Estimation and Prediction for Spatial Generalized Linear Mixed Models Using High Order Laplace Approximation

Evangelos Evangelou [*†] *Zhengyuan Zhu* [‡] *Richard L. Smith* [§]

Abstract

Estimation and prediction in generalized linear mixed models are often hampered by intractable high dimensional integrals. This paper provides a framework to solve this intractability, using asymptotic expansions when the number of random effects is large. To that end, we first derive a modified Laplace approximation when the number of random effects is increasing at a lower rate than the sample size. Secondly, we propose an approximate likelihood method based on the asymptotic expansion of the log-likelihood using the modified Laplace approximation which is maximized using a quasi Newton algorithm. Finally, we define the second order plug-in predictive density based on a similar expansion to the plug-in predictive density and show that it is a normal density. Our simulations show that in comparison to other approximations, our method has better performance. Our methods are readily applied to non-Gaussian spatial data and as an example, the analysis of the rhizoctonia root rot data is presented.

Keywords: Generalized linear mixed models; Laplace approximation; Maximum likelihood estimation; Predictive inference; Spatial statistics.

1 Introduction

As an extension of the generalized linear model, the Generalized Linear Mixed Model (GLMM) is used to allow different sources of variability in the mean response. This is achieved by including random effects in the linear predictor in addition to the fixed effects. In their simplest form, the random effects are taken to be independent, but a more general covariance structure is often assumed; see the examples in Breslow and Clayton (1993) and in Diggle et al. (1998). For the estimation of the parameters, no analytical methods are available because the likelihood is expressed as an intractable integral over the random effects (Breslow and Clayton, 1993). Instead, several methods have been proposed for approximating the integral numerically. These include simulation-based methods, as in McCulloch (1997) and Zhang (2002), or approximation methods; see *inter alia* Breslow and Clayton (1993), Shun (1997), Raudenbush et al. (2000), and Noh and Lee (2007).

The Laplace approximation (Barndorff-Nielsen and Cox, 1989) is a method for approximating integrals of the form $\int e^{-g(\boldsymbol{z})}\,\mathrm{d}\boldsymbol{z}$, a form which can be associated with the GLMM likelihood, where $\boldsymbol{z}$ are the random effects and e^{-g} is the joint density of the observations and the random effects. The idea behind the Laplace approximation is to replace the exponent of the integrand by its Taylor expansion around the point where it is maximized. The (first order) Laplace approximation requires a second order Taylor expansion and has relative error $O(n^{-1})$, n being the sample size. This method has been successfully used in Bayesian statistics for approximating posterior expectations (Tierney and Kadane, 1986; Tierney et al., 1989) and by Breslow and Clayton (1993) to approximate the GLMM likelihood. Although practical for many cases, Breslow and Clayton's method yields a non-negligible bias when applied to binary clustered data. Breslow and Lin (1995),

[*]Department of Mathematical Sciences, University of Bath, Bath, UK

[†]Corresponding author. Addr: Department of Mathematical Sciences, University of Bath, Bath, BA2 7AY, UK; Tel: 00441225385673; email: ee224@bath.ac.uk

[‡]Department of Statistics, Iowa State University, Ames IA, USA

[§]Department of Statistics and Operations Research, UNC Chapel Hill, Chapel Hill NC, USA

Lin and Breslow (1996), and Wolfinger and Lin (1997) give further improvements and alternatives on this idea. Even so, the approximation is not always effective: as indicated in Solomon and Cox (1992), it becomes unreliable as the variance of the random effects increases.

Shun and McCullagh (1995) and Shun (1997) also use maximum likelihood with the assumption that the dimension of the random effects increases with the sample size. This assumption is necessary for the variance components to be estimated consistently but under this framework it is not clear if the remainder term in the classical Laplace approximation is bounded. In their paper Shun and McCullagh derived a formula that takes this into account by grouping terms according to their asymptotic order; an application of this methodology for independent crossed random effects is illustrated in Shun (1997). Although their method performs well for small sample sizes, it becomes slow even for moderate samples because it involves the summing over many terms. In fact, Shun suggests the exclusion of some terms from the likelihood to speed up the algorithm. Noh and Lee (2007) propose an effective way to include these terms when the design matrix of the random effects is sparse. Furthermore Raudenbush et al. (2000) derived higher order correction to the Laplace approximation following the asymptotic expansion in Shun and McCullagh (1995).

The Laplace approximation can also be a useful tool for approximating the predictive density of the random effects. From a frequentist point of view, Booth and Hobert (1998) suggest an optimal predictor in terms of the conditional mean square error and Vidoni (2006) gives approximation formulae for the plug-in predictive density. While useful, these methods apply only in the case where the random effects are independent. In a Bayesian framework, Rue et al. (2009) and Eidsvik et al. (2009) combine Laplace approximation with Gauss-Hermite quadrature to provide a fast and accurate method for approximating the predictive density. On the other hand, the computational advantages of their method are in effect when the inverse covariance matrix for the random effects is sparse and the number of parameters is small.

A popular application of GLMM is to model geostatistical data in which observations are drawn from a number of different locations. In the simplest case, the observations are considered to be Gaussian with correlation depending on the distance between them (Cressie, 1993; Stein, 1999). Diggle et al. (1998) extend the Gaussian geostatistical model to include parametric families depending only on the mean of the spatial process in the same way that the classical linear model is extended to the generalized linear mixed model. Their model assumes that the observations are driven by an unobserved Gaussian random field with mean depending linearly on a set of covariates and covariance depending on a small number of parameters. They use Bayesian MCMC methods for the estimation of the parameters as well as for prediction at unsampled locations. Further suggestions toward the Bayesian MCMC approach were made in Christensen et al. (2000) and Christensen and Waagepetersen (2002). From a frequentist point of view, Zhang (2002) proposes a combination of the MCMC and the EM algorithm for estimation and prediction while Christensen (2004) suggests simulated maximum likelihood estimation via MCMC and Kuk (1999) and Booth and Hobert (1999) recommend using importance sampling based on Laplace approximation to approximate the maximum likelihood estimates. Considerably less attention has been given to approximation methods, because maximization of the approximate likelihood requires optimization of a function of as many variables as the number of locations for several values of the parameters, making it unsuitable for large samples. The likelihood approximation proposed in this paper does not require to carry large optimization problems but instead reduces them to several univariate ones.

Motivated by examples such as in Diggle et al. (2004) and Crainiceanu et al. (2008) where inference on large datasets is required in a short amount of time, in this paper we propose methods that can be used for estimation and prediction of spatial GLMM. Our development is based on a high order Laplace approximation derived by generalizing the formula by Shun and McCullagh (1995). Concerning estimation, we observe that the joint likelihood of the parameters can be written as an integral of a product of two functions with different orders of magnitude, simplifying calculation if one chooses to expand the integral around the optimum point of the function with the highest order only. The advantage is that for the application of Laplace approximation, we need to optimize only over univariate functions, improving the speed of the estimation without reducing the performance. Furthermore, we are able to derive closed form expressions for the first and second derivatives of the approximate likelihood with respect to the parameters and hence we propose a quasi-Newton algorithm for obtaining the approximate likelihood estimates.

For prediction we apply a high order asymptotic expansion to the predictive density and notice that the approximation has the form of a Normal density. Using this result we construct prediction intervals for the random effects at the unsampled locations. This method resembles Vidoni (2006) in the case of correlated

random effects and for random effects prediction instead of response prediction. We present a simulation study where we compare our predictions with other methods and show that our method achieves comparable accuracy and it is much faster. An application of our methods is also presented.

The remainder of this paper is organized as follows. In section 2 we introduce the spatial GLMM, in section 3 we derive asymptotic formulae for Laplace approximation to approximate integrals which we use in the subsequent sections for maximum likelihood estimation (section 4) and for prediction (section 5). Our simulations are presented in section 6 and the application to the rhizoctonia root rot data in section 7. Section 8 features a discussion and concludes.

2 Model and notation

The vector of the response variable is denoted by $\boldsymbol{Y}$ with components $\{Y_{il},\, i = 1,\dots,k, l = 1,\dots,n_i\}$ repeatedly sampled at k different sampling sites, $s_1,\dots,s_k$, within a domain $\mathbb{S}$. Depending on the application, $\mathbb{S}$ might represent, for example, a spatial region or a time domain. In practice, a fine grid covering the region of interest is considered and therepeated sampling corresponds to samples from locations around a point of the grid.

We assume the existence of an unobserved homogeneous random field $\mathcal{Z}$ defined on $\mathbb{S}$ such that conditioned on $\mathcal{Z}$ the observations are independent with distribution from the exponential family. We denote by $\boldsymbol{Z} = \{Z_i, i = 1,\dots,k\}$ the k-dimensional vector that consists of the components of $\mathcal{Z}$ that correspond to the k sampled sites and we refer to it as the random effects. Furthermore, the conditional mean $\mu_i = \mathrm{E}(Y_{il}|Z_i) = b(\theta_i)$ for some known differentiable function b, called the cumulant function, such that b' is strictly increasing, and conditional variance $\omega v(\mu_i)$ where v is a known function called the variance function and ω is an additional nuisance parameter called the dispersion parameter (McCullagh and Nelder, 1999). The parameter θ_i relates to the linear predictor $\eta_i = \boldsymbol{x}_i^\intercal\beta + Z_i$ through the relationship $\mu_i = b(\theta_i) = g^{-1}(\eta_i)$ for some function g called the link function. In our asymptotic analysis we consider the case in which k and n_i increase to infinity with the n_i's having the same order, $\min\{n_i\} = O(n)$ but k is increasing in a lower rate, $k/n \to 0$.

We assume that the finite dimensional distributions of the random field $\mathcal{Z}$ are Normal with mean $\mathbf{0}$ and covariance matrix parameterized by γ, i.e. $\boldsymbol{Z} \sim \mathrm{N}_k(\mathbf{0}, \Sigma(\gamma))$ where $\Sigma(\cdot)$ is a known function and its probability density function is denoted by $\phi(\boldsymbol{z};\gamma)$. Conditional on $\boldsymbol{Z}$, the density of $\boldsymbol{Y}$ has the form

$$f(\boldsymbol{y}|\boldsymbol{z};\beta) = \exp\left[\left\{\sum_{i=1}^{k} y_i(\boldsymbol{x}_i^\intercal\beta + z_i) - \sum_{i=1}^{k} n_i b(\boldsymbol{x}_i^\intercal\beta + z_i)\right\}\cdot\frac{1}{\omega} + \sum_{i=1}^{k} c(y_i,\omega)\right] \tag{1}$$

where $y_i = \sum_{j=1}^{n_i} y_{ij}$, and for known functions b and c. Although in (1) we implicitly used the canonical link for the distribution of $\boldsymbol{y}$, $\theta_i = \boldsymbol{x}_i^\intercal\beta + z_i$, the results that follow don't necessarily require this restriction.

The goal of the paper is to estimate (β,γ) and to predict a component Z_0 of $\mathcal{Z}$ that corresponds to an unsampled site. To that end, note that the likelihood based on $\boldsymbol{y}$ is

$$L(\beta,\gamma|\boldsymbol{y}) = \int f(\boldsymbol{y}|\boldsymbol{z};\beta)\phi(\boldsymbol{z};\gamma)\,\mathrm{d}\boldsymbol{z} \tag{2}$$

In addition, writing the distribution function of $Z_0|\boldsymbol{Z}$ as $\Phi(z_0|\boldsymbol{z};\gamma)$ and its density function as $\phi(z_0|\boldsymbol{z};\gamma)$, the predictive distribution function for Z_0 given the data $\boldsymbol{y}$ is

$$F(z_0|\boldsymbol{y};\beta,\gamma) = \frac{\int \Phi(z_0|\boldsymbol{z};\gamma) f(\boldsymbol{y}|\boldsymbol{z};\beta)\phi(\boldsymbol{z};\gamma)\,\mathrm{d}\boldsymbol{z}}{\int f(\boldsymbol{y}|\boldsymbol{z};\beta)\phi(\boldsymbol{z};\gamma)\,\mathrm{d}\boldsymbol{z}}. \tag{3}$$

Similarly, the predictive density is written as

$$f(z_0|\boldsymbol{y};\beta,\gamma) = \frac{\int \phi(z_0|\boldsymbol{z};\gamma) f(\boldsymbol{y}|\boldsymbol{z};\beta)\phi(\boldsymbol{z};\gamma)\,\mathrm{d}\boldsymbol{z}}{\int f(\boldsymbol{y}|\boldsymbol{z};\beta)\phi(\boldsymbol{z};\gamma)\,\mathrm{d}\boldsymbol{z}} \tag{4}$$

None of the integrals appearing in (2), (3) and (4) have an analytic expression and hence maximum likelihood estimation or prediction cannot be performed using standard numerical optimization procedures. To overcome this problem, the next section derives a formula that allows us to approximate the likelihood and the predictive density when the sample size is large.

3 Asymptotic expansions of integrals

For the derivations illustrated here, we follow the notation of McCullagh (1987) and use indices to denote components of arrays, derivatives and summations. For sums, an index that appears as a subscript and as a superscript implies a summation over all possible values of that index. Therefore, we will denote the components of a vector sometimes by subscripts and sometimes by superscripts. For example, the components of the three dimensional vector $\boldsymbol{x}$ will be written as x_1, x_2 and x_3 or as x^1, x^2 and x^3 depending on the expression i.e. $x_i x^i = x^i x_i = \sum_{i=1}^3 (x_i)^2$ but $x^i x^i$ is the square of the ith element of $\boldsymbol{x}$: $(x_i)^2$. The (i,j) component of a matrix A will be written as a_{ij} and its inverse (when exists) will have components a^{ij}.

For any real function $f(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^k$, its derivative with respect to the i^{th} component of $\boldsymbol{x}$ is denoted by a subscript i.e. $f_i(\boldsymbol{x}) = \frac{\partial f(\boldsymbol{x})}{\partial x_i}$ and $f_{ij}(\boldsymbol{x}) = \frac{\partial^2 f(\boldsymbol{x})}{\partial x_i \partial x_j}$. Furthermore, $f_{\boldsymbol{x}}$ is the gradient of f and $f_{\boldsymbol{xx}}$ is the Hessian matrix. Based on our notation on matrix inversion, f^{ij} is the (i,j) element of $f_{\boldsymbol{xx}}^{-1}$: the inverse of the Hessian matrix.

3.1 Modified Laplace approximation

Shun and McCullagh (1995) proposed a modification of Laplace approximation that can be used for evaluating integrals of the form

$$I_1 = \int e^{-g(\boldsymbol{z})} \, \mathrm{d}\boldsymbol{z} \tag{5}$$

where $g = O(n)$. Assuming that g has a unique minimum at $\hat{\boldsymbol{z}}$, Shun and McCullagh suggest an expansion of the integral around that minimum. They derive the identities

$$\log I_1 = -\hat{g} - \frac{1}{2} \log \left| \frac{\hat{g}_{\boldsymbol{zz}}}{2\pi} \right| + \sum_{m=1}^{\infty} \sum_{\substack{P,Q \\ P \vee Q = 1}} \frac{(-1)^t}{(2m)!} \hat{g}_{p_1} \dots \hat{g}_{p_t} \hat{g}^{q_1} \cdots \hat{g}^{q_m} \tag{6}$$

$$I_1 = e^{-\hat{g}} \left| \frac{\hat{g}_{\boldsymbol{zz}}}{2\pi} \right|^{-1/2} \sum_{m=0}^{\infty} \sum_{P,Q} \frac{(-1)^t}{(2m)!} \hat{g}_{p_1} \dots \hat{g}_{p_t} \hat{g}^{q_1} \cdots \hat{g}^{q_m} \tag{7}$$

where the second sum in each of (6) and (7) is over all partitions P, Q such that $P = p_1 | \dots | p_t$ is a partition of $2m$ indices into t blocks, each of size 3 or more and $Q = q_1 | \dots | q_m$ is a partition of the same indices into m blocks, each of size 2. $P \vee Q = 1$ means that the union of the graphs produced by joining elements in the same block of the two partitions is connected e.g. $Q = i_1 i_2 | i_3 i_4$ is connected with $P_1 = i_1 | i_2 i_3 | i_4$ but not with $P_2 = i_1 | i_2 | i_3 i_4$ (see Figure 1). The summation over all the possible values of the $2m$ indices is also implicit. For example, the first terms, up to $m = 3$, of (7) are

$$\begin{aligned} I_1 = e^{-\hat{g}} \left| \frac{\hat{g}_{\boldsymbol{zz}}}{2\pi} \right|^{-1/2} \Bigg\{ & 1 - \frac{3}{4!} \sum \hat{g}_{i_1 i_2 i_3 i_4} \hat{g}^{i_1 i_2} \hat{g}^{i_3 i_4} + \frac{90}{6!} \sum \hat{g}_{i_1 i_2 i_3} \hat{g}_{i_4 i_5 i_6} \hat{g}^{i_1 i_2} \hat{g}^{i_3 i_4} \hat{g}^{i_5 i_6} \\ & + \frac{60}{6!} \sum \hat{g}_{i_1 i_2 i_3} \hat{g}_{i_4 i_5 i_6} \hat{g}^{i_1 i_4} \hat{g}^{i_2 i_5} \hat{g}^{i_3 i_6} - \frac{15}{6!} \sum \hat{g}_{i_1 i_2 i_3 i_4 i_5 i_6} \hat{g}^{i_1 i_2} \hat{g}^{i_3 i_4} \hat{g}^{i_5 i_6} + \dots \Bigg\} \end{aligned}$$

where the summations are over all indices $i_1, i_2, \dots$ ranging from 1 to k. For $k = 3$ the first sum in the previous equation is

$$\hat{g}_{1111} \hat{g}^{11} \hat{g}^{11} + \hat{g}_{1112} \hat{g}^{11} \hat{g}^{12} + \hat{g}_{1113} \hat{g}^{11} \hat{g}^{13} + g_{1121} \hat{g}^{11} \hat{g}^{21} + \dots + \hat{g}_{3333} \hat{g}^{33} \hat{g}^{33} \quad (3^4 \text{ terms})$$

These formulae require expressing the integrand in a fully exponential form while for the results here we require the integrand to be written in the standard form (Tierney et al., 1989).

In our approach, we consider approximating the following integral:

$$I_2 = \int \exp\{-g(\boldsymbol{z})\} \times f(\boldsymbol{z}) \, \mathrm{d}\boldsymbol{z} \tag{8}$$

where f is not necessarily positive. Suppose that $\boldsymbol{z} \in \mathbb{R}^k$, $g(\boldsymbol{z}) = O(n)$ has a minimum at $\mathbf{0}$ and f and its derivatives are $O(1)$. A Taylor expansion of g around $\mathbf{0}$ gives

$$g(\boldsymbol{z}) = \hat{g} + \frac{1}{2!} z^{i_1} z^{i_2} \hat{g}_{i_1 i_2} + \frac{1}{3!} z^{i_1} z^{i_2} z^{i_3} \hat{g}_{i_1 i_2 i_3} + \frac{1}{4!} z^{i_1} z^{i_2} z^{i_3} z^{i_4} \hat{g}_{i_1 i_2 i_3 i_4} + \dots \tag{9}$$

Fig. 1: Connected partitions Q and P_1 (left) and unconnected Q and P_2 (right)

where the subscripts of g imply differentiation with respect to the indicated component of $\boldsymbol{z}$ and the hats imply that the function or its derivatives are evaluated at $\mathbf{0}$. The indices range from 1 to k and the sums are over all indices. Let $\hat{g}_{\boldsymbol{zz}}$ denote the hessian matrix of g evaluated at $\mathbf{0}$.

A similar expansion of f around the same point gives

$$f(z) = \hat{f} + \hat{f}_{j_1} z^{j_1} + \frac{1}{2}\hat{f}_{j_1 j_2} z^{j_1} z^{j_2} + \dots \tag{10}$$

Thus, letting for $r \geq 3$, $\hat{g}_{[i_1 \dots i_r]} = \sum_P \hat{g}_{p_1} \dots \hat{g}_{p_t}$ where P ranges over all partitions of $i_1 \dots i_r$ with blocks of size 3 or more, we have

$$\begin{aligned}
I_2 &= e^{-\hat{g}} \int e^{-\frac{1}{2}\boldsymbol{z}^\mathsf{T}\hat{g}_{\boldsymbol{zz}}\boldsymbol{z}} \exp\left\{-\frac{1}{3!}\hat{g}_{i_1 i_2 i_3} z^{i_1} z^{i_2} z^{i_3} - \frac{1}{4!}\hat{g}_{i_1 i_2 i_3 i_4} z^{i_1} z^{i_2} z^{i_3} z^{i_4} - \dots\right\} \\
&\qquad \times \left(\hat{f} + \hat{f}_{j_1} z^{j_1} + \frac{1}{2}\hat{f}_{j_1 j_2} z^{j_1} z^{j_2} + \dots\right) \mathrm{d}\boldsymbol{z} \\
&= e^{-\hat{g}} \int e^{-\frac{1}{2}\boldsymbol{z}^\mathsf{T}\hat{g}_{\boldsymbol{zz}}\boldsymbol{z}} \left(1 - \frac{1}{3!}\hat{g}_{[i_1 i_2 i_3]} z^{i_1} z^{i_2} z^{i_3} - \frac{1}{4!}\hat{g}_{[i_1 i_2 i_3 i_4]} z^{i_1} z^{i_2} z^{i_3} z^{i_4} - \dots\right) \\
&\qquad \times \left(\hat{f} + \hat{f}_{j_1} z^{j_1} + \frac{1}{2}\hat{f}_{j_1 j_2} z^{j_1} z^{j_2} + \dots\right) \mathrm{d}\boldsymbol{z} \\
&= e^{-\hat{g}} \left|\frac{\hat{g}_{\boldsymbol{zz}}}{2\pi}\right|^{-1/2} \mathrm{E}\left[\left(1 - \frac{1}{3!}\hat{g}_{[i_1 i_2 i_3]} W^{i_1} W^{i_2} W^{i_3} - \frac{1}{4!}\hat{g}_{[i_1 i_2 i_3 i_4]} W^{i_1} W^{i_2} W^{i_3} W^{i_4} - \dots\right)\right. \\
&\qquad \left. \times \left(\hat{f} + \hat{f}_{j_1} W^{j_1} + \frac{1}{2}\hat{f}_{j_1 j_2} W^{j_1} W^{j_2} + \dots\right)\right]
\end{aligned}$$

where $\boldsymbol{W}$ is a normally distributed random variable with mean $\mathbf{0}$ and covariance matrix $\hat{g}_{\boldsymbol{zz}}^{-1}$.

Then,

$$I_2 = e^{-\hat{g}} \left|\frac{\hat{g}_{\boldsymbol{zz}}}{2\pi}\right|^{-1/2} \sum_{r \in \{0,3,4,\dots\}} \sum_{s=0}^{\infty} (-1)^r \frac{1}{r!s!} \hat{g}_{[i_1 \dots i_r]} \hat{f}_{j_1 \dots j_s} \mathrm{E}\left[W^{i_1} \cdots W^{i_r} \cdot W^{j_1} \cdots W^{j_s}\right]$$

where we make the convention if $r = 0$ then $\hat{g}_{[i_1 \dots i_r]} = 1$, if $s = 0$ then $\hat{f}_{j_1 \dots j_s} = \hat{f}$ and if $r = s = 0$ then $\mathrm{E}\left[W^{i_1} \cdots W^{i_r} \cdot W^{j_1} \cdots W^{j_s}\right] = 1$

Using equation (2.8) from McCullagh (1987), I_2 becomes

$$I_2 = e^{-\hat{g}} \left|\frac{\hat{g}_{\boldsymbol{zz}}}{2\pi}\right|^{-1/2} \sum_{m=0}^{\infty} \sum_{s=0}^{2m} \sum_{P,Q} \frac{(-1)^t}{(2m)!} \hat{f}_{j_1 \dots j_s} \hat{g}_{p_1} \cdots \hat{g}_{p_t} \hat{g}^{q_1} \cdots \hat{g}^{q_m} \tag{11}$$

where P is a partition of $2m - s$ indices into t blocks each of size 3 or more and Q is a partition of the same indices together with $\{j_1, \dots, j_s\}$ into m blocks of size 2. Note that P and Q do not need to be connected.

In the special case where $f(\boldsymbol{z}) > 0$, say $f(\boldsymbol{z}) = \exp\{h(\boldsymbol{z})\}$, then from (11),

$$\log I_2 = -\hat{g} + \hat{h} - \frac{1}{2}\log\left|\frac{1}{2\pi}\hat{g}_{\boldsymbol{zz}}\right| + \sum_{m=1}^{\infty} \frac{1}{(2m)!} \sum_{\substack{P,Q \\ P \vee Q = 1}} \chi_{p_1} \cdots \chi_{p_t} \cdot \hat{g}^{q_1} \dots \hat{g}^{q_m} \tag{12}$$

where

$$\chi_{i_1\cdots i_s} = \begin{cases} \hat{h}_{i_1\cdots i_s} & \text{if } s \le 2 \\ \hat{h}_{i_1\cdots i_s} - \hat{g}_{i_1\cdots i_s} & \text{if } s \ge 3 \end{cases}$$

3.2 Approximation to the ratio of two integrals

In the following sections we will need to approximate ratios of integrals e.g. when we want to approximate conditional densities. Suppose we want to approximate

$$\frac{I_2}{I_1} = \frac{\int \exp\{-g(\boldsymbol{z})\} \times f(\boldsymbol{z})\,\mathrm{d}\boldsymbol{z}}{\int e^{-g(\boldsymbol{z})}\,\mathrm{d}\boldsymbol{z}} \tag{13}$$

Using equations (7) and (11),

$$\frac{I_2}{I_1} = \frac{\sum_{m=0}^{\infty}\sum_{s=0}^{2m}\sum_{P,Q}\frac{(-1)^t}{(2m)!}\hat{f}_{j_1\cdots j_s}\hat{g}_{p_1}\cdots\hat{g}_{p_t}\hat{g}^{q_1}\cdots\hat{g}^{q_m}}{\sum_{m=0}^{\infty}\sum_{P,Q}\frac{(-1)^t}{(2m)!}\hat{g}_{p_1}\cdots\hat{g}_{p_t}\hat{g}^{q_1}\cdots\hat{g}^{q_m}} \tag{14}$$

To illustrate the use of (14), suppose $kn^{-1} \to 0$ and that f and its derivatives are $O(1)$ as $k \to \infty$. In addition, suppose that g and its derivatives are $O(n)$ when the differentiation is performed with respect to the same component of $\boldsymbol{z}$, otherwise they are $O(1)$. As we will show later in Lemma 1, under the aforementioned assumptions on the derivatives of g, the inverse Hessian matrix of g is $O(n^{-1})$ at the diagonal and $O(n^{-2})$ at the off diagonal elements as $k \to \infty$. This a typical situation which we encounter in the subsequent sections. The numerator of (13) is approximated by

$$\hat{f} - \frac{1}{8}\hat{f}\hat{g}_{i_1i_2i_3i_4}\hat{g}^{i_1i_2}\hat{g}^{i_3i_4} + \frac{1}{8}\hat{f}\hat{g}_{i_1i_2i_3}\hat{g}_{i_4i_5i_6}\hat{g}^{i_1i_2}\hat{g}^{i_3i_4}\hat{g}^{i_5i_6} + \frac{1}{12}\hat{f}\hat{g}_{i_1i_2i_3}\hat{g}_{i_4i_5i_6}\hat{g}^{i_1i_4}\hat{g}^{i_2i_5}\hat{g}^{i_3i_6} - \frac{1}{2}\hat{f}_{i_1}\hat{g}_{i_2i_3i_4}\hat{g}^{i_1i_2}\hat{g}^{i_3i_4} + \frac{1}{2}\hat{f}_{j_1j_2}\hat{g}^{j_1j_2} + O(kn^{-2}) \tag{15}$$

where besides the first term: $\hat{f}$, all the other terms in (15) are $O(k\,n^{-1})$. A similar expansion exists for the denominator by replacing f in (15) by 1. Thus (14) becomes after we take $\hat{f}$ as a common factor

$$\begin{aligned}\frac{I_2}{I_1} = \hat{f}\Bigg(&1 - \frac{1}{8}\hat{g}_{i_1i_2i_3i_4}\hat{g}^{i_1i_2}\hat{g}^{i_3i_4} + \frac{1}{8}\hat{g}_{i_1i_2i_3}\hat{g}_{i_4i_5i_6}\hat{g}^{i_1i_2}\hat{g}^{i_3i_4}\hat{g}^{i_5i_6} + \frac{1}{12}\hat{g}_{i_1i_2i_3}\hat{g}_{i_4i_5i_6}\hat{g}^{i_1i_4}\hat{g}^{i_2i_5}\hat{g}^{i_3i_6} \\ &- \frac{1}{2}\frac{\hat{f}_{i_1}}{\hat{f}}\hat{g}_{i_2i_3i_4}\hat{g}^{i_1i_2}\hat{g}^{i_3i_4} + \frac{1}{2}\frac{\hat{f}_{j_1j_2}}{\hat{f}}\hat{g}^{j_1j_2} + O(kn^{-2})\Bigg) \\ &\times\Bigg(1 - \frac{1}{8}\hat{g}_{i_1i_2i_3i_4}\hat{g}^{i_1i_2}\hat{g}^{i_3i_4} \\ &+ \frac{1}{8}\hat{g}_{i_1i_2i_3}\hat{g}_{i_4i_5i_6}\hat{g}^{i_1i_2}\hat{g}^{i_3i_4}\hat{g}^{i_5i_6} + \frac{1}{12}\hat{g}_{i_1i_2i_3}\hat{g}_{i_4i_5i_6}\hat{g}^{i_1i_4}\hat{g}^{i_2i_5}\hat{g}^{i_3i_6} + O(kn^{-2})\Bigg)^{-1}\end{aligned} \tag{16}$$

Employing the identity $(1-\epsilon)^{-1} = 1 + \epsilon + O(\epsilon^2)$ we have in (16) after canceling between the numerator and the denominator

$$\begin{aligned}\frac{I_2}{I_1} &= \hat{f}\left(1 - \frac{1}{2}\frac{\hat{f}_{i_1}}{\hat{f}}\hat{g}_{i_2i_3i_4}\hat{g}^{i_1i_2}\hat{g}^{i_3i_4} + \frac{1}{2}\frac{\hat{f}_{j_1j_2}}{\hat{f}}\hat{g}^{j_1j_2} + O(kn^{-2})\right) \\ &= \hat{f} - \frac{1}{2}\hat{f}_{i_1}\hat{g}_{i_2i_3i_4}\hat{g}^{i_1i_2}\hat{g}^{i_3i_4} + \frac{1}{2}\hat{f}_{j_1j_2}\hat{g}^{j_1j_2} + O(kn^{-2})\end{aligned} \tag{17}$$

4 Approximate likelihood estimation in GLMM

Define

$$\ell(\beta, \gamma|\boldsymbol{y}, \boldsymbol{z}) = \log f(\boldsymbol{y}|\boldsymbol{z}; \beta) + \log \phi(\boldsymbol{z}; \gamma) \tag{18}$$

to be the log-likelihood when the *complete* dataset $(\boldsymbol{y}, \boldsymbol{z})$ is observed. Then, the likelihood based only on $\boldsymbol{y}$ is defined by integrating over the unobserved random effects:

$$\ell(\beta, \gamma|\boldsymbol{y}) = \log \int \exp\{\ell(\beta, \gamma|\boldsymbol{y}, \boldsymbol{z})\} \, \mathrm{d}\boldsymbol{z} \tag{19}$$

Joint maximization of (19) with respect to (β, γ) yields the maximum likelihood estimates for those parameters based on the data $\boldsymbol{y}$. Unfortunately, there is no direct way of evaluating (19) for the different values of the parameters because the integration cannot be carried out analytically. To derive the order of the asymptotic approximations, we need to know the order of the elements of $\ell^{-1}_{\boldsymbol{zz}}(\beta, \gamma|\boldsymbol{y}, \boldsymbol{z})$, the inverse Hessian matrix of the log-likelihood of the complete data. To that end, we show the following lemma:

Lemma 1. *If $k = o(n)$ then the diagonal elements of $\ell^{-1}_{\boldsymbol{zz}}(\beta, \gamma|\boldsymbol{y}, \boldsymbol{z})$ are $O(n^{-1})$ and the off diagonal are $O(n^{-2})$.*

Proof. Keeping only the terms that depend on $\boldsymbol{z}$, the Hessian of the complete log-likelihood has the form $\ell_{\boldsymbol{zz}} = n\, D - \Sigma^{-1}$, where D and Σ^{-1} are $k \times k$ matrices of order $O(1)$ and D is diagonal. Let I be the identity matrix. Using $(I - \varepsilon\, A)^{-1} = I + \varepsilon\, A + \varepsilon^2 A^2 + \varepsilon^3 A^3 + \ldots$, we have

$$\begin{aligned}
\ell^{-1}_{\boldsymbol{zz}} &= (n\, D - \Sigma^{-1})^{-1} \\
&= n^{-1} D^{-1} \{I - n^{-1} (D\Sigma)^{-1}\}^{-1} \\
&= n^{-1} D^{-1} \{I + n^{-1} (D\Sigma)^{-1} + O(k\, n^{-2})\} \\
&= n^{-1} D^{-1} + n^{-2} (D\Sigma D)^{-1} + O(k\, n^{-3})
\end{aligned}$$

where the diagonal elements of $\ell^{-1}_{\boldsymbol{zz}}(\beta, \gamma|\boldsymbol{y}, \boldsymbol{z})$ are $O(n^{-1})$ and the off diagonal are $O(n^{-2})$. □

4.1 Approximate likelihood

Our first approximation consists of writing (19) as

$$\ell(\beta, \gamma|\boldsymbol{y}) = \log \int \exp\{-h(\boldsymbol{y}, \boldsymbol{z}; \beta, \gamma)\} \, \mathrm{d}\boldsymbol{z} \tag{20}$$

and defining $\hat{\boldsymbol{z}} = \operatorname{argmin}_{\boldsymbol{z}} h(\boldsymbol{y}, \boldsymbol{z}; \beta, \gamma)$. Then by (6), ignoring the terms that don't depend on the parameters,

$$\begin{aligned}
\ell(\beta, \gamma; \boldsymbol{y}) = -\hat{h} - \frac{1}{2} \log \left| \hat{h}_{\boldsymbol{zz}} \right| \\
- \frac{1}{8} \hat{h}_{iiii} \hat{h}^{ii} \hat{h}^{ii} + \frac{1}{12} \hat{h}_{iii} \hat{h}_{iii} \hat{h}^{ii} \hat{h}^{ii} \hat{h}^{ii} + \frac{1}{8} \hat{h}_{i_1 i_1 i_1 i_1} \hat{h}_{i_2 i_2 i_2 i_2} \hat{h}^{i_1 i_1} \hat{h}^{i_2 i_2} \hat{h}^{i_1 i_2} + O(k\, n^{-2})
\end{aligned} \tag{21}$$

where the functions in the right hand side are evaluated at $\hat{\boldsymbol{z}}$.

The terms $\hat{h}_{iiii} \hat{h}^{ii} \hat{h}^{ii}$ and $\hat{h}_{iii} \hat{h}_{iii} \hat{h}^{ii} \hat{h}^{ii} \hat{h}^{ii}$ appearing in (21) have order $O(k\, n^{-1})$ and the term $\hat{h}_{i_1 i_1 i_1 i_1} \hat{h}_{i_2 i_2 i_2 i_2} \hat{h}^{i_1 i_1} \hat{h}^{i_2 i_2} \hat{h}^{i_1 i_2}$ has order $O(k^2\, n^{-2})$. The remainder terms which are excluded from (21), such as $\hat{h}_{iiiiii} \hat{h}^{ii} \hat{h}^{ii} \hat{h}^{ii}$ and $\hat{h}_{iiii} \hat{h}_{iiii} \hat{h}^{ii} \hat{h}^{ii} \hat{h}^{ii} \hat{h}^{ii}$, have order $O(k\, n^{-2})$. Parameter estimation may be carried out by maximizing the right hand side of (21). From a practical point of view, if k is too large, obtaining $\hat{\boldsymbol{z}}$ efficiently can be numerically challenging and this has to be performed for several values of the parameters.

A second approach is to write the likelihood in the form of (8) as

$$L(\beta, \gamma|\boldsymbol{y}) = |\Sigma|^{-1/2} \int \exp\{\lambda(\boldsymbol{z}; \gamma)\} \exp\{-\xi(\boldsymbol{y}|\boldsymbol{z}; \beta)\} \, \mathrm{d}\boldsymbol{z} \tag{22}$$

where now

$$\xi(\boldsymbol{y}|\boldsymbol{z}; \beta) = -\sum y_i (\boldsymbol{x}_i^\intercal \beta + z_i) + \sum n_i b(\boldsymbol{x}_i^\intercal \beta + z_i) + \frac{1}{2} \sum \sigma^{ii} z^i z^i \tag{23}$$

$$\lambda(\boldsymbol{z}; \gamma) = -\frac{1}{2} \sum_{i_1 \neq i_2} \sigma^{i_1 i_2} z^{i_1} z^{i_2} \tag{24}$$

and note that ξ has order $O(kn)$ while λ has order $O(k)$. Let $\hat{\boldsymbol{z}}$ be the value of $\boldsymbol{z}$ that minimizes (23). Substituting for λ in the place of h and for ξ in the place of g in (12), we have

$$\ell(\beta,\gamma|\boldsymbol{y}) = -\frac{1}{2}\log|\Sigma| - \hat{\xi} + \hat{\lambda} - \frac{1}{2}\log|\hat{\xi}_{\boldsymbol{zz}}| + \frac{1}{2}\hat{\lambda}_i\hat{\lambda}_i\hat{\xi}^{ii} - \frac{1}{2}\hat{\lambda}_i\hat{\xi}_{iii}\hat{\xi}^{ii}\hat{\xi}^{ii} - \frac{1}{8}\hat{\xi}_{iiii}\hat{\xi}^{ii}\hat{\xi}^{ii} + \frac{5}{24}\hat{\xi}_{iii}\hat{\xi}_{iii}\hat{\xi}^{ii}\hat{\xi}^{ii}\hat{\xi}^{ii} + O(kn^{-2}) \tag{25}$$

There is a significant computational advantage when using (25) instead of (21) because each component of $\hat{\boldsymbol{z}}$ is obtained separately from the others by solving

$$-y_i + n_i b'(\boldsymbol{x}_i\beta + \hat{z}_i) + \sigma^{ii}\hat{z}^i = 0 \tag{26}$$

for each i. Below we describe how the parameters can be estimated.

4.2 An algorithm for obtaining the approximate likelihood estimates

For the rest of this section we drop the summation convention for the indices.

The approximation in (25) can be written as a sum

$$\hat{\ell}(\beta,\gamma|\boldsymbol{y}) = -\frac{1}{2}\log|\Sigma| + \sum_{i=1}^{k} T(\beta,\gamma|y_i,\hat{\boldsymbol{z}}) \tag{27}$$

where

$$\begin{aligned} T(\beta,\gamma|y_i,\boldsymbol{z}) &= y_i(\boldsymbol{x}_i^\intercal\beta + z_i) - n_i b(\boldsymbol{x}_i^\intercal\beta + z_i) - \frac{1}{2}\sigma^{ii}z_i^2 + \frac{1}{2}\lambda_i z_i - \frac{1}{2}\log\xi_{ii} \\ &\quad + \frac{1}{2}\lambda_i^2/\xi_{ii} - \frac{1}{2}\lambda_i\xi_{iii}/\xi_{ii}^2 - \frac{1}{8}\xi_{iiii}/\xi_{ii}^2 + \frac{5}{24}\xi_{iii}^2/\xi_{ii}^3 \\ \xi_{ii} &= n_i b''(\boldsymbol{x}_i^\intercal\beta + z_i) + \sigma^{ii} \qquad \xi_{iii} = n_i b^{(3)}(\boldsymbol{x}_i^\intercal\beta + z_i) \\ \xi_{iiii} &= n_i b^{(4)}(\boldsymbol{x}_i^\intercal\beta + z_i) \qquad \lambda_i = -\sum_{i'=1}^{k}\sigma^{ii'}z_{i'} + \sigma^{ii}z_i \end{aligned} \tag{28}$$

The *approximate* score function, $\hat{\boldsymbol{u}}(\beta,\gamma)$, has components:

$$\begin{aligned} \frac{\mathrm{d}\hat{\ell}}{\mathrm{d}\beta_l} &= \sum\frac{\mathrm{d}}{\mathrm{d}\beta_l}T(\beta,\gamma|y_i,\hat{\boldsymbol{z}}) + \sum\frac{\mathrm{d}}{\mathrm{d}z_i}T(\beta,\gamma|y_i,\hat{\boldsymbol{z}})\frac{\mathrm{d}\hat{z}_i}{\mathrm{d}\beta_l} \\ \frac{\mathrm{d}\hat{\ell}}{\mathrm{d}\gamma_j} &= -\frac{1}{2}\operatorname{tr}(\Sigma^{-1}\Sigma_j) + \sum\frac{\mathrm{d}}{\mathrm{d}\gamma_j}T(\beta,\gamma|y_i,\hat{\boldsymbol{z}}) + \sum\frac{\mathrm{d}}{\mathrm{d}z_i}T(\beta,\gamma|y_i,\hat{\boldsymbol{z}})\frac{\mathrm{d}\hat{z}_i}{\mathrm{d}\gamma_j} \end{aligned}$$

and the approximate Hessian has components expressed in a similar fashion but not written explicitly here.

Note that by differentiating (26) with respect to the parameters we obtain analytical expressions for the derivatives of $\hat{z}_i$, i.e.

$$\frac{\mathrm{d}\hat{z}_i}{\mathrm{d}\beta_l} = -\frac{n_i b''(\boldsymbol{x}_i\beta + \hat{z}^i)x_{il}}{n_i b''(\boldsymbol{x}_i\beta + \hat{z}^i) + \sigma^{ii}}, \qquad \frac{\mathrm{d}\hat{z}_i}{\mathrm{d}\gamma_j} = -\frac{\frac{\mathrm{d}\sigma^{ii}}{\mathrm{d}\gamma_j}\hat{z}^i}{n_i b''(\boldsymbol{x}_i\beta + \hat{z}^i) + \sigma^{ii}}$$

and similarly for the second derivatives, where $\frac{\mathrm{d}\sigma^{ii}}{\mathrm{d}\gamma_j}$ is the ith diagonal element of $\frac{\mathrm{d}}{\mathrm{d}\gamma_j}\Sigma^{-1}$. The approximate likelihood estimates are defined as those values $(\hat{\beta},\hat{\gamma})$ that maximize (27). These are obtained using a quasi-Newton iteration scheme (e.g. Byrd et al., 1995) by taking advantage of the closed form expressions for the score function. Similar arguments can be made for the approximation in (21).

This procedure can be considered an alternative to other computational intensive methods such as MCMC. A concern for applying the method proposed here is that if the sample size is small, the bias is not negligible because the approximate likelihood is not very close to the true likelihood. A bias corrected estimate can be calculated using parametric bootstrap (Efron and Tibshirani, 1993).

5 Prediction

Consider the problem of predicting the random effect at site s_0, Z_0 say, based on observations $y_1, \ldots, y_k$ corresponding to the sampling sites $s_1, \ldots, s_k$.

A solution to this problem is given by the *predictive density* $f(z_0|\boldsymbol{y}; \beta, \gamma)$. Using the fact that conditional on $\boldsymbol{Z}$, the random effect at site s_0 is independent of the observations at the sampled sites, we write

$$f(z_0|\boldsymbol{y}; \beta, \gamma) = \frac{\int \phi(z_0|\boldsymbol{z}; \gamma) f(\boldsymbol{y}|\boldsymbol{z}; \beta) \phi(\boldsymbol{z}; \gamma) \, \mathrm{d}\boldsymbol{z}}{\int f(\boldsymbol{y}|\boldsymbol{z}; \beta) \phi(\boldsymbol{z}; \gamma) \, \mathrm{d}\boldsymbol{z}} \tag{29}$$

As the likelihood, the predictive density does not have a closed form expression. Furthermore, the predictive density depends on the unknown parameters β and γ. A common approach to overcome this last problem would be to replace the unknown parameters by some consistent estimates $(\hat{\beta}, \hat{\gamma})$, giving rise to the so-called *plug-in predictive density*.

5.1 Second order corrected plug-in Predictive Density

Suppose that, based on the sample $\boldsymbol{y} = (y_1, \ldots, y_k)^\intercal$ drawn from the sampling sites $s_1, \ldots, s_k$, we estimate the parameters β and γ by $\hat{\beta}$, respectively $\hat{\gamma}$. The plug-in predictive density is given by

$$f(z_0|\boldsymbol{y}; \hat{\beta}, \hat{\gamma}) = \frac{\int \phi(z_0|\boldsymbol{z}; \hat{\gamma}) f(\boldsymbol{y}|\boldsymbol{z}; \hat{\beta}) \phi(\boldsymbol{z}; \hat{\gamma}) \, \mathrm{d}\boldsymbol{z}}{\int f(\boldsymbol{y}|\boldsymbol{z}; \hat{\beta}) \phi(\boldsymbol{z}; \hat{\gamma}) \, \mathrm{d}\boldsymbol{z}} \tag{30}$$

The expression in (17) allows us to construct an approximation to the predictive distribution of Z_0. Write $\exp\{-h(\boldsymbol{y}, \boldsymbol{z}; \beta, \gamma)\}$ to be the density of $(\boldsymbol{Y}, \boldsymbol{Z})$ and

$$f(z_0|\boldsymbol{y}; \hat{\beta}, \hat{\gamma}) = \frac{\int \phi(z_0|\boldsymbol{z}; \hat{\gamma}) \exp\{-h(\boldsymbol{y}, \boldsymbol{z}; \hat{\beta}, \hat{\gamma})\} \, \mathrm{d}\boldsymbol{z}}{\int \exp\{-h(\boldsymbol{y}, \boldsymbol{z}; \hat{\beta}, \hat{\gamma})\} \, \mathrm{d}\boldsymbol{z}} \tag{31}$$

and define $\hat{\boldsymbol{z}} = \operatorname{argmin}_{\boldsymbol{z}} h(\boldsymbol{y}, \boldsymbol{z}; \hat{\beta}, \hat{\gamma})$. Then, letting μ and τ be the conditional mean and variance of $Z_0|\boldsymbol{Z}$, by (17)

$$\log f(z_0|\boldsymbol{y}; \hat{\beta}, \hat{\gamma}) = \log \hat{\phi} - \frac{1}{2} \frac{\hat{\phi}_{i_1}}{\hat{\phi}} \hat{h}_{i_2 i_2 i_2} \hat{h}^{i_1 i_2} \hat{h}^{i_2 i_2} + \frac{1}{2} \frac{\hat{\phi}_{i_1 i_2}}{\hat{\phi}} \hat{h}^{i_1 i_2} + O(k\, n^{-2}) \tag{32}$$

where $\log \hat{\phi} = -\frac{1}{2} \log(2\pi\, \hat{\tau}) - \frac{1}{2} \left(\frac{z_0 - \hat{\mu}}{\hat{\tau}} \right)^2$, $\frac{\hat{\phi}_{i_1}}{\hat{\phi}} = \hat{\tau}^{-1} \hat{\mu}_i \left(\frac{z_0 - \hat{\mu}}{\hat{\tau}} \right)$, $\frac{\hat{\phi}_{i_1 i_2}}{\hat{\phi}} = \hat{\tau}^{-2} \hat{\mu}_{i_1} \hat{\mu}_{i_2} \left\{ \left(\frac{z_0 - \hat{\mu}}{\hat{\tau}} \right)^2 - 1 \right\}$ and the subscripts at $\hat{\mu}$ denote differentiation with respect to the components of $\boldsymbol{z}$. Notice that the right hand side of (32) is a second degree polynomial in z_0, suggesting that the predictive density constructed by omitting the terms of order $O(kn^{-2})$ in the right hand side of (32) is normal. Consequently, we define the *second order corrected plug-in predictive density* by

$$\hat{f}(z_0|\boldsymbol{y}) = \exp\left\{ \log \hat{\phi} - \frac{1}{2} \frac{\hat{\phi}_{i_1}}{\hat{\phi}} \hat{h}_{i_2 i_2 i_2} \hat{h}^{i_1 i_2} \hat{h}^{i_2 i_2} + \frac{1}{2} \frac{\hat{\phi}_{i_1 i_2}}{\hat{\phi}} \hat{h}^{i_1 i_2} \right\} \tag{33}$$

while the first order Laplace approximation is $\hat{\phi}$, i.e. normal with mean $\hat{\mu}$ and variance $\hat{\tau}^2$. Notice that the coefficient of z_0^2 in the exponent of (33) is

$$-\frac{1}{2\hat{\tau}^2} \left(1 - \hat{\tau}^{-2} \mu_{i_1} \mu_{i_2} \hat{h}^{i_1 i_2} \right) \tag{34}$$

therefore, in order for (33) to be a proper density, (34) has to be negative. Since $\hat{h}_{\boldsymbol{zz}}$ is evaluated at $\hat{\boldsymbol{z}}$, it is positive definite, hence, so is $\hat{h}^{-1}_{\boldsymbol{zz}}$, therefore $\mu_{i_1} \mu_{i_2} \hat{h}^{i_1 i_2} > 0$ so the prediction variance using the higher order correction is bigger than the first order Laplace approximation. In fact, the first order Laplace approximation underestimates the variance which can be seen by writing

$$\begin{aligned} \operatorname{Var}(Z_0|\boldsymbol{Y}) &= \operatorname{E}(\operatorname{Var}(Z_0|\boldsymbol{Z})|\boldsymbol{Y}) + \operatorname{Var}(\operatorname{E}(Z_0|\boldsymbol{Z})|\boldsymbol{Y}) \\ &> \operatorname{E}(\operatorname{Var}(Z_0|\boldsymbol{Z})|\boldsymbol{Y}) = \tau^2. \end{aligned} \tag{35}$$

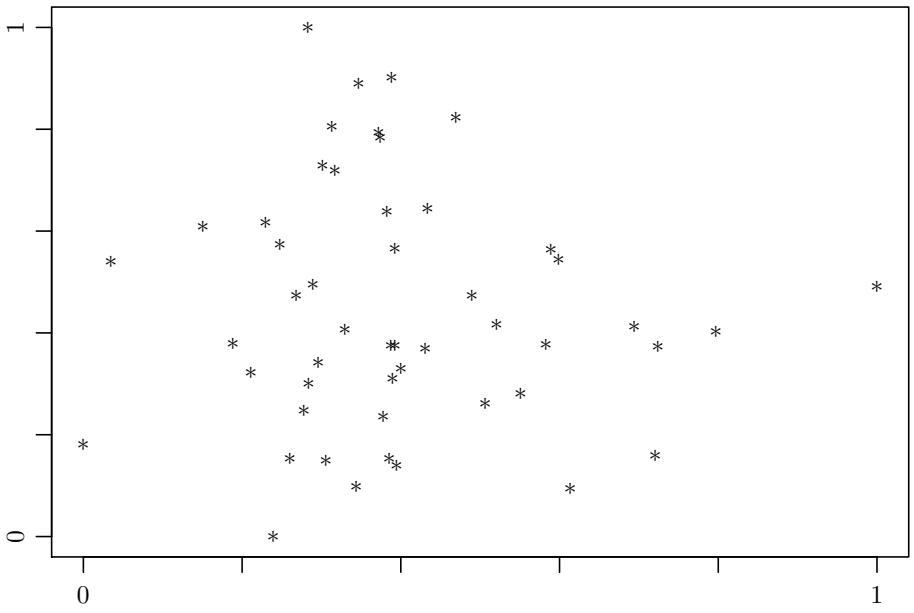

Fig. 2: Locations for the simulations

On the other hand, if $\hat{\tau}^{-2}\mu_{i_1}\mu_{i_2}\hat{h}^{i_1 i_2} > 1$, then (33) cannot be defined because (34) becomes positive. In this case, the approach can be modified as we explain below. Note though that since $\mu_{i_1}\mu_{i_2}\hat{h}^{i_1 i_2} = O(k\,n^{-1})$, then the coefficient of z_0^2 should be negative if the sample size is sufficiently large.

The variance of (33) is

$$\hat{\sigma}_c^2 = \hat{\tau}^2(1 - \hat{\tau}^{-2}\hat{\mu}_{i_1}\hat{\mu}_{i_2}\hat{h}^{i_1 i_2})^{-1} \tag{36}$$

and its mean is

$$\begin{aligned}\hat{\mu}_c &= \hat{\mu} - \frac{1}{2}(\hat{\mu}_{i_1}\hat{h}_{i_2 i_2 i_2}\hat{h}^{i_1 i_2}\hat{h}^{i_2 i_2})\,(1 - \hat{\tau}^{-2}\hat{\mu}_{i_1}\hat{\mu}_{i_2}\hat{h}^{i_1 i_2})^{-1} \\ &= \hat{\mu} - \frac{1}{2}(\hat{\mu}_{i_1}\hat{h}_{i_2 i_2 i_2}\hat{h}^{i_1 i_2}\hat{h}^{i_2 i_2})\frac{\hat{\sigma}_c^2}{\hat{\tau}^2}\end{aligned} \tag{37}$$

therefore, the α-quantile of the distribution of $Z_0|\{\boldsymbol{Y} = \boldsymbol{y}\}$ is estimated by $\hat{z}_\alpha = \hat{\mu}_c + \hat{\sigma}_c\Phi^{-1}(\alpha)$ where $\Phi^{-1}(\alpha)$ is the α-quantile of the standard Normal distribution. Additionally, by transforming the predictive quantile, the plug-in approach can be used to compute quantiles of monotone transformations of Z_0, such as $b'(Z_0)$ which corresponds to the probability of success and the rate in the binomial and Poisson cases respectively.

As we mentioned above, when $\hat{\tau}^{-2}\hat{\mu}_{i_1}\hat{\mu}_{i_2}\hat{h}^{i_1 i_2} > 1$, (33) is not a proper density. In this case we propose using $\hat{\sigma}_c^2 = \hat{\tau}^2(1 + \hat{\tau}^{-2}\hat{\mu}_{i_1}\hat{\mu}_{i_2}\hat{h}^{i_1 i_2})$, thus modifying (36) and (37) without changing the order of the approximation while making it positive. This can also be justified by observing that the recommended expression is approximately equal to (35), the true conditional variance, by noting that $\mathrm{Var}(\boldsymbol{Z}|\boldsymbol{Y}) \approx \hat{h}^{-1}_{\boldsymbol{zz}}$ (see Booth and Hobert, 1998).

6 Simulations

We perform simulations to compare the performance of the Laplace approximation against other methods. Our simulations consist of 500 realizations from the binomial spatial generalized linear mixed model under logit link from $k = 50$ locations (see Figure 2) selected within $[0, 1] \times [0, 1]$ and with $n = 100$ observations at each location. We used constant mean $\beta = -1.5$ and exponential covariance structure with parameters $\gamma = (0.10, 0.25, 0.10)$ corresponding to nugget, partial sill, and range.

		β	γ_1	γ_2	$\log(\gamma_3)$	Time (s)
trans	Bias	-0.02536	0.03122	0.05715	-0.44510	22
	SE	0.17039	0.11002	0.14633	0.96667	
	RMSE	0.17209	0.11426	0.15696	1.06334	
LA1	Bias	-0.01865	0.04988	-0.07109	-0.48934	21
	SE	0.16290	0.07298	0.09828	0.93504	
	RMSE	0.16380	0.08833	0.12122	1.05452	
LA2a	Bias	-0.00988	0.00661	-0.02579	-0.41685	21
	SE	0.15994	0.07116	0.10811	0.93232	
	RMSE	0.16008	0.07140	0.11104	1.02041	
LA2b	Bias	0.00645	-0.03040	0.01025	-0.37360	54
	SE	0.16276	0.07902	0.11835	0.89852	
	RMSE	0.16273	0.08459	0.11868	0.97227	
PQL	Bias	0.10562	1.53218	3.98771	-0.46565	179
	SE	0.16001	1.51238	2.09051	0.94923	
	RMSE	0.19159	2.15181	4.50148	1.05643	
SML	Bias	0.00607	-0.02806	0.01029	-0.40046	2039
	SE	0.16347	0.08293	0.12158	0.89982	
	RMSE	0.16342	0.08747	0.12189	0.98408	

Tab. 1: Comparison between different methods for estimation.

6.1 Estimation

First we investigate the performance of approximate likelihood estimation by comparing six methods:

- a transformation method which assumes that the logit transformation of the observed probabilities follows a normal distribution (trans),
- a first order Laplace approximation consisting of the first four terms of (25) (LA1),
- the second order Laplace approximation given by (25) (LA2a),
- the second order Laplace approximation given by (21) (LA2b),
- the penalized quasi likelihood (PQL) of Breslow and Clayton (1993),
- the simulated maximum likelihood method of Christensen (2004) implemented in the `R` package `geoRglm` (R Development Core Team, 2008; Christensen and Ribeiro, 2002) (SML). The burn-in size was 10000, and the subsequent iteration size 10000 with thinning 50.

Table 1 shows the average bias, the standard error (SE), and the root mean square error (RMSE) of the estimated mean and covariance parameters for each method. Because the distribution of the estimates of the range parameter, γ_3, is highly skewed, we chose to compare the estimates for its logarithm. Regarding the estimation of the mean parameter β, the SML has smaller bias, and second order approximation has smaller RMSE. For the covariance parameters, the second order approximation has the smallest mean square error and low bias. Between the two LA2 methods, we observe that the estimates for nugget and partial sill are quite different but the sums which correspond to the total variance at a given site are close. LA2a has smaller RMSE for the mean, nugget and partial sill parameters. It also appears that LA2b is closer in agreement to SML. Note however that LA2b is about three times slower than LA2a. The SML also has low bias but higher mean square error than the second order approximations and is about 100 times slower than the LA2a. The transformation and the first order approximation both have higher bias and mean square error and the PQL is very unreliable. This shows evidence of better performance for the second order approximation compared to the transformation and the first order approximation.

A known problem when estimating the covariance parameters in spatial models is the unidentifiability between the partial sill and range parameters if the data is highly correlated (Stein 1999). In another simulation study under the same setting but with range parameter being 0.8 instead of 0.1, we observed that

		β	γ_1	γ_2	$\log(\gamma_3)$
LA1	Bias	-1.85013	5.67380	8.35044	-7.23875
	SE	2.00346	4.93444	9.86196	3.7590
LA2a	Bias	0.02964	0.48570	0.24108	-7.52795
	SE	0.28454	1.98961	2.81356	3.2292
LA2b	Bias	-0.10171	0.91834	2.18871	-3.85622
	SE	0.38457	2.91127	5.22497	4.6177
SML	Bias	0.02821	-0.04686	-0.01232	-0.87623
	SE	0.26725	0.06776	0.15292	1.1358

Tab. 2: Comparison between different methods for estimation for binary data.

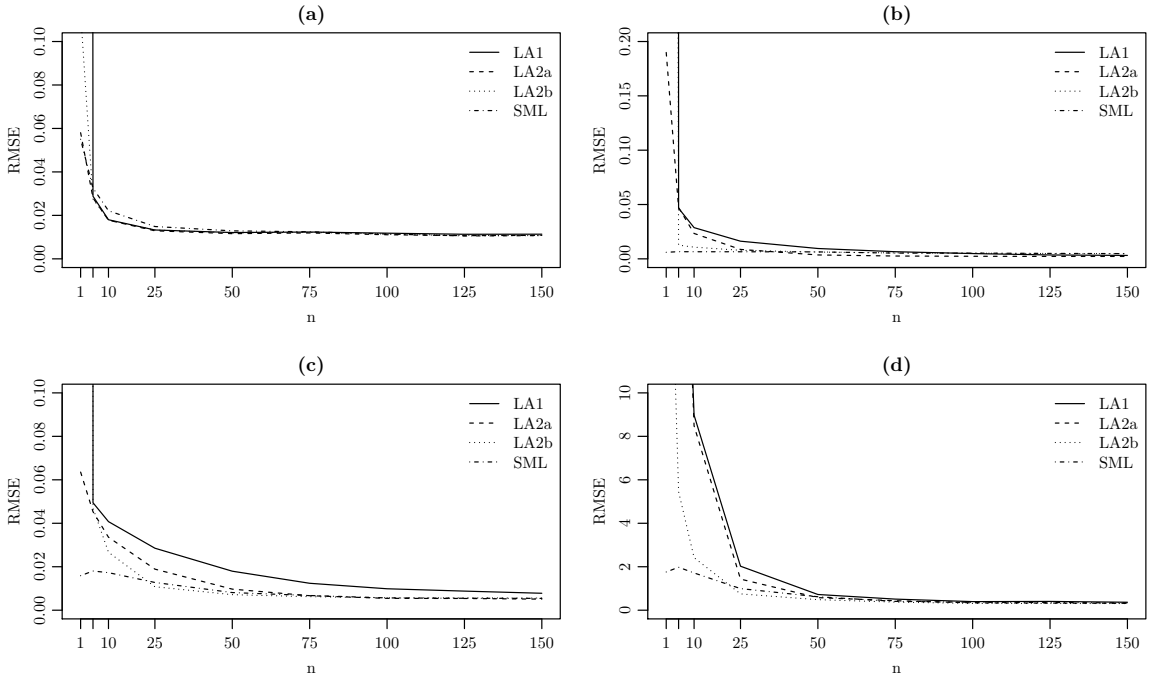

Fig. 3: RMSE for different methods as n varies for parameters (a) mean, (b) nugget, (c) partial sill, and (d) logarithm of range.

some of the estimates for the partial sill and range among the three methods were unreasonably large, which happened when the variability in the observations for that particular realisation was low. In the simulation study presented here the estimates obtained were within reasonable bounds. In applications where data is highly correlated, we suggest incorporating prior knowledge for the values of the covariance parameters and use a constrained optimization for the estimation.

An anonymous referee raised the question of the performance of the Laplace approximation when the assumption $k/n \to 0$ as $k \to \infty$ fails as in the case of binary data. We performed a number simulations under the same setting described in the beginning of this section but for different n each time. In particular for the case $n = 1$ we observe that the bias for the mean parameter β remains low but the variability increases. With respect to covariance parameters, both the bias and variability increase while SML was more accurate (see Table 2). In fact, for the case of binary data, SML seems to have better performance in estimating the covariance parameters even for larger k. As n increases the RMSE is reduced as shown in Figure 3. Overall, we observe that for n at least 50 the approximate likelihood produces reasonable estimates for the covariance parameters, however, even for n as low as 25 the reduction in the RMSE of the Laplace approximation is large.

6.2 Prediction

Using the same simulated data we compare the second order plug-in predictive density (LA 2) against the first order Laplace approximation and the Monte Carlo prediction as described in section 1.9.1 of Diggle et al. (2003) with burn-in 10000 and subsequent iteration size 10000 and thinning 50. The predictive density at 58 equally-spaced locations within the convex hull of the sampled locations was constructed using each method. For fairness and to reduce the variability in the measures we used for comparison, we assumed that the true parameter values are known. The total computation times for the first and second Laplace approximation were around 2 seconds while the MCMC needed around 4 minutes.

As a first measure of comparison we used three scoring rules that appeared in Gneiting and Raftery (2007). They are defined as follows: Let $U \in \{0, 1\}$ be an unobserved random variable and $p_j = \Pr(U = j)$, $j = 0, 1$. Suppose that a particular prediction method gives estimates $\hat{p}_j$ for p_j respectively. The following quantities give a measure of how close the estimated distribution is to the actual distribution of the predicted random variable. If the event $\{U = i\}$ is observed Gneiting and Raftery (2007) defined

- The negative Brier score: $2(1 - \hat{p}_i)^2$,
- the negative spherical score: $-\hat{p}_i(\hat{p}_0^2 + \hat{p}_1^2)^{-1/2}$,
- the negative logarithmic score: $-\log \hat{p}_i$

The lower the score, the better the performance.

Let $s(i, \hat{\boldsymbol{p}})$ denote the value of a particular score when the event $\{U = i\}$ is observed when the estimated distribution is $\hat{\boldsymbol{p}} = (\hat{p}_0, \hat{p}_1)$. The expected score is given by

$$\mathrm{E}s(U, \hat{\boldsymbol{p}}) = \mathrm{E}[p_0 s(0, \hat{\boldsymbol{p}}) + p_1 s(1, \hat{\boldsymbol{p}})]$$

where the expectation is over the joint distribution of $(Z_0, \boldsymbol{Y})$. In practice, since U is typically unobserved, the expected score provides an appropriate criterion.

In our simulations the variable U corresponds to one observation from an unsampled location with probability of "success" $p_1 = p_1(z_0) = e^{\beta + z_0}/(1 + e^{\beta + z_0})$ where z_0 is the value of the random field at that location. The true probability of success is replaced by its expectation with respect to the conditional distribution Z0—Z which is evaluated by one-dimensional Gaussian quadrature (see Demidenko, 2004, section 7.1.2). For computing the estimated probability, $\hat{p}_1$, we set $\hat{p}_1 = \int p_1(z_0)\tilde{f}(z_0|\boldsymbol{y})\,\mathrm{d}z_0$. For the first and second order Laplace approximation $\tilde{f}(z_0|\boldsymbol{y})$ is the normal density with parameters $(\hat{\mu}, \hat{\tau}^2)$ and $(\hat{\mu}_c, \hat{\sigma}_c^2)$ respectively and the integral is also evaluated by one-dimensional Gaussian quadrature. For the MCMC, a random sample from $Z_0|\boldsymbol{Y}$ is generated and the integral is evaluated by Monte-Carlo integration.

The total score over all locations create a measure for comparison between the three methods. In addition, for each simulation we compute the Mahalanobis distance with covariance matrix equal to the conditional variance of the random effects between the predicted random field and its conditional mean given the random field at the sampled locations. The logarithm of the Mahalanobis distance gives a measure of how close the prediction is from the one obtained under the assumption that the random field Z is observed. In this case, if the distance is smaller for one method, this method should be preferred. The averages of these measures along with their standard errors are shown in Table 3.

As we observe from Table 3, the averages over all measures is very small but the second order approximation has consistently better performance and in general gives more reliable predictions. Regarding the computational times, the approximation methods are again significantly faster than MCMC. For all 500 simulations the first and second order approximations needed about 2 seconds to finish while the MCMC took about 4.5 minutes.

As a second measure of comparison, we compared the coverage probabilities between the first and second order approximate predictive densities. Let $\hat{q}_\alpha$ denote the α-quantile of one of the approximate predictive densities obtained by inverting the approximate predictive distribution function for data $\boldsymbol{y}$. For the first order approximation $\hat{q}_\alpha = \hat{\mu} + \hat{\tau}\Phi^{-1}(\alpha)$. For the second order approximation, $\hat{q}_\alpha = \hat{\mu}_c + \hat{\sigma}_c\Phi^{-1}(\alpha)$, where $\Phi^{-1}(\alpha)$ stands for the α-quantile of the standard normal distribution. The coverage probability is defined as the probability $\Pr(Z_0 \leq \hat{q}_\alpha|\boldsymbol{Y} = \boldsymbol{y})$. This probability is computed empirically using a random observation from $Z_0|\boldsymbol{Z}$ and considering the proportion of times that $z_0 < \hat{q}_\alpha$ over simulations. If this probability is close to α then the prediction method is good.

	Brier	Spherical	Logarithmic	Mahalanobis	Time (s)
LA1	18.23452	-47.992734	28.56126	0.69074	2
	(1.2259)	(0.74217)	(1.4308)	(0.3317)	
LA2	18.23427	-47.992841	28.56084	0.67931	2
	(1.2263)	(0.74236)	(1.4316)	(0.3222)	
MCMC	18.23801	-47.991124	28.56663	0.82818	266
	(1.2268)	(0.74264)	(1.4320)	(0.2875)	

Tab. 3: Comparison of different methods for prediction. Showing the average of the four measures used for comparison along with their standard errors in parentheses.

Quantile	LA1	LA2	Quantile	LA1	LA2
0.01	0.01179	0.00979	0.8	0.80386	0.80121
	(0.00412)	(0.00413)		(0.01936)	(0.01956)
0.025	0.02852	0.02441	0.9	0.90010	0.90028
	(0.00657)	(0.00588)		(0.01404)	(0.01421)
0.05	0.05624	0.05000	0.95	0.94855	0.94941
	(0.01195)	(0.01019)		(0.01044)	(0.01011)
0.1	0.11128	0.10038	0.975	0.97317	0.97362
	(0.01707)	(0.01609)		(0.00774)	(0.00766)
0.2	0.21448	0.20241	0.99	0.98890	0.98941
	(0.01947)	(0.01895)		(0.00465)	(0.00453)
0.5	0.51238	0.50186			
	(0.02369)	(0.02347)			

Tab. 4: Average coverage of predictive quantiles obtained from first and second order Laplace approximation to the predictive density and its standard error in parenthesis.

The average coverage probability for each method over all locations along with its standard error is computed for a range of $0 < \alpha < 1$ and is shown in Table 4. From the table we see that the coverage of the second order Laplace approximation is overall closer to the target coverage than in the first order Laplace approximation.

In conclusion, we argue that the second order Laplace approximation should be preferred over other approximations and, when there are time constraints, it provides a fast alternative over MCMC.

7 Analysis of the rhizoctonia disease data

The rhizoctonia root rot is a disease that attaches on the roots of plants and hinders their process of absorbing water and nutrients. In this example examined by Zhang (2002), 15 plants were pulled from each of 100 randomly chosen locations in a farm and the number of crown roots and infected crown roots were counted. Similar to Zhang we assume constant mean and spherical covariance structure for the underlying Gaussian random field and treat the data as samples from binomial distribution. Because $\gamma_j > 0$, in our optimization we estimate the $\log \gamma_j$ and then exponentiate. The estimates using each method are summarized in Table 5. The standard errors were obtained by inverting the approximate Hessian matrix. We observe that there is an agreement between the different methods, in particular between the second order Laplace approximation and the SML; however, SML is much slower.

It is important to know the severity of the disease at every location in the field in order to be able to allocate treatment efficiently. Prediction at 3177 equally spaced locations within the convex hull of the sampled locations was performed using the first and second order Laplace approximation and the MCMC method. The predictions of the random field using the second order Laplace approximation are shown in Figure 4 and agree closely with those of Zhang (2002) although the range of our predictions is smaller.

We compare the same three methods considered for prediction by cross-validation. One location was removed each time and the remaining 99 locations were used to estimate the parameters and predict the

	β	nugget	partial sill	range	Time (s)
trans	-1.7601	0.6360	0.1254	151.2	0.2
	(0.1066)	(0.1492)	(0.1405)	(50.72)	
LA1	-1.7248	0.4901	0.0818	149.1	0.2
	(0.0945)	(0.1204)	(0.1019)	(53.63)	
LA2a	-1.7188	0.4723	0.1021	148.6	0.2
	(0.0970)	(0.1266)	(0.1146)	(47.02)	
LA2b	-1.7185	0.4681	0.1065	148.8	0.6
	(0.0977)	(0.1275)	(0.1169)	(45.49)	
SML [a]	-1.7187	0.4716	0.1048	148.3	30
MCEMG [b]	-1.6152	0.3451	0.1754	145.11	
	(0.0023)	(0.0898)	(0.1086)	(73.33)	

Tab. 5: Estimates of the parameters of the rhizoctonia example using different methods. [a]Standard errors not provided by software. [b]Quoted from Zhang (2002).

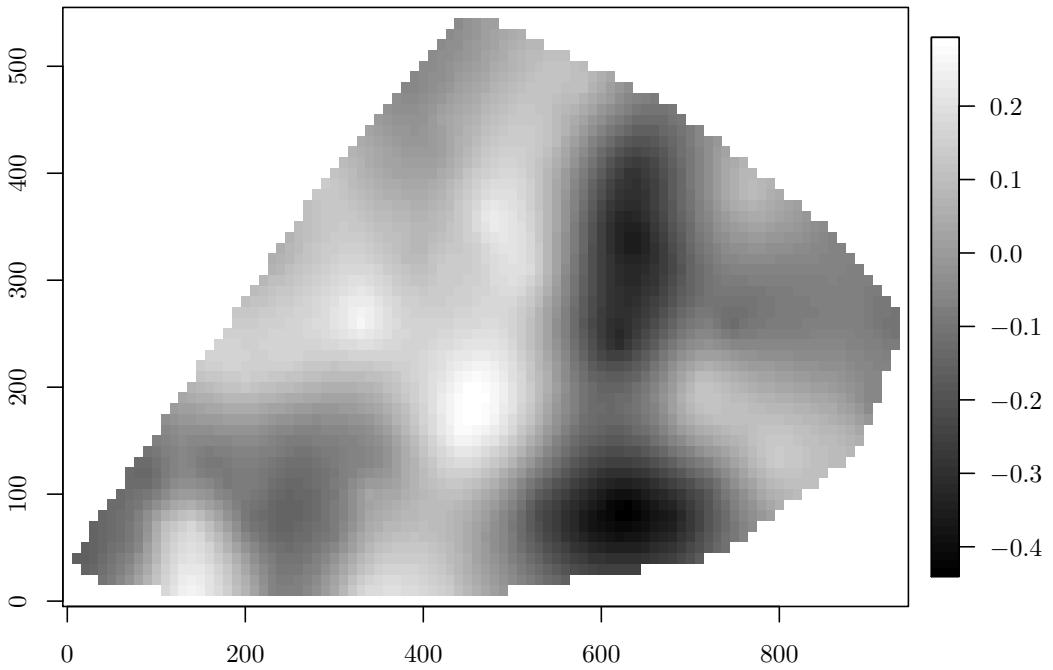

Fig. 4: Map of the predicted random effects (disease severity) using the second order plug-in predictive density.

random effect at the removed location. The predicted random effect was used to estimate the probability of an infected root and compute the same three scoring rules used in the simulations. In the end, for each of the 100 locations we have a score for the prediction using each method. The histograms of these scores show that their distribution over locations is non-symmetric with no significant differences. This shows consistency between the approximate methods and the MCMC. The total computation time for the two approximation methods was about the same while the MCMC method was about 40 times slower.

8 Summary

In this paper we demonstrated the use of Laplace approximation for estimation and prediction of spatial GLMM when the sample size is large. We found that the approximation becomes more accurate when higher order terms are included in the asymptotic expansion and we were able to achieve good accuracy in less computational time.

For estimating the parameters of the model, an approximate likelihood method was proposed. Primarily this consists of applying the Laplace approximation to the log-likelihood and maximizing this approximation. Our simulations showed that, in comparison with other approximations, this approach gives better results.

Regarding the prediction of the random field, we derived a normal approximation to the conditional density of the random effect at a certain location given the observations. The advantage of this approximation is that the prediction intervals can be computed from the quantiles of the normal distribution. Our simulations showed that this method gives very good accuracy and is very fast.

Finally, we note that although these methods are applied to spatial data, the results can be generalized to any type of clustered data where there exists one random effect in each cluster. In addition, the asymptotic expansions derived are general enough to be applied to other types of models where integral approximation is required.

Acknowledgments

We are grateful to an anonymous referee for helpful comments. This research was partially supported by NSF DMS grant 0605434.

References

Barndorff-Nielsen, O. E. and Cox, D. R. (1989). *Asymptotic Techniques for Use in Statistics*. Chapman & Hall Ltd.

Booth, J. G. and Hobert, J. P. (1998). Standard errors of prediction in generalized linear mixed models. *Journal of the American Statistical Association*, 93:262–272.

Booth, J. G. and Hobert, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 61:265–285.

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

Breslow, N. E. and Lin, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika*, 82:81–91.

Byrd, R., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal for Scientific Computing*, 16(5):1190–1208.

Christensen, O. F. (2004). Monte Carlo maximum likelihood in model-based geostatistics. *Journal of Computational and Graphical Statistics*, 13(3):702–718.

Christensen, O. F., Møller, J., and Waagepetersen, R. (2000). Analysis of spatial data using generalized linear mixed models and Langevin-type markov chain monte carlo. Technical report, Department of Mathematical Sciences, Aalborg University.

Christensen, O. F. and Ribeiro, P. J. (2002). GeoRglm: A package for generalised linear spatial models. *R News*, 2(2):26–28.

Christensen, O. F. and Waagepetersen, R. (2002). Bayesian prediction of spatial count data using generalized linear mixed models. *Biometrics*, 58(2):280–286.

Crainiceanu, C. M., Diggle, P. J., and Rowlingson, B. (2008). Bivariate Binomial Spatial Modeling of Loa loa Prevalence in Tropical Africa. *Journal of the American Statistical Association*, 103(481):21–37.

Cressie, N. A. (1993). *Statistics for Spatial Data*. John Wiley & Sons.

Demidenko, E. (2004). *Mixed Models: Theory and Applications*. Wiley-Interscience.

Diggle, P., Knorr-Held, L., Rowlingson, B., Su, T., Hawtin, P., and Bryant, T. (2004). On-line monitoring of public health surveillance data. In Brookmeyer, R. and Stroup, D., editors, *Monitoring the Health of Populations: Statistical Methods for Public Health Surveillance*, pages 233–266. Oxford University Press.

Diggle, P. J., Moyeed, R. A., and Tawn, J. A. (1998). Model-based geostatistics. *Journal of the Royal Statistical Society, Series C: Applied Statistics*, 47(3):299–326.

Diggle, P. J., Ribeiro Jr, J. P., and Christensen, O. F. (2003). An Introduction to Model-Based Geostatistics. In Møller, J., editor, *Spatial Statistics and Computational Methods*, volume 173 of *Lecture notes in statistics*, pages 43–86. Springer-Verlag, New York.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall Ltd.

Eidsvik, J., Martino, S., and Rue, H. (2009). Approximate Bayesian inference in spatial generalized linear mixed models. *Scandinavian journal of statistics*, 36(1):1–22.

Gneiting, T. and Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Kuk, A. Y. C. (1999). Laplace importance sampling for generalized linear mixed models. *Journal of Statistical Computation and Simulation*, 63:143–158.

Lin, X. and Breslow, N. E. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association*, 91:1007–1016.

McCullagh, P. (1987). *Tensor Methods in Statistics*. Chapman & Hall Ltd.

McCullagh, P. and Nelder, J. A. (1999). *Generalized Linear Models*. Chapman & Hall Ltd.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, 92:162–170.

Noh, M. and Lee, Y. (2007). REML estimation for binary data in GLMMs. *Journal of Multivariate Analysis*, 98(5):896–915.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Raudenbush, S. W., Yang, M.-L., and Yosef, M. (2000). Maximum likelihood for generalized linear models with nested random effects via high-order, multivariate Laplace approximation. *Journal of Computational and Graphical Statistics*, 9(1):141–157.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society, Series B, Methodological*, 71:1–35.

Shun, Z. (1997). Another look at the Salamander mating data: A modified Laplace approximation approach. *Journal of the American Statistical Association*, 92:341–349.

Shun, Z. and McCullagh, P. (1995). Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society, Series B, Methodological*, 57:749–760.

Solomon, P. J. and Cox, D. R. (1992). Nonlinear component of variance models. *Biometrika*, 79:1–11.

Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer-Verlag Inc.

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

Vidoni, P. (2006). Response prediction in mixed effects models. *Journal of Statistical Planning and Inference*, 136(11):3948–3966.

Wolfinger, R. and Lin, X. (1997). Two taylor-series approximation methods for nonlinear mixed models. *Computational Statistics and Data Analysis*, 25:465–490.

Zhang, H. (2002). On estimation and prediction for spatial generalized linear mixed models. *Biometrics*, 58(1):129–136.