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

J. Møller[1], A.N. Pettitt[2], K. K. Berthelsen[1] and R.W. Reeves[2]

Abstract

We present new methodology for drawing samples from a posterior distribution when (i) the likelihood function or (ii) a part of the prior distribution is only specified up to a normalising constant. In the case (i), the novelty lies in the introduction of an auxiliary variable in a Metropolis-Hastings algorithm and the choice of proposal distribution so that the algorithm does not depend upon the unknown normalising constant. In the case (ii), similar ideas apply and the situation is even simpler as no auxiliary variable is required. Our method is "on-line" as compared with alternative approaches to the problem which require "off-line" computations. Since it is needed to simulate from the "unknown distribution", e.g. the likelihood function in case (i), perfect simulation such as the Propp-Wilson algorithm becomes useful. We illustrate the method in case (i) by producing posterior samples when the likelihood is given by an Ising model and by a Strauss point process.

Keywords: Autologistic model; Auxiliary variable method; Hierarchical model; Ising model; Markov chain Monte Carlo; Metropolis-Hastings algorithm; Normalising constant; Partition function; Perfect simulation; Strauss point process.

1 Introduction

Unnormalised probability distributions with intractable normalising constants arise in a number of statistical problems, including the definition of Gibbs distributions such as Markov random fields (Besag, 1974; Cressie, 1993) and Markov point processes (Ripley and Kelly, 1977; Møller and Waagepetersen, 2003b). A simple example 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 (e.g. Bartolucci and Besag, 2002; Pettitt et al., 2003; Reeves and Pettitt, 2003).

We consider the problem of wanting to draw 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 known. By "not known", we mean that Z_θ is not available analytically and/or that exact computation is not feasible. A related problem occurs for

[1]Department of Mathematical Sciences, Aalborg University, Denmark
[2]School of Mathematical Sciences, Queensland University of Technology, Australia

Bayesian hierarchical models, or more generally directed graphical models, with an unobserved layer: Let $\theta = (\theta_1, \theta_2)$ and $\pi(\theta) = \pi(\theta_1)\pi(\theta_2|\theta_1)$, where

$$\pi(\theta_2|\theta_1) = q_{\theta_1}(\theta_2)/Z_{\theta_1} \tag{3}$$

depends on an unknown normalising constant Z_{θ_1}, while now $\pi(y|\theta)$ is known. For specificity, unless otherwise stated, we consider the setting of (1) with Z_θ in (2) unknown, but the ideas presented later in this paper also apply when Z_{θ_1} in (3) is unknown. In Section 2.2 we take up the issue of using our methodology in a hierarchical model.

It is not straightforward to generate samples from (1) by Markov chain Monte Carlo (MCMC) algorithms. In a Metropolis-Hastings algorithm, if θ is the current state of the chain and a proposal θ' with density $p(\theta'|\theta)$ is generated, θ' is accepted as the new state with probability $\alpha(\theta'|\theta) = \min\{1, H(\theta'|\theta)\}$, and otherwise we retain θ, where

$$H(\theta'|\theta) = \frac{\pi(\theta'|y)p(\theta|\theta')}{\pi(\theta|y)p(\theta'|\theta)}$$

is the Hastings ratio (e.g. Tierney, 1994). By (2),

$$H(\theta'|\theta) = \frac{\pi(\theta')q_{\theta'}(y)p(\theta|\theta')}{\pi(\theta)q_\theta(y)p(\theta'|\theta)} \Bigg/ \frac{Z_{\theta'}}{Z_\theta} \tag{4}$$

is unknown, since it depends on the ratio of unknown normalising constants $Z_{\theta'}/Z_\theta$. Ratios of unknown normalising constants also appear in other types of algorithms, including the sampling/importance resampling (SIR) algorithm (Rubin, 1987).

Because of their intractability, earlier work attempted to avoid algorithms involving unknown normalising constants. Three different examples of this are:

1. Besag et al. (1991), who considered an ad hoc procedure for a Bayesian hierarchical model for location of archaeological sites, which at one layer included an Ising prior (this may be considered as an example of (3) above with a uniform hyperprior on θ_1, though Besag et al. did not specify any such hyperprior). They adjusted the posterior density in an iterative MCMC scheme where the parameter of the Ising model (here θ_1) was estimated by pseudo likelihood (Besag, 1975).

2. Heikkinen and Högmander (1994) approximated the likelihood term (2) in (1) by a pseudo likelihood function of easily derivable full conditional distributions. The approximation involves improperly normalised distributions, and Heikkinen & Högmander admit that it is unclear what the distribution is that they are actually simulating.

3. Instead of estimating the entire posterior distribution, Heikkinen and Penttinen (1999) focused on finding the maximum a posteriori estimate for the interaction function in a Bayesian model where the likelihood function is given by a pairwise interaction point processes and its normalising constant is unknown (for a more detailed Bayesian analysis, see Berthelsen and Møller, 2003).

More recent papers dealing with a ratio of unknown normalising constants in a Metropolis-Hastings algorithm, use a Monte Carlo technique to estimate it:

4. Green and Richardson (2002), Sebastiani and Sørbye (2002), Dryden et al. (2003), and Berthelsen and Møller (2003) use this 'off-line' approach. Specifically, they use path sampling or, as it is known in statistical physics, thermodynamic integration. This is based on the path sampling identity
$$\log \frac{Z_{\theta''}}{Z_{\theta'}} = \int_0^1 v(\gamma(t))^T \frac{\mathrm{d}}{\mathrm{d}t} \gamma(t)\, \mathrm{d}t \tag{5}$$
where γ is a path in the parameter space, with $\theta' = \gamma(0)$ and $\theta'' = \gamma(1)$, and
$$v(\theta) = E_{y|\theta} \left\{ \frac{\mathrm{d}}{\mathrm{d}\theta} \log q_\theta(y) \right\} \tag{6}$$
(for details, see Gelman and Meng, 1998). Here $v(\theta)$ is found by generating values of y from $\pi(y|\theta)$, which is possible using a regular Metropolis-Hastings algorithm, for example, since the normalising constant Z_θ is fixed and cancels in the Hastings ratio for proposals.

The approach introduced in this paper avoids approximations such as those in points 1 to 4 above. Instead, we introduce an auxiliary variable x into a Metropolis-Hastings algorithm for (θ, x) so that ratios of normalising constants no longer appear but the posterior distribution for θ is retained. Generally, auxiliary variables have been introduced into MCMC schemes in order to improve mixing of the chains, and ease simulation of variables (Swendsen and Wang, 1987; Edwards and Sokal, 1988; Besag and Green, 1993; Higdon, 1998). Another motivation for auxiliary variables is the improvement of particle filters which are generally sensitive to outliers in the particle sample; see Pitt and Shephard (1999). Our auxiliary variable method is inspired by a certain kinship to path sampling, when the expectation in (6) is replaced by a single sample and the path in (5) is considered to be an infinitesimal one, thus collapsing the differentiation and integration. Furthermore, access to algorithms for making perfect (or exact) simulations (Propp and Wilson, 1996) from (2) or (3) is another motivation as explained later.

Section 2 presents the method in a general setting. Section 3 applies the method to the autologistic model, and compares MCMC posteriors for θ to the analytically obtained posterior modes for lattices where the normalising constant is tractable using recursion methods (Reeves and Pettitt, 2003). We also show MCMC posteriors for larger lattices where the normalising constant is intractable. Section 4 illustrates the method applied to a Strauss point process. Section 5 concludes with some further discussion.

2 General methods

In this section we develop methods for eliminating unknown normalising constants in the Hastings ratio. Section 2.1 consider the basic case (1) where the normalising constant Z_θ in the likelihood (2) is unknown. Section 2.2 considers the case of a hierarchical model with a hidden layer which is specified by a density (3) with an unknown normalising constant.

2.1 Auxiliary variable method

In case of (1) when Z_θ in (2) is unknown, we introduce an auxiliary variable x defined on the same space as the state space of y. Assume that x has conditional density $f(x|\theta, y)$, so that the joint density of (θ, x, y) is given by

$$\pi(\theta, x, y) = f(x|\theta, y)\pi(y|\theta)\pi(\theta).$$

The posterior density with $\pi(y|\theta)$ given by (2),

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

still involves the unknown Z_θ.

A Metropolis-Hastings algorithm for drawing from $\pi(\theta, x|y)$ has a Hasting ratio given by

$$H(\theta', x'|\theta, x) = \frac{\pi(\theta', x'|y)p(\theta, x|\theta', x')}{\pi(\theta, x|y)p(\theta', x'|\theta, x)} = \frac{f(x'|\theta', y)\pi(\theta')q_{\theta'}(y)p(\theta, x|\theta', x')}{f(x|\theta, y)\pi(\theta)q_\theta(y)p(\theta', x'|\theta, x)} \Bigg/ \frac{Z_{\theta'}}{Z_\theta} \tag{8}$$

where $p(\theta', x'|\theta, x)$ is the proposal density for (θ', x'). The proposal density can be factorized as

$$p(\theta', x'|\theta, x) = p(x'|\theta', \theta, x)p(\theta'|\theta, x) \tag{9}$$

and the choice of proposal distribution is arbitrary from the point of view of the equilibrium distribution of the chain of θ-values. We take the proposal density for the auxiliary variable x' to be the same as the likelihood, but depending on θ', rather than θ,

$$p(x'|\theta', \theta, x) = p(x'|\theta') = q_{\theta'}(x')/Z_{\theta'}. \tag{10}$$

Then

$$H(\theta', x'|\theta, x) = \frac{f(x'|\theta', y)\pi(\theta')q_{\theta'}(y)q_\theta(x)p(\theta|\theta', x')}{f(x|\theta, y)\pi(\theta)q_\theta(y)q_{\theta'}(x')p(\theta'|\theta, x)}. \tag{11}$$

does not depend on $Z_{\theta'}/Z_\theta$, and the marginalisation over x of the equilibrium distribution $\pi(\theta, x|y)$, gives the desired distribution $\pi(\theta|y)$. In contrast to (4) we now have a much simpler problem of finding the ratio of the distributions of the proposed and current auxiliary variable, $f(x'|\theta', y)/f(x|\theta, y)$, the other factors in (11) presenting no difficulty in evaluation.

The reader may wonder why we have not proposed using a Metropolis within Gibbs algorithm (also known as hybrid Metropolis-Hastings), i.e. a Metropolis-Hastings algorithm which shifts between updates of θ and x. In such an algorithm, when updating θ given x (and y), let $p(\theta'|\theta, x)$ be the proposal density for θ'. By (7),

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

so the Hasting ratio is given by

$$H(\theta'|\theta, x) = \frac{f(x'|\theta', y)\pi(\theta')q_{\theta'}(y)p(\theta|\theta', x)}{f(x|\theta, y)\pi(\theta)q_\theta(y)p(\theta'|\theta, x)} \Bigg/ \frac{Z_{\theta'}}{Z_\theta}.$$

Here it seems difficult to choose $f(x|\theta, y)$ and $p(\theta'|\theta, x)$ so that the ratio of normalising constants cancels. It is the choice (10) which makes the Metropolis-Hastings algorithm given by (8) practicable, but this choice is not available for the Metropolis within Gibbs step, with x fixed.

Henceforth, for simplicity, we assume that

$$p(\theta'|\theta, x) = p(\theta'|\theta) \tag{12}$$

does not depend on x. For simulation from the proposal density (9) we suppose that it is straightforward to make simulations from $p(\theta'|\theta)$ but not necessarily from $p(x'|\theta', \theta, x)$; for $p(x'|\theta', \theta, x)$ given by (10) appropriate perfect simulation algorithms are desirable to avoid convergence questions of straightforward MCMC algorithms.

A critical design issue for the algorithm is to choose an appropriate auxiliary density $f(x|\theta, y)$ and proposal density $p(\theta'|\theta)$ so that the algorithm has good mixing and convergence properties. Assume for the moment that Z_θ is known and the algorithm based on (4) has good mixing properties. If we let $f(x|\theta, y) = q_\theta(x)/Z_\theta$, then by (12), (11) reduces to (4), and so the mixing and convergence properties of the two Metropolis-Hastings algorithms using (4) and (11) are the same. Recommendations on how to tune Metropolis-Hastings algorithms to obtain optimal acceptance probabilities may exist in the case of (4) (see e.g. Gelman et al. (1996), Roberts et al. (1997), Roberts and Rosenthal (1998) and Breyer and Roberts (2000)). This suggests that the auxiliary distribution should approximate the distribution given by q_θ,

$$f(x|\theta, y) \approx q_\theta(x)/Z_\theta. \tag{13}$$

Sections 3 and 4 consider cases where

$$f(x|\theta, y) = q_{\tilde{\theta}}(x)/Z_{\tilde{\theta}} \tag{14}$$

and $\tilde{\theta}$ is fixed, for example $\tilde{\theta} = \tilde{\theta}(y)$ is an approximate maximum likelihood estimate or maximum pseudo likelihood estimate for θ based on the data y. Then, since $f(x|\theta, y)$ does not depend on θ, the normalising constant $Z_{\tilde{\theta}}$ in

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

conveniently cancels. The choice (14) may work well if θ is expected to be close to $\tilde{\theta}$ or if $q_\theta(\cdot)/Z_\theta$ does not strongly depend on θ. Another choice is considered in Section 4.2 where $f(x|\theta, y)$ depends on θ (but not on y) in a way so that (13) is expected to hold.

2.2 Hierarchical models

When the unnormalised distribution appears as an unobserved layer in a hierarchical model, cf. (3), while $\pi(y|\theta)$ is known, we consider the following idea inspired by the auxiliary variable technique. Assume that we update θ using the proposal density

$$p(\theta'|\theta) = q_{\theta'_1}(\theta'_2)p(\theta'_1|\theta_1)/Z_{\theta'_1}. \tag{15}$$

A proposal is then generated by first generating $\theta'_1 \sim p(\theta'_1|\theta_1)$ and next $\theta'_2 \sim q_{\theta'_1}(\theta'_2)/Z_{\theta'_1}$. Using (15) the Hastings ratio is

$$H(\theta'|\theta) = \frac{\pi(y|\theta')\pi(\theta'_1)p(\theta_1|\theta'_1)}{\pi(y|\theta)\pi(\theta_1)p(\theta'_1|\theta_1)},$$

where the unknown normalising constants and terms in $q_{\theta_1'(\cdot)}$ and $q_{\theta_1'(\cdot)}$ all cancel.

Instead of updating θ_1 and θ_2 simultaneously, an alternative which may improve mixing is to use a Metropolis within Gibbs algorithm where we shift between updating θ_1 and θ_2. Updates of θ_1 then involve a ratio of unknown normalising constants, but introducing an auxiliary variable θ_3 corresponding to θ_2 and with conditional density $f(\theta_3|\theta_1, \theta_2, y)$, the ratio can be eliminated. Specifically, in analogy with (10)–(12),

- for the update of (θ_1, θ_3) (when θ_2 is fixed), generate a proposal $\theta_1' \sim p(\theta_1'|\theta_1)$ and $\theta_3' \sim q_{\theta_1'}(\theta_3')/Z_{\theta_1'}$, and calculate the Hastings ratio

$$H(\theta_1', \theta_3'|\theta_1, \theta_3) = \frac{f(\theta_3'|\theta_1', \theta_2, y)\pi(\theta_1')q_{\theta_1'}(\theta_2)\pi(y|\theta_1', \theta_2)q_{\theta_1}(\theta_3)p(\theta_1|\theta_1')}{f(\theta_3|\theta_1, \theta_2, y)\pi(\theta_1)q_{\theta_1}(\theta_2)\pi(y|\theta_1, \theta_2)q_{\theta_1'}(\theta_3')p(\theta_1'|\theta_1}$$

- the update of θ_2 (when (θ_1, θ_3) is fixed) is straightforward, using e.g. a random walk proposal for θ_2', and calculating the Hastings ratio (where the unknown normalising constants cancel).

Here, in analogy with (13), we want

$$f(\theta_3|\theta_1, \theta_2, y) \approx q_{\theta_1}(\theta_3)/Z_{\theta_1}$$

which may be achieved by

$$f(\theta_3|\theta_1, \theta_2, y) = q_{\tilde{\theta}_1}(\theta_3)/Z_{\tilde{\theta}_1},$$

where $\tilde{\theta}_1$ is fixed (compare with (14)), or by other suitable methods (see e.g. Section 4.2). For example, $\tilde{\theta}_1$ may be an approximate MLE, considering θ_2 as missing data and

$$\pi(y|\theta_1) = h_{\theta_1}(y)/Z_{\theta_1} = \left[\int \pi(y|\theta_1, \theta_2)q_{\theta_1}(\theta_2)\,\mathrm{d}\theta_2\right] \Big/ Z_{\theta_1}$$

as the likelihood, and using MCMC methods for estimating ratios $h_{\theta_1'}(y)/h_{\theta_1}(y)$ and $Z_{\theta_1'}/Z_{\theta_1}$ when finding the approximate MLE, see e.g. Geyer (1999) and Møller and Waagepetersen (2003b).

In the case where both (2) and (3) are unknown we introduce an auxiliary variable x with density $f(x|\theta, y)$ chosen as discussed in Section 2.1. Then, if (θ, x) is the current state of our Metropolis-Hastings algorithm for drawing from $\pi(\theta, x|y)$, we may first generate a proposal θ' from the proposal density (15), then generate $x' \sim q_{\theta'}(x')/Z_{\theta'}$, and finally calculate the Hastings ratio

$$H(\theta', x'|\theta, x) = \frac{f(x'|\theta', y)q_{\theta'}(y)q_{\theta_1'}(\theta_2')\pi(\theta_1')q_\theta(x)q_{\theta_1}(\theta_2)p(\theta_1|\theta_1')}{f(x|\theta, y)q_\theta(y)q_{\theta_1}(\theta_2)\pi(\theta_1)q_{\theta'}(x')q_{\theta_1'}(\theta_2')p(\theta_1'|\theta_1)}$$

Another possibility, instead of generating θ' from (15) we may introduce yet another auxiliary variable θ_3 in line with the Metropolis within Gibbs algorithm above.

3 Application to the autologistic model

The autologistic model (Besag, 1972; Cressie, 1993) is an example of a distribution for which the normalising constant is difficult to compute for problems of reasonable size (e.g. Pettitt et al., 2003). We apply the method of Section 2.1 to the autologistic model, and show how posterior distributions of the autologistic parameters are obtained for fairly large problems.

3.1 The autologistic model

An autologistic model for $y = (y_1, \ldots, y_k)$ where $k > 0$ is a given integer, has unnormalised density

$$q_\theta(y) = \exp\left(\sum_{i=1}^{k} \theta_i y_i + \sum_{1 \le i < j \le K} \theta_{i,j} y_i y_j\right), \quad y \in \{\pm 1\}^k \tag{16}$$

where the θ_i and $\theta_{i,j}$ are real parameters (usually, most $\theta_{i,j}$ are set to zero). A particular important example is the Ising model considered in Section 3.2. For compatibility with the Ising model we let $y \in \{\pm 1\}^k$; in the literature, it is also common to let $y \in \{0,1\}^k$ in (16); the two cases lead to equivalent classes of autologistic models.

We exploit that the full conditionals are logistic distributions: Let $y_{-i} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_k)$ denote the collection of all y_j with $j \neq i$. Then

$$(y_i + 1)/2 \,|\, y_{-i} \sim \text{Bernoulli}\big((y_i+1)/2; \text{logit}(2(\theta_i + \sum_{j:j \neq i} \theta_{i,j} y_j))\big) \tag{17}$$

where $\theta_{i,j} = \theta_{j,i}$, $\text{logit}(a) = e^a/(1+e^a)$, and $\text{Bernoulli}(z;p) = p^z(1-p)^{1-z}$ for $a \in \mathbb{R}$, $p \in (0,1)$, and $z \in \{0,1\}$.

Various perfect simulation procedures exist for the autologistic model. One method is the Propp-Wilson algorithm (Propp and Wilson, 1996) or its extensions (Häggström and Nelander, 1998; Møller, 1999), which are all based on coupling from the past (CFTP). The most tractable cases are if either all $\theta_{i,j} \ge 0$ (the Gibbs sampler based on the full conditionals (17) is then monotone) or all $\theta_{i,j} \le 0$ (Gibbs sampling is then anti-monotone). Another method is Fill's algorithm, which applies both in the monotone case (Fill, 1998) and in the anti-monotone case (Møller and Schladitz, 1999).

Inference methods for the autologistic model are well established. For convenience we use in Section 3.2 the maximum pseudo likelihood estimate (MPLE), which maximises the pseudo likelihood function given by

$$\text{PL}(\theta; y) = \prod_{i=1}^{k} \text{Bernoulli}\big((y_i+1)/2; \text{logit}(2(\theta_i + \sum_{j:j \neq i} \theta_{i,j} y_j))\big) \tag{18}$$

(Besag, 1975). The MPLE is easily found using standard software packages, since by (17) and (18), the MPLE is formally equivalent to the maximum likelihood estimate (MLE) in a logistic regression model (Possolo, 1986; Strauss and Ikeda, 1990). The MLE of the autologistic model is a more efficient estimate (Geyer, 1991; Friel and Pettitt, 2003), and it can be approximated by MCMC computation of the log normalising constant ratio (e.g. Geyer and Thompson, 1992; Gelman and Meng, 1998; Gu and Zhu, 2001).

3.2 Results for the auxiliary variable method

3.2.1 Model assumptions and other details

This section specifies the likelihood, prior, and auxiliary variable density for our experiments. Further details for the auxiliary variable algorithm are given in Appendix A.

The autologistic model, in its simplest form, is an Ising model, which models interactions on a binary $m \times n$ lattice. Its unnormalised density is given by

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

with real parameters θ_0 and θ_1 and sufficient statistics

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

where i and j now index the rows and columns of the lattice and $y_{i,j} \in \{-1, 1\}$ denotes a response at location (i, j). Here θ_1 can be regarded as an association parameter; as in most statistical applications, we exclude negative values of θ_1 in Sections 3.2.2 and 3.2.3. For $\theta_1 = 0$, the $y_{(i,j)}$ are i.i.d., and for θ_1 large, the values of the $y_{(i,j)}$ tend to associate together, so that 'clumps' of '1' and '-1' are expected to appear. When $\theta_0 = 0$ and the model is extended to the infinite lattice $\mathbb{Z}^2$, phase transition happens at the so-called critical value of $\theta_1 \approx 0.44$. This marks the introduction of long range dependence into the model, and realisations drawn from the Ising model with θ_1 in the vicinity of, or greater than this critical value, tend to 'crystallise', that is most lattice locations will be either '1' or '-1': When $\theta_0 = 0$, there is an equal chance of crystallising toward '1' or '-1', while if e.g. $\theta_0 > 0$, the probability of crystallising around '1' is greater than the probability of crystallising at '-1'.

We use the MPLE for $\tilde{\theta}$ in (14) and also use it as the initial state for θ in the Metropolis-Hastings algorithm of Section 2. 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 $p(\theta|\theta')/p(\theta'|\theta) = 1$. The standard deviations of these proposal distributions can be adjusted to give the best mixing of the chain. 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 (an improper uniform prior on $\theta \in (-\infty, \infty) \times [0, \infty)$ leads to an improper posterior in the extreme cases where the $y_{(i,j)}$ are equal or form a chess board pattern). 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 of them. Ranges for θ_0 of ± 1 and for θ_1 of $[0, 1)$ are quite adequate for the examples we consider. Then $\pi(\theta')/\pi(\theta) = \mathbf{1}[\theta' \in \Theta]$ (the indicator function that $\theta' \in \Theta$), and the Metropolis-Hastings ratio (11) reduces to

$$H(\theta', x'|\theta, x) = \mathbf{1}[\theta' \in \Theta]\frac{q_{\tilde{\theta}}(x')q_{\theta'}(y)q_\theta(x)}{q_{\tilde{\theta}}(x)q_\theta(y)q_{\theta'}(x')}. \tag{19}$$

3.2.2 Analytic and empirical results for a small lattice

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 this size lattice, the posterior modes can be computed exactly using a forward recursion algorithm for the normalising constant (Reeves and Pettitt, 2003). We can also analytically estimate the posterior standard deviation using Laplace's method (e.g. Gelman et al., 1995, p. 306), which entails fitting a quadratic to the 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 MCMC algorithm after 100,000 iterations. The respective posterior mode and mean are rather close, the standard deviations are of the same magnitude, and each posterior mode or mean departs from the true values of θ_0 or θ_1 with at most one or two times the posterior standard deviation, results consistent with adequate convergence. The MCMC standard errors were computed using the "CODA" package (Best et al., 1995).

True		Analytic Posterior θ_0		MCMC Posterior θ_0		
θ_0	θ_1	Mode	Est. STD	Mean	STE	STD
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	6.1×10^{-5}	0.036
0.0	0.3	0.015	0.023	0.023	6.9×10^{-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	8.9×10^{-5}	0.038

True		Analytic Posterior θ_1		MCMC Posterior θ_1		
θ_0	θ_1	Mode	Est. STD	Mean	STE	STD
0.0	0.1	0.057	0.042	0.059	5.0×10^{-5}	0.034
0.0	0.2	0.223	0.038	0.219	5.1×10^{-5}	0.037
0.0	0.3	0.320	0.034	0.311	7.8×10^{-5}	0.033
0.1	0.1	0.109	0.042	0.109	8.7×10^{-5}	0.042
0.1	0.2	0.264	0.038	0.258	9.9×10^{-5}	0.038

Table 1: Summary of analytic and MCMC estimates for the posteriors of θ_0 and θ_1 for five different Ising models on a 10×30 lattice. Data was simulated using CFTP at the true values of θ_0 and θ_1. The standard deviations for the analytically estimated posterior modes, were estimated by Laplace's method. The means and standard deviations of the MCMC 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.

Figure 1 shows traces of parameters θ_0 and θ_1 for the MCMC posterior simulations. Some stickiness is apparent in isolated areas of the traces, and this becomes increasingly prevalent for higher parameter values of θ_1. The reasons for this and possible ways to overcome it are discussed further in Section 3.2.3.

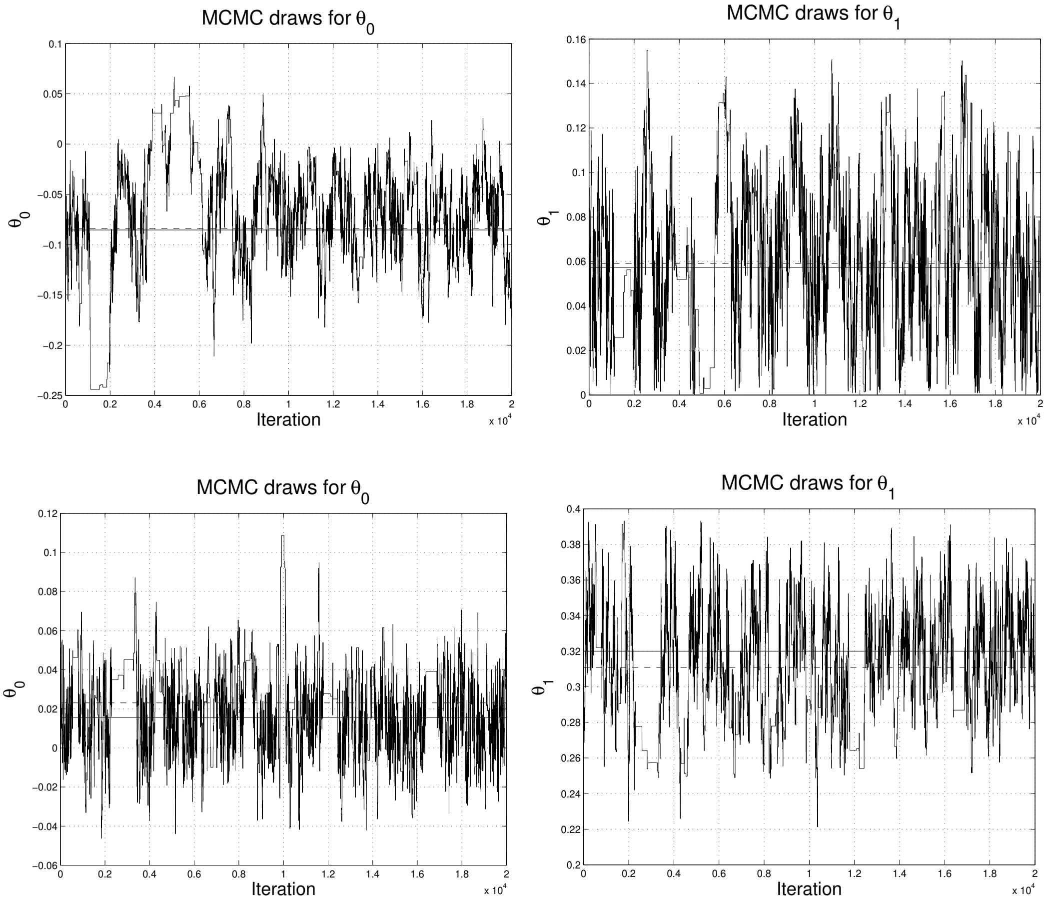

Figure 1: The first 20,000 MCMC draws of θ_0 and θ_1, when the data are simulated as in Table 1 at nominal parameter values $\theta_0 = 0.0$ and $\theta_1 = 0.1$ (top) and $\theta_0 = 0.0$ and $\theta_1 = 0.3$ (bottom). The analytically computed posterior mode appears as an unbroken line, while the simulation average (over 100,000 iterations) is shown as a dashed line.

3.2.3 Empirical results for larger lattices

We have also increased the lattice size considerably, to sizes typically encountered in statistical analyses. Applications that we have in mind are similar to recent analyses by Green and Richardson (2002), Sebastiani and Sørbye (2002), and Dryden et al. (2003) (although these analyses use the closely related Potts model); Sebastiani and Sørbye and Dryden et al. consider images of 64 $\times$ 64 pixels, while Green and Richardson consider the spatial arrangement of 94 local areas. While we now cannot compare MCMC methods of computing posterior summaries with an analytic equivalent, we can demonstrate that the chains appear to be mixing reasonably well by viewing histograms of the posteriors, the parameter traces, and examining the mean acceptance rates and lagged autocorrelations. This is illustrated in Table 2 and Figure 2.

For θ_1 approaching the critical value 0.44 (when $\theta_0 = 0$), the perfect sampling algorithm begins taking a much longer time, and mixing of the Metropolis-Hastings chain becomes problematic. While tuning the variance of the proposals for parameters θ_0 and θ_1 can go some way toward improving the mixing, such tuning becomes less effectual as θ_1 becomes greater than 0.4. For both these reasons, the auxiliary variable

		100×100	50×50	50×50	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
	$\sigma_{\bar{\theta}_0}$	7.7×10^{-6}	6.4×10^{-5}	2.6×10^{-5}	7.9×10^{-6}
	σ_{θ_0}	0.0083	0.023	0.023	0.007
	$\bar{\theta}_1$	0.199	0.108	0.107	0.312
	$\sigma_{\bar{\theta}_1}$	4.6×10^{-6}	3.2×10^{-5}	1.4×10^{-5}	9.8×10^{-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
	MAcP	0.278	0.387	0.329	0.347
	Extr	0.085	0.020	0.057	0.041

Table 2: Summary of MCMC posteriors for θ_0 and θ_1 for different lattices where posterior modes are unavailable analytically. Data were simulated from Ising models at the true parameter values indicated, using CFTP. The maximum pseudo likelihood estimates (MPLEs) of the parameters are shown, for interest. The MCMC calculations are based on 100,000 iterations of the chain, with no burn in, and the following 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; "MAcP", the mean acceptance probability; and "Extr", the proportion of acceptance ratios below exp(-10).

method appears to be inapplicable for θ_1 greater than approximately 0.4. If θ_0 is also particularly high, then this further reduces the allowable range of θ_1. Since the MPLEs are typically further from the true posterior modes as θ_1 approaches the critical value (Geyer, 1991), the density of the auxiliary variable may also less adequately approximate the true likelihood, $q_\theta(\cdot)/Z_\theta$, contributing to the poor mixing. A partial corrective action is to run the chain twice, the second time using the posterior means as parameter estimates to replace the MPLEs in the auxiliary variable distribution. This improves the mixing, but the chain is still slow to compute due to the length of time taken to draw a perfect sample. Alternatively, a better approximate MLE could be used.

Fortunately, for the construction of statistical models to study spatial association, the useful parameter range for θ_1 is probably from 0 up to about 0.4; otherwise the Ising model tends toward a predominance of one value over the other on the lattice. If it is required to extend the useful range of θ_1 above 0.4, it is possible to use perfect simulation for the random cluster model with which the Ising model has an equivalence via the Fortuin-Kastelyn-Swendsen-Wang representation (e.g. Swendsen and Wang, 1987). CFTP for the random cluster model is known to work in the vicinity of the critical value, and also for much larger lattices than considered in the present paper (Propp and Wilson, 1996, 1998).

4 Application to the Strauss process

We now consider the auxiliary variable method in the setting of spatial point processes. We restrict attention to a Strauss point process (Strauss, 1975; Kelly and Ripley, 1976).

4.1 The Strauss process

Let the Strauss process be defined on a bounded region $S \subset \mathbb{R}^2$ by a density

$$\pi(y|\theta) = \frac{1}{Z_\theta} \beta^{n(y)} \gamma^{s_R(y)} \tag{20}$$

with respect to μ which denotes a homogeneous Poisson point process on S with intensity one. Here y is a point configuration, i.e. a finite subset of S; $\theta = (\beta, \gamma, R)$, with $\beta > 0$ (known as the chemical activity in statistical physics), $0 < \gamma \leq 1$ (the interaction parameter), and $R > 0$ (the interaction range); $n(y)$ is the cardinality of y; and

$$s_R(y) = \sum_{\{\xi,\eta\} \subseteq y : \xi \neq \eta} \mathbf{1}[\|\eta - \xi\| \leq R]$$

is the number of pairs of points in y within a distance R from each other. Figure 3 shows a realisation y of a Strauss point process, where $s_R(y)$ is given by the number of pairs of overlapping discs. For $\gamma = 1$, we obtain a homogeneous Poisson process on S with intensity β. For $\gamma < 1$, typical realisations look more regular than in the case $\gamma = 1$. This is due to inhibition between the points, and the inhibition gets stronger as γ decreases or R increases. The normalising constant is unknown when $\gamma < 1$ (e.g. Møller and Waagepetersen, 2003b).

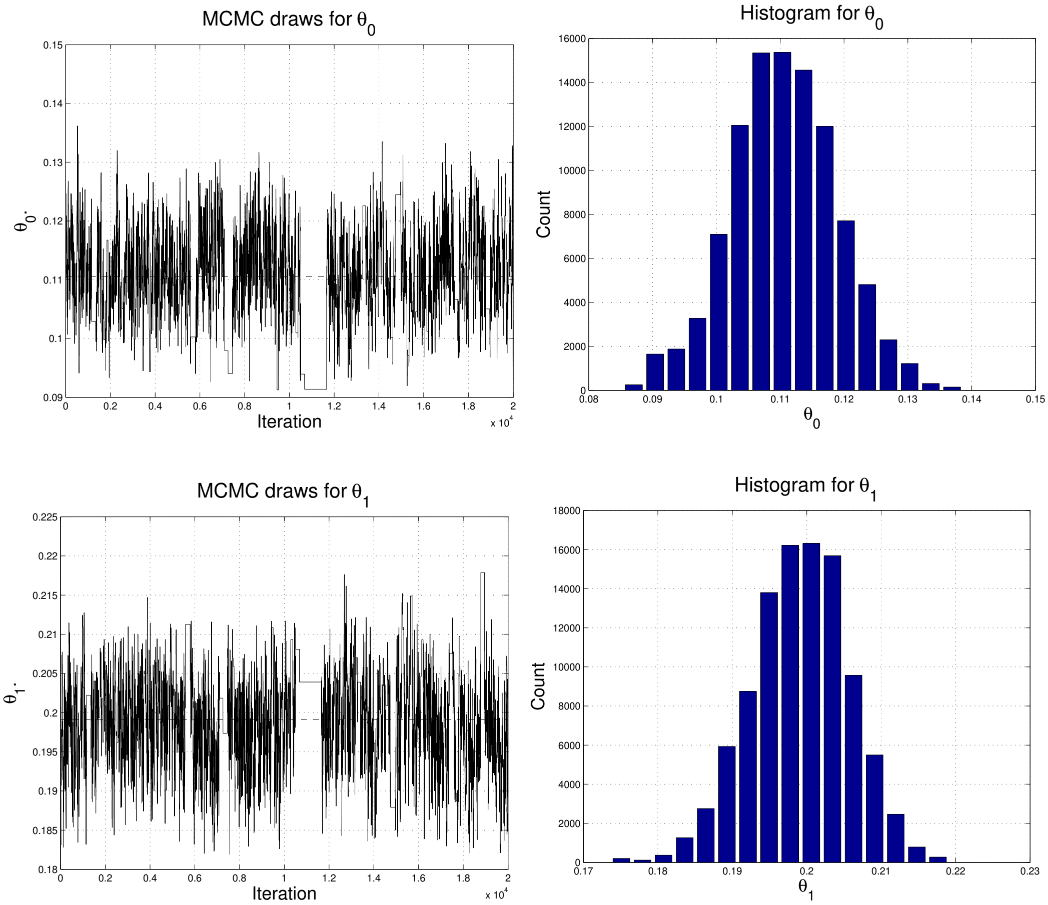

Figure 2: Traces of the first 20,000 iterations (left) and posterior histograms (right) for θ_0 and θ_1 based on 100,000 iterations. Data was simulated for an Ising model with $\theta_0 = 0.1$, $\theta_1 = 0.2$, and lattice size 100×100.

A Strauss point process is an example of a so-called locally stable point process. Such point processes can be simulated perfectly by an extension of the Propp-Wilson CFTP algorithm, called dominated CFTP, see Kendall and Møller (2000). Maximum likelihood and maximum pseudo likelihood estimation for spatial point processes and particularly the Strauss process is well established (Besag, 1977; Jensen and Møller, 1991; Geyer and Møller, 1994; Geyer, 1999; Baddeley and Turner, 2000; Møller and Waagepetersen, 2003a,b).

4.2 Specification of auxiliary point processes

In Section 4.3 we consider results for three different kinds of auxiliary variables with densities $f = f_1, f_2, f_3$ with respect to μ. In the sequel, for simplicity, we fix R, though our method extends to the case of varying interaction radius, but at the expense of further calculations.

The simplest choice is a homogeneous Poisson point process on S. We let its intensity be given by the MLE $n(y)/|S|$ based on the data y, where $|S|$ is the area of

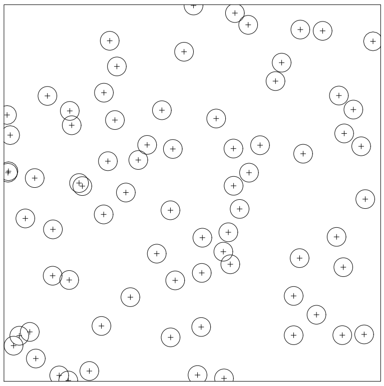

Figure 3: Realisation of a Strauss point process on the unit square, with $(\beta, \gamma, R) = (100, 0.5, 0.05)$, and generated by dominated CFTP. Circles centred at points have radii 0.025.

S. Then the auxiliary point process has density

$$f_1(x|\theta, y) = e^{|S|-n(y)}(n(y)/|S|)^{n(x)} \tag{21}$$

(e.g. Møller and Waagepetersen, 2003b). We refer to (21) as the fixed Poisson process.

The second choice takes the interaction into account. Its density is given by

$$f_2(x|\theta, y) \propto \hat{\beta}^n(x)\hat{\gamma}^{s_R(x)} \tag{22}$$

where $(\hat{\beta}, \hat{\gamma})$ is the MLE based on y and approximated by MCMC methods (Geyer and Møller, 1994; Møller and Waagepetersen, 2003b). We refer to (22) as the fixed Strauss process.

The densities f_1 and f_2 do not depend on the parameters β and γ, and they are both of the type (14). The third choice we consider takes both interaction and parameters into account, but not the data y. Its density is more complicated to present, but it is straightforward to make a simulation in a sequential way: Choose a subdivision C_i, $i = 1, \ldots, m$ of S into, say, square cells C_i of equal size. The simulation is then done in a single sweep, where the cells are visited once in some order. Each visit to a cell involves updating the point configuration within the cell in a way that only depends on the point configuration within the cells already visited.

Specifically, let $I = \{1, \ldots, m\}$ be the index set for the subdivision and for each $i \in I$ let X_i be a point process on C_i. Furthermore, we introduce a permutation $\rho : I \mapsto I$ of I; we shall later let ρ be random but for the moment we condition on ρ. Then, let $X_{\rho(1)}$ be a Poisson point process on $C_{\rho(1)}$ with intensity κ_1 and for $i = 2, \ldots, m$, conditional on $X_{\rho(1)} = x_1, \ldots, X_{\rho(i-1)} = x_{i-1}$, let $X_{\rho(i)}$ be a Poisson point process on $C_{\rho(i)}$ with intensity κ_i, where κ_i may depend on $x_1, \ldots, x_{i-1}$ (which is the case below). Then $X = \cup_{i=1}^m X_i$ is a point process which is an example of a so-called partially ordered Markov model (POMM).

POMMs were introduced by Cressie and Davidson (1998) and Davidson et al. (1999) who applied POMMs in the analysis of grey scaled digital images. POMMs have the attractive properties that their normalising constants are known (and equal one), and that they can model some degree of interaction. Cressie et al. (2000) consider what they call directed Markov point processes (DMPP) as limits of POMMs. Such POMMs and our POMM point process X are of the same type.

When specifying κ_i, $i \in I$ we want to approximate a Strauss point process. To do so we introduce the following concepts and notation. To each cell C_i, $i \in I$ we associate a reference point $\xi_i \in C_i$. Two cells C_i and C_j, $i \neq j$, are said to be neighbour cells if $\|\xi_i - \xi_j\| \leq R_P$, where $R_P > 0$ is the POMM interaction range (to be specified below). Further, for a given point configuration $x \subset S$, let $n_i(x) = n(x \cap C_{\rho(i)})$ denote the number of points in cell $C_{\rho(i)}$, and let $s_{i,R_P,\rho}(x) = \sum_{j \in I: j<i} n_j(x)\mathbf{1}[\|\xi_j - \xi_i\| \leq R_P]$ be the number of points in the cells C_j, $j < i$, which are neigbours to C_i (setting $s_{1,R_P,\rho}(x) = 0$). Note that we have suppressed the dependence on $\{C_i : i \in I\}$ and $\{\xi_i : i \in I\}$ in the notation. Setting $\kappa_i = \beta_P \gamma_P^{s_{i,R_P,\rho}(x)}$ we have that X is a POMM point process with density

$$f_P(x|\beta_P, \gamma_P, R_P, \rho) = \exp\big(-\beta_P \sum_{i \in I} |C_i| \gamma_P^{s_{i,R_P,\rho}(x)}\big) \beta_P^{n(x)} \prod_{i \in I} \gamma_P^{n_i(x) s_{i,R_P,\rho}(x)} \tag{23}$$

with respect to μ.

Cressie et al. (2000) use a Strauss like DMPP which suffers from clear directional effects (incidentally this does not show up in the examples they consider). Since our POMM point process (when ρ is fixed) resembles the POMM point processes used in Cressie et al. (2000), we will now consider ρ as an additional random variable and assume the following in an attempt to reduce any order dependent bias: We assume that ρ is independent of (θ, y) and is uniformly distributed over all permutations of I, and in the auxiliary variable method we use a uniform proposal ρ'. Further, we assume that x given (θ, y, ρ) has density f_3 as specified below. Then the Hastings ratio (11) in the auxiliary variable method is modified by replacing $f(x'|\theta', y)/f(x|\theta, y)$ with $f_3(x'|\theta', \rho', y)/f_3(x|\theta, \rho, y)$ when (θ, x, ρ) is the current state of the chain and (θ', x', ρ') is the proposal; for details, see Appendix B.

It remains to specify f_3 and (β_P, γ_P, R_P) in terms of $\theta = (\beta, \gamma, R)$. Let $(\beta_P, \gamma_P, R_P) = g(\theta) \equiv (g_1(\theta), g_2(\theta), g_3(\theta))$ where $g : (0, \infty) \times (0, 1] \times (0, \infty) \mapsto (0, \infty) \times (0, 1] \times (0, \infty)$. Conditional on (θ, ρ, y), the POMM auxiliary point process has density

$$f_3(x|\theta, \rho, y) = f_P(x|g(\theta), \rho). \tag{24}$$

When specifying g we note that for point configurations x (except for a null set with respect to a homogeneous Poisson process), $\sum_{i \in I} s_{i,R_P,\rho}(x)$ tends to $s_{R_P}(x)$ as $m \to \infty$. This motivates setting $g_3(\theta) = R$ when the cell size is small compared to R. We would like that

$$(g_1(\theta), g_2(\theta)) = \mathbb{E}[\operatorname{argmax}_{(\tilde{\beta}, \tilde{\gamma})} f_P(Y|\tilde{\beta}, \tilde{\gamma}, R, \rho)] \tag{25}$$

where Y is a Strauss process with parameter $\theta = (\beta, \gamma, R)$ and ρ is uniformly distributed and independent of Y. As this expectation is unknown to us, it is approximated as explained in Appendix C. In Table 3, Section 4.3, we refer to (25) as the "MLE". For comparison, we also consider the identity mapping $g(\theta) = \theta$ in Section 4.3, where we in Table 3 refer to this case as the "identity".

Aux.proc.	g	Prop σ_β	Prop σ_γ	MAcP	Extr	c_β	c_γ
Fixed Poisson		2	0.05	0.128	0.151	0.88	0.53
POMM (N=100)	identity	2	0.05	0.171	0.127	0.86	0.54
POMM (N=200)	identity	2	0.05	0.213	0.064	0.85	0.47
POMM (N=50)	MLE	2	0.05	0.246	0.055	0.85	0.46
Fixed Strauss		2	0.05	0.393	0.031	0.79	0.46
POMM (N=100)	MLE	4	0.1	0.298	0.030	0.52	0.21
POMM (N=200)	MLE	4	0.1	0.366	0.014	0.41	0.14
POMM (N=100)	MLE	2	0.05	0.321	0.013	0.79	0.38
POMM (N=200)	MLE	2	0.05	0.406	0.002	0.75	0.33

Table 3: Empirical results: For each auxiliary process considered, one million updates were generated. "Aux.Proc." is the type of auxiliary process used; g is the type of mapping used for each POMM point process (see the end of Section 4.2); "Prop σ_β" and "Prop σ_γ" are the proposal standard deviations for β and γ; "MAcP" is the mean acceptance probability; "Extr" is the fraction of acceptance ratios below $\exp(-10)$; c_β and c_γ are the lag 100 autocorrelation for β and γ.

4.3 Results for the auxiliary variable method

In our simulation study, the data y is given by the perfect simulation in Figure 3, where $S = [0,1]^2$, $\beta = 100$, $\gamma = 0.5$, $R = 0.05$, $n(y) = 75$, and $s_R(y) = 10$. For the MLE, we obtained $\hat{\beta} = 108$ and $\hat{\gamma} = 0.4$. A priori we assume that $R = 0.05$ is known and β and γ are independent and uniformly distributed on $(0, 150]$ and $(0, 1]$, respectively; perfect simulations for $\beta > 150$ can be slow (Berthelsen and Møller, 2002, 2003). For the POMM point process we divide S into $m = N^2$ square cells of side length $1/N$. Below we consider the values $N = 50$, 100, 200, or in comparison with $R = 0.05$, $1/N = 0.02$, 0.01, 0.005. Further details on the auxiliary variable method can be found in Appendix B.

The results are summarised in Table 3 for the different auxiliary processes, and in the POMM case, for different choices of N, the function g in (24), and proposal distributions. Experiments with the algorithm for the fixed Poisson and Strauss and the POMM processes with smaller values of N showed that trace plots of $n(x)$ and $s_R(x)$ (not shown here) may exhibit seemingly satisfactory mixing properties for several million updates and then get stuck — sometimes for more than 100,000 updates. Therefore, as in Table 2, we consider the fraction of acceptance probabilities below $\exp(-10)$ as an indicator for the mixing properties of the chain. Table 3 also shows the mean acceptance probability and the lag 100 autocorrelation of β and γ.

The different cases of auxiliary processes in Table 3 are ordered by the values of "Extr" (the fraction of extremely low acceptance probabilities). Seemingly the results for the autocorrelations depend predominantly on the choice of proposal standard deviations for β and γ. Using the POMM point process with $N = 200$ and $g = \text{MLE}$ appears to give the best mixing. Figure 4 shows the marginal posterior distribution for β and γ when using the POMM process with $N = 200$, $g = \text{MLE}$, and proposal standard deviations for β and γ equal to 2 and 0.05. Apart from Monte Carlo errors,

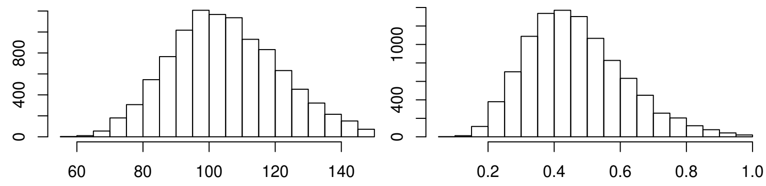

Figure 4: Empirical posterior distribution of β (left plot) and γ (right plot) generated using a POMM auxiliary process with $N = 200$ and $g =$ MLE.

the posterior mode and the MLE are expected to agree: by Figure 4, the MLE $\hat{\gamma} = 0.4$ is not far from the marginal posterior mode whereas $\hat{\beta} = 108$ is around 10% higher.

In conclusion, to obtain a significant improvement by using a POMM auxiliary process with $g =$ MLE compared to using a fixed Strauss process, a cell side length less than about $R/10$ is needed. Computer times show that using the POMM with $N = 100$ are not much slower than using the fixed Strauss process. For $N = 200$ the POMM takes twice as long as for $N = 100$.

5 Concluding remarks

The technique proposed in this paper adds significantly to the ability of statisticians to analyse models that have previously been subject to one or another approximate analysis. By using the auxiliary variable method presented here in conjunction with perfect sampling, we remove the need for stochastic estimation of normalising constant ratios, which require extensive MCMC runs to estimate before the analysis can begin (e.g. Berthelsen and Møller, 2003). We expect therefore that our method will be easier to setup for many problems as, apart from the perfect sampling, it involves only a single Markov chain.

A major issue with this method is the choice of auxiliary variable distribution. Experience shows that this has a major influence on mixing and, if poorly chosen, the chain may not mix at all. Mixing is also affected by the variance of the proposal distributions for the parameters, in the usual way. The only question here is whether the auxiliary variable given data and parameters can be simulated perfectly. If not, the requirement for generating a perfect sample can be relaxed somewhat, though this would introduce an additional undesirable stochasticity into the algorithm, which would place it in the same category as current methods. We have found that if ordinary Gibbs sampling replaces perfect sampling in the example of the Ising model, adequate convergence in each draw of the auxiliary variable is critical. The overall algorithm run considerably more slowly as a result.

We have demonstrated that a workable auxiliary variable distribution has the attribute of closely matching the unnormalised likelihood, while not requiring the computation of a normalising constant. Perhaps the most important consequence of this is that the proposal for the auxiliary variable is then very similar to its full conditional density, which we expect to promote good mixing. The auxiliary variable distribution must be a legitimate probability distribution, ruling out direct use of pseudo likelihoods. An auxiliary variable density based on the unnormalised likelihood evaluated at pseudo likelihood or approximate maximum likelihood parameter estimates, proved to be a suitable choice for both the autologistic model and Strauss point processes. However, it is by no means the only possible auxiliary variable density, as we have demonstrated in the case of Strauss point processes where we applied a partially ordered Markov model as the auxiliary variable density.

Future work should address the potential extensions of our method, particularly the following:

- The construction of useful auxiliary variable distributions which take both the parameter θ and the data y into account (for the examples of applications in this paper, $f(x|\theta, y)$ depends on either θ or y but not both).

- The application of our method to hierarchical models where the unnormalised distribution appears as an unobserved layer (Section 2.2) should be demonstrated. A simple example to consider would be an Ising prior for the archaeological sites studied in Besag et al. (1991), imposing a hyperprior on the parameter of the Ising model. In reference to the hierarchical models in Green and Richardson (2002), Sebastiani and Sørbye (2002), and Dryden et al. (2003), there seems no impediment to extending the technique to the Potts model, which through its connection with the random cluster model, is amenable to perfect sampling.

- The extension of the methods in Sections 2.1 and 2.2 to the more general setting of graphical models, e.g. when the joint density factorizes according to a directed acyclic graph (e.g. Lauritzen, 1996).

- While these extensions seem theoretically feasible, they remain to be practically demonstrated, particularly in regard to adequate mixing.

Appendix A

The auxiliary variable method applied to the Ising model in Section 3.2 can be summarised as follows.

1. Estimate the MPLEs for parameters θ_0 and θ_1, from the data y. Use these as initial values of the Markov chain, as well as retaining for the evaluation of the Metropolis-Hastings ratio in step 4. As initial value of x use a perfect simulation from the Ising model with the MPLE parameter estimates.

2. Draw proposals θ'_0 and θ'_1 from independent normal distributions with means θ_0 and θ_1.

3. Use perfect sampling to draw the proposal x' from $q_{\theta'}(\cdot)$.

4. With probability $\min\{1, H\}$, with H given by (19), set $(\theta, x) = (\theta', x')$.

5. Repeat from step 2.

The standard deviations of the normal distributions in step 2 can be adjusted to give the best mixing of the chain.

Appendix B

We now give details for the auxiliary variable method considered in Sections 4.2 and 4.3.

Consider first the Metropolis-Hastings algorithm for (θ, x) updates using either a fixed Poisson or a fixed Strauss auxiliary variable distribution, see (21) and (22). Recall that $\theta = (\beta, \gamma, R)$ where $R = 0.05$ is fixed. As initial values we choose $\theta = (n(y), 1, 0.05)$ and x is a realisation of a Poisson point process on $S = [0,1]^2$ with intensity $n(y)$. Then, if (θ, x) comprises the current state of the Metropolis-Hastings algorithm with $\theta = (\beta, \gamma, R)$, the next state is generated as follows with f in step 3 replaced by either f_1 (fixed Poisson case) or f_2 (fixed Strauss case).

1. Draw proposals β' and γ' from independent normal distributions with means β and γ.

2. Generate a realisation x' from a Strauss process specified by $\theta' = (\beta', \gamma', R)$ and using dominated CFTP.

3. With probability

$$\min\left\{1, \mathbf{1}[0 < \beta' \leq 150, 0 < \gamma' \leq 1] \left(\frac{\beta'}{\beta}\right)^{n(y)} \left(\frac{\gamma'}{\gamma}\right)^{s_R(y)} \frac{f(x'|y,\theta')}{f(x|y,\theta)} \frac{\beta^{n(x)}\gamma^{s_R(x)}}{\beta'^{n(x')}\gamma'^{s_R(x')}}\right\}$$

   set $\theta = \theta'$ and $x = x'$, otherwise do nothing.

The standard deviations of the normal distributions in step 2 can be adjusted to give the best mixing of the chain.

Consider next using a POMM auxiliary process. Then an extra auxiliary variable, the random permutation ρ, and an additional step is required in the update above. If the current state consists of (β, γ), ρ, and x, then steps 1 and 2 above are followed by

3. Generate a uniform random permutation ρ'.

4. With probability

$$\min\Big\{1, \mathbf{1}[0 < \beta' \leq 150, 0 < \gamma' \leq 1] \times \\ \left(\frac{\beta'}{\beta}\right)^{n(y)} \left(\frac{\gamma'}{\gamma}\right)^{s_R(y)} \frac{f_3(x'|y,\theta',\rho')}{f_3(x|y,\theta,\rho)} \frac{\beta^{n(x)}\gamma^{s_R(x)}}{\beta'^{n(x')}\gamma'^{s_R(x')}}\Big\}$$

   set $(\theta, \rho, x) = (\theta', \rho', x')$, otherwise do nothing.

Here f_3 is given by (24).

Appendix C

When the mapping g in Sections 4.2 and 4.3 is not the identity, it is specified as follows.

Based on the range of the empirical posterior distributions in the fixed Strauss case (not shown here) we define a grid $G = \{50, 52, \ldots, 150\} \times \{0.1, 0.2, \ldots, 1.0\} \times \{0.05\}$. For each grid point $\theta = (\beta, \gamma, R) \in G$, using dominated CFTP, we generate

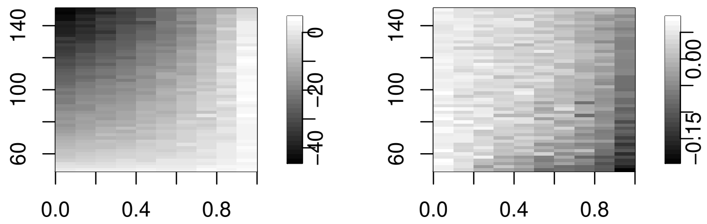

Figure 5: Plot of difference between $g(\theta)$ and θ for $\theta \in G$: $g_1(\beta, \gamma, R) - \beta$ (left) and $g_2(\beta, \gamma, R) - \gamma$ (right).

10 independent realisations $x^{(1)}, \ldots, x^{(10)}$ of a Strauss point process with parameter θ together with the generation of 10 independent random permutations $\rho^{(1)}, \ldots, \rho^{(10)}$. For $\theta \in G$, $g(\theta)$ is given by

$$(g_1(\theta), g_2(\theta)) = \frac{1}{10} \sum_{i=1}^{10} \text{argmax}_{(\tilde{\beta}, \tilde{\gamma})} f_P(x^{(i)} | \tilde{\beta}, \tilde{\gamma}, R, \rho^{(i)}),$$

and $g_3(\theta) = R$. For $(\beta, \gamma, 0.05) \notin G$, we set $g(\beta, \gamma, 0.05) = g(\tilde{\beta}, \tilde{\gamma}, 0.05)$ where $(\tilde{\beta}, \tilde{\gamma}, 0.05) \in G$ is the grid point closest to $(\beta, \gamma, 0.05)$.

Figure 5 shows $g_1(\beta, \gamma, R) - \beta$ and $g_2(\beta, \gamma, R) - \gamma$ for a range of β and γ values when $N = 200$. Results for $N = 50$ and $N = 100$ are almost identical to those for $N = 200$. In cases of strong interaction, i.e. for combinations of low values of γ and high values of β, the parameters $\beta_P = g_1(\beta, \gamma, R)$ and $\gamma_P = g_2(\beta, \gamma, R)$ in the POMM process are much smaller than β and γ in the Strauss process. This is explained by the fact that the interaction in the POMM auxiliary process is weaker than in the Strauss process with the same values of β, γ, and R.

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

Baddeley, A. and Turner, R. (2000), "Practical Maximum Pseudolikelihood for Spatial Point Patterns," *Australian and New Zealand Journal of Statistics*, 42, 283–322.

Bartolucci, F. and Besag, J. (2002), "A Recursive Algorithm for Markov Random Fields," *Biometrika*, 89, 724–730.

Berthelsen, K. K. and Møller, J. (2002), “A Primer on Perfect Simulation for Spatial Point Processes,” *Bulletin of the Brazilian Mathematical Society*, 33, 351–367.

— (2003), “Likelihood and Non-parametric Bayesian MCMC Inference for Spatial Point Processes Based on Perfect Simulation and Path Sampling,” *Scandinavian Journal of Statistics*, 30, 549–564.

Besag, J. (1975), “Statistical Analysis of Non-lattice Data,” *The Statistician*, 25, 179–195.

— (1977), “Some Methods of Statistical Analysis for Spatial Data,” *Bulletin of the International Statistical Institute*, 47, 77–92.

Besag, J. and Green, P. J. (1993), “Spatial Statistics and Bayesian Computation (with discussion),” *Journal of the Royal Statistical Society, Series B*, 16, 395–407.

Besag, J., York, J., and Mollié, A. (1991), “Bayesian Image Restoration, with two Applications in Spatial Statistics,” *Annals of the Institute of Statistical Mathematics*, 43, 1–59.

Besag, J. E. (1972), “Nearest-neighbour Systems and the Auto-logistic Model for Binary Data,” *Journal of the Royal Statistical Society, Series B*, 34, 75–83.

— (1974), “Spatial Interaction and the Statistical Analysis of Lattice Systems (with discussion),” *Journal of the Royal Statistical Society, Series B*, 36, 192–236.

Best, N. G., Cowles, M. K., and Vines, S. K. (1995), “CODA Convergence Diagnosis and Output Analysis Software for Gibbs Sampler Output: Version 0.3,” Tech. rep., Cambridge: MRC Biostatistics Unit.

Breyer, L. A. and Roberts, G. O. (2000), “From Metropolis to Diffusions: Gibbs States and Optimal Scaling,” *Stochastic Processes and their Applications*, 90, 181–206.

Cressie, N. and Davidson, J. L. (1998), “Image Analysis with Partially Ordered Markov Models,” *Computational Statistics and Data Analysis*, 29, 1–26.

Cressie, N., Zhu, J., Baddeley, A. J., and Nair, M. G. (2000), “Directed Markov Point Processes as Limits of Partially Ordered Markov Models,” *Methodology and Computing in Applied Probability*, 2, 5–21.

Cressie, N. A. C. (1993), *Statistics for Spatial Data*, Wiley, New York, 2nd ed.

Davidson, J. L., Cressie, N., and Hua, X. (1999), “Texture Synthesis and Pattern Recognition for Partially Ordered Markov Models,” *Pattern Recognition*, 32, 1475–1505.

Dryden, I. L., Scarr, M. R., and Taylor, C. C. (2003), “Bayesian Texture Segmentation of Weed and Crop Images Using Reversible Jump Markov Chain Monte Carlo Methods,” *Applied Statistics*, 52, 31–50.

Edwards, R. G. and Sokal, A. D. (1988), "Generalisation of the Fortuin-Kastelyn-Swendsen-Wang Representation and Monte Carlo Algorithms," *Physical Review Letters*, 38, 2009–2012.

Fill, J. A. (1998), "An Interruptible Algorithm for Perfect Sampling via Markov Chains," *Annals of Applied Probability*, 8, 131–162.

Friel, N. and Pettitt, A. N. (2003), "Likelihood Estimation and Inference for the Autologistic Model," *Journal of Computational and Graphical Statistics*, to appear.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995), *Bayesian Data Analysis*, Boca Raton: Chapman and Hall/CRC.

Gelman, A. and Meng, X.-L. (1998), "Simulating Normalizing Contants: From Importance Sampling to Bridge Sampling to Path Sampling," *Statistical Science*, 13, 163–185.

Gelman, A., Roberts, G. O., and Gilks, W. (1996), "Efficient Metropolis Jumping Rules," in *Bayesian Statistics V*, ed. Bernardo, J., Oxford University Press, pp. 599–608.

Geyer, C. J. (1991), "Markov Chain Monte Carlo Maximum Likelihood," in *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, pp. 156–163.

— (1999), "Likelihood Inference for Spatial Point Processes," in *Stochastic Geometry: Likelihood and Computation*, eds. Barndorff-Nielsen, O. E., Kendall, W. S., and van Lieshout, M. N. M., Boca Raton, Florida: Chapman & Hall/CRC, pp. 79–140.

Geyer, C. J. and Møller, J. (1994), "Simulation Procedures and Likelihood Inference for Spatial Point Processes," *Scandinavian Journal of Statistics*, 21, 359–373.

Geyer, C. J. and Thompson, E. A. (1992), "Constrained Monte Carlo Maximum Likelihood for Dependent Data (with discussion)," *Journal of the Royal Statistical Society, Series B*, 54, 657–699.

Green, P. J. and Richardson, S. (2002), "Hidden Markov Models and Disease Mapping," *Journal of the American Statistical Association*, 97, 1055–1070, Theory and Methods.

Gu, M. G. and Zhu, H.-T. (2001), "Maximum Likelihood Estimation for Spatial Models by Markov Chain Monte Carlo Stochastic Approximation," *Journal of the Royal Statistical Society, Series B*, 63, 339–355.

Häggström, O. and Nelander, K. (1998), "Exact Sampling from Anti-monotone Systems," *Statistica Neerlandia*, 52, 360–380.

Heikkinen, J. and Högmander, H. (1994), "Fully Bayesian Approach to Image Restoration with an Application in Biogeography," *Applied Statistics*, 43, 569–582.

Heikkinen, J. and Penttinen, A. (1999), "Bayesian Smoothing in the Estimation of the Pair Potential Function of Gibbs Point Processes," *Bernoulli*, 5, 1119–1136.

Higdon, D. M. (1998), "Auxiliary Variable Methods for Markov Chain Monte Carlo with Applications," *Journal of the American Statistical Association*, 93, 585–595.

Jensen, J. L. and Møller, J. (1991), "Pseudolikelihood for Exponential Family Models of Spatial Point Processes," *Annals of Applied Probability*, 3, 445–461.

Kelly, F. P. and Ripley, B. D. (1976), "A Note on Strauss's Model for Clustering," *Biometrika*, 63, 357–360.

Kendall, W. S. and Møller, J. (2000), "Perfect Simulation Using Dominating Processes on Ordered Spaces, with Application to Locally Stable Point Processes," *Advances in Applied Probability*, 32, 844–865.

Lauritzen, S. L. (1996), *Graphical models*, Oxford: Clarendon Press.

Møller, J. (1999), "Perfect Simulation of Conditionally Specified Models," *Journal of the Royal Statistical Society Series B*, 61, 251–264.

Møller, J. and Schladitz, K. (1999), "Extensions of Fill's Algorithm for Perfect Simulation," *Journal of the Royal Statistical Society Series B*, 61, 955–969.

Møller, J. and Waagepetersen, R. P. (2003a), "An Introduction to Simulation-based Inference for Spatial Point Processes," in *Spatial Statistics and Computational Methods*, ed. Møller, J., Springer-Verlag, New York, Lecture Notes in Statistics 173, pp. 143–198.

— (2003b), *Statistical Inference and Simulation for Spatial Point Processes*, Chapman and Hall/CRC.

Pettitt, A. N., Friel, N., and Reeves, R. (2003), "Efficient Calculation of the Normalising Constant of the Autologistic and Related Models on the Cylinder and Lattice," *Journal of the Royal Statistical Society, Series B*, 65, 235–246.

Pitt, M. K. and Shephard, N. (1999), "Filtering via Simulation: Auxiliary Particle Filters," *Journal of the American Statistical Association*, 94, 590–599.

Possolo, A. (1986), "Estimation of Binary Markov Random Fields," Tech. Rep. 77, Department of Statistics, University of Washington.

Propp, J. and Wilson, D. (1996), "Exact Sampling with Coupled Markov Chains and Applications to Statistical Mechanics," *Random Structures and Algorithms*, 9, 223–252.

Propp, J. G. and Wilson, D. B. (1998), "How to get a Perfectly Random Sample from a Generic Markov Chain and Generate a Random Spanning Tree to a Directed Graph," *Journal of Algorithms*, 27, 170–217.

Reeves, R. and Pettitt, A. N. (2003), "Efficient Recursions for General Factorisable Models," manuscript submitted for publication.

Ripley, B. D. and Kelly, F. P. (1977), "Markov Point Processes," *Journal of the London Mathematical Society*, 15, 188–192.

Roberts, G. O., Gelman, A., and Gilks, W. (1997), "Weak Convergence and Optimal Scaling of Random Walk," *Annals of Applied Probability*, 7, 110–120.

Roberts, O. O. and Rosenthal, J. S. (1998), "Optimal Scaling of Discrete Approximations to Langevin Diffusions," *Journal of the Royal Statistiscal Society Series B*, 60, 255–268.

Rubin, D. B. (1987), "Comment on "The Calculation of Posterior Distributions by Data Augmentation", by M.A. Tanner and W.H. Wong," *Journal of the American Statistical Association*, 82, 543–546.

Sebastiani, G. and Sørbye, S. H. (2002), "A Bayesian Method for Multispectral Image Data Classification," *Nonparametric Statistics*, 14, 169–180.

Strauss, D. J. (1975), "A Model for Clustering," *Biometrika*, 62, 467–475.

Strauss, D. J. and Ikeda, M. (1990), "Pseudolikelihood Estimation of Social Networks," *Journal of the American Statistical Association*, 85, 204–212.

Swendsen, R. H. and Wang, J.-S. (1987), "Nonuniversal Critical Dynamics in Monte Carlo Simulations," *Physical Review Letters*, 58, 86–88.

Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions," *Annals of Statistics*, 22, 1701–1728.