NORGES TEKNISK-NATURVITENSKAPELIGE UNIVERSITET

Approximate Bayesian Inference in Spatial Generalized Linear Mixed Models

by

Jo Eidsvik, Sara Martino and Håvard Rue

PREPRINT
STATISTICS NO. 2/2006

NORWEGIAN UNIVERSITY OF SCIENCE AND TECHNOLOGY
TRONDHEIM, NORWAY

This report has URL http://www.math.ntnu.no/preprint/statistics/2006/S2-2006.ps
Jo Eidsvik has homepage: http://www.math.ntnu.no/∼joeid
E-mail: joeid@stat.ntnu.no
Address: Department of Mathematical Sciences, Norwegian University of Science and Technology,
N-7491 Trondheim, Norway.

Approximate Bayesian Inference in Spatial Generalized Linear Mixed Models

Jo Eidsvik, Sara Martino and Håvard Rue
Department of Mathematical Sciences, NTNU, Norway.
Address: Department of Mathematical Sciences, NTNU, 7491 Trondheim, NORWAY.
Corresponding author: Jo Eidsvik (joeid@math.ntnu.no)

ABSTRACT

In this paper we propose fast approximate methods for computing posterior marginals in spatial generalized linear mixed models. We consider the common geostatistical special case with a high dimensional latent spatial variable and observations at only a few known registration sites. Our methods of inference are deterministic, using no random sampling.

We present two methods of approximate inference. The first is very fast to compute and via examples we find that this approximation is 'practically sufficient'. By this expression we mean that the results obtained by this approximate method do not show any bias or dispersion effects that might affect decision making. The other approximation is an improved version of the first one, and via examples we demonstrate that the inferred posterior approximations of this improved version are 'practically exact'. By this expression we mean that one would have to run Markov chain Monte Carlo simulations for longer than is typically done to detect any indications of bias or dispersion error effects in the approximate results.

The two methods of approximate inference can help to expand the scope of geostatistical models, for instance in the context of model choice, model assessment, and sampling design. The approximations take seconds of CPU time, in sharp contrast to overnight Markov chain Monte Carlo runs for solving these types of problems. Our approach to approximate inference could easily be part of standard softwares.

KEYWORDS: *Approximate inference, spatial GLM, circulant covariance matrix, Newton-Raphson, MCMC.*

1. INTRODUCTION

Several statistical problems include the analysis of data acquired at various spatial locations. In Bayesian geostatistical modeling one typically treats these data as indirect measurements of a smooth latent spatial variable, with priors for the parameters of the model. Among the popular applications are geophysics, mining, meteorology and disease mapping, see e.g. Cressie (1991), Diggle, Tawn and Moyeed (1998) and Banerjee, Carlin and Gelfand (2004). For Bayesian analysis of spatial data there are mainly two objectives; i) inference of model parameters, for example the standard deviation (or precision) and the range of the latent spatial variable, and ii) prediction of the latent variable at any spatial location.

One topic that has received much attention lately is inference and prediction for the spatial generalized linear mixed model (GLMM), see e.g. Diggle et al. (1998) and Christensen, Roberts and Sköld (2006). This common model can briefly be described as follows: Let $\boldsymbol{x}$ represent a latent spatial variable on a grid of n regularly spaced gridnodes on a two dimensional lattice. Suppose that $\boldsymbol{x}$ has a stationary Gaussian prior distribution specified by some covariance model parameters $\boldsymbol{\theta}$. Suppose next that observations $\boldsymbol{y}$ are made at k of the n nodes. These observations are modeled as an exponential family distribution with parameters given by the latent variable $\boldsymbol{x}$ at the sites where the data is acquired. Typical examples of this model include Poisson counts or binomial proportions registrered at some known locations in space, with the objective of predicting the underlying intensity or (log odds) risk surface across the spatial domain of interest, and inferring model parameters. The most common case is probably the situation where one wants to predict across a large spatial domain, but only a few locations register data, i.e. $n \gg k$. For example, this situation occurs in spatial data acquisition for weather forecasting (Gel, Raftery and Gneiting 2004) and in reserve site selection for predicting the presence of a certain type of species (Polasky and Solow 2001). Both examples in Diggle et al. (1998) are also of this type with $n \gg k$.

Bayesian analysis of spatial GLMMs have been considered difficult since the spatial problem is of high dimension and because of the lack of closed form solutions. The current state of the art is to generate realizations of parameter $\boldsymbol{\theta}$ and latent spatial variable $\boldsymbol{x}$ using Markov chain Monte Carlo (MCMC) algorithms. Since MCMC algorithms have grown mature over the last few decades, see e.g. Robert and Casella (2004), there is a number of fit-for-purpose algorithmic techniques for doing iterative Markov chain updates. Some of these algorithms are more relevant for spatial GLMMs (Diggle et al. 2003), but problems remain with convergence and mixing properties of the Markov chain, which in some cases are remarkably slow. Because of these challenges fast inference methods suitable for special cases are needed, possibly avoiding the problems with sampling methods.

The main contribution of this paper is a new method for approximate inference in spatial GLMMs with $n \gg k$. In particular, this paper provides a recipe for fast approximate Bayesian inference using the marginals $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ and $\pi(x_j|\boldsymbol{y})$, $j = 1, \ldots, n$, i.e. the marginal posterior density of the model parameters and the marginal posterior density of the latent variable at any spatial location. We also illustrate how the marginal likelihood $\pi(\boldsymbol{y})$ can be estimated within our framework. The examples show that the fast approximate method is 'practically sufficient'. By this expression we mean that results obtained by the approximate method and tedious Monte Carlo algorithms are hard to distinguish and that for most practical decision purposes the approximate method obtains in almost no time what the Monte Carlo approach would need much computation time to achieve. This result is important for general practitioners of geostatistics that can possibly avoid iterative Monte Carlo simulations which are hard to monitor, and rather do the direct calculation at almost no computational cost. Moreover, the approximate method could easily be incorporated in standard softwares. Fast approximate inference for this model can further help to expand the scope of geostatistical modeling. Possible applications include geostatistical design (Diggle and Lophaven 2006), and model choice (Clyde and George 2004) or model assessment (Johnson 2004) in a geostatistical setting.

Another contribution of this paper is an improved approximation for spatial prediction, going beyond our basic direct approximate solution. While the direct approximation uses the joint posterior mode for $\boldsymbol{x}$ to do spatial prediction, the improved approximation splits the joint into the marginal of x_j and the conditional for the remaining spatial nodes. The improved version provides a natural correction to the direct approximation. Improvements become important at spatial locations j where the direct approximation to $\pi(x_j|\boldsymbol{y}, \boldsymbol{\theta})$ or $\pi(x_j|\boldsymbol{y})$ is not sufficiently accurate, typically spatial sites where there are lots of non-Gaussian data. In our examples we find that this improved approximation is 'practically exact'. By this expression we mean that results obtained by the approximate approach and tedious Monte Carlo algorithms are indistinguishable, and that it is very hard to detect if small differences indicate limited run time in the Monte Carlo method or slight bias in the improved approximate approach.

Other recent approaches to approximate inference for spatial generalized linear models include Breslow and Clayton (1993) who studied penalized quasi likelihood, Ainsworth and Dean (2006) who compared penalized quasi likelihood with a Bayesian solution based on MCMC computations, and Rue and Martino (2006a) who used the Laplace approximation that we consider here, but with a Gaussian Markov random field for the latent variable. The above references all used examples where k is of the same order as n.

The outline is as follows: In Section 2 we define the special case of spatial GLMMs considered in this paper. The proposed method of approximate inference and prediction is described in Section 3, while Section 4 uses the Rongelap radionuclide dataset as an example of this method. After the ideas and results of the direct approximation have been presented, we describe the improved approximation for spatial prediction in Section 5, and use the Lancashire infection dataset as an example of this in Section 6. We discuss and conclude in Section 7. The computational aspects of our methods are postponed to the Appendix.

2. SPATIAL GLMM

Let $\boldsymbol{x} = (x_1, \ldots, x_n)'$ represent the latent field on a regular grid of size $n = n_1 n_2$, where n_1 and n_2 are the grid sizes in the North and East directions. In an application with binomial proportions data, the spatial variable $\boldsymbol{x}$ would denote the latent risk or log odds surface, while it would denote the latent log intensity surface for Poisson count data. Suppose $\boldsymbol{x}$ has a stationary Gaussian prior with $\pi(\boldsymbol{x}|\boldsymbol{\theta}, \mu) = N[\boldsymbol{x}; \mu \mathbf{1}_n, \boldsymbol{\Sigma}(\boldsymbol{\theta})]$, where $\mathbf{1}_n$ denotes an $n \times 1$ vector of ones, and $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$ is a positive definite, block circulant covariance matrix with $\boldsymbol{\theta}$ indicating the model parameters. For example, $\boldsymbol{\theta}$ could include $\sigma =$ pointwise standard deviation, and $\nu =$ spatial correlation range. Block circulant covariance structure means that the $n_1 \times n_2$ grid is wrapped on a torus. As a simple example of a covariance function we give the exponential defined by

$$\Sigma_h(\sigma, \nu) = \sigma^2 \exp(-\delta h/\nu), \qquad h = \sqrt{h_1^2 + h_2^2}, \tag{1}$$

where (h_1, h_2) are the (North, East) gridsteps between two nodes on the torus surface, while δ specifies the spacing on the grid. Many others are possible, such as the Matern covariance function which is often recommended, see e.g. Cressie (1991) and Stein (1999). We discuss this more general class of covariance functions in the context of an application. The covariance function imposes dependence in the latent variable, and in practice this means a smooth underlying risk or intensity surface, as one could expect. A Bayesian view is taken here with priors $\pi(\mu) = N(\mu; \beta_0, \tau^2)$ and $\pi(\boldsymbol{\theta})$ for the parameters. The μ mean parameter can then be integrated out to obtain $\pi(\boldsymbol{x}|\boldsymbol{\theta}) = N(\boldsymbol{x}; \boldsymbol{\beta}, \boldsymbol{C})$, $\boldsymbol{\beta} = \beta_0 \mathbf{1}_n$ and $\boldsymbol{C} = \mathbf{1}_n \tau^2 \mathbf{1}_n' + \boldsymbol{\Sigma}$, a block circulant matrix. Block circulant covariance structure means that the fast Fourier transform can be used as an efficient computational tool, see Appendix.

Suppose next that measurements y_i, $i = 1, \ldots, k$, are conditionally independent with likelihood

$$\pi(y_i | x_{s_i}) = \exp\{[y_i x_{s_i} - b(x_{s_i})]/a(\phi) + c(\phi, y_i)\}, \quad i = 1, \ldots, k, \tag{2}$$

where $\boldsymbol{x}_s = (x_{s_1}, \ldots, x_{s_k})' = \boldsymbol{A}\boldsymbol{x}$ and the $k \times n$ matrix $\boldsymbol{A}$ has entries

$$A_{ij} = I(s_i = j) = \left\{ \begin{array}{ll} 1 & \text{if} \quad s_i = j \\ 0 & \text{else} \end{array} \right. , \quad i = 1, \ldots, k, \quad j = 1, \ldots, n, \tag{3}$$

i.e. s_i is the gridnode of measurement i. With $n \gg k$ the matrix $\boldsymbol{A}$ consists mostly of 0 values, but it has one 1 value for each row / observation. For the exponential family distribution in equation (2) we have simple functional relationships for $b(x)$ and $a(\phi)$. For example, the Poisson, Binomial and Gaussian distributions can be defined by

$$\begin{array}{r|lll} & \text{Poisson} & \text{Binomial} & \text{Gaussian} \\ \hline b(x) & m\exp(x) & m\log[1+\exp(x)] & x^2/2 \\ a(\phi) & 1 & 1 & \phi^2 \end{array} , \tag{4}$$

where m is a fixed parameter and ϕ is a fixed standard deviation of the Gaussian likelihood model. These relationships are commonly used in generalized linear models (McCullagh and Nelder 1989). For example, Poisson is a case of log link function for count data, while the Binomial likelihood is a case of logit link function for proportions. The model treated in this paper is different from the standard GLMM setting because the data $\boldsymbol{y}$ are acquired at various spatial sites and because of the spatially correlated latent variable $\boldsymbol{x}$. Hence the term spatial GLMM.

3. APPROXIMATE INFERENCE AND PREDICTION

We will now outline our methods for fast approximate analysis of Bayesian spatial GLMMs with $n \gg k$. The first step is to use a Gaussian approximation at the mode of the conditional density $\pi(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})$. We next use this full conditional to approximate the densities of main interest; $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ and $\pi(x_j|\boldsymbol{y})$. The marginal likelihood $\pi(\boldsymbol{y})$ is also approximated. We have chosen to postpone the technical details to the Appendix.

3.1 Gaussian approximation for $\pi(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})$ at the posterior mode

The full conditional density of the latent spatial variable is

$$\pi(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}) \propto \exp\{-\frac{1}{2}\boldsymbol{x}'\boldsymbol{C}^{-1}\boldsymbol{x} + \boldsymbol{x}'\boldsymbol{C}^{-1}\boldsymbol{\beta} + \sum_{i=1}^{k}[y_i x_{s_i} - b(x_{s_i})]/a(\phi)\}. \tag{5}$$

A Gaussian approximation to this density is constructed by linearizing the likelihood part of equation (5) at a fixed location $\boldsymbol{x}_s^0 = \boldsymbol{A}\boldsymbol{x}^0$. For each $i = 1, \ldots, k$ this involves

$$[y_i x_{s_i} - b(x_{s_i})] \approx [y_i x_{s_i}^0 - b(x_{s_i}^0)] + (x_{s_i} - x_{s_i}^0)[y_i - b'(x_{s_i}^0)] - \frac{1}{2}(x_{s_i} - x_{s_i}^0)^2 b''(x_{s_i}^0), \qquad (6)$$

where the first and second derivatives can be written in closed form using equation (4). Inserting this into equation (5) gives a Gaussian approximation $N[\boldsymbol{x}; \hat{\boldsymbol{\mu}}_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{x}^0), \hat{\boldsymbol{V}}_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}}(\boldsymbol{x}^0)]$. The conditional covariance equals

$$\hat{\boldsymbol{V}}_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}}(\boldsymbol{x}^0) = \boldsymbol{C} - \boldsymbol{C}\boldsymbol{A}'\boldsymbol{R}^{-1}\boldsymbol{A}\boldsymbol{C}, \qquad \boldsymbol{R} = \boldsymbol{A}\boldsymbol{C}\boldsymbol{A}' + \boldsymbol{P}, \qquad (7)$$

where $\boldsymbol{P}$ is diagonal with entries $P_{i,i} = a(\phi)/b''(x_{s_i}^0)$, $i = 1, \ldots, k$. The conditional mean is

$$\begin{aligned} \hat{\boldsymbol{\mu}}_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}}(\boldsymbol{y}, \boldsymbol{x}^0) &= [\boldsymbol{C}^{-1} + \boldsymbol{A}'\boldsymbol{P}^{-1}\boldsymbol{A}]^{-1}[\boldsymbol{C}^{-1}\boldsymbol{\beta} + \boldsymbol{A}'\boldsymbol{P}^{-1}\boldsymbol{z}(\boldsymbol{y}, \boldsymbol{x}_s^0)], \\ &= \boldsymbol{\beta} + \boldsymbol{C}\boldsymbol{A}'\boldsymbol{R}^{-1}[\boldsymbol{z}(\boldsymbol{y}, \boldsymbol{x}_s^0) - \boldsymbol{A}\boldsymbol{\beta}], \qquad (8) \end{aligned}$$

where we use

$$z_i(y_i, x_{s_i}^0) = [y_i - b'(x_{s_i}^0) + x_{s_i}^0 b''(x_{s_i}^0)]/b''(x_{s_i}^0), \qquad i = 1, \ldots, k. \qquad (9)$$

Rather than fitting a Gaussian approximation at any $\boldsymbol{x}^0$, the above process is iterated using the Newton-Raphson algorithm. After a few iterations we have then fitted a Gaussian approximation $\hat{\pi}(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})$ at the argument of the posterior mode denoted by $\hat{\boldsymbol{m}}_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}}$, see Appendix. These Newton-Raphson calculations are similar to the traditional ones used for generalized linear models (McCullagh and Nelder 1989), but note that the dimension n of the latent variable $\boldsymbol{x}$ is large here and regular optimization would often be impossible for spatial models, except in small situations such as a limited number of counties (Breslow and Clayton 1993). One contribution of the current paper is to demonstrate that the Newton-Raphson optimization can be done in a fast manner using benefits of the Fourier domain. The effective method applies to our special case with $n \gg k$ and stationary prior density for the latent variable. This indicates that large problems of this common type can be handled with modest cost.

We remark some relevant features of the Gaussian approximation $\hat{\pi}(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})$. The computational details of these remarks are presented in the Appendix.

- The latter expression in equation (8) involves the inversion of a $k \times k$ matrix $\boldsymbol{R}$ and is clearly preferable in our case with $n \gg k$.

- Equation (8) is computed efficiently in the Fourier domain since $\boldsymbol{C}$ is block circulant.

- The Gaussian approximation $\hat{\pi}(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})$ is evaluated using Bayes theorem and benefits of the Fourier domain.

- Newton-Raphson optimization involves iterative calculation of equation (8) and is hence also available in the Fourier domain.

The quality of the Gaussian approximation depends on the particular situation. Intuitively one would expect it to be quite good since $n \gg k$ and hence the smooth Gaussian prior has much influence.

3.2 Parametric inference using $\pi(\boldsymbol{\theta}|\boldsymbol{y})$

The marginal density of model parameters $\boldsymbol{\theta}$ is

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\pi(\boldsymbol{y}|\boldsymbol{x})\pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{y})\pi(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})} \propto \frac{\pi(\boldsymbol{y}|\boldsymbol{x})\pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})}, \tag{10}$$

which is valid for any value of the spatial variable $\boldsymbol{x}$, for example $\boldsymbol{x} = \hat{\boldsymbol{m}}_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}}$, the argument at the posterior mode for fixed $\boldsymbol{\theta}$. The challenging part of equation (10) is the denominator $\pi(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})$ which is not available in closed form. We choose to approximate this denominator using the fitted Gaussian density in Section 3.1. The approximate density for the model parameters is then

$$\hat{\pi}(\boldsymbol{\theta}|\boldsymbol{y}) \propto \left. \frac{\pi(\boldsymbol{y}|\boldsymbol{x})\pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\hat{\pi}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})} \right|_{\boldsymbol{x}=\hat{\boldsymbol{m}}_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}}}. \tag{11}$$

Equation (11) is the Laplace approximation, see e.g. Tierney and Kadane (1986) and Carlin and Louis (2000). The relative error is $O(k^{-3/2})$ (Tierney and Kadane 1986). Relative error is advantageous, especially in the tails of the distribution. Equation (11) can be evaluated efficiently using properties of the Gaussian approximation, see Appendix.

To calculate the approximation in equation (11) we first locate the mode of $\hat{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ and fit the dispersion from the Hessian at this mode. In the relevant domain of the sample space for $\boldsymbol{\theta}$ we then evaluate equation (11) for a finite set of parameter values $\boldsymbol{\theta}_l = (\sigma_{l_1}, \nu_{l_2})$, $l_1 = 1, \ldots, L_1$, $l_2 = 1, \ldots, L_2$, normalized so that

$$\sum_l \Delta_\sigma \Delta_\nu \hat{\pi}(\boldsymbol{\theta}_l|\boldsymbol{y}) = \Delta_\sigma \Delta_\nu \sum_{l_1=1}^{L_1} \sum_{l_2=1}^{L_2} \hat{\pi}(\sigma_{l_1}, \nu_{l_2}|\boldsymbol{y}) = 1. \tag{12}$$

In this equation Δ_σ and Δ_ν are the spacings in the defined grid for σ and ν values. The approximate marginal densities are

$$\hat{\pi}(\sigma_{l_1}|\boldsymbol{y}) = \Delta_\sigma \sum_{l_2=1}^{L_2} \hat{\pi}(\sigma_{l_1}, \nu_{l_2}|\boldsymbol{y}), \qquad \hat{\pi}(\nu_{l_2}|\boldsymbol{y}) = \Delta_\nu \sum_{l_1=1}^{L_1} \hat{\pi}(\sigma_{l_1}, \nu_{l_2}|\boldsymbol{y}). \tag{13}$$

The density function $\hat{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ can also be approximated differently, for example by a parametric fit to the density or by numerical quadrature (Press, Teukolsky, Vetterling and Flannery 1996). We do not consider this here.

For the case with a Gaussian likelihood in equation (2) we can evaluate $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ exactly on the grid of $\boldsymbol{\theta}$ values (Diggle et al. 2003), whereas for the non-Gaussian case we merely obtain an approximation $\hat{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ on this set of parameter values. Note that constructing $\hat{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ can be done using only $\boldsymbol{x}_s = \boldsymbol{A}\boldsymbol{x}$, i.e. latent values at the registration sites, but in this presentation we have chosen to use the entire spatial variable $\boldsymbol{x}$ because this brings together the inference and prediction steps.

3.3 Estimation of marginal likelihood $\pi(\boldsymbol{y})$

For assessing the statistical model one often uses the marginal likelihood $\pi(\boldsymbol{y})$, see e.g. Hsiao, Huang and Chang (2004). The marginal likelihood is given by

$$\pi(\boldsymbol{y}) = \frac{\pi(\boldsymbol{y}|\boldsymbol{x})\pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta}|\boldsymbol{y})\pi(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})}. \tag{14}$$

The denominator in equation (14) can be approximated using the Gaussian approximation for latent variable and the approximate marginal for model parameters. These are both available. The simplest way to evaluate the marginal likelihood is however to use the normalizing constant required for computing the approximate density in equation (11) and (12). We denote this estimate of the marginal likelihood by $\hat{\pi}(\boldsymbol{y})$. Efficient computation of the marginal likelihood becomes important for example in model choice (Clyde and George 2004). For spatial models one natural model check is related to the choice of covariance function. We demonstrate this in one of the examples below.

3.4 Spatial prediction using $\pi(x_j|\boldsymbol{y})$

For approximate Bayesian spatial prediction we use marginals $\hat{\pi}(x_j|\boldsymbol{y}) = \sum_l \hat{\pi}(x_j|\boldsymbol{y}, \boldsymbol{\theta}_l)\hat{\pi}(\boldsymbol{\theta}_l|\boldsymbol{y})$, $j = 1, \dots, n$. This is a mixture of Gaussian distibutions with the weights denoting the posterior for each model parameter. The approximate marginal means become

$$\hat{\mu}_{x_j|\boldsymbol{y}} \approx \sum_l \hat{m}_{x_j|\boldsymbol{y},\boldsymbol{\theta}_l} \hat{\pi}(\boldsymbol{\theta}_l|\boldsymbol{y}), \quad j = 1, \dots, n, \tag{15}$$

where we pick element j of the joint conditional mode obtained at the last Newton-Raphson optimization step. The approximate marginal variances are

$$\hat{V}_{x_j|\boldsymbol{y}} \approx \sum_l \hat{V}_{x_j|\boldsymbol{y},\boldsymbol{\theta}_l} \hat{\pi}(\boldsymbol{\theta}_l|\boldsymbol{y}) + \sum_l (\hat{m}_{x_j|\boldsymbol{y},\boldsymbol{\theta}_l} - \hat{\mu}_{x_j|\boldsymbol{y}})^2 \hat{\pi}(\boldsymbol{\theta}_l|\boldsymbol{y}), \quad j = 1, \dots, n, \tag{16}$$

where $\hat{V}_{x_j|\boldsymbol{y},\boldsymbol{\theta}} = \hat{V}_{x_j|\boldsymbol{y},\boldsymbol{\theta}}(\hat{\boldsymbol{m}}_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}})$ denotes diagonal element j of the conditional covariance in equation (7), evaluated at the argument of the posterior mode. For these variance terms we need to calculate the diagonal entries of $\boldsymbol{C}\boldsymbol{A}'\boldsymbol{R}^{-1}\boldsymbol{A}\boldsymbol{C}$ given by

$$(\boldsymbol{C}\boldsymbol{A}'\boldsymbol{R}^{-1}\boldsymbol{A}\boldsymbol{C})_{jj} = \sum_{i=1}^{k} \sum_{i'=1}^{k} C_{j,s_i} R^{-1}_{ii'} C_{s_{i'},j}, \quad j = 1, \dots, n. \tag{17}$$

3.5 Approximations as proposal distributions in Monte Carlo sampling

We now discuss Monte Carlo methods for checking the quality of the approximations. The approximations are then used as proposals in a Metropolis–Hastings (MH) algorithm or in importance sampling.

Consider an independent proposal MH algorithm with proposed value $\boldsymbol{x}'$ from the Gaussian approximation $\hat{\pi}(\boldsymbol{x}'|\boldsymbol{y}, \boldsymbol{\theta})$, keeping $\boldsymbol{\theta}$ fixed. The proposal $\boldsymbol{x}'$ is generated efficiently in the Fourier domain, see Appendix. The acceptance rate becomes

$$\min\left\{1, \frac{[\prod_{i=1}^{k} \pi(y_i|x'_{s_i})]\pi(\boldsymbol{x}'|\boldsymbol{\theta})}{[\prod_{i=1}^{k} \pi(y_i|x_{s_i})]\pi(\boldsymbol{x}|\boldsymbol{\theta})} \frac{\hat{\pi}(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})}{\hat{\pi}(\boldsymbol{x}'|\boldsymbol{y}, \boldsymbol{\theta})}\right\}, \tag{18}$$

otherwise we keep the current value $\boldsymbol{x}$ in the Markov chain. The acceptance probability in equation (18) can be evaluated using properties of the Gaussian approximation, see Appendix. This MH scheme might be effective since large changes are proposed at every iteration, see e.g. Robert and

Casella (2004). Simultaneous MH updates of $(\boldsymbol{x}', \boldsymbol{\theta}')$ are also possible using the approximation $\hat{\pi}(\boldsymbol{\theta}'|\boldsymbol{y})$ followed by $\hat{\pi}(\boldsymbol{x}'|\boldsymbol{y}, \boldsymbol{\theta}')$, and then checking for acceptance.

One can alternatively use importance sampling (Shepard and Pitt 1997) based on the approximations presented above. This is done by first generating $\boldsymbol{x}^b$, $b = 1, \ldots, B$, independently from $\hat{\pi}(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})$. The B samples are next weighted based on the ratio of the posterior and the Gaussian approximation. It is again possible to sample $(\boldsymbol{x}^b, \boldsymbol{\theta}^b)$ using $\hat{\pi}(\boldsymbol{\theta}^b|\boldsymbol{y})$ followed by $\hat{\pi}(\boldsymbol{x}^b|\boldsymbol{y}, \boldsymbol{\theta}^b)$ for joint Monte Carlo estimation.

The Monte Carlo error of estimates based on B samples is additive and $O_p(B^{-1/2})$. The error can thus be reduced by increasing the number of samples, but the additive nature might cause problems in the tails of the distribution, especially compared to the relative error of the Laplace approximation.

4. EXAMPLE OF APPROXIMATE INFERENCE

In this Section we redo one of the examples used in Diggle et al. (1998). The data are made at only a few registration sites and the large latent spatial variable is modeled by a stationary prior distribution. The dataset consists of $k = 157$ measurements of y_i = radionuclide counts for various time durations m_i, $i = 1, \ldots, k$. All 157 registration sites are displayed on the map in Figure 1. The data are modeled as a spatial GLMM with a Poisson distribution in equation (2). The goal is to assess the latent intensity surface $\boldsymbol{x}$ and the model parameters $\boldsymbol{\theta} = (\sigma, \nu)$ given the data. For the spatial variable we construct a grid with interval spacing $\delta = 40$m covering the region from $(-4180, -6800)$ to $(640, 700)$ in the (North, East) coordinates displayed in Figure 1. The gridsize is then $n_1 = 103$ (North) and $n_2 = 187$ (East), in total $n \approx 23000$. Following Christensen et al. (2006) we use an exponential covariance function, see equation (1). We use a flat prior for $\boldsymbol{\theta}$, and as hyperparameters we use $\beta_0 = 1.5$ and $\tau^2 = 1$. The value for β_0 is directly calculated as the logarithm of all data scaled with the individual time intervals. Ten Newton-Raphson iterations are used to locate the mode of the Gaussian approximation $\hat{\pi}(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})$. After ten iterations the machine precision is reached.

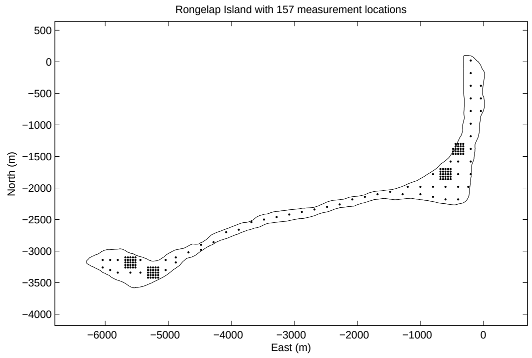


Figure 1: Rongelap island with the 157 registration sites for observations of radionuclide concentrations.

The set of parameter values defined in equation (12) covers $\sigma \in \{0.3, 1\}$ and $\nu \in \{50, 300\}$ with gridsize $L_1 = 50$ and $L_2 = 50$. The approximate density $\hat{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ obtained by equation (11) and (12) is shown in Figure 2 (left) along with marginals $\hat{\pi}(\sigma|\boldsymbol{y})$ and $\hat{\pi}(\nu|\boldsymbol{y})$ in Figure 2 (right,

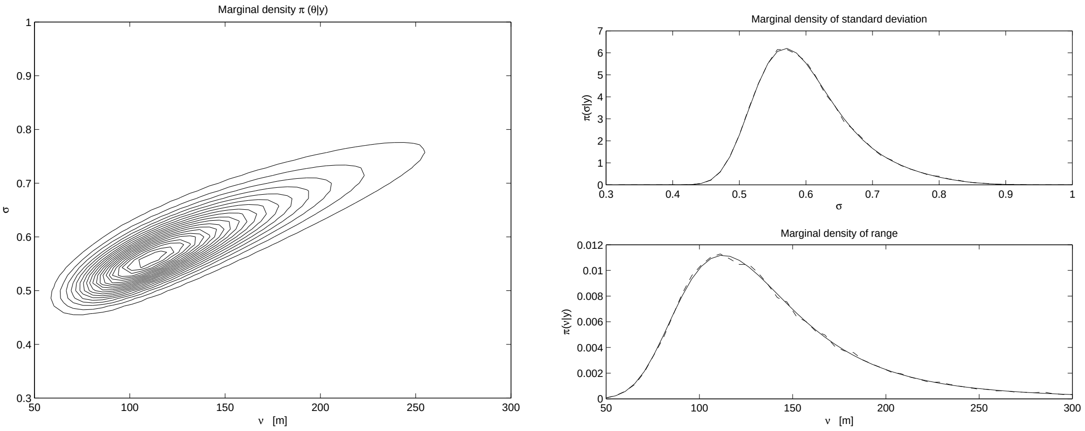


Figure 2: Rongelap dataset. Left) Direct approximation of joint density of parameters $\boldsymbol{\theta} = (\sigma, \nu)$. Right) Direct approximation of marginal densities for standard deviation σ and spatial correlation range ν (solid). MH approximation of marginal densities (dashed).

solid curve). Figure 2 (left) seems similar to results obtained earlier by MCMC sampling, see e.g. Christensen et al. (2006). This is also visible from Figure 2 (right, dashed curves) which displays estimates of the marginals using MH sampling with joint updating of $\boldsymbol{x}$ and $\boldsymbol{\theta}$, as described in Section 3.5. Since the solid and dashed curves in Figure 2 (right) are hard to distinguish, the direct Laplace approximation appears to be very good.

In Figure 3 we show the marginal predictions in equation (15) and standard deviations given by equation (16). We recognize the measurement locations in the standard deviations and see that the standard deviations climb to a level of about 0.6 as one goes about 500m away from measurement locations. Similarly, the spatial predictions in Figure 3 (left) are near the prior mean as we go away from the island with several measurement locations. The main trends are similar to the ones obtained by MCMC sampling in Diggle et al. (1998).

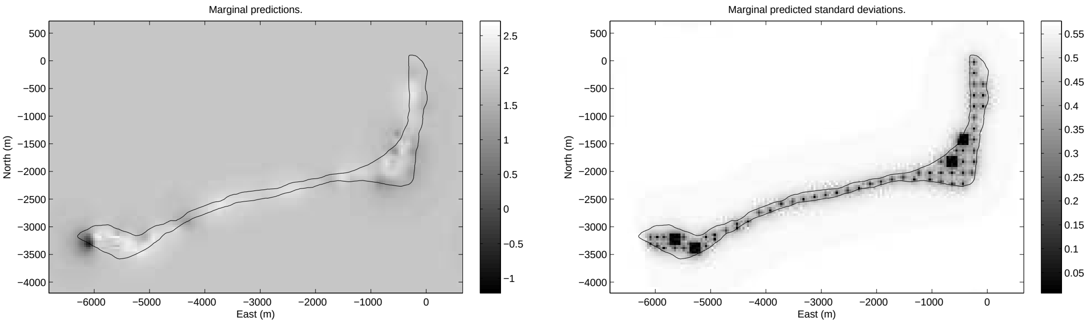


Figure 3: Rongelap dataset. Left) Predicted spatial variable. Right) Marginal standard deviation of spatial variable.

We go on to check the Gaussian approximation that we use for the latent variable via $\hat{\pi}(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})$. For this purpose we use Monte Carlo sampling as described in Section 3.5 above. We choose to evaluate three approximations (direct Gaussian approximation, MH method and importance sampling) at spatial locations $(-1800, -600)$ and $(-1850, -1000)$. These two are chosen since they represent locations near and far from registration sites, respectively, see Figure 1. We perform the testing for parameter $\boldsymbol{\theta} = (0.6, 152)$, regarded to be a likely parameter value (Figure 2). In Figure 4 (solid) we show the approximate densities $\hat{\pi}(x_j|\boldsymbol{y}, \boldsymbol{\theta})$, where j correponds to the two spatial locations. Also displayed in Figure 4 (dashed and dotted) are approximations based on independent proposal MH (dashed) and importance sampling (dotted). The results of the various approximations displayed in Figure 4 are very similar. Specifically we note that the result of direct approximate inference is hardly distinguishable from the two Monte Carlo approximations. The small fluctuations in Figure 4 (solid and dashed) are caused by Monte Carlo error. This error itself is larger than the differences between the direct approximation and the Monte Carlo estimates.

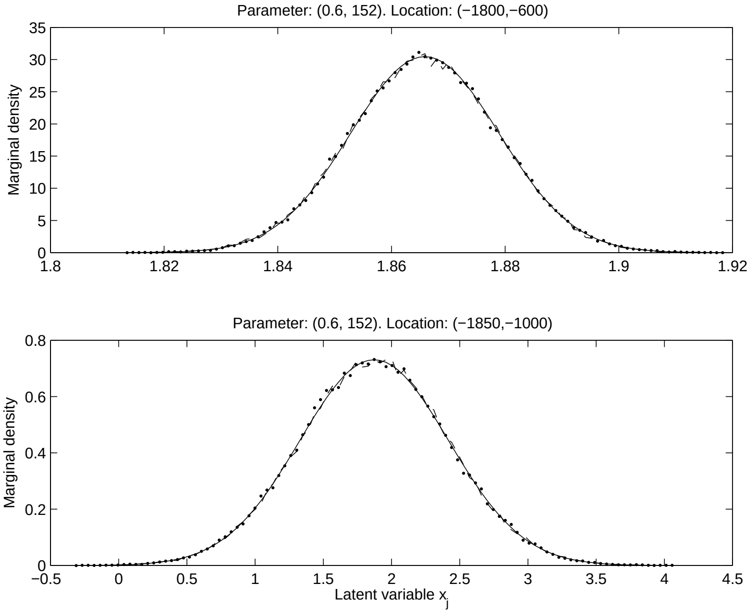


Figure 4: Rongelap dataset. Conditional density $\hat{\pi}(x_j|\boldsymbol{y}, \boldsymbol{\theta})$ obtained by approximate inference at two spatial locations and for parameter values fixed at ($\sigma = 0.6$, $\nu = 152$m). Solid is Gaussian approximation, dashed is approximation obtained by MH sampling, and dotted is approximation from importance sampling.

Figure 2 and 4 show that the results obtained by approximate inference are very similar to Monte Carlo results, and we note that approximate inference is sufficiently accurate for all practical purposes in this example. Possible bias effects are so small that it would have no impact on the decisions made concerning this application. We hence use the term 'practically sufficient' for our approximation.

The Monte Carlo algorithms used 100000 proposals. The acceptance probability of the MH algorithm with joint updating was 0.8, indicating that the approximation is quite good. Importance sampling resulted in an effective sample size of 90000, indicative of small variability in the weights for different proposals. For the plotting of Monte Carlo approximations in Figure 4 we split the sample space of x_j into 100 disjoint regions in the interval $\hat{\mu}_{x_j|\boldsymbol{y},\boldsymbol{\theta}} \pm 4\sqrt{\hat{V}_{x_j|\boldsymbol{y},\boldsymbol{\theta}}}$, and thus created the estimated density curve (dashed and dotted). We avoided smoothing this curve estimate in order to visualize some of the Monte Carlo error.

We finally study marginal likelihood values $\pi(\boldsymbol{y})$ for various spatial covariance functions. The choice of covariance model for the latent variable is hence checked. For this purpose we implement the more general Matern class of covariance functions which is defined by

$$\Sigma_h(\sigma, \nu, \kappa) = \sigma^2 \frac{\tau^\kappa K(\tau, \kappa)}{2^{\kappa-1}\Gamma(\kappa)}, \quad \tau = \alpha_\kappa \delta h/\nu, \quad h = h_1^2 + h_2^2, \tag{19}$$

where $K(\cdot, \kappa)$ denotes a modified Bessel function of order κ and $\Gamma(\cdot)$ is the Gamma function. In equation (19) the α_κ parameter is set so that the correlation is approximately 0.05 at spatial distance $h\delta = \nu$. The Matern family contains the exponential covariance in equation (1) as a special case when $\kappa = 0.5$, while it reduces to the Gaussian (squared exponential) covariance function when $\kappa = \infty$. We calculate the marginal likelihood estimate $\hat{\pi}(\boldsymbol{y})$ in Section 3.3 for four different Matern covariance models. These models are i) exponential covariance ($\kappa = 0.5$), ii) $\kappa = 1$, iii) $\kappa = 2$, and iv) Gaussian covariance ($\kappa = \infty$). The marginal likelihood is largest for the exponential covariance in this dataset. The difference in log marginal likelihood is 5.6 when comparing the exponential with case ii), 9.9 when comparing the exponential with case iii), while it is 19.6 in favor of the exponential over the Gaussian covariance. Hence, there is evidence of a steep exponential decline in covariance at zero distance for this dataset.

Our current prototype for direct approximate inference is implemented in Matlab. We are working on an implementation in C where we avoid using the $L_1 \times L_2$ grid over parameter space and use less iterations in the Newton–Raphson search for the conditional mode. The estimated run time to get the entire inference and prediction solution is then 10-15 seconds.

5. IMPROVED APPROXIMATION FOR $\pi(x_j|\boldsymbol{y}, \boldsymbol{\theta})$

The direct approximation to the density of x_j conditional on data $\boldsymbol{y}$ and fixed hyperparameters $\boldsymbol{\theta}$ is $\hat{\pi}(x_j|\boldsymbol{y}, \boldsymbol{\theta}) = N(x_j; \hat{m}_{x_j|\boldsymbol{y},\boldsymbol{\theta}}, \hat{V}_{x_j|\boldsymbol{y},\boldsymbol{\theta}})$, defined by picking the j-th index of the argument at the conditional mode $\hat{\boldsymbol{m}}_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}}$ and the j-th diagonal entry of the covariance $\hat{\boldsymbol{V}}_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}} = \hat{\boldsymbol{V}}_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}}(\hat{\boldsymbol{m}}_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}})$. In this Section we illustrate a method for constructing a more accurate approximation $\tilde{\pi}(x_j|\boldsymbol{y}, \boldsymbol{\theta})$. The improved version is valuable for two reasons: i) It is more accurate than the direct approach, and ii) If it is indistinguishable from the direct approximation, the direct one is checked and confirmed without Monte Carlo sampling. We present the improved version within the context of our geostatistical special case. A more general description of this approach is presented in Rue and Martino (2006b).

The improved version is based on

$$\pi(x_j|\boldsymbol{y}, \boldsymbol{\theta}) \propto \frac{\pi(\boldsymbol{y}|\boldsymbol{x})\pi(\boldsymbol{x}|\boldsymbol{\theta})}{\pi(\boldsymbol{x}_{-j}|x_j, \boldsymbol{y}, \boldsymbol{\theta})}, \quad j = 1, \ldots, n, \tag{20}$$

where $\boldsymbol{x}_{-j} = (x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n)'$. For the improved approximation we use a Gaussian approximation for $\boldsymbol{x}_{-j}$ conditional on x_j in the denominator of equation (20). The approximate marginal, denoted $\tilde{\pi}(x_j|\boldsymbol{y}, \boldsymbol{\theta})$, can be evaluated at a set of x_j values and normalized. Note that this marginal in equation (20) is based on conditioning on x_j and using a Laplace approximation to cancel out the remaining variables $\boldsymbol{x}_{-j}$. It is hence more accurate than the direct approach which fits a Gaussian as the joint distribution for all variables. In the choice of evaluation points for x_j in equation (20) we are guided by $\hat{m}_{x_j|\boldsymbol{y},\boldsymbol{\theta}}$ and $\sqrt{\hat{V}_{x_j|\boldsymbol{y},\boldsymbol{\theta}}}$ which are already avaiable from the direct Gaussian approximation in Section 3.

We fit $\tilde{\pi}(\boldsymbol{x}_{-j}|x_j, \boldsymbol{y}, \boldsymbol{\theta})$ in the denominator of equation (20) by introducing a fictitious measurement $\tilde{x}_j$ defined by $\pi(\tilde{x}_j|x_j) = N(\tilde{x}_j; x_j, \zeta^2)$, where $\zeta = 1^{-6}$, a very small number. In practice this means that x_j is fixed at the value of the fictitious measurement $\tilde{x}_j$. The denominator in equation (20) is then interpreted as

$$\tilde{\pi}(\boldsymbol{x}_{-j}|x_j, \boldsymbol{y}, \boldsymbol{\theta}) = \tilde{\pi}(\boldsymbol{x}|\tilde{\boldsymbol{z}}, \boldsymbol{\theta}), \quad \tilde{\boldsymbol{z}} = [\boldsymbol{z}(\boldsymbol{y}, \hat{\boldsymbol{m}}_{\boldsymbol{x}_s|\boldsymbol{y},\boldsymbol{\theta}}), \tilde{x}_j]. \tag{21}$$

We choose to evaluate equation (20) at the mean of this density given by

$$\tilde{\boldsymbol{\mu}}_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}} = \boldsymbol{\beta} + \boldsymbol{C}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{R}}^{-1}(\tilde{\boldsymbol{z}} - \tilde{\boldsymbol{A}}\boldsymbol{\beta}), \quad \tilde{\boldsymbol{R}} = \tilde{\boldsymbol{A}}\boldsymbol{C}\tilde{\boldsymbol{A}}' + \tilde{\boldsymbol{P}}, \tag{22}$$

where the matrices $\tilde{\boldsymbol{A}}$ and $\tilde{\boldsymbol{P}}$ are

$$\tilde{\boldsymbol{A}} = \begin{bmatrix} \boldsymbol{A} \\ \boldsymbol{a}_j \end{bmatrix} \qquad \tilde{\boldsymbol{P}} = \begin{bmatrix} \boldsymbol{P} & 0 \\ 0 & \zeta^2 \end{bmatrix}, \tag{23}$$

and $\boldsymbol{a}_j$ is a $1 \times n$ vector of zeros except for entry j which equals one. The mean in equation (22) is computed efficiently in the Fourier domain using that $\boldsymbol{C}$ is block circulant, see Appendix. The computationally demanding part of the improved approximation is factorizing the term involving $(k+1) \times (k+1)$ covariance matrix $\tilde{\boldsymbol{R}}$.

For the evaluation step we use that

$$\tilde{\pi}(\boldsymbol{x}|\tilde{\boldsymbol{z}}, \boldsymbol{\theta}) = \frac{\tilde{\pi}(\tilde{\boldsymbol{z}}|\boldsymbol{x}, \boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})}{\tilde{\pi}(\tilde{\boldsymbol{z}}|\boldsymbol{\theta})}, \quad \boldsymbol{x} = \tilde{\boldsymbol{\mu}}_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}}. \tag{24}$$

Each term in equation (24) is available, and in particular, $\tilde{\pi}(\tilde{\boldsymbol{z}}|\boldsymbol{x}, \boldsymbol{\theta}) = N(\tilde{\boldsymbol{z}}; \tilde{\boldsymbol{A}}\boldsymbol{x}, \tilde{\boldsymbol{P}})$ and $\tilde{\pi}(\tilde{\boldsymbol{z}}|\boldsymbol{\theta}) = N(\tilde{\boldsymbol{z}}; \tilde{\boldsymbol{A}}\boldsymbol{\beta}, \tilde{\boldsymbol{R}})$. The prior for the latent variable, which is also included in the last equation, cancels when we plug in equation (24) to evaluate $\tilde{\pi}(x_j|\boldsymbol{y}, \boldsymbol{\theta})$ in equation (20).

Recall that the improved approximation $\tilde{\pi}(x_j|\boldsymbol{y}, \boldsymbol{\theta})$ is an additional calculation after the direct Gaussian approximation has been fitted at the joint mode. The additional calculations are i) finding the conditional mean for fixed x_j, see equation (22), and ii) evaluating the approximate marginal in equation (20) at this conditional mean. Alternatively, we could evaluate the improved approximation at the new conditional mode, but using the mean is faster and could be as accurate (Hsiao et al. 2004).

An improved approximation for the marginal $\pi(x_j|\boldsymbol{y})$ can be obtained by integrating out the model parameter $\boldsymbol{\theta}$, like we did in Section 3.4. Our experience is that the difference between direct and improved approximations is larger for fixed $\boldsymbol{\theta}$.

6. EXAMPLE OF IMPROVED APPROXIMATION

For illustrating the improved approximation we consider the other example in Diggle et al. (1998) consisting of $k = 238$ measurements of campylobacter infection in Lancashire district, see Figure 5. The observations y_i = number of campylobacter infection out of enteric infections m_i,

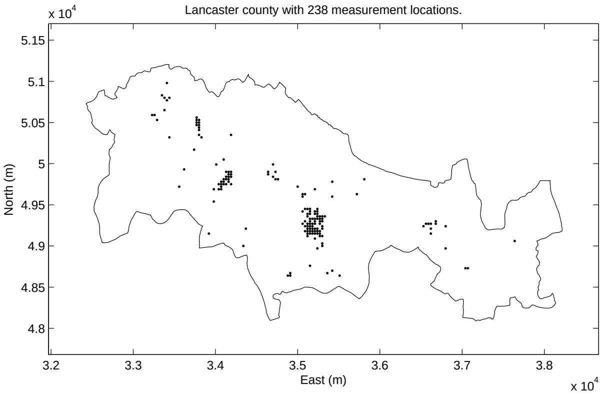


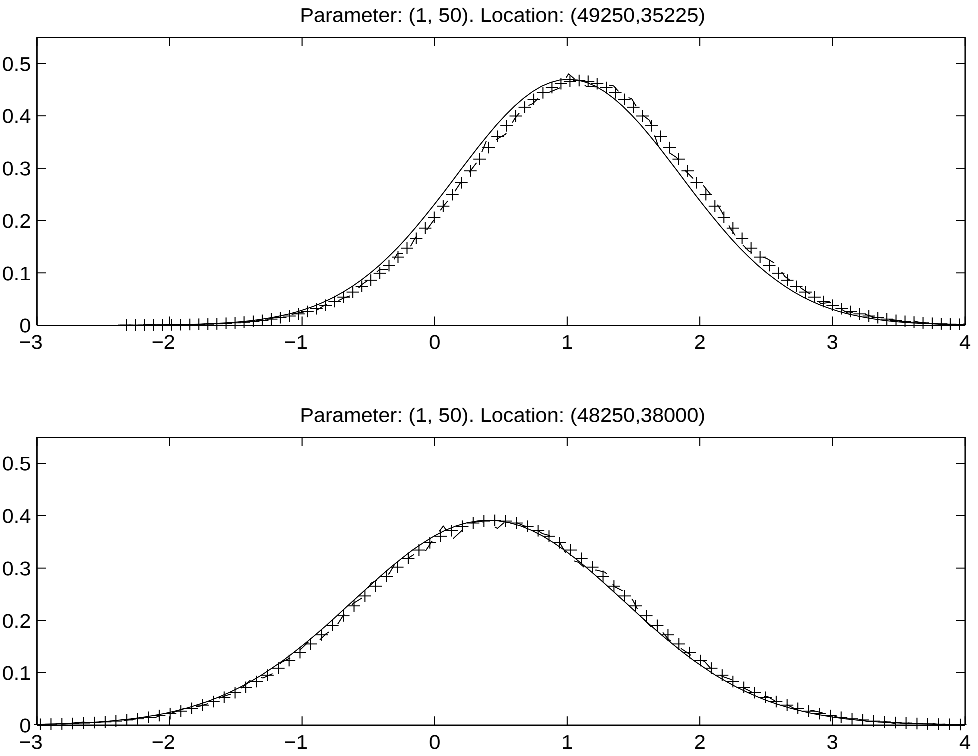
Figure 5: Lancashire county with 238 measurement locations of campylobacter infection.

$i = 1, \ldots, k$. Each observation is tied to a postal code at a spatial registration site. The infection data are modeled as a spatial GLMM with a binomial distribution in equation (2). For the spatial latent variable a grid with interval spacing $\delta = 30$m is constructed. This grid covers the region from $(31970, 47680)$ to $(38660, 51700)$ in the (North, East) coordinates displayed in Figure 5. Hence, the gridsize is $n_1 = 135$ (North) and $n_2 = 224$ (East), in total $n \approx 30000$. Following Rue, Steinsland and Erland (2004) and Steinsland (2006) an exponential covariance function is used, see equation (1). We fix the hyperparameter $\boldsymbol{\theta} = (\sigma, \nu) = (1, 50)$ in this example. This corresponds to quite likely parameter values (Steinsland 2006). Again, ten Newton-Raphson iterations are used to locate the mode of the Gaussian approximation $\hat{\pi}(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})$

In Figure 6 we show plots of the marginal density for x_j at two different spatial locations j. The two locations are $(49250, 35225)$ and $(48250, 38000)$, representing near and far away from registration sites, respectively. Figure 6 shows three approximations to the marginal density $\pi(x_j|\boldsymbol{y}, \boldsymbol{\theta})$ for each of the two locations. The three approximations are as follows: Direct Gaussian approximation $\hat{\pi}(x_j|\boldsymbol{y}, \boldsymbol{\theta})$ (solid), improved Gaussian approximation as presented in Section 5 and denoted $\tilde{\pi}(x_j|\boldsymbol{y}, \boldsymbol{\theta})$ (crossed), and an independent proposal MH approximation obtained by 100000 iterations (dashed) using the joint direct approximation as proposal distribution. In Figure 6 (top), which displays results of a location only one node (30m) from registration sites, we see that the direct Gaussian approximation (solid) is slightly biased to the left, while the improved approximation (crossed) and the MH approximation (dashed) are very similar. In this case the improved approximation does have an effect on the approximate marginal, possibly since we are near data nodes and there is much non-Gaussian influence. Hence, the joint Gaussian at the mode does not capture all features of the marginal density. In Figure 6 (below), which displays results of a location about 800m from the nearest registration site, the three plots are almost identical. This indicates that the direct approximation is accurate when there is less non-Gaussian influence.

In this example the direct Gaussian approximation is again 'practically sufficient', meaning that for most purposes the slight differences between the solid curve and the others in Figure 6 would not have any effect. The improved approximation lies on top of the Monte Carlo solution and is 'practically exact', meaning that the MH sampler cannot detect any possible differences between the improved approximation and the exact solution which is obtained by MH sampling in the limit.

As pointed out by other authors, see e.g. Diggle et al. (1998) and Steinsland (2006), the data



Parameter: (1, 50). Location: (49250,35225)

Parameter: (1, 50). Location: (48250,38000)

Figure 6: Lancashire dataset. Conditional density $\pi(x_j|\boldsymbol{y},\boldsymbol{\theta})$ obtained by approximate inference at two spatial locations and for parameter $\boldsymbol{\theta}$ fixed at ($\sigma = 1$, $\nu = 50$m). Solid is direct Gaussian approximation, crossed is improved Gaussian approximation, and dashed is approximation obtained by MH sampling.

does not carry much information about the model parameters in this example. We have hence chosen to treat the hyperparameters as fixed. An interesting line of research is spatial placement of registeration sites for reliable prediction and parameter estimation (Diggle and Lophaven 2006).

7. CONCLUSIONS

In this paper we present fast approximations of marginal posteriors for latent variables and model parameters for a very common geostatistical model. For this spatial model it has been hard to design MCMC algorithms that mix adequately. Based on the result of the current paper, we recommend the approximate solutions which are accurate and fast to compute. They can easily be part of standard softwares. Fast approximate inference can help to expand the scope of geostatistical modeling. Possible applications of this method include geostatistical design (Diggle and Lophaven 2006), model choice (Clyde and George 2004) and model assessment (Johnson 2004) in a geostatistical setting. We have presented the paper in the light of the exponential family likelihood model. Other distributions are also possible. The Gaussian approximation for the latent posterior become worse with extreme likelihoods such as measuring only the absolute value of the latent variable. Moreover, the measurements do not have to be at a single node, but can be aggregated at several nodes.

We show results of two methods of approximate inference in the paper. The first method is ’practically sufficient’ in the examples we studied, meaning that for purposes regarding decisions or model assessment the approximation is accurate enough. The improved version, which provides a correction to the first approximation, is ’practically exact’ in the examples we studied, meaning that we only confirmed the approximation when using tedious Markov chain Monte Carlo methods. In our opinion one would have to run Markov chains for much longer than is typically done to verify any possible bias of the improved version. However, further research is needed to assess the quality of each approximation. One might consider computing better, higher order improved approximations and checking if corrections are still relevant.

We briefly discuss the computational costs and limitations of the direct approximation. Newton-Raphson optimization to locate the posterior mode requires a matrix inversion of order $O(k^3)$ at each iteration step, finding all conditional variances requires $O(nk^2)$, while the fast Fourier transform requires the order of $O(n \log n)$. For the case that we consider with $n \sim 10000$, $k \sim 100$, the variance computation is the most computer intensive part. The limitation of our approach is the value of k. If k becomes large, say $k > 500$, the calculations become intractable. The improved approximation is typically needed only at $O(k)$ cells, i.e. grid nodes near registration sites. The CPU time of the approximations is in seconds, in constrast to standard Markov chain Monte Carlo algorithms which typically run overnight.

The special case considered in this paper includes a stationary prior model for the latent variable. This is a very common assumption in geostatistics. Yet, this assumption is easily violated in two ways; adding a trend in the prior mean or using a covariance matrix that is not block circulant. One must then include the trend parameters as part of $\boldsymbol{\theta}$ and do parametric inference on a larger sample space. Alternatively one might use a Gaussian Markov random field instead of a Gaussian random field and include trend parameters as part of the latent variable (Rue and Held 2005).

APPENDIX: COMPUTATIONAL ASPECTS

A.1 Stationary prior distribution:

Let $\boldsymbol{x} = (x_1, \ldots, x_n)'$ be a Gaussian variable represented on a regular grid of size $n = n_1 n_2$ with block circulant covariance matrix $\boldsymbol{C}$. The matrix $\boldsymbol{C}$ might be a function of model parameters $\boldsymbol{\theta}$ but this is suppressed here to simplify notation. We refer to the $n_1 \times n_2$ matrix $\boldsymbol{x}^m = (x^m_{0,0}, x^m_{0,1}, \ldots, x^m_{n_1-1,n_2-1})$ as the matrix associate of length n vector $\boldsymbol{x}$. The matrix $\boldsymbol{C}$ is defined by the covariance between $x^m_{0,0}$ and all other variables as they are positioned on a torus. We arrange these n covariance entries in an $n_1 \times n_2$ matrix which we denote by $\boldsymbol{c}^m$. We can collect the n eigenvalues of $\boldsymbol{C}$ in an $n_1 \times n_2$ matrix $\boldsymbol{\lambda}^m = \text{dft2}(\boldsymbol{c}^m)$ (Gray 2006). Here dft2 denotes the two dimensional discrete Fourier transform

$$\text{dft2}(c^m)_{j_1',j_2'} = \sum_{j_1=0}^{n_1-1} \sum_{j_2=0}^{n_2-1} c^m_{j_1',j_2'} \exp[-2\pi\iota(\frac{j_1 j_1'}{n_1} + \frac{j_2 j_2'}{n_2})], \quad j_1' = 1, \ldots, n_1, j_2' = 1, \ldots, n_2, \tag{25}$$

with $\iota = \sqrt{-1}$. The determinant of $\boldsymbol{C}$ is the product of all entries in $\boldsymbol{\lambda}^m$. We denote by idft2$(\boldsymbol{d}^m)$ the two dimensional inverse discrete Fourier transform of $n_1 \times n_2$ matrix $\boldsymbol{d}^m$.

Consider first matrix $\boldsymbol{C}$ multiplied with length n vector $\boldsymbol{v}$. The $n_1 \times n_2$ matrix associate of vector $\boldsymbol{w} = \boldsymbol{C}\boldsymbol{v}$ can be evaluated by

$$\boldsymbol{w}^m = \text{Re}\{\text{dft2}[\text{dft2}(\boldsymbol{c}^m) \odot \text{idft2}(\boldsymbol{v}^m)]\}, \tag{26}$$

where $\odot$ represents elementwise multiplication. Further, $\boldsymbol{w} = \boldsymbol{C}^a \boldsymbol{v}$ is given by

$$\boldsymbol{w}^m = \text{Re}\{\text{dft2}[\text{dft2}(\boldsymbol{c}^m)^{\owedge a} \odot \text{idft2}(\boldsymbol{v}^m)]\}, \tag{27}$$

where $(\boldsymbol{c}^m)^{\owedge a}$ means taking every element of $\boldsymbol{c}^m$ to the power of a. This last relation is useful for evaluation and sampling (Rue and Held 2005). For *evaluation* of the quadratic form we use that $\boldsymbol{v}'\boldsymbol{C}^{-1}\boldsymbol{v} = \boldsymbol{v}'\boldsymbol{w}$, where $\boldsymbol{w}^m$ is evaluated in equation (27) with $a = -1$. For *sampling* we let $\boldsymbol{v}^m$ denote a $n_1 \times n_2$ matrix of independent variables with mean zero and standard deviation 1, and with vector associate $\boldsymbol{v}$. A variable $\boldsymbol{w} \sim N(\boldsymbol{w}; 0, \boldsymbol{C})$ can be obtained via its matrix associate using (27) with $a = 1/2$ (Chan and Wood 1997).

A.2 Conjugate Gaussian posterior distribution:

Suppose we have prior distribution $\pi(\boldsymbol{x}) = N(\boldsymbol{x}; \boldsymbol{\beta}, \boldsymbol{C})$ and likelihood $\pi(\boldsymbol{z}|\boldsymbol{x}) = N(\boldsymbol{z}; \boldsymbol{A}\boldsymbol{x}, \boldsymbol{P})$, $\boldsymbol{z} = (z_1, \ldots, z_k)'$, where $\boldsymbol{A}$ denotes the sparse $k \times n$ matrix of 0s and 1s in equation (3) and assume that $n \gg k$. The posterior is $\pi(\boldsymbol{x}|\boldsymbol{z}) = N(\boldsymbol{x}; \boldsymbol{\mu}_{\boldsymbol{x}|\boldsymbol{z}}; \boldsymbol{V}_{\boldsymbol{x}|\boldsymbol{z}})$, where the conditional mean and covariance are

$$\boldsymbol{\mu}_{\boldsymbol{x}|\boldsymbol{z}} = \boldsymbol{\beta} + \boldsymbol{C}\boldsymbol{A}'\boldsymbol{R}^{-1}(\boldsymbol{z} - \boldsymbol{A}\boldsymbol{\beta}), \quad \boldsymbol{V}_{\boldsymbol{x}|\boldsymbol{z}} = \boldsymbol{C} - \boldsymbol{C}\boldsymbol{A}'\boldsymbol{R}^{-1}\boldsymbol{A}\boldsymbol{C}, \quad \boldsymbol{R} = \boldsymbol{A}\boldsymbol{C}\boldsymbol{A}' + \boldsymbol{P}. \tag{28}$$

For *evaluation* of this posterior we use that

$$\pi(\boldsymbol{x}|\boldsymbol{z}) = \frac{\pi(\boldsymbol{z}|\boldsymbol{x})\pi(\boldsymbol{x})}{\pi(\boldsymbol{z})}, \quad \pi(\boldsymbol{z}) = N(\boldsymbol{z}; \boldsymbol{A}\boldsymbol{\beta}, \boldsymbol{R}). \tag{29}$$

The prior is evaluated using the relationship following equation (27), while the other expressions only involve $k \times k$ matrices and with small k these are fast to evaluate. We can *sample* from the posterior as follows: i) Draw a sample from the unconditional density, $\boldsymbol{v} \sim N(\boldsymbol{v}; \boldsymbol{\beta}, \boldsymbol{C})$ using the relationship following equation (27). ii) Draw a sample $\boldsymbol{w} \sim N(\boldsymbol{w}; \boldsymbol{z}, \boldsymbol{P})$. iii) Set

$$\boldsymbol{x} = \boldsymbol{v} + \boldsymbol{C}\boldsymbol{A}'\boldsymbol{R}^{-1}(\boldsymbol{w} - \boldsymbol{A}\boldsymbol{v}) = \boldsymbol{v} + \boldsymbol{u}, \quad \boldsymbol{u}^m = \text{Re}\{\text{dft2}[\text{dft2}(\boldsymbol{c}^m) \odot \text{idft2}(\boldsymbol{t}^m)]\}, \tag{30}$$

where we use equation (26) and $\boldsymbol{t}^m$ is the matrix associate of $\boldsymbol{t} = \boldsymbol{A}'\boldsymbol{R}^{-1}(\boldsymbol{w} - \boldsymbol{A}\boldsymbol{v})$ calculated by

$$t_j = \left\{ \begin{array}{ll} \sum_{i'=1}^{k} R^{-1}_{i,i'}(w_{i'} - v_{s_{i'}}) & \text{if} \quad s_i \in j \\ 0 & \text{else.} \end{array} \right. , \quad i = 1, \ldots, k, \;\; j = 1, \ldots, n, \tag{31}$$

using the properties of $k \times n$ matrix $\boldsymbol{A}$. The matrix inversion in equation (31) is for $k \times k$ matrix $\boldsymbol{R}$ and we assume that k is of moderate size.

A.3 Newton-Raphson optimization:

Consider our linearization of the likelihood part in equation (6). For this non-Gaussian case we find the posterior mode of $\pi(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})$ for fixed $\boldsymbol{\theta}$ by iterative linearization using the Newton-Raphson algorithm. Note first that by setting $\boldsymbol{v} = \boldsymbol{\beta}$ in step i) and $\boldsymbol{w} = \boldsymbol{z}$ in step ii) of the sampling step before equation (30), we obtain the posterior mean in step iii) using equation (30). This is identical to the mean in equation (28). For the approximate Gaussian case we have that $\boldsymbol{z} = \boldsymbol{z}(\boldsymbol{y}, \boldsymbol{x}_s^0)$ as in equation (9), using some initial linearization point $\boldsymbol{x}_s^0$. Let next $\boldsymbol{x}_s^1$ denote the approximate posterior mean in equation (28) obtained by one application of Newton-Raphson in equation (30). This process can then be iterated, getting a new transformed measurement $\boldsymbol{z} = \boldsymbol{z}(\boldsymbol{y}, \boldsymbol{x}_s^1)$ as in equation (9), then a new posterior mean $\boldsymbol{x}^2$, and so on, until one reach the argument at the posterior mode denoted by $\hat{\boldsymbol{m}}_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}}$.

A.4 Evaluation of the Laplace approximation and the acceptance rate:

Let us now bring $\boldsymbol{\theta}$ into our notation and study the computational aspects regarding the Laplace approximation and the acceptance rate of the MH algorithm. We can evaluate the approximate Gaussian posterior using Bayes formula in a similar manner as in equation (29). This gives

$$\hat{\pi}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}) = N(\boldsymbol{x};\hat{\boldsymbol{m}}_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}},\hat{\boldsymbol{V}}_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}}) = \frac{\hat{\pi}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})}{\hat{\pi}(\boldsymbol{z}|\boldsymbol{\theta})}, \quad \boldsymbol{z} = \boldsymbol{z}(\boldsymbol{y},\hat{\boldsymbol{m}}_{\boldsymbol{x}_s|\boldsymbol{y},\boldsymbol{\theta}}), \tag{32}$$

where $\hat{\pi}(\boldsymbol{z}|\boldsymbol{\theta}) = N(\boldsymbol{z};\boldsymbol{A\beta},\boldsymbol{R})$, $\hat{\pi}(\boldsymbol{z}|\boldsymbol{x}) = N(\boldsymbol{z};\boldsymbol{A\beta},\boldsymbol{P})$. The $k \times k$ matrices $\boldsymbol{R}$ and $\boldsymbol{P}$ are now evaluated at the argument at the posterior mode $\hat{\boldsymbol{m}}_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}}$ from the last Newton-Raphson step. For the Laplace approximation in equation (11) this means that

$$\hat{\pi}(\boldsymbol{\theta}|\boldsymbol{y}) \propto \frac{\pi(\boldsymbol{y}|\boldsymbol{x})\pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\hat{\pi}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})} = \frac{\pi(\boldsymbol{y}|\boldsymbol{x})\pi(\boldsymbol{\theta})\hat{\pi}(\boldsymbol{z}|\boldsymbol{\theta})}{\hat{\pi}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta})}, \quad \boldsymbol{z} = \boldsymbol{z}(\boldsymbol{y},\hat{\boldsymbol{m}}_{\boldsymbol{x}_s|\boldsymbol{y},\boldsymbol{\theta}}). \tag{33}$$

The expression only involves $k \times k$ matrices and with small k these are fast to evaluate. For the acceptance rate in equation (18), treating $\boldsymbol{\theta}$ as fixed, this means that

$$\min\left\{1, \frac{[\prod_{i=1}^{k}\pi(y_i|x'_{s_i})]\pi(\boldsymbol{x}'|\boldsymbol{\theta})}{[\prod_{i=1}^{k}\pi(y_i|x_{s_i})]\pi(\boldsymbol{x}|\boldsymbol{\theta})}\frac{\hat{\pi}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})}{\hat{\pi}(\boldsymbol{x}'|\boldsymbol{y},\boldsymbol{\theta})}\right\} = \min\left\{1, \frac{[\prod_{i=1}^{k}\pi(y_i|x'_{s_i})]\hat{\pi}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta})}{[\prod_{i=1}^{k}\pi(y_i|x_{s_i})]\hat{\pi}(\boldsymbol{z}|\boldsymbol{x}',\boldsymbol{\theta})}\right\}, \tag{34}$$

which again only involves $k \times k$ matrices and the expression is fast to evaluate.

REFERENCES

Ainsworth, L. M., and Dean, C. B., (2006), “Approximate inference for disease mapping”, *Computational Statistics & Data Analysis*, 50, 2552-2570.

Banerjee, S., Carlin, B. P., and Gelfand, A. E., (2004), “Hierarchical modeling and analysis for spatial data”, Chapman & Hall.

Breslow, N. E., and Clayton, D. G., (1993), “Approximate inference in generalized linear mixed models”, *Journal of the American Statistical Association*, 88, 9-25.

Carlin, B. P., and Louis, T. A., (2000), “Bayes and empirical Bayes methods for data analysis”: Chapman and Hall.

Chan, G., and Wood, A. T. A., (1997), “An algorithm for simulating stationary Gaussian random fields”, *Applied Statistics*, 46, 171-181.

Christensen, O. F., Roberts, G. O., and Sköld, M., (2006), “Robust MCMC methods for spatial GLMM’s”, *Journal of Graphical and Computational Statistics*, 15, 1-17.

Clyde, M., and George, E. I., (2004), “Model uncertainty”, *Statistical Science*, 19, 81-94.

Cressie, N. A. C., (1991), “Statistics for spatial data”, Wiley.

Diggle, P. J., Tawn, J. A., and Moyeed, R. A., (1998), “Model-based geostatistics”, *Journal of the Royal Statistical Society, Ser. C*, 47, 299-350.

Diggle, P. J., Ribeiro Jr., P. J., and Christensen, O. F., (2003), “An introduction to model-based geostatistics”, In J. Møller (Ed.) Spatial Statistics and Computational Methods, Lecture notes in Statistics; **173**, 43-86, Berlin: Springer-Verlag.

Diggle, P. J., and Lophaven, S., (2006), “Bayesian Geostatistical Design”, *Scandinavian Journal of Statistics*, 33, 53-64.

Gel, Y., Raftery, A. E., and Gneiting, T., (2004), “Calibrated probabilistic mesoscale weather field forecasting: The geostatistical output perturbation method”, *Journal of the American Statistical Association*, 99, 575-583.

Gray, R. M., (2006), “Toeplitz and circulant matrices: A review”, Free book, available from http://ee.stanford.edu/ ∼gray.

Hsiao, C. K., Huang, S. Y., and Chang, C. W., (2004), “Bayesian marginal inference via candidate’s formula”, *Statistics and Computing*, 14, 59-66.

Johnson, V. E., (2004), “A Bayesian χ^2 test for goodness-of-fit”, *The Annals of Statistics*, 32, 2361-2384.

McCullagh, P., and Nelder, J. A., (1989), “Generalized linear models”, Chapman & Hall.

Polasky, A., and Solow, A. R., (2001), “The value of information in reserve site selection”, *Biodiversity and Conservation*, 10, 1051-1058.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P., (1996), “Numerical Recipes in C: The art of Scientific Computing”, Cambridge University Press.

Robert, C. P., and Casella, G., (2004), “Monte Carlo Statistical Methods”, Springer.

Rue, H., Steinsland, I., and Erland, S., (2004), “Approximating hidden Gaussian Markov random fields”, *Journal of the Royal Statistical Society, Ser. B*, 66, 877-892.

Rue, H., and Held, L., (2005), “Gaussian Markov random fields, Theory and applications”, Chapman & Hall.

Rue, H., and Martino, S., (2006a), “Approximate Bayesian inference for hierarchical Gaussian Markov random fields”, *Journal of Statistical Planning and Inference*, To appear.

Rue, H., and Martino, S., (2006b), “Approximate Bayesian inference for latent Gaussian models using a nested integrated Laplace-approximation”, In Prep.

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

Stein, M. L., (1999), “Interpolation of spatial data: Some theory for Kriging”, Springer.

Steinsland, I., (2006), “Parallel exact sampling and evaluation of Gaussian Markov random fields”, *Computational Statistics & Data Analysis*, To appear.

Tierney, L., and Kadane, J. B., (1986), “Accurate approximations for posterior moments and marginal densities”, *Journal of the American Statistical Association*, 81, 82-86.