Biometrika (2006), **93,** 2, *pp.* 451–458

Printed in Great Britain

Miscellanea

An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants

By J. MØLLER

Department of Mathematical Sciences, Aalborg University, Fredrik Bajers Vej 7G, 9220 Aalborg E, Denmark

jm@math.aau.dk

A. N. PETTITT, R. REEVES

School of Mathematical Sciences, Queensland University of Technology, GPO Box 2434, Brisbane, Queensland 4001, Australia

a.pettitt@qut.edu.au r.reeves@qut.edu.au

and K. K. BERTHELSEN

Department of Mathematical Sciences, Aalborg University, Fredrik Bajers Vej 7G, 9220 Aalborg E, Denmark

kkb@math.aau.dk

Summary

Maximum likelihood parameter estimation and sampling from Bayesian posterior distributions are problematic when the probability density for the parameter of interest involves an intractable normalising constant which is also a function of that parameter. In this paper, an auxiliary variable method is presented which requires only that independent samples can be drawn from the unnormalised density at any particular parameter value. The proposal distribution is constructed so that the normalising constant cancels from the Metropolis–Hastings ratio. The method is illustrated by producing posterior samples for parameters of the Ising model given a particular lattice realisation.

Some key words: Auxiliary variable method; Ising model; Markov chain Monte Carlo; Metropolis-Hastings algorithm; Normalising constant; Partition function.

1. Introduction

Intractable normalising constants arise in a number of statistical problems, including the definition of Markov random fields (Besag, 1974), image analysis (Ibanez & Simo, 2003), Markov point processes (Møller & Waagepetersen, 2003), Gaussian graphical models (Roverato, 2002), neural networks (Lee, 2002) and in large dimensional Gaussian multivariate models (Wong et al., 2003). Analytical expressions for normalising constants in such cases may be known, but are computationally infeasible for many typical problems. Maximum likelihood parameter estimation, or finding Bayesian posterior distributions for parameters of these distributions, is consequently impossible by straightforward methods. A wide range of approximate techniques and stochastic approximations have been proposed for circumventing this problem. These include pseudo-likelihood (Besag, 1975), various Markov chain Monte Carlo estimation techniques (Chen & Shao,

1997) and path sampling (Gelman & Meng, 1998), with recent examples of Berthelsen & Møller (2003) and Green & Richardson (2002).

The approach proposed in this paper avoids such approximations by introducing an auxiliary variable into a Metropolis–Hastings algorithm for the posterior of the parameters of the unnormalised distribution. By judicious choice of the proposal distribution, the normalising constants are made to cancel from the Metropolis–Hastings ratio. This is achieved by proposing the auxiliary variable from the unnormalised distribution at the proposed parameter value.

2 Auxiliary variable method

We consider the problem of drawing from a posterior density

$$\pi(\theta|y) \propto \pi(\theta)\pi(y|\theta) \tag{1}$$

when the likelihood

$$\pi(y|\theta) = q_\theta(y)/Z_\theta \tag{2}$$

is given by an unnormalised density $q_\theta(y)$ but its normalising constant, or partition function, Z_θ is not available analytically or exact computation is not feasible. To generate samples from (1) by a Metropolis–Hastings algorithm, we need to calculate a Metropolis–Hastings ratio

$$H(\theta'|\theta) = \frac{\pi(\theta')q_{\theta'}(y)p(\theta|\theta')}{\pi(\theta)q_\theta(y)p(\theta'|\theta)} \Big/ \frac{Z_{\theta'}}{Z_\theta}, \tag{3}$$

where $p(\theta'|\theta)$ is the proposal density. However (3) depends on the ratio of unknown normalising constants $Z_{\theta'}/Z_\theta$. The usual approach is to replace $Z_{\theta'}/Z_\theta$ by an estimate calculated by Markov chain Monte Carlo methods, hoping that the algorithm then has an equilibrium distribution close to (1).

Our Metropolis–Hastings algorithm works as follows. Assume that x is an auxiliary variable, defined on the same state space as that of y, which has conditional density $f(x|\theta, y)$, and consider the posterior

$$\pi(\theta, x|y) \propto \pi(\theta, x, y) = f(x|\theta, y)\pi(\theta)q_\theta(y)/Z_\theta,$$

which of course still involves the unknown Z_θ. Obviously, marginalisation over x of $\pi(\theta, x|y)$ gives the desired distribution $\pi(\theta|y)$. Now, if (θ, x) is the current state of the algorithm, propose first θ' with density $p(\theta'|\theta, x)$ and next x' with density $p(x'|\theta', \theta, x)$. As usual the choice of these proposal densities is arbitrary from the point of view of the equilibrium distribution of the chain of θ-values; the choice of $f(x|\theta, y)$ is also arbitrary in this respect. We take the proposal density for the auxiliary variable x' to be the same as the likelihood, but depending on θ' rather than θ:

$$p(x'|\theta', \theta, x) = \pi(x'|\theta') = q_{\theta'}(x')/Z_{\theta'}. \tag{4}$$

Then the Metropolis–Hastings ratio

$$H(\theta', x'|\theta, x) = \frac{f(x'|\theta', y)\pi(\theta')q_{\theta'}(y)q_\theta(x)p(\theta|\theta', x')}{f(x|\theta, y)\pi(\theta)q_\theta(y)q_{\theta'}(x')p(\theta'|\theta, x)} \tag{5}$$

does not depend on $Z_{\theta'}/Z_\theta$. With probability $\min\{1, H(\theta', x'|\theta, x)\}$ we accept (θ', x') as the next state; otherwise we retain (θ, x).

For simplicity, we assume henceforth that

$$p(\theta'|\theta, x) = p(\theta'|\theta)$$

does not depend on x.

Comparing (3) and (5), we see that Z_θ has been replaced by $q_\theta(x)/f(x|\theta, y)$ and $Z_{\theta'}$ has been replaced by $q_{\theta'}(x')/f(x'|\theta', y)$. An analogy may be made with importance sampling: the importance sampling identity is

$$Z_\theta = E\{q_\theta(x)/g_\theta(x)\}, \tag{6}$$

where the expectation is taken with respect to an 'instrumental' density $g_\theta(x)$ such that $g_\theta(x) > 0$ whenever $q_\theta(x) > 0$ (Robert & Casella, 2005, p. 92). Thus our method is analogous to two single sample importance sampling estimates of the normalising constants Z_θ and $Z_{\theta'}$ in each iteration of the Markov chain.

An appropriate auxiliary density $f(x|\theta, y)$ and proposal density $p(\theta'|\theta)$ must be chosen so that the algorithm has good mixing and convergence properties. The analogy with importance sampling suggests that the auxiliary distribution should approximate the distribution given by q_θ:

$$f(x|\theta, y) \simeq q_\theta(x)/Z_\theta.$$

If $f(x|\theta, y) = q_\theta(x)/Z_\theta$ exactly, which we assume is impractical, then (3) and (5) agree and the mixing properties of the two algorithms are the same. Recommendations on how to tune Metropolis–Hastings algorithms to obtain optimal acceptance probabilities may exist in the case of (3); see for example Breyer & Roberts (2000). This may suggest how to tune our Metropolis–Hastings algorithm when $f(x|\theta, y)$ approximates $q_\theta(x)/Z_\theta$.

A simple approximation is

$$f(x|\theta, y) = q_{\tilde\theta}(x)/Z_{\tilde\theta}, \tag{7}$$

where $\tilde\theta$ is fixed and, for example, $\tilde\theta = \tilde\theta(y)$ is an estimate for θ based on the data y. Then the normalising constant $Z_{\tilde\theta}$ cancels in

$$f(x'|\theta', y)/f(x|\theta, y) = q_{\tilde\theta}(x')/q_{\tilde\theta}(x).$$

Thus the Metropolis–Hastings ratios (3) and (5) agree if $Z_\theta/Z_{\theta'}$ is estimated by bridge sampling with single sample expectations and 'bridge' $q_{\tilde\theta}$ (Meng & Wong, 1996). The choice (7) may hence be expected to work well if the posterior distribution of θ is concentrated near $\tilde\theta$, or, clearly, if $q_\theta(.)/Z_\theta$ does not strongly depend on θ.

Another approach is to make $f(x|\theta, y)$ some tractable distribution which approximates the intractable $q_\theta(x)/Z_\theta$. Partially ordered Markov models (Cressie et al., 2000) have proven suitable auxiliary functions for approximating Markov point processes (Berthelsen & Møller, 2006) and in on-going work with Ising models.

3. Application to the Ising model

3·1. *Model and algorithm*

A simple example of a distribution with an intractable normalising constant is given by the Ising model on a rectangular lattice. For large lattices and most neighbourhood structures the computation of the normalising constant is not feasible, although a number of special results are available; see for example Bartolucci & Besag (2002) and Reeves & Pettitt (2004).

Consider an Ising model with a constant external field parameter θ_0 and a constant, isotropic association parameter θ_1. The unnormalised density is given by

$$q_\theta(y) = \exp(\theta_0 V_0 + \theta_1 V_1),$$

with

$$V_0 = \sum_{i=1}^{m} \sum_{j=1}^{n} y_{i,j}, \quad V_1 = \sum_{i=1}^{m-1} \sum_{j=1}^{n} y_{i,j} y_{i+1,j} + \sum_{i=1}^{m} \sum_{j=1}^{n-1} y_{i,j} y_{i,j+1},$$

where i and j index the rows and columns of an $m \times n$ rectangular lattice and $y_{i,j} \in \{-1, 1\}$ denotes a response at location (i, j). As in most statistical applications, where only positive spatial association is a consideration, we exclude negative values of θ_1.

We use the maximum pseudolikelihood estimate (Besag, 1975) for $\tilde{\theta}$ in the auxiliary variable distribution (7), and also use it as the initial state for θ in the Metropolis–Hastings algorithm for drawing from $\pi(\theta, x|y)$. Furthermore, if $\theta = (\theta_0, \theta_1)$ is the current state of the algorithm, we draw proposals θ'_0 and θ'_1 from independent normal distributions with means θ_0 and θ_1, so that $p(\theta|\theta')/p(\theta'|\theta) = 1$. The standard deviations of these proposal distributions can be adjusted to give the best mixing of the chain. The auxiliary variable x' is then drawn from (4) by perfect simulation (Propp & Wilson, 1996). Also we assume a uniform prior on

$$\theta \in \Theta = [\min \theta_0, \max \theta_0] \times [0, \max \theta_1],$$

where $\min \theta_0 < 0$, $\max \theta_0 > 0$ and $\max \theta_1$ are large but finite numbers. Then $\pi(\theta')/\pi(\theta)$ is the indicator function $1[\theta' \in \Theta]$, and the Metropolis–Hastings ratio (5) reduces to

$$H(\theta', x'|\theta, x) = 1[\theta' \in \Theta] \frac{q_{\tilde{\theta}}(x') q_{\theta'}(y) q_{\theta}(x)}{q_{\tilde{\theta}}(x) q_{\theta}(y) q_{\theta'}(x')}.$$

In practice, the exact values of $\min \theta_0 < 0$, $\max \theta_0 > 0$ and $\max \theta_1$ have very little influence on the chain, as long as they are large enough so that proposals very rarely fall outside them. Ranges for θ_0 of ± 1 and for θ_1 of $[0, 1)$ are quite adequate for the examples we consider.

3·2. *Analytical and empirical results*

Table 1 summarises some results for a 10×30 lattice with data simulated, by perfect simulation, from Ising models at five different values of θ. For a lattice of this size, the posterior modes can be computed exactly using a forward recursion algorithm for the normalising constant (Reeves & Pettitt, 2004). We can also estimate the posterior standard deviation analytically using Laplace's method (Gelman et al., 2004, p. 341), which entails fitting a quadratic to the logarithm of the

Table 1. *Summary of analytical and Markov chain Monte Carlo estimates for the posteriors of θ_0 and θ_1 for five different Ising models on a 10×30 lattice. The standard deviations for the analytically estimated posterior modes were estimated by Laplace's method. The means and standard deviations of the Markov chain Monte Carlo draws are calculated from 100 000 iterations of the chain, with no burn-in. The standard errors,* STE, *of the means were computed taking into account the correlation within samples*

True θ_0	True θ_1	Analytical posterior Mode	Analytical posterior Est. STD	MCMC posterior Mean	MCMC posterior STE	MCMC posterior STD
		Parameter of interest θ_0				
0·0	0·1	−0·085	0·054	−0·084	16×10^{-5}	0·055
0·0	0·2	0·020	0·034	0·021	$6{\cdot}1 \times 10^{-5}$	0·036
0·0	0·3	0·015	0·023	0·023	$6{\cdot}9 \times 10^{-5}$	0·027
0·1	0·1	0·085	0·050	0·084	11×10^{-5}	0·047
0·1	0·2	0·074	0·039	0·079	$8{\cdot}9 \times 10^{-5}$	0·038
		Parameter of interest θ_1				
0·0	0·1	0·057	0·042	0·059	$5{\cdot}0 \times 10^{-5}$	0·034
0·0	0·2	0·223	0·038	0·219	$5{\cdot}1 \times 10^{-5}$	0·037
0·0	0·3	0·320	0·034	0·311	$7{\cdot}8 \times 10^{-5}$	0·033
0·1	0·1	0·109	0·042	0·109	$8{\cdot}7 \times 10^{-5}$	0·042
0·1	0·2	0·264	0·038	0·258	$9{\cdot}9 \times 10^{-5}$	0·038

MCMC, Markov chain Monte Carlo; Est. STD, estimated standard deviation; STE, standard error; STD, standard deviation.

posterior in the region of the mode, from which the Hessian is estimated. In Table 1, we compare the analytically-obtained posterior mode and estimated posterior standard deviation to the posterior mean and posterior standard deviation given by the Markov chain Monte Carlo algorithm after 100 000 iterations. The respective posterior mode and mean are rather similar, the standard deviations are of the same magnitude, and each posterior mode or mean departs from

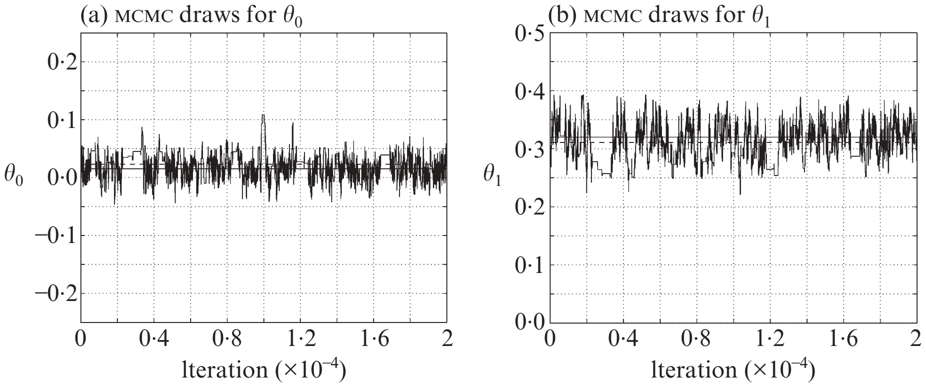

Fig. 1. The first 20 000 Markov chain Monte Carlo draws of (a) θ_0 and (b) θ_1, with the lattice data simulated as in Table 1 at parameter values $\theta_0 = 0{\cdot}0$ and $\theta_1 = 0{\cdot}3$. The analytically computed posterior mode appears as an unbroken line, while the simulation average, over 100 000 iterations, is shown as a dashed line.

Table 2. *Summary of Markov chain Monte Carlo posteriors for* θ_0 *and* θ_1 *for different lattices where posterior modes are unavailable analytically. The Markov chain Monte Carlo calculations are based on* 100 000 *iterations of the chain, with no burn-in, and the following are shown*: 'Prop σ', *the proposal standard deviation for* θ_0 *and* θ_1; *the posterior means* $\bar{\theta}_0$ *and* $\bar{\theta}_1$ *and their standard errors* $\sigma_{\bar{\theta}_0}$ *and* $\sigma_{\bar{\theta}_1}$; *the posterior standard deviations* σ_{θ_0} *and* σ_{θ_1}; c_{θ_0} *and* c_{θ_1}, *the corresponding lag*-100 *autocorrelations*; *and* 'Extr', *the proportion of acceptance ratios below* $\exp(-10)$

		Case 1 100 × 100	Case 2 50 × 50	Case 3 50 × 50	Case 4 50 × 50
True	θ_0	0·1	0·2	0·2	0·0
	θ_1	0·2	0·1	0·1	0·3
MPLE	$\hat{\theta}_0$	0·115	0·225	0·217	−0·001
	$\hat{\theta}_1$	0·195	0·105	0·108	0·309
MCMC	Prop σ	0·005	0·005	0·01	0·005
	$\bar{\theta}_0$	0·111	0·220	0·220	0·000
	$\sigma_{\bar{\theta}_0}$	$7{\cdot}7 \times 10^{-6}$	$6{\cdot}4 \times 10^{-5}$	$2{\cdot}6 \times 10^{-5}$	$7{\cdot}9 \times 10^{-6}$
	σ_{θ_0}	0·0083	0·023	0·023	0·007
	$\bar{\theta}_1$	0·199	0·108	0·107	0·312
	$\sigma_{\bar{\theta}_1}$	$4{\cdot}6 \times 10^{-6}$	$3{\cdot}2 \times 10^{-5}$	$1{\cdot}4 \times 10^{-5}$	$9{\cdot}8 \times 10^{-6}$
	σ_{θ_1}	0·0066	0·015	0·015	0·011
	c_{θ_0}	0·192	0·502	0·208	0·089
	c_{θ_1}	0·132	0·431	0·183	0·125
	Extr	0·085	0·020	0·057	0·041

MPLE, maximum pseudolikelihood estimate; MCMC, Markov chain Monte Carlo.

the true values of θ_0 or θ_1 by at most two times the posterior standard deviation. These results are consistent with adequate convergence. The Markov chain Monte Carlo standard errors were computed using the 'coda' package described in a technical report by N. G. Best, M. K. Cowles and S. K. Vines, from the MRC Biostatistics unit at the University of Cambridge.

Figure 1 shows traces of parameters θ_0 and θ_1 for the first 20 000 iterations of the Markov chain Monte Carlo posterior simulations. Some stickiness is apparent in isolated areas of the traces, and this becomes increasingly prevalent for higher parameter values of θ_1.

In Table 2 we show results for lattice sizes of up to 100×100. While we cannot compare these Markov chain Monte Carlo posterior summaries with an analytical equivalent, the histograms of the posteriors and the parameter traces of Fig. 2 appear consistent with an adequate degree of convergence.

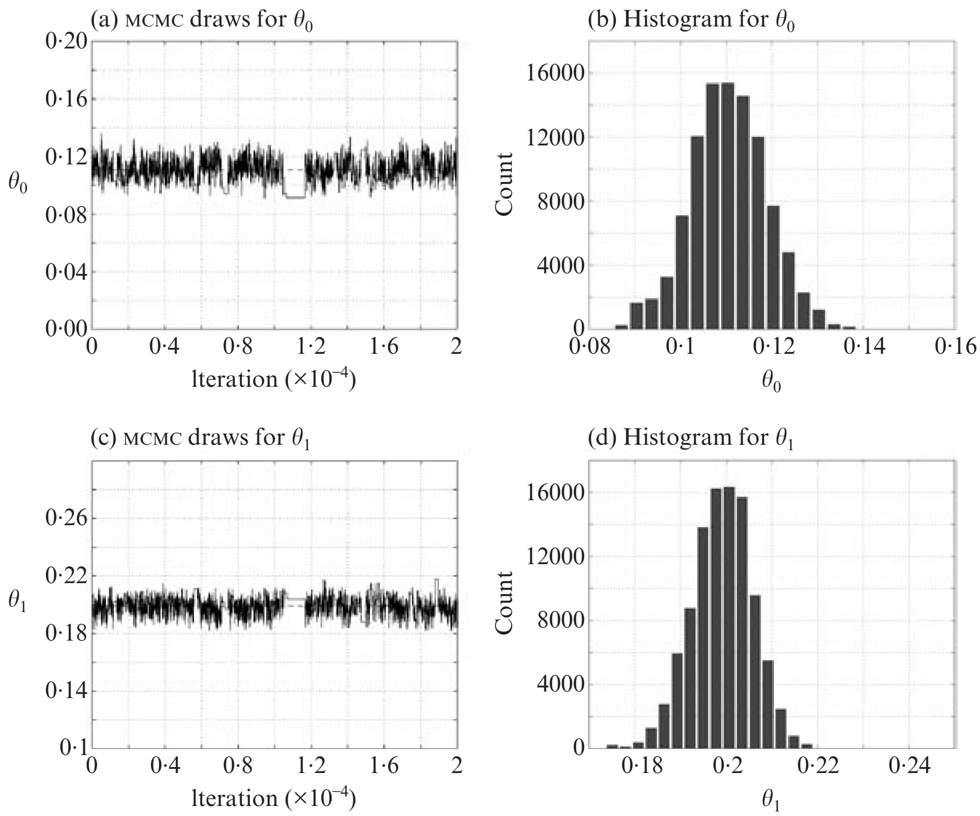

Fig. 2. Traces of the first 20 000 iterations, (a) and (c), and posterior histograms, (b) and (d), for θ_0 and θ_1 based on 100 000 iterations. Data were simulated for an Ising model with $\theta_0 = 0{\cdot}1$, $\theta_1 = 0{\cdot}2$, and lattice size 100×100, Case 1 of Table 2.

4. Discussion

A characteristic of this algorithm appears to be the occasional appearance of long runs where no proposal is accepted, as for example in the vicinity of iteration 11 000 in Fig. 2. The results presented in Table 2 and Fig. 2 suggest that the tendency to produce these long runs becomes worse for large lattices, and for increasing values of the association parameter θ_1. One expects the acceptance rate to fall as the dimensionality of the Metropolis–Hastings step increases, as is typically the case with block updates. The dependence on the association parameter is perhaps explained by the mismatch between the auxiliary variable distribution $f(x|\theta, y)$ and the likelihood $q_\theta(.)/Z_\theta$, which with our choice of auxiliary variable distribution becomes worse as θ_1 increases towards criticality. This is because the maximum pseudolikelihood parameter estimate is increasingly poor as θ_1 increases (Geyer & Thompson, 1992), and, as θ_1 becomes larger, the fixed parameter value may be less effective at approximating both the current and proposed θ values. Replacing the maximum pseudolikelihood estimate of the parameters with the maximum likelihood estimate, based on stochastic estimation of the normalising constants, may lead to improvement, although this introduces a substantially greater computational burden. An auxiliary variable distribution which better approximates the likelihood $q_\theta(.)/Z_\theta$ may also lead to improvement.

Partly as a consequence of long runs of nonacceptance, and partly as a consequence of requiring a small proposal standard deviation for the parameters in order to keep these long runs to a minimum, the autocorrelations of the parameter samples can be quite high. For example, in Table 2, Case 2, corresponding to a 50×50 lattice, with a proposal standard deviation of 0·005, has lag-100 autocorrelations of 0·50 and 0·43. This indicates a need for a large number of iterations in order to be assured of convergence.

In the symmetrical case $\theta_0 = 0$, as θ_1 approaches the critical value of about 0·44 for the Ising model, the perfect sampling algorithm becomes time-consuming. While more sophisticated perfect sampling algorithms avoid this critical slowdown (Propp & Wilson, 1996), the region near criticality, in which the Ising model tends toward a predominance of one value over the other, is generally not particularly useful in statistical models for studying spatial association.

The ability to draw perfect samples of the auxiliary variable from the likelihood ensures that the posterior for the parameters of interest arises exactly from marginalising the equilibrium distribution of the Markov chain. It is not necessary to use Markov chain methods such as coupling from the past and its developments for this, if simpler, direct methods are available. Any method of drawing from the likelihood is acceptable, with the proviso that Markov chain methods must have converged adequately to the equilibrium distribution to avoid introducing additional undesirable stochasticity.

The technique proposed in this paper is applicable to Markov chain Monte Carlo methods for inference in which a normalising constant is unknown, whenever samples may be drawn from the likelihood without approximation, by perfect simulation for example. It overcomes the need to resort to some computationally demanding approximate analysis, such as stochastic estimation of normalising constant ratios. Our method eliminates such sources of error in posterior inference, as well as being more easily implemented for those problems for which it is applicable.

Recent research explores connections with approximate Bayesian computation in which an auxiliary variable is used to eliminate intractable likelihoods from a Metropolis–Hastings algorithm (Reeves & Pettitt, 2005). This is an example of the method in a hierarchical model setting as suggested in an unpublished Aalborg University technical report by the authors.

Acknowledgement

The research of J. Møller was supported by the Danish Natural Science Research Council and the Network in Mathematical Physics and Stochastics, MaPhySto, funded by grants from the Danish National Research Foundation. The research of A. N. Pettitt and R. W. Reeves was supported by the Australian Research Council and a Queensland University of Technology Post Doctoral Fellowship.

References

Bartolucci, F. & Besag, J. (2002). A recursive algorithm for Markov random fields. *Biometrika* **89**, 724–30.

Berthelsen, K. K. & Møller, J. (2003). Likelihood and non-parametric Bayesian MCMC inference for spatial point processes based on perfect simulation and path sampling. *Scand. J. Statist.* **30**, 549–64.

Berthelsen, K. K. & Møller, J. (2006). Bayesian analysis of Markov point processes. In *Case Studies in Spatial Point Process Modeling*, Ed. A. Baddeley, P. Gregori, J. Mateu, R. Stoica and D. Stoyan, pp. 85–97, New York: Springer.

Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician* **24**, 179–95.
Besag, J. E. (1974). Spatial interaction and the statistical analysis of lattice systems (with Discussion). *J. R. Statist. Soc.* B **36**, 192–236.
Breyer, L. A. & Roberts, G. O. (2000). From Metropolis to diffusions: Gibbs states and optimal scaling. *Stoch. Proces. Applic.* **90**, 181–206.
Chen, M.-H. & Shao, Q.-M. (1997). On Monte Carlo methods for estimating ratios of normalizing constants. *Ann. Statist.* **25**, 1563–94.
Cressie, N., Zhu, J., Baddeley, A. J. & Nair, M. G. (2000). Directed Markov point processes as limits of partially ordered Markov models. *Methodol. Comp. Appl. Prob.* **2**, 5–21.
Gelman, A. & Meng, X.-L. (1998). Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statist. Sci.* **13**, 163–85.
Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (2004). *Bayesian Data Analysis*, 2nd ed. Boca Raton, FL: Chapman and Hall/CRC.
Geyer, C. J. & Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data (with Discussion). *J. R. Statist. Soc.* B **54**, 657–99.
Green, P. J. & Richardson, S. (2002). Hidden Markov models and disease mapping. *J. Am. Statist. Assoc.* **97**, 1055–70.
Ibanez, M. & Simo, A. (2003). Parametric estimation in Markov random field image modeling with imperfect observations. A comparative study. *Pat. Recog. Lett.* **24**, 2377–89.
Lee, H. K. H. (2002). Difficulties in estimating the normalizing constant of the posterior for a neural network. *J. Comp. Graph. Statist.* **11**, 222–35.
Meng, X. L. & Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statist. Sinica* **6**, 831–60.
Møller, J. & Waagepetersen, R. P. (2003). *Statistical Inference and Simulation for Spatial Point Processes*. Boca Raton, FL: Chapman and Hall/CRC.
Propp, J. & Wilson, D. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Struct. Algor.* **9**, 223–52.
Reeves, R. & Pettitt, A. (2005). A theoretical framework for approximate Bayesian computation. In *20th International Workshop on Statistical Modelling*, Ed. A. Francis, K. Matawie, A. Oshlack and G. Smyth, pp. 393–6. Sydney, Australia: University of Western Sydney.
Reeves, R. & Pettitt, A. N. (2004). Efficient recursions for general factorisable models. *Biometrika* **91**, 751–7.
Robert, C. & Casella, G. (2005). *Monte Carlo Statistical Methods*, 2nd ed. New York: Springer-Verlag.
Roverato, A. (2002). Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scand. J. Statist.* **29**, 391–411.
Wong, F., Carter, C. K. & Kohn, R. (2003). Efficient estimation of covariance selection models. *Biometrika* **90**, 809–30.

[*Received February* 2004. *Revised November* 2005]