A Bayesian MCMC method for point process models with intractable normalising constants*

Kasper K. Berthelsen and Jesper Møller
Aalborg University, Department of Mathematical Sciences,
F. Bajers Vej 7G, DK-9220 Aalborg, Denmark.

Abstract

We present new methodology for drawing samples from a posterior distribution when the likelihood function is only specified up to a normalising constant. 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", perfect simulation algorithms become useful. We illustrate the method in cases where the likelihood is given by a Markov point process model. Particularly, we consider semi-parametric Bayesian inference in connection to both inhomogeneous Markov point process models and pairwise interaction point processes.

Keywords: Inhomogeneous Markov point processes; Markov chain Monte Carlo; Pairwise interaction point processes; Perfect simulation; Semi-parametric Bayesian inference; Strauss point process.

1 Introduction

Unnormalised probability distributions with intractable normalising constants arise in a number of statistical problems, including the definition of Markov point processes [11, 14, 16]. 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. Indeed this is the case

* Work partly based on [2, 3, 13]. 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.

when we consider a likelihood function for a parametric family of Markov point process models.

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, see e.g. [17]. By (2),

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

is unknown, since it depends on the ratio of unknown normalising constants $Z_{\theta'}/Z_{\theta}$.

Because of their intractability, earlier work attempted to avoid algorithms involving unknown normalising constants. An example in connection to spatial point processes is Heikkinen and Penttinen [8], who instead of estimating the entire posterior distribution, 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. Recently, Berthelsen and Møller [2] performed a more detailed Bayesian analysis, using path sampling [7] or, as it is known in statistical physics, thermodynamic integration, for estimating the ratio of normalising constants.

The approach introduced in Section 2 avoids approximations such as those discussed 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. Access to algorithms for making perfect (or exact) simulations [1, 9, 10, 14] from (2) becomes an important ingredient as explained later. Application to the Strauss process is discussed in Section 3. Particularly, we discuss semi-parametric Bayesian inference for inhomogeneous Markov point processes in Section 4 and for pairwise interaction points processes in Section 5.

2 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),

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

still involves the unknown Z_θ.

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

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

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

$$p(\theta', x'|\theta, x) = p(x'|\theta', \theta, x)p(\theta'|\theta, x) \tag{4}$$

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,

$$p(x'|\theta', \theta, x) = p(x'|\theta') = q_{\theta'}(x')/Z_{\theta'}. \tag{5}$$

Then

$$H(\theta', x'|\theta, x) = \frac{f(x'|\theta', y)\pi(\theta')q_{\theta'}(y)q_\theta(x)p(\theta|\theta', x')}{f(x|\theta, y)\pi(\theta)q_\theta(y)q_{\theta'}(x')p(\theta'|\theta, x)}. \tag{6}$$

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 (3) 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 (6) presenting no difficulty in evaluation.

Henceforth, for simplicity, we assume that

$$p(\theta'|\theta, x) = p(\theta'|\theta) \tag{7}$$

does not depend on x. For simulation from the proposal density (4) 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 (5) appropriate perfect simulation algorithms [1, 9, 10, 14] are used 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 (3) has good mixing properties. If we let $f(x|\theta, y) = q_\theta(x)/Z_\theta$, then by (7), (6) reduces to (3), and so the mixing and convergence properties of the two Metropolis-Hastings algorithms using (3) and (6) are the same. Recommendations on how to tune Metropolis-Hastings algorithms to obtain optimal acceptance probabilities may exist in the case of (3). This suggests that the auxiliary distribution should approximate the distribution given by q_θ,

$$f(x|\theta, y) \approx q_\theta(x)/Z_\theta. \tag{8}$$

It is interesting to notice that if equality holds in (8), then the states from the chain for the auxiliary variable x can be interpreted as posterior predictions. Choices where (8) are satisfied will be discussed in the following.

3 Application to the Strauss process

Consider the Strauss point process defined on a bounded region $W \subset \mathbb{R}^2$ by a density

$$\pi(y|\theta) = \frac{1}{Z_\theta} \beta^{n(y)} \gamma^{s_R(y)} \tag{9}$$

with respect to a unit rate Poisson point process on W. Here $\theta = (\beta, \gamma, R)$, with $\beta > 0$, $0 < \gamma \leq 1$, and $R > 0$; $n(y)$ 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. The normalising constant Z_θ is unknown when $\gamma < 1$, cf. [14] and the references therein.

When specifying the auxiliary variable a general choice is

$$f(x|\theta, y) = q_{\tilde\theta}(x)/Z_{\tilde\theta} \tag{10}$$

where $\tilde\theta = \tilde\theta(y)$ is an estimate, for example an (approximate) MLE or a maximum pseudo likelihood estimate, see e.g. [14]. Note that the unknown normalising constant $Z_{\tilde\theta}$ cancels in (6). The choice (10) is expected to work well if the posterior distribution of θ is concentrated near $\tilde\theta$. In [13] we consider two cases of (10) with R fixed. One is a homogeneous Poisson process (i.e. $\gamma = 1$ in (9)) where $\tilde\beta$ is the MLE. The other is a Strauss process, where $\tilde\theta$ is an approximate MLE obtained by MCMC methods.

In [13] we also consider another type of auxiliary point process, which in contrast to (10) takes the parameter θ 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

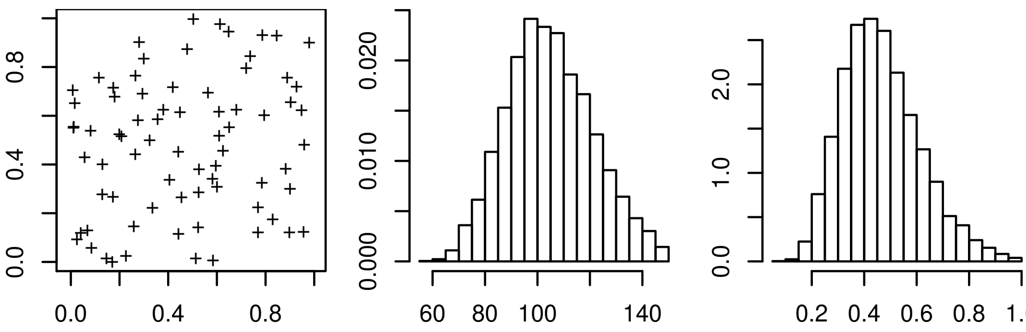

Figure 1: Results from [13] concerning a Bayesian analysis when "data" is given by a perfect simulation of a Strauss point process with $\beta = 100$, $\gamma = 0.5$ and $R = 0.05$ (left plot). A priori we assume that $R = 0.05$ is fixed, and β and γ are independent and uniformly distributed on $[0, 150]$ and $[0, 1]$, respectively. Plots show the resulting empirical posterior densities of β (middle plot) and γ (right plot) obtained by the auxiliary variable technique using a POMM point process as the auxiliary variable.

way: Choose a subdivision $C_1, \ldots, C_m$ of W. First generate uniformly at random a permutation $(\sigma_1, \ldots, \sigma_m)$ of $(1, \ldots, m)$. Next generate a Poisson point process X_1 on C_{σ_1} with intensity κ_{σ_1} (as defined in [13]). Finally, for each $i = 2, \ldots, m$, conditional on $X_1, \ldots, X_{i-1}$, generate a Poisson point process X_i on C_{σ_i} with intensity κ_{σ_i}, where κ_{σ_i} depends on $X_1, \ldots, X_{i-1}$ (as described in [13]). 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) [4, 5]. Note that the random permutation is introduced to avoid any order dependent bias.

The three choices of auxiliary variable considered are compared in a simulation study in [13]. The conclusion is that the best choice is a POMM point process where the "width" of subsets C_i are no larger than one tenth of the interaction range. Figure 1 illustrates the kind of results provided by a Bayesian analysis as in [13].

4 Application to semi-parametric inhomogeneous Markov point processes

The left plot of Figure 2 shows an inhomogeneous point pattern. Assume that such a point pattern can be modelled by an inhomogeneous pairwise interaction point

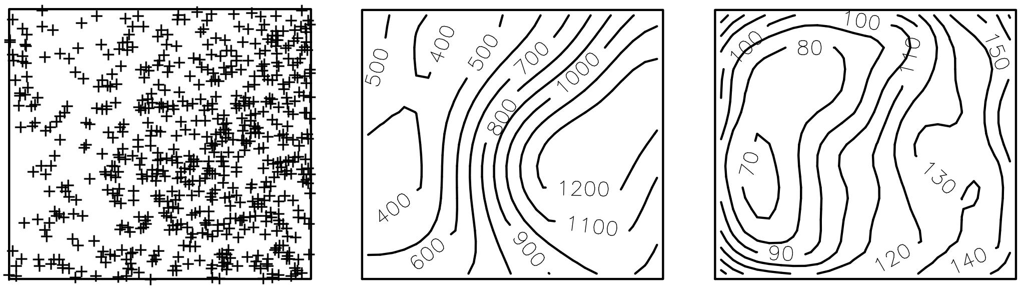

Figure 2: Left plot: locations of 617 cells in a 2D section of the mocous membrane of the stomach of a healthy rat. Middle and right plot: posterior mean and standard deviation for the first order term $\beta(\cdot)$ in (11).

process on a bounded region $W \subset \mathbb{R}^2$ with normalised density

$$\pi(x|\beta,\varphi) = \frac{1}{Z_{(\beta,\varphi)}} \prod_{\xi \in x} \beta(\xi) \prod_{\{\xi,\eta\} \subseteq x : \xi \neq \eta} \varphi(\|\xi - \eta\|) \tag{11}$$

w.r.t. a unit rate Poisson process on W. Here $Z_{(\beta,\varphi)}$ is the normalising constant, $\beta : \mathbb{R}^2 \mapsto [0,\infty)$ is describing the inhomogeneity, and $\varphi : [0,\infty) \mapsto [0,1]$ is an interaction function.

For specificity we choose

$$\varphi(r) = \begin{cases} 0 & \text{if } r \leq R_0 \\ 1 - \left(1 - \left(\frac{r-R_0}{R-R_0}\right)^m\right)^m & \text{if } R_0 < r \leq R \\ 1 & \text{if } r > R \end{cases}$$

where the parameter $R \geq 0$ specifies the interaction range for φ, $R_0 \in [0,R)$ is a hard core parameter, and $m > 0$ controls the "smoothness" of $\varphi(r)$ in the interval $R_0 \leq r \leq R$.

A priori we assume that the intensity function β can be approximated by

$$\beta(\xi;\psi) = \sum_j \lambda_j \mathcal{K}(\xi; c_j)$$

where

- $\mathcal{C} = \{c_1, c_2, \ldots, c_n\}$ is a homogeneous Poisson process on a bounded region $W_{\text{ext}} \supseteq W$ with intensity κ;

- conditional on $\mathcal{C}$, the λ_j are i.i.d. gamma distributed with scale parameter $\alpha_1 > 0$ and shape parameter $\alpha_2 > 0$;
- $\mathcal{K}(\cdot; c)$ is the density of a bivariate normal distribution with mean c and covariance matrix $\sigma^2 I$, $\sigma > 0$.

The specific choice of W_{ext} is guided by results in [12]. See [3] for more details.

Ideally we want to specify non-informative priors on all model parameters. In the analysis in [3] we have for computational convenience used a fixed value of $(\kappa, \alpha_1, \alpha_2)$ and specified independent uniform priors on R_0, R and m on bounded intervals. It is then straightforward to generate posterior simulations by the auxiliary variable technique, using a POMM point process (similar to the one considered in Section 3) as the auxiliary variable. The middle and right plots in Figure 2 show some results from the Bayesian analysis in [3].

5 Application to semi-parametric pairwise interaction point process models

Berthelsen and Møller [2] consider a semi-parametric Bayesian analysis for the pairwise interaction point process (11) when $\beta(\xi) = \beta$ is constant and φ is approximated by a step function,

$$\varphi(r) = \sum_{i=1}^{p} \gamma_i \mathbf{1}[r_{i-1} < r \leq r_i] + \mathbf{1}[r > r_p]$$

where $r_0 = 0 < r_1 < \ldots < r_p$, $0 < \gamma_1 < \ldots < \gamma_p < 1$ and $p \in \{0, 1, 2, \ldots\}$ is the number of change-points. This is an extension of (9) called the multiscale process [15]. Imposing a prior for (β, ψ), where $\psi = \{(r_1, \gamma_1), \ldots, (r_p, \gamma_p)\}$ is viewed as a marked point process, allowing the number of change points $r_1, \ldots, r_p$ to vary, we perform in [2] a fully Bayesian MCMC analysis where ratios of unknown normalising constants are approximated by path sampling [7]. Some results of the analysis in [2] are shown in Figure 3.

Alternatively the full Bayesian analysis can be carried out using the auxiliary variable technique. In this setup, using a POMM point process seems appropriate as the auxiliary variable, and again perfect simulation is needed for generating proposals. We plan to include in [3] an analysis of this kind.

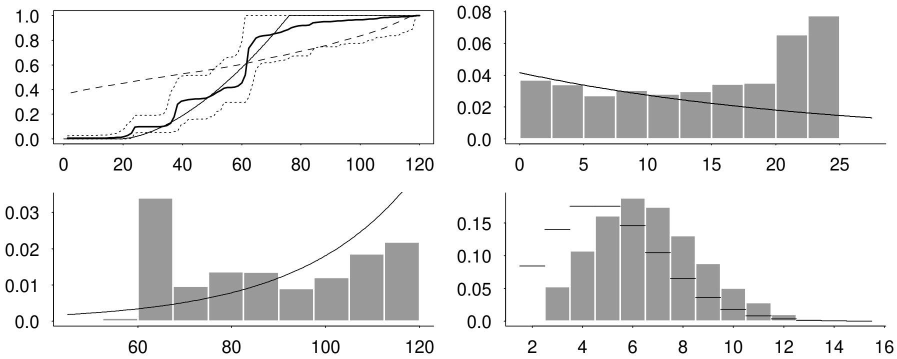

Figure 3: Top left: posterior mean (thick solid line) together with a 95% posterior interval (dotted lines), prior mean (dashed line), and an ad hoc parametric estimate from [6] (thin solid line). Top right, bottom left and bottom right: posterior and prior density (solid line) of r_1, r_p, and p. See [2] for more details.

References

[1] K. K. Berthelsen and J. Møller. A primer on perfect simulation for spatial point processes. *Bulletin of the Brazilian Mathematical Society*, 33: 351–367, 2002.

[2] K. K. Berthelsen and J. Møller. 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, 2003.

[3] K. K. Berthelsen and J. Møller. Semi-parametric Bayesian inference for inhomogeneous Markov point processes. In preparation.

[4] N. Cressie and J. L. Davidson. Image analysis with partially ordered Markov models. *Computational Statistics and Data Analysis*, 29: 1–26, 1998.

[5] N. Cressie, J. Zhu, A. J. Baddeley and M. G. Nair. Directed Markov point processes as limits of partially ordered Markov models. *Methodology and Computing in Applied Probability*, 2: 5–21, 2000.

[6] P. Diggle and R. Gratton. Monte Carlo methods of inference for implicit statistical models (with discussion). *Journal of the Royal Statistical Society Series B*, 46: 193–227, 1984.

[7] A. Gelman and X.-L. Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, 13: 163–185, 1998.

[8] J. Heikkinen and A. Penttinen. Bayesian smoothing in the estimation of the pair potential function of Gibbs point processes. *Bernoulli*, 5: 1119–1136, 1999.

[9] O. Häggström, M. N. M. van Lieshout and J. Møller. Characterization results and Markov chain Monte Carlo algorithms including exact simulation for some spatial point processes. *Bernoulli*, 5: 641–659, 1999.

[10] W. S. Kendall and J. Møller. Perfect simulation using dominating processes on ordered spaces, with application to locally stable point processes. *Advances in Applied Probability*, 32: 844–865, 2000.

[11] M. N. M. van Lieshout. *Markov Point Processes and Their Applications*. Imperial College Press, London, 2000.

[12] J. Møller. Shot noise Cox processes. *Advances in Applied Probability*, 35: 614–640, 2003.

[13] J. Møller, A. N. Pettitt, K. K. Berthelsen and R. W. Reeves. An efficient MCMC method for distributions with intractable normalising constants. Research Report R-2004-02, Department of Mathematical Sciences, Aalborg University, 2004.

[14] J. Møller and R. P. Waagepetersen. *Statistical Inference and Simulation for Spatial Point Processes*. Chapman and Hall/CRC, Boca Raton, 2003.

[15] A. Penttinen. *Modelling Interaction in Spatial Point Patterns: Parameter Estimation by the Maximum Likelihood Method*. Number 7 in Jyväskylä Studies in Computer Science, Economics, and Statistics, 1984.

[16] B. D. Ripley and F. P. Kelly. Markov point processes. *Journal of the London Mathematical Society*, 15: 188–192, 1977.

[17] L. Tierney. Markov chains for exploring posterior distributions. *Annals of Statistics*, 22: 1701–1728, 1994.