J. R. Statist. Soc. B (2004)
66, *Part* 4, *pp.* 877–892

Approximating hidden Gaussian Markov random fields

Håvard Rue, Ingelin Steinsland and Sveinung Erland

Norwegian University of Science and Technology, Trondheim, Norway

[Received April 2003. Revised March 2004]

Summary. Gaussian Markov random-field (GMRF) models are frequently used in a wide variety of applications. In most cases parts of the GMRF are observed through mutually independent data; hence the full conditional of the GMRF, a hidden GMRF (HGMRF), is of interest. We are concerned with the case where the likelihood is non-Gaussian, leading to non-Gaussian HGMRF models. Several researchers have constructed block sampling Markov chain Monte Carlo schemes based on approximations of the HGMRF by a GMRF, using a second-order expansion of the log-density at or near the mode. This is possible as the GMRF approximation can be sampled exactly with a known normalizing constant. The Markov property of the GMRF approximation yields computational efficiency. The main contribution in the paper is to go beyond the GMRF approximation and to construct a class of non-Gaussian approximations which adapt automatically to the particular HGMRF that is under study. The accuracy can be tuned by intuitive parameters to nearly any precision. These non-Gaussian approximations share the same computational complexity as those which are based on GMRFs and can be sampled exactly with computable normalizing constants. We apply our approximations in spatial disease mapping and model-based geostatistical models with different likelihoods, obtain procedures for block updating and construct Metropolized independence samplers.

Keywords: Block sampling; Conditional autoregressive model; Gaussian Markov random field; Hidden Markov models; Markov chain Monte Carlo methods; Metropolized independence sampler; Sequential Monte Carlo methods

1. Introduction

Gaussian Markov random fields (GMRFs), or conditional autoregressions, are finite Gaussian fields with a Markov property: the density of one component conditioned on all the other components depends only on its neighbours (Mardia, 1988; Cressie, 1993; Besag and Kooperberg, 1995). GMRFs are frequently used in a wide variety of models including dynamic linear models (West and Harrison, 1997), semiparametric regression (Fahrmeir and Lang, 2001a, b) and spatial and spatiotemporal models (Besag *et al.*, 1991; Heikkinen and Arjas, 1998; Knorr-Held and Besag, 1998; Wikle *et al.*, 1998; Besag and Higdon, 1999; Knorr-Held, 2000; Fernández and Green, 2002; Rue and Tjelmeland, 2002). GMRFs are analytically tractable and have excellent computational properties owing to fast numerical algorithms for sparse matrices (Rue, 2001; Rue and Follestad, 2003). These algorithms include those based on the Kalman filter that were developed for dynamic linear models (Carter and Kohn, 1994; Shephard, 1994; Frühwirth-Schnatter, 1994) as a special case; see the appendix of Knorr-Held and Rue (2002).

In applications the GMRF $\mathbf{x} = (x_1, \ldots, x_n)$ usually depends on hyperparameters $\boldsymbol{\theta}$ with prior $\pi(\boldsymbol{\theta})$. Further, we assume that the data $\mathbf{y}$ are mutually independent, indirect observations of

Address for correspondence: Håvard Rue, Department of Mathematical Sciences, Norwegian University of Science and Technology, N-7491 Trondheim, Norway.
E-mail: Havard.Rue@math.ntnu.no

 1369–7412/04/66877

some components $x_i, i \in \mathcal{I}$, with likelihood $\pi(y_i|x_i)$. The full conditional density $\mathbf{x}$ is then

$$\pi(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y}) \propto \pi(\mathbf{x}|\boldsymbol{\theta}) \prod_{i \in \mathcal{I}} \pi(y_i|x_i). \tag{1}$$

We shall refer to stochastic variables with posterior densities of this form as hidden GMRFs (HGMRFs). The HGMRF is of interest by itself but more often as part of the joint density

$$\pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}) \propto \pi(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})\, \pi(\boldsymbol{\theta}|\mathbf{y}) \tag{2}$$

which is the posterior for $(\mathbf{x}, \boldsymbol{\theta})$ given the data $\mathbf{y}$. Since the HGMRF is a GMRF when the likelihood is Gaussian, we focus on non-Gaussian likelihoods giving non-Gaussian HGMRFs. The number of hyperparameters is usually small (1–3) whereas n is large (10^3–10^4). We aim to do inference about $(\boldsymbol{\theta}, \mathbf{x})$ from density (2).

Inference from density (2) is often made by using single-site Markov chain Monte Carlo (MCMC) algorithms (Robert and Casella, 1999); see for example Besag *et al.* (1991), Diggle *et al.* (1998) and Besag and Higdon (1999). Single-site MCMC sampling has been found to perform well in many cases. But can we rely on the results as the complexity and size of statistical models increase? Regarding sampling from density (2), Shephard (2000) stated that naïve single-site MCMC algorithms are poorly suited, whereas Christensen *et al.* (2003) warned against being overoptimistic about mixing and convergence using single-site algorithms. This may explain why Christensen *et al.* (2000), section 5.2, were unable to reproduce the result for one example in Diggle *et al.* (1998) by using improved MCMC algorithms and long runs. See also Knorr-Held and Rue (2002) and Gamerman *et al.* (2003) for similar comparisons. Further Christensen *et al.* (2003) noted that the performance of many MCMC algorithms can be highly sensitive to the particular data sets that are under study. Improved and robust MCMC algorithms therefore seem to be needed for high dimensional models of the form (2). There is undoubtedly a price to be paid both with respect to the complexity of implementation and computational cost. We believe that this is strictly necessary to obtain reliable inference and to maintain credibility regarding the statistical results.

A known remedy for improving slow converging single-site MCMC algorithms is to update more than one variable simultaneously. This is known as block sampling or grouping (Liu *et al.*, 1994; Roberts and Sahu, 1997). Concerning expression (2) one might expect that a two-block Gibbs sampler which alternately updates $\boldsymbol{\theta}$ and $\mathbf{x}$ from their full conditionals has rapid convergence. Rue and Follestad (2003) showed that this is not always so. Let θ (with a Gaussian prior) be the common mean for the elements in a GMRF. Then the marginal θ-chain is a Gaussian autoregressive process of order 1 with correlation function decaying with rate $1 - \mathcal{O}(1/n)$. The number of iterations to convergence is $\mathcal{O}(n)$, where n is the dimension of the GMRF. This result is supported by empirical findings in Knorr-Held and Rue (2002) in the non-Gaussian case. The slow convergence is due to the strong interaction between $\boldsymbol{\theta}$ and $\mathbf{x}$, but this is resolved if $(\boldsymbol{\theta}, \mathbf{x})$ can be updated jointly.

We consider the following two-step procedure for producing joint updates of $(\boldsymbol{\theta}, \mathbf{x})$ in a Metropolis–Hastings algorithm. Let $q(\cdot|\cdot)$ denote a generic proposal density. We sample a new proposal for the hyperparameter $\boldsymbol{\theta}' \sim q(\boldsymbol{\theta}'|\boldsymbol{\theta}, \mathbf{y})$ followed by $\mathbf{x}' \sim q(\mathbf{x}'|\boldsymbol{\theta}', \mathbf{x}, \mathbf{y})$ and then accept or reject $(\boldsymbol{\theta}', \mathbf{x}')$ jointly. The optimal scheme uses $q(\boldsymbol{\theta}'|\boldsymbol{\theta}, \mathbf{y}) = \pi(\boldsymbol{\theta}'|\mathbf{y})$ and $q(\mathbf{x}'|\boldsymbol{\theta}', \mathbf{x}, \mathbf{y}) = \pi(\mathbf{x}'|\boldsymbol{\theta}', \mathbf{y})$ in density (1), which gives immediate convergence. The density $\pi(\mathbf{x}'|\boldsymbol{\theta}', \mathbf{y})$ of the HGMRF is essential in both these proposals since we may use that

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{\pi(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})}{\pi(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})} \tag{3}$$

for any value of $\mathbf{x}$. Although the optimal scheme is infeasible, we can apply equation (3) if we can construct a good approximation to the HGMRF. The approximation to the HGMRF can be used in equation (3) to construct (a numerical) approximation $q(\boldsymbol{\theta}'|\boldsymbol{\theta}, \mathbf{y})$ to $\pi(\boldsymbol{\theta}'|\mathbf{y})$. Next, sample $\mathbf{x}'$ using the HGMRF approximation conditioned on $\boldsymbol{\theta}'$, and finally accept or reject $(\boldsymbol{\theta}', \mathbf{x}')$ jointly. The result is a Metropolized independence sampler. As the number of hyperparameters is small, the choice of $q(\boldsymbol{\theta}'|\boldsymbol{\theta}, \mathbf{y})$ is not so critical. Therefore, we may also choose a simple (log-)random walk.

The requirements for the HGMRF approximations that are discussed are that they must be sufficiently accurate, and it must be possible to draw samples from and to evaluate the normalized density. The HGMRF is non-Gaussian but has a Gaussian flavour, and this has motivated the use of GMRF approximations to HGMRFs (Shephard and Pitt, 1997; Gamerman, 1998; Knorr-Held, 1999; Durbin and Koopman, 2000; Knorr-Held and Rue, 2002). GMRF approximations are typically found by expanding the logarithm of density (1) to second order at or near the mode. We can sample from this GMRF and compute the normalized density, but we cannot control the accuracy of the approximation. As the dimension n increases or the likelihood term in expression (1) becomes more dominant, the GMRF approximation is not sufficiently accurate to be used as a proposal in a Metropolis–Hastings algorithm but leads to (extremely) low acceptance rates (Shephard and Pitt, 1997; Gamerman, 1998; Pitt, 2000).

In those cases where the GMRF approximation to the HGMRF is too crude, we might look towards more general approaches for constructing block updates. Langevin–Hastings algorithms (Besag, 1994; Roberts and Tweedie, 1996) propose a small change to the variables of interest. A Gaussian approximation to the HGMRF is important also for Langevin-Hastings algorithms as it provides a reparameterization into independent variables of equal variance (Christensen and Waagepetersen, 2002; Christensen *et al.*, 2000, 2003). The reparameterization is crucial for Langevin–Hastings algorithms, but even reasonable reparameterizations can turn out to be insufficiently accurate, leading to slow convergence when applied to some data sets (Christensen *et al.* (2000), section 5.2). It seems difficult to find good reparameterizations for $(\boldsymbol{\theta}, \mathbf{x})$ which are sufficiently accurate for a wide range of $\boldsymbol{\theta}$. One approach might be to let the parameterization depend on $\boldsymbol{\theta}$, but this leads to other complications.

The main contribution in this paper is the development of a new class of approximations to HGMRFs. These approximations adapt automatically to the particular HGMRF that is under study (under certain assumptions). The accuracy of the approximations can be tuned by intuitive parameters to nearly any precision, and the new non-Gaussian approximations share the same computational complexity as GMRFs. They can be sampled exactly from and have a computable normalizing constant. Both the construction of the approximations and our implementation are designed for general GMRFs in the HGMRF (1). The computational complexity depends on the structure of the Markov properties. It is $\mathcal{O}(n)$ for models in time and between $\mathcal{O}(n^{3/2})$ and $\mathcal{O}(n^2)$ for spatial models. Using our HGMRF approximations we can update $(\boldsymbol{\theta}, \mathbf{x})$ jointly and construct Metropolized independence samplers for density (2).

The outline of the paper is as follows. In Section 2 we present some background for the problem that is 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 Metropolized independence samplers for some models that are 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 labelled undirected graph with n nodes, where a node for example is a spatial region, a pixel in a lattice or a tile in a tessellation. Two nodes, i and j, may be 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 (proper) GMRF with respect to $\mathcal{G}$, meaning that its $n \times n$ precision matrix (inverse covariance) $\mathbf{Q}$ has the property that $Q_{ij} \neq 0$ if and only if $i \sim j$ or $i = j$. The Markov property of $\mathbf{x}$ is given by $\mathcal{G}$ as x_i and x_j are conditionally independent given the rest if and only if $i \not\sim j$. The precision matrix often depends on further parameters $\boldsymbol{\theta}$, which we denote by $\mathbf{Q}(\boldsymbol{\theta})$. Let $\mathbf{x}_{i:j}$ be $(x_i, x_{i+1}, \ldots, x_j)^{\mathrm{T}}$ for $i \leqslant j$ and empty otherwise.

An HGMRF is a GMRF observed through data $\mathbf{y}$. We assume throughout that the observations $\mathbf{y}$ are mutually independent given $\mathbf{x}$, and that each y_i for $i \in \mathcal{I}$ only depends on x_i with likelihood $\pi(y_i|x_i)$. Further, we assume that $\pi(y_i|x_i)$ as a function of x_i is strictly positive and absolutely continuous with respect to the Lebesgue measure, such that the posterior density in expression (1) of the HGMRF is

$$\pi(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y}) \propto \exp\left\{-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\,\mathbf{Q}(\boldsymbol{\theta})\mathbf{x} - \sum_{i \in \mathcal{I}} g_i(x_i, y_i)\right\} \tag{4}$$

for some functions $g_i(x_i, y_i)$. In Section 3 we show how to construct approximations to density (4) when 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 criterion for density (4) to be log-concave and thereby unimodal is that $-g_i(x_i, y_i)$ as a function of x_i is concave for all $i \in \mathcal{I}$.

3. Approximations to a hidden Gaussian Markov random field

3.1. A Gaussian Markov random-field approximation

A GMRF approximation to $\pi(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ in density (4) can be constructed by finding the mode $\mathbf{x}^{\mathrm{m}} = \mathbf{x}^{\mathrm{m}}(\boldsymbol{\theta}, \mathbf{y})$, and replacing $g_i(x_i, y_i)$ by the Taylor series expansion at the mode, $a_i + b_i x_i + c_i x_i^2/2$. The GMRF approximation $\pi_{\mathrm{G}}(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ has precision matrix $\mathbf{Q}_{\mathrm{G}} = \mathbf{Q} + \mathrm{diag}(\mathbf{c})$ and the mode $\mathbf{x}^{\mathrm{m}}$ as expected value. For computational and presentational reasons we assume that the indices are permuted so $\mathbf{Q}_{\mathrm{G}}$ is a band matrix with small bandwidth b_w. We return to a discussion of computational issues in Section 4.

Let $\mathbf{L}_{\mathrm{G}}$ be the Cholesky factorization of $\mathbf{Q}_{\mathrm{G}} = \mathbf{L}_{\mathrm{G}}\mathbf{L}_{\mathrm{G}}^{\mathrm{T}}$. The Cholesky factor $\mathbf{L}_{\mathrm{G}}$ is a lower triangular matrix with the same bandwidth b_w as $\mathbf{Q}$. A sequential representation of the GMRF approximation π_{G} with mean $\mathbf{x}^{\mathrm{m}}$ is directly available by

$$\pi_{\mathrm{G}}(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y}) = \prod_{t=n}^{1} \pi_{\mathrm{G}}(x_t|\mathbf{x}_{(t+1):n}, \boldsymbol{\theta}, \mathbf{y}),$$

where

$$\pi_{\mathrm{G}}(x_t|\mathbf{x}_{(t+1):n}, \boldsymbol{\theta}, \mathbf{y}) = \mathcal{N}\left\{x_t;\, x_t^{\mathrm{m}} - \frac{1}{L_{\mathrm{G},tt}} \sum_{j=t+1}^{\min\{t+b_w, n\}} L_{\mathrm{G},jt}(x_j - x_j^{\mathrm{m}}),\, \frac{1}{L_{\mathrm{G},tt}^2}\right\}, \tag{5}$$

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

In Section 3.4 we show how the GMRF approximation π_{G} can be computed quickly and sampled exactly from, and that the normalizing constant can be found easily. The approximation does, however, have a major drawback; the accuracy cannot be tuned.

3.2. Improved approximations

When constructing improved approximations based on π_{G}, note that density (4) can be written in the following two ways:

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

$$\propto \prod_{t=1}^{n} \pi_{\mathrm{G}}(x_t|\mathbf{x}_{(t+1):n},\boldsymbol{\theta},\mathbf{y}) \exp\{-h_t(x_t,y_t)\} \tag{7}$$

where π_{G} is defined in equation (5) and

$$h_t(x_t,y_t) = g_t(x_t,y_t) - (a_t + b_t x_t + c_t x_t^2/2)$$

for $t \in \mathcal{I}$ and $h_t(x_t,y_t)$ is 0 otherwise. For $t \notin \mathcal{I}$, y_t is empty.

Each of the terms in the sequential representation (6) can be represented by means of density (7) as

$$\pi(x_t|\mathbf{x}_{(t+1):n},\boldsymbol{\theta},\mathbf{y}) \propto \pi_{\mathrm{G}}(x_t|\mathbf{x}_{(t+1):n},\boldsymbol{\theta},\mathbf{y}) \exp\{-h_t(x_t,y_t)\} \times \int \pi_{\mathrm{G}}(\mathbf{x}_{1:(t-1)}|\mathbf{x}_{t:n},\boldsymbol{\theta},\mathbf{y}) \exp\left\{-\sum_{j=1}^{t-1} h_j(x_j,y_j)\right\} \mathrm{d}\mathbf{x}_{1:(t-1)}, \tag{8}$$

where all the conditional densities of π_{G} can be easily derived from density (5). This representation has the important property that the mode of the integrand usually is close to the mode of $\pi_{\mathrm{G}}(\mathbf{x}_{1:(t-1)}|\mathbf{x}_{t:n},\boldsymbol{\theta},\mathbf{y})$. This is because $\pi_{\mathrm{G}}(\mathbf{x}|\boldsymbol{\theta},\mathbf{y})$ has the same mode as density (4) which is assumed to be unimodal, and the modes of the conditional densities are usually close to this mode. In what follows this enables us to use importance sampling to produce accurate approximations to the integral.

If we neglect the dependence of $\mathbf{y}$ in π_{G} and the possibility that $\exp\{-h_i(x_i,y_i)\}$ is not integrable in y_i, the right-hand side of expression (7) can be interpreted as the posterior of $\mathbf{x}$ with GMRF prior $\pi_{\mathrm{G}}(\mathbf{x})$ 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 expression (8), denoted by $\tilde{\pi}(x_t|\mathbf{x}_{(t+1):n},\boldsymbol{\theta},\mathbf{y})$, and to join them together into an approximation to $\pi(\mathbf{x}|\boldsymbol{\theta},\mathbf{y})$ based on equation (6):

$$\tilde{\pi}(\mathbf{x}|\boldsymbol{\theta},\mathbf{y}) = \prod_{t=1}^{n} \tilde{\pi}(x_t|\mathbf{x}_{(t+1):n},\boldsymbol{\theta},\mathbf{y}). \tag{9}$$

Note that equation (9) can be sampled sequentially backwards in time, and its normalizing constant is the product of n univariate normalizing constants. We shall now discuss how to construct these univariate approximations, by removing what can be considered as less important terms on the right-hand side of expression (8).

(a) The crudest approximation is to neglect both the $h_t(x_t,y_t)$ term and the integral term in expression (8),

$$\tilde{\pi}_{(\mathrm{a})}(x_t|\mathbf{x}_{(t+1):n},\boldsymbol{\theta},\mathbf{y}) = \pi_{\mathrm{G}}(x_t|\mathbf{x}_{(t+1):n},\boldsymbol{\theta},\mathbf{y}). \tag{10}$$

This gives the GMRF approximation in Section 3.1.

(b) A simple, but often significant, improvement to equation (10) is to include the $h_t(x_t,y_t)$ term, which can be considered as the second most important term in expression (8),

$$\pi_{(\mathrm{b})}(x_t|\mathbf{x}_{(t+1):n}, \boldsymbol{\theta}, \mathbf{y}) \propto \pi_{\mathrm{G}}(x_t|\mathbf{x}_{(t+1):n}, \boldsymbol{\theta}, \mathbf{y}) \ \exp\{-h_t(x_t, y_t)\}. \tag{11}$$

The density in expression (11) is univariate and can usually be sampled from in many ways. However, to be able to compute the normalizing constant, we choose to approximate $\pi_{(\mathrm{b})}$ by a density $\tilde{\pi}_{(\mathrm{b})}$ that is constructed by log-quadratic splines, i.e. compute the logarithm of the right-hand side of expression (11) in some evaluation points $\{\check{x}_t\}$, and interpolate using piecewise quadratic polynomials. $\tilde{\pi}_{(b)}$ can be sampled exactly from by use of the real and complex complementary error function, and the normalizing constant is easily found. Details are provided in Appendix A.

$\pi_{(\mathrm{b})}$ in expression (11) can be written as the conditional density of an HGMRF with $\mathbf{y}_{1:(t-1)}$ as missing values. The approximation can be a significant improvement to the GMRF approximation in equation (10). If for instance $\mathbf{Q} = \kappa\mathbf{P}$, for a scalar κ and $\kappa \to 0$, the likelihood dominates in density (4). Then the GMRF approximation can be quite poor, and the error by using the approximation in expression (11) mainly depends on how accurate the log-spline representation is.

(c) The next class of approximations takes the integral term in expression (8) into account. $\tilde{\pi}_{(\mathrm{c})}$ is constructed by log-quadratic splines based on evaluations in the points $\{\check{x}_t\}$ of

$$\pi_{(\mathrm{c})}(x_t|\mathbf{x}_{(t+1):n}, \boldsymbol{\theta}, \mathbf{y}) \propto \pi_{\mathrm{G}}(x_t|\mathbf{x}_{(t+1):n}, \boldsymbol{\theta}, \mathbf{y}) \exp\{-h_t(x_t, y_t)\} \ \hat{I}(x_t). \tag{12}$$

Here $\hat{I}(x_t)$ is an approximation to the integral term in expression (8), which can be written as

$$I(x_t) = E\left[\exp\left\{-\sum_{j=1}^{t-1} h_j(x_j, y_j)\right\}\right]. \tag{13}$$

The expectation is with respect to $\pi_{\mathrm{G}}(\mathbf{x}_{1:(t-1)}|\mathbf{x}_{t:n}, \boldsymbol{\theta}, \mathbf{y})$. Since the likelihood consists of independent observations conditioned on the GMRF prior, the HGMRF inherits the Markov properties from the GMRF. We therefore expect neighbouring sites to be most correlated. Hence, as a function of x_t, the important terms in equation (13) are those js which are neighbours to t or have a common neighbour with t and so on. Let $\mathcal{J}(t)$ be the set of sites which we want to include in our approximation to equation (13). We estimate this approximation by an empirical average, using M samples from $\pi_{\mathrm{G}}(\mathbf{x}_{1:(t-1)}|\mathbf{x}_{t:n}, \boldsymbol{\theta}, \mathbf{y})$,

$$\hat{I}(x_t) = \frac{1}{M}\sum_{i=1}^{M} \exp\left\{-\sum_{j\in\mathcal{J}(t)} h_j(x_j^i, y_j)\right\}. \tag{14}$$

Here, x_j^i is the jth component of the ith sample, which is obtained by successively drawing from density (5) from time t until $\min_t\{\mathcal{J}(t)\}$. If density (4) is not unimodal, the estimate (14) is less accurate, since samples from $\pi_{\mathrm{G}}(\mathbf{x}_{1:(t-1)}|\mathbf{x}_{t:n}, \boldsymbol{\theta}, \mathbf{y})$ might not cover all the modes. Note that the approximation in expression (9) is determined by the sequence of random numbers that is used in equation (14), and by keeping this sequence fixed we can produce several samples from the same approximation. With the number of knots K and this sequence of random numbers being fixed, both $\tilde{\pi}_{(\mathrm{b})}(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ and $\tilde{\pi}_{(\mathrm{c})}(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ are continuous, and they are differentiable almost everywhere. In contrast, if the sequence is not fixed, $\hat{I}(x_t)$ is stochastic, and the density of $\tilde{\pi}_{(\mathrm{c})}$ is difficult to compute.

A proper $\pi(\mathbf{x})$ is sufficient for both $\pi(\mathbf{x}|\mathbf{y})$ and all our approximations to be proper. In the case where $\pi(\mathbf{x})$ is improper and $\pi(\mathbf{x}|\mathbf{y})$ is proper, the approximations are proper under some mild conditions on the likelihood.

3.3. Geometrical ergodicity

In Section 4 we use $\tilde{\pi}(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$ as a proposal density in a Metropolized independence sampler. For MCMC algorithms, geometrical ergodicity can be crucial to ensure reliable inference (Roberts and Rosenthal, 1998). A sufficient and necessary condition for geometrical ergodicity is that the supremum of the ratio of the target density and the proposal density is finite. Concerning $\tilde{\pi}(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$, we need that

$$M = \sup_{\mathbf{x}}\{\pi(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})/\tilde{\pi}(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})\} < \infty$$

and $\tilde{\pi}(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$ is one of our approximations. Assume that the likelihood $\pi(\mathbf{y}|\mathbf{x})$ is bounded in $\mathbf{x}$ for fixed $\mathbf{y}$. We can provide sufficient conditions for $M < \infty$ but the conditions are rather technical and do not give much insight. An easy way out is instead to use the mixture, $p\,\pi(\mathbf{x}) + (1-p)\,\tilde{\pi}(\mathbf{x}|\mathbf{y}, \boldsymbol{\theta})$, with $0 < p < 1$, as the proposal density. With this choice $M < \infty$.

3.4. Computational issues

Before discussing computational issues, we explain how to draw samples from a GMRF. A zero-mean GMRF $\mathbf{x}$ with precision $\mathbf{Q}$ can be sampled by taking $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and solving $\mathbf{L}^{\mathrm{T}}\mathbf{x} = \mathbf{z}$, where $\mathbf{L}$ is the Cholesky factorization of $\mathbf{Q}$ satisfying $\mathbf{Q} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$ (Rue, 2001). If the mean of $\mathbf{x}$ is non-zero, a sample is obtained by adding the mean to a zero-mean GMRF with the same $\mathbf{Q}$. For those cases where the mean is given implicitly by $\mathbf{Q}\boldsymbol{\mu} = \mathbf{b}$, it is found efficiently by solving $\mathbf{L}\mathbf{u} = \mathbf{b}$ and $\mathbf{L}^{\mathrm{T}}\boldsymbol{\mu} = \mathbf{u}$. The normalizing constant is available from $\mathbf{L}$, since

$$\log|\mathbf{Q}| = 2\sum_{i}\log(L_{ii}).$$

The computational cost of sampling from both the GMRF and the other approximations can be significantly reduced by exploiting 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 that there is 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 concerning GMRFs are given in Rue (2001). In the spatial case $b_w = \mathcal{O}(\sqrt{n})$, and computation of the Cholesky factorization $\mathbf{Q} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$ can then be computed using only $\mathcal{O}(nb_w^2) = \mathcal{O}(n^2)$ operations compared with $\mathcal{O}(n^3)$ in the general case (see also the comment below). Note that $\mathbf{L}$ is a lower triangular matrix with the same bandwidth as $\mathbf{Q}$.

The computation of equation (14) is potentially quite costly and must therefore be done carefully. If $\mathcal{J}(t)$ are those neighbours to t that are smaller than t, we need $\mathcal{O}(\sqrt{n})$ evaluations of density (5), 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 factorizing $\mathbf{Q}_{\mathrm{G}}$. However, this cost can be reduced to $\mathcal{O}(Mn^2)$: note that the conditional mean of $\pi_{\mathrm{G}}(x_j, j \in \mathcal{J}(t)|\mathbf{x}_{t:n}, \boldsymbol{\theta}, \mathbf{y})$ is linear in x_t, and the conditional covariance does not depend on x_t. By first computing the conditional mean in expression (5) for two values of x_t, the conditional mean for all other values of x_t is given as a linear combination of these two. The samples for estimating equation (14) are then obtained by sampling M independent samples with zero mean and adding the conditional mean depending on x_t to each of the samples. Since the same stream of random numbers is used for each value of x_t, equation (14) is continuous with respect to x_t.

Antithetic ideas (Durbin and Koopman, 1997) can be used to improve the estimate in equation (14): let $\mathbf{v}$ be a sample from a zero-mean Gaussian distribution, u a sample from uniform(0, 1), f_1 the u-quantile of a χ_n-distribution and f_2 the $(1-u)$-quantile. Then $\pm f_1\mathbf{v}/\sqrt{(\mathbf{v}^{\mathrm{T}}\mathbf{v})}$ and $\pm f_2\mathbf{v}/\sqrt{(\mathbf{v}^{\mathrm{T}}\mathbf{v})}$ will be from the same Gaussian distribution. Without extra costs, this provides

three dependent, hopefully negatively correlated extra samples, for each of the M samples. This will reduce the variance of the empirical average in equation (14).

Another improvement to equation (14) is to use approximation (b) in Section 3.2 as the sampling distribution instead of π_{G}. This requires some obvious changes in expressions (13) and (14) and costs $\mathcal{O}(2KMn^2)$ operations. We do not discuss this option further.

There are more efficient factorizations of $\mathbf{Q}$ than those based on band algorithms (Rue (2001), appendix); see Rue and Follestad (2003) for details. The computational cost in the spatial case is $\mathcal{O}(n^{3/2})$ operations, but the non-zero pattern in $\mathbf{L}_{\mathrm{G}}$ is quite complex so a detailed presentation is omitted in favour of the simpler band matrix approach. Further, the complexity prevents a precise operation count for approximation (c). Although we believe that approximation (c) can be computed in $\mathcal{O}(n^{3/2})$ operations, we do not have any fail-safe argument. We conclude that the computational complexity of approximations (a) and (b) in Section 3.2 is $\mathcal{O}(n^{3/2})$, and for approximation (c) it is somewhere between $\mathcal{O}(n^{3/2})$ and $\mathcal{O}(n^2)$.

We have implemented the algorithm in C as a part of the open source `GMRFLib` library (Rue and Follestad, 2002), which is available from `http://www.math.ntnu.no/~hrue`. 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 to construct Metropolized independence samplers for such models. The first is motivated from a Bayesian model for mapping disease (Besag *et al.*, 1991; Mollié, 1996), whereas 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 $i = 1, \ldots, n$. 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

$$\pi(\mathbf{x}, \kappa|\mathbf{y}) \propto \kappa^{(n-1)/2} \exp\Big[-\frac{\kappa}{2} \sum_{i \sim j} (x_i - x_j)^2 - \sum_i \{e_i \exp(x_i) - y_i x_i\} \Big] \pi(\kappa) \tag{15}$$

where κ is the precision of the GMRF prior, with prior $\pi(\kappa)$. The full conditional of $\mathbf{x}$ is of the form (4), 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 of Besag *et al.* (1991) also includes an additional unstructured heterogeneity term in the log-relative-risk. This term should always be included in density (15) when applied to data. We ignore it here only to avoid unnecessary complications in illustrating our approximations. We shall illustrate our approximations on some data on oral cavity cancer mortality for males in Germany (1986–1990), analysed by Knorr-Held and Raßer (2000).

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

We shall now demonstrate how improved approximations compare with the GMRF approximation when κ is fixed by constructing various approximations for $\kappa = 0.1, 1, 10$. 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 $\pi(\mathbf{x}|\kappa, \mathbf{y})$: approximation (a) in Section 3.2 is the GMRF approximation, (b) the approximation including only the likelihood term (8), (c)(i) the approximation including also equation (14) with $\mathcal{J}(t)$ as the neighbours to node t less than t, using $M = 1$, and (c)(ii) the same as (c)(i) but with $M = 100$. Approximations (c)(i) and (c)(ii) also make use of three extra antithetic variables for each sample, as described in Section 3.2. We use $K = 20$ knots in the log-spline approximation.

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

Table 1 displays an estimate of the acceptance rate for the four approximations averaged over 1000 iterations. For approximations (c)(i) and (c)(ii) we use different random numbers to generate each of the 1000 approximations; hence we average over that source of randomness as well. The results that were obtained are quite typical. When κ is small, $\pi(\mathbf{x}|\kappa, \mathbf{y})$ is dominated by the non-Gaussian likelihood, and the acceptance rate for approximation (a) increases for increasing κ.

The inclusion of the likelihood term in approximation (b) raises the acceptance rate from 0.01 to 0.94 for $\kappa = 0.1$. For increasing κ, approximation (a) becomes better, whereas approximation (b) has a slight decrease in the acceptance rate. This is due to the increase of the relative influence of the GMRF prior. Approximations (c)(i) and (c)(ii) demonstrate further improvements, by accounting for the spatial dependence in addition to the likelihood by including equation (14). Increasing the number of samples from 1 to 100 improves the approximations. However, the improvement of approximations (c)(i) and (c)(ii) over (b) is less than how much approximation (b) improves over (a). The higher the acceptance rate, the more difficult it seems to improve the approximation. For increasing κ, the acceptance rate for all approximations eventually tends to 1.

In this example a 1200-MHz laptop computer needed 0.06 s per iteration for approximation (a), whereas approximations (b), (c)(i) and (c)(ii) required 10, 30 and 1900 times this respectively. Each iteration requires the construction of two approximations and two optimizations. The computational efficiency that is obtained by approximations (b) and (c)(i) compared with (a) is to us quite impressive.

Table 1. Average acceptance rate for four different approximations: GMRF (a) and improvements (b), (c)(i) and (c)(ii)

Approximation	*Average acceptance rates for the following values of* κ*:*		
	$\kappa = 0.1$	$\kappa = 1$	$\kappa = 10$
(a)	0.01	0.11	0.47
(b)	0.94	0.80	0.78
(c)(i)	0.96	0.87	0.86
(c)(ii)	0.99	0.96	0.90

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 acceptance rate is high. In most cases 10–20 knots are sufficient. The inclusion of the likelihood term in expression (11) can give a huge improvement compared with the GMRF approximation. Correcting using equation (14) generally helps but is less important compared with the likelihood term in expression (11). If approximation (b) gives too low an acceptance rate then equation (14) is required. Computing equation (14) can be expensive, as demonstrated in this example. We have good experience using only one sample ($M = 1$) in equation (14), and letting this be the conditional mean (computed under the GMRF approximation) or mode. This correction usually gives a positive influence on the acceptance rate whereas further improvements require relatively much more computing. We generally recommend using $\mathcal{J}(t)$ as the neighbours of node t less than t, but an increase in speed can be gained if $\mathcal{J}(t) = \{t-1, t-2\}$, say, is sufficient to obtain a reasonable acceptance rate. The computational cost is between $\mathcal{O}(n^{3/2})$ and $\mathcal{O}(n^2)$ for the first choice (see Section 3.4), but only $\mathcal{O}(n^{3/2})$ for the second.

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

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

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

$$\pi(\kappa|\mathbf{y}) = \frac{\pi(\mathbf{x}, \kappa|\mathbf{y})}{\pi(\mathbf{x}|\kappa, \mathbf{y})}, \tag{16}$$

which is valid for any $\mathbf{x}$ such that the denominator is non-zero; see also Besag (1989). The implication of equation (16) is that we can replace integration over $\mathbf{x}$ in $\pi(\mathbf{x}, \kappa|\mathbf{y})$ 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 $\tilde{\pi}\{\mathbf{x}|\kappa, \mathbf{y}, \mathbf{x}^{\mathrm{m}}(\kappa)\}$ be an approximation to $\pi(\mathbf{x}|\kappa, \mathbf{y})$ around the mode $\mathbf{x}^{\mathrm{m}}(\kappa)$. A natural candidate for an approximation to $\pi(\kappa|\mathbf{y})$ is

$$\tilde{\pi}(\kappa|\mathbf{y}) \propto \left. \frac{\pi(\kappa)\, \pi(\mathbf{x}|\kappa)\, \pi(\mathbf{y}|\mathbf{x})}{\tilde{\pi}\{\mathbf{x}|\kappa, \mathbf{y}, \mathbf{x}^{\mathrm{m}}(\kappa)\}} \right|_{\mathbf{x}=\mathbf{x}^{\mathrm{m}}(\kappa)}. \tag{17}$$

Note that the denominator contains a κ-dependent normalizing constant, which is computable for our approximations. The right-hand side is evaluated in $\mathbf{x}^{\mathrm{m}}$, the point which we think gives the most accurate result, following Tierney *et al.* (1989). We fix the random numbers that are used in the approximation to make the denominator continuous with respect to κ. A Metropolized independence sampler can now be constructed, by sampling κ from a log-quadratic-spline approximation to expression (17) and then sampling $\mathbf{x}$ from $\tilde{\pi}\{\mathbf{x}|\kappa, \mathbf{y}, \mathbf{x}^{\mathrm{m}}(\kappa)\}$.

Fig. 1 shows our approximations to the marginal posterior 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 posterior appear as one curve and suggest that approximation (a) might be sufficient

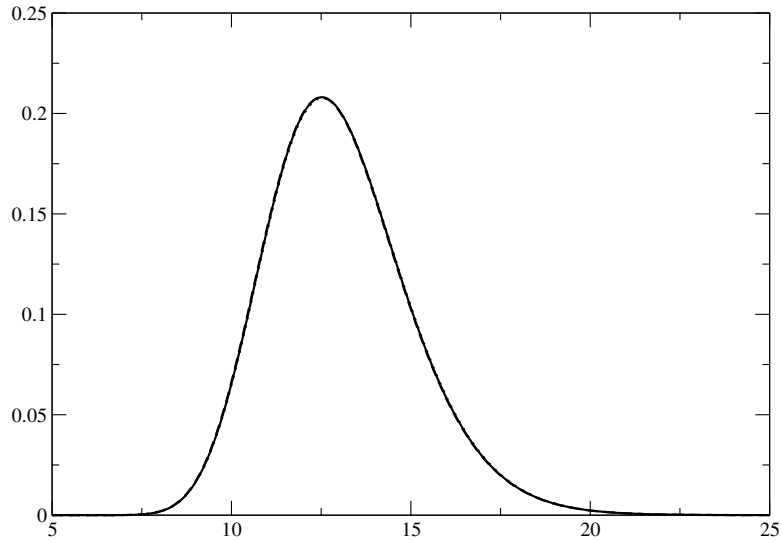

Fig. 1. Approximated marginal posterior density for κ, computed by using expression (17) and approximations (a), (b) and (c)(i): the three approximations nearly coincide

for this purpose. This contrasts with the acceptance rate in Table 1 which varies with κ and the approximation used. The interpretation is that the denominator in expression (17) has about the same functional form in κ (evaluated at $\mathbf{x}^{\mathrm{m}}(\kappa)$) for the various approximations. The constant of proportionality for this function cancels when normalizing expression (17).

A Metropolized independence sampler using approximations (a), (b) and (c)(i) gave an acceptance rate of 0.43, 0.82 and 0.86 averaged over 1000 iterations respectively. The autocorrelation for κ at lag $k \geqslant 0$ is approximately $(1-\alpha)^k$ where α is the average acceptance rate. Hence, the sampler seems to converge quite quickly for all three approximations.

The delayed rejection algorithm (Mira, 2001) could also have been used here, using approximation (a) to sample the first proposal, and then using for example approximation (b) if the first proposal is rejected. No extra cost is involved as the GMRF approximation is needed in any case.

To obtain more insight into the convergence of the Metropolized independence sampler in this example, we use the empirical supremum rejection sampler that was introduced by Caffo *et al.* (2002). Their algorithm is the standard rejection sampler, but where the supremum of $\pi(\mathbf{x}, \kappa|\mathbf{y})/\tilde{\pi}(\mathbf{x}, \kappa|\mathbf{y})$, denoted by C, is replaced by the largest value observed so far. Let C_m denote this quantity after m trials. They studied the convergence rate of C_m towards C as $m \to \infty$ and on the basis of this they argued that we can treat the output of this algorithm as random samples from the target when the samples are used to estimate expectations with respect to it. We applied their algorithm and estimated C to be 25.0, 1.47 and 1.39 for the joint approximation based on approximations (a), (b) and (c)(i), after 1000 iterations. We also ran the algorithm based on approximation (c)(i) 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 is always more questionable. If we believe in the estimated Cs, we can sample exactly from the joint posterior by using rejection sampling.

The convergence of the Metropolized 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 (b) is about three times more efficient compared with the GMRF approximation, taking the computation cost into account.

4.4. Model-based geostatistics

Diggle *et al.* (1998) discussed Bayesian models which combine traditional geostatistical methods with those of generalized linear models. The common setting is a spatial Gaussian field with some unknown parameters $\boldsymbol{\theta}$ (mean, precision and correlation length) which is observed at some locations with a non-Gaussian likelihood. The goal is to estimate $\boldsymbol{\theta}$ and to estimate the Gaussian field.

Consider one of the examples in Diggle *et al.* (1998), where the data are reported outbreaks of campylobacter, salmonella and cryptosporidia in north Lancaster (UK) between April and December 1994. The data come 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 p_{l_i} that an enteric infection at position l_i is campylobacter is linked to the spatial field by $\text{logit}(p_{l_i}) = x_{l_i}$. The spatial latent surface measures the risk that an outbreak is from campylobacter.

We 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 Metropolized 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 $\mathbf{x}$ with a 5×5 neighbourhood and coefficients as computed by their method, for a finite set of ranges with step 0.05. This reduces the computational cost by a factor between n and $n^{3/2}$, when predictions for non-observed locations as well as parameter estimates for $\boldsymbol{\theta}$ are required. If only $\boldsymbol{\theta}$ is of interest, we may use only the set of sites where we have observed data, but this option is not considered here.

Fig. 2 shows our approximation to the scaled marginal posterior for $\boldsymbol{\theta} = (\log(\tau), r) \in [-2, 7] \times [0, 50]$ using $\mu = 0.35$ (an estimate that was obtained from the data). The plot is scaled with the prior for $(\log(\tau), r)$ and to have maximum value equal to 1. Here we used expression (17) and approximation (b). Using approximation (a) gave similar results. Each evaluation in the grid of selected $(\log(\tau), r)$ values required about 30 s of computing. We could have included μ in our '$\mathbf{x}$' by giving it a Gaussian prior at essentially no extra cost (Rue (2001), appendix), but the implementation (in the `GMRFLib` library) did not support this option at the time of writing.

On the basis of the scaled marginal posterior in Fig. 2 we can construct a log-spline approximation to the marginal posterior density of $\boldsymbol{\theta}$ and then construct a Metropolized independence sampler as in Section 4.4. Here, using a triangularization of the area of interest and log-linear splines within each triangle is perhaps the simplest choice. We easily obtain an acceptance 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 to sample the spatial field (Knorr-Held and Rue, 2002).

We also investigate the case where data similar to those observed are added to all pixels in the 200×100 lattice. There are no problems constructing approximations for the spatial field with an acceptance rate above 50%. It requires about 1 min to construct the GMRF approximation and slightly more for improved approximations. As long as the likelihood is reasonably close to Gaussian, sufficently good approximations seem easy to construct.

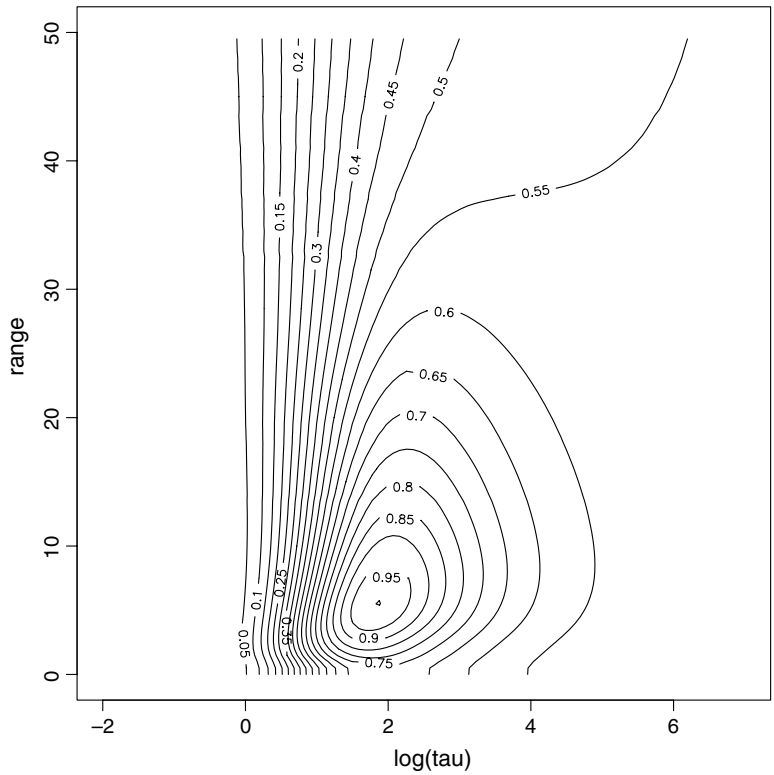

Fig. 2. Our approximation to the scaled marginal posterior of $(\log(\tau), r) \in [-2,7] \times [0,50]$ for fixed $\mu = 0.35$ for the example in Section 4.4: the plot is scaled with the prior for $(\log(\tau), r)$ and to have maximum value 1

A more challenging example is obtained when the likelihood is double exponential (Dryden *et al.*, 2002),

$$\pi(y_i|x_i) \propto \exp(-|x_i - y_i|). \tag{18}$$

This makes density (4) strongly non-Gaussian and quite an accurate approximation using procedure (c) is needed for the Metropolized independence sampler to produce an acceptance rate above 0, using simulated data. (A technical report that is available from the authors discusses this example in more detail.) It is encouraging that computing seems to be the practical limit, not our approach to construct approximations. A marginal likelihood for $\boldsymbol{\theta}$ computed with our approximations is an alternative to the asymptotic-motivated approximations that were studied by Dryden *et al.* (2002).

5. Discussion

In this paper we have introduced a new approach of constructing approximations to unimodal HGMRFs on general graphs. We can sample exactly from these approximations, which have computable normalizing constants. The examples have demonstrated how to construct joint updates and Metropolized independence samplers for spatial models. Such sampling schemes are major improvements compared with the single-site schemes that are commonly used. Which of the approximations to be used for a specific case is a trade-off between adequate mixing and convergence and computational cost.

An interesting special case is GMRF models in time where the computational complexity is $\mathcal{O}(n)$. Our experience is that good approximations are much easier to construct than in the spatial case. For GMRF models in time, or dynamic models in general, there is an extensive literature on sequential Monte Carlo methods (Doucet *et al.*, 2001). These methods can also be used to construct Metropolized 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). However, the dynamic nature of these models makes it more natural to focus on filtering and prediction. Our approach has some similarities with these methods, but we do not depend on the forward filtering–backward sampling recursions that are inherent in sequential Monte Carlo methods. This recursion requires approximations to densities of dimension b_w which is difficult for $b_w > 3$, say. Our approach works fine even for $b_w = \mathcal{O}(\sqrt{n})$ and also for general HGMRF models where there is no natural time ordering of the GMRF.

Acknowledgements

The research was funded in part by doctoral grants from the Research Council of Norway and the 'Computational mathematics in applications' programme of the Research Council of Norway. We acknowledge the comments and suggestions of the referees, the Associate Editor and the Joint Editor.

Appendix A: Constructing log-quadratic-spline approximations

Let $f(x)$ be the univariate density that we want to be sampled from. In this section we describe how to construct a log-quadratic-spline approximation $\tilde{f}(x) \approx f(x)$ and how to sample from it.

The approximation $\tilde{f}(x)$ is constructed in the following way: let the knots $\{x_k\}_{k=1}^{K+1}$ be points where $f(x)$ are known. Define the intervals $I_k = [x_k, x_{k+1})$ for $k = 1, \ldots, K$, and let $I_0 = (-\infty, x_1)$ and $I_{K+1} = [x_{K+1}, \infty)$. We define the log-quadratic spline approximation $\tilde{f}(x)$ as the mixture

$$\tilde{f}(x) = \sum_{k=0}^{K+1} p_k \, s_k(x) \, \mathbf{1}_{[x \in I_k]} \tag{19}$$

where $s_k(x)$ is the approximated density conditioned on $x \in I_k$. We compute $\{p_k\}$ and $\{s_k\}$ as follows. Let $s_0(x)$ and $s_K(x)$ be simple exponentially decaying functions. This ensures infinite support and not too light tails. On the bounded intervals let the unnormalized density s_k^u be a log-quadratic-spline, where (a_k, b_k, c_k) are chosen such that

$$\log\{s_k^u(x)\} \equiv a_k + b_k x + c_k x^2/2 \tag{20}$$

interpolates $\log\{f(x_k)\}$, $\log\{f(x_{k+1})\}$ and $\log\{f(x_{k+1/2})\}$ in the respective points. The additional knot $x_{k+1/2}$ can be set to either $(x_k + x_{k+1})/2$ or to the point which divides I_k such that the corresponding areas under a log-linear-spline interpolating $\log\{f(x_k)\}$ and $\log\{f(x_{k+1})\}$ are equal. The log-quadratic-spline (19) is completed by computing

$$s_k(x) = s_k^u(x) \Big/ \int_{I_k} s_k^u(z) \, \mathrm{d}z,$$

$$p_k = \int_{I_k} s_k^u(z) \, \mathrm{d}z \Big/ \sum_{j=0}^{K+1} \int_{I_j} s_j^u(z) \, \mathrm{d}z,$$

for $k = 0, \ldots, K+1$. The integrals can be computed by means of the real and imaginary error functions, which are available in numerical libraries.

A.1. Choice of knots

In the construction above, we need to decide how many knots to use and where to place them. We aim to construct a continuous approximation to the joint density (4). In our applications, this requires that

the conditional density $\tilde{f}(x|\mathbf{x}_{t+1:n})$ is constructed such that it is a continuous function of $\mathbf{x}_{t+1:n}$. This rules out both automatic methods for this purpose (Denison *et al.*, 1998) and adaptive approaches. Instead, we make use of information about $f(x)$. Note that in our applications we can write

$$f(x|\mathbf{x}_{t+1:n}) \propto \phi(x; \mu, \sigma^2) \exp\{c(x)\},$$

where ϕ is the Gaussian density and $c(x)$ is a correction function in the log-scale given implicitly in expressions (11) and (12) such that $c(\mu) = 0$. Now, fix the number of knots $K > 1$, and let them be equidistant points on a bounded range of interest. In our examples we centre this range at μ and let $|x_{K+1} - x_1|/2 = 6\sigma$ with $K = 20$. Note that, when $f(x)$ is Gaussian (i.e. $c(x) \equiv 0$), the log-quadratic-spline approximation is exact for any choice of knots.

A.2. Sampling

To sample from the log-quadratic-spline approximation (19), we first sample the region I_k by using the probabilities $\{p_k\}$. Within region I_k, we apply rejection sampling using a log-linear approximation $\tilde{s}_k(x)$ to $s_k(x)$. The bound $\max\{s_k(x)/\tilde{s}_k(x)\}$ that is required for rejection sampling is easily found explicitly. An alternative is to sample from a truncated Gaussian distribution when $c_k < 0$.

References

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

Besag, J. (1994) Discussion on 'Representations of knowledge in complex systems' (by U. Grenander and M. I. Miller). *J. R. Statist. Soc.* B, **56**, 591–592.

Besag, J. and Higdon, D. (1999) Bayesian analysis of agricultural field experiments (with discussion). *J. R. Statist. Soc.* B, **61**, 691–746.

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

Besag, J., York, J. and Mollié, A. (1991) Bayesian image restoration with two applications in spatial statistics (with discussion). *Ann. Inst. Statist. Math.*, **43**, 1–59.

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

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

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

Christensen, O. F., Møller, J. and Waagepetersen, R. P. (2000) Analysis of spatial data using generalized linear mixed models and Langevin-type Markov chain Monte Carlo. *Research Report R-00-2009*. Department of Mathematical Sciences, Aalborg University, Aalborg.

Christensen, O. F., Roberts, G. O. and Sköld, M. (2003) Robust MCMC for spatial GLMM's. *Preprints in Mathematical Sciences 23*. Department of Mathematical Sciences, Lund University, Lund.

Christensen, O. F. and Waagepetersen, R. P. (2002) Bayesian prediction of spatial count data using generalized linear mixed models. *Biometrics*, **58**, 280–286.

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

Denison, D. G. T., Mallick, B. K. and Smith, A. F. M. (1998) Automatic Bayesian curve fitting. *J. R. Statist. Soc.* B, **60**, 333–350.

Diggle, P. J., Tawn, J. A. and Moyeed, R. A. (1998) Model-based geostatistics (with discussion). *Appl. Statist.*, **47**, 299–350.

Doucet, A., de Freitas, N. and Gordon, N. (eds) (2001) *Sequential Monte Carlo Methods in Practice*. New York: Springer.

Dryden, I. L., Ippoliti, L. and Romagnoli, L. (2002) Adjusted maximum likelihood and pseudo-likelihood estimation for noisy Gaussian Markov random fields. *J. Comput. Graph. Statist.*, **11**, 370–388.

Durbin, J. and Koopman, S. J. (1997) Monte carlo maximum likelihood estimation for non-Gaussian state space models. *Biometrika*, **84**, 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 (with discussion). *J. R. Statist. Soc.* B, **62**, 3–56.

Fahrmeir, L. and Lang, S. (2001a) Bayesian inference for generalized additive mixed models based on Markov random field priors. *Appl. Statist.*, **50**, 201–220.

Fahrmeir, L. and Lang, S. (2001b) Bayesian semiparametric regression analysis of multicategorical time-space data. *Ann. Inst. Statist. Math.*, **53**, 11–30.

Fernández, C. and Green, P. J. (2002) Modelling spatially correlated data via mixtures: a Bayesian approach. *J. R. Statist. Soc.* B, **64**, 805–826.

Frühwirth-Schnatter, S. (1994) Data augmentation and dynamic linear models. *J. Time Ser. Anal.*, **15**, 183–202.

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

Gamerman, D., Moreira, A. R. B. and Rue, H. (2003) Space-varying regression models: specifications and simulations. *Comput. Statist. Data Anal.*, **42**, 513–533.
Heikkinen, J. and Arjas, E. (1998) Non-parametric Bayesian estimation of a spatial Poisson intensity. *Scand. J. Statist.*, **25**, 435–450.
Knorr-Held, L. (1999) Conditional prior proposals in dynamic models. *Scand. J. Statist.*, **26**, 129–144.
Knorr-Held, L. (2000) Bayesian modelling of inseparable space-time variation in disease risk. *Statist. Med.*, **19**, 2555–2567.
Knorr-Held, L. and Besag, J. (1998) Modelling risk from a disease in time and space. *Statist. Med.*, **17**, 2045–2060.
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. *Scand. J. Statist.*, **29**, 597–614.
Liu, J. S., Wong, W. H. and Kong, A. (1994) Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, **81**, 27–40.
Mardia, K. V. (1988) Multidimensional multivariate Gaussian Markov random fields with application to image processing. *J. Multiv. Anal.*, **24**, 265–284.
Mengersen, K. L. and Tweedie, R. L. (1996) Rates of convergence of the Hastings and Metropolis algorithms. *Ann. Statist.*, **24**, 101–121.
Mira, A. (2001) On Metropolis-Hastings algorithms with delayed rejection. *Metron*, **59**, 231–241.
Mollié, A. (1996) Bayesian mapping of disease. In *Markov Chain Monte Carlo in Practice* (eds W. R. Gilks, S. Richardson and D. J. Spiegelhalter), pp. 359–379. London: Chapman and Hall.
Pitt, M. K. (2000) Discussion on 'Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives' (by J. Durbin and S. J. Koopman). *J. R. Statist. Soc.* B, **62**, 38–39.
Robert, C. P. and Casella, G. (1999) *Monte Carlo Statistical Methods*. New York: Springer.
Roberts, G. O. and Rosenthal, J. S. (1998) Markov-chain Monte Carlo: some practical implications of theoretical results (with discussion). *Can. J. Statist.*, **26**, 5–31.
Roberts, G. O. and Sahu, S. K. (1997) Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *J. R. Statist. Soc.*, B, **59**, 291–317.
Roberts, G. O. and Tweedie, R. L. (1996) Exponential convergence of Langevin diffusions and their discrete approximations. *Bernoulli*, **2**, 341–363.
Rue, H. (2001) Fast sampling of Gaussian Markov random fields. *J. R. Statist. Soc.* B, **63**, 325–338.
Rue, H. and Follestad, T. (2002) GMRFLib: a C-library for fast and exact simulation of Gaussian Markov random fields. *Statistics Report 1*. Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim.
Rue, H. and Follestad, T. (2003) Gaussian markov random field models with applications in spatial statistics. *Statistics Report 6*. Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim.
Rue, H. and Tjelmeland, H. (2002) Fitting Gaussian Markov random fields to Gaussian fields. *Scand. J. Statist.*, **29**, 31–50.
Shephard, N. (1994) Partial non-Gaussian state space. *Biometrika*, **81**, 115–131.
Shephard, N. (2000) Discussion on 'Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives' (by J. Durbin and S. J. Koopman). *J. R. Statist. Soc.* B, **62**, 30–32.
Shephard, N. and Pitt, M. K. (1997) Likelihood analysis of non-Gaussian measurement time series. *Biometrika*, **84**, 653–667.
Tierney, L., Kass, R. E. and Kadane, J. B. (1989) Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *J. Am. Statist. Ass.*, **84**, 710–716.
West, M. and Harrison, J. (1997) *Bayesian Forecasting and Dynamic Models*, 2nd edn. New York: Springer.
Wikle, C. K., Berliner, L. M. and Cressie, N. A. (1998) Hierarchical Bayesian space-time models. *Environ. Ecol. Statist.*, **5**, 117–154.