NORGES TEKNISK-NATURVITENSKAPELIGE
UNIVERSITET

Approximating Hidden Gaussian Markov Random Fields

by

Håvard Rue, Ingelin Steinsland & Sveinung Erland

PREPRINT
STATISTICS NO. 1/2003

NORWEGIAN UNIVERSITY OF SCIENCE AND
TECHNOLOGY
TRONDHEIM, NORWAY

This report has URL `http://www.math.ntnu.no/preprint/statistics/2003/S1-2003.ps`

Håvard Rue has homepage: `http://www.math.ntnu.no/~hrue`
E-mail: `hrue@math.ntnu.no`
Address: Department of Mathematical Sciences, Norwegian University of Science and Technology, N-7491
Trondheim, Norway.

Approximating Hidden Gaussian Markov Random Fields

Håvard Rue, Ingelin Steinsland & Sveinung Erland
Department of Mathematical Sciences
NTNU, Norway

March 5, 2003

Abstract

This paper discusses how to construct approximations to a unimodal hidden Gaussian Markov random field on a graph of dimension n when the likelihood consists of mutually independent data. We demonstrate that a class of non-Gaussian approximations can be constructed for a wide range of likelihood models. They have the appealing properties that exact samples can be drawn from them, the normalisation constant is computable, and the computational complexity is only $\mathcal{O}(n^2)$ in the spatial case. The non-Gaussian approximations are refined versions of a Gaussian approximation. The latter serves well if the likelihood is near-Gaussian, but it is not sufficiently accurate when the likelihood is not near-Gaussian or if n is large. The accuracy of our approximations can be tuned by intuitive parameters to near any precision.

We apply our approximations in spatial disease mapping and model-based geostatistics models with different likelihoods. We also present procedures for block-updating and construction of Metropolised independence samplers for such models. These sampling schemes are major improvements compared to the single-site schemes commonly used.

KEYWORDS: Block-sampling, Conditional Autoregressive model, Gaussian Markov random field, Hidden Markov Models, Sequential Monte Carlo, Markov chain Monte Carlo, Metropolised Independence Sampler.

ADDRESS FOR CORRESPONDENCE: H. Rue, Department of Mathematical Sciences, The Norwegian University for Science and Technology, N-7491 Trondheim.

E-MAIL: Havard.Rue@math.ntnu.no

WWW-ADDRESS: http://www.math.ntnu.no/~hrue

ACKNOWLEDGEMENTS: The research was funded in part by PhD-grants from the Research Council of Norway and the BeMatA program of the Research Council of Norway.

1 Introduction

Gaussian Markov Random fields (GMRF), or conditional autoregressions, is finite Gaussian fields, where the field has a Markov property; the conditional density for one component given the rest, depends only on its neighbours (Cressie, 1993; Besag and Kooperberg, 1995). In this paper we will focus on spatial GMRFs, although our results also apply to GMRF on general graphs. GMRFs have the convenient property that general fast algorithms based on sparse numerical algebra exist (Rue, 2001). This makes fast sampling and computation of the normalisation constant possible. An important application is the case when the GMRF is observed indirectly through additive Gaussian noise. The conditional density for the hidden GMRF (HGMRF) is still a GMRF, and the same algorithms apply. Conditioning on the observations can also be extended to conditioning on other parameters in the model, as GMRFs are often used as building blocks in spatial models, see for example Heikkinen and Arjas (1998), Wikle et al. (1998), Besag and Higdon (1999) and Fernández and Green (2002).

A Gaussian likelihood is not a requirement for doing inference, since this also can be done using MCMC-methods (Robert and Casella, 1999). Single-site MCMC-algorithms often mix quite slowly in such problems due to the strong interaction within the HGMRF, and in particular between the HGMRF and (some of) its (hyper-)parameters (Knorr-Held and Rue, 2002; Gamerman et al., 2003). MCMC-algorithms performing block-updating on the HGMRF, or better, the HGMRF and its hyper-parameters jointly, may have much better convergence compared to single-site algorithms (Carter and Kohn, 1994; Gamerman, 1998; Knorr-Held and Rue, 2002). Regarding block-updates, it is required that the HGMRF, or an approximation to it, can be sampled exactly. For joint updates, we also need to know the normalisation constant as this is required in the acceptance probability in the MCMC-algorithm. In practise, it is quite hard to draw samples from the HGMRF or construct approximations to it. The only known candidate that we are aware of in the spatial case, is a GMRF approximation (Knorr-Held, 1999; Knorr-Held and Rue, 2002; Rue and Tjelmeland, 2002). This is however not always accurate enough, as it may produce near zero acceptance-rate in MCMC-algorithms when the likelihood is not near-Gaussian or the dimension n is large.

This paper demonstrate how to construct a class of non-Gaussian approximations to a unimodal HGMRF that have the important properties that they can be drawn from and have computable normalisation constants. The approximations are all based on a GMRF approximation. The algorithm for constructing such approximations is fast and of same order as sampling from a GMRF, see Rue (2001). Another advantage is that the same computer code for constructing our approximations can be applied to GMRF-models on general graphs, like models in time or space-time. The computational cost is (most often) $\mathcal{O}(n)$ in time and $\mathcal{O}(n^2)$ in space.

The outline of the paper is as follows. In Section 2 we present some background for the problem considered, and in Section 3 our class of approximations is introduced with discussion of some computational issues. In Section 4, we discuss how to do block-updating and construct Metropolised independence samplers for some models relevant for applications in spatial disease mapping and model-based geostatistics. We conclude with a discussion in Section 5.

2 Background

Let $\mathcal{G}$ be a graph with n nodes, where for example a node denotes a spatial region, a pixel in a lattice or a tile in a tessellation. Two nodes i, j are defined as neighbours, $i \sim j$, if they share a common edge or pixel i is close to pixel j. Let $\mathbf{x}$ denote a zero mean Gaussian Markov Random Field (GMRF) on $\mathcal{G}$, meaning that its $n \times n$ precision matrix (inverse covariance) $\mathbf{Q}$, has the property that $Q_{ij} \neq 0$ iff $i \sim j$ or $i = j$. The Markov properties of $\mathbf{x}$ is given by $\mathcal{G}$ as x_i and x_j are conditionally independent given the rest iff $i \not\sim j$. The precision matrix often depends on further parameters θ, which we denote by $\mathbf{Q}(\theta)$. Define $\mathbf{x}_{i:j}$ as $(x_i, x_{i+1}, \ldots, x_j)^T$.

A HGMRF is a GMRF observed through data $\mathbf{y}$. We assume throughout that the likelihood is such that y_i

only depends on x_i and $\mathbf{y}$ are mutually independent given $\mathbf{x}$, so

$$\pi(\mathbf{y} \mid \mathbf{x}) = \prod_i \pi(y_i \mid x_i).$$

We assume further that $\pi(y_i \mid x_i)$ as a function of x_i, is strictly positive and absolutely continuous wrt the Lebesgue measure, such that the posterior density for the HGMRF is

$$\pi(\mathbf{x}|\theta,\mathbf{y}) \propto \exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{Q}(\theta)\mathbf{x} - \sum_i g_i(x_i,y_i)\right), \tag{1}$$

for some functions $g_i(x_i,y_i)$. In Section 3 we show how to construct approximations to (1) when the it is unimodal. If not, our approximations may still be good if the different modes are close or one of the modes is dominant in terms of probability mass. Otherwise, our approximations are less accurate. A sufficient criteria for (1) to be unimodal, is that $-g_i(x_i,y_i)$ as a function of x_i, is concave for all i.

3 Approximations to a HGMRF

3.1 A GMRF Approximation

Before discussing approximations, we make some assumptions. Note that $\mathbf{Q}$ is a sparse matrix; if each site has a fixed number of neighbours, there are only $\mathcal{O}(n)$ non-zero terms in $\mathbf{Q}$. We assume there exists a permutation of the indices, such that $\mathbf{Q}$ is a band-matrix with a small bandwidth b_w, and that $\mathbf{x}$ is indexed according to this permutation. The motivation for such a permutation, algorithms and further details and motivation, are given in Rue (2001). In the spatial case $b_w = \mathcal{O}(\sqrt{n})$, and computation of the Cholesky factorisation $\mathbf{Q} = \mathbf{L}\mathbf{L}^T$ can then be computed using only $\mathcal{O}(nb_w^2) = \mathcal{O}(n^2)$ operations compared to $\mathcal{O}(n^3)$ in the general case. Note that $\mathbf{L}$ is a lower triangular matrix with the same bandwidth as $\mathbf{Q}$.

A zero mean GMRF $\mathbf{x}$ with precision $\mathbf{Q}$ can be sampled by sampling $\mathbf{z} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$, and then solve $\mathbf{L}^T\mathbf{x} = \mathbf{z}$ (Rue, 2001). If the mean of $\mathbf{x}$ is non-zero, we need to add the mean to $\mathbf{x}$. For those cases where the mean is given implicit by $\mathbf{Q}\boldsymbol{\mu} = \mathbf{b}$, we solve $\mathbf{L}\mathbf{u} = \mathbf{b}$, $\mathbf{L}^T\boldsymbol{\mu} = \mathbf{u}$. The normalisation constant is available from $\mathbf{L}$, since $\log|\mathbf{Q}| = 2\sum_i \log(L_{ii})$.

We want to find an approximation to $\pi(\mathbf{x}|\theta,\mathbf{y})$ in (1). The natural candidate is a GMRF, which can be constructed in the following way. First find the mode $\mathbf{x}^m = \mathbf{x}^m(\theta,\mathbf{y})$ in (1). We assume that $\mathbf{x}^m = \mathbf{0}$, as it simplifies the notation later on. Replace $g_i(x_i,y_i)$ by the Taylor expansion in the mode, $a_i + c_i x_i^2/2$. Our GMRF approximation $\pi_G(\mathbf{x}|\theta,\mathbf{y})$ has precision matrix $\mathbf{Q}_G = \mathbf{Q} + \text{diag}(\mathbf{c})$ and the mode $\mathbf{x}^m$ as mean. Note that $\mathbf{Q}_G$ has bandwidth b_w, the same as $\mathbf{Q}$. Let $\mathbf{L}_G$ be the Cholesky factorisation of $\mathbf{Q}_G$.

The GMRF approximation π_G, can be computed fast, sampled exactly from and the normalisation is known and computable. The approximation does however have a major drawback; we cannot tune the accuracy.

Since $\mathbf{L}_G$ is a lower triangular matrix with bandwidth b_w, a sequential representation of π_G is also directly available by

$$\pi_G(\mathbf{x} \mid \theta,\mathbf{y}) = \prod_{t=n}^{1} \pi_G(x_t \mid \mathbf{x}_{(t+1):n},\theta,\mathbf{y}),$$

where

$$\pi_G(x_t \mid \mathbf{x}_{(t+1):n},\theta,\mathbf{y}) \;=\; \mathcal{N}\left(x_t; \quad -\frac{1}{L_{G,tt}} \sum_{j=t+1}^{\min\{t+b_w,n\}} L_{G,tj}x_j, \quad \frac{1}{L_{G,tt}^2}\right), \tag{2}$$

and $\mathcal{N}(x_t;\mu,\sigma^2)$ is the Gaussian density. This is a non-homogeneous autoregressive process of order b_w defined backward in time. This representation will prove useful in the next section.

3.2 Improved Approximations

When constructing improved approximations based on π_G, note that (1) can be written in the following two ways:

$$\begin{aligned} \pi(\boldsymbol{x} \mid \theta, \boldsymbol{y}) &= \prod_{t=1}^{n} \pi(x_t \mid \boldsymbol{x}_{(t+1):n}, \theta, \boldsymbol{y}) && (3) \\ &\propto \prod_{t=1}^{n} \pi_G(x_t \mid \boldsymbol{x}_{(t+1):n}, \theta, \boldsymbol{y}) \exp\left(-h_t(x_t, y_t)\right) && (4) \end{aligned}$$

where π_G is defined in (2), and

$$h_t(x_t, y_t) = g_t(x_t, y_t) - c_t x_t^2/2.$$

Each of the terms in the sequential representation (3) can be represented by means of (4) as

$$\begin{aligned} \pi(x_t \mid \boldsymbol{x}_{(t+1):n}, \theta, \boldsymbol{y}) &\propto \pi_G(x_t \mid \boldsymbol{x}_{(t+1):n}, \theta, \boldsymbol{y}) \exp\left(-h_t(x_t, y_t)\right) \\ &\times \int \pi_G(\boldsymbol{x}_{1:(t-1)} \mid \boldsymbol{x}_{t:n}, \theta, \boldsymbol{y}) \exp\left(-\sum_{j=1}^{t-1} h_j(x_j, y_j)\right) d\boldsymbol{x}_{1:(t-1)}, && (5) \end{aligned}$$

where all the conditional densities of π_G can easily be derived from (2). This representation has the important property that the mode of the integrand is reasonably close to the mode of $\pi_G(x_{1:(t-1)} \mid \boldsymbol{x}_{t:n}, \theta, \boldsymbol{y})$, since (1) is assumed to be unimodal. In the sequel this enables us to produce accurate, sample-based approximations to the integral as a function of x_t.

If we neglect the dependency of $\boldsymbol{y}$ in π_G and the possible non-boundedness of $-h_i(x_i, y_i)$, the rhs of (4) can be interpreted as the posterior of $\boldsymbol{x}$ with a GMRF prior $\pi_G(\boldsymbol{x})$ and mutually independent observations y_i with log-likelihood $-h_i(x_i, y_i)$. These log-likelihood terms are neglected in the GMRF approximation π_G. Our improved approximations rectify this approximation error.

Our approach is to construct univariate approximations to (5), denoted by $\widetilde{\pi}(x_t \mid \boldsymbol{x}_{(t+1):n}, \theta, \boldsymbol{y})$, and join them together into an approximation to $\pi(\boldsymbol{x} \mid \theta, \boldsymbol{y})$ based on (3):

$$\widetilde{\pi}(\boldsymbol{x} \mid \theta, \boldsymbol{y}) = \prod_{t=1}^{n} \widetilde{\pi}(x_t \mid \boldsymbol{x}_{(t+1):n}, \theta, \boldsymbol{y}). \tag{6}$$

Note that (6) can be sampled sequentially backward in time, and its normalising constant is the product of n univariate normalising constants. We will now discuss how to construct these univariate approximations, by removing what can be considered as less important terms in the rhs of (5).

$A1$) The crudest approximation is to neglect both the $h_t(x_t, y_t)$-term and the integral-term in (5),

$$\widetilde{\pi}_{A1}(x_t \mid \boldsymbol{x}_{(t+1):n}, \theta, \boldsymbol{y}) = \pi_G(x_t \mid \boldsymbol{x}_{(t+1):n}, \theta, \boldsymbol{y}). \tag{7}$$

This gives the GMRF approximation in Section 3.1.

$A2$) A simple, but significant improvement to (7) is to include the $h_t(x_t, y_t)$-term, which can be considered as the second most important term in (5),

$$\widetilde{\pi}_{A2}(x_t \mid \boldsymbol{x}_{(t+1):n}, \theta, \boldsymbol{y}) \propto \pi_G(x_t \mid \boldsymbol{x}_{(t+1):n}, \theta, \boldsymbol{y}) \, \exp\Big(-h_t(x_t, y_t)\Big). \tag{8}$$

Eq. (8) can be well approximated using log-quadratic splines; compute the logarithm of the rhs of (8) for x_t in some evaluation points $\{\check{x}_t\}$, and interpolate using piecewise quadratic polynomials. This spline is easily integrated and can be sampled exactly from by use of the real and complex complementary error-function. We choose the evaluation points $\{\check{x}_t\}$ in the following manner; let $\widetilde{\mu}_t$ be the conditional mean and $\widetilde{\sigma}_t^2$ the conditional variance in the GMRF approximation, then choose $\{\check{x}_t\}$ as the set $\{\widetilde{\mu}_t \pm kf\widetilde{\sigma}_t\}_{k=0}^{K}$ for some fixed factor f and number of knots K. Beyond $\widetilde{\mu}_t \pm Kf\widetilde{\sigma}_t$, we extrapolate log-linearly in order to ensure infinite support and not too light tails.

We can show that $\pi(\mathbf{x}|\mathbf{y},\theta)/\widetilde{\pi}(\mathbf{x}|\mathbf{y},\theta)$ is bounded, if the likelihood $\pi(\mathbf{y}|\mathbf{x})$ is bounded in $\mathbf{x}$ for fixed $\mathbf{y}$, and we replace $\widetilde{\sigma}_t^2$ with $\min\{\widetilde{\sigma}_t^2, S^2\}$ where S^2 is a fixed but finite constant. This is the case for all our examples in Section 4 and is needed for geometrically ergodicity of the Metropolised independence sampler.

The improved approximation (8) can be a significant improvement to the GMRF approximation. Assume $\mathbf{Q} = \kappa\mathbf{P}$, for a scalar κ. As $\kappa \to 0$, the likelihood dominates in (1) and the GMRF approximation can be quite poor. (This is illustrated in Figure 2 in Section 4.2.) The error of the approximation (8) depends almost only on how accurate the log-spline representation is.

$A3)$ In all further improvements to (8), we include the integral term in (5) which may be written as

$$\mathrm{E}\Big[\exp\Big(-\sum_{j=1}^{t-1} h_j(x_j, y_j)\Big)\Big], \tag{9}$$

where the expectation is wrt $\pi_G(\mathbf{x}_{1:(t-1)} \mid \mathbf{x}_{t:n}, \theta, \mathbf{y})$. We need estimates of (9) as a function of x_t, but only for $x_t \in \{\check{x}_t\}$, ie. in the $2K+1$ points the log-spline approximation is based on. As $\pi_G(\mathbf{x}|\theta,\mathbf{y})$ is a GMRF, we expect neighbouring sites to be most correlated. Hence, as a function of x_t, we expect important terms in (9) to be those j's that are neighbours to t smaller than t, or have a common neighbour and so on. Let $\mathcal{J}(t)$ be the set of sites which we want to include in our approximation to (9). We estimate this approximation using the average computed from M samples from $\pi_G(\mathbf{x}_{1:(t-1)} \mid \mathbf{x}_{t:n}, \theta, \mathbf{y})$. Our estimate of (9) is up to a multiplicative constant,

$$\frac{1}{M}\sum_{i=1}^{M} \exp\Big(-\sum_{j\in\mathcal{J}(t))} h_j(x_j^i, y_j)\Big). \tag{10}$$

Here, x_j^i, is the j'th component of the i'th sample from π_G, which is obtained by successively using (2) from time t until $\min_t \mathcal{J}(t)$. If (1) is not unimodal, the estimate (10) will be less accurate.

The computation of (10) is potentially quite costly and must therefore be done carefully. If $\mathcal{J}(t)$ are those neighbours to t smaller than t, we need $\mathcal{O}(\sqrt{n})$ evaluations of (2), each containing $\mathcal{O}(\sqrt{n})$ terms in the sum. Repeating all n nodes requires $\mathcal{O}(2KMn^2)$ operations. This is the same order as factorising $\mathbf{Q}_G$. Two adjustments reduce this cost to $\mathcal{O}(Mn^2)$. First, note that the conditional mean in π_G is linear in x_t and the conditional covariance does not depend on x_t. Secondly, use the same stream of random numbers to make (10) continuous wrt x_t. The estimation of (10) is then done as follows. Compute the conditional mean for $j \in \mathcal{J}(t)$ using (2) for two values of x_t. The conditional mean for all other values of x_t is a linear combination of these two. Sample M independent samples with zero mean, then and add the conditional mean, depending on x_t, to it.

Additionally, we make use of antithetic ideas which provide three extra samples (for each of our M samples) for free (Durbin and Koopman, 1997); Let $\boldsymbol{\nu}$ be a sample from a zero mean Gaussian vector, u a sample from uniform$(0,1)$, f_1 the u-quantile in a χ_n-distribution and f_2 the $1-u$ quantile, then $\pm f_1\boldsymbol{\nu}/\sqrt{\boldsymbol{\nu}^T\boldsymbol{\nu}}$ and $\pm f_2\boldsymbol{\nu}/\sqrt{\boldsymbol{\nu}^T\boldsymbol{\nu}}$ have the correct distribution.

An improvement to (10) is to use $A2$ as the sampling distribution instead of π_G. This require some obvious changes in (9) and (10) and costs $\mathcal{O}(2KMn^2)$ operations. We do not discuss this option further.

The approximation (6) is indexed by the sequence of random numbers used in (10), and by keeping this sequence fixed we can produce several samples from the same approximation.

We have implemented the algorithm in `C` as a part of the open source `GMRFLib`-library (Rue and Follestad, 2002), which is available from the first author's homepage. The algorithm is written for general graphs and great effort was made to make the algorithm run efficiently.

4 Examples

We now demonstrate our new approximations on three spatial models with different likelihoods, showing how to do joint updating and construct Metropolised independence samplers for such models. The first is motivated from a Bayesian model for mapping of disease (Besag et al., 1991; Mollié, 1996), while the other two are model-based geostatistical models (Diggle et al., 1998). In the last example an additional feature is introduced to construct approximations.

4.1 Bayesian mapping of disease

A spatial region (land or part of it) is divided into n contiguous areas labelled $i = 1, \dots, n$. In each area we observe y_i; the number of deaths from the disease of interest during the study period. When the disease is non-contagious and rare, we assume that the deaths in each area are mutually independent and Poisson distributed with mean $e_i \exp(x_i)$. Here, e_i is the known "expected" counts assuming constant risk for all areas, and x_i the log-relative risk. To estimate $\mathbf{x}$, we borrow strength from spatial neighbouring areas and assuming an intrinsic GMRF model for $\mathbf{x}$, defining area i to be a neighbour of j, $i \sim j$, if they share a border. The full posterior reads

$$\pi(\mathbf{x}, \kappa \mid \mathbf{y}) \propto \kappa^{(n-1)/2} \exp\left(-\frac{\kappa}{2} \sum_{i \sim j} (x_i - x_j)^2 - \sum_i (e_i \exp(x_i) - y_i x_i) \right) \pi(\kappa) \qquad (11)$$

where κ is the precision of the GMRF prior, with prior $\pi(\kappa)$. The full conditional of $\mathbf{x}$ is on the form (1), with $Q_{ij} = -\kappa$ if $i \sim j$, Q_{ii} is κ times the number of neighbours of i, and $g_i(x_i, y_i) = e_i \exp(x_i) - y_i x_i$.

The model known as the BYM-model (Besag et al., 1991) also include an additional unstructured heterogeneity term in the log-relative risk. This term should always be included in (11) when applied to data. We ignore it here only for the purpose of avoiding unnecessary complications illustrating our approximations. We will illustrate our approximations on some data on oral cavity cancer mortality for males in Germany ($1986 - 1990$), analysed by Knorr-Held and Raßer (2000) and shown in Figure 1.

4.2 Approximating $\pi(\mathbf{x}|\kappa, \mathbf{y})$

We will now demonstrate how various improved approximations compare to the GMRF approximation when κ is fixed. We construct various approximations for $\kappa = 0.1, 1$ and 10. These choices correspond to very small, small and reasonable values of κ, which will become apparent in Section 4.3. For each of these values of κ, we construct four different approximations to $\pi(\mathbf{x}|\kappa, \mathbf{y})$: $A1$ is the GMRF approximation, $A2$ the one including only the likelihood term (5), $A3a$ the one including also (10) with $\mathcal{J}(t)$ as the neighbours to node t less than t, using $M = 1$, and $A3b$ the same as $A3a)$ but with $M = 100$. Approximations $A3a$ and $A3b$ also use extra antithetic variables for each sample, as described in Section 3.2. We use $K = 20$ knots and $f = 6$ in the log-spline approximation.

The accuracy of the approximations is measured by the accept-rate using the approximation in a Metropolised independence sampler for $\mathbf{x}$. This is advocated by Robert and Casella (1999, Section 6.4.1), but they also note that the expected accept-rate does not give any upper bound on $\sup_{\mathbf{x}} \pi(\mathbf{x}|\kappa, \mathbf{y}) / \widetilde{\pi}(\mathbf{x}|\kappa, \mathbf{y})$, which controls the convergence of the algorithm.

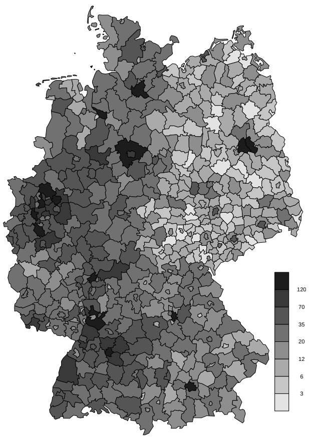


Figure 1: The map of Germany with $n = 544$ regions displaying the number of oral cavity cancer cases in each region males in the period $1986-1990$. The data has 1st quantile 9, median 19, mean 28 ad 3rd quantile 33. The graph has average 5.2, minimum 1 and maximum 11 neighbours.

Average accept-rate	$\kappa = 0.1$	$\kappa = 1$	$\kappa = 10$
Approximation $A1$	0.01	0.11	0.47
Approximation $A2$	0.94	0.80	0.78
Approximation $A3a$	0.96	0.87	0.86
Approximation $A3b$	0.99	0.96	0.90

Table 1: The average accept-rate for four different approximations; The GMRF ($A1$) and improved ones ($A2$, $A3a$ and $A3b$).

Table 1 displays an estimate of the accept-rate for the four approximations averaged over $1\,000$ iterations. For $A3a$ and $A3b$, we use different random numbers to generate each of the $1\,000$ approximations, hence we average over that source of randomness as well. The results obtained are quite typical. When κ is small, $\pi(\boldsymbol{x}|\kappa, \boldsymbol{y})$ is dominated by the non-Gaussian likelihood, and the accept-rate for $A1$ will decrease for decreasing κ. This is illustrated in Figure 2 which shows the joint posterior for $n = 2$, $\boldsymbol{y} = (1, 0)$, $\boldsymbol{e} = (3, 5)$, for $\kappa = 0.1$ and $\kappa = 1$.

The inclusion of the likelihood term in $A2$, raise the accept-rate from 0.01 to 0.94. For increasing κ, $A1$ becomes better, while $A2$ have a slight decrease in the accept-rate. This is due to the increase of the relative influence of the GMRF prior. Approximation $A3a$ and $A3b$ demonstrate further improvements, by accounting for the spatial dependency in addition to the likelihood by including (10). Increasing the number of samples

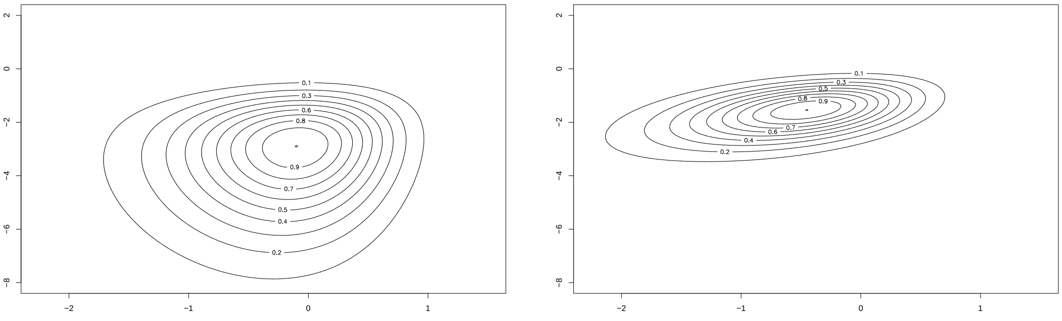


Figure 2: The contour-plot of $\pi(\boldsymbol{x}|\kappa, \boldsymbol{y})$ when $n = 2$, $\boldsymbol{y} = (1, 0)$, $\boldsymbol{e} = (3, 5)$, for $\kappa = 0.1$ (left) and $\kappa = 1$ (right). Increasing κ makes the density closer to the Gaussian.

from 1 to 100 improves the approximations. However, the improvement of $A3a$ and $A3b$ over $A2$, is less than how much $A2$ improve over $A1$. The higher the accept-rate, the harder it seems to improve the approximation. For increasing κ, the accept-rate for all approximations eventually tends to one.

The computational cost in this example on a $1\,200$MHz laptop is 0.06 seconds pr iteration for $A1$, while $A2$, $A3a$ and $A3b$ require 10, 30 and 1900 times this, respectively. Each iteration requires the construction of two approximations and two optimisations. The computational efficiency obtained by $A2$ and $A3a$ compared to $A1$, is to us quite impressive.

Although this example is typical, it does not demonstrate the effect of the parameters controlling the approximation. Our experience is as follows. A higher number of knots K generally improves the approximation and most notably when the accept-rate is high. In most cases 10 to 20 knots are sufficient. The inclusion of the likelihood-term in (8) can give a huge improvement compared to the GMRF approximation. Correcting using (10) generally helps, but is less important compared to the likelihood. If $A2$ gives too low accept-rate (10) is required. Computing (10) can be expensive, as demonstrated in this example. We have good experience using only one sample ($M = 1$) in (10), and letting this be the conditional mean (computed under the GMRF approximation) or mode. This correction usually gives a positive influence on the accept-rate while further improvements require relative much more computing. We generally recommend using $\mathcal{J}(t)$ as the neighbours of node t less than t, but speed-up can be gained if $\mathcal{J}(t) = \{t-1, t-2\}$, say, is sufficient to obtain a reasonable accept-rate. The computational cost is $\mathcal{O}(n^2)$ for the first choice, but only $\mathcal{O}(n^{3/2})$ for the second one.

It is our experience that parameters can be selected to fit the application in hand and tuned to near any required accept-rate. The cost however, can be relatively high if we require an accept-rate close to 1, while cheap approximations can produce a reasonable accept-rate and can give significant improvements compared to the GMRF approximation.

4.3 Approximating $\pi(x, \kappa|y)$

This section demonstrate how our approximations to $\pi(x|\kappa, y)$ can be used to construct a Metropolised independence sampler for x and κ, jointly. We do this by constructing an approximation to $\widetilde{\pi}(\kappa|y)$, and then combine it with $\widetilde{\pi}(x|\kappa, y)$. We start by stating the seemingly obvious,

$$\pi(\kappa \mid y) = \frac{\pi(x, \kappa \mid y)}{\pi(x \mid \kappa, y)}, \tag{12}$$

which is valid for any x such that the denominator is non-zero, see also Besag (1989). The implication of (12) is that we can replace integration over x in $\pi(x, \kappa \mid y)$ with conditioning. The commonly used Laplace approximation for integration (Tierney et al., 1989), is the same as constructing a Gaussian approximation to the denominator in our case. Let $\widetilde{\pi}(x|\kappa, y, x^m(\kappa))$ be an approximation to $\pi(x|\kappa, y)$ around the mode $x^m(\kappa)$. A natural candidate for an approximation to $\pi(\kappa \mid y)$, is

$$\widetilde{\pi}(\kappa \mid y) \propto \left. \frac{\pi(\kappa)\pi(x|\kappa)\pi(y|x)}{\widetilde{\pi}(x|\kappa, y, x^m(\kappa))} \right|_{x = x^m(\kappa)} \tag{13}$$

An important ingredient in (13) is the κ-dependent computable normalisation constant in the denominator. The rhs is evaluated in x^m, the point we think gives the most accurate result, following Tierney et al. (1989). We fix the random numbers used in the approximation to make the denominator continuous wrt κ. A Metropolised independence sampler can now be constructed, by sampling κ from a log-quadratic spline approximation to (13) and then sampling x from $\widetilde{\pi}(x|\kappa, y, x^m(\kappa))$.

Figure 3 shows the estimated posterior marginal for κ for a $\Gamma(0.0001, 0.0001)$-prior using three of the four approximations in Section 4.2. The three approximations for the marginal appear as one curve. This contrasts the accept-rate in Table 1, which vary with κ and which approximation is used. The interpretation is that

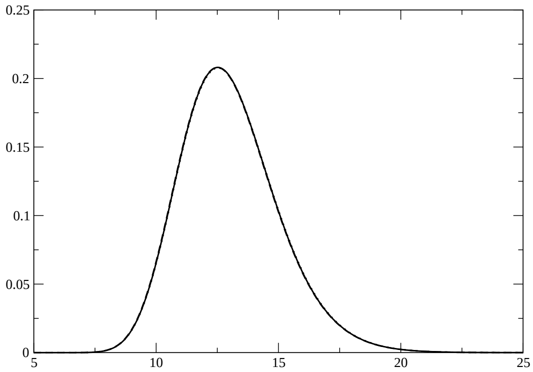


Figure 3: The estimated posterior marginal density for κ, computed using (13) and $A1$, $A2$ and $A3a$. The three estimates nearly coincide.

the denominator in (13) have about the same functional form of κ (evaluated in $\mathbf{x}^m(\kappa)$) for the different approximations. The constant of proportionality of this function will cancel when normalising (13).

A Metropolised independence sampler using $A1$, $A2$ and $A3a$, gave an accept-rate of 0.43, 0.82 and 0.86 averaged over $1\,000$ iterations, respectively. The auto-correlation for κ at lag $k \geq 0$ is approximately $(1-\alpha)^k$ where α is the average accept-rate. Hence, the sampler seems to converge quite fast for all three approximations.

The delayed rejection algorithm (Mira, 2001) could also have been used here, using $A1$ to sample the first proposal, and then use $A2$, say, if the first proposal is rejected. There is no extra cost involved as the GMRF approximation is needed in any case.

To get more insight into the convergence of the Metropolised independence sampler in this example, we will use the empirical supremum rejection sampler as introduced by Caffo et al. (2002). Their algorithm is the standard rejection sampler, but where the supremum of $\pi(\mathbf{x}, \kappa|\mathbf{y})/\widetilde{\pi}(\mathbf{x}, \kappa|\mathbf{y})$ is replaced with the largest value observed so far. Let C_m denote this quantity after m trials. They study the convergence rate of C_m as $m \to \infty$, and based on this argue that we can treat the output of this algorithm as random samples from the target when the samples are used to estimate expectations wrt to it. We applied their algorithm, and estimated C to be 25.0, 1.47 and 1.39 for the joint approximation based on $A1$, $A2$ and $A3a$, after $1\,000$ iterations. We also ran the one based on $A3a$ for a very long time, with virtually no change in the estimated C. Although these estimates are surely somewhat optimistic, they give anyway an estimate of the accuracy of the approximations in the most important areas and the ability to produce exact samples in these areas. The behaviour of the approximations in areas with low probability are always more questionable. If we believe in the estimated C's, we can sample exactly from the joint posterior using rejection sampling.

The convergence of the Metropolised independence sampler in total variation norm, is bounded by $(1 - 1/C)^{\#\text{iterations}}$ (Mengersen and Tweedie, 1996). Comparing this bound with our estimated values of C, we note that Approximation 2 is about 3 times more efficient compared to the GMRF one, taking the computation cost into account.

4.4 Model-based Geostatistics

Diggle et al. (1998) discuss Bayesian models which combine traditional geostatistical methods with those of generalised linear models. The common setting is a spatial Gaussian field with some unknown parameters θ (mean, precision and correlation-length) which is observed at some locations with a non-Gaussian likelihood. The goal is to estimate θ and estimate the Gaussian field.

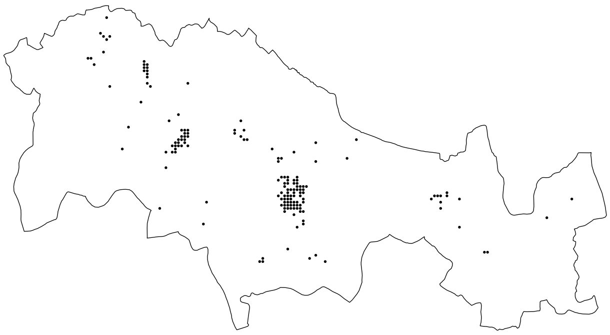


Figure 4: The outline of the campylobacter, salmonella and cryptosporidia infections data in north Lancaster (Diggle et al., 1998). Each point is the location (given as post-codes) of the enteric outbreak.

4.4.1 Binomial likelihood

Consider the following example taken from Diggle et al. (1998). Figure 4 shows the position of reported outbreaks of campylobacter, salmonellae and cryptosporidia in north Lancaster (UK) between April and December 1994. Two or more persons can obtain the disease from the same source, and infections reported at the same location in a five-day period is considered as the same outbreak. The data consists of 399 outbreaks in 236 different locations where 234 of them are campylobacter. The problem considered is to estimate a spatial latent surface measuring the risk that an outbreak is campylobacter. The data comes in triplets (l_i, n_i, y_i), $i = 1, \ldots, 236$ where l_i is the location, n_i the number of enteric infections and y_i the number of them being campylobacter. The probability of an enteric outbreak at position l_i in the binomial likelihood, p_{l_i}, is linked to the spatial field by $\text{logit}(p_{l_i}) = x_{l_i}$.

Diggle et al. (1998) analysed this model using single-site MCMC algorithms. There are reasons to believe that such an approach encounters severe problems in mixing between θ and x, at least for cases with more data. Although the number of lattice points (n) covering the region of interest is large, the number of data is small. We will now demonstrate how our approximations can be used to construct a joint approximation for the spatial field and its hyperparameters following the approach in Section 4.1. This joint approximation can then be applied as a Metropolised independence sampler or used in an empirical supremum rejection sampler to estimate expectations.

We follow Diggle et al. (1998) and use for the isotropic spatial Gaussian field, an exponential correlation function with unknown precision τ, range r (in pixels) and common mean μ. Our modification is to use GMRF proxies for the Gaussian field on a fine 200×100 lattice covering the region of interest, introduced by Rue and Tjelmeland (2002). Hence, we use a GMRF, x, with a 5×5 neighbourhood and coefficients as computed by their method, for a finite set of ranges with step of 0.05. This reduces the computational cost with a factor n, when predictions for non-observed locations as well as parameter estimates for θ are required. If only θ are of interest, we may use only the set of sites where we have observed data, but this option is not considered here.

Figure 5 shows the scaled marginal likelihood for $(\log(\tau), r) \in [-2, 7] \times [0, 50]$ where $\mu = 0.35$, the empirical mean from the data. Here, we used (13) with an obvious correction and $A2$ as described in Section 4.1. Using $A1$ gave similar results. Each evaluation in the grid of selected $(\log(\tau), r)$ values, required about 30 seconds of computing. The marginal likelihood is quite flat in a huge region demonstrating a small content of information in the data regarding $(\log\tau, r)$. We could have included μ in our "x" by giving it a Gaussian prior at essentially no extra cost (Rue, 2001, Appendix), but our implementation does not support this option at the time of the writing.

Based on the marginal likelihood in Figure 5, we can construct a log-spline approximation to the marginal density of θ and then construct a Metropolised independence sampler as in Section 4.4. Here, using a triangu-

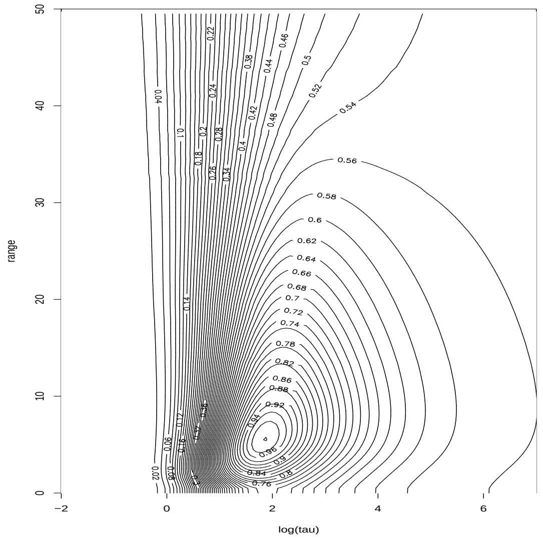


Figure 5: The marginal likelihood of $(\log(\tau), r) \in [-2,7] \times [0,50]$ for fixed $\mu = 0.35$ for the example in Section 4.4. The plot is scaled to have maximum value equal one.

larisation of the area of interest and log-linear splines within each triangle, is perhaps the simplest choice. We easily get an accept-rate exceeding 40% all depending on how well we tune the approximation. An alternative, is to do a joint (log-)random-walk proposal on the hyperparameters, and conditionally on these values sample the spatial field (Knorr-Held and Rue, 2002).

We also investigate the case where similar data to the observed ones, is added to all pixels in the 200×100 lattice. There are no problems constructing approximations for the spatial-field with accept-rate above 50%. It requires about one minute to construct the GMRF approximation and slightly more for improved ones. As long as the likelihood is reasonably close to a Gaussian, good enough approximations seem easy to construct.

4.4.2 Double exponential likelihood

A more serious challenge is motivated from one of the examples in Dryden et al. (2002), where the likelihood is double exponential, ie

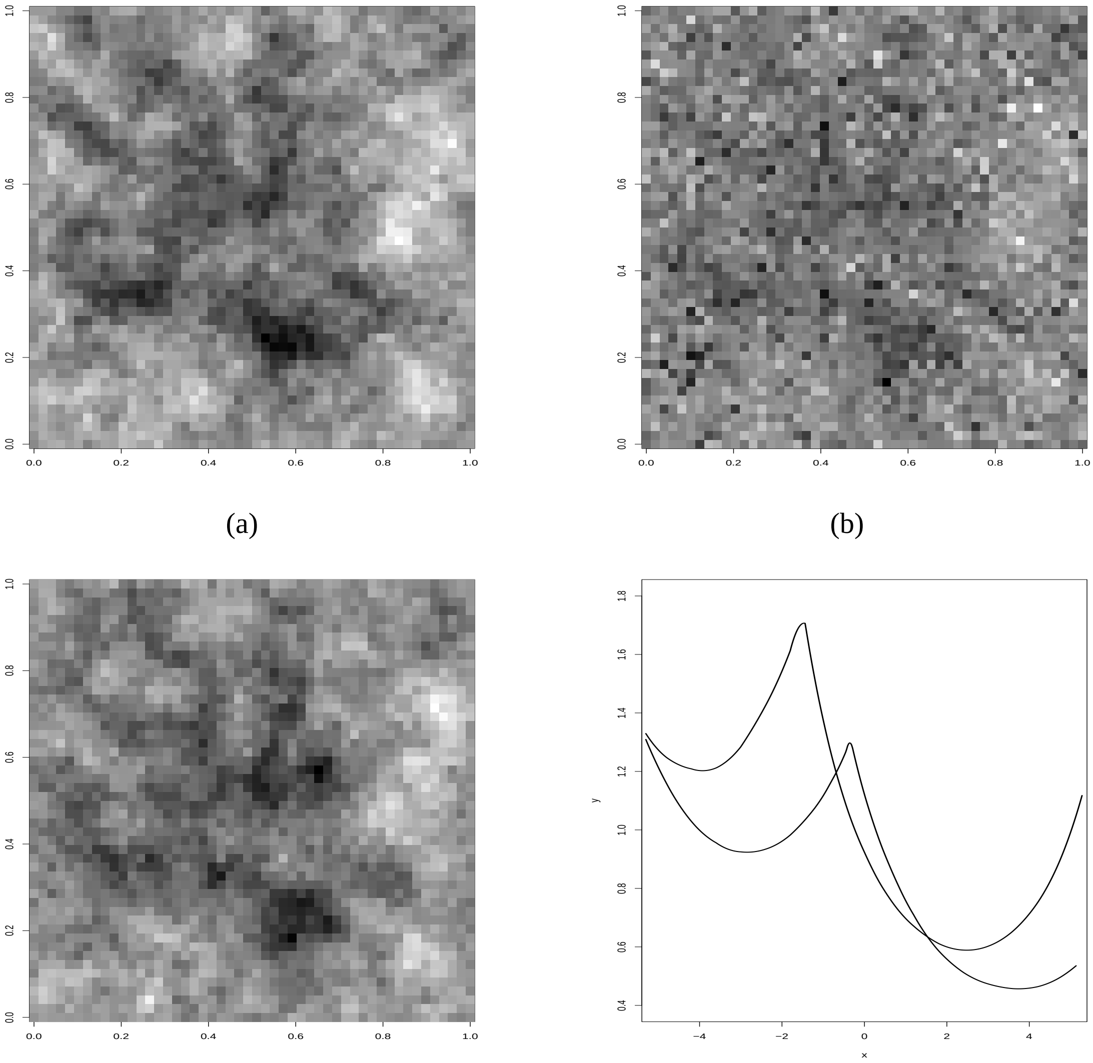
$$\pi(y_i \mid x_i) \propto \exp(-\mid x_i - y_i \mid). \tag{14}$$

This makes (1) strongly non-Gaussian. We note in passing, that the marginal likelihood for θ computed with our approximations is an alternative to the asymptotic motivated approximations studied by Dryden et al. (2002).

We sample a true spatial field on a 50×50 lattice with exponential correction function with range one third of the horizontal length of the lattice, unit precision and zero mean. We then add independent noise according to (14). Only the spatial field is treated as unknown in this example. The parameters selected, balance the likelihood and the prior and makes the construction of good approximations harder.

Two practical problems arise due to the likelihood (14). The second derivative of the log-likelihood is zero hence locating the mode is hard(er). For the same reason, the GMRF approximation constructed using Taylor-expansion is quite poor. Both these problems are solved using the approximation idea in Rue (2001); the Taylor-expansion is replaced by a quadric expansion fitting the log-likelihood more accurately over a larger range around $\mathbf{x}^m$, say in a range ± 2 of x_i^m for all i.

Figure 6 shows the true field in (a) and the data in (b). We use the same parameters as in A3a but with $M = 10$ samples. Increasing M was needed to get a reasonable accept-rate of about 30%, where the spatial field $\mathbf{x}$



(a) (b)

(c) (d)

Figure 6: Image (a) shows the true spatial field, (b) the observed data, (c) a sample from the Metropolised independence sampler, and (d) two examples of the ratio of the estimated conditional in (5) and the conditional density using the Gaussian approximation, for the centre pixel in the image. In (d) the horizontal axis is standardised with the conditional mean and standard deviation of the Gaussian approximation.

was sampled using a Metropolised independence sampler, at the cost of 50 seconds/iteration. Figure 6(c) shows a sample from the Metropolised independence sampler. Using $A1$ gave essentially zero accept-rate. The reason is displayed in (d), showing two examples of the ratio of the estimated conditional in (5) and the conditional density using the Gaussian approximation, for the centre pixel in the image. The horizontal axis is standardised with the conditional mean and standard deviation of the Gaussian approximation. The conditional density is skewed and the mode is slightly shifted. It is obvious that $A1$ cannot be sufficiently accurate in this case.

We now increase the lattice to 100×100. Quite accurate approximations is needed for the Metropolised independence sampler to produce an accept-rate above zero. About 8 minutes of computing for each iteration is needed to produce an accept-rate of about 30%. However, it is encouraging that computing seems to be the practical limit, not our approach to construct approximations.

5 Discussion

In this paper we have presented an approach to construct approximations to a unimodal hidden Gaussian Markov Random field (HGMRF) on general graphs, which can be sampled exactly from and have computable normalisation constants. The examples have demonstrated how to construct joint updates and Metropolised independence samplers for spatial models. Such sampling schemes are major improvements compared to the single-site schemes commonly used. Our approach can also be applied when the precision matrix is full, but the computational cost is then $\mathcal{O}(n^3)$.

Another interesting case is GMRF models in time. Here, the cost is only $\mathcal{O}(n)$. As our method and computer code are valid also in this case, we have experimented also with such models with various kinds of observation models. Good approximations are much easier to construct compared to spatial ones. For GMRF models in time, or dynamic models in general, there exists an extensive literature on sequential Monte Carlo methods, see Doucet et al. (2001) for an overview. These methods can also be used to construct Metropolised independence samplers (although Gaussian approximations are often used, see Durbin and Koopman (2000) and Shephard and Pitt (1997)), and to analyse non-dynamic models (Chopin, 2002), but the dynamic nature of these models makes it more natural to focus on filtering and prediction. Our approach have some similarities with these methods, but we do not rely on the forward-filtering backward-sampling recursions that are inherent in sequential Monte Carlo. This recursion requires densities of dimension b_w to be approximated. This is hard for $b_w > 3$, say, but our approach works fine even for $b_w = \mathcal{O}(\sqrt{n})$ and also for HGMRF models in general where there is no natural time-ordering of the GMRF, as is the case for spatial GMRF models.

References

Besag, J. (1989). A candidate's formula: A curious result in Bayesian prediction. *Biometrika*, 76(1):183.

Besag, J. and Higdon, D. (1999). Bayesian analysis of agricultural field experiments (with discussion). *Journal of the Royal Statistical Society, Series B*, 61(4):691–746.

Besag, J. and Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika*, 82(4):733–746.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics*, 43(1):1–59.

Caffo, B. S., Booth, J. G., and Davison, A. C. (2002). Empirical supremum rejection sampling. *Biometrika*, 89(4):745–754.

Carter, C. K. and Kohn, R. (1994). On Gibbs sampling for state space models. *Biometrika*, 81(3):541–543.

Chopin, N. (2002). A sequential particle filter method for static models. *Biometrika*, 89(3):539–552.

Cressie, N. A. C. (1993). *Statistics for spatial data*. John Wiley, New York, 2nd edition.

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998). Model-based geostatistics (with discussion). *Journal of the Royal Statistical Society, Series C*, 47(3):299–350.

Doucet, A., de Freitas, N., and Gordon, N., editors (2001). *Sequential Monte Carlo methods in practice*. Statistics for Engineering and Information Science. Springer-Verlag, New York.

Dryden, I. L., Ippoliti, L., and Romagnoli, L. (2002). Adjusted maximum likelihood and pseudo-likelihood estimation for noisy Gaussian Markov random fields. *Journal of Computational and Graphical Statistics*, 11:370–388.

Durbin, J. and Koopman, S. J. (1997). Monte carlo maximum likelihood estimation for non-Gaussian state space models. *Biometrika*, 84(3):669–684.

Durbin, J. and Koopman, S. J. (2000). Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives. *Journal of the Royal Statistical Society, Series B*, 62(1):3–56. With discussion and a reply by the authors.

Fernández, C. and Green, P. J. (2002). Modelling spatially correlated data via mixtures: a Bayesian approach. *Journal of the Royal Statistical Society, Series B*, 64(4):805–826.

Gamerman, D. (1998). Markov chain Monte Carlo for dynamic generalised liner models. *Biometrika*, 85(1):215–227.

Gamerman, D., Moreira, A. R. B., and Rue, H. (2003). Space-varying regression models: specifications and simulations. *Computational Statistics and Data Analysis*, xx(xx):xx–xx. (to appear in Spessial Issue on Computational Economics.

Heikkinen, J. and Arjas, E. (1998). Non-parametric Bayesian estimation of a spatial Poisson intensity. *Scandinavian Journal of Statistics*, 25(3):435–450.

Knorr-Held, L. (1999). Conditional prior proposals in dynamic models. *Scandinavian Journal of Statistics*, 26(1):129–144.

Knorr-Held, L. and Raßer, G. (2000). Bayesian detection of clusters and discontinuities in disease maps. *Biometrics*, 56:13–21.

Knorr-Held, L. and Rue, H. (2002). On block updating in Markov random field models for disease mapping. *Scandinavian Journal of Statistics*, 29(4):597–614.

Mengersen, K. L. and Tweedie, R. L. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *The Annals of Statistics*, 24(1):101–121.

Mira, A. (2001). On Metropolis-Hastings algorithms with delayed rejection. *Metron*, 59:231–241.

Mollié, A. (1996). Bayesian mapping of disease. In Gilks, W. R., Richardson, S., and Spiegelhalter, D. J., editors, *Markov Chain Monte Carlo in Practice*, pages 359–379. London: Chapman & Hall.

Robert, C. P. and Casella, G. (1999). *Monte Carlo statistical methods*. Springer-Verlag New York.

Rue, H. (2001). Fast sampling of Gaussian Markov random fields. *Journal of the Royal Statistical Society, Series B*, 63(2):325–338.

Rue, H. and Follestad, T. (2002). GMRFLib: a C-library for fast and exact simulation of Gaussian Markov random fields. Statistics report no. 1, Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway.

Rue, H. and Tjelmeland, H. (2002). Fitting Gaussian Markov random fields to Gaussian fields. *Scandinavian Journal of Statistics*, 29(1):31–50.

Shephard, N. and Pitt, M. K. (1997). Likelihood analysis of non-Gaussian measurement time series. *Biometrika*, 84(3):653–667.

Tierney, L., Kass, R. E., and Kadane, J. B. (1989). Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *Journal of the American Statistical Association*, 84(407):710–716.

Wikle, C. K., Berliner, L. M., and Cressie, N. A. (1998). Hierarchical Bayesian space-time models. *Environmental and Ecological Statistics*, 5(2):117–154.