Analysis of spatial data using generalized linear mixed models and Langevin-type Markov chain Monte Carlo

Ole F. Christensen* Jesper Møller* Rasmus Waagepetersen*

May 29, 2000

Abstract

Markov chain Monte Carlo methods are useful in connection with inference and prediction for spatial generalized linear mixed models, where the unobserved random effects constitute a spatially correlated Gaussian random field. We point out that so-called Langevin-type updates are useful for Metropolis-Hastings simulation of the posterior distribution of the random effects given the data. Furthermore, we discuss the use of improper priors in Bayesian analysis of spatial generalized linear mixed models with particular emphasis on the so-called Poisson-log normal model. For this and certain other models non-parametric estimation of the covariance function of the Gaussian field is also studied. The methods are applied to various data sets including counts of weed plants on a field.

Keywords: Bayesian statistics; generalized linear mixed model; geostatistics; improper prior; Langevin-Hastings update; Markov chain Monte Carlo; Metropolis-adjusted Langevin algorithm; non-parametric covariance estimate; Poisson-log normal distribution; spatial statistics.

1 Introduction

Conventional geostatistical methods such as kriging and trans-Gaussian kriging (Cressie, 1993; Stein, 1999) solve the problem of estimation and prediction for a random field provided it is Gaussian (possibly after transformation). This assumption clearly fails in many practical applications. For example, if the data are binary or counts, normality cannot be obtained by means of transformation. Then generalized linear mixed models (GLMMs) with spatially correlated random effects become useful.

*Department of Mathematical Sciences, Aalborg University, Fredrik Bajersvej 7E, DK-9220 Aalborg. email: `olefc@math.auc.dk, jm@math.auc.dk, rw@math.auc.dk`

This paper partly follows up and expands the work in Diggle *et al.* (1998) on model-based geostatistics using spatial GLMMs and Markov chain Monte Carlo (MCMC) methods. In particular, we focus on the following four points:

(I) Diggle *et al.* (1998) considered Bayesian inference using ordinary single-site Metropolis-Hastings algorithms. We advocate the use of Langevin-Hastings (or Metropolis-adjusted Langevin) updates which are simultaneous updates based on gradient information. The Langevin-Hastings algorithm was introduced in the statistical community by Besag (1994) and earlier in the physics literature by Rossky *et al.* (1978). The Langevin-Hastings algorithm is further studied in Roberts and Tweedie (1996), Roberts and Rosenthal (1998b), Møller *et al.* (1998), and Christensen *et al.* (2000).

(II) It may sometimes be appealing to use flat improper priors for some model parameters in a Bayesian analysis of a spatial GLMM. We discuss to what extend flat priors can be used while maintaining posterior propriety.

(III) Diggle *et al.* (1998) discuss the empirical variogram of the observed random field. For the Poisson-log normal model (Aitchison and Ho, 1989) and other spatial GLMMs with a logarithmic link function we present alternatively a non-parametric estimate of the covariance function of the unobserved random effects.

(IV) We investigate the Bayesian approach to inference for spatial GLMMs by considering the analysis of weed count data with covariate information and the Rongelap data studied in Diggle *et al.* (1998).

The paper is organized as follows. Section 2 surveys spatial GLMMs and discusses the use of improper priors in Bayesian inference for spatial GLMMs. Section 3 considers hybrid MCMC algorithms with Langevin-Hastings updates. The non-parametric estimate of the random effects covariance function is presented in Section 4. Section 5 contains analyses of count data sets and numerical examples. Finally, Section 6 contains some concluding remarks.

2 Generalized linear mixed models with spatially correlated random effects

Sections 2.1 and 2.2 survey spatial GLMMs and various approaches to inference for spatial GLMMs. The question of posterior propriety when flat improper priors are used in Bayesian inference is discussed in Section 2.3.

2.1 Setup

Generalized linear mixed models (GLMMs) (Breslow and Clayton, 1993; Lee and Nelder, 1996) are extensions of generalized linear models (GLMs) (McCullagh and Nelder, 1989) that allow additional sources of variability due to unobservable random effects. In this article we consider spatial GLMMs where the random effects are modelled by a spatial Gaussian field. Such models and the notation used throughout this paper are briefly described below.

Let $\mathcal{S} = \{S(x) : x \in I\}$ denote a Gaussian field of random effects with mean 0 and index set $I \subseteq \mathbb{R}^2$. We assume that conditionally on $\mathcal{S}$, the random variables $\mathcal{Y} = \{Y(x) : x \in I\}$, are mutually independent, and the error distribution $Y(x)|\mathcal{S}$ has a density $f(\cdot; M(x))$ which only depends on the conditional mean $M(x) = \mathrm{E}[Y(x)|S(x)]$. We restrict attention to the case where the density $f(\cdot;\mu)$ is with respect to counting or Lebesgue measure, and it is of an exponential family form

$$f(z;\mu) = \exp\left(zg_c(\mu) + b(z) - a(g_c(\mu))\right), \quad z \in \Omega, \tag{1}$$

where $\Omega \subseteq \mathbb{R}$ is the support of the density, μ is the mean parameter, and a, b, g_c are real functions; g_c is called the canonical link function.

The conditional mean $M(x)$ is assumed to be related to $S(x)$ by a link function g so that

$$g(M(x)) = S(x) + d(x)^{\mathrm{T}}\beta, \tag{2}$$

where $d(x) \in \mathbb{R}^p$ is a vector of covariates associated with the location x and $\beta \in \mathbb{R}^p$ is a vector of regression parameters. The superscript T denotes transposition of vectors and matrices. The link function is assumed to be continuous differentiable and strictly increasing; these conditions are satisfied in the special case where $g = g_c$. By (2) we cannot choose an arbitrary link function as the range of $g(M(x))$ must be the entire real line. The mean parameter space is the open interval $\mathcal{M} = g^{-1}(\mathbb{R})$.

We focus on the Poisson-log normal model where

$$f(z;\mu) = \exp\left(z\log\mu - \log(z!) - \mu\right), \quad z = 0, 1, \ldots, \tag{3}$$

is a Poisson density, $\mathcal{M} =]0;\infty[$, and $g(\mu) = g_c(\mu) = \log\mu$ is the canonical log-link. In Christensen *et al.* (2000) and in the Appendix we consider two other examples: the binomial density with the canonical logit-link, $g(\mu) = \log(\mu/(N-\mu))$; and the exponential density, with log-link $g(\mu) = \log\mu$ ($g = g_c$ is not valid for the exponential density since the range of $g_c(\mu) = -1/\mu$ is strictly contained in $\mathbb{R}$).

Suppose that we have observed $\mathcal{Y}$ at distinct locations $x_i \in I$, $i = 1, \ldots, n$, so the data $y = (y_1, \ldots, y_n)$ is a realization of $Y = (Y_1, \ldots, Y_n)$, where $Y_i = Y(x_i)$ for $i = 1, \ldots, n$. We set $S_i = S(x_i)$, $d_i^{\mathrm{T}} = d(x_i)^{\mathrm{T}}$, $M_i = g^{-1}(S_i + d_i^{\mathrm{T}}\beta)$, and let $D = (d_1 \ldots d_n)^{\mathrm{T}}$ denote the design matrix of covariates at the locations where we have observations. Then the conditional density of Y given $\mathcal{S}$ depends only on $\mathcal{S}$ through $S = (S_1, \ldots, S_n)$, and it is given by

$$f(y|S) = \prod_{i=1}^{n} f(y_i; M_i). \tag{4}$$

Finally, the Gaussian field $\mathcal{S}$ is assumed to be isotropic and stationary, i.e., the covariance function

$$C(u) = \mathrm{E}\left[S(x)S(x')\right]$$

depends only on the distance $u = ||x - x'||$ between locations $x, x' \in I$ (most of the ideas and results presented easily extend to the anisotropic case). The covariance function is modelled using a positive semi-definite function as in traditional geostatistics, see e.g. Cressie (1993). We consider parametric models

$$C(u) = \sigma^2 \rho(u/\alpha), \quad u \geq 0, \tag{5}$$

where ρ is a known correlation function and the parameter $(\sigma, \alpha) \in]0; \infty[^2$ is unknown; α is a correlation scale parameter and σ^2 is the variance. For example, correlation functions of the form

$$\rho(u) = \exp(-u^\delta), \tag{6}$$

where $0 < \delta \leq 2$, includes the exponential correlation function ($\delta = 1$) and the Gaussian correlation function ($\delta = 2$). Indeed many other correlation functions than (6) may be of relevance in applications, cf. the discussion in Diggle *et al.* (1998).

An alternative would be to model $\mathcal{S}$ by a conditional autoregression (Besag, 1974; Besag and Kooperberg, 1995). This requires the index set I to be countable and equipped with a neighbourhood structure and further introduces problems with edge-effects; see the discussion in Besag and Kooperberg (1995), Besag and Higdon (1999), and Møller and Waagepetersen (1999).

There are of course alternatives to the use of Gaussian random effects. For example, smoothed gamma-random fields are used to model correlation in spatial point patterns and spatial count data in Wolpert and Ickstadt (1998) and Best *et al.* (1998).

2.2 Inference for spatial GLMMs

Several possibilities are available for inference in GLMMs. The most frequently used methods are either pseudo-likelihood (also called penalized quasi-likelihood or h-likelihood) estimation (Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993; Lee and Nelder, 1996) or Bayesian inference (Zeger and Karim, 1991; Diggle *et al.*, 1998). These and two other methods are briefly discussed below.

Pseudo-likelihood has the advantage of being fast since it is not required to integrate out the unobserved random effects. The random effects are instead treated as unknown parameters, which together with the regression and covariance parameters are estimated by maximizing the joint density of observations and random effects. Lee and Nelder (1996) establish consistency and asymptotic normality, but their setup does not cover our situation where there is only one observation for each random effect and the random effects are correlated.

In the Bayesian framework it is easy to account for uncertainty of parameter estimates in the calculation of variance of predictors for unobserved spatial variables. The drawback of the Bayesian approach is that it may on one hand be difficult to elicit informative priors, and the use of flat improper priors may on the other hand lead to improper posterior distributions, see Section 2.3. We discuss Bayesian inference in more detail in Sections 2.3 and 5.

Maximum likelihood estimates for GLMMs may be obtained using Monte Carlo Newton-Raphson, Monte Carlo EM (McCulloch, 1997; Booth and Hobert, 1999) or by Monte Carlo estimation of the likelihood (Geyer and Thompson, 1992; McCulloch, 1997). However, parametric models as in (5) typically lead to a complicated curved exponential family model for S. This makes computations more complex than in the examples in the above-mentioned references, since samples of S cannot be reduced to samples of low dimensional sufficient statistics. It seems not known whether asymptotic normality of the maximum likelihood estimate is valid in the context of spatial GLMMs.

Conditional simulation of $S|Y = y$ is relevant in connection with Bayesian inference, including prediction of a functional of the random effects, as well as with Monte Carlo maximum likelihood estimation. MCMC algorithms for conditional simulation of $S|Y = y$ are discussed in Section 3.

In certain cases the covariance function of S can be expressed in terms of the first and second order moments of the data Y whereby a non-parametric estimate of the covariance function can be constructed. The non-parametric estimate may be useful for choosing an appropriate parametric model as in (5), and the parameters α and σ can be estimated by a minimum contrast method; see Section 4.

2.3 Flat priors in Bayesian inference for spatial GLMMs

Flat priors should be used with caution in Bayesian analysis of GLMMs as demonstrated in Natarajan and McCulloch (1995). They show that the use of an improper prior on the variance for the random effects may lead to an improper posterior. The problem is even more significant for spatial GLMMs, since an improper prior on the correlation scale parameter α in (5) will in general result in an improper posterior as exemplified below.

Suppose for the moment that $\beta = 0$, (5) holds with $\sigma = 1$, and α has prior density π_a. Then posterior propriety is equivalent to

$$\int_0^\infty \mathrm{E}_\alpha[f(y|S)]\pi_a(\alpha)d\alpha < \infty.$$

The assumptions in Section 2.1 imply that $f(y|\cdot)$ is a continuous function, and in many applications, including the example (3), $f(y|\cdot)$ is bounded. Typically, S converges in distribution to a multivariate standard normal distribution as $\alpha \to 0$, and to a vector $(U, U, \dots, U)$ as $\alpha \to \infty$, where U is univariate standard normal. These properties taken together imply that $\lim_{\alpha\to 0} \mathrm{E}_\alpha[f(y|S)]$ and $\lim_{\alpha\to\infty} \mathrm{E}_\alpha[f(y|S)]$ exist and are strictly positive. Thereby the tail-behaviour of the posterior for α is determined by the tails of the prior (a related observation is made by Stein, 1998, concerning the prior for $1/\alpha$ used in Diggle *et al.*, 1998). Thus posterior propriety in general requires propriety of π_a. Note that the posterior variance for α can be made arbitrarily large by choosing a sufficiently diffuse proper prior.

Conditions for proper posteriors in Bayesian analysis of GLMMs with a known singular correlation matrix for the random effects are given in Sun *et al.* (1999), while general conditions for posterior propriety with an improper prior for β in a GLM are studied in Gelfand and Sahu (1999). These results do not cover cases like the Bayesian analysis described in Section 5.1, where we use the Poisson-log normal model with exponential correlation function $\rho(u) = \exp(-u)$ and the following independent priors for the parameters:

$$\pi_a(\alpha) \propto 1/\alpha, \log\alpha \in [a_1; a_2]; \;\; \pi_b(\beta) \propto 1,\, \beta \in \mathbb{R}^8; \;\; \pi_c(\sigma) \propto \sigma^{-1}\exp(-\eta/\sigma), \sigma > 0. \quad (7)$$

That is, a log-uniform prior for α on a finite interval $[a_1; a_2]$, an improper uniform prior for β, and an improper inverse gamma prior for σ. Box and Tiao (1992) suggest to use a flat prior for $\log\sigma$; in Section 5.1 the parameter $\eta > 0$ is chosen so that the prior of $\log\sigma$ is essentially flat on $]0;\infty[$. The posterior is proper according to the following proposition, which is verified in the Appendix.

Proposition 1. *Consider a realization $y = (y_1, \ldots, y_n)$ of the Poisson-log normal model (3), and assume that $y_1, \ldots, y_m$ are positive and $y_{m+1}, \ldots, y_n$ are zero. Let $\kappa_+(\alpha)$ denote the correlation matrix of $S_+ = (S_1, \ldots, S_m)$ and $D_+ = (d_1 \ldots d_m)^{\mathrm{T}}$ the corresponding $m \times p$ design matrix. Suppose that α, β, σ are a priori independent with densities π_a, π_b, π_c, where $\pi_b(\beta) \propto 1$ for all $\beta \in \mathbb{R}^p$. Then the posterior is proper if*

1. *D_+ has rank p,*
2. *$\kappa_+(\alpha)$ is invertible for all $\alpha \in \operatorname{supp}\pi_a$,*
3. *$(|D_+^{\mathrm{T}}\kappa_+^{-1}(\alpha)D_+||\kappa_+(\alpha)|)^{-1/2}\pi_a(\alpha)$ is integrable on $]0;\infty[$,*
4. *$\int_0^\infty \sigma^{p-m}\pi_c(\sigma)d\sigma < \infty$.*

Other conditions may be relevant for other models in order to establish posterior propriety; see the discussion in the Appendix. The condition 3 is trivially satisfied if $\operatorname{supp}\pi_a$ is compact and the mapping $\alpha \mapsto \kappa_+(\alpha)$ is continuous, but the condition is in general not easy to verify when $\operatorname{supp}\pi_a$ is unbounded.

To shed some light on condition 3, consider the case where π_a is proper and S_+ consists of evenly spaced Gaussian random variables on the line with the exponential correlation structure. Then $\kappa_+(\alpha)$ converges to the identity matrix as $\alpha \to 0$, so it suffices to consider the case $\alpha \to \infty$. The precision matrix $\kappa_+(\alpha)^{-1}$ has a simple tridiagonal structure and we can verify that $|\kappa_+(\alpha)|^{-1/2}$ is $O(\alpha^{(m-1)/2})$. Assuming e.g. $D_+ = (1, \ldots, 1)^{\mathrm{T}}$, then $|D_+^{\mathrm{T}}\kappa_+^{-1}(\alpha)D_+|^{-1/2}$ is $O(\alpha^{-1/2})$ and therefore $|D_+^{\mathrm{T}}\kappa_+^{-1}(\alpha)D_+|^{-1/2}|\kappa_+(\alpha)|^{-1/2}$ is $O(\alpha^{m/2-1})$. The condition 3 thus holds if the tail of $\pi_a(\alpha)$ decreases as a polynomial of order less than $-m/2$.

3 Posterior simulation using Langevin-Hastings updates

In this section we discuss MCMC algorithms for posterior simulation in a spatial GLMM.

Recall that $S = (S(x_1), \ldots, S(x_n))$ is the vector of random effects associated with the n locations where we have observations. Suppose that $S^* = (S(x_{n+1}), \ldots, S(x_{n+q}))$ are $q \geq 0$ additional locations of interest for prediction.

Diggle *et al.* (1998) use a fixed scan hybrid algorithm (in the terminology of Roberts and Rosenthal, 1997) where the covariance parameters, the regression parameters, and each of the random effects $S_1, \ldots, S_{n+q}$ are updated in turn in each scan. The update of a random effect S_i is computationally demanding, since it involves the calculation of the conditional variance given the $n + q - 1$ other random effects.

We instead consider a fixed scan hybrid algorithm where the random effects are updated simultaneously using either random walk Metropolis or Langevin-Hastings updates. These two types of updates are discussed in Section 3.1 in the context of posterior simulation of the random effects for fixed values of the model parameters. In Section 3.2 we describe the hybrid MCMC algorithm used in the Bayesian analysis in Section 5.1. Finally, Section 3.3 deals with some computational issues. For background material on MCMC we refer the reader to Besag *et al.* (1995).

3.1 Posterior simulation of random effects

In the following we discuss posterior simulation of the random effects (S, S^*) for fixed values of the model parameters.

Let Σ denote the covariance matrix of (S, S^*) and let $\Sigma^{1/2}$ be a $(n+q) \times d$ 'square root' of Σ so that $\Sigma = \Sigma^{1/2}(\Sigma^{1/2})^{\mathrm{T}}$ — in Section 3.3 we present two different methods for constructing $\Sigma^{1/2}$, where in one case $d = n + q$ while $d \gg n + q$ in the other case. We can assume that $(S, S^*)^{\mathrm{T}} = \Sigma^{1/2}\Gamma$, where Γ follows a d-dimensional standard multivariate Gaussian distribution. By (4), the log density of Γ given $Y = y$ is

$$\log f(\gamma|y) = \mathrm{const}(y) - \frac{1}{2}\|\gamma\|^2 + \sum_{i=1}^{n} \log f(y_i; \mu_i), \tag{8}$$

with

$$\mu_i = \mu_i(\gamma) = g^{-1}(s_i + d_i^{\mathrm{T}}\beta) \tag{9}$$

where $(s_1, \ldots, s_n)^{\mathrm{T}} = Q\gamma$ and Q denotes the upper $n \times d$ submatrix of $\Sigma^{1/2}$.

Posterior simulations of (S, S^*) can be obtained by transforming MCMC samples of the conditional distribution of Γ given $Y = y$. This is obviously advantageous when Σ is not positive definite. For the Langevin-Hastings algorithm considered in Møller *et al.*

(1998) for the case of log Gaussian Cox processes, samples of (S, S^*) given by transforming conditional samples of Γ given $Y = y$ were less auto-correlated than when the algorithm was applied directly to (S, S^*). This may be explained by the fact that the algorithm uses uncorrelated proposals whose correlation structure is in better accordance with the marginal distribution of Γ than that of (S, S^*).

3.1.1 Gaussian random walk Metropolis

The Gaussian random walk Metropolis update is given by two steps. First a proposal γ' is generated from a multivariate normal distribution with mean vector γ equal to the current state and covariance matrix hI, where $h > 0$ is a user-specified parameter. Secondly we return γ' with probability

$$\alpha(\gamma, \gamma') = 1 \wedge \frac{f(\gamma'|y)}{f(\gamma|y)};$$

otherwise the state γ is retained.

3.1.2 Langevin-Hastings algorithm

More efficient algorithms can be obtained by adapting the proposal kernel to the target distribution of interest. Let

$$\nabla(\gamma) = \frac{\partial}{\partial \gamma} \log f(\gamma|y) = -\gamma + Q^{\mathrm{T}} \left\{ (y_i - \mu_i) \frac{g_c'(\mu_i)}{g'(\mu_i)} \right\}_{i=1}^{n} \tag{10}$$

denote the gradient of the log target density. In the Langevin-Hastings update the proposal distribution is a multivariate normal distribution with mean vector $\xi(\gamma) = \gamma + (h/2)\nabla(\gamma)$ and covariance matrix hI, $h > 0$, and the acceptance probability is

$$\alpha(\gamma, \gamma') = 1 \wedge \frac{f(\gamma'|y) \exp(-\frac{1}{2h} \|\gamma - \xi(\gamma')\|^2)}{f(\gamma|y) \exp(-\frac{1}{2h} \|\gamma' - \xi(\gamma)\|^2)}. \tag{11}$$

Using the gradient to adapt the proposal kernel to the target density may lead to much better convergence and mixing properties than for an ordinary random walk Metropolis chain. By Roberts *et al.* (1997) and Roberts and Rosenthal (1998b), the number of iterations required to obtain convergence is $O(d^{-1})$ for the random walk algorithm and $O(d^{-1/3})$ for the Langevin-Hastings algorithm, so the benefit of using Langevin-Hastings increases as the dimension increases.

3.1.3 Geometric ergodicity

Geometric ergodicity of MCMC-algorithms is a desirable property which ensures the validity of central limit theorems for the Monte Carlo estimates, and justifies assessment of the precision of a Monte Carlo estimate by estimation of the asymptotic variance in the limiting normal distribution (see e.g. Roberts and Rosenthal, 1998a).

It is verified in Christensen *et al.* (2000) that the random walk Metropolis algorithm is geometrically ergodic for the model (3) and several other GLMMs. The Langevin-Hastings algorithm is not geometrically ergodic for the model (3) (Proposition 1 in Christensen *et al.*, 2000). The problem is that $\|\nabla(\gamma)\|$ increases very fast when $\|\gamma\| \to \infty$ in some directions. If one replaces $\nabla(\gamma)$ in the Langevin-Hastings proposal kernel with a "truncated" gradient $\nabla(\gamma)^{\text{trunc}}$, then a geometrically ergodic algorithm is obtained for all spatial GLMMs (Theorem 2 in Christensen *et al.*, 2000). For model (3) the gradient is of a simple form

$$\nabla(\gamma) = -\gamma + Q^{\mathrm{T}} \{y_i - \mu_i\}_{i=1}^{n}, \tag{12}$$

and one may take

$$\nabla(\gamma)^{\text{trunc}} - \gamma + Q^{\mathrm{T}} \{y_i - \mu_i \wedge H\}_{i=1}^{n}$$

where $0 < H < \infty$ is a truncation constant.

Christensen *et al.* (2000) consider a numerical study where the truncated Langevin-Hastings algorithm performs much better than the random walk Metropolis algorithm, when performance is measured in terms of asymptotic variances of Monte Carlo estimates of the random effects. The practical importance of using truncation is also discussed in Christensen *et al.* (2000).

3.2 Extension with updates of model parameters

For the Bayesian analysis in Section 5.1 we apply a Poisson-log normal model and priors as in (7). The MCMC computations are carried out using a fixed scan hybrid algorithm where γ, β, σ, and $\log\alpha$ are updated in turn in each scan. As the advantage of using Langevin-Hastings updates is most significant in high dimensions, we just use Gaussian random walk updates when updating the one-dimensional parameters σ and $\log\alpha$ while truncated Langevin-Hastings updates are used for the eight-dimensional regression parameter β and the high-dimensional vector γ. The truncated gradient for β is

$$\frac{\partial}{\partial\beta} \log f(y|\gamma;\beta,\sigma,\alpha) = D^{\mathrm{T}} \{y_i - \mu_i \wedge H\}_{i=1}^{n}.$$

We do not know whether the resulting Markov chain is geometrically ergodic. Geometric ergodicity for hybrid algorithms is studied in Roberts and Rosenthal (1997), but their results are not easily applicable in our situation.

3.3 Calculation of a square root of the covariance matrix

In general the calculation of $\Sigma^{1/2}$ and the transformation of Γ into (S, S^*) is time-consuming when $n+q$ is large. The Cholesky factorization of Σ requires $O(d^3)$ operations, where

$d = n + q$. The complexity of this factorization may not be a problem if it is needed only once, but in a fully Bayesian analysis (like in Section 5) $\Sigma^{1/2}$ must be calculated for each step of the Markov chain. The complexity of transforming Γ into (S, S^*) is in this connection a minor problem, as it only requires $O(d^2)$ operations.

An alternative approach is based on the two-dimensional discrete fast Fourier transform (FFT). Suppose that the locations $(x_1, \ldots, x_{n+q})$ can be embedded in a rectangular $M \times N$ grid. Then Σ_{ext} is a certain extension of Σ, defined on a $M_{ext} \times N_{ext}$ grid chosen sufficient large so that Σ_{ext} becomes positive semi-definite, where $M_{ext} \geq 2(M-1)$ and $N_{ext} \geq 2(N-1)$ are powers of 2 (or 3 or 4); see Dietrich and Newsam (1993) and Wood and Chan (1994) for details. The FFT is first used to obtain a 'square root' $\Sigma_{ext}^{1/2}$, where now $d = N_{ext}M_{ext}$. Letting S^{**} be the auxiliary variables associated with the extra locations on the extended grid,

$$(S, S^*, S^{**})^{\mathrm{T}} = \Sigma_{ext}^{1/2}\Gamma$$

is normally distributed with mean 0 and covariance matrix Σ_{ext} — this transformation is also computed using the FFT. In particular the subvector (S, S^*) is normally distributed with mean 0 and covariance matrix Σ (so we can let $\Sigma^{1/2}$ be the upper $(n+q) \times d$ submatrix of $\Sigma_{ext}^{1/2}$). The FFT requires $O(d \log_2 d)$ operations.

4 Non-parametric estimation of the covariance function

In this section we consider non-parametric estimation of the covariance function for the Poisson-log normal model and other spatial GLMMs where the link function is $g(\mu) = \log \mu$.

4.1 Derivation of the non-parametric estimate

Similarly to ideas used in Aitchison and Ho (1989) and Møller *et al.* (1998), we use below certain mean value relations to obtain an estimator for the covariance function C.

For any distinct locations $x^{(1)}, \ldots, x^{(M)}$, we have using $g(\mu) = \log \mu$ that

$$\mathrm{E} \prod_{i=1}^{M} Y(x^{(i)}) = \exp\left(M\sigma^2/2 + \sum_{i=1}^{M} \sum_{j=i+1}^{M} C(\|x^{(i)} - x^{(j)}\|) + \sum_{i=1}^{M} d(x^{(i)})^{\mathrm{T}}\beta \right). \tag{13}$$

Thereby

$$C(\|x_i - x_j\|) = \log\left(\frac{\mathrm{E}[Y_i Y_j]}{\mathrm{E}\, Y_i \,\mathrm{E}\, Y_j} \right), \quad x_i \neq x_j. \tag{14}$$

Furthermore, it follows that in the special case where the error distribution is Poisson,

$$\mathrm{E}[Y_i(Y_i - 1)] = \exp(2\sigma^2 + 2d_i^{\mathrm{T}}\beta)$$

and so by (13),

$$\sigma^2 = \log\left(\frac{\mathrm{E}[Y_i(Y_i-1)]}{(\mathrm{E}\,Y_i)^2}\right). \tag{15}$$

Now, assume that $d_i^{\mathrm{T}}\beta$ is known (or estimated) for all $i = 1, \ldots, n$; or at least that each term

$$\psi_i = d_i^{\mathrm{T}}\beta - \beta_0$$

is known, where β_0 is a common parameter (intercept) for the mean in the error distribution. Note that if $\hat{\sigma}^2$ denotes an estimator for the variance, we obtain by (13) an estimator for β_0,

$$\hat{\beta}_0 = \log\left(\frac{1}{n}\sum_{i=1}^{n} Y_i \exp(-\psi_i)\right) - \frac{\hat{\sigma}^2}{n}.$$

The first order moments $\mathrm{E}[Y_i \exp(-\psi_i)]$, $i = 1, \ldots, n$ are all equal; similarly for the second order moments $\mathrm{E}[Y_i \exp(-\psi_i) Y_j \exp(-\psi_j)]$. This suggests to rewrite the ratio of mean values in (14) as

$$\frac{\mathrm{E}[Y_i Y_j]}{\mathrm{E}\,Y_i\,\mathrm{E}\,Y_j} = \frac{\mathrm{E}[Y_i \exp(-\psi_i) Y_j \exp(-\psi_j)]}{\mathrm{E}[Y_i \exp(-\psi_i)]\,\mathrm{E}[Y_j \exp(-\psi_j)]}.$$

Hence we may estimate $C(u)$ by

$$\hat{C}(u) = \log\left(\frac{\frac{1}{\mathrm{card}(W_u^\Delta)}\sum_{(i,j)\in W_u^\Delta} Y_i \exp(-\psi_i) Y_j \exp(-\psi_j)}{\left(\frac{1}{n}\sum_{i=1}^{n} Y_i \exp(-\psi_i)\right)^2}\right), \quad u > 0, \tag{16}$$

where

$$W_u^\Delta = \{(i,j) : i,j \in \{1,\ldots,n\}, \|x_i - x_j\| \in [u-\Delta; u+\Delta]\}.$$

and $0 \le \Delta < u$. Similarly, for the Poisson case an estimate of the variance is obtained from (15),

$$\hat{\sigma}^2 = \log\left(\frac{\frac{1}{n}\sum_{i=1}^{n} Y_i \exp(-\psi_i)(Y_i - 1)\exp(-\psi_i)}{\left(\frac{1}{n}\sum_{i=1}^{n} Y_i \exp(-\psi_i)\right)^2}\right). \tag{17}$$

The non-parametric estimate $\hat{C}$ is not necessarily positive semi-definite. A formal approach for fitting a valid parametric covariance function to $\hat{C}$ would be minimum contrast estimation described in Møller *et al.* (1998).

The performance of $\hat{C}$ as an exploratory and diagnostic tool is studied on simulated and real data in Section 4.2 and Section 5.3.

4.2 Simulation experiment

In order to study the performance of $\hat{C}$ we made simulations on a square 21×21 grid with spacing 0.05 under the Poisson-log normal model with all $\psi_i = 0$, using exponential and Gaussian covariance functions with different values of $(\alpha, \beta_0, \sigma)$.

Figure 1 illustrates our general conclusion that $\hat{C}(u)$ is biased downwards with a rather symmetric distribution. The variation of $\hat{C}(u)$ becomes large when $u \geq 0.7$, which is about half the maximal distance on the square grid. In Figure 1, the exponential covariance function and the values $(\alpha, \beta_0, \sigma) = (0.1, -1, 1)$ are used. This value of β_0 gives a large proportion of simulated observations equal to zero. The estimates $\hat{C}(u)$ are calculated for $(\Delta, u) = (0, 0)$, $(\Delta, u) = (0, 0.05)$ (corresponding to nearest horizontal/vertical grid points), $(\Delta, u) = (0, \sqrt{2} \times 0.05)$ (corresponding to nearest diagonal grid points), and for $\Delta = 0.025$ and $u = 0.10, 0.15, \ldots, 1$.

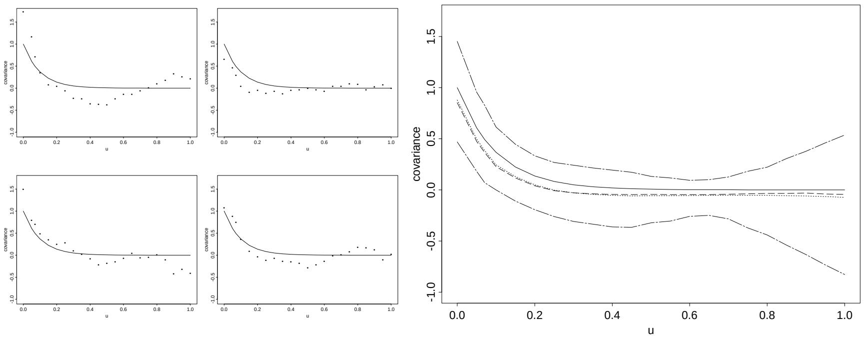

Figure 1: Simulation study when using the exponential covariance function (solid line) and parameters $(\alpha, \beta_0, \sigma) = (0.1, -1, 1)$. Left: four independent simulated estimates of $\hat{C}(u)$ (dots). Right: means, medians, and 2.5 % and 97.5 % quantiles for $\hat{C}(u)$ estimated from 10000 simulations.

In the simulation experiment $\hat{C}$ performs rather well and would be useful for suggesting a parametric model for the covariance and suitable values for the covariance parameters. In Section 5.3 we consider models where the sampling distribution of $\hat{C}$ is more dispersed and $\hat{C}$ consequently less useful.

5 Examples

Section 5 concerns Bayesian inference and computational aspects. In Section 5.1 we consider a data set with counts of weed plants which was collected in a Danish project on precision farming. At many locations the counts equal zero, so assuming a normal distribution would certainly be inappropriate here, even for transformed data. Section 5.2 briefly considers the radionuclide concentration data set studied in Diggle *et al.* (1998). In Section 5.3 the non-parametric estimate of the covariance is used for model validation in connection with the data sets in Sections 5.1 and 5.2.

5.1 Bayesian analysis of weed count data

In a Danish project on precision farming counts of weed plants on a field are recorded in 1993, 1994 and 1995; the data are partly presented and analyzed in Walter *et al.* (1997). One objective of the project was to investigate whether weed occurrence could be predicted from observations of soil texture and soil chemical properties. Here we model the relation between counts of the specie *Viola arvensis* in year 1994 and certain soil properties using a Poisson-log normal model (3), an exponential correlation function as in (6) with $\delta = 1$, and a prior for (α, β, σ) given by (7) with $[a_1; a_2] = [-6.91; -0.29]$ and $\eta = 10^{-6}$. Section 5.1.1 provides a short description of the data and Sections 5.1.2–5.1.3 are concerned with computational issues. The posterior analysis and model assumptions are discussed in Sections 5.1.4–5.1.6.

5.1.1 Description of data

The weed counts are displayed in Figure 2, using the actual values and gray scales. The horizontal axis in Figure 2 corresponds to the ploughing direction, and the counts are observed within $0.25\,m^2$ circular frames with spacing $20\,m$. Except for some missing sites in the first row, the centers of the frames form a rectangular grid, which in the computations was scaled to a rectangle of sidelength 1×0.7.

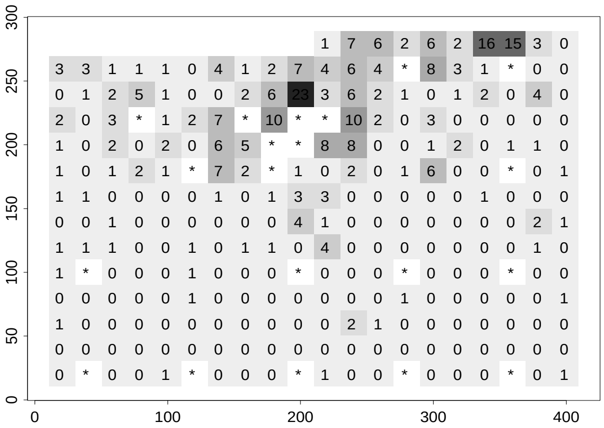

Figure 2: Counts of weed plants.

For most of the locations/centers we have additional information about eleven explanatory variables: 1–2) the two coordinates of the location, 3) organic matter, 4) clay, 5) silt, 6) coarse sand, 7) medium coarse sand, 8) coarse silt, 9) reaction number (pH), 10) content of phosphorus and 11) content of potassium. Observations at locations with missing explanatory variables (indicated with $*$ in Figure 2) are excluded in the analysis below. The five contiguous locations with missing explanatory variables belong to a peatbog where the amount of organic matter is extraordinarily high, and where the texture variables 4–8) have not been measured. The six texture variables 3–8), whose sum equals 100, have been transformed using the generalized logit-transform (Aitchison, 1986) so that the ith transformed texture variable ($i = 3, \ldots, 7$) is given by the log ratio between the ith texture variable and the last texture variable (coarse silt). All explanatory variables are further standardized by first substracting the mean and secondly scaling to have the maximal absolute value equal to one. Thereby only ten explanatory variables $d_{i1}, \ldots, d_{i10}$ are obtained for a location x_i. We include an intercept β_0 in the regression parameter $\beta = (\beta_0, \beta_1, \ldots, \beta_{10})^{\mathrm{T}}$ so that $d_i = (1, d_{i1}, \ldots, d_{i10})^{\mathrm{T}}$ is the covariate vector associated to x_i.

5.1.2 Computations

For the posterior simulations we used 500,000 scans of the hybrid algorithm in Section 3.2 with Cholesky decomposition of the covariance matrix, see Section 3.3. The output was studied by plotting time series and autocorrelations for different statistics. Some representative time series are shown in Figure 3. One may note that equilibrium is reached quickly and that σ has a heavy-tailed posterior distribution.

Theoretical results in Roberts *et al.* (1997), Roberts and Rosenthal (1998b), and Breyer and Roberts (2000) suggest that one should tune the proposal variances to obtain acceptance rates around 0.23 for random walk updates and 0.57 for Langevin-Hastings updates. The overall acceptance rates for the updates of γ, β, σ and $\log\alpha$ were 0.57, 0.56, 0.23, and 0.27, respectively. We observed that the acceptance probability for updates of γ decreased when α increased. As discussed in Sections 3.1.3 and 3.2 it may be advantageous to truncate the conditional means in the gradients for γ and β; we used $H = 50$ (see also Christensen *et al.*, 2000).

5.1.3 Comparison of algorithms

As an alternative to the hybrid Langevin-Hastings/random walk (LH/RW) algorithm applied in Section 5.1 we considered another hybrid algorithm (RW) where random walk Metropolis updates are used for both γ, β, σ, and $\log\alpha$ (with acceptance rates 0.23, 0.22, 0.23, and 0.29 respectively). Figure 4 shows estimated autocorrelations for various parameters and the two algorithms. The autocorrelations decrease much faster when Langevin updates are used for γ and β instead of random walk updates.

The computing times on a 400 Mhz workstation for generating 1000 scans are 146 CPU seconds for the LH/RW-hybrid algorithm and 141 for the RW-hybrid algorithm. If

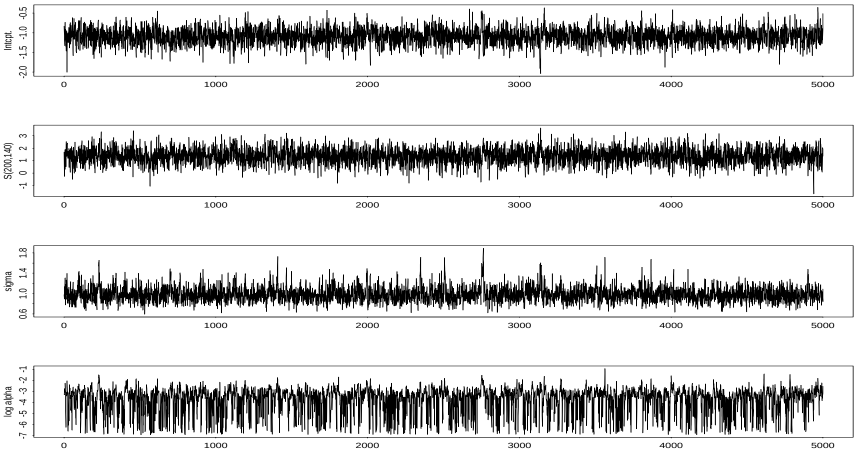

Figure 3: From top to bottom: time series (each 100th scan) for β_0, S_i with $x_i = (200, 140)$, σ, and $\log \alpha$.

one applies the FFT factorization, then an extended grid of size 128×128 is required in order to obtain a positive definite extended covariance matrix for all values of $\log \alpha$ in $[-6.91, -0.29]$. When using the FFT implementation, the computing time for 1000 scans from the LH/RW-hybrid algorithm is 161 CPU seconds.

5.1.4 Posterior distributions

Posterior histograms for the intercept and the covariance parameters are shown in Figure 5. The rightmost plot shows the posterior distribution for $\log \alpha$ obtained with the type of prior used in Diggle *et al.* (1998), i.e. when $\pi(\alpha) \propto 1/\alpha^2$, $\log \alpha \in [-6.91; -0.29]$. Note that the posterior distribution of $\log \alpha$ is heavily influenced by both the shape of the prior and the chosen lower limit for the support of the prior. The other parameters and the random effects have nearly symmetrical posteriors with well-defined modes and supports of moderate size (plots omitted). Posterior means of $\beta_0, \ldots, \beta_{10}$, σ, and $\log \alpha$ are given in Table 1 together with the posterior probabilities for being less than zero. The weed counts seem to depend significantly on the second spatial coordinate (β_2) and on the organic matter explanatory variable (β_3).

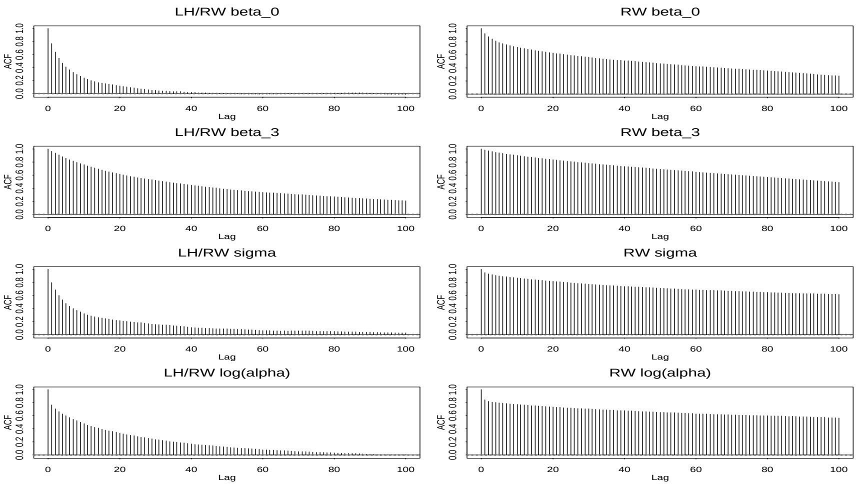

Figure 4: Estimated autocorrelations for various parameters using each 10th scan and either the LH/RW-hybrid algorithm (left) or the RW-hybrid algorithm (right). From top to bottom: parameters β_0, β_3, σ, and $\log\alpha$.

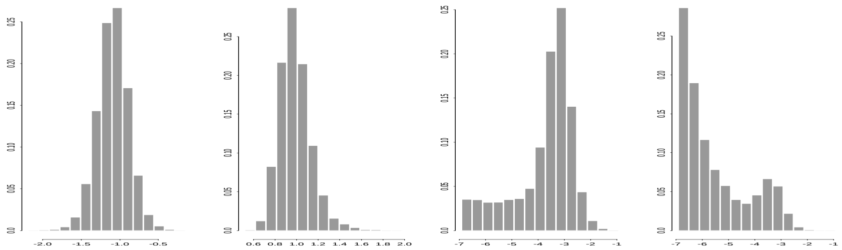

Figure 5: Left to right: marginal posterior distributions for the intercept β_0, σ, $\log\alpha$ (with prior given in (7)), and $\log\alpha$ (with prior as in Diggle *et al.*, 1998).

5.1.5 Comparison with GLIMMIX

As discussed in Section 2.2, an alternative to the Bayesian analysis is to use pseudo-likelihood estimation which is implemented in the SAS macro GLIMMIX. Table 1 shows that the posterior means from the Bayesian analysis and the GLIMMIX estimates are rather similar, and 0 is an extreme value in the posterior distributions for exactly those parameters which are significant at level 5 % according to the Wald-tests from GLIMMIX.

Parameter	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	σ	$\log\alpha$
Post. mean	-1.09	-0.20	2.00	1.70	-0.29	-0.49	0.37	0.19	-0.39	-0.68	0.04	0.99	-3.76
Post. prob.	1.00	0.76	0.00	0.01	0.75	0.81	0.27	0.34	0.83	0.74	0.48		
GLIMMIX	-0.75^*	-0.18	1.83^*	1.50^*	-0.27	-0.38	0.35	0.15	-0.37	-0.53	0.06	0.87	-3.21

Table 1: Monte Carlo posterior means and GLIMMIX estimates. The middle row contains the posterior probabilities that the different parameters are less than 0. A $*$ in the third row indicates that the corresponding variable is significant according to the Wald test computed by GLIMMIX (no test is made for σ and $\log\alpha$).

The posterior mean of the random effects and the GLIMMIX estimate of the random effects are also quite similar, see the left plot in Figure 6. Pseudo-likelihood is computationally much less demanding than the Bayesian/MCMC approach, but is in general considered as less reliable (see the discussion in Lee and Nelder, 1996).

5.1.6 Model validation

If we predict a new observation of Y_i by the fitted value $\hat{\lambda}_i = \mathrm{E}(\exp(d_i^{\mathrm{T}}\beta + S_i)|Y = y)$, we obtain a residual $r_i = (y_i - \hat{\lambda}_i)/\sqrt{\hat{\lambda}_i}$. The residuals versus fitted values are given in the middle plot in Figure 6. The residuals are positively biased when the values of $\hat{\lambda}_i$ are large. A histogram of the posterior means of the random effects is given in the right plot in Figure 6. At present we are not sure what to conclude from the middle and right plot in Figure 6, since we have no knowledge concerning the distribution of the residuals and the posterior means under the assumed Poisson-log normal model. Simulation studies seem to be required to obtain such knowledge.

Another approach, which is more in the spirit of Bayesian inference, is to consider posterior predictive distributions (Rubin, 1984; Gelman *et al.*, 1996). If we condition on β and S_i and let $\lambda_i = \exp(d_i^{\mathrm{T}}\beta + S_i)$, then the magnitude of a standardized residual $(y_i - \lambda_i)/\sqrt{\lambda_i}$ may be assessed by considering the p-value

$$P_i(\beta, S_i) = P\left((Y_i - \lambda_i)^2/\lambda_i \geq (y_i - \lambda_i)^2/\lambda_i |\, \beta, S_i\right). \tag{18}$$

The idea is then to average over the posterior distribution of the unknown S and β given $Y = y$, whereby posterior predictive p-values $p_i = E(P_i(\beta, S_i)|Y = y)$ are obtained (Gelman *et al.*, 1996). The p_i's can easily be calculated along with the other posterior characteristics. The posterior predictive p-values do not give reason to reject the chosen model for the weed count data since the minimum value of p_i is 0.10.

5.2 Radionuclide concentrations on Rongelap Island

The radionuclide concentration data set studied in Diggle *et al.* (1998) consists of measurements of γ-ray counts at $n = 157$ locations. Diggle *et al.* (1998) assume a Poisson-log

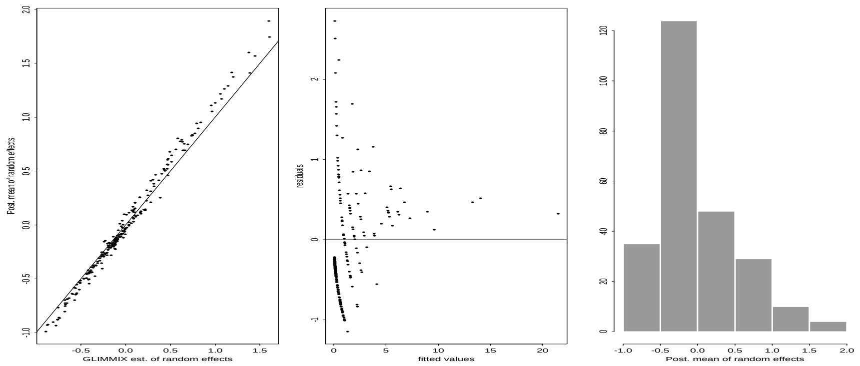

Figure 6: Left: posterior mean of random effects vs. GLIMMIX estimates. Middle: residuals vs. fitted values. Right: histogram for posterior means of random effects.

normal model (3) with $d(x_i)^{\mathrm{T}}\beta = \beta_0 + \log \tau_i$, where τ_i is the length of the recording period at location x_i. Furthermore, the correlation function is assumed to be of the form (6), where uniform priors on bounded intervals are imposed on the parameters σ^2, $1/\alpha$, δ, and β_0 (note that δ was fixed in the previous sections).

We extend the LH/RW-hybrid algorithm with random walk updates of δ and generate MCMC samples from the posterior model considered in Diggle *et al.* (1998). The algorithm does not mix so well as the algorithm in Section 5.1.2 due to high posterior correlation between the model parameters. Estimated posterior means for β_0, σ^2, $6702/\alpha$, and δ based on 5 million scans are 1.84, 0.31, 61.90, 0.84, respectively. These estimates differ from the values 1.7, 0.65, 22.8, 0.7 obtained by Diggle *et al.* (1998) who only employed 50,000 scans of a Metropolis-Hastings chain with single-site updating of the random effects. In order to study this difference one would need to take into account the Monte Carlo error. A formal justification for calculating asymptotic variances is however missing as we do not know if the two algorithms are geometrically ergodic.

5.3 Model validation based on the nonparametric estimator for the covariance function

In this section we supplement the simulation study in Section 4.2 with a discussion on the performance of the non-parametric estimator for the covariance function (Section 4.1) when applied to the weed count data and the radionuclide concentration data.

We consider first the model in Section 5.2. In order to check the chosen covariance structure for S it would be desirable to compare the observed $\hat{C}$ with the sampling distribution $\mathcal{D}(\hat{C}|\beta_0, \sigma, \alpha, \delta)$ of $\hat{C}$ given the model parameters. Note that this is partly different from the situation considered in (18) where we condition on S also. Figure 7 shows the observed $\hat{C}$ together with Monte Carlo estimates of the mean and the 5% and 95% quantiles for the sampling distribution of $\hat{C}$, when the unknown model parameters are replaced by the posterior means. The value $\hat{C}(0)$ is below the 5% quantile and it is actually less than the minimal simulated value of $\hat{C}(0)$ for the 10,000 independent simulated realizations used for the Monte Carlo estimates of the mean and quantiles in Figure 7. This shows a lack of fit of the used model; several possibilities for modification of the model are mentioned in the discussion part of Diggle *et al.* (1998).

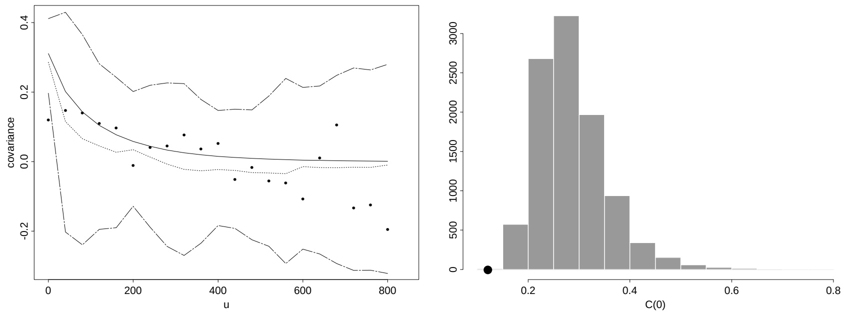

Figure 7: Left: $\hat{C}$ based on the radionuclide concentration data (dots), estimated mean (dotted line), 5% quantiles, and 95% quantiles for the sampling distribution of $\hat{C}$ (dotted and dashed lines), and the parametric covariance function with parameters equal to the posterior means (solid line). Right: The simulated distribution of $\hat{C}(0)$ and the estimated value of $\hat{C}(0)$ (dot).

We have also simulated the sampling distribution of $\hat{C}$ with $(\beta_0, \sigma, \log\alpha, \delta)$ estimated by the posterior mean found in Diggle *et al.* (1998). Similar to Section 5.1.6 another approach is to consider the posterior predictive distribution of $\hat{C}$, but now obtained by averaging $\mathcal{D}(\hat{C}|\beta_0, \sigma, \alpha, \delta)$ over the posterior distribution of the model parameters given $Y = y$. The results obtained with the posterior means in Diggle *et al.* (1998) and with the posterior predictive distribution are qualitatively similar to Figure 7.

By comparing the curves in Figure 7 for the estimated parametric covariance function and the mean of $\hat{C}(u)$, we see that $\hat{C}(u)$ is biased downwards. This is in accordance with the experience in Section 4.2.

We also applied the non-parametric estimate in a similar way for checking the appropriateness of a stationary model $\beta_1 = \ldots = \beta_7 = 0$ for the weed count data in Section 5.1. From this analysis we were not able to reject the fitted stationary model with remaining parameters given by estimated posterior means. This is possibly due to strong correlation in the fitted model which lead to a very dispersed sampling distribution of $\hat{C}$.

6 Discussion

6.1 Inference

The discussion in Section 2.3 concerning the posterior for the correlation parameter α and the results in Section 5.1.4 raise some questions concerning the use of Bayesian inference for α. The posterior of α is heavily sensitive to the choice of prior and the posterior variance can in fact be arbitrary large by choosing the prior diffuse enough. In the absence of prior knowledge for α we can therefore not rely on using a very diffuse prior. It would be interesting to compare the Bayesian inference with results obtained using maximum likelihood estimation. We believe that the Langevin-Hastings algorithm discussed in Section 3.1.2 would be useful as the simulation component of a procedure for obtaining Monte Carlo maximum likelihood estimates.

In the example in Section 5.1 we have not considered prediction of unobserved weed counts. This is straightforward using our MCMC algorithm if the explanatory variables are available at the locations where predictions are required. If this is not the case, one could in principle include Gaussian random field models for the continuous explanatory variables and extend the MCMC algorithm with conditional simulations of the unobserved explanatory variables.

Further studies on simulated and real data seem required to assess the usefulness of the residuals r_i and the non-parametric estimate $\hat{C}$ for model validation. Crossvalidation is another but computationally intensive possibility for model checking.

6.2 Algorithms and computations

We have demonstrated the advantages of using Langevin updates over random walk updates. The proposal kernel in Section 3.1.2 is based on the Euler-discretization of the Langevin diffusion for $f(\cdot \mid y)$, see Roberts and Tweedie (1996). An alternative is to construct proposal kernels based on more refined discretizations of the Langevin diffusion as suggested in Stramer and Tweedie (1999) and further studied in the multidimensional case in Roberts and Stramer (2000).

As in Section 3.1.2, let $\nabla(\gamma)$ denote the gradient of the log posterior density. Further, let $J(\gamma)$ be the second derivative of the log posterior density. The so-called local linearization scheme (Ozaki, 1992; Shoji and Ozaki, 1998; Stramer and Tweedie, 1999) applied to the Langevin diffusion gives rise to a proposal kernel of the form $N(\mu_\gamma, K_\gamma)$ where

$$\mu_\gamma = \gamma + J(\gamma)^{-1}(\exp(hJ(\gamma)/2) - I)\nabla(\gamma)$$

and

$$K_\gamma = J(\gamma)^{-1}(\exp(hJ(\gamma)) - I).$$

The examples studied in Stramer and Tweedie (1999) and Roberts and Stramer (2000) show that much faster convergence to the equilibrium distribution may be obtained when using an algorithm based on local linearization instead of the simple Euler-discretization. However, in the context of this paper, the evaluation of $J(\gamma)^{-1}$ and the matrix exponential $\exp(hJ(\gamma)/2)$ is computationally very demanding.

Considering the hybrid algorithm in Section 3.2 we believe that further development is needed. The Langevin-Hastings algorithm works very well for fixed model parameters but the use of Langevin-Hastings updates in the hybrid algorithm does not necessarily solve mixing problems due to high posterior correlation between model parameters. It is also important to investigate the geometric ergodicity properties of the hybrid algorithm (or an improved version of it).

Acknowledgement: This research was supported by the European Union's network "Statistical and Computational Methods for the Analysis of Spatial Data. ERB-FMRX-CT96-0096", by MaPhySto – Centre for Mathematical Physics and Stochastics, funded by a grant from The Danish National Research Foundation – and by the Danish Informatics Network in the Agricultural Sciences, funded by a grant from the Danish Research Councils. We wish to thank Svend Elsnab Olesen, Svend Erik Simmelsgaard, and Anne Mette Walter from the Danish Institute of Agricultural Sciences for providing the data used in Section 5.1. We are also grateful to Gareth Roberts and Osnat Stramer for showing us their preprint and to Peter Diggle for helpful comments to the manuscript.

Appendix: proof of posterior propriety

We start by verifying Proposition 1.

Set

$$I(y,\alpha,\sigma) = \int_{\mathbb{R}^p} \mathrm{E}_{\alpha,\beta,\sigma}\left[f(y|S)\right]\, d\beta$$

and

$$I(y) = \int_0^\infty \int_0^\infty I(y,\alpha,\sigma)\pi_a(\alpha)\pi_c(\sigma) d\alpha d\sigma.$$

Since $\pi_b(\beta)$ is constant, the posterior is proper if $I(y) < \infty$.

Let $p_+(\cdot;\alpha,\beta,\sigma)$ denote the density of the multivariate normal distribution with mean $D_+\beta$ and covariance matrix $\sigma^2\kappa_+(\alpha)$. Recalling (2) and (4) and considering the model (3), we obtain that

$$I(y,\alpha,\sigma) \le \int_{\mathbb{R}^m}\int_{\mathbb{R}^p} \prod_{i=1}^m f(y_i; g^{-1}(s_{+i}))p_+(s_+;\alpha,\beta,\sigma)d\beta ds_+.$$

Letting $\hat\beta(s_+) = (D_+^{\mathrm{T}}\kappa_+^{-1}(\alpha)D_+)^{-1}D_+^{\mathrm{T}}\kappa_+^{-1}(\alpha)s_+$ be the maximum likelihood estimate of β

based on s_+ when (α, σ) is fixed, it is well-known that

$$p_+(s_+; \alpha, \beta, \sigma) = p_+(s_+; \alpha, \hat{\beta}(s_+), \sigma) \exp\left(-\frac{1}{2\sigma^2}(\beta - \hat{\beta}(s_+))^{\mathrm{T}} D_+^{\mathrm{T}} \kappa_+^{-1}(\alpha) D_+ (\beta - \hat{\beta}(s_+))\right).$$

Since $(2\pi)^{-m/2}|\sigma^2\kappa_+(\alpha)|^{-1/2}$ is an upper bound on the density p_+, we obtain that

$$I(y, \alpha, \sigma) \leq \frac{|(D^{\mathrm{T}}\kappa_+^{-1}(\alpha)D)^{-1}|^{1/2}\sigma^{p-m}}{|\kappa_+(\alpha)|^{1/2}(2\pi)^{(m-p)/2}} \int_{\mathbb{R}^m} \prod_{i=1}^{m} f(y_i; g^{-1}(s_{+i}))ds_+ ,$$

where the latter integral is finite. Thereby $I(y) < \infty$, so the posterior is proper as asserted. $\square$

Next we investigate to what extend the conditions in Proposition 1 are needed.

(I) The proof of Proposition 1 relies on $f(y_i; g^{-1}(s_{+i}))$ being integrable as a function of s_i for $i = 1, \ldots, m$. Therefore Proposition 1 also holds with $m = n$ when the conditional density is exponential and the log-link is used. For the binomial distribution with $N > 1$ and the logit-link, Proposition 1 holds with $y_1, \ldots, y_m \notin \{0, N\}$ and $y_{m+1}, \ldots, y_n \in \{0, N\}$. The case $N = 1$ is not covered by the results of Proposition 1.

(II) We can verify that $\int_1^\infty \sigma^{p-m}\pi_c(\sigma)d\sigma < \infty$ is necessary, but not that $\int_0^1 \sigma^{p-m}\pi_c(\sigma)d\sigma < \infty$ is necessary.

(III) We have also considered situations where the prior for β is proper and established posterior propriety under various conditions; we omit these details as the relevant conditions can be rather problem specific.

References

Aitchison, J. (1986). *The statistical analysis of compositional data*. Chapman and Hall, London.

Aitchison, J. and Ho, C. (1989). The multivariate Poisson-log normal distribution. *Biometrika* **76**, 643–653.

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

Besag, J. E. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. R. Statist. Soc.* B **36**, 192–236.

Besag, J. E. (1994). Discussion on the paper by Grenander and Miller. *J. R. Statist. Soc.* B **56**, 591–592.

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

Besag, J. E., Green, P. J., Higdon, D. and Mengersen, K. (1995). Bayesian computation and stochastic systems. *Statist. Sci.* **10**, 3–66.

Best, N. G., Ickstadt, K. and Wolpert, R. L. (1998). Spatial Poisson regression for health and exposure data measured at disparate resolutions. *Discussion paper 98-36*, Institute of Statistics and Decision Sciences, Duke University.

Booth, J. G. and Hobert, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *J. R. Statist. Soc.* B **61**, 265–285.

Box, G. E. P. and Tiao, G. C. (1992). *Bayesian Inference in Statistical Analysis*. Wiley, New York.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Am. Statist. Assoc.* **88**, 9–25.

Breyer, L. A. and Roberts, G. O. (2000). From Metropolis to diffusions: Gibbs states and optimal scaling. *Stoch. Proc. Appl.* (to appear).

Christensen, O. F., Møller, J. and Waagepetersen, R. (2000). Geometric ergodicity of Metropolis Hastings algorithms for conditional simulation in generalised linear mixed models. *Research Report R-00-2010*, Department of Mathematics, Aalborg University.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. Revised edition. Wiley, New York.

Dietrich, C. R. and Newsam, G. N. (1993). A fast and exact method for multidimensional Gaussian stochastic simulation. *Water Resources Research* **29**, 2861–2869.

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

Gelfand, A. E. and Sahu, S. K. (1999). On the proprierity of posteriors and Bayesian identifiability in generalized linear models. *J. Am. Statist. Assoc.* **94**, 247–253.

Gelman, A., Xiao-Li, M. and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica* **6**, 733–807.

Geyer, C. J. and Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *J. R. Statist. Soc.* B **54**, 657–699.

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models. *J. R. Statist. Soc.* B **58**, 619–678.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*. Chapman and Hall, London, 2nd edition.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *J. Am. Statist. Ass.* **92**, 162–170.

Møller, J. and Waagepetersen, R. (1999). Discussion on the paper by Besag and Higdon. *J. R. Statist. Soc.* B **61**, 735.

Møller, J., Syversveen, A. R. and Waagepetersen, R. (1998). Log Gaussian Cox processes. *Scand. J. Statist.* **25**, 451–482.

Natarajan, R. and McCulloch, C. E. (1995). A note on the existence of the posterior distribution for a class of mixed models for binomial responses. *Biometrika* **82**, 639–643.

Ozaki, T. (1992). A bridge between nonlinear time series models and nonlinear stochastics dynamical systems: a local linearization approach. *Stat. Sin.* **2**, 113–135.

Roberts, G. O. and Rosenthal, J. S. (1997). Geometric ergodicity and hybrid Markov chains. *Electron. Commun. Probab.* **2**, 13–25.

Roberts, G. O. and Rosenthal, J. S. (1998a). Markov chain Monte Carlo: some practical implications of theoretical results. *Canad. J. Statist.* **26**, 5–31.

Roberts, G. O. and Rosenthal, J. S. (1998b). Optimal scaling of discrete approximations to Langevin diffusions. *J. R. Statist. Soc.* B **60**, 255–268.

Roberts, G. O. and Stramer, O. (2000). Tempered Langevin diffusions and algorithms. In preparation.

Roberts, G. O. and Tweedie, R. L. (1996). Exponential convergence of Langevin diffusions and their discrete approximations. *Bernoulli* **2**, 341–363.

Roberts, G. O., Gelman, A. and Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.* **7**, 110–120.

Rossky, P. J., Doll, J. D. and Friedman, H. L. (1978). Brownian dynamics as smart Monte Carlo simulation. *J. Chemical Physics* **69**, 4628–4633.

Rubin, D. B. (1984). Bayesian justifiable and relevant frequency calculations for the applied statistician. *Ann. Stat.* **12**, 1151–1172.

Shoji, T. and Ozaki, T. (1998). A statistical method of estimation and simulation for systems of stochastic differential equations. *Biometrika* **85**, 240–243.

Stein, M. (1998). Discussion on the paper by Diggle, Tawn and Moyeed. *Appl. Statist.* **47**, 341.

Stein, M. (1999). *Spatial interpolation, some theory for kriging*. Springer Verlag, New York.

Stramer, O. and Tweedie, R. L. (1999). Langevin-type models II: self-targeting candidates for MCMC algorithms. *Methodol. Comput. Appl. Probab.* **1**, 307–328.

Sun, D., Speckman, P. L. and Tsutakawa, R. K. (1999). Random effects in generalized linear mixed models (GLMMs). In: *Generalized linear models: A Bayesian perspective* (eds. D. K. Dey and S. K. Ghosh), Marcel Dekker, 23–40.

Walter, A. M., Heisel, T. and Christensen, S. (1997). Shortcuts in weed mapping. In: *Precision Agriculture 1997* (ed. J. V. Stafford), BIOS Scientific Publishers Ltd., 777–784.

Wolfinger, R. and O'Connell, M. (1993). Generalized linear mixed models: a pseudo-likelihood approach. *J. Statist. Comput. Simul.* **48**, 233–243.

Wolpert, R. L. and Ickstadt, K. (1998). Poisson/gamma random field models for spatial statistics. *Biometrika* **85**, 251–267.

Wood, A. T. A. and Chan, G. (1994). Simulation of stationary Gaussian processes in $[0,1]^d$. *J. Comput. Graph. Statist.* **3**, 409–432.

Zeger, S. L. and Karim, M. R. (1991). Generalized linear models with random effects; a Gibbs sampling approach. *J. Am. Statist. Ass.* **86**, 79–86.