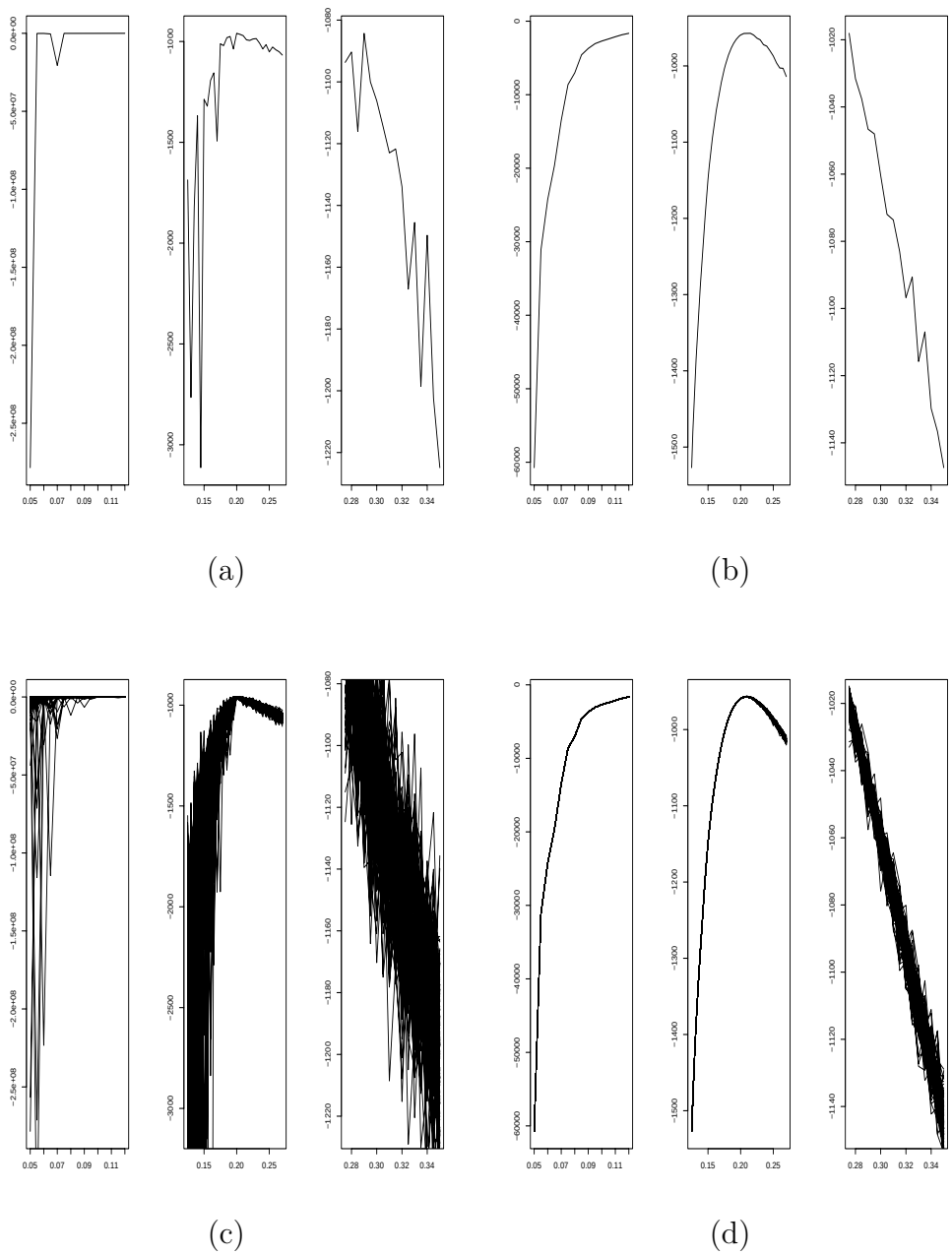


Figure 6.16: $\log \pi(X \mid \lambda, \theta_T, \mu_T, \tilde{\Psi})$ as a function of λ for dataset 6 for the MPP-THIN-NCP ((a) and (c)) and the MPP-CDF-NCP ((b) and (d)). The function for a single realisation of $\tilde{\Psi}$ is plotted on the top panel ((a) and (b)). The bottom panel superimposes this function calculated for 100 different realisations of $\tilde{\Psi}$. In both algorithms $\tilde{\Psi}$ is transformed to the same Ψ when the parameters take the values λ_T, θ_T. All draws of $\tilde{\Psi}$ have been simulated from its conditional distribution given the data and the true parameter values.

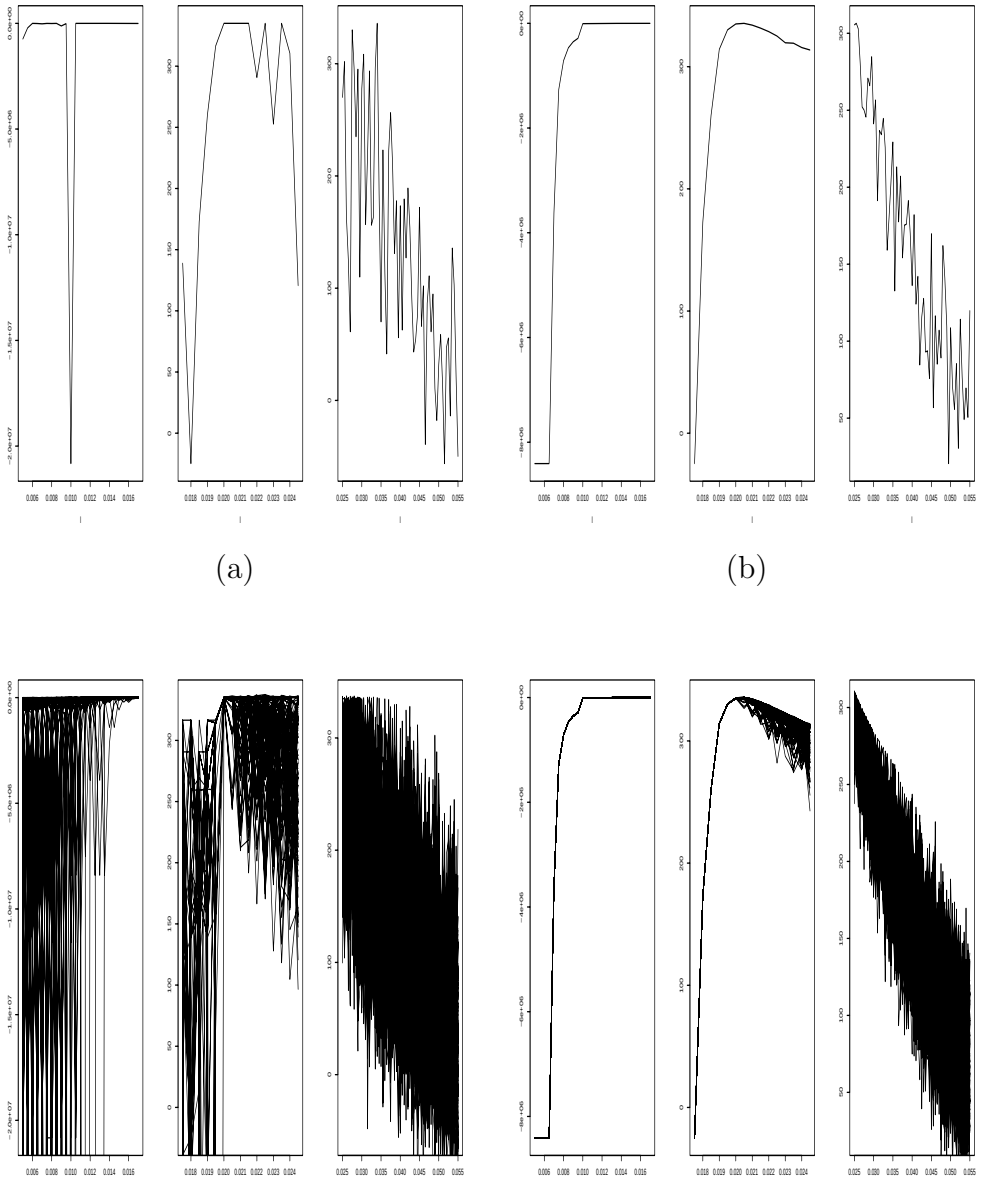


Figure 6.17: $\log \pi(X \mid \lambda, \theta_T, \mu_T, \tilde{\Psi})$ as a function of λ for dataset 1 for the MPP-THIN-NCP ((a) and (c)) and the MPP-CDF-NCP ((b) and (d)). The function for a single realisation of $\tilde{\Psi}$ is plotted on the top panel ((a) and (b)). The bottom panel superimposes this function calculated for 100 different realisations of $\tilde{\Psi}$. In both algorithms $\tilde{\Psi}$ is transformed to the same Ψ when the parameters take the values λ_T, θ_T. All draws of $\tilde{\Psi}$ have been simulated from its conditional distribution given the data and the true parameter values.

data.

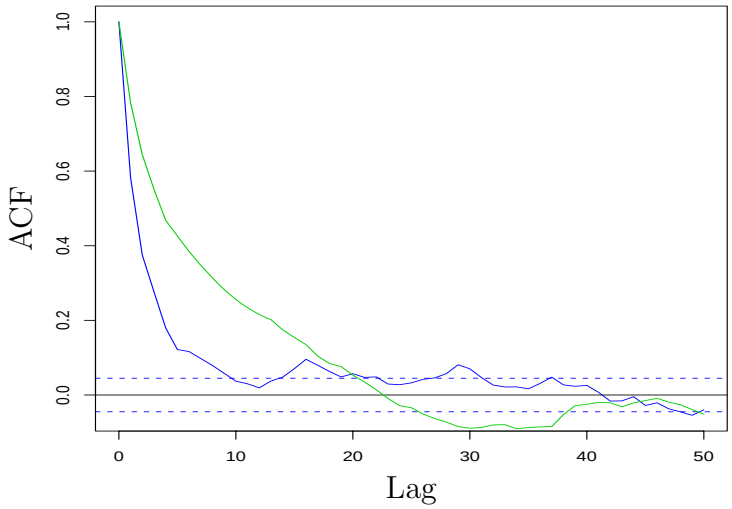


Figure 6.18: Estimates of the autocorrelation function of the marginal chain corresponding to $\lambda = \nu\mu$ for the MPP-CDF-NCP and the THIN-NCP for simulated dataset 1. The two algorithms were run performing 100 updates for Ψ for every update of the parameters. The blue line corresponds to the estimated ACF of λ for the THIN-NCP and the green line for the MPP-CDF-NCP. The estimates were calculated after thinning each of the chains one every fiftieth and discarding the 100 initial points. Runs of length 10^5 were used for both algorithms.

We have just scratched the surface here. The following question needs to be investigated and is currently part of ongoing work: whether the varying efficiency between different NCPs is due to the different dependence between the parameters and the transformed missing data, or due to the efficiency of the Metropolis-Hastings step used to update the parameters. If the latter is true, then other unresolved issues raised in this thesis could be clarified. For example, it would become clearer why the state-space expanded NCPs converge much slower than the scale NCPs for the models in Section 4.3. Moreover, the finding that the MPP-THIN-NCP works worse for the two-OU than for the single-OU model, as we shall see in Section 6.13, could more easily be explained. Intuitively, we would expect the NCP to be even better, since the more refined latent structure is harder to be identified from the observed data, thus the NCP is likely to be even more successful for the two-OU model. On the other hand, since more parameters are introduced the Metropolis-Hastings step has a higher dimensional discontinuous target density, and its performance is bound to get worse.

6.13 Augmentation and non-centered parameterisation for the superposition of OU processes

In two-OU model of Section 6.5.1 the parameters of interest are $\nu_1, \nu_2, \mu_1, \mu_2, \theta$ and the augmentation scheme of Section 6.6 is adopted, where the missing data are the initial volatilities $v_1(0)$ and $v_2(0)$ and the two marked Poisson processes Ψ_1 and Ψ_2 on S where each $\Psi_i,\ i = 1, 2$ contains the points $\{(C_{ij}, E_{ij})\}$. The MCMC mixing problems based on a centered parameterisation described for the single-OU model in Section 6.8 are expected to be even more profound for the model based on the superposition of such processes. Therefore we directly proceed to construct a non-centered parameterisation for this augmentation scheme.

We propose a non-centered parameterisation which is a direct extension of the MPP-THIN-NCP which we constructed for the single-OU model in Section 6.9. It works with a marked Poisson process $\tilde{\Psi}$ with points $\{(C_j, M_j)\}$ on $[0, T] \times (-\infty, \infty)$ and marks $\tilde{E}_j$ on $(0, \infty)$ independent of each other and independent of the points. We take

$$\tilde{\Lambda}(dc \times dm \times d\tilde{\epsilon}) = e^{-\tilde{\epsilon}} dc\ dm\ d\tilde{\epsilon}$$

to be the mean measure of $\tilde{\Psi}$. Notice that the marked Poisson process introduced in Section 6.10 is the same as $\tilde{\Psi}$ defined above, but restricted on the subset $[0, T] \times (0, \infty) \times (0, \infty)$. Transformation of $(\lambda_1, \lambda_2, \theta, \tilde{\Psi})$ to (Ψ_1, Ψ_2) might be done as follows (see also Figure 6.19).

MPP-THIN-NCP transformation $(\tilde{\Psi}, \lambda_1, \lambda_2, \theta) \to (\Psi_1, \Psi_2)$

Select all points from $\tilde{\Psi}$ for which $-\lambda_2 < M_j < \lambda_1$.

Denote those with positive M_j as $\{(C_{1j}, M_{1j}, \tilde{E}_{1j})\}$ and the rest as $\{(C_{2j}, M_{2j}, \tilde{E}_{2j})\}$.

Project these points to $[0, T] \times (0, \infty)$.

$\Psi_i = \{(C_{ij}, E_{ij}), j = 1, 2, \ldots\}$, where $E_{ij} = \tilde{E}_{ij}/\theta$ for $i = 1, 2$.

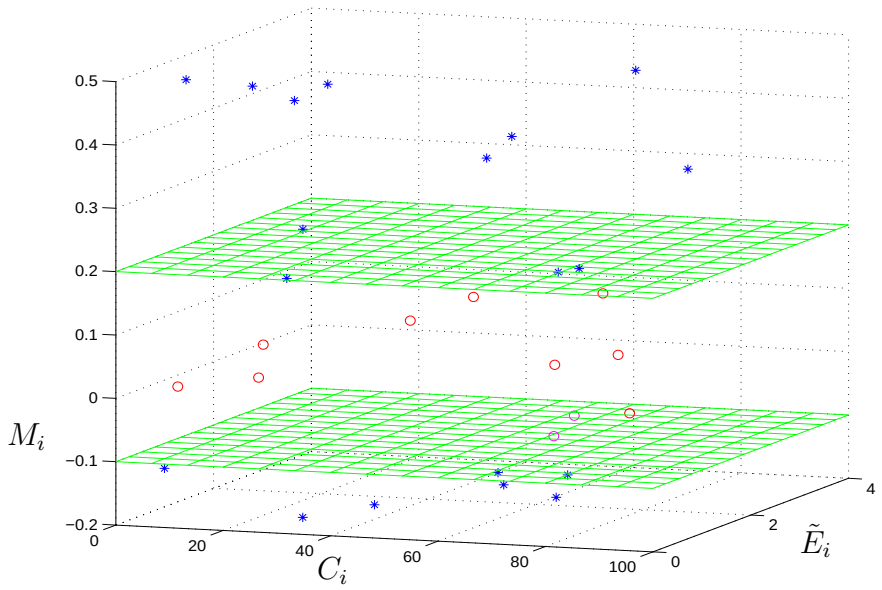


Figure 6.19: The MPP-THIN-NCP of $(\Psi_1, \Psi_2, \lambda_1, \lambda_2, \theta)$. Current values of the parameters are assumed to be $\lambda_1 = 0.2, \lambda_2 = 0.1, \theta = 1$ and $T = 100$. $\tilde{\Psi}$ is a Poisson process on $[0, T] \times (-\infty, \infty) \times (0, \infty)$ with mean measure $e^{-\tilde{\epsilon}} dc\, dm\, d\tilde{\epsilon}$; choose all $(C_i, M_i, \tilde{E}_i) \in \tilde{\Psi}$ with $-\lambda_2 < M_i < \lambda_1$ (denoted by circles as opposed to the points with $M_i > \lambda_1$ or $M_i < -\lambda_2$ denoted by asterisks).

We also transform $v_i(0)$ to $\tilde{v}_i(0) = v_i(0)/\theta,\ i = 1, 2$ and the resulting posterior density is

$$\begin{aligned}
&\pi(\nu_1, \nu_2, \mu_1, \mu_2, \theta, \tilde{\Psi}, \tilde{v}_1(0), \tilde{v}_2(0) \mid X) \\
&\quad \propto \quad \pi(X \mid \tilde{\Psi}, \tilde{v}_1(0), \tilde{v}_2(0), \nu_1, \nu_2, \mu_1, \mu_2, \theta) \frac{\tilde{v}_1(0)^{\nu_1 - 1}}{\Gamma(\nu_1)} \frac{\tilde{v}_2(0)^{\nu_2 - 1}}{\Gamma(\nu_2)} e^{-\tilde{v}_1(0) - \tilde{v}_2(0)} \\
&\quad \times \quad \pi(\tilde{\Psi}) \pi(\nu_1, \nu_2, \mu_1, \mu_2, \theta) \mathbb{1}[\mu_1 > \mu_2] \qquad (6.51)
\end{aligned}$$

where $\pi(\nu_1, \nu_2, \mu_1, \mu_2, \theta)$ denotes the prior density of the parameters; see Section 6.5.2 for guidelines on how to construct this prior, but also Section 6.14. Recall that we assume that $\theta \sim \text{Ga}(\alpha_\theta, \beta_\theta)$, which is shown to be computationally convenient below, and that a bivariate prior for (μ_1, μ_2) is used constrained on $\mu_1 > \mu_2$.

We use a component-wise Metropolis-Hastings algorithm to sample from the posterior in (6.51), which is largely based on the MCMC algorithm described in Section 6.10 for the single-OU model. We update the parameters $(\nu_1, \nu_2, \mu_1, \mu_2, \theta)$ all in one block, mainly to reduce the amount of likelihood evaluations which are computationally expensive. The full conditional distribution of θ is known, as in the case of the single-OU model,

$$\theta \mid \cdot \sim \text{Ga}(n/2 + \alpha_\theta, k(\nu, \mu, \tilde{\Psi}, \tilde{v}(0), X) + \beta_\theta)$$

since the jumps of two the Lévy processes are identically distributed with common parameter θ and we use a gamma prior. In the expression above k is computed as in (6.49). Therefore we can perform a blocking updating scheme as described in Section 6.10, with d given by (6.49), but with g_i replaced by the sum $g_{i1} + g_{i2}$ and each $g_{ij}, j = 1, 2$ computed separately for each process as in Section 6.10.

Although not considered in this thesis, alternative NCPs exist for the superposition of the two-OU model, which are immediate extensions of those proposed in Section 6.11

6.13.1 Examples using simulated data

The efficiency of the MCMC algorithm described in the previous section is tested using the two different simulated datasets described in Table 6.2. Both consist of 2000 daily data but the parameter values used in the simulations differ across the datasets. The first dataset was used in the simulation study contained in Roberts et al. (2003). The second dataset is simulated using the parameter estimates obtained by Roberts et al. (2003) when fitting the model to the US Dollar-Deutsch Mark exchange rate; see Section 6 of Roberts et al. (2003) and Section 6.16 of this thesis.

Dataset	θ	ν_1	ν_2	μ_1	μ_2	Volatility stationary law
1	10	0.25	0.5	0.8	0.01	Ga$(0.75, 10)$
2	25	0.66	0.62	3	0.04	Ga$(1.28, 25)$

Table 6.2: Information about simulated datasets for the two-OU model

For each of the datasets the non-centered algorithm of Section 6.13 was used to sample from the posterior distribution of the parameters. The algorithm was run for 6 million iterations, the first 50,000 were removed as a burn-in period and then parameters were stored every 100th iteration. The priors we used were chosen to be relatively flat in the posterior modal area, a Ga$(1, 0.1)$ for $\nu = \nu_1 + \nu_2$, a Ga$(1, 0.01)$ for θ, a Un$[0, 1]$ for $w_2 = \nu_2/\nu$, a Ga$(1, 1)$ for μ_2 and a Ga$(1, 0.01)$ for $\mu_1 - \mu_2$ (see Section 6.5.2 for a discussion on the prior specification for this model).

A significant feature is the slow mixing of the chains for the second dataset. Generally, we would expect the non-centered algorithm to be much more suitable for the superposition of the OU processes than for the single OU model, since the more complex latent structure should be more difficult to be identified by the data. A potential explanation of the slow convergence is the point made in Section 6.12.2 regarding the effect of the smoothness of the target densities in the Metropolis-Hastings step used to update the parameters. We have already seen that both the MPP-THIN-NCA and the THIN-NCA perform not very satisfactorily when the underlying Poisson rate λ is high (see for example Figure 6.11), thus

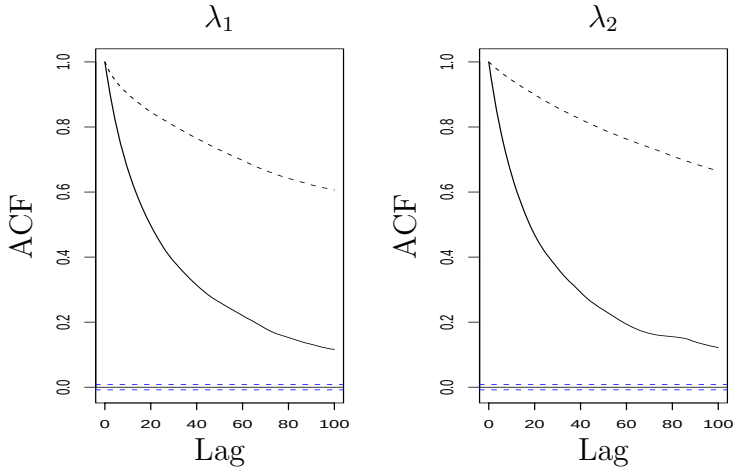


Figure 6.20: Estimates of the autocorrelation function of the marginal chain corresponding to $\lambda_1 = \nu_1\mu_1$ and $\lambda_2 = \nu_2\mu_2$ for the dataset 1 (solid) and the dataset 2 (dashed), described in Table 6.2. The estimates were calculated after thinning each of the chains one every hundredth and discarding the 1000 initial points.

it is not surprising to see the MPP-THIN-NCA to perform poorly for the superposition where λ_1 is very high, and actually its performance to deteriorate the higher the λ_1. We are currently investigating these issues and exploring whether the MPP-CDF-NCP could provide an improvement.

6.14 Posterior inference and sensitivity analysis

We now refrain from the computational issues and focus on posterior inference for the gamma-OU models of Section 6.5 and Section 6.5.1. We initially consider the single-OU and then the two-OU model.

Histograms of the posterior distributions of the parameters (ν, θ, μ) for the simulated datasets 1,5 and 6 (see Table 6.1) are shown in Figure 6.21, Figure 6.22 and Figure 6.23 respectively. Superimposed are the values that were used in the simulations and the prior density for each of the parameters. The histograms were produced using the function `truehist` of `R`, thus the area under each histogram is 1.

It can be seen that the parameters are well identified in all datasets and the posterior distributions are becoming more peaked around the true values the longer the time series. An interesting feature is that the highly correlated datasets (simulated using small values of μ) are very informative about the memory and less about the stationary parameters. The opposite is true for datasets simulated using large values of μ. For example, the posterior variances of $\log(\mu)$ are 8.62×10^{-3} and 3.94×10^{-2} for datasets 1 and 6 respectively, although the latter contains four times more data than the former. On the other hand, the posterior variances of $\log(\xi)$ (where $\xi = \nu/\theta$ is the stationary mean of the volatility) are 5.03×10^{-2}

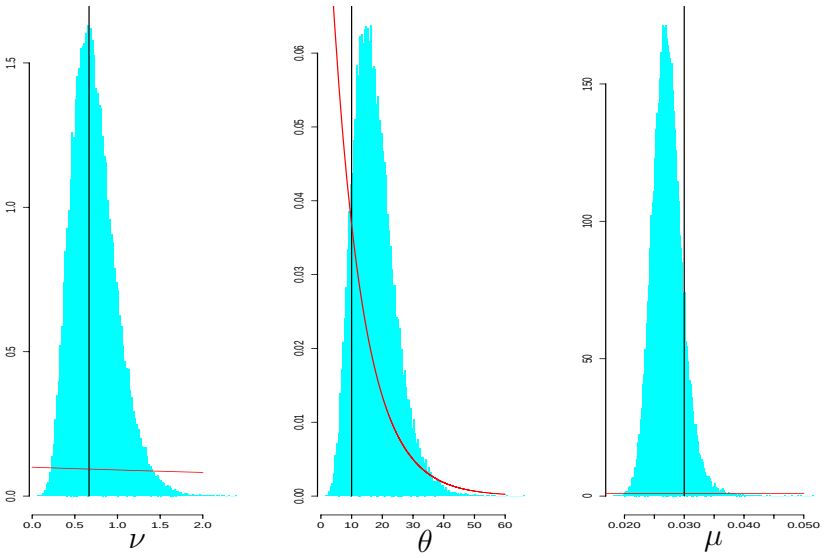


Figure 6.21: Histograms of the posterior distribution for the parameters ν, θ, μ under dataset 1 (see Table 6.1). The black vertical lines indicate the values of the parameters used in the data simulation. The prior density for each of the parameters is plotted in red.

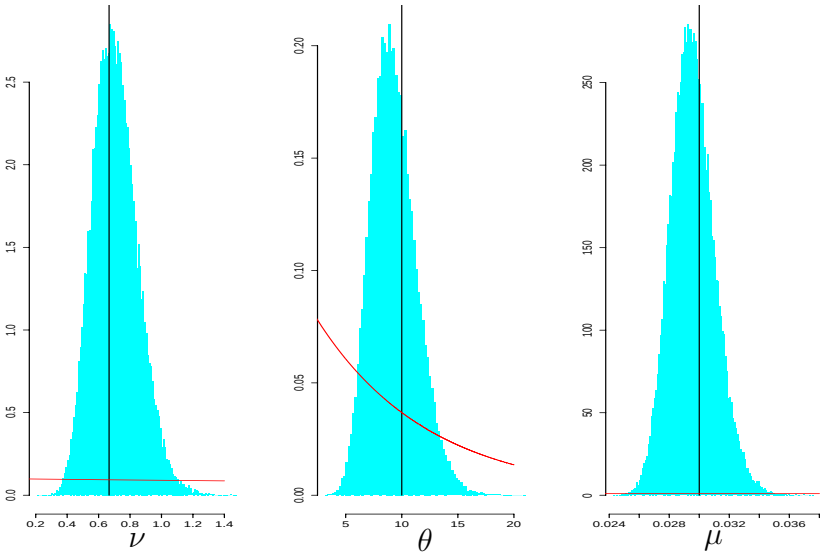


Figure 6.22: Histograms of the posterior distribution for the parameters ν, θ, μ under dataset 5 (see Table 6.1). The black vertical lines indicate the values of the parameters used in the data simulation. The prior density for each of the parameters is plotted in red.

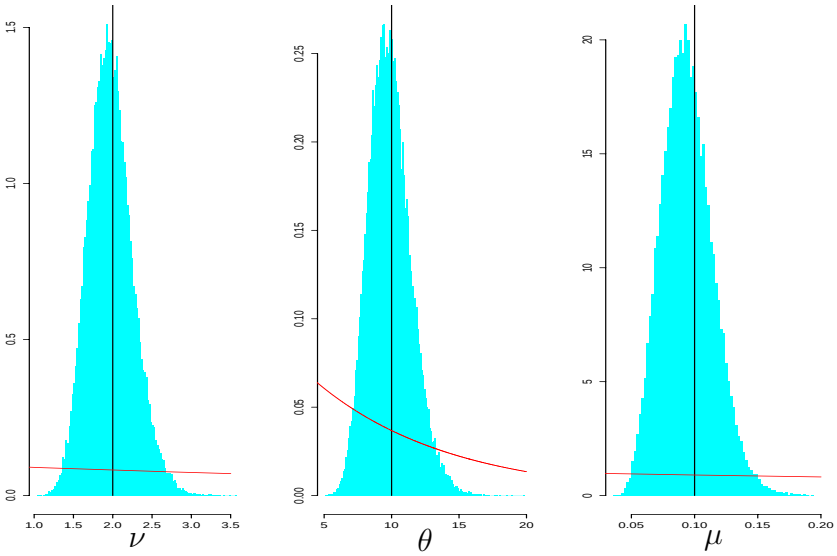


Figure 6.23: Histograms of the posterior distribution for the parameters ν, θ, μ under dataset 6 (see Table 6.1). The black vertical lines indicate the values of the parameters used in the data simulation. The prior density for each of the parameters is plotted in red.

and 7.12×10^{-3} for datasets 5 and 6 respectively. These results are very reasonable, since when $\mu \to \infty$ the observed data are IID observations, thus they are very informative about ν and θ, which control their common marginal distribution, while μ is not identifiable.

The priors can be seen to be flat in the posterior modal area for all the parameters except for θ. The informativity of the prior on θ is mostly apparent in dataset 1. It is relatively easy to choose flat gamma priors for ν and μ when analysing financial time series. The former being the shape parameter of the stationary distribution of the volatility is expected to take values no larger than 7 or 8, since the distribution is expected to be skewed and highly non-Gaussian. Our experience suggests that when the single-OU model is fitted to real data the long memory component is identified (see Section 6.16). Therefore, since financial series exhibit volatility clustering, μ is typically very small, in the range $0.01 - 0.15$. Thus, it is feasible to choose the hyperparameters of the gamma priors so that the densities are flat in the posterior modal area. On the contrary, this task is much harder for θ, since its magnitude depends on the scaling of the data.

A problem with the prior specification for this model is that default improper priors cannot be used blindly. For example, if ν, θ are kept fixed and $\mu \to \infty$, the likelihood does not converge to 0; instead it corresponds to a model where the log-returns are independent with the appropriate mixture of Gaussian distributions. Therefore, it is necessary to choose a proper prior for μ. This necessity is well known in latent variable models and has attracted the interest of many authors, especially in the area of finite mixture of densities; see Diebolt and Robert (1994) and Roeder and Wasserman (1997). In this case, since even the use of arbitrary vague proper priors is not adequate, there has been a series of attempts to propose

"default" priors that are, somehow, data-dependent (Richardson and Green (1998), Robert (1996)). On the other hand, it can be argued (see for example Section 2.5 of Griffin and Steel (2002)) that a $\text{Ga}(1,1)$ is a non-informative prior for the memory parameter μ.

In order to to assess the sensitivity of the posterior distributions of the parameters to the prior specification for θ, dataset 1 was re-analysed using a flatter prior for θ in the area of its posterior mode. The posterior histograms shown in Figure 6.24 confirm that the $\text{Ga}(1, 0.1)$ prior used in the previous analysis for θ was quite informative about the right tail of its posterior distribution. Notably, the inference for ν and especially for μ is quite robust to the prior elicitation of θ.

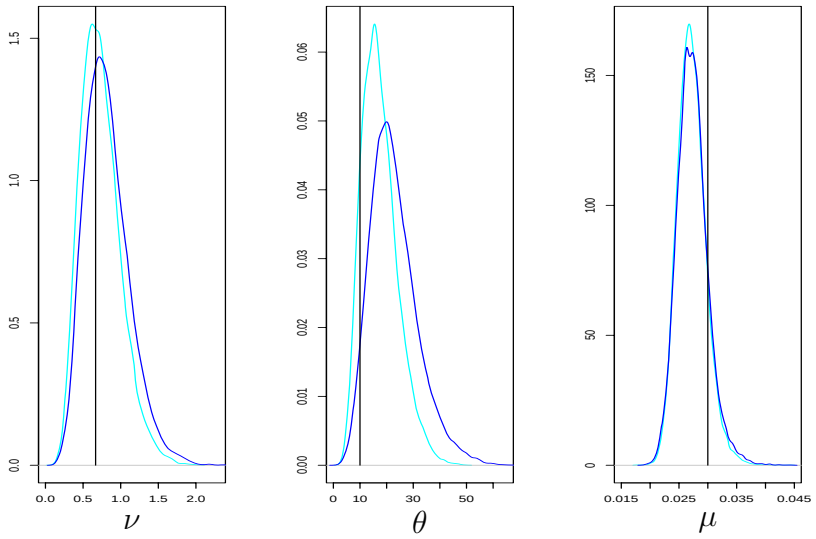


Figure 6.24: Kernel estimates of the posterior density for the parameters ν, θ, μ under dataset 1 (see Table 6.1), for two different prior specifications: $\nu \sim \text{Ga}(1, 0.1)$, $\mu \sim \text{Ga}(1, 1)$ and $\theta \sim \text{Ga}(1, 0.1)$ (light blue) and $\theta \sim \text{Ga}(1, 0.01)$ (dark blue). Notice that the light blue lines are density estimates of the same posterior distribution that is represented by histograms in Figure 6.21. The black vertical lines indicate the values of the parameters used in the data simulation.

We also analysed dataset 1 using exponential priors for all parameters with mean equal to their true values. The estimates of the corresponding posterior densities for ν and μ plotted in Figure 6.24, indicate reasonable robustness of the posteriors to very different priors. Notice that results for θ are not included in the figure, since the new prior coincides with the one used in the original analysis and the new posterior density is essentially identical to the one plotted in Figure 6.21.

All density estimators have been produced using the R function density, with Gaussian kernel and Silverman's optimal bandwidth. This is a sensible choice, since the effect of the rejections in our Hastings algorithms is eliminated by the thinning of the chains, however see Sköld and Roberts (2003) for some results on optimal bandwidth selection for Metropolis-Hastings Markov chains.

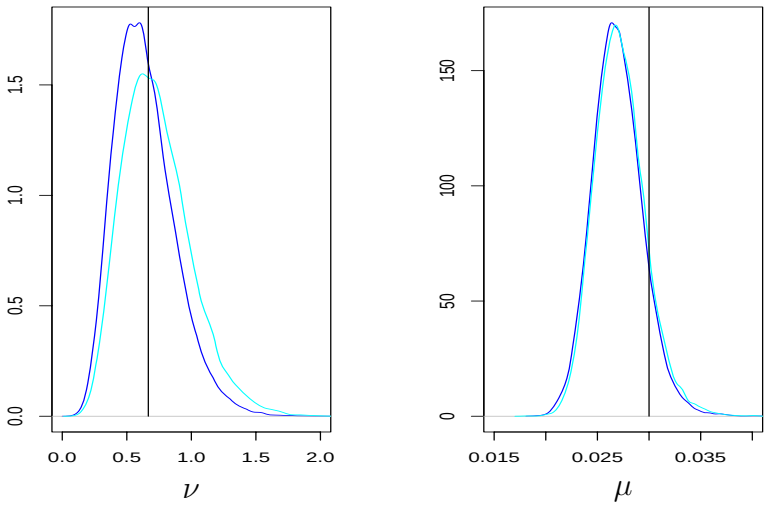


Figure 6.25: Kernel estimates of the posterior density for the parameters ν, μ under dataset 1 (see Table 6.1), for two different prior specifications: $\nu \sim \text{Ga}(1, 0.1),\ \mu \sim \text{Ga}(1, 1)$ (light blue) and $\nu \sim \text{Ga}(1, 3/2),\ \mu \sim \text{Ga}(1, 100/3)$ (dark blue). Notice that the light blue lines are density estimates of the same posterior distribution that is represented by histograms in Figure 6.21. The black vertical lines indicate the values of the parameters used in the data simulation.

Figure 6.26 and Figure 6.27 show the posterior density estimates (using again the R function density) for the parameters of the two-OU model for each of the datasets 1 and 2 of Table 6.2. We have used the priors described in Section 6.5.2 and Section 6.13.1. As it can be seen the priors are flat in the area of high posterior probability.

6.15 Model diagnostic tools

It is interesting to investigate the strength of the prior assumptions regarding the latent structure, and how well it is identified from the data. To this end, this section introduces some simple graphical diagnostics which use the MCMC output of Ψ to assess the Poisson and exponential assumptions about the jump times and the jump sizes respectively under the OU models in Section 6.5 and Section 6.5.1. Our diagnostics are based on ergodic properties of the model, therefore they are informative when λT is rather large. These diagnostics are the subject of current research, therefore we will not go into too much detail. Instead we just present the main idea below and illustrate the method by applying it to simulated data.

Let $\tilde{C} = \{\lambda C_1, \lambda C_2, \ldots\}$ and $\tilde{E} = \{\theta E_1, \theta E_2, \ldots\}$ where $C_j,\ E_j$ are defined in Section 6.4. Using MCMC samples from the posterior distribution of $\tilde{C}$ and $\tilde{E}$, we graphically examine whether they are consistent with the prior assumptions, namely that $\lambda(C_j - C_{j-1}) \sim \text{Ex}(1)$ and that $\theta E_j \sim \text{Ex}(1)$. Our diagnostic plots rely on the property that if $E \sim \text{Ex}(1)$ then $-\log P[E > t] = t,\ t > 0$. Thus, systematic deviations from a straight line can inspected

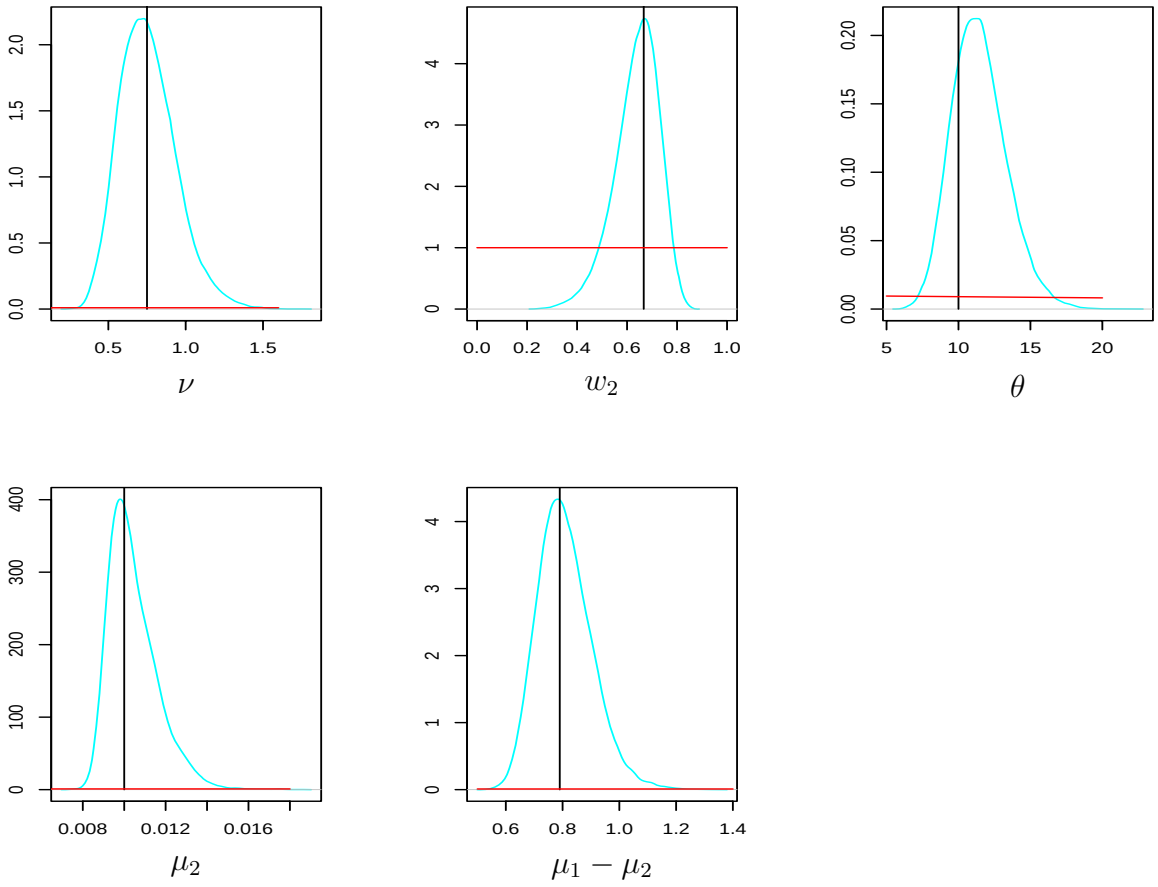


Figure 6.26: Posterior density estimates for the parameters $\nu, w_2, \theta, \mu_2, \mu_1 - \mu_2$ under dataset 1 (see Table 6.2). The black vertical lines indicate the values of the parameters used in the data simulation. The prior density for each of the parameters is plotted in red.

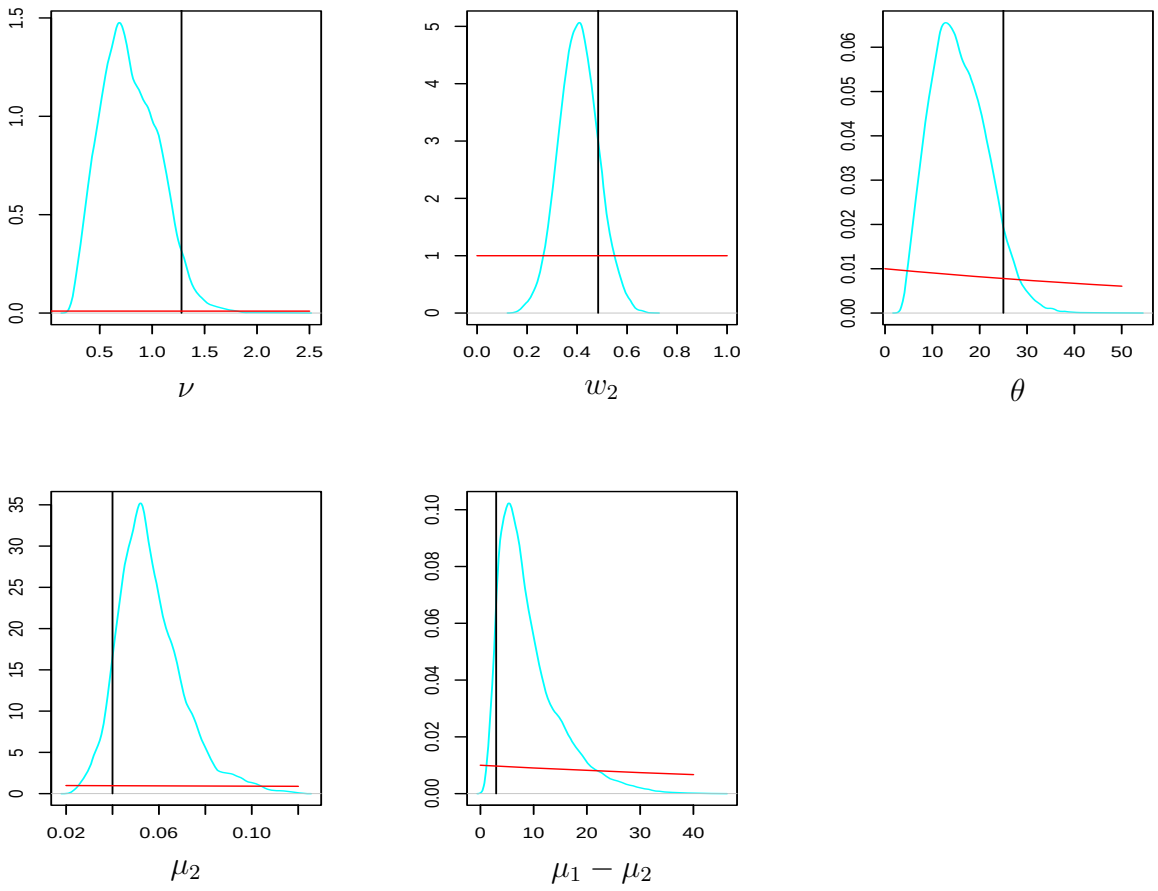


Figure 6.27: Posterior desnity estimates for the parameters $\nu, w_2, \theta, \mu_2, \mu_1 - \mu_2$ under dataset 2 (see Table 6.2). The black vertical lines indicate the values of the parameters used in the data simulation. The prior density for each of the parameters is plotted in red.

by plotting an empirical estimate of $-\log P[\epsilon > t]$ against t. If the model is the "correct" one, then due to the ergodic properties of the time series model, we would expect the sample of the "residuals" in $\tilde{C}$ and $\tilde{E}$ to be drawn from the prior distribution, since we implicitly average over the empirical distribution of the observed data. We aim at two things: first, to apply our diagnostics to real data as a method of assessing model fit; second, to apply them to data simulated from a different model and examine whether the misspecification is reflected in the posterior distribution of the parameters and the missing data.

In Figure 6.28 the diagnostics are tried for different simulated time series in $[0, 2000]$. Dataset 6 (Table 6.1) is used in the left column. The data used in the middle and right columns have been simulated using the same underlying OU process for both but different sampling frequency. The OU process has memory parameter $\mu = 0.1$ and it is driven by a compound Poisson process as in (6.24) where $\lambda = 0.2$ but the jump sizes have been simulated from a $\text{Ga}(0.1, 1)$. Daily and 100 data per day have been used for the middle and right columns respectively. For daily data, the diagnostics provide weak evidence about mispesification and we have found that the evidence varies considerably across different datasets simulated for the same realisation of the OU process. This suggests that for this observation frequency some aspects of the model cannot be identified. On the other hand, the diagnostics strongly indicate model mispecification if high frequency data are used. These diagnostic tools are adopted in the analysis of exchange rates data in Section 6.16. Nevertheless, careful analysis of the applicability and interpretation of these diagnostics is ongoing work and will be reported elsewhere.

6.16 A real data example

We fitted the models of Section 6.5 to the series of US dollar (US$) - Deutsch Mark (DM) exchange rate. The data were obtained from JP Morgan and are daily closing prices that span the period from 01/01/1986 to 01/01/1996 (2614 data points in total); they are plotted in Figure 6.1. We have scaled the original log-prices by a multiplicative factor of $1000^{1/2}$. This transformation affects only the parameter that controls the distribution of the jump sizes, that is θ in the models considered in this chapter.

This FX-market has been studied in detail by Andersen et al. (2001) and by Barndorff-Nielsen and Shephard (2002a) for a similar period of time (01/12/86-30/11/96). However, they use the Olsen high-frequency data (see for example Andersen et al. (2001)).

For the single-OU model, a $\text{Ga}(1, 0.01)$ prior is chosen for θ, a $\text{Ga}(1, 0.1)$ for ν and a $\text{Ga}(1, 1)$ for μ. For the two-OU model, we follow the suggestion made in Section 6.5.2 and parameterise in terms of $\nu = \nu_1 + \nu_2$ and $w_2 = \nu_2/(\nu_1 + \nu_2)$. We choose a $\text{Ga}(1, 0.1)$ prior for ν, a $\text{Un}[0, 1]$ for w_2, a $\text{Ga}(1, 0.01)$ prior for θ, a $\text{Ga}(1, 1)$ for μ_2 and $\mu_1 - \mu_2$ is assumed to be a $\text{Ga}(1, 0.01)$ random variable.

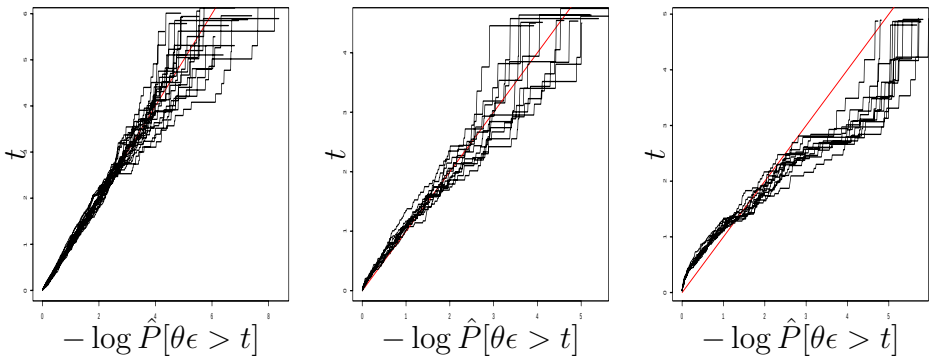


Figure 6.28: Model diagnostic plots for the single-OU model of Section 6.5: For each draw of $\tilde{C}$ and $\tilde{E}$ from the posterior distribution, we estimate empirically $-\log P[\lambda(c_j - c_{j-1}) > t]$ and $-\log P[\theta\epsilon > t]$. Each jagged line in the figures corresponds to such an estimator. If the model is "true" the estimator has to coincide with the straight 45^o line. Dataset 6 (Table 6.1) was used in the left column. The underlying OU process used in the middle and right columns has been generated under the specification $\lambda = 0.2,\ \mu = 0.1, T = 2000$ but the jump sizes have been generated from a Ga$(0.1, 1)$, instead of an Ex(10) used for Dataset6. Daily data were used in the middle column and 100 data per day in the right column. The posterior means for the parameters (ν, θ, μ) are (1.97,9.86,0.09) (left), (0.59,2.43,0.089) (middle) and (0.68,2.7,0.098) (right).

Tables 6.3 and 6.4 provide some posterior summaries for the parameters from fitting the single-OU and the two-OU models respectively. In particular, we report summaries for the mean and the variance of the volatility process, ξ and ω^2 respectively ($\xi = \nu/\theta, \omega^2 = \nu/\theta^2$, where $\nu = \nu_1 + \nu_2$ in the two-OU model), the decaying rates and w_2.

Parameter	Mean	Median	Standard deviation	95% Credible interval
ξ	5.08×10^{-2}	5.07×10^{-2}	3.49×10^{-3}	$(4.43, 5.81) \times 10^{-2}$
ω^2	8.55×10^{-4}	8.39×10^{-4}	1.72×10^{-4}	$(5.77, 12.5) \times 10^{-4}$
μ	7.1×10^{-2}	6.93×10^{-2}	1.83×10^{-2}	$(4.07, 11.58) \times 10^{-2}$

Table 6.3: Posterior parameter summaries for the US$/DM series under the single-OU model

Our results can be contrasted with those obtained by Barndorff-Nielsen and Shephard (2002a). They fit superposition-based OU stochastic volatility models to the realised volatility time series constructed from high frequency data. The fit is based solely on second-order characteristics using quasi-likelihood methods. Under this framework, OU and constant elasticity of variance models are indistinguishable and Barndorff-Nielsen and Shephard (2002a) make no specific assumptions about the form of the Lévy process in (6.7), and consequently about the stationary distribution of the volatility process. However, the results summarised

Parameter	Mean	Median	Standard deviation	95% Credible interval
ξ	5.19×10^{-2}	5.16×10^{-2}	4.55×10^{-3}	$(4.34, 6.17) \times 10^{-2}$
ω^2	2.43×10^{-3}	2.32×10^{-3}	6.7×10^{-4}	$(1.42, 3.99) \times 10^{-3}$
μ_1	4.012	3.789	1.43	$(1.86, 7.36)$
μ_2	5.43×10^{-2}	5.18×10^{-2}	1.16×10^{-2}	$(2.99, 9.04) \times 10^{-2}$
w_2	0.467	0.469	8.27×10^{-2}	$(0.30, 0.62)$

Table 6.4: Posterior parameter summaries for the US$/DM series under the two-OU model

in Table 3 of Barndorff-Nielsen and Shephard (2002a) are in agreement with ours. There, μ_1 is estimated as 3.74 and μ_2 as 0.043, although w_2 is estimated around 0.2 whereas our estimated posterior median is much larger, around 0.48. Therefore, Barndorff-Nielsen and Shephard (2002a) estimate a much sharper initial drop in the volatility autocorrelation than us.

Figure 6.29 contains a collection of plots which assist in assessing the model adequacy and fit. Recall the definitions of the series $\{v_n^*,\ n = 1, \ldots, T\}$ and $\{y_n, n = 1, \ldots, T\}$ given in Section 6.2 and Section 6.1 respectively. Figure 6.29a shows y_n^2 as points and the posterior median of v_n^* under the single-OU (dashed line) and the two-OU (solid line) models, for $n = 1000, \ldots, 1200$. Figure 6.29b shows this smoothing for the whole period of 2614 days. Figure 6.29c applies the diagnostics of Section 6.15 to the jump sizes from the single-OU model. Our diagnostics, applied also to the two-OU model but not reproduced here, indicate no significant model inadequacy. Figure 6.29d draws a kernel estimate of the log-density of the predictive distribution of $\log(v_n^*)$. It is interesting to compare this with similar estimates plotted in Figures 4b and 5b of Barndorff-Nielsen and Shephard (2002a). See also Figure 1 of Andersen et al. (2001) for some non-parametric estimates of the distribution of v_n^*. Finally, Figure 6.29e plots the posterior median of the autocorrelation function of the series $\{v_n^*\}$ from lag one onwards. Their theoretical forms as a function of the parameters have been derived by Barndorff-Nielsen and Shephard (2002a) and is given for the single-OU model in (6.15). Figure 6.29e shows that the two-OU model results in faster initial and slower subsequent autocorrelation decay than the single-OU model.

All results were obtained using the non-centered algorithms of Section 6.9 and Section 6.13. The mixing of the algorithm for the superposition of the OU processes is not very satisfactory. This is not very surprising given that we have already commented (in Section 6.13.1) that the algorithm mixes slowly for parameter values as those estimated for the exchange rate data.

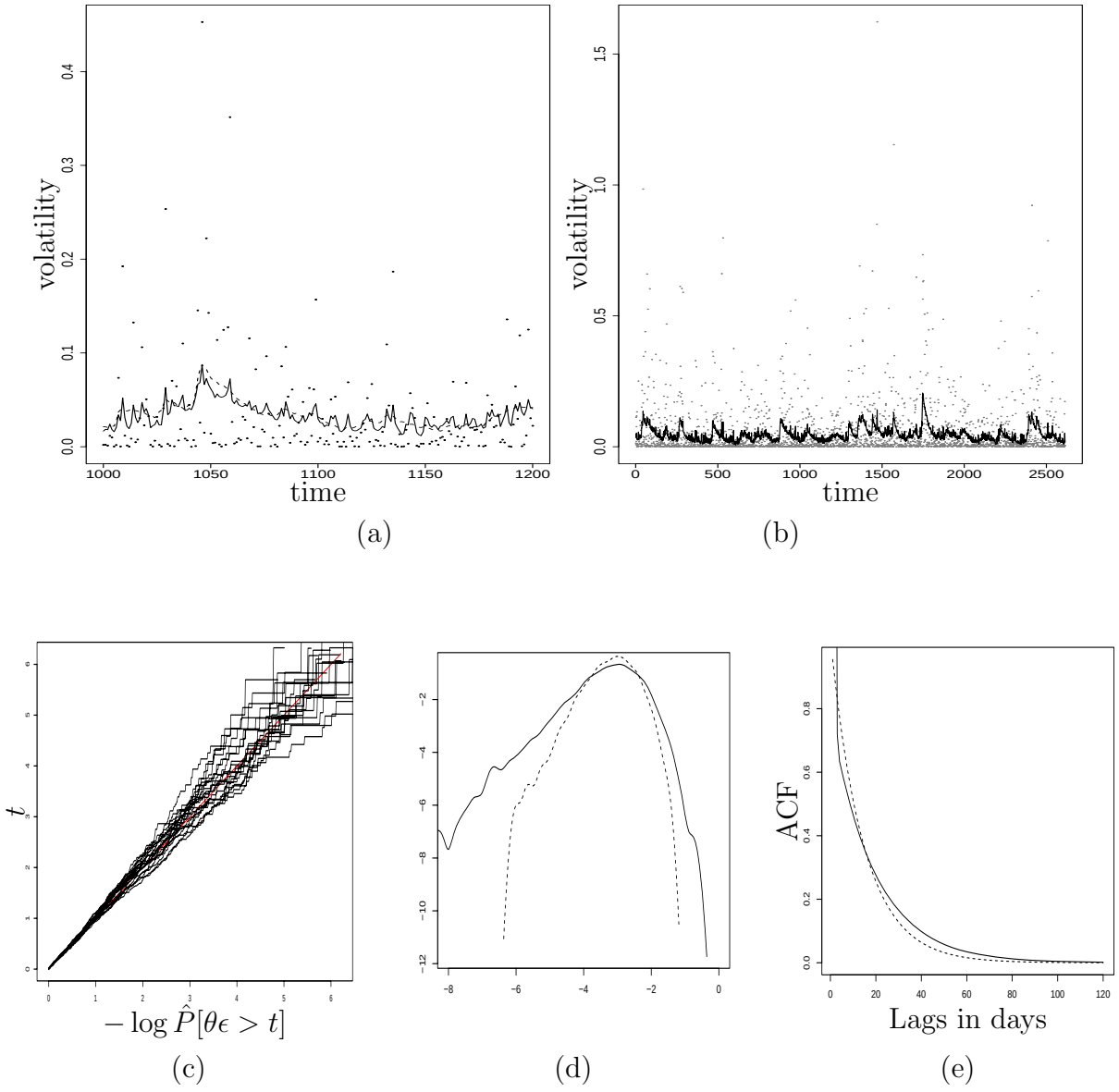


Figure 6.29: Results from fitting the OU models to the US$/DM data. In all plots dashed and solid lines corresponds to results obtained under the single-OU and the two-OU models respectively. Top: pointwise posterior median v_n^*: (a) short period of time for the single-OU and the two-OU models. (b) long period of time for the two-OU model; the grey dots correspond to y_n^2. Bottom: (c) diagnostic of Section 6.14 applied to the jump sizes from the single-OU model. (d) kernel estimates of the log-predictive density of $\log(v_n^*)$. (e) posterior median of the ACF of the series $\{v_n^*\}$.

6.17 Extensions and further work

This section deals with issues not covered in this chapter. We discuss how the methodology we have developed can be extended to cope with these problems.

6.17.1 Drift and risk premium

Our MCMC methodology can easily cope with the extension where the drift and risk premium terms have been added to the model. Thus, the SDE of the log-price conditionally on the volatility process is given in (6.3) with non-zero α, β. For reasons of interpretability and parameter orthogonality we consider the reparameterisation $(\alpha, \beta) \rightarrow (u, \beta)$ where $u = \alpha + \beta\xi$ and ξ is the mean of the volatility process. The SDE of the price process can now be written as

$$dx(t) = udt + \beta(v(t) - \xi)dt + v(t)^{1/2}dB(t),\ t \in [0, T]. \tag{6.52}$$

This reparameterisation is common in regression models, see for example Chapter 9 of O'Hagan (1994). It follows that (see Section 6.2 for definitions of the quantities appearing below)

$$y_n = u\Delta + \beta(v_n^* - \mathsf{E}\{v_n^*\}) + (v_n^*)^{1/2}\epsilon_n \tag{6.53}$$

where $\epsilon_n,\ n = 1, 2, \ldots$ is a sequence of independent standard normal random variables and independent of the volatility process. Notice that the distribution of

$$v_n^* = \int_{(n-1)\Delta}^{n\Delta} v(s)ds$$

depends on the time lag Δ chosen. Therefore,

$$\begin{aligned} \mathsf{E}(y_n) &= u\Delta \\ \text{Var}\,(y_n) &= \mathsf{E}(v_n^*) + \beta^2\text{Var}\,(v_n^*) \\ \text{Cov}(y_n, y_k) &= \beta^2\text{Cov}(v_n^*, v_k^*),\ k \neq n \end{aligned}$$

Notice that when the β term is not zero then the returns are positively correlated. This dependence is implicitly induced by their correlation with the actual volatility process. Specifically,

$$\text{Cov}(y_n, v_n^*) = \beta\text{Var}\,(v_n^*) \tag{6.54}$$

thus

$$\text{Cov}(y_n, y_k) = \beta \text{Cov}(y_n, v_n^*)\text{Corr}(v_n^*, v_k^*), \ n \neq k.$$

(6.54) relates with the interpretation of β as a risk premium. The risk premium is a measure of discrepancy between the marginal and the conditional characteristics of the returns and it can be defined as the following function of the unobserved volatility:

$$\text{risk premium} \quad = \quad \frac{\mathsf{E}(y_n) - \mathsf{E}(y_n \mid v_n^*)}{\mathsf{E}(v_n^*) - v_n^*}. \tag{6.55}$$

In (6.52) the risk premium coincides with β. Under this formulation the risk premium is not invariant under scaling of the data. In the discussion of Barndorff-Nielsen and Shephard (2001) the authors argue that they use this specification (instead for example using the standard deviation in the definition above) for reasons of analytical tractability; see Barndorff-Nielsen and Shephard (2001) for more references and alternative models.

β also controls the skewness of the marginal distribution of the returns. The coefficient of skewness of a random variable X is

$$\text{skew}(X) = \frac{\mathsf{E}\{(X - \mathsf{E}(X))^3\}}{(\text{Var}\,(X))^{3/2}}$$

hence it can be shown by first conditioning on v_n^* that

$$\text{skew}(y_n) = \beta \, \frac{\beta^2 \mathsf{E}[(v_n^* - \xi\Delta)^3] + 3\text{Var}\,(v_n^*)}{(\mathsf{E}(v_n^*) + \beta^2 \text{Var}\,(v_n^*))^{2/3}}$$

which shows that the returns distribution is skewed in the presence of $\beta \neq 0$. Moreover, the sign of the skewness is that of β.

It was first noted in Section 6.1 that economic theory suggests $\beta > 0$, so that investors are compensated for the risks they undertake. On the other hand, the empirical studies reviewed in Section 6.1 reveal negative skewness for many financial returns, which implies a $\beta < 0$ for the model (6.52). Thus, it seems reasonable to assign a prior on β which gives mass on the whole of the real line. We specify Gaussian priors for both u and β, a choice which turns out to be computationally convenient as well.

Under the full model (6.52), together with the OU specification for the volatility (6.9), Bayesian inference is concerned with the joint posterior distribution of the parameters of the volatility process, u and β. When working with background driving compound Poisson processes (Section 6.4) we adopt the augmentation scheme proposed in Section 6.6. We are interested in designing an MCMC algorithm which samples from the joint distribution of the missing data and the drift and volatility parameters. This can be done in a componentwise-updating algorithm which updates the missing data and volatility parameters conditionally

upon the drift parameters, and the drift parameters conditionally upon the rest. The first step can be implemented as described in Section 6.7 or Section 6.9, depending on whether a centered parameterisation is preferred or not. The second step can be done by direct simulation. The missing data and the volatility parameters uniquely determine the integrated volatility series $\{v_n^*\}$. Conditionally on this series, (6.53) defines a regression model with heteroscedastic errors and where $\{v_n^*\}$ is the vector of covariates. Under Gaussian priors, the posterior distribution of u, β, which are the regression coefficients, is Gaussian and can be easily derived as shown in Lindley and Smith (1972).

6.17.2 Leverage effect and non-integrable Lévy measures

Section 6.1 discussed the leverage effect observed and explained by Black (1976), according to which low returns in equities markets increase future volatility. In terms of the observables negative correlation exists between y_n and subsequent $y_k^2, k > n$. In an SV framework it is not possible for the price process to feed into the variance process, therefore the leverage effect has to be incorporated into the model in the reverse way: by modifying the model to allow the feedback of the innovations of the volatility process into the price process. To this end, Barndorff-Nielsen and Shephard (2001) propose the following model

$$dx(t) = udt + \beta(v(t) - \xi)dt + v(t)^{1/2}dB(t) + \rho d\overline{z}(t),\ t \in [0, T]$$

where

$$\overline{z}(t) \quad = \quad z(t) - \mathbb{E}(z(t)).$$

It is expected that $\rho < 0$, so that large positive innovations lead to negative returns. Notice that $\overline{z}(\cdot)$ is a Lévy process and a martingale. Since this construction works directly with the BDLP, our augmentation methodology and parameterisations could naturally be extended to cope with the leverage effect model.

A fascinating problem is inference for OU models driven by Lévy processes with non-integrable Lévy measures. These processes do not correspond to compound Poisson processes, thus our methodology needs to be carefully modified. The main complication is due to the fact that the Poisson process Ψ corresponding to the Lévy process $z(\cdot)$ is not locally-finite, which means that its mean measure is not σ-finite. It is not straightforward to handle computationally these objects, neither is it simple to find the conditional distribution of the parameters given the Lévy process. Often, Ψ will contain infinite information about some of the parameters. This implies that an NCP is necessary, but Section 5.8.1 discussed the difficulties which arise when constructing NCPs for Poisson processes with non-σ-finite mean measure. The usual practice is to approximate Ψ with a finite process (see Wolpert

and Ickstadt (1998) for an example) and Rosinski (2002) reviews some related methods. Our methodology is directly applicable if such an approximation is adopted. Actually, the findings of Section 6.12.2 are promising, since they suggest that an MPP-CDF-NCP might be very efficient.

Chapter 7

Partially Non-centered parameterisations

7.0 Introduction

This chapter introduces partially non-centered parameterisations (PNCPs). This is a new class of parameterisations which lie on a continuum between the CP and the NCP. This construction is motivated by the aspiration to construct MCMC algorithms robust to the information content of the data, so that the user does not have to choose beforehand whether to use a CP or an NCP. In fact we show that in the context of normal hierarchical models, there exist partially non-centered Gibbs sampling algorithms (PNCAs) which outperform both the CA and the NCA and there is one which is the optimal Gibbs sampling algorithm producing IID draws from the posterior distribution of (X, Θ). Moreover, PNCPs for general non-Gaussian models are constructed. A discussion is provided on the relevance of this methodology to other augmentation techniques proposed in the literature. In particular we establish and explore the connections with parameterisations which minimise the posterior correlation between X and Θ, and with the conditional and marginal augmentation. We address the issue of optimisation of a PNCP and conclude the chapter with some examples where this methodology has successfully been applied to. Some of the material in this chapter is based on Section 4 of Papaspiliopoulos et al. (2003).

7.1 Partial non-centering of hierarchical models

We have extensively discussed two alternative parameterisations for hierarchical models: the centered and the non-centered. The corresponding graphical models are shown in Figure 1.3 and Figure 1.6. These parameterisations were designed to be used in conjunction with data augmentation methods, and we termed the corresponding Hastings-within-Gibbs sampling

algorithms the centered (CA) and the non-centered (NCA) algorithm, see Section 4.1. The CA is optimal in the limiting case where the data Y are very informative about the missing data X. On the contrary, the NCA is optimal in the limiting case where there are no observed data at all, since it produces IID observations from the joint prior distribution of the transformed missing data $\tilde{X}$ and Θ.

A natural question is whether it is possible at all to construct a family of parameterisations which contains the CP and the NCP as special cases. We wish that this family contains parameterisations which for a given dataset are preferable to both the CP and the NCP. It is also desirable to develop techniques which allow the data to choose somehow automatically a particular member of this family, so that for example the user does not have to choose beforehand whether to use a CP or an NCP.

This chapter introduces such a family of parameterisations which lie on a continuum between the CP and the NCP. We call the members of this family partially-non-centered parameterisations (PNCP). Section 7.2 describes the methodology for the normal hierarchical model, where convergence rates of the proposed parameterisation can be analytically computed. However, this model is used mostly for pedagogical purposes, since our aim is to apply this methodology to non-Gaussian models and this is discussed in Section 7.5.

7.2 PNCP for the normal hierarchical model

We have already defined the CP and NCP for the normal hierarchical model in Section 2.3 (see (2.14) and (2.18) respectively). Section 2.3 showed that the rate of convergence of the associated Gibbs sampler algorithms, the CA and the NCA respectively, depends on the angle (in an $\mathcal{L}^2$ sense) between Θ and X (in the CA) and Θ and $\tilde{X}$ (in the NCA). It is easy to show that when Θ has the improper uniform prior then $\mathrm{Cov}(X, \tilde{X} \mid Y) = 0$, therefore X and $\tilde{X}$ are orthogonal *a posteriori* . Hence, the angle between Θ and X is complementary to that between Θ and $\tilde{X}$ (see Figure 7.1) and as a consequence $\rho_c = 1 - \rho_{nc}$. This geometric interpretation motivates the partially non-centered parameterisation (PNCP): some linear combination between X and $\tilde{X}$, $\tilde{X}^{(w)}$ say, will be orthogonal to Θ *a posteriori* , and the corresponding Gibbs sampler will produce IID samples; see Figure 7.1.

We now proceed to show how we can obtain such a reparameterisation. Consider the following alternative parameterisation for the normal hierarchical model,

$$\begin{aligned} Y_i &= w\Theta + \tilde{X}_i^{(w)} + \sigma_y \epsilon_i \\ \tilde{X}_i^{(w)} &= (1-w)\Theta + \sigma_x z_i, \ i = 1, \ldots, m. \end{aligned} \tag{7.1}$$

where w is a fixed number in $[0, 1]$. We will refer to w as the working parameter, borrowing

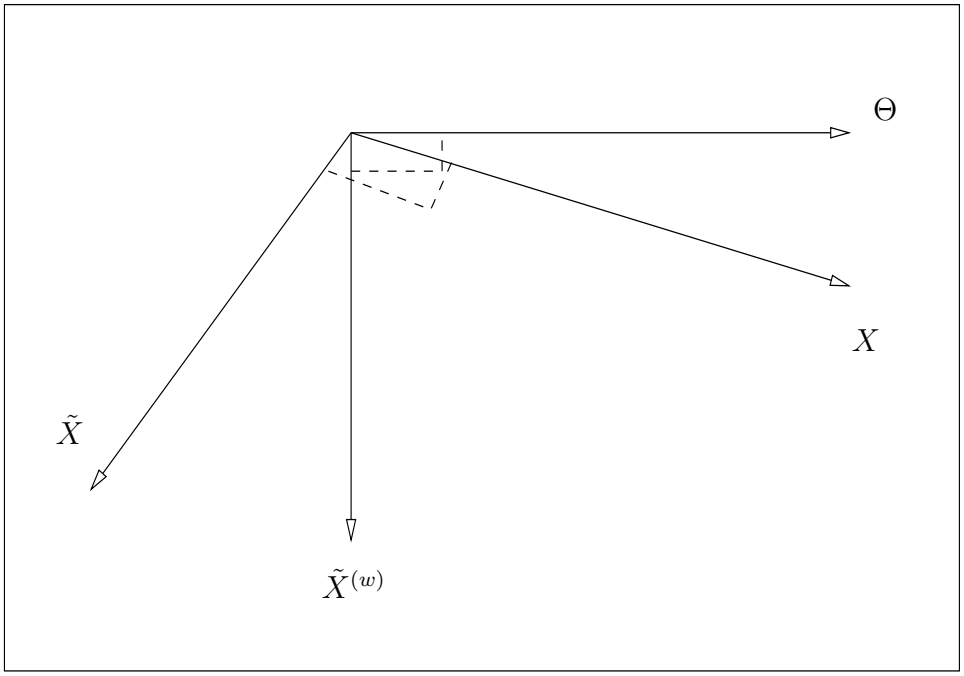


Figure 7.1: The geometry of the normal hierarchical model.

the terminology from the conditional augmentation literature, see Section 7.9.1. Therefore

$$\tilde{X}_i^{(w)} = X_i - w\Theta \tag{7.2}$$

and

$$\tilde{X}_i^{(w)} = (1-w)X_i + w\tilde{X}_i \tag{7.3}$$

where X_i and $\tilde{X}_i$ are defined in Section 2.3, thus (7.1) defines a family of parameterisations with the CP at one extreme, for $w = 0$ and the NCP at the other, for $w = 1$. Notice that (7.1) is over-parameterised, since w is not identifiable from the observed data, in the sense that $Y \mid \Theta, w$ is the same for all values of w.

The joint posterior distribution of $\tilde{X}^{(w)} = (\tilde{X}_1^{(w)}, \ldots, \tilde{X}_m^{(w)})$ and Θ is still Gaussian, since $\tilde{X}^{(w)}$ is the linear transformation of X and Θ given in (7.2); see also Section 7.3.1 for some useful relevant expressions. Therefore we can calculate the rate of convergence of the Gibbs sampler under this parameterisation using the general results of Section 2.1.1, which is given below and plotted against w (for a specific value of κ) in Figure 7.2:

$$\rho_{pnc}(w) = \frac{(w-(1-\kappa))^2}{w^2\kappa + (1-w)^2(1-\kappa)}; \tag{7.4}$$

κ was defined in (2.16) as $\kappa = \sigma_x^2/(\sigma_y^2 + \sigma_x^2)$. Recall from (2.15) and (2.19) respectively that $\rho_c = 1 - \kappa$ and $\rho_{nc} = \kappa$. These two rates correspond, as we would expect from (7.3), to

$\rho_{pnc}(0)$ and $\rho_{pnc}(1)$ respectively. It can be easily shown and is also depicted in Figure 7.2

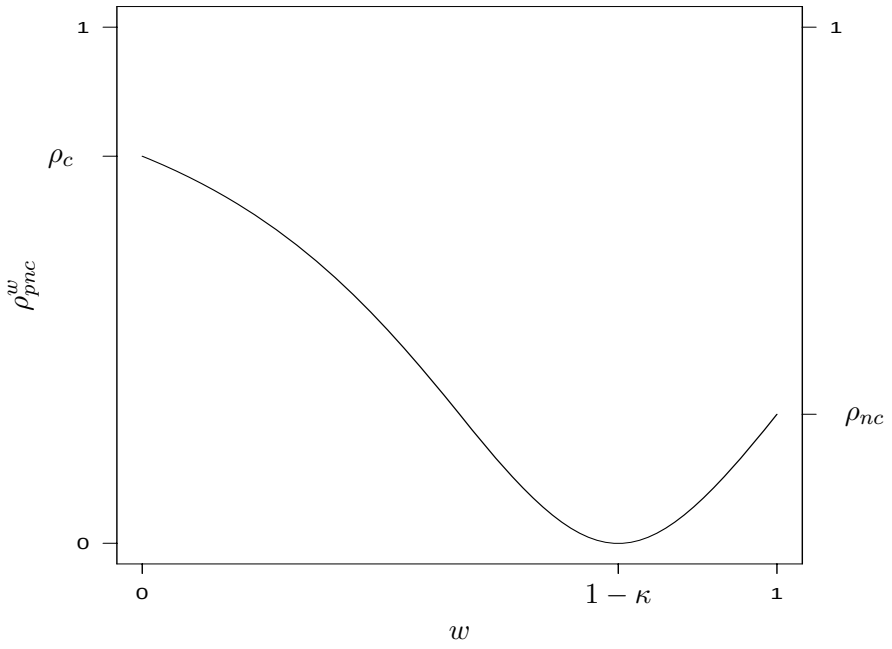


Figure 7.2: Rate of convergence $\rho_{pnc}(w)$ for the PNCP on the normal hierarchical model. In this example we have taken $\sigma_x^2 = 1,\ \sigma_y^2 = 3$, thus $\kappa = 1/4$.

that

$$\rho_{pnc}(w) \leq \max(\rho_c, \rho_{nc}),\ \forall w \in (0,1)$$

and

$$\rho_{pnc}(w) = 0 \text{ for } w = 1-\kappa.$$

Therefore the PNCA (7.1) can outperform both CP and NCP, but also it can be tuned appropriately to produce IID samples, by setting $w = 1-\kappa$. Moreover, $\rho_{pnc}(w) \geq \min(\rho_c, \rho_{nc})$ for all values of w outside the unit interval.

Clearly, since $(\tilde{X}^{(w)}, \Theta)$ are jointly Gaussian there exists linear transformations which make them uncorrelated and consequently independent. One such transformation is exactly described by (7.2), when w is set equal to $1-\kappa$. The fact that $w = 1-\kappa$ makes $\tilde{X}^{(w)}$ and Θ independent is not surprising taking into account the weighed average form of the expectation of X_i conditional on Y_i and Θ given in (2.17).

Therefore, it may seem that the PNCP is just an alternative way to achieve the diagonalisation of the covariance matrix of a Gaussian vector. However, the novelty of this method lies in the way that the optimal algorithm is found by "interpolating" between the CA and the NCA, as in (7.3) or in (7.2). This construction can easily be extended to non-Gaussian models, as we shall see in Section 7.5. In general, the PNCP will involve non-linear transfor-

mations of the missing data X and Θ and will be suggested by the way $(\tilde{X}, \Theta)$ is transformed to X.

7.3 PNCP for the general normal hierarchical model

We now extend the results of Section 7.2 to the general normal hierarchical model introduced in Section 2.4. Consider the following alternative parameterisation for the model in (2.21)

$$\begin{aligned} Y_i &= C_{1i}(\tilde{X}_i^{(w)} + W_i\Theta) + (\sigma_i^2 I_{n_i})^{1/2}\epsilon_i \\ \tilde{X}_i^{(w)} &= (I_p - W_i)C_2\Theta + D^{1/2}z_i \end{aligned} \tag{7.5}$$

where W_i is a $p \times p$ matrix, which implies that

$$\tilde{X}_i^{(w)} = X_i - W_i C_2 \Theta \tag{7.6}$$

When W_i is the identity (7.5) becomes the non-centered parameterisation while when it is the null (7.5) becomes the centered. Using the results of (2.22) and (2.26) it can be easily seen that

$$\begin{aligned} \mathrm{Cov}(\tilde{X}_i^{(w)}, \Theta \mid Y) &= 0, \text{ for all } i = 1, \ldots, m, \text{ when} \\ B_i^{-1} &= \sigma_y^{-2} C_{1i}^T C_{1i} + D^{-1} \\ W_i &= B_i D^{-1} \end{aligned} \tag{7.7}$$

which suggests that the corresponding Gibbs sampler will produce IID samples for this choice of the weight matrix. This can be formally proved by showing first that the inverse variance matrix of $(\Theta, \tilde{X}^{(w)})$ is a block diagonal and therefore the corresponding B-matrix (defined in Section 2.1.1) is the null matrix.

Section 2.4 highlighted the importance of the W_i matrix defined in (7.7) in assessing the convergence properties of the CA for the general normal hierarchical model.

Notice that in (7.5) the proportion of Θ subtracted from each X_i varies with i unlike (7.1), reflecting the varying informativity of each Y_i about the underlying X_i present in (7.5).

7.3.1 Full conditional distributions

For completeness we give here the full conditional distributions that are essential for Gibbs sampling under the PNCP for the general normal hierarchical model.

The $\tilde{X}_i^{(w)}$ are conditionally independent given Θ and their distributions are readily available using (7.6) and the conditionals for the X_is given in (2.25). More involved is the derivation of the conditional distribution for Θ. We begin by giving the precision matrix of

$(\tilde{X}^{(w)}, \Theta)$:

$$Q^{pnc} = \begin{pmatrix} Q_1 & \mathbf{0} & \dots & \mathbf{0} & -(Q_1 W_1 + D^{-1}) C_2 \\ \mathbf{0} & Q_2 & \dots & \mathbf{0} & -(Q_2 W_2 + D^{-1}) C_2 \\ \vdots & \vdots & \ddots & \dots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & Q_m & -(Q_m W_m + D^{-1}) C_2 \\ -C_2^t (W_1^t Q_1^t + D^{-1}) & -C_2^t (W_2^t Q_2^t + D^{-1}) & \dots & -C_2^t (W_m^t Q_m^t + D^{-1}) & Q_\Theta \end{pmatrix}. \tag{7.8}$$

where

$$Q_\Theta \;=\; C_2^t \sum_{i=1}^{m} \left\{ W_i^t Q_i W_i + D^{-1} W_i + W_i^t D^{-1} + D^{-1} \right\} C_2; \tag{7.9}$$

see Section 2.4 for definitions of the matrices involved in the above expressions. This result can be obtain by first noticing that $(\tilde{X}^{(w)}, \Theta)$ is a linear transformation of (X, Θ), whose precision matrix is given in (2.30), and then using the standard results for deriving covariance matrices of linear transformations. The conditional posterior of Θ is Gaussian, with precision Q_Θ and mean vector M_Θ which is derived from (7.8), (2.22) and (2.26) after some algebra.

7.4 PNCP with proper priors

An issue which was raised in the discussion of the paper of Papaspiliopoulos et al. (2003), where the PNCP was first introduced, concerns the effect of the prior on Θ on the construction and the rate of convergence of the PNCP. This issue is addressed here in the context of the simple normal hierarchical model of Section 2.3. The prior we assumed for Θ in Section 7.2 was the improper uniform, $\pi(\Theta) \propto 1$. Suppose that we choose the proper and conjugate prior

$$\Theta \sim \mathrm{N}(\mu, \tau^2)$$

instead, where by convention the limit of this distribution as $\tau^{-2} \to 0$ is taken to be the improper uniform. See Section 2.3.2 for a convergence rate analysis of the CA and the NCA for this model.

The PNCP remains the same as in (7.2), but the rate of convergence of the Gibbs sampler under this parameterisation changes with respect to (7.4), and becomes

$$\rho_{pnc}(w) = \frac{(w - (1-\kappa))^2}{w^2 \kappa + (1-w)^2 (1-\kappa) + \sigma_x^2 (1-\kappa)(1/\tau^2)} \tag{7.10}$$

where κ is defined in (2.16). From this it follows that the optimal value of w, which is $1-\kappa$, is not affected by the choice of the prior, however the convergence rate is uniformly better for all w the smaller the prior variance of Θ. This is graphically depicted in Figure 7.3, where we plot $\rho_{pnc}(w)$ against w for fixed κ but different values of τ.

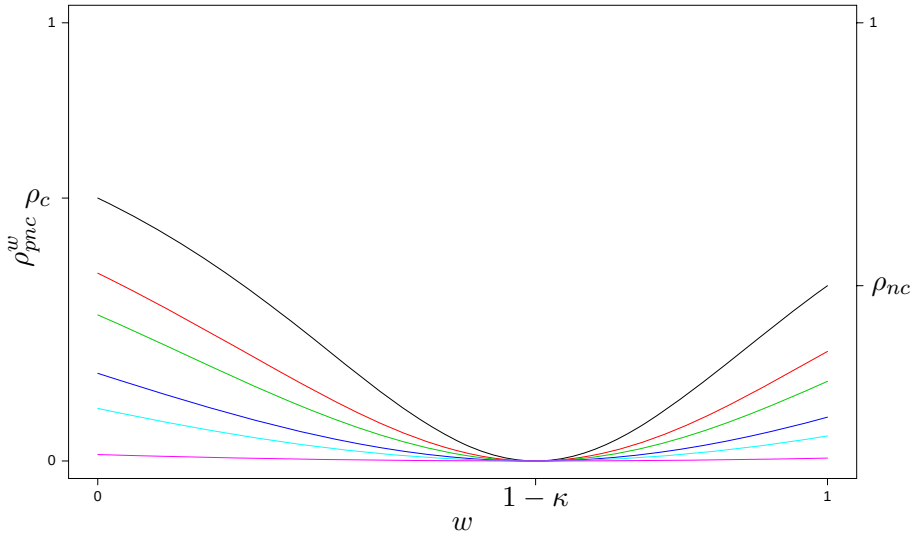


Figure 7.3: Rate of convergence $\rho_{pnc}(w)$ different values of $1/\tau^2 = 0, 1, 2, 5, 10, 100$, when $\sigma_x = 1,\ \kappa = 0.4$. Curves that lie above others correspond to smaller values of $1/\tau^2$.

7.5 PNCP outside the Gaussian context

Partial non-centering can be used for many models outside the Gaussian context. In general there is no unique way of defining a continuum of partial non-centering strategies. However often there will be a natural one suggested strongly by the model structure, and the way $(\tilde{X}, \Theta)$ is transformed to X. For example, when Θ is a location parameter for the prior distribution of X then we typically take

$$\begin{aligned} X &= \tilde{X} + \Theta \\ X &= \tilde{X}^{(w)} + w\Theta \end{aligned}$$

while if Θ is a scale parameter we could choose

$$\begin{aligned} X &= \Theta\tilde{X} \\ X &= \Theta^w\tilde{X}^{(w)} \end{aligned}$$

with $w \in [0, 1]$ in both cases. Such choices become less obvious when working with state-space expanded NCPs but Section 7.6 gives some ideas in that direction.

Outside the Gaussian context, it is rare that pure Gibbs sampling can be used in conjunction with a PNCP, so that as with the NCP and often the CP, appropriate Hastings-within-Gibbs strategies will be necessary. Thus the PNCP cannot be expected to produce IID observations from the target distribution for any w. The general MCMC single-component updating algorithm based on the PNCP is described below.

A Hastings-within-Gibbs to sample from $(\Theta, \tilde{X}^{(w)}) \mid Y$

Iterate the following steps:

1. Update Θ according to $\pi(\Theta \mid \tilde{X}^{(w)}, Y)$
2. Transform $(\Theta, \tilde{X}^{(w)}, w) \to X$
3. Update X according to $\pi(X \mid \Theta, Y)$
4. Transform $(\Theta, X, w) \to \tilde{X}^{(w)}$.

When working with non-Gaussian models, it turns out that the really challenging problem is how to find values of w which lead to substantially better algorithms than both the CA and the NCA. This choice is straightforward for the normal model, since we can carry out analytic calculations. In more general contexts though, this choice is less obvious. We note however, that sensitivity of algorithmic performance to data is very common in many classes of hierarchical models, since it is often the case that the information about X_i contained in Y_i will depend on Y_i. Therefore, to extend the PNCP to other models in an efficient way, we will need to allow w to vary across i, as in (7.5), and possibly be a function of the corresponding data Y_i. This problem will be investigated in Section 7.10 after we establish the connection between the PNCP and the so-called conditional augmentation. Examples of the PNCP methodology applied to non-Gaussian models will be given in Section 7.11.

We close this section by highlighting an advantage of seeking PNCPs for complex models, rather than just working with the CP or the NCP. For the simple normal hierarchical model of Section 2.4 we showed that $\rho_c = 1 - \rho_{nc}$, which implies that when the one algorithm is very good the other is very poor. Therefore in such a scenario, for example when κ is close to 0 or 1, there is not a big advantage in using the PNCP instead of the best between the CP and the NCP.

When the CA and the NCA have similar rate of convergence then they are both doing very well, since their convergence rates are around 0.5. Of course we are not interested in investing

a lot of set up and computational time to improve on an algorithm whose convergence rate is as good as 0.5. This is why we argued in Section 7.1 that the construction for the simple normal model serves mostly as an illustration and motivation to the PNCP, rather than as a means to itself. However, in other models it is not true that $\rho_c = 1 - \rho_{nc}$, even inside the Gaussian family, see Section 2.4 and in particular model (2.38) for an example. In these models, it might be the case that both algorithms are poor and therefore substantial improvements can be achieved by using a PNCP.

7.6 State-space expanded PNCPs

Designing parameterisations which lie on a continuum between a centered and a state-space-expanded non-centered parameterisation, is rather challenging as this section demonstrates. Moreover, "obvious" choices might lead to a reducible Gibbs sampler and such an example will be given in the sequel. The results of this section are at a preliminary stage, therefore we will just focus on a specific example.

Suppose that $X \sim \text{Ga}(\Theta, 1)$ and the state-space-expanded NCP proposed in Section 4.2 has been adopted, that is $\tilde{X}(\cdot)$ is a gamma process and $X = \tilde{X}(\Theta)$. This construction is thoroughly studied in Section 4.2, where details can be found. Therefore the NCP takes $\tilde{X}$ to be a standard gamma process and X is obtained as its value at the unknown time Θ. We could instead obtain X as the value of a gamma process with unknown shape parameter Θ at the fixed time 1. This observation suggests that a PNCP could be constructed by taking $\tilde{X}^{(w)}(\cdot)$ to be a gamma process with rate Θ^w and

$$X = \tilde{X}^{(w)}(\Theta^{1-w});$$

see also Figure 7.4. However, this parameterisation would lead to a reducible Gibbs sampler, since it can be shown that $\tilde{X}^{(w)}$ contains infinite information about Θ, see Section 1.8.

An alternative parameterisation which can be easily implemented is based on the observation that, due to the infinite divisibility of the gamma distribution,

$$X = X_1 + X_2,\ X_1 \sim \text{Ga}(w\Theta, 1),\ X_2 = \text{Ga}((1-w)\Theta, 1).$$

We then take $\tilde{X}(\cdot)$ be a standard gamma process, X_2 be as defined above and the PNCP is given by

$$X = \tilde{X}(w\Theta) + X_2.$$

Therefore, X is obtained as the value of a gamma process at time $w\Theta$ started from a random point X_2 at time zero; see Figure 7.4 for a graphical illustration of this method. Intuitively, if the information in the data Y about X is large compared to that about Θ, then there is

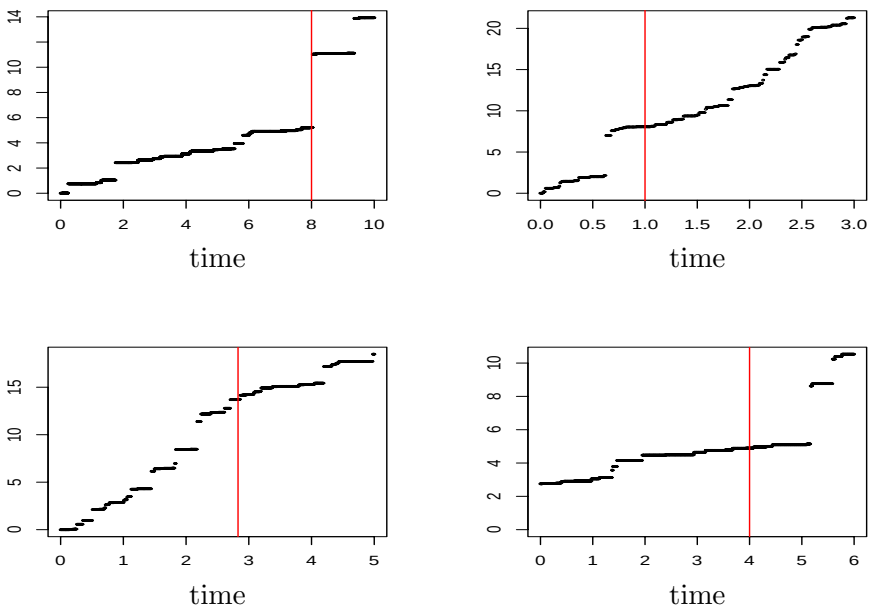


Figure 7.4: Different ways to simulate $X \sim \text{Ga}(\Theta, 1)$ using gamma processes. We use the general notation $\tilde{X}_{\alpha,\beta}$ for a gamma process with shape parameter α and scale parameter β. $X = \tilde{X}_{1,1}(\Theta)$ (top left), $X = \tilde{X}_{\Theta,1}(1)$ (top right), $X = \tilde{X}_{\Theta^w,1}(\Theta^{1-w})$ (bottom left), $X = \tilde{X}_{1,1}(w\Theta) + X_2$, $X_2 \sim \text{Ga}((1-w)\Theta, 1)$ (bottom right). In this example we have taken $\Theta = 8$ and $w = 1/2$.

much more information about where to start the gamma process from, that is about X_2, rather than about which time we should stop $\tilde{X}$, that is $w\Theta$. In this case a w close to 0 is preferable. If the marginal information about Θ is large compared with the information about the value of X, a value of w close to 1 will be preferred. Clearly, the CP is obtain for $w = 0$ and the NCP for $w = 1$.

Implementation of this scheme is simple; we would typically use a three-component Gibbs sampler, which updates in turn Θ, X_2 and $\tilde{X}$ from their full conditionals. The conditional distributions of $\tilde{X}$ and X_2 are known explicitly, and a Metropolis-Hastings step is used to update Θ. The algorithm is very similar to the one described in Section 4.2.

Choosing a w which results in an algorithm with faster convergence than both the CA and the NCA is rather difficult. We are currently at a preliminary stage, and although Section 7.10 makes some suggestions, considerable more work has to be done in this direction.

7.7 PNCP and correlation analysis

Section 7.2 constructed the PNCP for the simple normal model and showed that it is optimised when $w = 1 - \kappa$. For this value, $\tilde{X}_i^{(w)}$ is uncorrelated with Θ *a posteriori* . Thus, it is of interest to know how the PNCP relates to parameterisations which try to achieve

posterior uncorrelated-ness between the missing data and the parameters by means of linear transformations. This issue was raised in the discussion of Papaspiliopoulos et al. (2003), where it was suggested that in many cases it is sufficient to consider the posterior correlation structure when constructing a reparameterisation.

In this thesis, our approach is different. We are primarily interested in constructing and assessing the performance of CPs and NCPs. Of course there are many situations where neither of these methods is adequate, and when a CP and an NCP exist, we try to find ways to construct intermediate parameterisations that adapt according to the information in the data, so that the user does not have to choose *apriori* between the two extremes.

It is true that the PNCP for the normal hierarchical model in Section 7.2 and Section 7.3 can also be obtained using a posterior uncorrelated-ness argument because in the Gaussian context the maximal correlation described in Amit (1991) (see also Section 2.1 of this thesis) is attained by linear functions. The PNCP coincides with parameterisations which are based on *a posteriori* uncorrelated-ness in other applications as well. This is for example the case with the reparameterisation designed for the geostatistical model in Section 7.11, which mimics the construction for the normal hierarchical model.

However, outside the Gaussian context it is possible for posterior correlations to be very unreliable for the purposes of predicting convergence properties; see for example Roberts (1992), where it is shown that a two-component Gibbs sampler with uncorrelated components can be reducible, while if correlation is induced between the updated components irreducibility is achieved. Of course that example is extreme and contrived, since the target distribution for the Gibbs sampler is the uniform on two disjoint subsets of the $\mathbb{R}^2$.

Section 7.8 gives a much more realistic example where posterior uncorrelated-ness is not enough to improve on the convergence of the CA. In particular, we study reparameterisations for generalised hierarchical mixed models. Such models are described in Section 3 of Gelfand et al. (1996), for example. We show that there exist parameterisations under which the missing data and the parameters are uncorrelated, but which possess convergence properties inferior to the CP (due to higher order dependence between the random effects and the parameters).

As a final general remark, we find unclear how to generalise the construction based on minimising posterior correlation to models where the missing data live on non-Euclidean spaces, as for example those considered in Chapter 5 and Chapter 6.

7.8 Reparameterisations for GLHM

This section proposes a reparameterisation of (X, Θ) in the context of generalised linear hierarchical models (GLHM). It is based on the weighted average form that the posterior expectation of X_i given the parameters Θ and the data Y_i takes. Before giving the details of

this construction we review some relevant theory about exponential families and generalised linear hierarchical models.

7.8.1 Natural exponential family with quadratic variance function

In this section we will give a brief summary of the main ideas underlying the models we will be interested in, that is models in the Natural Exponential Family with Quadratic Variance Function (NEF-QVF). The material of this section is based on Morris (1983).

A random variable Y has a distribution in the NEF with *natural* parameter $\omega \in \Omega \subset \mathbb{R}$ if

$$P[Y \in A] = \int_A \exp\{y\omega - b(\omega)\} F(dy)$$

for some Stiltjes measure $F(\cdot)$ independent of ω and $A \subset R$.

$$\begin{aligned} \mathsf{E}(e^{uY}) &= \int_R \exp\{uy\} \exp\{y\omega - b(\omega)\} F(dy) \\ &= \int_R \exp\{y(\omega + u) - b(\omega)\} F(dy) \end{aligned}$$

therefore

$$\begin{aligned} K(u; Y) &= \log \mathsf{E}(e^{uY}) = b(u + \omega) - b(\omega) \\ \mathsf{E}(Y \mid \omega) &= b'(\omega) =: X \\ \mathrm{Var}\,(Y \mid \omega) &= b''(\omega) =: V(X). \end{aligned}$$

The Variance Function (VF) $V(\cdot)$ characterises the sub-family of NEF. The NEF-QVF is characterised by the property

$$V(X) = v_0 + v_1 X + v_2 X^2$$

and includes the following distributions: Gaussian, Poisson, gamma, binomial and negative binomial; see table 1 of Morris (1983). The family is closed under convolution and location and scale transformations.

7.8.2 GLHM

We will look at models where Y_i comes from the one-parameter NEF with convolution parameter n_i (thus, this setting incorporates the possibility of multiple observations) and

scale parameter $1/\phi_i$ considered either known or intrinsically specified

$$\pi(Y_i \mid z_i) \propto \exp\{n_i[Y_i z_i - b(z_i)]/\phi_i\}.$$

These models are considered in Gelfand et al. (1996). The random effects z_i are assigned the typical conjugate prior in the NEF

$$\pi(z_i \mid \Theta) \propto \exp\{n_0[\Theta z_i - b(z_i)] - g(\Theta, n_0)\}.$$

We transform $z_i \to X_i$, where $X_i = b'(z_i)$ and since this transformation is generally non-linear the resulting distribution for X_i is in the exponential family but not in the NEF. Then under some conditions (Theorem 5.2 of Morris (1983)) and assuming that Y_i is in the NEF-QVF, it follows that

$$\begin{aligned} \mathsf{E}(X_i \mid \Theta) &= \Theta \\ \text{Var}\,(X_i \mid \Theta) &= \mathsf{E}(\frac{1}{n_0} V(X_i)) = (n_0 - v_2)^{-1} V(\Theta). \end{aligned}$$

It is also true that the posterior mean of X_i conditional on Θ admits a weighted average form

$$\mathsf{E}(X_i \mid \Theta, Y_i) = w_i \Theta + (1 - w_i) Y_i \tag{7.11}$$

$$w_i = \frac{n_0}{n_0 + n_i/\phi_i} \tag{7.12}$$

while the posterior variance takes the form

$$\text{Var}\,(X_i \mid \Theta, Y_i) = V(w_i\Theta + (1 - w_i)Y_i)/(n_0 + n_i/\phi_i - v_2).$$

An example of the weighted average form for the posterior mean in (7.12) was given in Section 2.3 for Gaussian models, see (2.17) (where $w_i = 1 - \kappa$ and κ is given in (2.16)).

7.8.3 Reparameterisation based on posterior correlations

Gelfand et al. (1996) term (X, Θ), $(X - \Theta, \Theta)$ the centered and the non-centered parameterisation respectively, although the latter is clearly not an NCP according to our definition in Section 4.1; $X - \Theta$ is independent of Θ only for Gaussian models.

It is intriguing to consider the parameterisation $(\tilde{X}^{(w)}, \Theta)$ where

$$\tilde{X}_i^{(w)} = X_i - w_i \Theta \tag{7.13}$$

which is motivated by the PNCP for the normal hierarchical model proposed in Section 7.2.

(7.13) does not qualify as a PNCP, since the definition in Section 7.1 requires that it has an NCP as a limit, although we do recover the CP when $w_i = 0$. On the other hand, for the special case of normal hierarchical models it is exactly the PNCP proposed in Section 7.2. To simplify exposition and avoid the introduction of unnecessary terminology, we will refer to $(\tilde{X}^{(w)}, \Theta)$ as a PNCP in this section.

A natural choice, driven by its role in (7.12) is to set

$$w_i = \frac{n_0}{n_0 + n_i/\phi_i}. \tag{7.14}$$

We shall shortly show that the posterior correlation between the missing data $\tilde{X}^{(w)}$ and Θ is zero, for this choice of w_i.

Direct calculations show that

$$\text{Cov}(X_i, \Theta \mid Y) = w_i \text{Var}(\Theta \mid Y).$$

Defining $b(n_0) := \text{Var}(\Theta \mid Y)$, and $c_i(n_0) := \mathsf{E}(\text{Var}(X_i \mid \Theta, Y_i))$ then

$$\begin{aligned}
\text{Corr}(X_i, \Theta \mid Y) &= \frac{w_i \text{Var}(\Theta \mid Y)}{[\text{Var}(\Theta \mid Y)\text{Var}(X_i \mid Y)]^{1/2}} \\
&= \frac{w_i \text{Var}(\Theta \mid Y)}{[\text{Var}(\Theta \mid Y)(c_i(n_0) + w_i^2 \text{Var}(\Theta \mid Y))]^{1/2}} \\
&= \frac{1}{[c_i(n_0)/(b(n_0)w_i^2) + 1]^{1/2}}
\end{aligned}$$

which is formula (3.4) of Gelfand et al. (1996). We can easily derive that

$$\text{Cov}(\tilde{X}_i^{(w)}, \Theta \mid Y) = \text{Cov}(X_i, \Theta \mid Y) - w_i \text{Var}(\Theta \mid Y) = 0 \tag{7.15}$$

which shows that this parameterisation manages to make the transformed random effects and the parameter uncorrelated. Nevertheless, is this enough to guarantee better convergence properties of the resulting sampler than the CA in a non-Gaussian model? Analytic convergence rates cannot be computed therefore we will try to empirically investigate this question in the context of a specific example.

7.8.4 Simulation results

Suppose we are interested in the following model written in a canonical form

$$\pi(Y_i \mid z_i) \propto \exp\{c[Y_i z_i + \log(-z_i)]\}, \; z_i \leq 0$$

together with the natural conjugate prior

$$\pi(z_i \mid \Theta) \propto \exp\{n_0[z_i\Theta + \log(-z_i)] + (n_0+1)\log(\Theta)\}$$

and where Θ is assigned a Gamma prior with known hyperparameters α, β. If we define $X_i = -1/z_i$ then we can re-write the model in the more familiar form

$$\begin{aligned} Y_i &\sim \text{Ga}(c, c/X_i) \\ X_i^{-1} &\sim \text{Ga}(n_0+1, n_0\Theta) \\ V(X_i) &= X_i^2/c,\ i = 1, \ldots, m. \end{aligned} \tag{7.16}$$

Due to conditional conjugacy the CP is very easy to implement since

1 $X_i^{-1} \mid Y_i, \Theta \sim \text{Ga}(n_0 + 1 + c, cY_i + n_0\Theta)$

2 $\Theta \mid X \sim \text{Ga}(m(n_0+1) + \alpha, n_0 \sum_i (1/X_i) + \beta)$.

We also consider the $(\tilde{X}^{(w)}, \Theta)$ parameterisation, where $\tilde{X}_i^{(w)} = X_i - w\Theta$. We use a common w for all $i = 1, \ldots, m$, since for the model (7.16) the expression (7.14) does not depend on i. The joint posterior distribution of $(\tilde{X}^{(w)}, \Theta)$ can be derived by a simple change of variables. Implementation of the Gibbs sampler under the $(\tilde{X}^{(w)}, \Theta)$ parameterisation as described in Section 7.5 is not as straightforward as the one for the CP, since $\pi(\Theta \mid \tilde{X}^{(w)}, Y)$ is not of known form, neither is it log-concave, therefore we simply use a Metropolis-Hastings step to update Θ, although more efficient and sophisticated methods could potentially be employed. Notice that for this very simple model we can integrate the random effects out and derive the likelihood function

$$\log \pi(Y \mid \Theta) \propto \sum_{i=1}^{m} \{(n_0+1)\log(\Theta) - (c + n_o + 1)\log(cY_i + n_0\Theta). \tag{7.17}$$

The results of Figure 7.5 suggest that the convergence rate of CP is not just a function of w. We wish to investigate the extent to which w affects the relative performance of the CA and the PNCA, although as suggested earlier w does not characterise the convergence rate uniquely. Some simulation results are presented in Figure 7.6. From this analysis appears that the CA always outperforms the PNCA.

Of course when c and n_0 have high values so that the model is more "Gaussian-like", the PNCP is successful, especially when combined with multiple Metropolis-Hastings steps to imitate a Gibbs algorithm, as shown in Figure 7.7. Nevertheless, the poor behaviour of PNCA is not solely explained by the use of the Metropolis-Hastings step. This is brought out in Figure 7.8, where it is seen that although the multiple Metropolis-Hastings steps are improving the algorithm, it still is much worse than the CA.

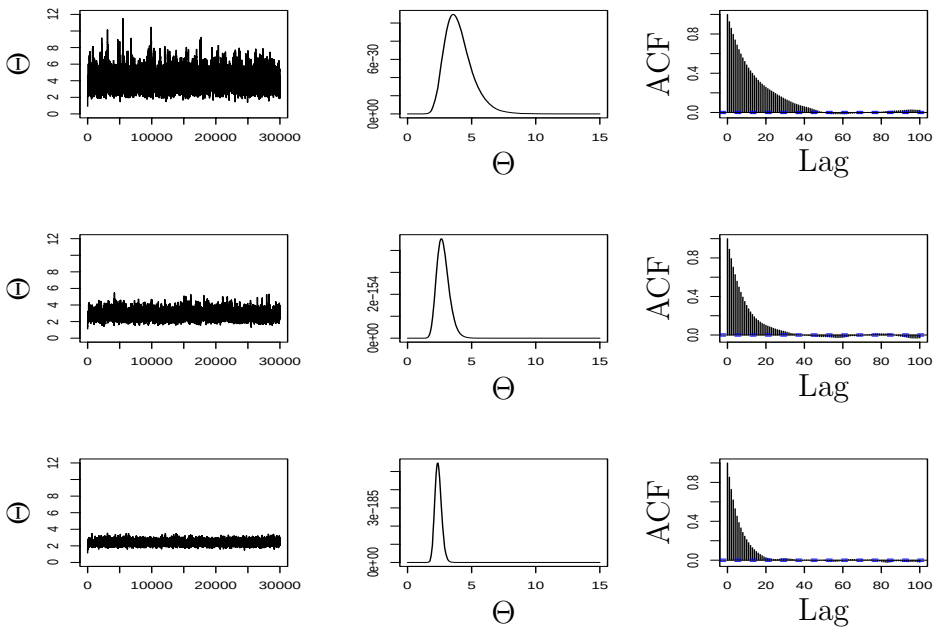


Figure 7.5: MCMC output for the CA for the GLHM (7.16). We have kept $w = 0.8$ fixed and take $n_0 = 1$ (top), $n_0 = 2$ (middle) and $n_0 = 5$ (bottom). Traces of Θ are drawn on the same scale. In the middle column we draw the likelihood function.

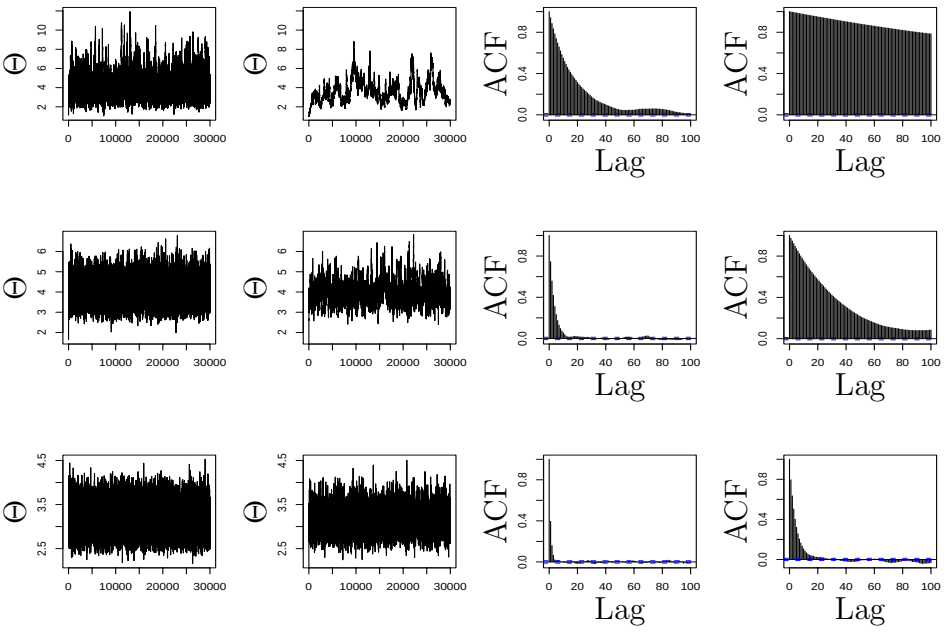


Figure 7.6: MCMC output for the CA (first and third columns) and the PNCA (second and fourth columns) for the GLHM (7.16). We have kept fixed $n_0 = 1$ and varied $w = 1/1.2$ (top), $w = 1/2$ (middle) and $w = 1/6$ (bottom). Notice that traces are drown on the same scale for both the CA and the PNCA.

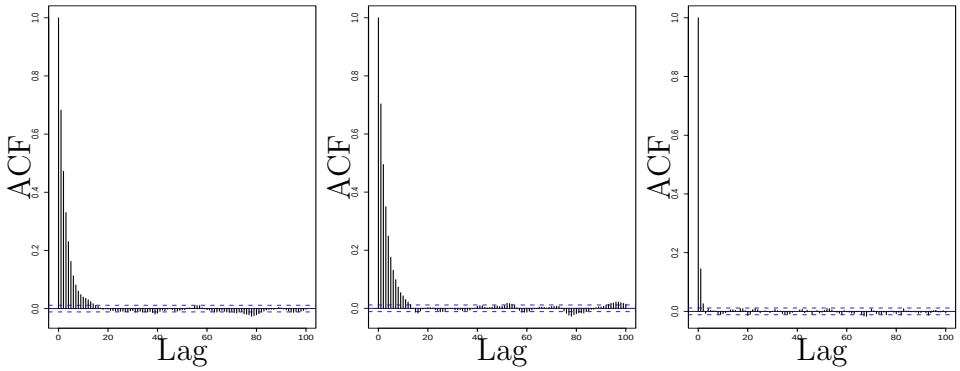


Figure 7.7: ACFs for CA (left) and PNCA (middle and right) when $n_0 = 20$ and $c = 10$, $(w = 0.67)$. The plot on the right corresponds to the PNCA algorithm that performs 30 Metropolis-Hastings steps for every update of the missing data.

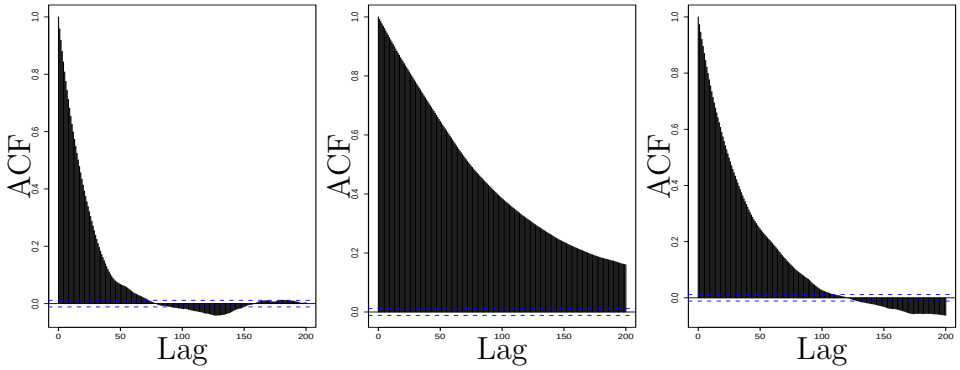


Figure 7.8: ACFs for the CA (left) and the PNCA (middle and right) when $n_0 = 3$ and $c = 0.2$ $(w = 0.9375)$. The plot on the right corresponds to the PNCA algorithm that performs 300 Metropolis-Hastings steps for every update of the missing data.

We have shown in (7.15) that Θ and $S^w = \sum_i \tilde{X}_i^{(w)}$ are uncorrelated, which is graphically illustrated verified in Figure 7.9. However, the maximal correlation in (2.7), which characterises the $\mathcal{L}^2$ rate of convergence, is clearly attained by functions of the missing data other than S^w.

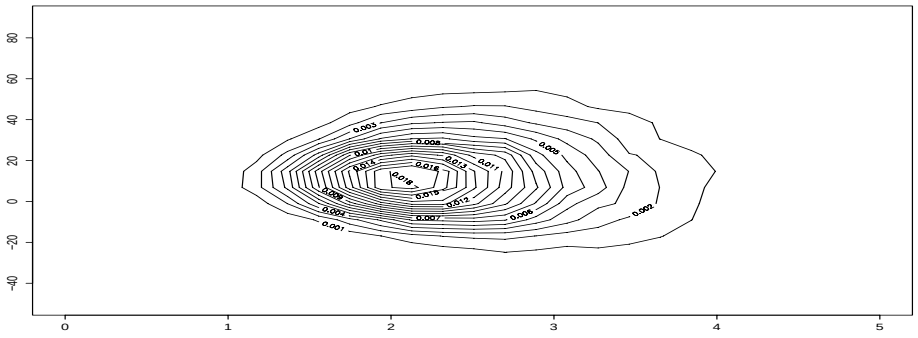


Figure 7.9: Contour plot of the estimated joint distribution of $\theta, S^w = \sum_i \tilde{X}_i^{(w)}$ when $n_0 = 3$ and $c = 0.2$ ($w = 0.9375$).

7.8.5 Higher order PNCP

As we have shown the parameterisation constructed in (7.13) based on posterior correlation analysis is not particularly successful. An improvement could be made if we tried to re-scale the $\tilde{X}_i^{(w)}$s so that their variance does not depend on Θ, i.e we set

$$\tilde{X}_i^{(w)} = \frac{1}{[\mathrm{Var}\,(X_i \mid \Theta, Y_i)]^{1/2}}(X_i - w_i\Theta) \tag{7.18}$$

$$= \left(\frac{n_0 + n_i/\phi_i - v2}{V(w_i\Theta + (1-w_i)Y_i)}\right)^{1/2}(X_i - w_i\Theta). \tag{7.19}$$

Clearly, $\tilde{X}_i^{(w)}$ and Θ are still uncorrelated but also the variance of $\tilde{X}_i^{(w)}$ does not depend on Θ.

We apply this parameterisation to the model (7.16) (assuming $n_0 + c > 1$)

$$\tilde{X}_i^{(w)} = \frac{\sqrt{n_0 + c - 1}}{w_i\Theta + (1-w_i)Y_i}(X_i - w_i\Theta) \tag{7.20}$$

$$w_i = w = \frac{n_0}{n_0 + c}. \tag{7.21}$$

We implement the Gibbs sampler based on this parameterisation and the simulation results plotted in Figure 7.10 suggest a big advantage over both the CA and the PNCA of the previous section.

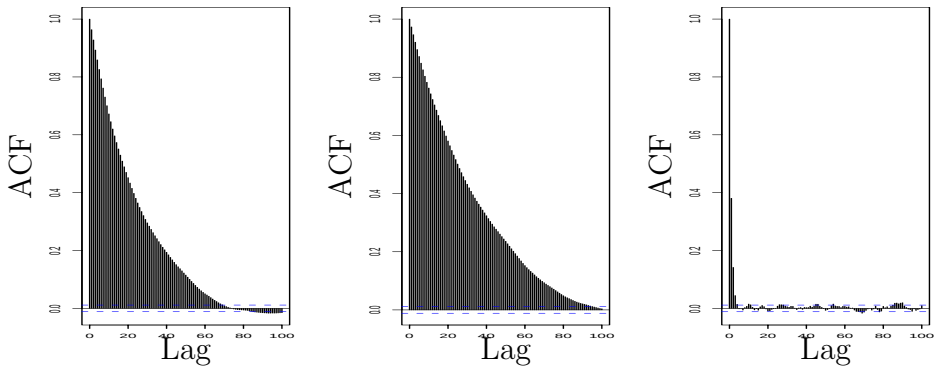


Figure 7.10: ACFs for the CA (left), the PNCA (middle) and the higher-order PNCA (right) when $n_0 = 3$ and $c = 0.2$ ($w = 0.9375$). 100 Metropolis-Hastings steps were performed for every update of the missing data in the two PNCAs.

7.9 Conditional and marginal augmentation

This section discusses two recently developed augmentation techniques. Our aim is to investigate their relevance to the PNCP. Both of these methods were initially used for the EM algorithm (Meng and van Dyk (1997), Meng and van Dyk (1999) and Liu et al. (1998)) however they have been successfully extended to Gibbs sampling as well (Meng and van Dyk (2001) and Liu and Wu (1999)). Before presenting the methods, we remind ourselves of the terminology introduced in Section 1.3 where missing, observed and augmented data correspond to X, Y, and (Y, X) respectively.

7.9.1 Conditional augmentation

Conditional augmentation is in many ways very similar to the PNCP. It is based on the introduction of a so-called (see p.5 of Meng and van Dyk (2001)) "working parameter", w say, which is not identifiable from the observed but only from the augmented data, therefore the marginal likelihood $\pi(Y \mid \Theta)$ is the same for all values of w. The introduction of w essentially implies a transformation of the missing data $X \to \tilde{X}^{(w)}$ and the Gibbs sampler corresponding to this parameterisation simulates iteratively from the conditional distribution of $\Theta \mid \tilde{X}^{(w)}, Y$ and $\tilde{X}^{(w)} \mid \Theta, Y$. The motivation behind this augmentation scheme is to find a w such that the maximal correlation between Θ and $\tilde{X}^{(w)}$ is smaller than between Θ and X. If this is achieved, then Section 2.1 shows that the Gibbs sampler based on the $(\tilde{X}^{(w)}, \Theta)$ parameterisation will converge faster than the one which updates X and Θ. Typically, as for the PNCP, an improvement in the convergence rate is not guaranteed for all possible values of w.

An example often used in the literature (see for example Section 5 of Meng and van Dyk (2001) and Section 2.2 of Meng and van Dyk (1997)) to illustrate the method is that of a t-model. Y given X is assumed to be normally distributed with variance $1/X$, and $X \sim$

$\text{Ga}(\nu/2, \Theta\nu/2)$ where ν is assumed to be known. Notice that we have expressed the model using a CP. Meng and van Dyk (1997) propose the reparameterisation $X \to \tilde{X}^{(w)}$, where $X = \Theta^w \tilde{X}^{(w)}$ for some w and for $w = 0$ we recover the original (centered) parameterisation. Thus, $Y \mid \tilde{X}^{(w)}, \Theta \sim \text{N}(0, \Theta^w/\tilde{X}^{(w)})$ and $\tilde{X}^{(w)} \mid \Theta \sim \text{Ga}(\nu/2, \Theta^{1-w}\nu/2)$.

As the example demonstrates, the conditional augmentation has many similarities to the PNCP, and in many cases the two augmentation schemes coincide. Specifically, in the example above the choice $w = 1$ results in a non-centered algorithm (see for example Section 4.1) and the conditional augmentation designed is exactly a PNCP. The connection between the non-centered methodology and the conditional augmentation has been established by Papaspiliopoulos et al. (2003). In our approach, $\tilde{X}^{(w)}$ is chosen so that it creates a continuum of parameterisations between the CP and the NCP, which was argued to be desirable in Section 7.1 and Section 7.2. On the contrary, this choice is more arbitrary in many of the examples where conditional augmentation has been used effectively. Moreover, conditional augmentation has solely been concerned with scale and location transformations. Our results suggest that it might be interesting and feasible to consider much more complicated transformations, as those suggested in Section 7.6. Such extensions will allow the conditional augmentation to be used for complex models with hidden stochastic processes, such as those considered in Chapter 5 and Chapter 6 for example, whereas up to now it has largely been applied to relatively simple random-effects type models. Therefore, we believe that our work on PNCP is complimentary to the conditional augmentation methodology.

An important issue in both the conditional augmentation and the PNCP is the choice of w. This will be tackled in Section 7.10. The next section describes an alternative to choosing a specific value for w: integrating out of the problem.

7.9.2 Marginal augmentation

Instead of conditioning on a specific value of the so-called working parameter in a conditional augmentation, it has been suggested that a prior distribution is assigned to it and then it is marginalised. This scheme is known as marginal augmentation, sometimes called parameter expanded data augmentation, and it has been proposed independently by Liu and Wu (1999) and Meng and van Dyk (1999), although this idea arose initially in the context of the EM algorithm, see for example Liu et al. (1998). Since the working parameter is first introduced in the augmentation and then it is integrated out, it might seem that nothing has been achieved, nevertheless this is not the case.

We first show how the method works for the simple normal hierarchical model. This example is used for pedagogical purposes by Liu and Wu (1999) and this section reproduces their construction.

The NCP for this model was constructed in Section 2.3 and is given in (2.18). To retain

consistency with Liu and Wu (1999) we assume that $\sigma_y^2 = 1$ and denote $\sigma_x^2 =: D$. Thus κ in (2.16) equals $D/(1+D)$ and the model writes as

$$\begin{aligned} Y_i &= X_i + \epsilon_i \\ X_i &= \Theta + D^{1/2} z_i,\ i = 1, \ldots, m. \end{aligned} \tag{7.22}$$

A consideration which we find motivating for the marginal augmentation is that the constraint imposed by the data that $\Theta + \tilde{X}_i = Y_i +$ "error", results in slow convergence of the Gibbs sampler, when D is large compared to 1, as it was shown in Section 2.3. Thus, it would be desirable if there was some added remaining uncertainty in this constraint, which could be achieved if for example $\tilde{X}_i$ was the sum of two random variables and we were conditioning on just one of those when updating Θ. Liu and Wu (1999) suggest using the over-parameterised model

$$\begin{aligned} Y_i &= \Theta + Z_i - \alpha + \epsilon_i \\ Z_i &\sim \mathrm{N}(\alpha, D),\ i = 1, \ldots, m \end{aligned} \tag{7.23}$$

where we have transformed

$$\tilde{X}_i \to Z_i, \text{ where } \tilde{X}_i =: Z_i - \alpha. \tag{7.24}$$

Moreover, suppose that we specify the prior $\alpha \sim N(0, B)$ and define $Z := (Z_1, \ldots, Z_m)$. Since Z is a linear transformation of $\tilde{X}$ we can easily find that (Z, Θ) conditionally on α and Y are normally distributed. Due to the conjugate prior of α, the same holds for the posterior distribution of (Z, Θ) when α is integrated out. This integration is done simply, since α is not identifiable and its conditional distribution given Y is the same as its prior. The marginal augmentation proceeds by performing Gibbs sampling on the joint posterior of (Z, Θ), that is α is integrated out rather than kept fixed throughout the simulation.

Abstracting from the specific example, the starting point for the marginal augmentation is to express the hierarchical model as a missing data problem, usually using either a centered or a non-centered parameterisation. For the sake of simplicity and clarity, this section violates the notation established in this thesis and uses $\tilde{X}$ as a generic notation for the missing data, even if a centered parameterisation is employed. Thus, the augmented data are $(Y, \tilde{X})$, which we refer to as the ordinary data augmentation. Marginal augmentation proceeds by finding a non-identifiable (by the observed data) parameter α, some function t_α and constructs a transformation of the original missing data $\tilde{X} \to Z$ where $\tilde{X} = t_\alpha(Z)$. For instance, in the normal example introduced earlier $t_\alpha(Z) = Z - \alpha$, see (7.24). A (possibly improper) prior is chosen for α, which however guarantees that the joint posterior of (Z, Θ) is proper. The

two main assumptions of the marginal augmentation methodology, which will be shortly motivated, are the following:

Condition (a): t_α is a one to one differentiable mapping.

Condition (b): α and Θ are *a priori* independent.

We can then easily find (by a change of variables argument) the joint posterior distribution of (Z, Θ, α)

$$\pi(Z, \Theta, \alpha \mid Y) \propto \pi(Y \mid Z, \Theta, \alpha)\pi(Z \mid \Theta, \alpha)\pi(\alpha)\pi(\Theta) \tag{7.25}$$

from which α is integrated out and Gibbs sampling is used to simulate from the joint posterior of (Z, Θ).

This is the main idea behind the marginal augmentation, we try to "collapse" some part of the missing data $\tilde{X}$, for example by taking $\tilde{X} = Z - \alpha$ and integrating out α, while still being able to do Gibbs sampling. Clearly these two tasks are competing, since the more we collapse the more difficult Gibbs sampling becomes. For example, if we collapse all of $\tilde{X}$ by integrating it out of the problem, then most likely we will not be able to sample from the posterior distribution of Θ. Actually, $\tilde{X}$ was originally introduced to simplify this simulation. There are many issues remaining to be resolved, such as the choice of the prior on α, the transformation t_α to be used, the implementation of the Gibbs sampling algorithm, the relevance of this methodology with the PNCP and a proof that the marginal is superior to the ordinary augmentation under *conditions (a)* and *(b)* above. We start by addressing the latter.

We will present a simple and intuitive argument which has not appeared in the literature before and it shows that under the conditions stated above the marginal is guaranteed to have better convergence rate than the ordinary augmentation. Our argument motivates these conditions naturally, however it assumes that a proper prior is chosen for α. Thus it cannot be used as a general proof, since it is not valid when an improper prior is assigned to α, a choice which turns out to be optimal in many cases. The proof of the general theorem, which is based on the characterisation of the convergence rate of the two-component Gibbs sampler as the maximal correlation between the updated components (see Section 2.1 and (2.7)), can be found in Liu and Wu (1999).

Notice that the ordinary augmentation corresponds to the Gibbs sampler which simulates iteratively from the conditional distributions

$$\begin{aligned} &1. \quad (Z, \alpha) \mid \Theta, Y \\ &2. \quad \Theta \mid Y, (Z, \alpha). \end{aligned} \tag{7.26}$$

Since $\tilde{X}$ is a one-to-one transformation of α and Z its value is uniquely determined by the

value of the pair (Z, α). Moreover, by *condition (b)* Θ is *a priori* independent of α, thus the distribution of $\Theta \mid Y, \tilde{X}, \alpha$ coincides with that of $\Theta \mid Y, \tilde{X}$. Therefore, the second step of (7.26) simulates Θ given Y and $\tilde{X}$, as in the ordinary data augmentation. Since α is non-identifiable and *a priori* independent of Θ, its distribution conditional on Θ and Y is the same as the prior. Since the latter is assumed to be proper, we can easily simulate a draw from it and conditionally on that value we can simulate from Z given Θ, Y, α. This can be achieved for example by drawing first $\tilde{X}$ given Θ, Y as in the ordinary augmentation, and then setting $Z = t_\alpha^{-1}(\tilde{X})$. The existence of t_α^{-1} is ensured by the *condition (a)* above. Once Z and α have been drawn, we set $\tilde{X} = t_\alpha(Z)$. Thus, the first step of (7.26) is the same as the first step of the ordinary data augmentation.

The algorithm described in (7.26) is known as blocked (Roberts and Sahu (1997)) or *grouped* (Liu et al. (1994)) Gibbs sampler, since the random variables Z and α are updated in one block. On the other hand, the marginal augmentation simulates iteratively from

$$\begin{aligned} &1. \quad Z \mid \Theta \\ &2. \quad \Theta \mid Z \end{aligned} \tag{7.27}$$

which is with respect to (7.26) a *collapsed* (Liu (1994a)) Gibbs sampler, since α has been integrated out. Hence, when a proper prior is used for α, and *conditions (a)* and *(b)* hold, our argument shows that the marginal has better convergence rate than the ordinary augmentation as long as the collapsed has better convergence rate than the grouped Gibbs sampler. However, Theorem 5.1 of Liu et al. (1994) proves that, for any Z, α, Θ, the spectral radius of the *grouped* Gibbs sampler (7.26) is greater or equal than the spectral radius of the *collapsed* Gibbs sampler (7.27). Therefore, they show what it is also intuitive, that scheme (7.27) is converging faster than scheme (7.26). Therefore, it is straightforward to prove that the marginal augmentation leads to a faster converging Gibbs sampler when $\pi(\alpha)$ is proper. The general proof can be found in Liu and Wu (1999). Notice that the result presupposes that exact simulations from the conditionals are feasible and no Metropolis-Hastings within Gibbs steps are used.

It turns out however that an improper prior is often optimal. We return to the example we introduced in the beginning of this section. Since (Z, Θ) is jointly Gaussian, it is easy to calculate the rate of convergence of the Gibbs sampler with this target, using the general results summarised in Section 2.1. This is done in p.1265 of Liu and Wu (1999) where it is found that

$$\rho_{pxda} = \frac{D - (B^{-1} + D^{-1})^{-1}}{1 + D}; \tag{7.28}$$

this is plotted against the prior precision B^{-1} for different values of D in Figure 7.11. Notice that for all B $\rho_{pxda} \leq \rho_{nc} = 1/(1 + D^{-1})$ and that as $B^{-1} \to 0$ the algorithm improves its rate, actually only for this limiting case does it give considerable improvement. Since by

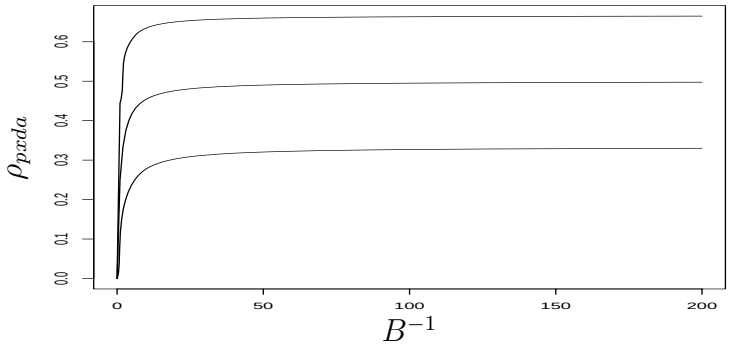


Figure 7.11: ρ_{pxda} for $D = 2, 1, 0.5$. The asymptote in each of these curves is ρ_{nc}.

convention we associate the limiting measure of a $\mathrm{N}(0, B)$ as $B \to \infty$ with the improper uniform, the results imply that this is the optimal prior for α in this problem. Liu and Wu (1999) argue intuitively that it might be better to "let the imputed data decide α at each iteration" (see p.1266 of their paper) and therefore a non-informative improper prior should generally be preferred. There are many results about the choice of such priors and the implementation of the corresponding marginal augmentation. Nevertheless, these considerations are outside the scope of this chapter and we simply refer to Liu and Wu (1999) for details and insightful discussion. We finish this review of the marginal augmentation by discussing its relevance to the PNCP.

Notice that the PNCP can be obtained as a special case of the marginal augmentation, where the prior distribution of α depends on Θ. For example, the PNCP of Section 7.2 used the transformation $\tilde{X} \to Z := \tilde{X} + (1-w)\Theta$. This can be put into the marginal augmentation framework discussed above, where $t_\alpha(Z) = Z - \alpha$ and the prior distribution of α is a point mass centered at $(1 - w)\Theta$. This prior violates *condition (b)*, which is necessary to show that the marginal is superior to ordinary augmentation. Actually, we have already seen in Section 7.2 that this prior can lead to algorithms which converge slower than the ordinary data augmentation. Generally, we find it inappropriate to construct parameterisations which are meaningful inside the conditional augmentation context and then assigning a prior to the working parameter with view to integrating it out. On the one hand, they are not guaranteed to lead to faster convergence since they typically violate *condition (b)*. On the other hand, our experience even with the simplest of models suggests that they are very difficult to implement.

7.10 Optimising the PNCP

The conditional augmentation literature has contributed with some suggestions about the choice of optimal values for the working parameter w. There are three most commonly used

methods, which are reviewed in Section 2 of Meng and van Dyk (2001). The first is to find the value of w which minimises the geometric rate of convergence of the Gibbs sampler, or equivalently (see Section 2.1) which minimises the maximal correlation between $\tilde{X}^{(w)}$ and Θ

$$\gamma_w(\tilde{X}^{(w)}, \Theta) := \sup_{h:\mathrm{Var}\,[h(\Theta)|Y,w]=1} \mathrm{Var}^{1/2}[\mathsf{E}[h(\Theta) \mid \tilde{X}^{(w)}, Y, w] \mid Y, w] \tag{7.29}$$

This is a very sensible criterion since it selects the w which leads to the fastest converging Gibbs sampler in an $\mathcal{L}^2$ sense, nevertheless it is hardly ever possible to perform this optimisation for non-Gaussian models. Theorem 2.1.1 states that the maximal correlation between the updated components equals the maximal lag-1 autocorrelation of the marginal chains. Thus the above method is equivalent to maximising the maximal lag-1 autocorrelation in the Θ marginal chain. The second method proposed by Meng and van Dyk (2001) as more feasible is to minimise the maximal lag-1 autocorrelation over linear combinations in the Θ marginal chain. When Θ is a scalar, this method is equivalent to optimising (7.29) over linear functions h only, which however can lead to undesirable results, as Section 7.8 showed. The third method is the most practically appealing. It consists of finding the w which minimises the convergence rate of the deterministic counterpart of the Gibbs sampler, i.e the EM algorithm. We mentioned in Section 7.9.1 that conditional augmentation was originally developed for the EM algorithm and this method, known as the EM criterion, has proved successful in that context. Given the augmentation scheme $(\tilde{X}^{(w)}, \Theta)$ the EM algorithm (Dempster et al. (1977), Meng and van Dyk (1997)) for locating the posterior mode of Θ, Θ^* say, has a theoretical convergence rate which when Θ is a scalar is given by (Dempster et al. (1977))

$$r_{EM}(w) = 1 - I_{obs} I_{aug}^{-1}(w) \tag{7.30}$$

where

$$I_{aug}(w) = \mathsf{E}\left[-\frac{\partial^2 \log \pi(\Theta \mid Y, \tilde{X}^{(w)}, w)}{\partial \Theta^2} \mid Y, \Theta, w \right] \Bigg|_{\Theta=\Theta^*} \tag{7.31}$$

is the expected augmented Fisher information and

$$I_{obs} = -\frac{\partial^2 \log \pi(\Theta \mid Y)}{\partial \Theta^2}\Big|_{\Theta=\Theta^*}$$

is the observed Fisher information for Θ. The EM criterion selects the w which minimises (7.30), or equivalently minimises (7.31). There is an added technical complexity when Θ is a vector. Although the above definitions are extended in a natural way, (7.30) defines a matrix whose the spectral radius has to be minimised. Moreover, minimising (7.31) is

not necessarily equivalent to minimising (7.30) in the vector case, see Meng and van Dyk (2001) for more details and references. An advantage of the EM criterion is that in many applications it is not necessary to compute Θ^* in order to calculate (7.31), see Sections 6-8 of Meng and van Dyk (2001).

All methods are equivalent for Gaussian models, but the last two are essentially based on a Gaussian approximation of any given model. It is not clear whether criteria and constructions which are suitable for mode-hunting algorithms such as the EM, are suitable for algorithms which explore distributions, particularly in the presence of heavy tails. Relevant discussion can be found in Meng and van Dyk (2001).

Section 7.5 argued that it might be desirable to let w depend upon the observed data. The following example illustrates how this might be achieved. Consider the random effects model:

$$\begin{aligned} Y_i &\sim \pi(\cdot|X_i) \\ X_i &= \Theta + \sigma_x z_i,\ i = 1, \ldots, m\ . \end{aligned} \tag{7.32}$$

for some class of densities or probabilities $\pi(\cdot|\cdot)$ and $z_i \sim \mathrm{N}(0,1)$. A quadratic expansion of the log-likelihood, $\ell = \log \pi$ gives a rough indication into the information content present in Y_i about X_i. We set $I(Y_i) = -\partial^2 \ell(Y_i|X_i)/\partial X_i^2$ evaluated at the MLE $\hat{X}_i$ (ignoring the latent structure). Other approximations of information may be more appropriate in certain cases. In the normal hierarchical model, $I(Y_i) = {\sigma_y}^{-2}$, but more generally $I(Y_i)$ will depend on Y_i. Data-dependent non-centering then sets

$$\tilde{X}_i^{(w)} = X_i - w_i(Y)\theta \tag{7.33}$$

where $w_i(Y) = (1 + I(Y_i)/\sigma_x^2)^{-1}$. This is a crude and easy method for selecting w, but as the geostatistical example of Section 7.11 shows, it can be rather effective. Another important example of a PNCP with data dependent w can be found in the spatial application of Higdon (1998).

7.11 Examples

This chapter has focused on developing methodology regarding the PNCP and establishing the connections with alternative augmentation schemes suggested in the literature. There have already been attempts to apply this methodology to complex statistical models, among which we describe an application in the modelling of spatial variation in some detail.

7.11.1 Spatial GLMM

The results of this section originally appeared in Papaspiliopoulos et al. (2003) but the problem is thoroughly investigated in Christensen et al. (2003).

We consider a special case of the generalised linear mixed model (GLMM) introduced in Breslow and Clayton (1993) and proposed in the spatial context by Diggle et al. (1998). Similar modeling approaches have received much attention recently but strong posterior correlation between parameters and latent variables makes a fully Bayesian approach difficult without the use of complex MCMC algorithms and careful reparameterisations. We will here consider a spatial Poisson log-Normal model also studied in Diggle et al. (1998) for modeling radioactive counts and in Christensen and Waagepetersen (2003) for modeling counts of weed. For a more detailed description of the model refer to Diggle et al. (1998).

The data consists of recorded observations $Y = (Y_1, \ldots, Y_m)$ with

$$\begin{aligned} Y_i &\sim \mathrm{Pn}(\exp(X_i)) \\ X &\sim \mathrm{N}(\Theta\mathbf{1}, \sigma^2 R). \end{aligned} \tag{7.34}$$

Here $X = (X_1, \ldots, X_m) = (X(t_1), \ldots, X(t_m))$ are (unobserved) values from a stationary isotropic Gaussian random field $X = \{X(t), t \in \mathbb{R}^2\}$ with mean Θ, standard deviation σ and correlation function $r(u) = \mathrm{Corr}(X(s_1), X(s_2)) = \exp(-u/\alpha),\ u = \|s_2 - s_1\|$ (Euclidean distance), and $\mathbf{1}$ is a $m \times 1$ vector of 1s. In the limiting case where $\alpha \to \infty$ this reduces to the random effects case described by (7.32), with $Y_i \sim \mathrm{Pn}(\exp(X_i))$.

Thus the unknown components in this model are the parameters Θ, α and σ, together with the underlying field X. It is beyond the scope of this section to fully describe partially non-centered methods which can be applied effectively to this problem; considerably more detail can be found in Christensen et al. (2003). We shall instead concentrate on part of the algorithm, in which the partial non-centering is applied to Θ while the remaining parameters σ and α remain fixed. Similar non-centering strategies can be applied to σ and α and also directly to X in order to break down the posterior correlation structure present in the field X. This is reported in Christensen et al. (2003).

Both Diggle et al. (1998) and Christensen and Waagepetersen (2003) use the NCP for Θ, i.e. they alternate between updating $\tilde{X} = X - \mathbf{1}\Theta$ and Θ. Diggle et al. (1998) use a single site Gibbs updating of $\tilde{X}$ and Christensen and Waagepetersen (2003) use the Metropolis adjusted Langevin algorithm (MALA, see for example Roberts and Tweedie (1996a)) which can give considerable convergence advantages for large m. Here we shall extend the data-dependent partial non-centering ideas of Section 7.3, Section 7.5 and Section 7.10 to the present case of spatially varying X.

In the absence of covariates, the partial non-centering parameters for the general normal

hierarchical model with $m = 1$ are given in Section 7.3 as

$$W = BD^{-1} = (\text{diag}(\sigma_x^{-2}) + D^{-1})^{-1}D^{-1}$$

where D is the $n \times n$ prior variance matrix and $\text{diag}(\sigma_x^{-2})$ is the $n \times n$ matrix with σ_x^{-2} on the diagonal and zeros elsewhere. The latter matrix being constant along the diagonal reflects the fact that the variance of the error distribution $Y|X$ is independent of X and thus the loss of information about Θ is equal along the components of Y. This is a feature not shared by model (7.34) since here the error-distribution depends on X in a nonlinear way through the logarithmic link-function and large values in Y tend to be more informative about the mean than small ones.

Outside the Gaussian context, there is no direct analogue of the B_i matrices in (7.7), but a quadratic expansion of the likelihood as in Section 7.10 suggests we set

$$\hat{B}^{-1} =: \frac{d^2}{dX^2} \log \pi(X|Y, \theta)|_{X=\hat{X}} = -(\text{diag}(\exp(\hat{X})) + \sigma^{-2}R^{-1}) = -(\text{diag}(Y) + \sigma^{-2}R^{-1}) \tag{7.35}$$

(where $\hat{X}$ is the MLE from the observation equation alone) leading to the partial non-centering for Θ

$$\tilde{X}^w = X - W\Theta. \tag{7.36}$$

with $W = \hat{B}\sigma^{-2}R^{-1}$.

A simulation study involving 100 observations equally spaced on the unit square was carried out. It involved using two different combinations of Θ and σ each under two levels of dependence. Although updates of $\tilde{X}^w$ need to be carried out by a suitable Hastings algorithm in this context for each of our parameterisations, our comparison is based on pure Gibbs updates (approximated by running multiple MALA steps for $\tilde{X}^w$ and Θ). This eliminates the possibility that any difference between simulation performance could be due to varying efficiency of the MALA updates. ACFs summarising our results are given in Figure 7.12.

The CP performs better in relation to the NCP as the dependence in X becomes stronger (as measured by α), which is in agreement with our results for the state-space model in Section 2.5 and for the OU volatility model in Section 6.8 and Section 6.12. Moreover, while the performance of the CP and NCP vary considerably as the parameters change, the data-dependent PNCP performs extremely well in all cases.

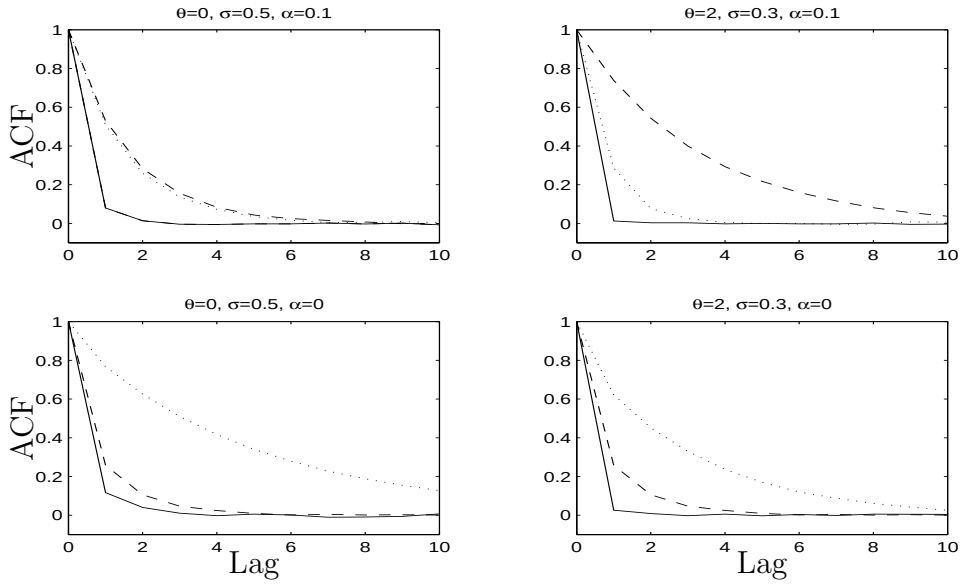


Figure 7.12: The spatial GLMM of Section 7.11.1 and its special case, the random effects model of Section 7.10 (showed in the bottom two plots). ACF for Θ using CA (dotted), NCA (dashed) and data-dependent PNCA (solid) for various parameter values.

Bibliography

Amit, Y. (1991) On rates of convergence of stochastic relaxation for Gaussian and non-Gaussian distributions. *J. Multivariate Analysis*, **38**, 82–99.

Amit, Y. and Grenander, U. (1991) Comparing sweep strategies for stochastic relaxation. *Journal of Multivariate Analysis*, **37**, 197–222.

Andersen, T., Bollerslev, T., Diebold, F. and Labys, P. (2001) The distribution of realised exchange rate volatility. *JASA*, **96**, 42–55.

Barndorff-Nielsen, O., Nicolato, E. and Shephard, N. (2002) Some recent developements in stochastic volatility modelling. *Quantitive Finance*, **2**, 11–23.

Barndorff-Nielsen, O. and Shephard, N. (2001) Non-Gaussian Ornstein-Ulhenbeck-based models and some of their uses in financial economics. *J. Roy. Stat. Soc., B*, **63**, 167–241.

— (2002a) Econometric analysis of realised volatility and its use in estimating stochastic volatility models. *J. Roy. Stat. Soc., B*, **64**, 253–280.

— (2002b) Modelling by Lévy processess for financial econometrics. In *Lévy Processes – Theory and Applications* (eds. O. E. Barndorff-Nielsen, T. Mikosch and S. Resnick). Birkhauser, Boston.

— (2003) Integrated OU processes and non-Gaussian OU-based stochastic volatility models. *Scandinavian Journal of Statistics*, **30**, 277–295.

— (2004) *Financial Volatility: stochastic volatility and Lévy based models.* Cambridge University Press.

Bernardo, J. M. and Smith, A. F. M. (1994) *Bayesian Theory.* Wiley.

Black, F. (1976) Studies of stock price volatility changes. *Proc. Bus. Econ. Statist. Sec. Am. Statist. Ass.*, 177–181.

Bondesson, L. (1982) On simulation from infinitely divisible distributions. *Advances in Applied Probability*, **14**, 855–869.

Breiman, L. and Friedman, J. H. (1985) Estimating optimal transformations for multiple regression and correlation (c/r: P598-619). *Journal of the American Statistical Association*, **80**, 580–598.

Breslow, N. E. and Clayton, D. G. (1993) Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25.

Brockwell, P. J. and Davis, R. A. (1991) *Time Series. Theory and Methods (Second Edition)*. Springer-Verlag.

Brooks, S. and Roberts, G. (1999) Assessing convergence of Markov chain Monte Carlo algorithms. *Statistics and Computing*, **8**, 319–335.

Campbell, J., Lo, A. and MacKinlay, A. (1997) *The econometrics of financial markets*. Princeton University Press.

Cappé, O., Robert, C. and Rydén, T. (2003) Reversible jump mcmc converging to birth-and-death mcmc and more general continuous time samplers. *Journal of the Royal Statistical Society, Series B, Methodological*, **to appear**.

Carter, D. S. and Prenter, P. M. (1972) Exponential spaces and counting processes. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **21**, 1–19.

Christensen, O., Roberts, G. and Sköld, M. (2003) Bayesian analysis of spatial GLMM using partially non-centered MCMC methods. *in preparation*.

Christensen, O. and Waagepetersen, R. (2003) Bayesian prediction of spatial count data using generalised linear mixed models. *Biometrics*, **58**, 280–286.

Cowles, M. K., and B.P., C. (1996) Markov chain monte carlo convergence diagnostics. *JASA*, **91**, 883–904.

Cox, D. and Isham, V. (1980) *Point Processes*. Chapman and Hall.

Damien, P., Laud, W. and Smith, A. (1995) Approximate Random Variate Generation from Infinitely Divisible Distributions woth Applications to Bayesian Inferene. *Journal of the Royal Statistical Society, series B*, **57**, 547–563.

Dawid, A. P. (1973) Posterior expectations for large observations. *Biometrika*, **60**, 664–667.

— (1979) Conditional independence in statistical theory (c/r: P15-31). *Journal of the Royal Statistical Society, Series B, Methodological*, **41**, 1–15.

Dellaportas, P., Forster, J. and Ntzoufras, I. (2002) On Bayesian model and variable selection using MCMC. *Statistics and Computing*, **12**, 27–36.

Dellaportas, P. and Smith, A. F. M. (1993) Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling. *Applied Statistics*, **42**, 443–459.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (c/r: P22-37). *Journal of the Royal Statistical Society, Series B, Methodological*, **39**, 1–22.

Devroye, L. (1986) *Non-uniform Random Variate Generation*. Springer-Verlag.

Diebolt, J. and Robert, C. (1994) Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society, Series B*, **56**, 363–375.

Diggle, P., Tawn, J. and Moyeed, R. (1998) Model-based geostatistics (*with discussion*). *Journal of the Royal Statistical Society, Series C*, **47**, 2991–350.

Diggle, P. J., Liang, K.-Y. and Zeger, S. L. (1994) *Analysis of Longitudinal Data (ISBN 0198522843)*. Clarendon Press.

Duffie, D. (1992) *Dynamic asset pricing theory*. Princeton University Press.

Embrechts, P., Kluppelberg, C. and Mikosch, T. (1997) *Modelling extremal events for insurance and finance*. Springer.

Feller, W. (1971) *An introduction to probability theory and its applications, Vol. II*. Wiley & Sons, New York.

Ferguson, T. S. (1973) A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, **1**, 209–230.

— (1974) Prior distributions on spaces of probability measures. *The Annals of Statistics*, **2**, 615–629.

Ferguson, T. S. and Klass, M. J. (1972) A representation of independent increment processes without Gaussian components. *The Annals of Mathematical Statistics*, **43**, 1634–1643.

Ferguson, T. S. and Phadia, E. G. (1979) Bayesian nonparametric estimation based on censored data. *The Annals of Statistics*, **7**, 163–186.

Gammerman, D. (1997) *Markov chain Monte Carlo: Stochastic simulation for Bayesian inference*. Chapman and Hall.

Gelfand, A., Sahu, S. and Carlin, B. (1995) Efficient parametrization for normal linear mixed models. *Biometrika*, **82**, 479–488.

— (1996) Efficient parameterizations for generalised linear models. In *Bayesian Statistics V*, 479–488. Oxford. Ed. J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith.

Gelfand, A. and Smith, A. (1990) Sampling-based approaches to sampling marginal densities. *JASA*, **85**, 398–409.

Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machines intelligence*, **6**, 721–741.

Geyer, C. and Møller, J. (1994) Simulation procedures and likelihood inference for spatial point processes. *Scand. J. Statist.*, **21**, 359–373.

Geyer, C. J. (1992) Practical Markov chain Monte Carlo (disc: P483-503). *Statistical Science*, **7**, 473–483.

Ghysels, E., Harvey, A. C. and Renault, E. (1996) Stochastic volatility. In *Statistical Method in Finance* (eds. C. Rao and G. Maddala). Amsterdam: North-Holland.

Gilks, W., Richardson, S. and Spiegelhalter, D. (eds.) (1996) *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London.

Gilks, W. R., Thomas, A. and Spiegelhalter, D. J. (1994) A language and program for complex bayesian modelling. *The Statistician*, **43**, 169–178.

Goodman, J. and Sokal, A. (1989) Multigrid monte carlo method: conceptual foundations. *Phys. Rev. D*, **40**, 2035–2071.

Green, P. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Griffin, J. and Steel, M. (2002) Inference with non-Gaussian Ornstein-Uhlenbeck processes for stochastic volatility. *mimeo Institute of Mathematics and Statistics, University of Kent at Cantenrbury*.

Harvey, A., Ruiz, E. and N., S. (1994) Multivariate stochastic variance models. *Review of Economic studies*, **61**, 247–264.

Hastings, W. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Higdon, D. (1998) Auxiliary variable methods for Markov Chain Monte Carlo with applications. *Journal of the American Statistical Assocoation*, **93**, 585–596.

Hills, S. and Smith, A. (1992) Parameterization issues in bayesian inference. In *Bayesian Statistics 4* (eds. J. Bernardo, J. Berger, A. Dawid and A. Smith), 227–246. Oxford.

Jacquier, E., Polson, N. G. and Rossi, P. E. (1994) Bayesian analysis of stochastic volatility models (disc: P389-417). *Journal of Business and Economic Statistics*, **12**, 371–389.

Karr, A. F. (1991) *Point Processes and Their Statistical Inference (Second Edition)*. Marcel Dekker.

Kim, S., Shephard, N. and Chib, S. (1998) Stochastic volatility: Likelihood inference and comparison with arch models. *Review of Economic studies*, **65**, 361–393.

Kingman, J. (1993) *Poisson Processes*. Oxford.

Kirkpatrick, S., Gelatt, C. and Vecchi, M. (1983) Optimization by simulated annealing. *Science*, **220**, 671–680.

Lancaster, H. O. (1958) The structure of bivariate distributions (corr: V35 p1388). *The Annals of Mathematical Statistics*, **29**, 719–736.

Lawler, G. and Sokal, A. (1988) Bounds on the $l2$ spectrum for markov chains and markov processes. *Transations of the AMS*, **309**, 557–580.

Lee, Y. and Nelder, J. (1996) Hierarchical generalized linear model (with discussion). *Journal of the Royal Statistical Society, Series B, Methodological*, **58**, 619–656.

Lindley, D. V. and Smith, A. F. M. (1972) Bayes estimates for the linear model (with discussion). *Journal of the Royal Statistical Society, Series B, Methodological*, **34**, 1–41.

Liu, J. (2001) *Monte Carlo strategies in scientific computing*. Springer-Verlag.

Liu, J., Rubin, D. B. and Wu, Y. (1998) Parameter expansion to accelerate EM - the PX-EM algorithm. *Biometrika*, **85**, 755–770.

Liu, J. S. (1994a) The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, **89**, 958–966.

— (1994b) Fraction of missing information and convergence rate of Data Augmentation. In *Proc. 26th Symp. Interface of Comp. Sci. and Statist.*, 490–496.

Liu, J. S., Wong, W. H. and Kong, A. (1994) Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, **81**, 27–40.

— (1995) Covariance structure and convergence rate of the Gibbs sampler with various scans. *Journal of the Royal Statistical Society, series B*, **57**, 157–169.

Liu, J. S. and Wu, Y. N. (1999) Parameter expansion for Data Augmentation. *Journal of the American Statistical Assocoation*, **94**, 1264–1274.

Meng, X.-L. and van Dyk, D. (1997) The EM algorithm – an old folk song sung to a fast new tune. *J. Roy. Stat. Soc., B*, **59**, 511–567.

— (1999) Seeking efficient data augmentation schemes via conditional and marginal augmentation. *Biometrika*, **86**, 301–320.

— (2001) The art of Data Augmentation. *Journal of Computational and Graphical Statistics*, **10**, 1–50.

Meng, X.-L. and Rubin, D. B. (1993) Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, **80**, 267–278.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. and Teller, E. (1953) Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–1091.

Meyn, S. P. and Tweedie, R. L. (1993) *Markov Chains and Stochastic Stability*. London: Springer-Verlag.

Mikosch, T. (2002) *Modeling dependence and tails of financial time series*. SEMSTAT Lecture notes, available at `http://www.math.ku.dk/~mikosch/Semstat/`.

Mills, T. (1999) *The econometric modelling of financial time series*. Cambridge University Press.

Morris, C. N. (1983) Natural exponential families with quadratic variance functions: Statistical theory. *The Annals of Statistics*, **11**, 515–529.

Neal, R. M. (2001) Defining priors for distributions using Dirichlet diffusion trees. *unpublished*.

Nielsen, B. and Shephard, N. (2003) Likelihood Analysis of a First Order Autoregressive Model with Exponential Innovations. To appear in Journal of Time Series Analysis.

Norris, J. (1997) *Markov chains*. Cambridge University Press.

O'Hagan, A. (1979) On outlier rejection phenomena in bayes inference. *Journal of the Royal Statistical Society, Series B, Methodological*, **41**, 358–367.

— (1994) *Kendall's Advanced Theory of Statistics. Volume 2B: Bayesian Inference*. Cambridge University Press.

Papaspiliopoulos, O., Roberts, G. and Sköld, M. (2003) Non-centered parameterisations for hierarchical models and data augmentation. In *Bayesian Statistics 7* (eds. J. Bernardo, J. Berger, A. Dawid and A. Smith), to appear. Oxford University Press.

Pericchi, L. and Smith, A. (1992) Exact and approximate posterior moments for a normal location parameter. *Journal of the Royal Statistical Society, Series B, Methodological*, **54**, 793–804.

Pitt, M. K. and Shephard, N. (1999) Analytic convergence rates and parameterisation issues for the Gibbs sampler applied to state space models. *J. Time Ser. Anal.*, **20**, 63–85.

Preston, C. (1975) Spatial birth-and-death processes. *Bull. Int. Statist. Inst.*, **46**, 371–391.

— (1976) *Random fields*. Springer-Verlag, Berlin.

Richardson, S. and Green, P. (1998) On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society, Series B*, **59**, 731–792.

Ripley, B. D. (1987) *Stochastic Simulation*. Wiley.

Robert, C. (1996) Mixtures of distributions: inference and estimation. In *MCMC in practice* (eds. W. Gilks, S. Richardson and D. Spiegelhalter), 46–57. Chapman and Hall.

Robert, C. and Casella, G. (1999) *Monte Carlo Statistical Methods*. Springer.

Roberts, G. (1996) Markov chain concepts related to sampling algorithms. In *MCMC in practice*, 46–57. Chapman and Hall. Ed. Gilks, W.R. and Richardson, S. and Spiegelhalter, D.J.

— (2003) Linking theory and practice of MCMC. In *Highly Structured Stochastic Systems* (eds. P. Green, N. Hjort and S. Richardson). Oxford University Press.

Roberts, G., Gelman, A. and Gilks, W. (1997) Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Applied Probability*, **7**, 110–120.

Roberts, G. and Papaspiliopoulos, O. (2003) Convergence of MCMC for linear hierarchical models with heavy tailed links. *in preparation*.

Roberts, G. and Rosenthal, J. (2001) Markov chains and de-initializing processes. *Scan. J. Statistics*, **28**, 489–504.

Roberts, G. and Stramer, O. (2001) Bayesian inference for incomplete observations of diffusion processes. *Biometrika*, **88**, 203–221.

Roberts, G. and Tweedie, R. (1996a) Exponential convergence of Langevin diffusions and their discrete approximations. *Bernoulli*, **2**, 341–364.

— (1996b) Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, **83**, 95–110.

— (2004) *Understanding MCMC*. Springer.

Roberts, G. and Yuen, J. (2003) Scaling of high dimensional Metropolis algorithms on discontinuous densities. Work on progress.

Roberts, G. O. (1992) Comment to Parameterization issues in Bayesian inference by S.E. Hills and A.F.M. Smith.'. In *Bayesian Statistics 4* (eds. J. Bernardo, J. Berger, A. Dawid and A. Smith), 241. Oxford.

Roberts, G. O., Papaspiliopoulos, O. and Dellaportas, P. (2003) Bayesian inference for Non-Gaussian Ornstein-Uhlenbeck Stochastic Volatility processes. *submitted for publication*.

Roberts, G. O. and Sahu, S. K. (1997) Updating schemes, correlation structure, blocking and parameterisation for the Gibbs sampler. *J. Roy. Statis. Soc., B*, **59**, 291–397.

— (2001) Rate of convergence of the gibbs sampler by gaussian approximation. *Journal of Computational and Graphical Statistics*, **10**.

Roeder, K. and Wasserman, L. (1997) Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, **92**, 894–902.

Rogers, L. C. G. and Williams, D. (1994) *Diffusions, Markov Processes, and Martingales. Volume 1: Foundations (Second Edition)*. Wiley.

Rosinski, J. (2002) Series representations of Lévy processes from the persective of point processes. In *Lévy Processes – Theory and Applications* (eds. O. E. Barndorff-Nielsen, T. Mikosch and S. Resnick). Birkhauser, Boston.

Rubin, D. B. (1987) *Multiple Imputation for Nonresponse in Surveys*. Wiley.

Rydberg, T. and Shephard, N. (1998) Dynamics of trade-by-trade price movements: decomposition and models. *Wrkshp Econometrics and Finance, Issac Newton Institute, Cambridge, Oct.*

Rynne, B. P. and Youngson, M. A. (2000) *Linear functional analysis*. Springer.

Sahu, S. and Roberts, G. (1999) On convergence of the em algorithm and the gibbs sampler. *Statistics and Computing*, **9**, 55–64.

Sato, K. (1999) *Lévy Processes and Infinitely Divisible Distributions*. Cambridge University Press.

Shephard, N. (1996) Statistical aspects of ARCH and Stochastic Volatility. In *Time Series Models In Econometrics, Finance and Other Fields* (eds. D. Cox, O. Barndorff-Nielsen and D. Hinkley), 1–67. London: Chapman and Hall.

Sköld, M. and Roberts, G. (2003) Density estimation for the Metropolis-Hastings algorithm. To appear in Scand. J. Statist.

Smith, A. F. M. and Roberts, G. O. (1993) Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods (disc: P53-102). *Journal of the Royal Statistical Society, Series B, Methodological*, **55**, 3–23.

Stephens, M. (2000) Bayesian analysis of mixture models with an unknown number of components - an alternative to reversible jump methods. *Annals of Statistics*, **28**, 40–74.

Stoyan, D., Kendall, W. and Mecke, J. (1995) *Stochastic geometry and its applications (Second Edition)*. Wiley.

Strauss, D. J. (1975) A model for clustering. *Biometrika*, **62**, 467–476.

Tanner, M. (1996) *Tools for Statistical inference: methods for exploration of posterior distributions and likelihood functions, 3rd edition*. Springer-Verlag.

Tanner, M. A. and Wong, W. H. (1987) The calculation of posterior distributions by data augmentation. *JASA*, **82**, 528–540.

Taylor, S. (1986) *Modelling Financial Time Series*. Wiley.

Tierney, L. (1994) Markov chains for exploring posterior distributions (disc: P1728-1762). *The Annals of Statistics*, **22**, 1701–1728.

— (1998) A note on Metropolis-Hastings kernels for general state-spaces. *Ann. Appl. Prob.*, **8**, 1–9.

Wakefield, J. C., Smith, A. F. M., Racine-Poon, A. and Gelfand, A. E. (1994) Bayesian analysis of linear and non-linear population models by using the Gibbs sampler. *Applied Statistics*, **43**, 201–221.

Walker, S. and Damien, P. (1998) A Full Bayesian Non-parametric Analysis Involving a Neutral to the Right Process. *Skandinavian Journal of Statistics*, **25**, 669–680.

— (2000) Representations of Lévy processes without Gaussian components. *Biometrika*, **87**, 477–483.

Walker, S., Damien, P., Laud, W. and Smith, A. (1999) Bayesian nonparametric inference for random distributions and related functions (with discussion). *Journal of the Royal Statistical Society, Series B, Methodological*, **61**, 485–527.

West, M. and Harrison, J. (1990) *Bayesian Forecasting and Dynamic Models*. Springer-Verlag.

Whittaker, J. (1990) *Graphical models in applied multivariate statistics*. Wiley.

Wild, P. and Gilks, W. R. (1993) Algorithm As 287: Adaptive rejection sampling from log-concave density functions. *Applied Statistics*, **42**, 701–708.

Williams, D. (1991) *Probability With Martingales*. Cambridge University Press.

Wolpert, R. and Ickstadt, K. (1998) Poisson/gamma random field models for spatial statistics. *Biometrika*, **85**, 251–267.

Wolpert, R., Ickstadt, K. and Hansen, M. (2003) A nonparametric Bayesian approach to inverse problems. In *Bayesian Statistics 7* (eds. J. Bernardo, J. Berger, A. Dawid and A. Smith), to appear. Oxford University Press.