NORGES TEKNISK-NATURVITENSKAPELIGE UNIVERSITET

Approximate inference for spatial GLMs

by

Jo Eidsvik 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 inference for latent spatial GLMs

Jo Eidsvik and Håvard Rue
Department of Mathematical Sciences, NTNU, Norway.

ABSTRACT

In this paper latent spatial generalized linear models are analyzed. A Bayesian model is used with a stationary prior for the latent variable, priors for model parameters, and a likelihood model of exponential family type, conditional on the latent variables. We consider the common geostatistical special case with large latent variable and observations at only a few known registration sites.

The method of inference is based on direct approximations of the marginal densities for latent spatial variable and model parameters using Newton-Raphson optimization and the Laplace approximation. For the geostatistical special case that we consider, optimization and evaluation can be done very efficiently in the Fourier domain. The approximations are also useful as proposal distributions in Monte Carlo simulations.

We study two examples; radionuclide concentrations on Rongelap Island and disease infections in Lancashire. For these examples the direct approach to inference is compared with time consuming Monte Carlo methods. The results are similar and we hence recommend the faster direct approximation for such spatial generalized linear models.

KEYWORDS: *Approximate inference, spatial GLM, circulant covariance matrix, Newton-Raphson, Metropolis–Hastings, 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 model. Among the popular applications are geophysics, mining, meteorology and disease mapping, see e.g. Cressie (1991), Diggle et al. (1998) and Banerjee et al. (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 model (GLM), see e.g. Breslow and Clayton (1993), Diggle et al. (1998) and Christensen et al. (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 in two dimensional space. Suppose that $\boldsymbol{x}$ has a stationary prior distribution specified by some 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. This hierarchical model is sometimes referred to as a spatial generalized linear mixed model (Breslow and Clayton, 1993) to clarify the random latent effects, but we have chosen to simply use spatial GLM here. Typical examples of this model include binomial proportions or Poisson counts registrered at a some known locations in space, with the objective being prediction of the underlying intensity or risk (log odds) surface across the spatial domain of interest, and inference for 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 with registrated data are available, i.e. $n \gg k$. For example, this situation often occurs in spatial registration for weather forecasting (Gel et al., 2004) and in reserve site selection for predicting the probability of presence for 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 GLMs 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 parameters $\boldsymbol{\theta}$ and latent spatial variables $\boldsymbol{x}$ using Markov chain Monte Carlo (MCMC) algorithms. Since MCMC algorithms have grown mature over the last few decades, see e.g. Liu (2001), 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 GLMs (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 techniques suitable for special cases are needed, possibly avoiding the problems with sampling methods.

The main contribution of this paper is a fast direct approximate Bayesian solution to the common special case of spatial GLMs with $n \gg k$. In particular, this paper provides a recipe for efficient calculation of 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 density of the latent variable at any spatial location. Via examples we show that the fast approximate method gives similar results as time consuming Monte Carlo methods. This 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 computations at almost no computational cost. Moreover, the direct method could quite

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), in a geostatistical setting. The outlined method of direct Bayesian inference and spatial prediction is based on Newton-Raphson optimization and the Laplace approximation (Tierney and Kadane, 1986). The dimension of n is typically very large, but nevertheless the approach becomes computationally attractive because tasks can be solved efficiently in the Fourier domain.

Other recent approaches to approximate inference for spatial GLMs 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 (2006) 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 with $n = k$ and n was of moderate size.

The outline is as follows: In Section 2 we define the special case of spatial GLMs considered in this paper. The proposed methods of approximate inference and prediction are described in Section 3. Two numerical examples are provided in Section 4. We discuss and conclude in Section 5. The computational aspects are postponed to the Appendix.

2. Spatial GLM

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 $\boldsymbol{x}$ would denote the latent / underlying risk or log odds surface, while it would denote the latent 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 model parameters. For example, $\boldsymbol{\theta}$ could include σ; pointwise standard deviation, and ν; spatial correlation range. Block circulant covariance structure means that the $n_1 \times n_2$ grid is wrapped on a torus. As an example of a covariance function we use the exponential defined by

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

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. We use this covariance function in the examples below, but many others are possible, see e.g. Cressie (1991). 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 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. The fast Fourier transform is an efficient tool for evaluating and sampling Gaussian fields with block circulant covariance matrices since the eigenvectors of a circulant matrix are provided by the Fourier basis, see e.g. Chan and Wood (1997), Rue and Held (2005), Gray (2006), and Appendix A.

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

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

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

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

i.e. s_i is the gridnode of measurement i. Assume that $n \gg k$ and that the matrix $\boldsymbol{A}$ consists mostly of 0 values, but it has one 1 value in 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 defined by

	Poisson	Binomial	Gaussian
$b(x)$	$m\exp(x)$	$m\log[1+\exp(x)]$	$x^2/2$
$a(\phi)$	1	1	ϕ^2

, (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 GLMs (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 GLM 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 GLM.

3. Approximate inference and prediction

We will now outline our methods for fast approximate analysis of Bayesian spatial GLMs 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 $\pi(x_j|\boldsymbol{y})$. 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 GLM 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

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

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

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

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

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

where we use

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

Rather than fitting a Gaussian approximation at any $\boldsymbol{x}^0$, the above process is iterated using the Newton-Raphson algorithm for $\boldsymbol{x}$. After a few iterations we have then fitted a Gaussian approximation $\hat{\pi}(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})$ at the posterior mode which denoted by $\boldsymbol{\mu}_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}}$, see Appendix B and C. These Newton-Raphson calculations are similar to the traditional ones used for GLMs (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 of the main contributions 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. 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 Gaussian prior has much influence. If we have small Poisson counts or few Binomial trials, the likelihood is far from Gaussian and the approximation might be worse.

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.
i) 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 $n \gg k$.
ii) Equation (8) can be computed efficiently in the Fourier domain using that $\boldsymbol{C}$ is block circulant.
iii) The Gaussian approximation $\hat{\pi}(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})$ can be evaluated using Bayes theorem and benefits of the Fourier domain.
iv) Newton-Raphson optimization involves iterative calculation of equation (8) and is hence also available in the Fourier domain.

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

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

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

Equation (10) is valid for any value of the spatial variable $\boldsymbol{x}$, for example $\boldsymbol{x} = \boldsymbol{\mu}_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}}$, the posterior mode for a fixed $\boldsymbol{\theta}$. The challenging part of equation (10) is the denominator which is unknown. We choose to approximate this denominator using the fitted Gaussian density from the previous Section. The approximated density for the model parameters then becomes

$$\hat{\pi}(\boldsymbol{\theta}|\boldsymbol{y}) \propto \frac{\pi(\boldsymbol{y}|\boldsymbol{x})\pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\hat{\pi}(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})}. \tag{11}$$

Equation (11) is a Laplace approximation, see e.g. Tierney and Kadane (1986) and Carlin and Louis (2000), since we approximate $\pi(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})$ by a Gaussian density at the posterior mode, and next use this to find the marginal of the $\boldsymbol{\theta}$ parameters. The relative error of this Laplace approximation is $O(k^{-3/2})$ (Tierney and Kadane, 1986). Relative error could be advantageous, especially in the tails of the distribution. Equation (11) can be evaluated efficiently using Bayes rule and properties of the Gaussian approximation, see Appendix D.

We evaluate the approximation in 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, see e.g. Press et al. (1996), but 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. Spatial prediction using $\pi(\boldsymbol{x}_j|\boldsymbol{y})$

For spatial prediction it is common to use marginals after integrating out the parameters (Diggle et al., 1998). Using that $\pi(\boldsymbol{x}_j|\boldsymbol{y}) = \int \pi(\boldsymbol{x}_j|\boldsymbol{y},\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}$, $j = 1, \ldots, n$, we can find the marginal distribution for the spatial variables. Each term in the integral is here approximated using the Gaussian approximation $\hat{\pi}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})$ and $\hat{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ in equation (11), respectively.

The approximate marginal means become

$$\mu_{x_j|\boldsymbol{y}} \approx \sum_l \mu_{x_j|\boldsymbol{y},\boldsymbol{\theta}_l} \hat{\pi}(\boldsymbol{\theta}_l|\boldsymbol{y}), \quad j = 1, \ldots, n. \tag{14}$$

The conditional means for fixed parameters $\boldsymbol{\theta}$, obtained on the last Newton-Raphson optimization step and denoted $\boldsymbol{\mu}_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}}$, are here weighted to obtain the marginal mean.

The approximate marginal variances are

$$V_{x_j|\boldsymbol{y}} \approx \sum_l V_{x_j|\boldsymbol{y},\boldsymbol{\theta}_l} \hat{\pi}(\boldsymbol{\theta}_l|\boldsymbol{y}) + \sum_l (\mu_{x_j|\boldsymbol{y},\boldsymbol{\theta}_l} - \mu_{x_j|\boldsymbol{y}})^2 \hat{\pi}(\boldsymbol{\theta}_l|\boldsymbol{y}). \tag{15}$$

In this equation $V_{x_j|\boldsymbol{y},\boldsymbol{\theta}} = V_{x_j|\boldsymbol{y},\boldsymbol{\theta}}(\boldsymbol{\mu}_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}})$ denotes the condtional variance in equation (7) evaluated at the posterior mode. For these terms we need to calculate the diagonal entries of $\boldsymbol{C}\boldsymbol{A}'\boldsymbol{R}^{-1}\boldsymbol{A}\boldsymbol{C}$

in equation (7) given by

$$(\boldsymbol{C}\boldsymbol{A}'\boldsymbol{R}^{-1}\boldsymbol{A}\boldsymbol{C})_{jj} = \sum_{i=1}^{k}\sum_{i'=1}^{k} C_{j,s_i} R^{-1}_{ii'} C_{s_{i'},j}, \quad j = 1,\ldots,n. \tag{16}$$

3.4. Approximations as proposal distributions in Monte Carlo sampling

Consider an independent proposal MH algorithm with proposals $\boldsymbol{x}'$ from the Gaussian approximation $\hat{\pi}(\boldsymbol{x}'|\boldsymbol{y},\boldsymbol{\theta})$, keeping $\boldsymbol{\theta}$ fixed. The proposal $\boldsymbol{x}'$ can be generated efficiently in the Fourier domain, see Appendix B. We denote the current state of the Markov chain by $\boldsymbol{x}$. The acceptance rate is

$$\min\left\{1, \frac{[\prod_{i=1}^{k}\pi(y_i|x'_{s_i})]\pi(\boldsymbol{x}'|\boldsymbol{\theta})}{[\prod_{i=1}^{k}\pi(y_i|x_{s_i})]\pi(\boldsymbol{x}|\boldsymbol{\theta})}\frac{\hat{\pi}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})}{\hat{\pi}(\boldsymbol{x}'|\boldsymbol{y},\boldsymbol{\theta})}\right\}, \tag{17}$$

otherwise we keep the current value $\boldsymbol{x}$. The acceptance probabilty in equation (17) can be evaluated using Bayes rule and propoerties of the Gaussian approximation, see Appendix D. This MH scheme might be effective since large changes are proposed at every iteration, see e.g. Liu (2001). On the other hand one might run into problems as the chain can get stuck in a state $\boldsymbol{x}$ with large posterior probability and relatively small approximate probability. The acceptance rate is then close to zero for most proposals. This problem can be avoided by using the independent proposal in conjunction with other MH samplers (Tierney and Mira, 1999). 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. This is a one-block MH sampler, accepting (or rejecting) variables jointly (Liu et al., 1994).

One can alternatively use importance sampling based on the approximations presented above. This is done by generating $\boldsymbol{x}^b$, $b = 1,\ldots,B$, independently from $\hat{\pi}(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})$, keeping $\boldsymbol{\theta}$ fixed. Any expectation $E[g(\boldsymbol{x})|\boldsymbol{y},\boldsymbol{\theta}]$ can then be estimated by a weighted sum (Shepard and Pitt, 1997). The quality of the approximation can be monitored using effective sample size (ESS), see e.g. Liu (2001). 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 $O(B^{-1/2})$. 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. Examples

In this section we analyze the two examples in Diggle et al. (1998). They are both special cases with data made at only a few registration sites and with a large latent spatial variable modeled using a stationary prior distribution.

4.1. Poisson distribution for concentrations on Rongelap Island

This 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 of the island in

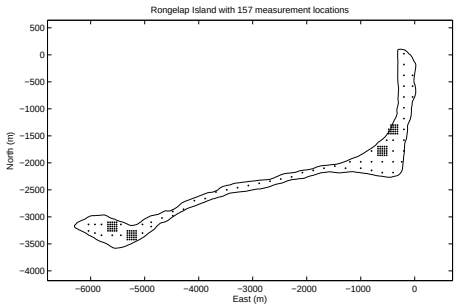

Figure 1: Rongelap island with the 157 registration sites for observations of radionuclide concentrations.

Figure 1. The data are modeled as a spatial GLM with a Poisson distribution in equation (2). The goal of the study is to assess the latent intensity surface and the model parameters. For the spatial intensity 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 σ and ν, and as hyperparameters we use $\beta_0 = 1.5$ and $\tau^2 = 1$. 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.

The grid of $\boldsymbol{\theta} = (\sigma, \nu)$ values defined in equation (12) covers $\sigma \in \{0.2, 1\}$ and $\nu \in \{50, 350\}$ 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, solid curve). Figure 2 (left) seems similar to results obtained earlier by MCMC sampling, see e.g.

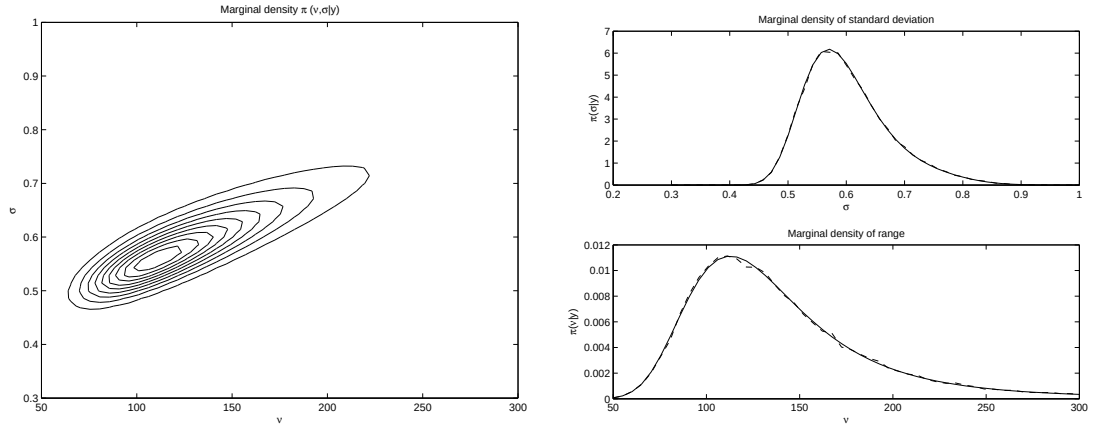

Figure 2: Rongelap dataset. Left) Direct approximation of joint density of parameters $\boldsymbol{\theta} = (\sigma, \nu)$. Right) Direct approximation of marginal densities for standard deviation σ and spatial correlation range ν (solid). MH approximation of marginal densities (dashed).

Christensen et al. (2006). This is also visible from Figure 2 (right, dashed curve) which displays an estimate of the marginals using MH sampling with joint updating of $\boldsymbol{x}$ and $\boldsymbol{\theta}$, as described

in Section 3.4. Since the solid and dashed curves in Figure 2 (right) are hard to distinguish, the direct approximation appears to be very good. The quality of the direct approximation for these parameter marginals is tied to the accuracy of the Gaussian approximation of the latent variable, which in turn gives an approximate relationship in equation (11). We will get back to this topic below.

In Figure 3 we show the marginal predictions in equation (14) and standard deviations given by equation (15). We recognize the measurement locations in the standard deviations and see that the

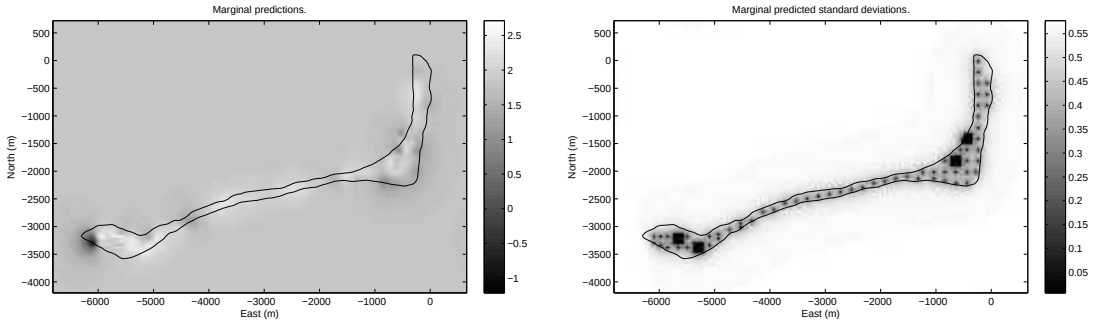

Figure 3: Rongelap dataset. Left) Predicted spatial variable. Right) Marginal standard deviation of spatial variable.

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 maximum radiation level is predicted to be at location $(-3140, -5560)$. The main trends are similar to the ones obtained by MCMC sampling in Diggle et al. (1998).

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 3.4 above. We choose to evaluate all approximations (direct Gaussian approximation, MH method and importance sampling) at spatial locations $(-1800, -600)$, $(-1650, -750)$ and $(-1500, -1500)$. These three locations are chosen since they are at, near and far from registration sites, see Figure 1. We perform the testing for parameter values $\boldsymbol{\theta} = (0.6, 142)$, 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 three 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 results of direct approximate inference are not too different from the two Monte Carlo approximations. When we use Monte Carlo methods the mode seems shifted about 0.003 to the left at location $(-1800, -600)$ (top), shifted about 0.05 to the right in the middle plot, while there is no visible shift at the location far from the Island (below) where the non-Gaussian influence is negligible. The small fluctuations in Figure 4 (solid and dashed) are caused by Monte Carlo error.

The details of the Monte Carlo algorithms were as follows: The independent proposal MH algorithms were implemented for 100000 iterations. The acceptance probability of the MH algorithms were about 0.7, which is large, indicating that the approximation is quite good. The largest number of iterations with no accepted proposals was about 100 for the case with joint updating, indicating that mixing is acceptable and that no domains are likely to have been over- or undersampled.

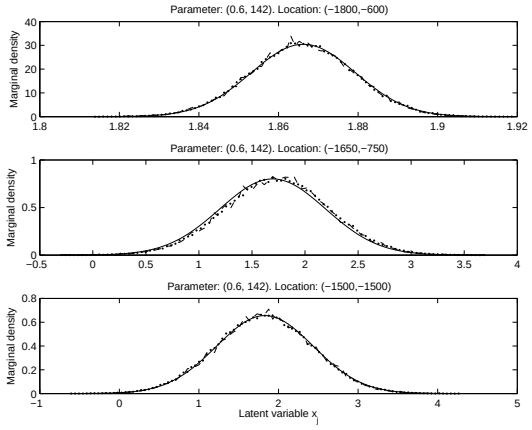

Figure 4: Rongelap dataset. Conditional density $\hat{\pi}(x_j|\boldsymbol{y},\boldsymbol{\theta})$ obtained by approximate inference at three spatial locations and for parameter values fi xed at $(0.6,142)$. Solid is Gaussian approximation, dashed is approximation obtained by MH sampling, and dotted is approximation from importance sampling.

Importance sampling resulted in ESS values about 75000, which is large and indicates that importance sampling works well in this case. 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 $\mu_{x_j|\boldsymbol{y},\boldsymbol{\theta}} \pm 4\sqrt{V_{x_j|\boldsymbol{y},\boldsymbol{\theta}}}$, and thus created the estimated density curve (dashed and dotted). We did not smooth this estimate in order to visualize some of the Monte Carlo error.

4.2. Binomial distribution for records of infections in Lancashire

This dataset consists of $k = 238$ postal code measurements of campylobacter infection in Lancashire district, see Figure 5. The observations y_i = number of campylobacter infection out

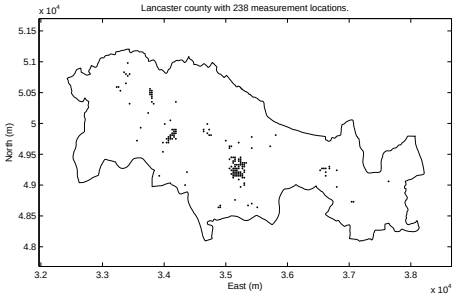

Figure 5: Lancashire county with 238 measurement locations of campylobacter infection.

of enteric infections m_i, $i = 1,\ldots,k$. These data are modeled as a spatial GLM with a binomial distribution in equation (2). The goal is to assess the latent risk surface for these data and the model parameters. 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

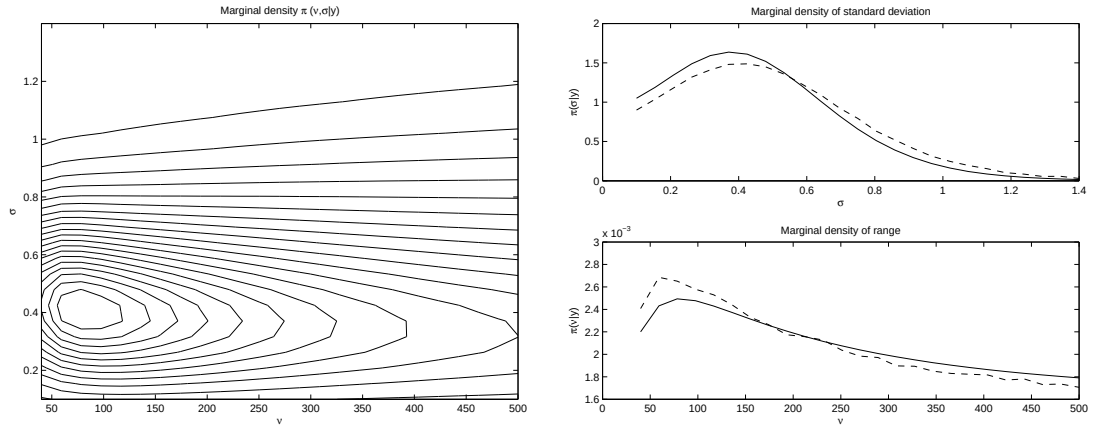

Figure 6: Lancashire dataset. Left) Direct approximation of joint density of parameters $\boldsymbol{\theta} = (\sigma, \nu)$. Right) Direct approximation of marginal densities for standard deviation σ and spatial correlation range ν (solid). MH approximation of marginal densities (dashed).

$n \approx 30000$. Following Rue et al. (2004) and Steinsland (2006) an exponential covariance function is used, see equation (1). We use a flat prior for $\boldsymbol{\theta}$, and as hyperparameters we use $\beta_0 = 0.35$ and $\tau^2 = 1$. 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 (left) we display $\hat{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ obtained by equation (11), along with marginals $\hat{\pi}(\sigma|\boldsymbol{y})$ and $\hat{\pi}(\nu|\boldsymbol{y})$ in Figure 6 (right, solid). We see that the standard deviation is about 0.4 for this dataset and that the density for the range parameter is very skewed with a mode near 80m. The grid of $\boldsymbol{\theta} = (\sigma, \nu)$ values defined in equation (12) covers $\sigma \in \{0.1, 1.4\}$ and $\nu \in \{40, 500\}$ with gridsize $L_1 = 25$ and $L_2 = 25$. We also display the results based on independent proposal MH sampling with joint updates for $\boldsymbol{\theta}$ and $\boldsymbol{x}$ in Figure 6 (right, dashed). The Laplace approximation and MH sampler give quite similar results, but for the standard deviation the MH results are slightly shifted to the right. In the MH estimate for the range parameter we clearly notice the Monte Carlo error.

Figure 7 (left) shows the marginal predictions in equation (14), while Figure 7 (right) displays the marginal standard deviations in equation (15). The spatial field fluctuates quite rapidly, also

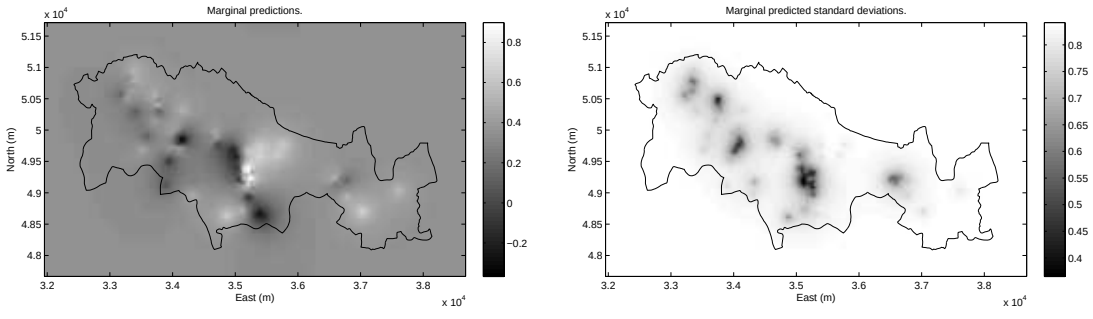

Figure 7: Lancashire dataset. Left) Predicted spatial variable. Right) Marginal standard deviation of spatial variable.

indicated by the small correlation range. The largest predicted value is 0.9 at $(49390, 35240)$, corresponding to a probability of about 0.7 for campylobacter infection.

In Figure 8 (solid) we show plots of the Gaussian approximation $\hat{\pi}(x_j|\boldsymbol{y}, \boldsymbol{\theta})$ for a fixed value of $\boldsymbol{\theta} = (0.26, 193)$ and for three different spatial locations j. The three locations are $(49500, 35000)$,

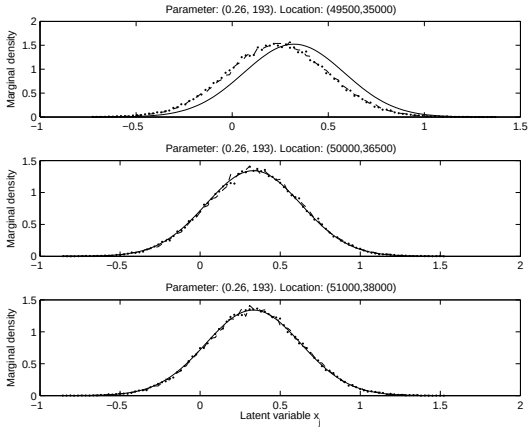

Figure 8: Lancashire dataset. Conditional density $\hat{\pi}(x_j|\boldsymbol{y}, \boldsymbol{\theta})$ obtained by approximate inference at three spatial locations and for parameter values fi xed at $(0.26,193)$. Solid is Gaussian approximation, dashed is approximation obtained by MH sampling, and dotted is approximation from importance sampling.

$(50000, 36500)$ and $(51000, 38000)$, representing near, mid and far away from registration sites. Also displayed in Figure 8 are the approximation based on a long MH simulation (dashed) and the results obtained by importance sampling (dotted). We note that the Gaussian approximation is quite close to the Monte Carlo approximations, but we see a bias to the right for the nearest location (top), a very small bias to the left for the mid-distance location (middle), and hardly no bias for the location far from observations (below). The latter two are possibly the most relevant cases since predictions are typically carried out at locations where no data is available. Fluctuations caused by Monte Carlo error are visible in Figure 8 (dashed and dotted).

The Binomial counts are small for this example (ranging from 0 to a maximum of 6), still the Gaussian approximation and the resulting Laplace approximation in Figure 6 are quite accurate, but not as accurate in the previous example with large Poisson counts. We also notice that the marginal spatial predictive densities in Figure 8 are good as we move away from the registration sites, but not that good when we are at the registration sites where there is much non-Gaussian influence. As pointed out by other authors, see e.g. Diggle et al. (1998) and Steinsland (2006), the data does not carry much information about the range parameter in this example, and an interesting line of research is spatial placement of registeration sites for both prediction and reliable parameter estimation (Diggle and Lophaven, 2006).

5. Conclusions

In this paper we have proposed a new approach to Bayesian inference and prediction for spatial GLMs. Our new approach is based on direct approximation involving Newton-Raphson optimiza-

tion and the Laplace approximation. The main contribution of this paper is that marginal posteriors for latent variables and model parameters can be calculated efficiently for large spatial models based on data at only a few known registration sites. In two examples we compare the results of direct approximate inference and approximations obtained by Monte Carlo sampling. Results are very similar, but the computational time of approximate inference is negligible compared to that of Monte Carlo algorithms. Moreover, two Monte Carlo runs would give different results depending on the random seed, while direct approximate inference returns one non-random answer.

The following arguments are relevant for the methods presented in this paper:
i) The geostatistical case with few data and large spatial variable with stationary prior is very common in practice, but no closed form solutions exist.
ii) For this spatial model it has been hard to design MCMC algorithms that mix adequately.
iii) The direct approximate solutions are fast and easy to code and can easily be part of standard softwares.
iv) Fast approximate inference can help to expand the scope of geostatistical modeling. Possible applications of this method include geostatistical design (Diggle and Lophaven, 2006), and model choice (Clyde and George, 2004), in a geostatistical setting.

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 the 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 $n \gg k$, 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 calculation become intractable. At the same time the Gaussian approximation of the full conditional for latent variable becomes less accurate because there is more non-Gaussian influence in the model. Based on simple testing the computational limit is reached before this Gaussian approximation becomes too poor.

The special case considered in this paper includes a stationary prior model for the latent variable. This is a very common assumption in geostatistical modeling. 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 possible remedy to this problem is to include the trend parameters in the parameter vector and do parametric inference on a larger sample space. Another solution is to use a Gaussian Markov random field instead of a Gaussian random field, and then integrate the parameters as part of the latent variable (Rue and Held, 2005).

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 Skold, 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, Series 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.

Liu, J. S., Wong, W. H., and Kong, A., 1994, Covariance structure of the Gibbs sampler with applications to the comparison of estimators and augmentation schemes, Biometrika, **81**, 27-40.

Liu, J. S., 2001, Monte Carlo Strategies in Scientific Computing, Springer.

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.

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

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

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

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

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.

Tierney, L., and Mira, A., 1999, Some adaptive Monte Carlo methods for Bayesian inference, Statistics in Medicine, **18**, 2507-2515.

Appendix: Computational aspects.

A: 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;

$$\text{dft2}(c^m)_{j_1',j_2'} = \sum_{j_1=0}^{n_1-1} \sum_{j_2=0}^{n_2-1} c^m_{j_1',j_2'} \exp[-2\pi\iota(\frac{j_1 j_1'}{n_1} + \frac{j_2 j_2'}{n_2})], \quad j_1' = 1, \ldots, n_1, j_2' = 1, \ldots, n_2, \quad (18)$$

with $\iota = \sqrt{-1}$. The determinant of $\boldsymbol{C}$ is the product of all entries in $\boldsymbol{\lambda}^m$. We denote by $\text{idft2}(\boldsymbol{d}^m)$ 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

$$\boldsymbol{w}^m = \text{Re}\{\text{dft2}[\text{dft2}(\boldsymbol{c}^m) \odot \text{idft2}(\boldsymbol{v}^m)]\}, \quad (19)$$

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

$$\boldsymbol{w}^m = \text{Re}\{\text{dft2}[\text{dft2}(\boldsymbol{c}^m)^{\circledcirc a} \odot \text{idft2}(\boldsymbol{v}^m)]\}, \quad (20)$$

where $(\boldsymbol{c}^m)^{\circledcirc a}$ 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 (20) with $a = -1$. For *sampling* we let $\boldsymbol{v}^m$ denote a $n_1 \times n_2$ matrix of complex independent variables with mean zero and standard deviation 1 for both real and imaginary parts, and with vector associate $\boldsymbol{v}$. A variable $\boldsymbol{x} \sim N(\boldsymbol{x}; 0, \boldsymbol{C})$ can be obtained via its matrix associate using (20) with $a = 1/2$ (Chan and Wood, 1997).

B: 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 $\boldsymbol{A}$ is 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

$$\boldsymbol{\mu}_{\boldsymbol{x}|\boldsymbol{z}} = \boldsymbol{\beta} + \boldsymbol{C}\boldsymbol{A}'\boldsymbol{R}^{-1}(\boldsymbol{z} - \boldsymbol{A}\boldsymbol{\beta}), \quad \boldsymbol{V}_{\boldsymbol{x}|\boldsymbol{z}} = \boldsymbol{C} - \boldsymbol{C}\boldsymbol{A}'\boldsymbol{R}^{-1}\boldsymbol{A}\boldsymbol{C}, \quad \boldsymbol{R} = \boldsymbol{A}\boldsymbol{C}\boldsymbol{A}' + \boldsymbol{P}. \quad (21)$$

For *evaluation* of this posterior we use that

$$\pi(\boldsymbol{x}|\boldsymbol{z}) = \frac{\pi(\boldsymbol{z}|\boldsymbol{x})\pi(\boldsymbol{x})}{\pi(\boldsymbol{z})}, \quad \pi(\boldsymbol{z}) = N(\boldsymbol{z}; \boldsymbol{A}\boldsymbol{\beta}, \boldsymbol{R}). \quad (22)$$

The prior is evaluated using the relationship following equation (20), 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 (20). ii) Draw a sample $\boldsymbol{w} \sim N(\boldsymbol{w}; \boldsymbol{z}, \boldsymbol{P})$. iii) Set

$$\boldsymbol{x} = \boldsymbol{v} + \boldsymbol{C}\boldsymbol{A}'\boldsymbol{R}^{-1}(\boldsymbol{w} - \boldsymbol{A}\boldsymbol{v}) = \boldsymbol{v} + \boldsymbol{u}, \quad \boldsymbol{u}^m = \text{Re}\{\text{dft2}[\text{dft2}(\boldsymbol{c}^m) \odot \text{idft2}(\boldsymbol{t}^m)]\}, \tag{23}$$

where we use equation (19) and $\boldsymbol{t}^m$ is the matrix associate of $\boldsymbol{t} = \boldsymbol{A}'\boldsymbol{R}^{-1}(\boldsymbol{w} - \boldsymbol{A}\boldsymbol{v})$ calculated by

$$t_j = \begin{cases} \sum_{i'=1}^{k} R^{-1}_{i,i'}(w_{i'} - v_{s_{i'}}) & \text{if} \quad s_i \in j \\ 0 & \text{else.} \end{cases}, \quad i = 1, \ldots, k, \;\; j = 1, \ldots, n, \tag{24}$$

using the properties of $k \times n$ matrix $\boldsymbol{A}$. The matrix inversion in equation (24) is for $k \times k$ matrix $\boldsymbol{R}$ and we assume that k is of moderate size.

C: Newton-Raphson optimization:

Consider our linearization of the GLM 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 in Appendix B, we obtain the posterior mean in step iii) using equation (23). For the approximate Gaussian case we have that $\boldsymbol{z} = \boldsymbol{z}(\boldsymbol{y}, \boldsymbol{x}^0)$ as in equation (9), for some initial linearization point $\boldsymbol{x}^0$. Let next $\boldsymbol{x}^1$ denote the approximate posterior mean in equation (21) obtained by one application of Newton-Raphson in equation (23). This process can then be iterated, getting a new transformed measurement $\boldsymbol{z} = \boldsymbol{z}(\boldsymbol{y}, \boldsymbol{x}^1)$ as in equation (9), then a new posterior mean $\boldsymbol{x}^2$, and so on, until the posterior mode $\boldsymbol{\mu}_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}}$ is located.

D: 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 (22). This gives

$$\hat{\pi}(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta}) = N(\boldsymbol{x}; \mu_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}}, \boldsymbol{V}_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}}) = \frac{\pi(\boldsymbol{z}|\boldsymbol{x})\pi(\boldsymbol{x}|\boldsymbol{\theta})}{\pi(\boldsymbol{z}|\boldsymbol{\theta})}, \quad \boldsymbol{z} = \boldsymbol{z}(\boldsymbol{y}, \boldsymbol{\mu}_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}}), \tag{25}$$

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

$$\hat{\pi}(\boldsymbol{\theta}|\boldsymbol{y}) \propto \frac{\pi(\boldsymbol{y}|\boldsymbol{x})\pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\hat{\pi}(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})} = \frac{\pi(\boldsymbol{y}|\boldsymbol{x})\pi(\boldsymbol{z}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{z}|\boldsymbol{x})}, \quad \boldsymbol{z} = \boldsymbol{z}(\boldsymbol{y}, \boldsymbol{\mu}_{\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta}}). \tag{26}$$

The expression only involves $k \times k$ matrices and with small k these are fast to evaluate. For the acceptance rate in equation (17) this means that

$$\min\left\{1, \frac{[\prod_{i=1}^{k} \pi(y_i|x'_{s_i})]\pi(\boldsymbol{x}'|\boldsymbol{\theta})}{[\prod_{i=1}^{k} \pi(y_i|x_{s_i})]\pi(\boldsymbol{x}|\boldsymbol{\theta})} \frac{\hat{\pi}(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})}{\hat{\pi}(\boldsymbol{x}'|\boldsymbol{y}, \boldsymbol{\theta})}\right\} = \min\left\{1, \frac{[\prod_{i=1}^{k} \pi(y_i|x'_{s_i})]\pi(\boldsymbol{z}|\boldsymbol{x})}{[\prod_{i=1}^{k} \pi(y_i|x_{s_i})]\pi(\boldsymbol{z}|\boldsymbol{x}')}\right\}, \tag{27}$$

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