The Linked Importance Sampler Auxiliary Variable Metropolis Hastings Algorithm for Distributions with Intractable Normalising Constants

The Linked Importance Sampler Auxiliary Variable Metropolis Hastings Algorithm for Distributions with Intractable Normalising Constants*

Johan H. Koskinen †

MelNet Social Networks Laboratory Technical Report 08–01

Department of Psychology, School of Behavioural Science

University of Melbourne, Parkville Victoria 3010, Australia

April 11, 2008 (Revised November 28)

*This work was supported by the Australian Defence Science and Technology Organisation (DSTO). The author is grateful to Philippa Pattison and Garry Robins for helpful comments and suggestions and to Dean Lusher for the use of the Australian school data.

†E-mail: johank@unimelb.edu.au.

Abstract

We consider parameter inference for the class of models where the likelihood function is analytically intractable as a result of a complicated normalising constant. This means that an MCMC algorithm for drawing from the posterior of the parameters of the model would involve evaluating an acceptance ratio containing a ratio of unknown normalising constants. We propose to improve on the recently proposed auxiliary variable MCMC by extending the variable space of the auxiliary variable. Whereas the auxiliary variable MCMC may be construed as a Metropolis-Hastings algorithm that estimates the acceptance ratio in each iteration using simple importance sample based on only one observation, we show how the algorithm proposed here can be seen as substituting the one-observation simple importance sample with the more efficient linked importance sampler (LIS). While retaining the properties of the Metropolis-Hastings algorithm the use of LIS is generally applicable and allows flexibility in tuning the mixing. In particular we show that the algorithm can be made to work for some social network models for which Bayesian analysis has not previously been feasible. While the auxiliary variable MCMC works for an Ising model it does not work for other models, for which we show that we can achieve reasonable mixing by using the tuning features of the proposed algorithm. The models mentioned include a social influence model for correlated binary outcomes for pupils in an Australian school class where sociometric data is available, and, (curved) exponential family models for the collaboration network of partners in a New England law firm.

Keywords: Linked importance sampler; auxiliary variable; Normalising constant; Ising model; Social network; Influence model; Exponential family random graph model (p-star)

1. Introduction

Markov chain Monte Carlo (MCMC) methods have proved a popular tool for performing Bayesian inference since they only require that the posterior distribution

is know up to a normalizing constant. There is however a large class of important statistical models for which MCMC is not a viable option since the likelihood function is not analytically tractable as a result of a complicated normalising constant in the model density. We deal here with models for a variable x defined on a finite space $\mathcal{X}$ with a probability mass function (pmf) of the form

$$p(x|\theta) = \frac{1}{c(\theta)} q(x;\theta), \tag{1}$$

where the normalising constant

$$c(\theta) = \sum_{y \in \mathcal{X}} q(y;\theta),$$

is a function of the parameter vector $\theta \in \Theta \subseteq R^p$ that assures that the pmf sums to unity. A basic premiss here is that q is easy to evaluate but that $c(\theta)$ is computationally hard to evaluate. In principle this makes it straightforward to simulate from the model using e.g. the Metropolis-Hastings algorithm (Hastings, 1970; Tierny, 1994; Chib and Greenberg, 1995). That $c(\theta)$ is hard to evaluate is however a major obstacle to Bayesian inference since although the normalising constant $\mathrm{m}(x)$ of the posterior distribution

$$\pi(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{\mathrm{m}(x)} \propto \frac{1}{c(\theta)} q(x;\theta)\pi(\theta), \tag{2}$$

is only a function of data x, $c(\theta)$ is a function of the parameter vector. As a consequence, in a Metropolis-Hastings sampler for drawing samples from $\pi(\theta|x)$, where a move from θ to θ^* is proposed from a distribution $g(\theta^*|\theta)$ (henceforth, g denotes a generic proposal density), the Hastings ratio in the acceptance probability $\min(1, H)$ would be

$$H = \frac{\pi(\theta^*|x)}{\pi(\theta|x)} \frac{g(\theta|\theta^*)}{g(\theta^*|\theta)} = \frac{q(x;\theta^*)/c(\theta^*)\pi(\theta^*)}{q(x;\theta)/c(\theta)\pi(\theta)} \frac{g(\theta|\theta^*)}{g(\theta^*|\theta)}. \tag{3}$$

While the marginal likelihood, $\mathrm{m}(x)$, cancels in the Hastings ratio we are left with a ratio

$$\lambda(\theta^*, \theta) = \frac{c(\theta)}{c(\theta^*)}$$

of unknown normalising constants.

Models of the form (1) include those where the variable x represents the spin configuration of particles on a binary $m \times n$ lattice (Besag, 1972); exponential family random graph (ERGM) distributions for the adjacency matrix describing the social interaction of individuals (Frank and Strauss, 1986; Snijders, Pattison, Robins, and Handcock, 2006; Hunter and Handcock, 2006); or activation indicators of voxels in functional magnetic resonance imaging data (Smith and Fahrmeir, 2007). The methods presented here do not rely on the assumption of $\mathcal{X}$ being finite but we limit our treatment here to the more transparent, finite case.

Given that non-Bayesian estimation until recently largely relied on maximisation of the pseudo likelihood (Besag, 1974, 1975; Strauss and Ikeda, 1990) rather than the likelihood function it is perhaps a natural approach to perform Bayesian inference using the pseudo likelihood rather than the true likelihood function as was done in Heikkinen and Högmander (1994). This transforms the problem into a regular inference issue and standard MCMC methods may be used but the distribution from which one is sampling is not known. Another way of avoiding having to evaluate the normalising constant is by limiting the analysis to finding a point estimate (Heikkinen and Penttinen, 1999).

Since there are numerous efficient algorithms for numerically calculating (approximating) the normalising constant (Gelman and Meng, 1998), MCMC schemes have been proposed for models with intractable normalising constants where a MCMC approximation to the normalising constant in the likelihood is substituted for the exact value. Normalising constants can be evaluated on a grid of parameter values and stored (Green and Richardson, 2002) or estimated repeatedly in the course of the MCMC (Berthelsen and Møller, 2003), using a sample from an importance distribu-

tion that is stored off-line or regenerated on-line. Calculating approximations of the partition function may be considerably harder when the parameter space is of higher dimensions.

Common to the previously employed Bayesian inference schemes is that it is hard to establish the properties of an MCMC procedure that relies on approximations to distributions relative to that of the exact expression. The properties of MCMC may not always carry over. The estimators of the normalising constant that are currently available are mostly constructed to estimate individual constants (or ratios of constants) and are not necessarily suited to repeated estimation. From the ergodic theorem (Tierney, 1994), we know that the estimate gets close to its true value as the number of iterations gets large, something we might not be able to allow for were we to take an estimate in each iteration of the MCMC, and the required number of iterations is likely to vary for different parameter values. Furthermore, while the estimate of the normalising constant is simulation consistent, the estimate of the acceptance ratio is not.

Møller, Pettitt, Berthelsen, and Reeves (2006) proposed the first truly "exact" or "pure" MCMC algorithm, namely, the auxiliary variable method (AVM) MCMC algorithm, for performing Bayesian inference for models with intractable normalising constants. The idea is to simulate jointly from the posterior of θ and an auxiliary variable density. Mixing is affected both by the posterior and the proposal distribution for θ as well as the choice of auxiliary distribution. For the simple form of AVM, the degree to which the mixing can be improved is limited to the choice of auxiliary distribution. Berthelsen and Møller, (2004a,b, 2006, 2007) give examples of ways of adopting the auxiliary distribution for a range of different models to improve mixing. The dependencies between variables for some models used in social science cause the AVM to mix so poorly as to prevent analysis. In particular the Bayesian analysis of ERGMs suffers the consequence of complex interdependencies to the same extent that non-Bayesian analysis does (Corander, Dahmström, and Dahmström, 1998, 2002; Snijders, 2002; Handcock, 2002; van Duijn, Gile, and Handcock, 2008). Building on

the advances of Møller et al. (2006) we propose a tunable MCMC sampler, the linked importance sampler auxiliary variable (LISA) Metropolis Hastings. This extension rests upon the observation that the poor mixing of the AVM that we see in the case of a couple of important models in social network analysis can be understood if AVM is formulated as a regular MCMC with an embedded importance sampler that estimates the normalising constant in each step using only one sample point. We suggest that this makes it natural to replace the one-sample simple importance sample by more elaborate variations of the importance sampler that have proved more efficient. As noted above, any sampler would not do. It turns out however, that a specific importance sampler, the linked importance sampler (LIS) (Neal, 2005), can be incorporated into the MCMC as an auxiliary variable when the space on which the importance distribution is defined is considered to be an extended sample space of the MCMC. This extended sample space is discrete but can be of high cardinality. However, we need not consider the variable defined on the extended state space explicitly in the sense that we need to save memory-consuming variables - the part of the LISA that concerns the auxiliary variable reduces to taking an importance sample. Anonymous reviewers alerted our attentions to a similar approach to modifying the auxiliary distribution by Murray, Ghahramani, and MacKay (2006).

The rest of the paper is structured as follows. The first part reviews known results that we feel are important for motivating and understanding the proposed approach. First we shall give a brief review of the auxiliary variable method of Møller et al. (2006) and its formulation as a Metropolis-Hastings algorithm with an embedded importance sampler. We review the principles of some standard importance samplers and show how the LIS is an attempt at combining the advantages of different importance samplers. This is followed by the main result where we show how LISA is constructed from combining LIS with the AVM. We then proceed to discuss the performance of the proposed algorithm in the context of some illustrative examples, namely using simulated data from an Ising model, fitting an influence model to data for a school class, and fitting a curved exponential family graph model to a well

known social networks data set. The last two models prove the main benefit of and motivation for the proposed approach as compared to AVM, as the AVM does not work for them.

2. The auxiliary variable method

To circumvent the need to evaluate $\lambda(\theta^*, \theta)$ in (3) while retaining the properties of the MCMC scheme, Møller et al. (2006) proposed to introduce an auxiliary variable y, defined on the same state space $\mathcal{X}$ as x, and to set up the MH to produce a sample $(\theta^{(k)}, y^{(k)})_{k=0}^{N}$ from the joint posterior distribution of θ and y. By letting y have the pmf $q(y;\psi)/c(\psi)$ for ψ fixed, the Hastings ratio for the joint acceptance of (θ^*, y^*) becomes

$$H = \frac{q(x;\theta^*)/c(\theta^*)}{q(x;\theta)/c(\theta)} \frac{q(y^*;\psi)/c(\psi)}{q(y;\psi)/c(\psi)} \frac{g(\theta, y|\theta^*, y^*)}{g(\theta^*, y^*|\theta, y)} \frac{\pi(\theta^*)}{\pi(\theta)}.$$

While we see that the normalising constant $c(\psi)$ in the pmf of y cancel, the problem of evaluating $\lambda(\theta^*, \theta)$ still remains. The trick employed in Møller et al. (2006) was to firstly factorise the proposal density $g(\theta^*, y^*|\theta, y) = g(y^*|\theta^*)g(\theta^*|\theta)$ so that y^* is drawn conditional on the proposed new value θ^*. Secondly, the conditional proposal distribution for y is set to $q(y;\theta^*)/c(\theta^*)$ (for some models they employed more elaborate proposal distributions). Doing this, and inserting the proposal distributions in the Hastings ratio we get

$$H = \frac{q(x;\theta^*)/c(\theta^*)}{q(x;\theta)/c(\theta)} \frac{q(y^*;\psi)}{q(y;\psi)} \frac{q(y;\theta)/c(\theta)}{q(y^*;\theta^*)/c(\theta^*)} \frac{g(\theta|\theta^*)}{g(\theta^*|\theta)} \frac{\pi(\theta^*)}{\pi(\theta)},$$

where we see that the normalising constants $c(\theta^*)$ and $c(\theta)$ of the proposal distributions cancel the respective normalising constants in the target distributions

$$H = \frac{q(x;\theta^*)}{q(x;\theta)} \frac{q(y^*;\psi)}{q(y;\psi)} \frac{q(y;\theta)}{q(y^*;\theta^*)} \frac{g(\theta|\theta^*)}{g(\theta^*|\theta)} \frac{\pi(\theta^*)}{\pi(\theta)}, \tag{4}$$

and thus $\lambda(\theta^*, \theta)$ disappears in the expression. Hence, using only a bit of algebra we have done away with the need to evaluate the normalising constant. (Note that the motivation for introducing an auxiliary variable differs somewhat from previous cases in the literature, e.g. Higdon, 1998.)

Møller et al. (2006) give some heuristic motivations for the choice of auxiliary density and observe that if the auxiliary distribution is poorly chosen the algorithm will mix poorly or not at all. In particular the algorithm may appear to be mixing well for many iterations before it gets stuck in one state for thousands of iterations (something we shall see examples of in Section 5). Since we have to accept both θ and y simultaneously in (4), the choice of distribution for y clearly is crucial to achieve an acceptable acceptance rate but is not immediately clear how (4) relates to the Hastings ratio (3).

For understanding AVM it is helpful to consider it in terms of importance sampling. If we inspect the part in (4) that pertains to the auxiliary variable and write

$$\lambda(\theta^*, \theta; y^*, y) = \frac{\lambda(\theta^*, \psi; y^*)}{\lambda(\theta, \psi; y)} \tag{5}$$

where

$$\lambda(\theta, \psi; y) = \frac{q(y; \psi)}{q(y; \theta)}, \tag{6}$$

we see that $\lambda(\theta, \psi; y)$ is an estimator of $\lambda(\theta, \psi)$ in the sense that the expected value

$$E_{y|\theta}\left\{\lambda(\theta, \psi; y)\right\} = \sum_{y \in \mathcal{X}} \left[\frac{q(y; \psi)}{q(y; \theta)}\right] \frac{1}{c(\theta)} q(y; \theta) = \frac{c(\psi)}{c(\theta)}.$$

This is the principle of the importance sampler (e.g. Ott, 1979; Geyer and Thompson, 1992; in the follwoing we use the acronym SIS, denoting "simple importance sample", to distinguish this standard importance sampler from other importance samplers). More specifically, if $y^{(1)}, \ldots, y^{(M)}$ is a sample from the importance distribution

$q(y;\theta)/c(\theta)$, the ergodic average

$$\bar{\lambda}(\theta,\psi) = \frac{1}{M}\sum_{m=1}^{M}\lambda(\theta,\psi;y^{(m)}) \tag{7}$$

is the SIS estimator of $\lambda(\theta,\psi)$. The SIS estimator is simulation consistent in the sense that $\bar{\lambda}(\theta,\psi)$ tends to $\lambda(\theta,\psi)$ as M gets large. The AVM may then be seen as a Metropolis-Hastings algorithm where an SIS is run with $M = 1$ in each iteration to approximate the true Hastings ratio through estimation of (6) in (5).

3. Using bridging distributions in importance samplers for estimating ratios of normalising constants

Even if the support of y under the different distributions defined by θ and ψ is (loosely speaking) the same, that is, $q(\cdot;\theta)$ is zero whenever $q(\cdot;\psi)$ is, it is common for the supports of $q(\cdot;\psi)$ and $q(\cdot;\theta)$ to be well separated in the sense that there is a region in $\mathcal{X}$ that has a low probability under both $q(\cdot;\psi)$ and $q(\cdot;\theta)$ that separates the regions of high probability under the respective distributions.

In this situation, we rarely produce y from $q(\cdot;\theta)$ with high probability under $q(\cdot;\psi)$. This typically manifests itself in high or infinite variance for $\bar{\lambda}(\theta,\psi)$. Note however that this also applies to the less extreme cases whenever the "overlap" between distributions is too small. If we were to perform SIS repeatedly for many different values of θ it would be hard to monitor how close the supports $q(\cdot;\psi)$ and $q(\cdot;\theta)$ are.

Assuming that there is a distribution indexed by a parameter $\theta_{1/2}$, whose support overlaps those of $q(\cdot;\psi)$ and $q(\cdot;\theta)$, we may use this as a bridging distribution. By expanding $\lambda(\theta,\psi)$

$$\lambda(\theta,\psi) = \frac{c(\psi)}{c(\theta)} = \frac{c(\psi)}{c(\theta_{1/2})}\times\frac{c(\theta_{1/2})}{c(\theta)}, \tag{8}$$

we may obtain a more stable estimate of $\lambda(\theta,\psi)$ by estimating the ratios of normalising

constants to the bridging distribution (referred to as "bridge sampling" by Meng and Wong, 1996, and the "acceptance ratio method" by Bennett, 1976). The rationale is that if the overlap between $q(\cdot;\psi)$ and $q(\cdot;\theta_{1/2})$, and $q(\cdot;\theta)$ and $q(\cdot;\theta_{1/2})$ is greater than the overlap between $q(\cdot;\psi)$ and $q(\cdot;\theta)$, then an estimator based on the expansion (8) would have smaller variance than the corresponding SIS (Neal, 1993). In practice we may be required to have more than one bridging distribution but the principle remains unchanged.

Path sampling is a generalisation of bridged importance sampling that draws on the principle of thermodynamic integration in statistical physics (Gelman and Meng, 1998; Neal, 1993). Consider extending the number of bridging distributions to "uncountably many" bridging distributions, indexed by parameters $\theta(t)$ for a smooth mapping $\theta : [0,1] \to \Theta$, e.g. linear $\theta(t) = t\psi + (1-t)\theta$, that connects $\theta(0) = \theta$ and $\theta(1) = \psi$. The estimator of the logarithm of LHS of (8) may then be derived from the path sampling identity: $\log \lambda(\theta,\psi) = \int_0^1 f(\theta(t))^{\mathrm{T}}(\mathrm{d}\theta(t)/\mathrm{d}t)\mathrm{d}t$; where $f(\theta) = E_{y|\theta}\{\mathrm{d}\log(q(y;\theta))/\mathrm{d}\theta\}$. MCMC sample equivalents of $f(\theta(t))$, may then be averaged over different values of t in the interval $[0,1]$, to obtain an estimate of $\log \lambda(\theta,\psi)$ (as in for example Hunter and Handcock, 2006).

3.1 Linked importance sampling (LIS)

Neal (2005) proposed a method he called linked importance sampling (LIS) that combines the merits of the SIS (being unbiased) with the advantages of using bridging distributions while not requiring more than one independent realisation from an importance distribution. The path sampler, though efficient, requires several independent draws of $y \in \mathcal{X}$. When we use MCMC to generate sample points in the data space this translates into having to wait for the MCMC to burn in between each sample point. LIS is best described as a sequence of MCMC samples as in Figure 1 (an applied example is given in connection with Figure 2), each with different target distributions, that are linked (as in share realisations (b) and (c)) with each other. The two schematically represented supports of distributions indexed by $\theta(0)$ and $\theta(2)$

both have overlap with the support of the bridging distribution $\theta(1)$. The state a is a realisation from the distribution $p(y|\theta(0))$. The Markov chains are generated using MCMC conditional on the linking and initial states (with the difference from the standard implementation being the need to simulate both forward in time from a as well as backwards in time). The LIS is related to the annealed importance sampler (Neal, 2001; Jarzynski, 1997), with the difference that the latter takes an initial sample point from the first distribution and then updates this according to a transition kernel defined by the bridging distribution.

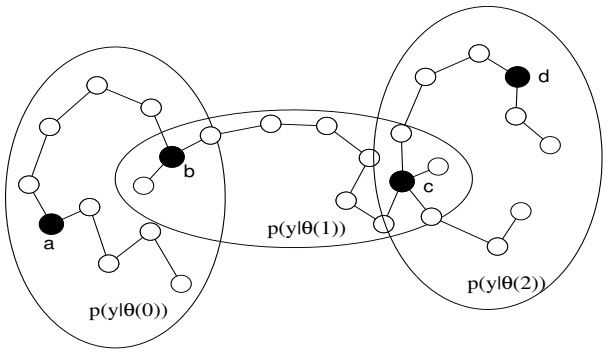

Figure 1: An illustration (based on Figure 1 in Neal, 2005) of LIS that starts in the vertex a and ends in vertex d

3.1.1 The sample

To describe LIS in more detail, we begin by considering MCMC sampling into the future and into the past. By assumption we may draw y from $p(y|\theta)$ using MCMC. Denote by $T_\theta(y^{(t)}, y^{(t+1)})$ the Markov chain transition probabilities that are used to update a state $y^{(t)}$ to a new state $y^{(t+1)}$ in the standard implementation of producing a Markov chain running forwards in time with target distribution $p(y|\theta)$.We may also produce a Markov chain running backwards in time operating such that if the present state is $y^{(t)}$, the next (or previous as it were) state $y^{(t-1)}$ is drawn using the reverse transition probabilities $\underline{T}_\theta(y^{(t)}, y^{(t-1)})$. For reversible MCMC, which includes

the one-block Metropolis algorithm, $\underline{T}_\theta(y,x) = T_\theta(y,x)$. For the Gibbs sampler with systematic updates of coordinates, $\underline{T}_\theta(y,x)$ corresponds to updating the coordinates in reversed order (Neal, 2005).

The LIS estimator is based on K sample points from Markov chains with m different target distributions $(y_1^{(1)},\ldots,y_1^{(K)})$, $(y_2^{(1)},\ldots,y_2^{(K)})$, ..., $(y_m^{(1)},\ldots,y_m^{(K)})$, drawn using Metropolis-Hasting transition probabilities $T_{\theta(t)}$ and $\underline{T}_{\theta(t)}$, for a smooth mapping connecting θ and ψ as in path sampling (Gelman and Meng, 1998). A convenient choice is to let $\theta(j) = (j-1)/(m-1)\psi + (1-(j-1)/(m-1))\theta$.

The m samples are connected in points

$$\mu_1,\ldots,\mu_m \text{ and } \nu_1,\ldots,\nu_m,\ \mu_i,\nu_i \in \{1,\ldots,K\},$$

such that given μ_j and $y_j^{(1)},\ldots,y_j^{(K)}$, we let the state $y_{j+1}^{(\nu_{j+1})} := y_j^{(\mu_j)}$. In Figure 1, b is such a linking state, linking the first chain to the second. Given ν_j and $y_j^{(\nu_j)}$ we create the chain $y_j^{(1)},\ldots,y_j^{(K)}$ by simulating forward from $y_j^{(\nu_j)}$ using $T_{\theta(j)}(y_j^{(\nu_j)}, y_j^{(\nu_j+1)})$, $T_{\theta(j)}(y_j^{(\nu_j+1)}, y_j^{(\nu_j+2)})$, etc., until we have produced $y_j^{(K)}$. We also simulate backwards from $y_j^{(\nu_j)}$ using the reversed transition kernels $\underline{T}_{\theta(j)}(y_j^{(\nu_j)}, y_j^{(\nu_j-1)})$, $\underline{T}_{\theta(j)}(y_j^{(\nu_j-1)}, y_j^{(\nu_j-2)})$, etc., until we have produced $y_j^{(1)}$. The implied pmf of a chain $y_j = (y_j^{(i)})_{i=1}^K$ conditional on the insertion point and the linking state is

$$P(y_j|\nu_j, y_j^{(\nu_j)}) = \prod_{i=1}^{\nu_j-1} \underline{T}_{\theta(j)}(y_j^{(i+1)}, y_j^{(i)}) \prod_{i=\nu_j}^{K} T_{\theta(j)}(y_j^{(i)}, y_j^{(i+1)}).$$

To choose which of the K sample points that should provide the link to the next chain, we choose μ_j with probabilities

$$\eta(\mu_j|y_j) = \frac{w(y_j^{(\mu_j)};\theta(j),\theta(j+1))}{\sum_{i=1}^K w(y_j^{(i)};\theta(j),\theta(j+1))},$$

where

$$w(y;\theta,\theta^*) = q(y;\theta)^{-1/2} q(y;\theta^*)^{1/2} \tag{9}$$

and insertion points ν_j are chosen uniformly on $\{1,\ldots,K\}$. The initial state $y_1^{(\nu_1)}$ (a in Figure 1) of the first chain is chosen according to $p(y_1^{(\nu_1)}|\theta)$.

3.1.2 The estimator

Given a sample $\omega = (y, \mu, \nu)$, the LIS estimate of $\lambda(\theta,\psi)$ is given by

$$\lambda_{LIS}(\theta,\psi;\omega) = \prod_{j=1}^{m-1} \frac{\sum_{i=1}^{K} w(y_j^{(i)};\theta(j),\theta(j+1))}{\sum_{i=1}^{K} w(y_{j+1}^{(i)};\theta(j+1),\theta(j))}. \tag{10}$$

At this point we may stop to consider the form of the weights (9). These ratios are in fact the square roots of the quantities $\lambda(\theta,\theta^*;y)$ used in the simple importance sampler. In this case, it derives from the implicit use of a geometric linking distribution but Neal (2005) also suggests other possible forms for these weights.

The sampling scheme outlined above defines an unnormalised distribution

$$Q^F_{\theta,\psi}(\omega) = q(y_1^{(\nu_1)};\theta) \prod_{j=1}^{m} \frac{1}{K} P(y_j|\nu_j, y_j^{(\nu_j)}) \eta(\mu_j|y_j) \tag{11}$$

on $\Omega \subseteq \prod_{j=1}^{m} \mathcal{X}^K \times \{1,\ldots,K\} \times \{1,\ldots,K\}$. For each $\omega \in \Omega$ we may also define the algorithm in reverse, i.e. starting in $y_m^{(\mu_m)}$, treating this as $y_1^{(\nu_1)}$ and proceeding as above but swapping roles for ν and μ. In Figure 1, this corresponds to starting in b, rather than a, and proceed "backwards". This analogously defines an unnormalised pmf

$$Q^B_{\psi,\theta}(\omega) = q(y_m^{(\mu_m)};\theta) \prod_{j=1}^{m} \frac{1}{K} P(y_j|\mu_j, y_j^{(\mu_j)}) \eta(\nu_j|y_j). \tag{12}$$

Inspecting the unnormalised forwards and backwards distributions, $Q^F_{\theta,\psi}$ and $Q^B_{\psi,\theta}$, it is clear that their normalising constants must be $c(\theta)$ and $c(\psi)$ respectively. Furthermore, using a little algebra it can be shown (the details are given in Neal, 2005) that

the LIS estimator (10) can be written

$$\lambda_{LIS}(\theta, \psi; \omega) = \frac{Q^B_{\psi,\theta}(\omega)}{Q^F_{\theta,\psi}(\omega)}. \tag{13}$$

If we let $P^F_{\theta,\psi}(\omega) = Q^F_{\theta,\psi}(\omega)/c(\theta)$, and ω be a variate from this distribution, then $\lambda_{LIS}(\theta, \psi; \omega)$ is the SIS estimator of the ratio of normalising constants for $P^F_{\theta,\psi}(\omega)$ and $P^B_{\theta,\psi}(\omega) = Q^B_{\psi,\theta}(\omega)/c(\psi)$, namely $\lambda(\theta, \psi)$. As opposed to the SIS and the estimator (7), LIS employs bridging distributions to span the supports corresponding to θ and ψ, but in contrast to bridge sampling the variates from the bridging distributions are generated dependent on each other.

4. Proposed approach

4.1 Combining importance sampling and auxiliary variable MCMC

The question is now whether we can improve on the performance of AVM by getting a better estimate of $\lambda(\theta^*, \psi)$ than the SIS with $M = 1$? There are a few aspects of the importance samplers presented that prevent immediate incorporation in the AVM. For example, here $\bar{\lambda}(\theta, \psi) \to \lambda(\theta, \psi)$ only as M gets large and we have to get an estimate in every iteration. If the distributions indexed by θ and ψ have little or no overlap there could be a severe bias or infinite variance. As we have seen this can be remedied by introducing bridging distributions but in general when an importance sampler is used to approximate the Hastings ratio H by $\hat{H}$, while $E_{y|\theta^*}\{\hat{H}\} = H$, we have that $E_{y|\theta^*}\{\hat{H}\} > E_{y|\theta^*}\{\min(1, \hat{H})\}$. Consequently, if we use importance samplers with an approximation $\hat{H}$ inplace of H we may accept updates in the Metropolis-Hastings with the wrong probabilities on average.

4.2 LISA - extended state space

In AVM we performed draws from the joint distribution of the parameters and the auxiliary variable $y \in \mathcal{X}$. Consider now as an auxiliary variable $\omega \in \Omega$, and a

distribution $P^B_{\psi,\theta}(\omega)$ that depends on both θ and ψ. The linked importance sampler (LISA) MCMC is a Metropolis-Hastings algorithm that performs draws from the joint distribution

$$\begin{aligned}\pi(\omega,\theta|x) &= P^B_{\psi,\theta}(\omega)\pi(\theta|x)\\ &\propto Q^B_{\psi,\theta}(\omega)\frac{q(x;\theta)}{c(\theta)}\pi(\theta).\end{aligned}$$

It is straightforward to show that θ has the desired marginal distribution $\sum_{\omega\in\Omega}\pi(\omega,\theta|x) = \pi(\theta|x)$. The Hastings ratio in the Metropolis updating step still contains the ratio $\lambda(\theta^*,\theta)$

$$H = \frac{q(x;\theta^*)/c(\theta^*)}{q(x;\theta)/c(\theta)}\frac{Q^B_{\psi,\theta^*}(\omega^*)}{Q^B_{\psi,\theta}(\omega)}\frac{g(\theta,\omega|\theta^*,\omega^*)}{g(\theta^*,\omega^*|\theta,\omega)}\frac{\pi(\theta^*)}{\pi(\theta)}.$$

Assume now that we condition on θ^* proposed from $g(\theta^*|\theta)$, and use $P^F_{\theta^*,\psi}(\omega^*)$ to propose ω^*. Substituting this pmf into the Hastings ratio we get

$$\frac{q(x;\theta^*)/c(\theta^*)}{q(x;\theta)/c(\theta)}\frac{Q^B_{\psi,\theta^*}(\omega^*)}{Q^B_{\psi,\theta}(\omega)}\frac{Q^F_{\theta,\psi}(\omega)/c(\theta)}{Q^F_{\theta^*,\psi}(\omega^*)/c(\theta^*)}\frac{g(\theta|\theta^*)}{g(\theta^*|\theta)}\frac{\pi(\theta^*)}{\pi(\theta)}$$

where we see that $\lambda(\theta^*,\theta)$ cancel against its reciprocal stemming from the ratio of proposal distributions $P^F_{\theta,\psi}(\omega)/P^F_{\theta^*,\psi}(\omega^*)$. Since $q(\cdot;\theta)$ is easy to evaluate by assumption and $Q^B_{\psi,\theta}$ and $Q^F_{\theta,\psi}$ consist of simple functions of q according to (11) and (12), we are left only with known quantities for updating θ and ω.

In order to interpret this algorithm we may note that upon rearranging the Hastings ratio it can be expressed in terms of estimates λ_{LIS} using the identity (13)

$$H = \frac{q(x;\theta^*)}{q(x;\theta)}\frac{\lambda_{LIS}(\theta^*,\psi;\omega^*)}{\lambda_{LIS}(\theta,\psi;\omega)}\frac{g(\theta|\theta^*)}{g(\theta^*|\theta)}\frac{\pi(\theta^*)}{\pi(\theta)}. \tag{14}$$

As the number of sample points gets large (for more detailed results on asymptotics on K, see Neal, 2005), (14) will approach the true (marginal) acceptance probability as $\lambda_{LIS}(\theta^*,\psi;\omega^*)$ and $\lambda_{LIS}(\theta,\psi;\omega)$ tend to their means so that $\lambda_{LIS}(\theta^*,\psi;\omega^*)/\lambda_{LIS}(\theta,\psi;\omega) \rightarrow$

$\lambda(\theta, \theta^*)$, the ratio of normalising constants in (3).

4.3 LISA in summary

We may now make the following observations regarding LISA. LISA is a Metropolis-Hastings MCMC defined on Ω and Θ, where in each iteration we propose a move to a new parameter value $\theta \in \Theta$ and, conditional on this θ and a fixed ψ, an element $\omega \in \Omega$ is proposed from a distribution $P^F_{\theta,\psi}$. These proposed moves are either jointly accepted or rejected with the Hastings ratio (14) corresponding to the joint target distribution $P^B_{\psi,\theta}(\omega)\pi(\theta|x)$. As seen in (14) there is no need to save or keep track of the realisations ω. Since the LISA procedure is equivalent to a Metropolis-Hastings for the target distribution $\pi(\theta|x)$ where we estimate the ratio $\lambda(\theta, \psi)$ using the LIS estimator, we need only to save the estimate $\lambda_{LIS}(\theta, \psi; \omega)$ for the current state θ. The mixing of LISA may be tuned using the constants m and K, corresponding to the number of bridging distributions and chain lengths respectively. The algorithm of Møller et al. (2006) may be considered the special case $K = 1$ and $m = 1$, i.e. when we only produce the initial state $y_1^{(\nu_1)}$ and there are no bridging states.

LISA is a one-block Metropolis-Hastings MCMC on the joint state space of Ω and Θ for all choices of K and m as long as $y_1^{(\nu_1)}$ is an independent realisation. We have made the assumption that we may produce draws from the model $p(\cdot|\theta)$ using MCMC but not explicitly that we may draw "directly" from $p(\cdot|\theta)$. For some models we may use MCMC to produce an "exact" sample (Propp and Wilson, 1996) from $p(\cdot|\theta)$ and in the cases where we may not, performing draws from $p(\cdot|\theta)$ using MCMC is so computationally efficient and cheap as to be equivalent to producing independent draws to all intents and purposes.

5. Examples

We shall now proceed to illustrate the effects of different choices for the tuning constants K and m and for the distribution, as defined by ψ, for some illustrative

data sets. Neal (2005) provide some heuristics for the effect of different K and m in the general case and some more detailed results for a specific model. An efficient choice of ψ is likely to belong to a region with high mass in the posterior of θ but the derivation of an optimal choice of ψ may be difficult (Møller et al., 2006). As we shall see, $K = m = 1$, i.e. AVM, works for a model with a simple dependence structure, as in the Ising model, but K and m have to be increased for models for data with more complex dependences steming from realisic assumtions for individuals. In the examples to come, improper constant priors were used (the conditions under which the posteriors are well-defined are treated in Koskinen, Robins, and Pattison 2008b, drawing on the results of Diaconis and Ylvisaker, 1979).

5.1 Ising model on binary 50×50 lattice

A well-known case of an autologistic model is the Ising model (Besag, 1972; Cressie, 1993). This model was used for illustration of the AVM by Møller et al. (2006) and we generate data according to some of the parameter specifications used there in order to provide a comparison. Apart from the Ising model being a well known model, it has a relatively simple dependence structure and allows for perfect sampling of data. The Ising model on a binary $u \times n$ lattice has been used to model how the charges of particles interact and in the simplest case it is assumed that the particles can have either of two spins, up or down. Outside of statistical mechanics the model has for example been used for modelling spatial autocorrelation in functional magnetic resonance imaging of the brain (Smith and Fahrmeir, 2007) and related models, e.g. the Potts, have been used in geographical modelling (e.g. Green and Richardson, 2002). The spin of a given particle depends on the general tendency towards spin up and the spins of its neighbours on the lattice. The neighbourhood of a particle (i, j) is defined as $\{(k, \ell) : |i - k| + |j - \ell| = 1\}$. We define the model for data $x = (x_{ij} : i = 1, \ldots, u, \text{and, } j = 1, \ldots, n)$, where the element x_{ij} is equal to 1 or -1, according to whether the corresponding particle has spin up or spin down. The

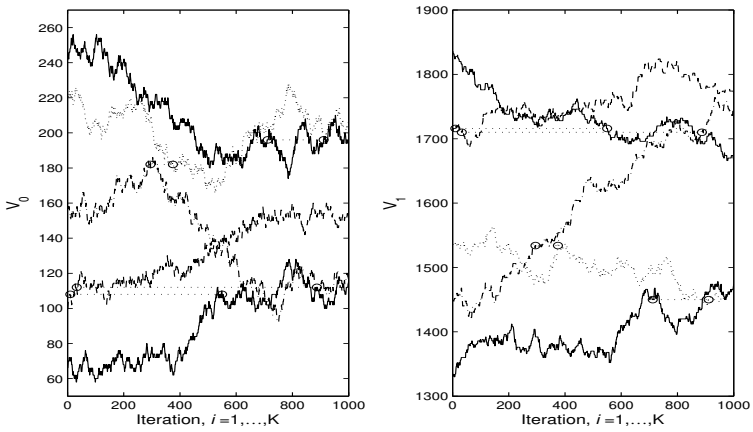

Figure 2: A linked importance sample for the Ising model on a binary 50×50 lattice. Traces for V_0 (left) and V_1 (right) for chains generated by Metropolis-Hastings for $\theta(1)$, '$-$', $\theta(2)$, '$--$', $\theta(3)$, '$-\cdot$', $\theta(4)$, '$\cdot\cdot$', $\theta(5)$, '$-$'; where $\theta(1) = (0, 0.3)^{\mathrm{T}}$ and $\theta(5) = (0.1, 0.2)^{\mathrm{T}}$. Connections in μ_j and ν_{j+1} indicated by circles

pmf is defined as in (1) with

$$q(x;\theta) = \exp(\theta_0 V_0 + \theta_1 V_1),$$

where

$$V_0 = \sum_{i=1}^{u}\sum_{j=1}^{n} x_{ij}, \text{ and } V_1 = \sum_{i=1}^{u-1}\sum_{j=1}^{n} x_{ij}x_{(i+1)j} + \sum_{i=1}^{u}\sum_{j=1}^{n-1} x_{ij}x_{i(j+1)}$$

When generating data according to the model we have to rely on MCMC since there is no direct way of drawing data from an Ising model. In the case of the Ising model it is possible to "sample perfectly" from the model, that is, to take as an output a state that we know to have been produced after the Markov chain has converged to the target distribution. Here we have used Wilson's (2000) modification to the Propp and Wilson (1996) algorithm, coupling from the past using a read once only source of randomness (CFTPRO).

To illustrate how the chains $(y_j^{(1)}, \ldots, y_j^{(K)})$ are connected for $j = 1, \ldots, m$ in the

	$K=1$		$K=3000$		$K=7000$	
	$m=1$		$m=5$		$m=9$	
Posterior mean	0.201	0.105	0.199	0.105	0.199	0.105
Posterior std	0.021	0.014	0.022	0.014	0.022	0.015
Lag 50 SACF	0.646	0.512	0.526	0.362	0.460	0.313
Lag 100 SACF	0.434	0.329	0.292	0.195	0.201	0.109
ESS	451	645	526	662	1099	1351
Mean acc. prob	0.393		0.589		0.725	
Prop. min(1,H) $\leq e^{-10}$	0.018		0.0006		0.0000	

Table 1: Comparisons for LISA applied to a data set simulated from an Ising model with $\theta = (.2, .1)^{\mathrm{T}}$ on a binary 50×50 lattice for different tuning constants K and m. The MPLE was $(0.196, 0.109)^{\mathrm{T}}$

LIS part of LISA, we have plotted the sufficient statistics V_0 and V_1 for $m = 5$ chains in Figure 2. The starting point $y_j^{(\mu_1)}$ is generated (using CFTPRO) from an Ising model defined by $\theta = (\theta_0, \theta_1)^{\mathrm{T}}$, with $\theta_0 = 0$ and $\theta_1 = 0.3$. A linear map is used with $\theta(5) = \psi = (.1, .2)^{\mathrm{T}}$. For the right hand panel, showing the traces of V_1, we expect the chains to progressively move downwards since the parameter θ_1 corresponding to the number of same spin sites is gradually lowered. The state connecting the first chain with the second is $y_1^{(\mu_1=550)}$, with $V_1 = 1716$, which is then set as the initial state $y_2^{(\nu_2=10)}$, in the second chain etc. until the last chain is started in $y_5^{(\nu_5=714)}$, whose $V_1 = 1450$ is considerably lower than the overall level of the number of same spin sites in the first chains. Thus the bridging chains have managed to link the supports of the two extreme distributions defined by $\theta(1)$ and $\theta(5)$.

To compare AVM with LISA and investigate the effect of different choices of K and m, we apply the estimation schemes to two simulated data sets, one with parameters $\theta = (0.2, 0.1)^{\mathrm{T}}$ and the other with $\theta = (0, 0.3)^{\mathrm{T}}$. The results for the first data set are given in Table 1 and the second in Table 2. A bivariate normal distribution with covariance matrix $.005I$ centred on the current parameter vector was used as the proposal for θ in all estimations and ψ was set to the maximum pseudo likelihood estimate (MPLE) (Besag, 1974, 1975). A total of 100,000 iterations was used and the

	$K=1$		$K=3000$		$K=7000$	
	$m=1$		$m=5$		$m=9$	
Posterior mean	−0.004	0.300	−0.003	0.299	−0.003	0.299
Posterior std	0.007	0.011	0.008	0.012	0.008	0.011
Lag 50 SACF	0.309	0.470	0.181	0.335	0.110	0.199
Lag 100 SACF	0.182	0.354	0.096	0.205	0.045	0.067
ESS	877	214	1235	521	1852	1563
Mean acc. prob	0.250		0.343		0.421	
Prop. min(1,H) $\leq e^{-10}$	0.0865		0.0203		0.0067	

Table 2: Comparisons for LISA applied to a data set simulated from an Ising model with $\theta = (0, .3)^{\mathrm{T}}$ on a binary 50×50 lattice for different tuning constants K and m. The MPLE was $(-0.005, 0.315)^{\mathrm{T}}$.

figures in the tables are based on the un-thinned entire sample without any burnin period.

If we compare the analyses from the three combinations of K and m, we may first note that the posterior means are not markedly different across the different tuning settings. The mixing does improve considerably when the tuning constants are increased as reflected by the sample autocorrelation functions (SACF). Though there are some differences in the lag 50 SACF, the real difference is seen in the lag 100 SACF. For $K = 1$ and $m = 1$ (AVM) the autocorrelation is still considerable at lag 100 for both parameters, roughly .43 and .33. These are reduced to .29 and .2 when K is increased to 3000 and m to five, and then again to .2 and .11, for $K = 7000$ and $m = 9$. The lag 100 SACF is hence reduced to a little less than a third for θ_1 when we compare AVM and LISA with $K = 7000$ and $m = 9$. The improvement in mixing is also reflected in the autocorrelation time and the effective sample sizes (ESS)(ESS is the total number of iterations divided by the autocorrelation time; autocorrelation time was calculated according to Kass, Carlin, Gelman, and Neal, 1998).

The reason for the relatively slow mixing for small tuning constants is that the Markov chain gets stuck in some states for long spells. From trace plots it looks as if the chain is mixing well for a long period of time before it gets stuck. A symptom

of this is that acceptance rate ("Mean acc. prob", the average of min(1,H)) for AVM looks reasonable but that the proportion of "extreme" proposals ("Prop. $\min(1, H) \leq e^{-10}$", the proportion of times the log probability of accepting a proposed move was less than -10) is close to two per cent according to Table 1.

The increased mixing comes at a price however. To improve mixing we need to increase the tuning constants and thus increase the number of iterations in the Metropolis-Hastings on $\mathcal{X}$. This increase may however be considered modest even for $K = 7000$ and $m = 9$, since the extra Km iterations needed to calculate LIS has to be compared to the number of iterations it takes to generate $y_1^{(\nu_1)}$. For the MCMCs in Table 1, an average of 5.6×10^4 iterations were required by the perfect sampler to generate $y_1^{(\nu_1)}$, and in each of these iterations 5 chains are updated in tandem.

For the data in Table 2, the differences between point estimates are again small but there is a great reduction in the lag 100 SACF. This is particularly the case for the interaction parameter θ_1 for which the lag 100 SACF decreases from .35 to .07. This is also what we would expect when we compare the first data set with the second since the second data set is generated from a model with higher spatial autocorrelation, a larger θ_1, and hence higher degree of deviation from independence of observations.

5.2 The social influence model

With the advent of multilevel models (e.g. Snijders and Bosker, 1999) it was recognized that outcomes of respondents belonging to the same geographical units, having the same institutional affiliations or being part of the same group cannot plausibly be treated as independent observations. Multilevel models take the interdependence between observations in the same units into account by incorporating random and fixed effects into generalised linear models. Although a range of work has been done on different kinds of interdependence such as crossed random effects and multiple membership models (Rashbash & Browne, 2002), little attention has been paid to incorporating dependencies stemming from social interaction (with some notable exceptions for multilevel modelling of the interaction itself, e.g. Snijders & Baerveldt,

2003). For example, if one makes allowances for non-independence of pupils in a school class it seems obvious that one should pay equal heed to the fact that some pupils interact more than others. Two models that explicitly incorporates the structure of social interaction are the network effects model (Ebring & Young, 1979; Doreian, 1982), for continuous response variables, and the social influence model (Robins, Pattison, and Elliot, 2001), for binary response (Brock and Durlauf, 2001, describe models where there is no dependence between the individual out-comes; for related models for longitudinal data see Snijders, Steglich, and Schweinberger, 2007). There are some obvious similarities between the social influence model and the Ising model. For the former the interdependence structure is however given by empirical observations, which is not homogeneous and usually quite complicated (for a derivation and motivation of the sufficient statistics see Robins et al., 2001).

Here we fit the social influence model to a data set for 106 pupils (all male) in a school class in Australia (Lusher, 2006). We let $x = (x_i : i = 1, \ldots, 106)$ be our response variable where, for each of the 106 pupils, x_i is equal to 0 if pupil i has gender equity attitudes and, 1 if pupil i has male dominance attitudes. We have 4 covariates of substantial interest: dominant culture, v_{i1}, which indicates if i has an Anglo-Australian ethno-cultural background (1) or not (0); v_{i2}, the socio economic status of i's household (as measured by standardised SES based on postcode); v_{i3}, the occupational score for the father of i (original range is 0 to 100 according to Jones & McMillan, 2001, but here it is standardised); v_{i4}, the occupational score for the mother of i. Each of the pupils were furthermore asked to nominate who their friends were and (after symmetrising) we have an adjacency matrix $a = (a_{ij} : 1 \leq i, j \leq 106)$, with elements a_{ij} equal to 1 if i has nominated j as his friend or if j has nominated i as his friend, and 0 otherwise.

The model we fit is defined as in (1) with

$$q(x;\theta) = \left\{ \theta_1 \sum_{i=1}^{106} x_i + \theta_2 \sum_{i=1}^{106}\sum_{j=1}^{106} x_i a_{ij} + \theta_3 \sum_{i=1}^{106}\sum_{j=1}^{106} x_i x_j a_{ij} + \sum_{k=1}^{4} \theta_{k+3} \sum_{i=1}^{106} x_i v_{ik} \right\}.$$

	MPLE		MCMCMLE		Posterior			
	EST	SE	EST	SE	MEAN	STD	95 HPD	
Intercept, θ_1	0.11	0.596	0.12	0.504	0.14	0.560	−1.15	1.45
Activity, θ_2	−0.24	0.083	−0.13	0.046	−0.11	0.055	−0.23	0.02
Contagion, θ_3	0.45	0.123	0.29	0.067	0.24	0.082	0.04	0.42
Dominant, θ_4	−0.03	0.472	−0.39	0.427	−0.44	0.451	−1.44	0.57
SES, θ_5	0.10	0.227	0.20	0.215	0.23	0.223	−0.27	0.72
Father, θ_6	−0.16	0.220	−0.17	0.210	−0.19	0.219	−0.69	0.30
Mother, θ_7	−0.05	0.224	0.08	0.206	0.08	0.215	−0.41	0.57

Table 3: Estimates for influence model fitted to Lusher's (2006) 106 school data. Bayes estimates based on LISA with $K = 2000$ and $m = 7$

The parameters θ_4 through θ_7 may be interpreted as in a regular logistic regression model and θ_1 as the intercept. The parameter θ_2 corresponds to what is called the activity effect, something which is meant to capture whether pupils with many ties are more likely ($\theta_2 > 0$) to have the response 1. The activity parameter also acts as a control for the "contagion effect" that is associated with the parameter θ_3. The latter is where the effect of social interaction on interdependency is taken into account and the interpretation of a positive θ_3 is that friends tend to have responses similar to each other controlling for everything else.

For different choices of tuning parameters LISA is run to estimate the parameters in the models, where in three cases ψ is set to the MPLE and in three cases set to the MLE (obtained from an MCMC approximation of Fisher scoring: Geyer and Thompson, 1992; Hunter and Handcock, 2006). The covariance matrix in the proposal distribution for the parameters was set to $.4/\sqrt{1+p}$ times an estimate, $\hat{\Sigma}$, of the covariance matrix in the target distribution (this is roughly in accordance with the guidelines offered by Tierney, 1994, and Roberts, Gelman and Giles, 1997). For LISA based on the MPLE, $\hat{\Sigma}$ was set to a diagonal matrix with diagonal elements being the squares of the MPLE standard errors. For results based on the MLE, $\hat{\Sigma}$ was set to the inverse of the estimated Fisher information matrix. The MPLE and MLE and their corresponding standard errors are found in Table 3. A total of 100,000 iterations

of LISA were performed for all simulation settings. For the influence model it is not straightforward to construct a monotonic chain such that we may implement a perfect sampling scheme (such as CFTPRO) for drawing $y_1^{(\nu_1)}$. Instead we have relied on the rule of thumb for the length of burn in that is $100n$, as suggested in Snijders (2002). (Inspection of Markov chains with over-dispersed initial states for parameter vectors in the tails of the posterior indicated that the burn in was sufficient.)

When ψ is set to the MPLE we see an improvement of the mixing in the trace plots of the first three parameters (first three columns in Figure 3) as we increase the tuning parameters but on the whole, judging by the histograms, these posterior samples are of little practical use for analysis. The traces for $K = 2000$ and $m = 7$, look reasonably good except for the run around iteration 40,000 where the sampler remains in the same state for roughly 6,000 iterations. Whereas the chain with $K = m = 1$ keeps within the confidence intervals for the MPLEs given in Figure 3 for all the 100,000 iterations, the chains increasingly move outside of the intervals as the tuning parameters are increased.

When ψ is set to the MLE the chains (the third through 6th column in Figure 3) with $K = 200$ and $m = 5$, and $K = 2000$ and $m = 7$, move freely in the parameter space but the chain with $K = m = 1$ exhibits strange behaviour. The latter appears to be mixing well for the first 20,000 iterations only to get stuck in a state close to the MLE for more than 50,000 iterations. This happens when the current estimate (13) is far from the true value (in the next section we shall see how the bias reduces as a function of the tuning constants).

To some extent the effect of using the MPLE for ψ is explained by the fact that these estimates are relatively different from the true MLEs (Table 3). In the case of the AVM ($K = m = 1$) this means that most of the time the proposed auxiliary variable will have low probability under the target distribution and hence the acceptance rate will be small. As the tuning constants are increased the state space of the auxiliary variable ω grows larger and the probability function becomes less sensitive to the distance between the target distribution $P^B_{\psi,\theta}$ and the proposal distribution $P^F_{\theta,\psi}$.

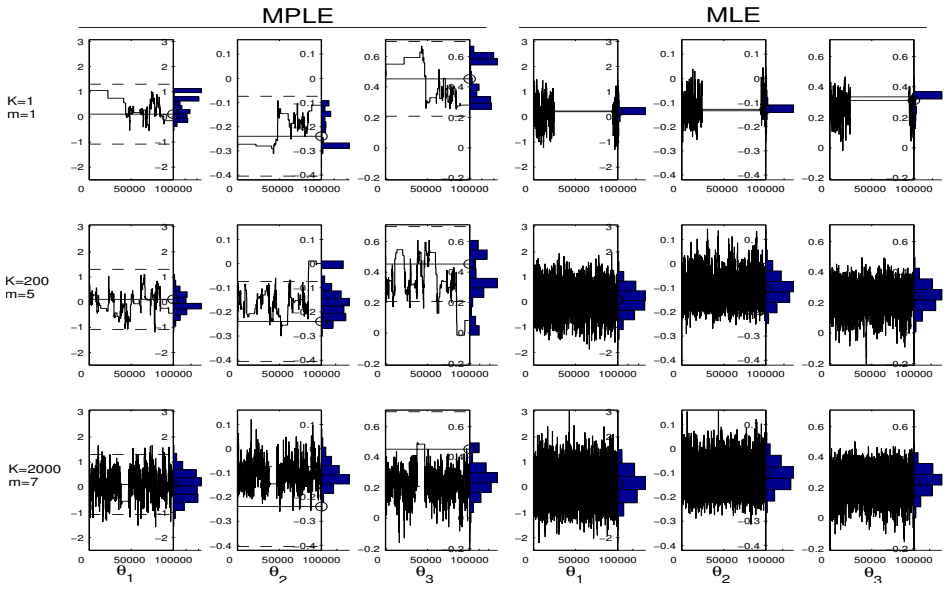

Figure 3: The effect of choice of ψ and tuning constants on mixing for influence model applied to Lusher's (2006) 106 school data. First three columns give trace plots and histograms for intercept, activity and contagion parameters for different choices of K and m when ψ is set to MPLE. Columns 4 through 6 give trace plots and histograms for intercept, activity and contagion parameters for different choices of K and m when ψ is set to MLE.

The most striking differences between the MPLEs and the MLEs are found for θ_2 and θ_3, something which echoes similar result for higher-order stars and triangles in exponential family models for random graphs (Snijders, 2002; for the social influence model, previously estimation has relied only on the MPLEs, Robins et al., 2001). The MLE and the Bayes estimates are similar on the other hand. If we were to compare the conclusions drawn by using Wald's test, on one hand and using .95 highest posterior density intervals (95 HPD) on the other, these would agree except for θ_1.

5.3 Curved exponential family models for networks

A (symmetric) network with n actors may be represented by an $n \times n$ binary adjacency matrix, x, the elements x_{ij} of which are 1 or 0 according to whether actors i and j are relationally tied to each other or not (see for example Wasserman and Faust, 1994). We shall refer to the elements x_{ij} as tie-variables since they indicate the presence of ties between pairs of nodes (actors) in the sociogram (Moreno, 1934) implied by the adjacency matrix. We employ the convention that ties from one actor to him/herself are not meaningful wherefore the diagonal of x is 0. Hence, for a network with n actors we make observations on $n(n-1)/2$ tie variables.

In a seminal paper, Frank and Strauss (1986) introduced a family of distributions for random graphs called Markov graphs. Markov graphs improved on the previous exponential family distributions proposed by Holland and Leinhardt (1981) and Fienberg and Wasserman (1981) in that it allowed for more elaborated dependencies between the tie variables. More specifically, building on results in spatial statistics (Besag, 1974), a dependence graph D was defined for the tie variables with edges in D between variables x_{ij} and $x_{k\ell}$ if and only if the intersection $\{i,j\} \cap \{k,\ell\}$ is non-empty. This has the interpretation that two tie variables are conditionally dependent (given everything else) if they pertain to the same actor or node. Given some assumptions regarding permutation invariance and model parsimony, the dependence assumptions imply an exponential distribution on adjacency matrices with a vector $z(x)$ of sufficient statistics. The Markov graphs have since been extended (Wasserman

& Pattison, 1996; Pattison & Wasserman, 1999; Robins, Pattison, and Wasserman, 1999; Snijders et al., 2006; Hunter and Handcock, 2006) and they are collectively referred to as exponential family random graph models (ERGM) or p^* because they generalise the p_1 model of Holland and Leinhardt (1981).

Here we are going to analyse the collaboration network of $n = 36$ partners in a New England law firm (Lazega, 2001). There is a tie between partner i and j, $x_{ij} = 1$, if i collaborates with j and 0 otherwise. In addition to the adjacency matrix there are 4 covariates: the seniority of i in terms of rank (divided by maximum rank) v_{i1}; a binary indicator of whether i is working in corporate law $v_{i2} = 1$, or litigation, $v_{i2} = 0$; whether i is female (male) $v_{i3} = 1(v_{i3} = 0)$; the location of the office of i, $v_{i4} = 1, 2, 3$ for Boston, Hartford and Providence, respectively.

The general form of an ERGM is according to (1) with

$$q(x;\theta) = \exp\left\{\sum_{k=1}^{p} \theta_k z_k(x)\right\},$$

where the statistics $z_k(x)$ may also include functions of the covariate values. We shall analyse two different models (previously fitted to the same data set by Snijders et al., 2006, and, Hunter and Handcock, 2006), the first one is a dyad-independent model but the second one assumes a more elaborate form of dependence. We include the number of edges in the graph $z_1(x) = \sum_{i<j} x_{ij}$ as a baseline effect, which we will refer to as density. The statistics z_2 and z_3 represent the main effects of seniority and practice $\sum_{i<j} x_{ij}(v_{ik} + v_{jk})$, for $k = 1, 2$. The following 3 statistics are the homophily effects of practice, sex and office, with statistics $\sum_{i<j} x_{ij}\mathbf{1}\{v_{ik} = v_{jk}\}$, where $\mathbf{1}$ is the indicator function, for k equal to 2, 3 and 4 respectively.

As the tie variables are independent of each other in the first model, the model reduces to what is computationally equivalent to a logistic regression of x on the statistics. In the social networks literature this type of model is often (particularly for directed networks) referred to as a dyad-independent model (Holland and Leinhardt, 1981; Fienberg and Wasserman, 1981; Wong, 1987; dependence may however be

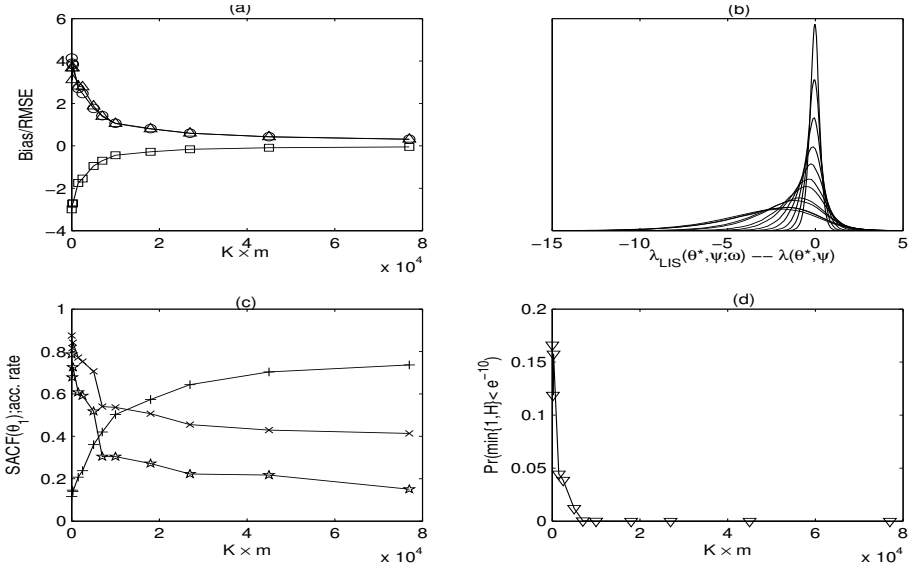

Figure 4: MCMC diagnostics for LISA applied to Lazega's lawyers with dyad independent model. From left to right: (a) Expected error $\lambda_{LIS}(\theta^*, \psi; \omega^*) - \lambda(\theta^*, \psi)$ ($\square$) and mean square error ($\circ$) with respect to $P^F_{\theta^*,\psi}(\omega^*)q(\theta^*|\theta)\pi(\theta|x)$, and mean square error of $\log \lambda_{LIS}(\theta, \psi; \omega) - \log \lambda(\theta, \psi)$ with respect to $\pi(\theta, \omega|x)$ ($\triangle$); (b) distributions of $\log \lambda_{LIS}(\theta^*, \psi; \omega^*) - \log \lambda(\theta^*, \psi)$ for different K and m; (c) SACF for θ_1 at lag 50 ($\times$) at lag 100 ($\star$); acceptance rate ($+$); (d) probability of extreme proposal (∇).

introduced in dyad-independent models using random effects as in van Duijn et al., 2004, or Hoff, 2002). Because of the independence of observations we can easily calculate $c(\theta)$ exactly as

$$c(\theta) = \prod_{i<j} \left\{ 1 + \exp\left[\sum_{k=1}^{p} \theta_k \Delta_{ij} z_k(x) \right] \right\},$$

where $\Delta_{ij} z_k(x)$ is the change in the corresponding statistic when the (i,j)th element of x is changed from 0 to 1, all other elements equal to those in the original matrix. Additionally, drawing from $p(x|\theta)$ is straightforward so that we can generate $y_1^{(\nu_1)}$ without the aid of MCMC.

Some MCMC diagnostics are provided in Figure 4 for posteriors obtained using LISA with ψ set to the MLE for different choices of K and m. When we inspect the error $\text{err}_{K,m} = \log(\lambda_{LIS}(\theta^*, \psi; \omega^*)) - \log(\lambda(\theta^*, \psi))$ as a function of the tuning constants we see that not only does the mean square error decrease with increasing tuning constants but also that there is a considerable systematic error for small K and m (Figure 4 (a)). Looking at the distributions of $\text{err}_{K,m}$, we also see that they are shifted and considerably skewed to the left for small K and m (Figure 4 (b)). This explains why the chains tend to get stuck in some states, namely if $\lambda_{LIS}(\theta, \psi; \omega)$ happens to be an overestimate, then the probability of generating a pair (θ^*, ω^*) with equally high $\lambda_{LIS}(\theta^*, \psi; \omega^*)$ is relatively low. That this tendency to get stuck tends to decrease with increasing tuning constants is reflected in the probability of an extremely small acceptance probability (Figure 4 (d)). As a result of the decreasing variance of $\text{err}_{K,m}$, the mixing improves as measured by the autocorrelation and the acceptance rate (Figure 4 (c)). Note that while $\lambda_{LIS}(\theta, \psi; \omega)$ is unbiased for fixed ψ and θ, the same is not necessarily true for $\log \lambda_{LIS}(\theta, \psi; \omega)$ when θ varies. There seems to be a dramatic drop in the autocorrelations when Km is about halfway between 1 and 2×10^4 (the marker for $K = 1000$, $m = 7$, $Km = 7000$) and 10000 ($K = 2000$, $m = 5$). Summary measures for the posteriors are given in Table 4.

Snijders et al. (2006) proposed to extend the class of Markov models by general-

	MCMCMLE		Bayes		MCMCMLE		Bayes	
	MLE	se	Mean	STD	MLE	s.e.	Mean	STD
Popularity	−6.501	0.727	−6.593	0.725	−6.510	0.637	−6.763	0.650
Main effect								
seniority	1.594	0.324	1.618	0.326	0.855	0.235	0.931	0.252
practice	0.902	0.163	0.910	0.157	0.410	0.118	0.474	0.130
Homophily								
practice	0.879	0.231	0.882	0.236	0.759	0.194	0.751	0.201
sex	1.129	0.349	1.161	0.359	0.704	0.254	0.765	0.277
office	1.654	0.254	1.671	0.249	1.146	0.195	1.211	0.206
Clustering								
GWEPS					0.897	0.304	1.006	0.338
$\log(\lambda)$					0.778	0.215	0.694	0.206

Table 4: Estimates for ERGM fitted to Lazega's (2001) New England Lawyers. Bayes estimates based on LISA with $K = 7000$ and $m = 11$ (dyad independent model) and $K = 3000$ and $m = 7$

ising the dependence graph D to allow for partial dependence among the variables. One of the new statistics they arrived at (the statistics may also be motivated from the point of view of model fit, see Hunter, Goodreau, and Handcock, 2008) was the alternating k-triangle

$$3t_1(x) - \frac{t_2(x)}{\lambda} + \cdots + (-1)^{n-3}\frac{t_{n-2}(x)}{\lambda^{n-3}}$$

which captures multiple clustering, where λ is a positive smoothing parameter and $t_k(x)$ counts the number of pairs of connected nodes that are connected to the same k other actors. Hunter and Handcock (2006) suggested that λ (which is typically arbitrarily chosen by the analyst, Robins & Morris, 2007) should be treated as yet another parameter to be estimated. In order to estimate λ, they reformulated the alternating k-triangle statistic in terms of shared partner statistics and set up the corresponding curved exponential family random graph model (CERGM) with the implied set of canonical parameters. Using well-known properties of curved exponential families (Efron, 1975) they proposed an importance sampler MCMC procedure for calculating

	$K=1$ $m=1$	$K=3000$ $m=5$	$K=7000$ $m=9$
SACF 50 θ_1	0.9554	0.7893	0.6370
SACF 100 θ_1	0.9232	0.6441	0.4314
acceptance rate	0.0462	0.1278	0.3240
ave. min(1,H)	0.0468	0.1274	0.3230
Prop. min(1,H) $\leq e^{-10}$	0.5025	0.1830	0.0248

Table 5: Some MCMC summaries for LISA applied for Model II to Lazega's lawyers

the MLE based on Geyer and Thompson (1992). Using their reparametrisation we fit the above previously dyad-independent, model with the additional shared partner statistics to Lazega's lawyers using LISA. When λ is estimated jointly with the alternating triangle parameter we refer to the latter as the geometrically weighted shared partner statistic (GWESP).

The MCMCMLE was obtained using the algorithm in Hunter and Handcock (2006). The MLE was used as ψ and the proposal covariance matrix set to $.22/\sqrt{1+p}$ times the inverse Fisher information matrix. For the CERGM, CFTPRO cannot be implemented (the algorithm of Corcoran and Tweedie, 2002, applies in principle but is too ineffective to be of practical use) and the time to convergence may vary a great deal depending on the parameter values. Instead of setting the burnin period according to some predetermined fixed value we used a variation on CFTPRO where the distances $|z_k(y^a) - z_k(y^b)|$ are used as indicators of approximate coalescence. This has the interpretation of running parallel chains with over-dispersed initial states employing a deterministic stopping criterion but where the restarts rid the chain of the dependence between the stopping time and the out-put state (details may be had from the authors upon request; in this case, the results are unaltered if a fixed burnin of $100n(n-1)/2$ iterations, as recommended by Snijders, 2002, is used).

The chain run with $K = m = 1$ has high autocorrelation and low acceptance rate (Table 5) primarily because of several runs, the longest of which lasts 30,000 iterations. With tuning constants $K = 7000$ and $m = 9$ the autocorrelation is still

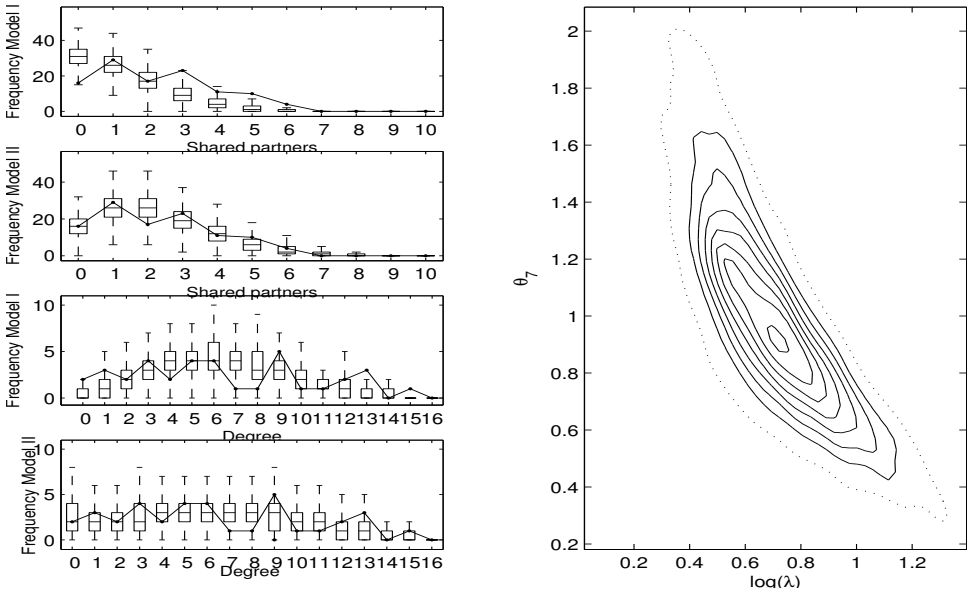

Figure 5: Posterior distributions for ERGM fitted to Lazega's (2001) Lawers. Posterior predictive distributions for the number of shared partners and the degree distribution for dyad-independent model (Model I) and full CERGM (Model II), against observed counts $(\cdot{-})$. On right-hand side the joint posterior distribution of the GWESP parameter (θ_7) and the "smoothing" parameter ($\theta_8 = \log(\lambda)$) with .95 HPD region $(\cdots)$

considerable but the acceptance rate is good. The former can to a large degree be put down to slightly too low proposal variance. The improvement in mixing as the tuning constants are increased is also reflected in the proportion of extremely low acceptance probabilities ("Prop. $\min(1, \mathrm{H}) \leq e^{-10}$"). When half of the proposed moves have a probability of being accepted less than $\exp\{-10\}$, as in the case of the AVM ($K = m = 1$), the chain gets stuck and the acceptance rate becomes low. The acceptance rate and the proportion of extreme proposals is greatly improved for $K = 3000$ and $m = 5$ but it is only for $K = 7000$ and $m = 9$ that we have a performance comparable to that of the simpler Ising models in Section 5.1.

Comparing the MCMCMLE with the corresponding Bayes estimates (Table 4) there is little difference for the first 6 parameters. For the GWESP and $\theta_8 = \log(\lambda)$,

the differences are more noticeable. Part of this has to do with the shape of the joint posterior of θ_7 and θ_8 (Figure 5) and the fact that the modal point in the (marginal) joint posterior clearly is different from the mean.

Because of the relative complexity of CERGMs, simulation of data from models has been proposed as a means of understanding different aspects of the model (Snijders, 2002; Robins, Pattison, and Woolcock, 2005; Hunter et al., 2008). When we have a draw from the posterior distribution of the parameters it is natural to use these parameters to make draws from the posterior predictive distributions. Recall that $y_1^{(\nu_1)}$ in ω is generated from the model defined by the current parameter θ. Hence, marginally the graphs $y_1^{(\nu_1)}$ constitute a sample from the distribution $\int p(y|\theta)\pi(\theta|x)d\theta$, and a draw from the posterior predictive distribution is thus readily available at the termination of LISA.

Some examples of posterior predictive distributions are given in Figure 5. The addition of GWESP improves the fit to the data (in part due to the re-parametrisation, see Hunter and Handcock, 2006) of the distribution of shared partners (Figure 5, top two left panels). Note that while Hunter et al. (2008) proposed using predictive distributions conditioned on the MLEs to asses goodness of fit, using the posterior predictive distributions has the advantage that these take the uncertainty of the parameter estimates into account, effectively marginalising with respect to the parameters.

The importance of the so called degree distribution is something that frequently has been brought up in recent years (Frank and Strauss, 1986; Snijders et al., 2006; Goodreau, 2007) and Hunter et al. (2008) also suggest that one should investigate how well the estimated model reproduces the degree distribution. While the dyad independent model does a reasonable job of reproducing the observed degree distribution, the model that includes GWESP does an even better job as is seen in Figure 5.

6. Concluding remarks

We have proposed an MCMC algorithm, LISA, for Bayesian inference for distributions where the likelihood function is analytically intractable because of a normalising constant. We have furthermore demonstrated its use in the case of three different types of data structures and models. LISA has the dual interpretation that it may either be considered a Metropolis algorithm that employs an importance sampler in each iteration to estimate the acceptance probability or simply a standard Metropolis algorithm on an extended state space. The cardinality of the extended state space may analogously be interpreted in terms of the tuning constants (K and m) that may be set arbitrarily to adjust the mixing of the chain. The auxiliary variable method MCMC introduced by Møller et al. (2006) may be considered a special case of LISA. The principle behind LISA is also surprisingly simple and the algorithm is easy to implement since it mostly only relies on sampling in the data space using MCMC.

As compared to the auxiliary variable method, the increase in computation time is almost negligible as the extra number of iterations due to $K, m > 1$ is small in comparison to the number of iterations needed to generate one realisation $y_1^{(\nu_1)} \in \mathcal{X}$. Compared to approximations such as those used by for example Smith and Fahrmeir (2007) however, the difference is big. The auxiliary distribution may also be talilored to specific models to increase efficiency (Berthelsen and Møller, 2004a,b, 2006, 2007). For ERGMs the computational complexity of ML fitting increases with network size (c.f. Goodreau, 2007; Hunter et al., 2008), mostly to do with the computational burden of generating independent draws from the model, and we would expect this to also be the case for LISA (LISA has thus far been successfully been implemented for directed ERGMs on as many as 106 actors). It is hard however to see how the accuracy of the inference drawn from approximations can be checked in a way other than by using an algorithm like LISA.

LISA opens up the possibility of using arbitrary prior distributions in the analysis to the extent that the performance of the algorithm is not adversely affected by

"bad" prior distributions in similar fashion to how the choice of ψ affects the mixing as illustrated in Figure 3. When applying LISA to an ERGM with latent class Koskinen, Robins and Pattison (2008a) use partially informative prior distributions to control label switching and partial model degeneracy (Handcock, 2002).

Among the future challenges is the question of how well LISA performs for models not investigated here. The approach is generally applicable and should apply with minor modifications to e.g. spatial smoothing for fMRI (Smith and Fahrmeir, 2007) and geographical autocorrelation (Green and Richardson, 2002) as well as CERGMs with more complex dependence structures. Other challenges are developing perfect sampling for in the case where this is not yet available and to develop methods for model selection. For the latter, in addition to the use of posterior predictive distributions a further extension could be to look at e.g. the posterior mean of the likelihood (Dempster, 1974), which requires only the additional computation of $c(\psi)$, or the posterior distribution of the deviance, which requires much future work.

References

Bennett, C. H. (1976), "Efficient Estimation of Free Energy Differences from Monte Carlo Data," *Journal of Computational Physics*, 22, 245–268.

Berthelsen, K. K., and Møller, J. (2007), "Non-parametric Bayesian inference for inhomogeneous Markov point processes," To appear in *Australian and New Zealand Journal of Statistics*.

— (2006), "Bayesian analysis of Markov point processes," in *Case Studies in Spatial Point Process Modeling*, eds. A. Baddeley, P. Gregori, J. Mateu, R. Stoica and D. Stoyan, Springer Lecture Notes in Statistics 185, Springer-Verlag: New York, pp. 85–97.

— (2004a), "A Bayesian MCMC method for point process models with intractable normalising constants," in *Spatial point process modelling and its applications*, eds. A. Baddeley, P. Gregori, J. Mateu, R. Stoica and D. Stoyan, Publicacions de la Universitat Jaume I, 7–15.

— (2004b), "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. 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.

— (1975), "Statistical analysis of non-lattice data," *The Statistician*, 24, 179–195.

Brock, W. A. and Durlauf, S. N. (2001) "Interactions-based Models," in *Handbook of Econometrics*, eds. J.J. Heckman & E.E. Leamer, ed 1, 5, ch 54, Amsterdam: North-Holland, pp. 3297–3380.

Chib, S., and Greenberg, E. (1995), "Understanding the Metropolis Algorithm," *The American Statistician*, 49, 327–335.

Corcoran, J. N., and Tweedie, R. L. (2002), "Perfect Sampling from Independent Metropolis-Hastings Chains," *Journal of Statistical Planning and Inference*, 104, 297–314.

Corander, J. and Dahmström, K. and Dahmström, P. (1998). Maximum likelihood estimation for Markov graphs, Research report, 1998:8, Stockholm University, Department of Statistics.

Corander, J., and Dahmström, K., and Dahmström, P. (2002). Maximum likelihood estimation for exponential random graph model, pp:1-17 in Jan Hagberg (ed.), Contributions to Social Network Analysis, Information Theory, and Other Topics in Statistics; A Festschrift in honour of Ove Frank. University of Stockholm: Department of Statistics.

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

Dempster, A. P. (1974), "The Direct Use of Likelihood for Significance Testing," in *Proceedings of Conference on Foundational Questions in Statistical Inference*, eds. O. Barndorff-Nielsen, P. Blaesild and G. Sihon, Department of Theoretical Statistics: University of Aarhus, pp. 335–352.

Diaconis, P., and Ylvisaker, D., (1979), "Conjugate Priors for Exponential Families," *Ann. Stat.*, 7, 269–281.

Doreian, P. (1982), "Maximum Likelihood Methods for Linear Models," *Sociological Methods and Research*, 10, 243–269.

Ebring, L., and Young, A. A. (1979), "Individuals and Social Structure: Contextual Effects as Endogeneous Feedback," *Sociological Methods and Research*, 7, 396–430.

Efron, B. (1975), "Defining the Curvature of a Statistical Problem (With Applications to Second Order Efficiency)" (with discussion), *The Annals of Statistics*, 3, 1189–1242.

Fienberg, S., and Wasserman, S. S. (1981), "Categorical Data Analysis of Single Sociometric Relations," in *Sociological Methodology*, ed. S. Leinhardt, San Francisco: Jossey-Bass, pp. 156–192.

Frank, O., and Strauss, D. (1986), "Markov Graphs," *Journal of the American Statistical Association*, 81, 832–842.

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

Geyer, C. J., and Thompson, E. (1992), "Constrained Monte Carlo maximum likelihood for dependent data," *Journal of the Royal Statistical Society*, Series B, 54, 657–699.

Goodreau, S. M. (2007), "Advances in exponential random graph (p*) models applied to a large social network," *Social Networks*, 29, 231–248.

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

Handcock, M. S. (2002), "Statistical Models for Social Networks: Inference and Degeneracy," in *Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers*, eds. R. Breiger, K. Carley, and P. E. Pattison, Washington, DC: The National Academies Press, pp. 229–240.

Hastings, W. K. (1970), "Monte Carlo Sampling Methods Using Markov Chains and their Application," *Biometrika*, 57, 97–109.

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.

Hoff, P. D. (2005), "Bilinear Mixed-Effects Models for Dyadic Data," *Journal of the American Statistical Association*, 100, 286 – 295.

Holland, P., and Leinhardt, S. (1981), "An exponential family of probability distributions for directed graphs" (with discussion), *Journal of the American Statistical Association*, 76, 33–65.

Hunter, D. R, Goodreau, S. M., and Handcock, M. S., (2008), "Goodness of Fit of Social Network Models," *Journal of the American Statistical Association*, 103, 248-258.

Hunter, D. R., and Handcock, M. S. (2006), "Inference in Curved Exponential Family Models for Networks," *Journal of Computational and Graphical Statistics*, 15, 565–583.

Jarzynski, C. (1997), "Nonequilibrium Equality for Free Energy Differences," *Physical Review Letters*, 78, 2690–2693.

Jones, F. L., and McMillan, J. (2001), "Scoring Occupational Categories for Social Research: A Review of Current Practice, with Australian Examples," *Work, Employment and Society*, 15, 539–563.

Kass, R. E., Carlin, B. P., Gelman, A., and Neal, R. M. (1998), "Markov Chain Monte Carlo in Practice: A Roundtable Discussion," *The American Statistician*, 52, 93–100.

Koskinen, J., Robins, G., Pattison, P. E., (2008a) "Missing data in social networks: Model-based inference," MelNet Social Networks Laboratory Technical Report 08–03, Department of Psychology, School of Behavioural Science, University of Melbourne, Australia.

— (2008b) "Analysing Exponential Random Graph (p-star) Models with Missing Data Using Bayesian Data Augmentation," MelNet Social Networks Laboratory Technical Report 08–04, Department of Psychology, School of Behavioural Science, University of Melbourne, Australia.

Lazega, E. (2001), *The Collegial Phenomenon: The Social Mechanisms of Cooperation Among Peers in a Corporate Law Partnership*, Oxford: Oxford University Press.

Lusher, D. (2006), *Masculinities in Local Contexts: Structural, Individual and Cultural Interdependencies*, unpublished PhD thesis, Department of Psychology, University of Melbourne, Australia.

Meng, X. L., and Wong, H. W. (1996), "Simulating Ratios of Normalizing Constants Via a Simple Identity: A Theoretical Exploration," *Statistica Sinica*, 6, 831–860.

Moreno, J. L. (1934), *Who Shall Survive? Foundations of Sociometry, Group Psychotherapy and Sociodrama*, Washington, D.C.: Nervous and Mental Disease Publishing Co.

Murray, I., Ghahramani, Z., and MacKay, D. J. C. (2006), "MCMC for doubly-intractable distributions," *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI)*.

Møller, J., Pettitt, A. N., Berthelsen, K. K., and Reeves, R.W. (2006), "An Efficient Markov Chain Monte Carlo Method for Distributions with Intractable Normalising Constants," *Biometrika*, 93, 451 – 458.

Neal, R. M. (1993), "Probabilistic Inference Using Markov Chain Monte Carlo Methods," Technical Report CRG-TR-93-1, University of Toronto, Department of Computer Science. Obtainable from http://www.cs.utoronto.ca/~radford/.

— (2001), "Annealed Importance Sampling," *Statistics and Computing*, 11, 125–139.

— (2005), "Estimating Ratios of Normalizing Constants Using Linked Importance Sampling," Technical Report No. 0511, Department of Statistics, University of Toronto. (available from http://arxiv.org/abs/math.ST/0511216).

Ott, J. (1979), "Maximum Likelihood Estimation by Counting Methods Under Polygenic and Mixed Models in Human Pedigrees," *American Journal of Human Genetics*, 31, 161–175.

Pattison, P. E., Wasserman, S. (1999), "Logit Models and Logistic Regressions for Social Networks: II. Multivariate Relations," *British Journal of Mathematical and Statistical Psychology*, 52, 169–193.

Propp, J., and Wilson, D. (1996), "Exact Sampling with Coupled Markov Chains and Applications to Statistical Mechanics," *Random Structures and Algorithms*, 9, 232-252.

Rasbash J., and Browne W. J. (2002), "Non-Hierarchical Multilevel Models," To appear in *Handbook of Quantitative Multilevel Analysis*, eds. De Leeuw, J. and Kreft, I.G.G., Boston: Kluwer Acad.

Roberts, G. O., Gelman, A., and Gilks, W. R. (1997), "Weak Convergence and Optimal Scaling of Random Walk Metropolis Algorithms," *Annals of Applied Probability*, 7, 1, 110–120.

Robins, G. L., and Morris, M. (2007), "Advances in Exponential Random Graph (p*) Models," *Social Networks*, 29, 169–172 .

Robins, G. L., Pattison, P. E., and Elliot, P. (2001), "Network Models for Social Influence Processes," *Psychometrika*, 66, 161–190.

Robins, G. L., Pattison, P. E., and Wasserman, S. (1999), "Logit Models and Logistic Regressions for Social Networks, III. Valued Relations," *Psychometrika*, 64, 371–394.

Robins, G. L., Pattison, P. E., and Woolcock, J. (2005), "Small and Other Worlds: Global Network Structures from Local Processes," *American Journal of Sociology*, 110, 894–936.

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

Smith, M., and Fahrmeir, L. (2007), "Spatial Bayesian Variable Selection with Application to Functional Magnetic Resonance Imaging," *Journal of the American Statistical Association*, 102, 417–431.

Snijders, T. A. B., (2002), "Markov chain Monte Carlo estimation of exponential random graph models," *Journal of Social Structure*, 3(2), April.

Snijders, T. A. B., and Baerveldt, C., (2003), "A Multilevel Network Study of the Effects of Delinquent Behavior on Friendship Evolution," *Journal of Mathematical Sociology*, 27, 123-151.

Snijders, T. A. B., and Bosker, R. J. (1999), *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, London: Sage Publishers.

Snijders, T. A. B., Pattison, P. E., Robins, G. L., and Handcock, M. S. (2006), "New Specifications for Exponential Random Graph Models," *Sociological Methodology*, 36, 99–153.

Snijders, T. A. B., Steglich, C. E. G., and Schweinberger, M. (2007), "Modeling the Co-evolution of Networks and Behavior," in *Longitudinal Models in the Behavioral and Related Sciences*, eds. K. van Montfort, H. Oud and A. Satorra, Lawrence Erlbaum, pp. 41–71.

Tanner, M. A., and Wong, W.H., (1987), "The calculation of posterior distributions by data augmentation (with discussion)," *Journal of the American Statistical Association*, 82, 528–550.

Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions" (with discussion and a rejoinder by the author), *Annals of Statistics*, 22, 1701–1762.

van Duijn, M. A. J., Gile, K. J., and Handcock, M. S. (2008), "A Framework for the Comparison of Maximum Pseudo Likelihood and Maximum Likelihood Estimation of Exponential Family Random Graph Models," *Social Networks*, Forthcoming.

Van Duijn, M. A. J., Snijders, T. A. B., and Zijlstra, B. H. (2004), "p2 : a Random Effects Model with Covariates for Directed Graphs," *Statistica Neerlandica*, 58, 234–254.

Wasserman, S., Faust, K. (1994), *Social Network Analysis: Methods and Applications*, Cambridge: Cambridge University Press.

Wasserman, S., and Pattison, P. E. (1996), "Logit Models and Logistic Regressions for Social Networks: I. An Introduction to Markov Graphs and p*," *Psychometrika*, 61, 401–425.

Wilson, D. (2000), "How to Couple From the Past Using a Read-once Source of Randomness," *Random Structures and Algorithms*, 16, 85–113.

Wong, G. Y. (1987), "Bayesian Models for Directed Graphs," *Journal of the American Statistical Association*, 82, 140–148.

Zellner, A. (1996), *An Introduction to Bayesian Inference in Econometrics*, New York; Chichester: John Wiley.