Web Working Papers
by

The Italian Group of Environmental Statistics

Gruppo di Ricerca per le Applicazione della Statistica
ai Problemi Ambientali

www.graspa.org

Pairwise likelihood inference in spatial generalized linear mixed models

Cristiano Varin, Gudmund Høst,. Øivind Skare

GRASPA Working paper n.19, Novembre 2003

Pairwise likelihood inference in spatial generalized linear mixed models

Cristiano Varin*[1], Gudmund Høst[2] and Øivind Skare[3]

[1]Department of Statistics, Padova (Italy)
[2]Norwegian Computing Center, Oslo (Norway)
[3]Institute of Biology, Oslo (Norway)

October 15, 2003

Abstract

Spatial generalized linear mixed models are flexible models for a variety of applications, where we have observations of spatially dependent and non-Gaussian random variables. Our focus is inference in spatial generalized linear mixed models for large data sets. Maximum likelihood or Bayesian Markov chain Monte Carlo approaches may in such cases be computationally very slow or even prohibitive. Alternatively, one may consider a composite likelihood, which is the product of likelihoods of subsets of data. Here, we use a composite likelihood based on pairs of observations. In order to maximize the pairwise likelihood, we introduce a new Expectation-Maximization type algorithm which uses numerical quadrature. We illustrate the method on simulated data and on data from air pollution effects for fish populations in Norwegian lakes. A comparison with alternative methods is given. We find that the proposed algorithm gives reasonable parameter estimates and that it is computationally efficient.

Keywords: composite likelihood, Expectation-Maximization algorithm, Gauss-Hermite quadrature, model-based geostatistics, window sub-sampling.

*Correspondence to: Cristiano Varin, Department of Statistics, via Cesare Battisti, 241, 35121 Padova, Italy. e-mail: sammy@stat.unipd.it

1 Introduction

We present methodology for computationally efficient parameter estimation in generalized linear mixed models (GLMMs) for spatial data. This class of models has found applications to a wide range of problems within spatial statistics. Modern spatial data sets, for example those collected by remote sensing or automatic sensors, can be very large. Inference for large data sets requires repeated high-dimensional integration and matrix inversion, which may be restrictive even for powerful computers. Our approach is based on a composite (or pseudo-) likelihood which reduces the high-dimensional integral to a sum of low dimensional integrals. These low dimensional integrals can be efficiently computed by numerical quadrature. The parameters are estimated iteratively by an Expectation-Maximization (EM) type of algorithm.

Spatial GLMMs are flexible models for a variety of applications where we have observations of spatially dependent and non-Gaussian random variables. Such applications include problems within epidemiology, ecology, agriculture and remote sensing.The spatial GLMM was described by Diggle, Tawn & Moyeed (1998). Here, the underlying random effects were modeled by a Gaussian random field (GRF). As in standard GLMM (Breslow & Clayton 1993), given the random effects, the observations at the measurement locations are conditionally independent and follow a generalized linear model.

Both Bayesian and frequentist methods have been developed for inference and forecasting in spatial GLMMs. Diggle et al. (1998) used a Bayesian Markov chain Monte Carlo (MCMC) framework with priors on the unknown regression parameters and the covariance parameters of the Gaussian random field. The computational burden increases with the number of observations, because the number of correlated random effects to be simulated is equal to the number of observations. A more efficient Langevin-Hastings MCMC algorithm was given by Christensen & Waagepetersen (2002).

Maximum likelihood estimation in spatial GLMMs generally involves numerical integration of a high dimensional integral. The integral may be computed by Monte Carlo integration. McCulloch (1997) reviews several Monte Carlo techniques for maximum likelihood estimation within GLMMs. Booth & Hobert (1999) describes a Monte Carlo EM (MCEM) algorithm for the spatial probit model, while Zhang (2002) used a maximum likelihood approach together with an MCEM algorithm to estimate parameters of a general spatial GLMM. Alternatively, the integral may be computed by

Laplace's method which uses a Gaussian approximation of the integrand. In GLMM inference, this method has been used by Breslow & Clayton (1993) and Skaug (2002).

Both Bayesian MCMC and maximum likelihood EM inference involves high-dimensional matrices that have to be inverted repeatedly. Convergence will be slower and more iterations will be needed as the dimension increases. Thus, none of these approaches are practical for large data sets.

To gain in computational efficiency, one may approximate the GRF random effects model and do inference under the approximate model. This was done by Rue & Tjelmeland (2002), who approximated the GRF by a Gaussian Markov random field. This allows for fast calculations drawing on methods for Markov fields (Rue 2001). Another approach of this type is to cast the model in the form of a tree structure. This allows for fast spatial prediction by using the methods of Huang, Cressie & Gabrosek (2002).

An alternative to model approximation, which will be followed in this paper, is to approximate the objective function. Instead of the likelihood, we consider a pairwise likelihood, which is the product of likelihoods for pairs of data and estimate parameters by maximizing this product. This reduces the computational effort from order N^3 to order N^2 operations. In practice, it is not necessary to use all possible pairs of observations, but rather a subset of neighboring pairs. This allows for further reduction in computational effort.

Pairwise likelihood is a special case of a more general class of pseudo likelihoods called composite likelihood (Lindsay 1988). Applications to correlated data include random set models in image analysis (Nott & Rydén 1999), correlated binary data (Kuk & Nott 2000), multivariate survival data analysis (Parner 2001), multilevel models (Renard, Molenberghs & Geys 2003) and frailty models for longitudinal data (Henderson & Shimakura 2003). Applications to Gaussian spatial data has been described by Hjort & Omre (1994). Heagerty & Lele (1998) used a pairwise likelihood approach to analyze binary spatial data in a spatial probit model, where the involved two-dimensional integrals could be expressed in closed form. In the more general situation to be considered here, these integrals require the use of two-dimensional numerical integration. This can be done efficiently by numerical quadrature techniques. Our proposed algorithm is a computationally efficient alternative to the MCEM algorithm, tuned to situations with many random effects.

The paper is organized as follows. In Section 2 we introduce notation for the spatial GLMM and define pairwise likelihood. In Section 3 we describe the algorithm and some theoretical properties, while the implemetation is

described in Section 4. Finally, in Sections 5 and 6, our approach is assessed through simulation studies and an application to fish data from Norwegian lakes.

2 Pairwise likelihood inference

Let $S \subseteq \mathbb{R}^d$ be some region of interest and denote by $\boldsymbol{s}$ a particular location within S. We define the spatial GLMM.

(i) Denote by $\{u(\boldsymbol{s}) : \boldsymbol{s} \in S\}$ a stationary GRF with zero mean and spatial covariance function $\mathrm{Cov}\big(u(\boldsymbol{s}), u(\boldsymbol{s}')\big) = \sigma^2 \rho(\boldsymbol{s} - \boldsymbol{s}'; \boldsymbol{\alpha})$. Here, $\rho(\cdot; \boldsymbol{\alpha})$ is a positive definite function and $\boldsymbol{\alpha}$ is a vector of correlation parameters.

(ii) Given $\boldsymbol{u} = \big(u(\boldsymbol{s}_1), \ldots, u(\boldsymbol{s}_n)\big)^{\mathrm{T}}$; $\boldsymbol{s_i} \in S$; $i = 1, \ldots, n$, the observations $\boldsymbol{y} = \big(y(\boldsymbol{s_1}), \ldots, y(\boldsymbol{s_n})\big)^{\mathrm{T}}$ are mutually independent.

(iii) The conditional mean of an observation at $\boldsymbol{s}$ is $\mathrm{E}\big(y(\boldsymbol{s})|u(\boldsymbol{s})\big)$ $= g^{-1}\big(\eta(\boldsymbol{s})\big)$, where $g(\cdot)$ is a differentiable and invertible link function with domain $\mathbb{R}$, $\eta(\boldsymbol{s}) = \boldsymbol{x}^{\mathrm{T}}(\boldsymbol{s})\boldsymbol{\beta} + u(\boldsymbol{s})$, $\boldsymbol{x}(\boldsymbol{s})$ is a p dimensional vector of known covariates and $\boldsymbol{\beta}$ is a vector of p unknown regression parameters.

(iv) Given a dispersal parameter ϕ, the conditional density of an observation $y_i = y(\boldsymbol{s}_i)$ given $u_i = u(\boldsymbol{s}_i)$, $i = 1, \ldots, n$, belongs to the exponential class

$$f(y_i|u_i; \boldsymbol{\beta}, \phi) = \exp\Big[\frac{1}{\phi}\big\{a(\mu_i)y_i - b(\mu_i)\big\}\Big] c\Big(\frac{1}{\phi}, y_i\Big).$$

Here, $\mu_i = \mathrm{E}(y_i|u_i)$ while $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are specific functions, see McCullagh & Nelder (1989). If $a(\cdot) \equiv g(\cdot)$ we have a canonical link function.

The likelihood of the GLMM is

$$L(\boldsymbol{\psi}; \boldsymbol{y}) \propto \int \cdots \int \prod_{i=1}^{n} f(y_i|u_i; \boldsymbol{\beta}, \phi)\, f(\boldsymbol{u}; \sigma^2, \boldsymbol{\alpha}) \mathrm{d}\boldsymbol{u}, \tag{1}$$

where $\boldsymbol{\psi} = (\boldsymbol{\beta}, \phi, \sigma^2, \boldsymbol{\alpha})$. Generally, the n-dimensional integral (1) cannot be factorized into low dimensional terms as is common in a non-spatial GLMM (McCulloch & Searle 2001). Unless the conditional density of y_i is Gaussian the integral will have to be evaluated by computational methods. We mitigate this problem through the use of pairwise likelihood.

Definition 1 (Pairwise likelihood). Let $\{L(\boldsymbol{\psi}; \boldsymbol{y}), \boldsymbol{\psi} \in \boldsymbol{\Psi}, \boldsymbol{y} \in \boldsymbol{Y}\}$ be a statistical model. The *pairwise likelihood* is the product of the bivariate likelihoods

$$\mathrm{PL}(\boldsymbol{\psi}; \boldsymbol{y}) = \prod_{(i,j)\in\mathcal{R}} \mathrm{L}(\boldsymbol{\psi}; y_i, y_j), \tag{2}$$

where $\mathcal{R}$ is a subset of all possible pairwise neighbors. The element in $\boldsymbol{\Psi}$ which maximizes the pairwise likelihood is $\widehat{\boldsymbol{\psi}}_{MPL}$, the *maximum pairwise likelihood* (MPL) estimator.

The pairwise likelihood in spatial GLMMs is a product of form (2) of bivariate likelihoods from (1). This gives the product of double integrals

$$\mathrm{PL}(\boldsymbol{\psi}; \boldsymbol{y}) \propto \prod_{(i,j)\in\mathcal{R}} \iint f(y_i|u_i; \boldsymbol{\beta}, \phi) f(y_j|u_j; \boldsymbol{\beta}, \phi) f(u_i, u_j; \sigma^2, \boldsymbol{\alpha}) \mathrm{d}u_i \mathrm{d}u_j.$$

Conditions for consistency and asymptotic normality of the MPL estimator for observations on a spatial grid were given by Heagerty & Lele (1998) and Heagerty & Lumley (2000).

3 EM for pairwise likelihood

The EM algorithm is a method for function maximization which alternates an expectation step and a maximization step. It is popular for likelihood inference and may also be used for pairwise likelihood.

Definition 2 (Pairwise EM). Choose a starting value ${}_0\boldsymbol{\psi}$ such that $\mathrm{PL}({}_0\boldsymbol{\psi}; \boldsymbol{y}) > 0$ and set $d = 0$. The *pairwise EM* (PEM) algorithm iterates the following steps until convergence.

(i) **Expectation step:** evaluate the sum of the conditional expectations

$$Q(\boldsymbol{\psi}|{}_d\boldsymbol{\psi}) = \sum_{(i,j)\in\mathcal{R}} \iint \log\{f(u_i, u_j, y_i, y_j; \boldsymbol{\psi})\} f(u_i, u_j|y_i, y_j; {}_d\boldsymbol{\psi}) \mathrm{d}u_i \mathrm{d}u_j. \tag{3}$$

(ii) **Maximization step:** choose ${}_{d+1}\boldsymbol{\psi}$ such that ${}_{d+1}\boldsymbol{\psi} = \arg\max_{\boldsymbol{\psi}} Q(\boldsymbol{\psi}|{}_d\boldsymbol{\psi})$.

(iii) Set $d = d + 1$.

PEM has similar properties as EM for full likelihood. The basic property of EM type algorithms is the *ascent property*, which says that for each iteration of the algorithm the likelihood will not decrease. This property also applies to PEM.

Proposition 1 (Ascent property). *Let* ${}_0\boldsymbol{\psi}, {}_1\boldsymbol{\psi}, {}_2\boldsymbol{\psi}, \ldots$ *be the sequence of iterates of PEM, then the pairwise likelihood do not decrease at each iteration of the PEM algorithm,* i.e. $\mathrm{PL}({}_d\boldsymbol{\psi}; \boldsymbol{y}) \leq \mathrm{PL}({}_{d+1}\boldsymbol{\psi}; \boldsymbol{y})$.

Proof. See Appendix A. □

PEM produces a monotonous sequence. Therefore, convergence properties of EM (Wu 1983) applies to PEM with suitable changes in notation.

It is not necessary that ${}_d\boldsymbol{\psi}$ maximizes $Q(\boldsymbol{\psi}|{}_d\boldsymbol{\psi})$ for the convergence of PEM. Indeed, the ascent property is still satisfied if ${}_{d+1}\boldsymbol{\psi}$ is chosen such that

$$Q({}_{d+1}\boldsymbol{\psi}|{}_d\boldsymbol{\psi}) \geq Q({}_d\boldsymbol{\psi}|{}_d\boldsymbol{\psi}). \tag{4}$$

Algorithms where the maximization step is substituted with (4) are called *generalized* EM algorithms (McLachlan & Krishnan 1997).

If the expectation can not be expressed in closed form it must be approximated numerically. We define an approximate version of EM by substituting Q by an approximation $\widehat{Q}$.

Definition 3 (Approximate pairwise EM algorithm). Choose a starting value ${}_0\boldsymbol{\psi}$ such that $\mathrm{PL}({}_0\boldsymbol{\psi}; \boldsymbol{y}) > 0$ and set $d = 0$. The *approximate pairwise EM* algorithm iterates the following steps until convergence.

(i) **Approximate expectation step:** approximate the expectation step (3) in PEM by $\widehat{Q}(\boldsymbol{\psi}; {}_d\boldsymbol{\psi})$.

(ii) **Generalized maximization step:** Choose ${}_{d+1}\boldsymbol{\psi}$ such that $\widehat{Q}({}_{d+1}\boldsymbol{\psi}|{}_d\boldsymbol{\psi}) \geq \widehat{Q}({}_d\boldsymbol{\psi}|{}_d\boldsymbol{\psi})$.

(iii) Set $d = d + 1$.

Here, the maximization step has been substituted by condition (4), because $\widehat{Q}(\boldsymbol{\psi}; {}_d\boldsymbol{\psi})$ cannot be maximized analytically in spatial GLMMs.

Booth & Hobert (1999) and McCulloch (1997) used Monte Carlo integration in the expectation step for GLMMs. In pairwise likelihood maximization the expectation step is a sum of double integrals. Double integrals are more

efficiently evaluated by Gauss-Hermite quadrature than by Monte Carlo integration. Thus, we suggest to use quadrature in the approximate expectation step. Our resulting quadrature pairwise EM (QPEM) algorithm is described in detail in Section 4.

Typically, QPEM converges to a stationary point. To ensure that this stationary point is a local maximum it is advisable to rerun the algorithm with perturbed starting values. Generally, assessment of convergence is simpler for QPEM than for MCEM, because QPEM is a deterministic algorithm.

4 Implementation of QPEM

The estimation step of QPEM involves approximating the sum of double integrals in (3) by Gauss-Hermite quadrature. Gauss-Hermite quadrature reduces the integral of a function with respect to a given kernel to a weighted sum of the integrand evaluated at M specific nodes. Details are given in Appendix B. The resulting approximation of $Q(\boldsymbol{\psi}; {}_d\boldsymbol{\psi})$ is

$$\widehat{Q}(\boldsymbol{\psi}; {}_d\boldsymbol{\psi}) = \sum_{(i,j)\in\mathcal{R}} \sum_{m=1}^{M} \log f(\boldsymbol{u}_{(i,j)}(m), y_i, y_j; \boldsymbol{\psi}) w\big(\boldsymbol{u}_{(i,j)}(m); {}_d\boldsymbol{\psi}\big). \qquad (5)$$

The bivariate nodes $\boldsymbol{u}_{(i,j)}(m) = (u_i, u_j)^{\mathrm{T}}(m)$ and the weights $w\big(\boldsymbol{u}_{(i,j)}(m); {}_d\boldsymbol{\psi}\big)$, $m = 1, \ldots, M$ are given in Appendix B. Now, $\widehat{Q}(\boldsymbol{\psi}; {}_d\boldsymbol{\psi})$ may be decomposed into

$$\widehat{Q}(\boldsymbol{\psi}; {}_d\boldsymbol{\psi}) = \widehat{Q}(\boldsymbol{\beta}, \phi; {}_d\boldsymbol{\psi}) + \widehat{Q}(\sigma^2, \boldsymbol{\alpha}; {}_d\boldsymbol{\psi}). \qquad (6)$$

Here, the two functions on the right hand side of (6) are defined

$$\widehat{Q}(\boldsymbol{\beta}, \phi; {}_d\boldsymbol{\psi}) = \sum_{(i,j)\in\mathcal{R}} \sum_{m=1}^{M} \log f\big(y_i, y_j | \boldsymbol{u}_{(i,j)}(m); \boldsymbol{\beta}, \phi\big) w\big(\boldsymbol{u}_{(i,j)}(m); {}_d\boldsymbol{\psi}\big),$$

and

$$\widehat{Q}(\sigma^2, \boldsymbol{\alpha}; {}_d\boldsymbol{\psi}) = \sum_{(i,j)\in\mathcal{R}} \sum_{m=1}^{M} \log f\big(\boldsymbol{u}_{(i,j)}(m); \sigma^2, \boldsymbol{\alpha}\big) w\big(\boldsymbol{u}_{(i,j)}(m); {}_d\boldsymbol{\psi}\big).$$

The advantage of the decomposition above is that the two terms may be maximized separately. The first term, $\widehat{Q}(\boldsymbol{\beta}, \phi; {}_d\boldsymbol{\psi})$, is the usual GLM term,

involving only the fixed effects. The second term, $\widehat{Q}(\sigma^2, \boldsymbol{\alpha}; {}_d\boldsymbol{\psi})$, involves only the random effects. The fixed effects term may be maximized by iterative weighted least squares as described in McCullagh & Nelder (1989).

The random effects term $\widehat{Q}(\sigma^2\boldsymbol{\alpha}; {}_d\boldsymbol{\psi})$ is a weighted sum of log-bivariate Gaussian densities with zero mean and covariance matrix

$$\sigma^2 \begin{pmatrix} 1 & \rho_{(i,j)}(\boldsymbol{\alpha}) \\ \rho_{(i,j)}(\boldsymbol{\alpha}) & 1 \end{pmatrix},$$

where $\rho_{(i,j)}(\boldsymbol{\alpha}) = \rho(\boldsymbol{s}_i - \boldsymbol{s}_j; \boldsymbol{\alpha})$.

The random effects term may also be maximized by Newton-Raphson. However, experimentation led us to use the more robust Nelder-Mead downhill simplex algorithm (Nelder & Mead 1965).

4.1 Practical issues

Reasonable starting values are important for fast convergence to the correct mode of the pairwise likelihood. Furthermore, we need to select a covariance function for the spatial random effects field. Our procedure for choosing realistic starting values and an appropriate covariance function is as follows.

First, neglect the random effects and estimate the regression parameters $\boldsymbol{\beta}$ under a fixed effects model. Next, transform the observations by the link function and fit empirical residuals *i.e.* $\widehat{r}(\boldsymbol{s}_i) = g(y_i) - \boldsymbol{x}_i^{\mathrm{T}}{}_0\widehat{\boldsymbol{\beta}}$, $i = 1, \ldots, n$. Now, the empirical variogram (Cressie 1993) for the residuals $\widehat{r}(\boldsymbol{s}_i)$; $i = 1, \ldots, n$ may be calculated. A plausible covariance function is fitted to the empirical variogram and starting values for the covariance parameters ${}_0\sigma^2$ and ${}_0\boldsymbol{\alpha}$ are estimated by least squares.

Some care is needed if the model includes Poisson data with log link or binary data with logit link. In the first case, a remedy is to add a small number to each observation before transformation. In the second case, we may aggregate the observations over spatial subregions and use the mean frequencies of the aggregated data.

Different pairs of observed data give different contributions to the pairwise likelihood product. Nott & Rydén (1999) observe that only distinct pairs that show significant spatial dependence need to be included in the product. They use a moving neighborhood and a fixed design mask to select pairs within the neighborhood. They also discuss various choices of design masks and weighting of pairs in the product.

In many spatial applications, the data are not regularly spaced, and the above ideas are not straightforward to implement. In practice, we have found that random sampling of pairs within a moving neighborhood works well. Using a moving window excludes pairs far apart that have with little spatial correlation, while the random sampling of pairs gives a reasonable coverage of the neighborhood.

4.2 Variance of parameter estimates

Variance estimates for maximum likelihood parameter estimates are often based on the information matrix. For pairwise likelihood we suggest corresponding variance estimates based on *pairwise information*. Write $pl(\boldsymbol{\psi}; \boldsymbol{y}) = \log \mathrm{PL}(\boldsymbol{\psi}; \boldsymbol{y})$, then a Taylor series argument, (Heagerty & Lele 1998, Nott & Rydén 1999), shows that the asymptotic variance of the MPL estimator is the inverse of the pairwise information

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\psi}) = \mathrm{E}_{\boldsymbol{y}}\big\{\nabla^2 pl(\boldsymbol{\psi}; \boldsymbol{y})\big\}\,\mathrm{Var}_{\boldsymbol{y}}^{-1}\big\{\nabla pl(\boldsymbol{\psi}; \boldsymbol{y})\big\}\,\mathrm{E}_{\boldsymbol{y}}\big\{\nabla^2 pl(\boldsymbol{\psi}; \boldsymbol{y})\big\}^{\mathrm{T}}.$$

To estimate $\mathcal{I}(\boldsymbol{\psi})$, we need the gradient and the Hessian of each bivariate log likelihood $\log f(y_i, y_j; \boldsymbol{\psi})$. These are given by the following proposition.

Proposition 2 (Derivatives of the pairwise likelihood). *Consider a pair of locations $(\boldsymbol{s}_i, \boldsymbol{s}_j)$ such that $f(y_i, y_j; \boldsymbol{\psi}) > 0$ and assume that the order of integration and differentiation may be inter-changed. Then, the gradient and the Hessian matrix of $\log f\big(y_i, y_j; \boldsymbol{\psi}\big)$ are given by*

$$\nabla \log f\big(y_i, y_j; \boldsymbol{\psi}\big) = \mathrm{E}_{\boldsymbol{u}|\boldsymbol{y}}\, \nabla \log f(u_i, u_j, y_i, y_j; \boldsymbol{\psi}),$$

and

$$\begin{aligned}\nabla^2 \log f(y_i, y_j; \boldsymbol{\psi}) = \mathrm{E}_{\boldsymbol{u}|\boldsymbol{y}}\, \nabla^2 \log f(u_i, u_j, y_i, y_j; \boldsymbol{\psi}) \\ + \mathrm{Var}_{\boldsymbol{u}|\boldsymbol{y}}\, \nabla \log f(u_i, u_j, y_i, y_j; \boldsymbol{\psi}).\end{aligned}$$

Here, $\mathrm{E}_{\boldsymbol{u}|\boldsymbol{y}}$ and $\mathrm{Var}_{\boldsymbol{u}|\boldsymbol{y}}$ are the expectation and the variance operators with respect to the conditional density $f(u_i, u_j | y_i, y_j; \boldsymbol{\psi})$.

Proof. The proof of the proposition is obtained by straightforward differentiation using the approach in Louis (1982) to pairs of indices (i, j). □

We see that the gradient and Hessian are byproducts of the expectation step of the QPEM algorithm. The mean $\mathrm{E}_{\boldsymbol{y}}\{\nabla^2 pl(\boldsymbol{\psi};\boldsymbol{y})\}$ and the variance $\mathrm{Var}_{\boldsymbol{y}}\{\nabla pl(\boldsymbol{\psi};\boldsymbol{y})\}$ are expectations with respect to the unknown true density of $\boldsymbol{y}$. These moments may be estimated using *window resampling*, as described by Heagerty & Lele (1998). In window resampling, we subdivide the study region into overlapping spatial windows and compute empirical estimates of $\mathrm{E}_{\boldsymbol{y}}\{\nabla^2 pl(\widehat{\boldsymbol{\psi}}_{MPL};\boldsymbol{y})\}$ and $\mathrm{Var}_{\boldsymbol{y}}\{\nabla pl(\widehat{\boldsymbol{\psi}}_{MPL};\boldsymbol{y})\}$ for each window. The final estimate is obtained by averaging window estimates with weights proportional to the area of the respective windows. Some theoretical considerations and a proof of consistency for window subsampling variance estimators for general estimating functions in spatial models are given in Heagerty & Lumley (2000).

5 Simulated data examples

We illustrate the use of our method on a spatial GLMM with Poisson errors, log-link and $\eta(\boldsymbol{s}) = \beta_0 + \beta_1 s_1 + u(\boldsymbol{s})$. We use a random effects field $u(\boldsymbol{s})$ with zero mean and spatial covariance function

$$\mathrm{Cov}\big(u(\boldsymbol{s}'), u(\boldsymbol{s}'')\big) = \sigma^2 \exp\big(-3\|\boldsymbol{s}' - \boldsymbol{s}''\|/\alpha\big).$$

Here, the constant 3 is introduced in accordance with geostatistics literature, giving negligible covariance for $|\boldsymbol{s}' - \boldsymbol{s}''| > \alpha$. In the first example, parameters were fixed at $(\beta_0, \beta_1, \sigma^2, \alpha) = (-2.0, 0.1, 1.5, 6.0)$. We simulated $n = 25 \times 25$ data on a regular grid of locations. A realization of simulated data from the model is shown in Figure 1. The large proportion of zero values is due to β_0 being negative in the present example.

For each observation we use a neighborhood of radius 4. Constructing pairs using all 48 neighbors gives $48n = 30,000$ pairs, neglecting border effects. This compares to a total number of possible pairs $n(n-1)/2 = 195,000$.

Now, parameters were fitted by QPEM with $M = 4 \times 4$ Gauss Hermite quadrature with starting values as described in Section 4. The maximization with respect to the mixed effects parameters was constrained using logit-like transformations in order to avoid singularities. The parameter σ^2 was constrained to the interval $(0.1, 5.0)$ and α to the interval $(0.1, 10.0)$. We used relative difference $\max_i |{}_{d+1}\boldsymbol{\psi}_i - {}_d\boldsymbol{\psi}_i|/|{}_d\boldsymbol{\psi}_i|$ as convergence criterion.

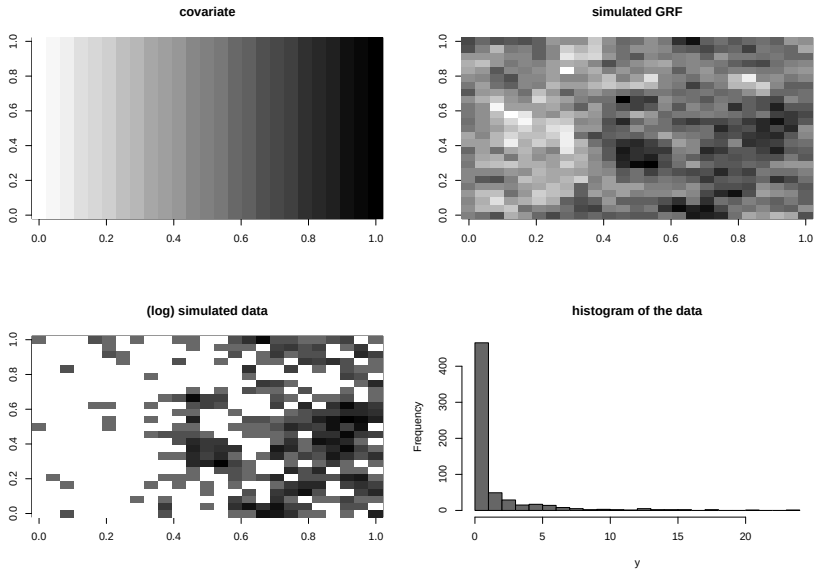

Figure 1: Realization from the spatial Poisson model. From top left to bottom right: covariate $x(\boldsymbol{s}) = \boldsymbol{s}_1$, simulated mixed effects field $u(\boldsymbol{s})$, (log) simulated data, histogram of the data.

Alternatively, the algorithm was stopped when $Q(\boldsymbol{\psi}; {}_d\boldsymbol{\psi})$ stopped increasing at the M-step.

Our version of the QPEM algorithm was implemented in C++ and run on 100 data sets simulated from the model. Figure 2 shows parameter values as function of iteration number for one simulated data set. The algorithm converged in less than 40 iterations for 80 of the data sets. For 10 data sets 40-100 iterations was needed, while 9 data sets needed more than 100 iterations of QPEM to converge. Cases with slow convergence were usually characterized by large values of $\widehat{\sigma^2}$.

The results are summarized in Table 1. We see that there is good correspondence between true and estimated values. The largest bias occurred for the range parameter α, which is typically difficult to estimate with any method.

The computational time may be reduced by thinning the number of pairs used within each moving neighborhood, as suggested in Section 4. We illustrate this by subsampling $r = 15$ random locations without replacement

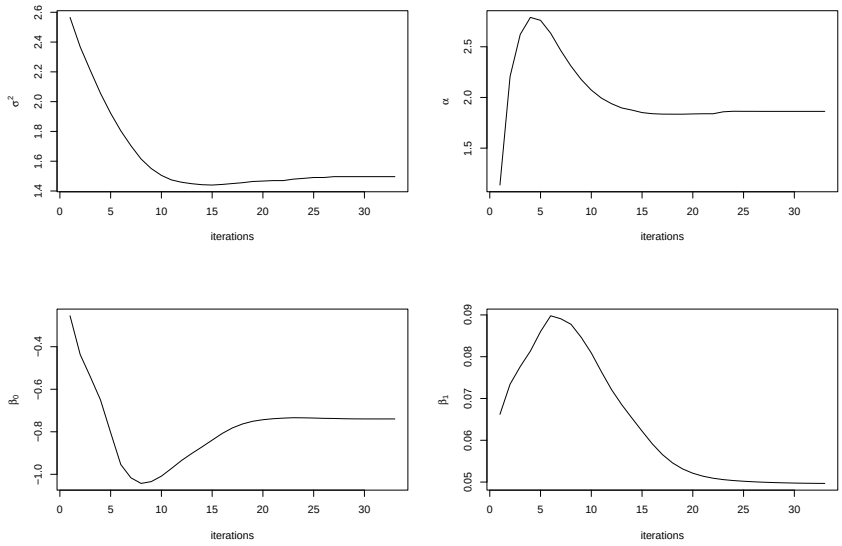

Figure 2: QPEM iterates for a typical data set. From top left to bottom right: σ^2, α, β_0 and β_1.

within a neighborhood of radius 4, as shown in Figure 3. Running through all data locations we obtained about $15n = 9,375$ pairs. The computing time for this exercise was about 30% of the computing time needed for the previous example, i.e. proportional to the reduction in the number of pairs used.

The results from the thinning exercise are summarized in Table 2. Again, there is good correspondence between true and estimated values. We see that very little information is lost by random subsampling of pairs in this model example.

We also ran the algorithm on the same data sets varying the neighborhood radius, the number of sampled pairs and the number of quadrature nodes. Increasing the number of quadrature nodes to 5×5 points gave very similar results, while decreasing to 3×3 was considerably worse. Furthermore, increasing the neighborhood radius and the number of sampled pairs did not have much effect on the estimates.

Next, we compare our pairwise likelihood approach with maximum likelihood (ML) estimation. The results from this study comparison will depend

Table 1: Results from QPEM estimation on 100 simulated data sets from the spatial Poisson model.

Parameter	True	Estimate (mean)	Bias	SD	MSE
σ^2	1.5	1.40	-7.14%	0.40	0.1672
α	6.0	5.33	-12.6%	0.68	0.9123
β_0	-2.0	-2.008	-0.4%	0.68	0.4618
β_1	0.1	0.105	4.8%	0.05	0.0024

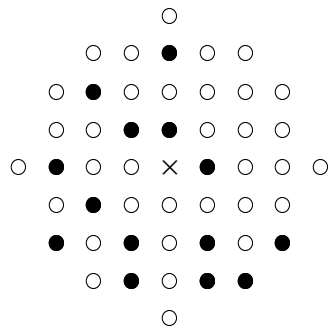

Figure 3: Sampling pairs within a neighborhood of radius 4. Here, x is the observation location and the filled circles are 15 neighbors sampled at random with replacement. The contributing pairs consist of x and each of the 15 sampled neighbors.

on the actual implementation of QPEM and ML, but may give an indication of the difference between the methods.

The computationally demanding tasks are integrating out the random effects and inverting the covariance matrix. Here, we compute the likelihood using the full covariance matrix and Laplace's approximation for integration (Shun & McCullogh 1995). A general implementation of Laplace's approximation for random effects models is available in the software package AD Model Builder (http://otter-rsch.com/admodel.htm). An advantage of this package is that derivatives are calculated automatically, see Skaug (2002).

Again, we simulated 100 data sets from a spatial Poisson model. Since inverting the covariance matrix in the likelihood is computationally demanding, we used a smaller spatial grid of only $n = 12 \times 12$ observations in this exercise. Parameters were fixed at $\beta_0 = -2.0$, $\beta_1 = 0.3$, $\sigma^2 = 1.0$ and $\alpha = 3.0$.

Table 2: Results from QPEM estimation with random subsampling of pairs on the simulated Poisson data.

Parameter	True	Estimate (mean)	Bias	SD	MSE
σ^2	1.5	1.44	-4.2%	0.44	0.1997
α	6.0	5.50	-9.1%	0.74	0.8001
β_0	-2.0	-2.026	-1.3%	0.66	0.4314
β_1	0.1	0.107	6.5%	0.05	0.0025

We ran QPEM using all neighboring pairs within a radius of 2.5 and with $M = 5 \times 5$ quadrature nodes. For both ML and QPEM, we used the true parameter values as starting values.

A summary of the results is given in Table 3. We see that the mean squared error for the estimated parameters σ^2, β_0 and β_1 are somewhat smaller for ML than for QPEM. For the range parameter α, ML estimation gives much larger bias and variance than QPEM. The present choice of β_0 and β_1 typically gives simulated data with very skewed likelihood. Laplace approximation is based on a Gaussian approximation to the likelihood which may not work well in this situation.

Table 3: Comparison between ADMB and QPEM for Poisson data on a regular 12x12 lattice. Corr is the correlation between the ADMB and the QPEM estimators.

		ADMB		QPEM		
Param	True	Mean	MSE	Mean	MSE	Corr
σ^2	1.0	0.93	0.059	1.04	0.185	0.74
α	3.0	4.72	10.717	2.83	1.670	-0.46
β_0	-2.0	-1.97	0.257	-2.23	0.808	0.87
β_1	0.3	0.29	0.003	0.33	0.013	0.84

The computing time for QPEM on a typical dataset using a 550 MHz Pentium III with 4GB RAM was 69 seconds. The ML estimates for the same data set was computed in 756 seconds. In this case, QPEM used 21MB of computer memory, while ML used 385 MB. In a different example, increasing the number of simulated observations with 30% increased QPEM memory use by 30% and ML memory use by 140%.

6 Acidification data example

Acid deposition from long range transportation of air pollutants has been of major concern in Norway for several decades. These pollutants contribute to the acidification of lakes and streams, which may kill fish populations. In particular, the trout populations are sensitive to acidification.

We use data on population status of trout from 542 lakes in Norway. The data were collected during 1986 from interviews with local fishermen. For each lake, the population status is coded as unaffected (0) or decreased/extinct (1). In addition to spatial location of each lake, we use the measured acid neutralizing capacity (ANC) as a covariate. ANC reflects local properties of geology and soils as well as the load from current and historic acid deposition. Our aim is to make spatial prediction of trout population status.

Figure 4 shows the observed population status. Here, green circles mark lakes unaffected by acidification, while red circles mark affected lakes. We see that trout in the southern and western parts of Norway are most affected by acidification.

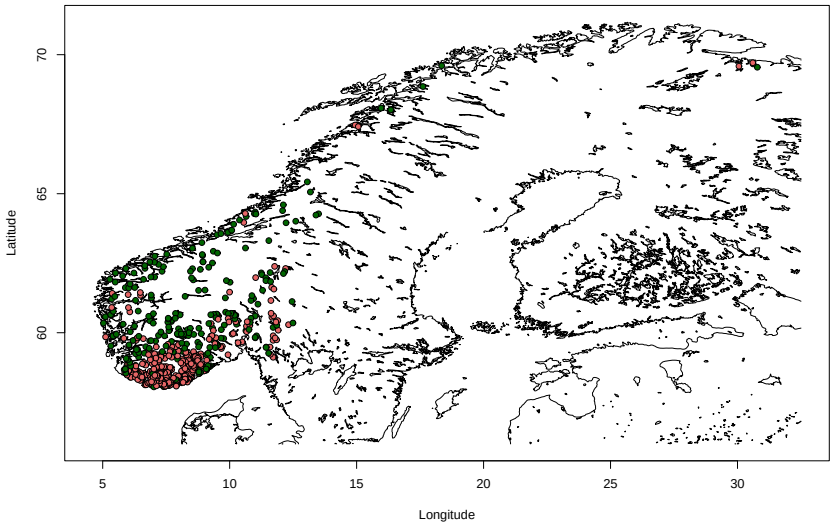

Figure 4: Trout data. Green circles denotes lakes where trout are not affected by acidification, red circles lakes where the trout population has decreased or is extinct.

For the purpose of this study, we randomly sampled 400 observations and reserved the remaining 142 observations for model validation. We use a model with Bernoulli data, log-link and two regression parameters, i.e. $\eta(\boldsymbol{s}) = \beta_0 + \beta_1 \mathrm{ANC}(\boldsymbol{s}) + u(\boldsymbol{s})$. To obtain starting values, we used the procedure described in Section 4. The 400 sample observations were fitted to a fixed effects model by using the package glm implemented in the R language (Ihaka & Gentleman 1996). The parameter estimates obtained by R glm were ${}_0\beta_0 = 0.222$ and ${}_0\beta_1 = -0.120$. As suggested in Section 4, we calculated an empirical variogram of transformed residuals. The shape of the variogram suggested an exponential covariance function, and a least squares fit provided starting values ${}_0\sigma^2 = 2.248$ and ${}_0\alpha = 171.8$ Km.

Data pairs for the QPEM algorithm were obtained by using $r = 20$ locations sampled from a neighborhood of radius 250 km around each of the 400 data locations. We used 4×4 points Gauss-Hermite quadrature and a tolerance of 10^{-4} for convergence. The algorithm converged after 14 iterations to the values $\widehat{\sigma}^2 = 2.217$, $\widehat{\alpha} = 188.2$ Km, $\widehat{\beta}_0 = 0.246$, $\widehat{\beta}_1 = -0.161$. Convergence to a local maximum was checked by re-running the algorithm with perturbed starting values. Standard deviations for the parameters were estimated by window subsampling, as described in Section 4. For each of the 400 locations, a window of 250 km and maximum 50 data locations were used. The estimated standard deviations were $(0.098, 26.0 \text{ Km}, 0.531, 0.036)$. We also estimated parameters by QPEM using 6×6 quadrature points and obtained similar estimates.

For validation, we predict the observations at the 142 validation locations using the QPEM estimates and compare with the original values. Following Zhang (2002), the best predictor of the random effect $u(\boldsymbol{s}_0)$ at some location $\boldsymbol{s}_0$ in the validation set is

$$\widehat{u}(\boldsymbol{s}_0) = \mathrm{E}\{u(\boldsymbol{s}_0)|\boldsymbol{y}\} = \sum_{i=1}^{400} c_i(\boldsymbol{s}_0)\, \mathrm{E}\{u(\boldsymbol{s}_i)|\boldsymbol{y}\}$$

where $\boldsymbol{y} = (y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_{400}))^{\mathrm{T}}$ are the data used for estimation, while $c_1(\boldsymbol{s}_0)$, $\ldots$, $c_{400}(\boldsymbol{s}_0)$ are ordinary kriging weights (Cressie 1993). The conditional means $\mathrm{E}\{u(\boldsymbol{s}_1)|\boldsymbol{y}\}, \ldots, \mathrm{E}\{u(\boldsymbol{s}_{400})|\boldsymbol{y}\}$ are predicted by MCMC through a single-component Metropolis-Hastings algorithm as suggested by Zhang (2002). Finally, writing $\widehat{\eta}(\boldsymbol{s}_0) = \widehat{\beta}_0 + \widehat{\beta}_1 x(\boldsymbol{s}_0) + \widehat{u}(\boldsymbol{s}_0)$, the predictor for $y(\boldsymbol{s}_0)$

was obtained by mean of the threshold

$$\widehat{y}(\boldsymbol{s}_0) \begin{cases} 1, & \text{if } \exp\{\widehat{\eta}(\boldsymbol{s}_0)\}/[1+\exp\{\widehat{\eta}(\boldsymbol{s}_0)\}] > 0.5, \\ 0, & \text{otherwise.} \end{cases}$$

This procedure was repeated for each location in the validation set. The predicted population status was correct at 129 of the 142 locations, i.e. about 90% of the cases.

We estimate parameter estimation by ML for this data set to need several hours of CPU time on our 550 MHz Pentium III PC. However, ML estimation was impossible due to the memory requirements of approximately 8 GB.

7 Discussion

The computational savings from using the QPEM algorithm are great compared to likelihood inference. In particular, this is important for large data sets, because the computing time for QPEM increases as a linear function of the number of observations, while likelihood inference is cubic in the number of observations. The pairwise likelihood function seems to capture much of the information in the data, and this function may be efficiently maximized by combining numerical quadrature and EM. The computational speed of QPEM may be further increased through subsampling of pairs.

Since likelihood inference is computationally constrained by $O(n^3)$, it is difficult to compare inference from ML and QPEM on large sets of data. For a moderately large simulated data set, QPEM gives parameter estimates that are comparable with ML, as measured by bias and variance. QPEM also seemed to give reasonable estimates for a data set of size 400 on fish health in Norwegian lakes.

For larger data sets where ML is impractical or impossible, QPEM may be a promising method for inference. It may be particularly useful in data mining of massive spatial data sets, such as those derived from remote sensing. In contrast, QPEM may also be a practical tool for finding starting values for maximum likelihood inference in data sets of moderate size.

The optimal tuning of our pairwise EM algorithm involves several topics for further research. One such topic is subsampling of pairs for the pairwise likelihood product. Another topic is related to the integration of the random effects. Alternative quadrature methods or methods based on Laplace approximation may be advantageous in various situations.

QPEM is applicable to a wide class of spatial models, because there is no restriction on the structure of the mixed effects covariance matrix. In particular, QPEM is also applicable to non-spatial mixed models. A future research topic of interest would be extension to non-Gaussian mixed effects models. Finally, pairwise likelihood methods may also have applications to model selection, and some work is in progress in this direction (Varin & Vidoni 2003).

Acknowledgements

The authors would like to thank Hans J. Skaug and Dave Fournier who made modifications to AD Model Builder software to allow for the model comparison study. Norwegian Institute of Water research kindly provided the data for the fish population example.

A Proof of the proposition 1

Choose a starting value ${}_0\boldsymbol{\psi}$ and write $pl(\boldsymbol{\psi};\boldsymbol{y}) = \log \mathrm{PL}(\boldsymbol{\psi};\boldsymbol{y})$. We have

$$\begin{aligned}
pl(\boldsymbol{\psi};\boldsymbol{y}) &= \sum_{(i,j)\in\mathcal{R}} \log L(\boldsymbol{\psi}; y_i, y_j) \\
&= \sum_{(i,j)\in\mathcal{R}} \log L(\boldsymbol{\psi}; y_i, y_j) \iint f(u_i, u_j | y_i, y_j; {}_d\boldsymbol{\psi}) \mathrm{d}u_i \mathrm{d}u_j \\
&= \sum_{(i,j)\in\mathcal{R}} \iint \log\big\{ f(u_i, u_j, y_i, y_j; \boldsymbol{\psi}) \big\} f(u_i, u_j | y_i, y_j; {}_d\boldsymbol{\psi}) \mathrm{d}u_i \mathrm{d}u_j \\
&\quad - \sum_{(i,j)\in\mathcal{R}} \iint \log\big\{ f(u_i, u_j | y_i, y_j; \boldsymbol{\psi}) \big\} f(u_i, u_j | y_i, y_j; {}_d\boldsymbol{\psi}) \mathrm{d}u_i \mathrm{d}u_j \\
&= \sum_{(i,j)\in\mathcal{R}} Q_{(i,j)}(\boldsymbol{\psi}|{}_d\boldsymbol{\psi}) - \sum_{(i,j)\in\mathcal{R}} H_{(i,j)}(\boldsymbol{\psi}|{}_d\boldsymbol{\psi}) \\
&= Q(\boldsymbol{\psi}|{}_d\boldsymbol{\psi}) - H(\boldsymbol{\psi}|{}_d\boldsymbol{\psi}).
\end{aligned}$$

Thus, the difference between log pairwise likelihoods in subsequent iterations is

$$pl({}_{d+1}\boldsymbol{\psi};\boldsymbol{y}) - pl({}_d\boldsymbol{\psi};\boldsymbol{y}) = Q({}_{d+1}\boldsymbol{\psi}|{}_d\boldsymbol{\psi}) - Q({}_d\boldsymbol{\psi}|{}_d\boldsymbol{\psi}) + \sum_{(i,j)\in\mathcal{R}} D_{(i,j)}({}_{d+1}\boldsymbol{\psi}|{}_d\boldsymbol{\psi}),$$

Here, $D_{(i,j)}$ is the Kullbach-Leibler distance between the bivariate densities $f(u_i, u_j|y_i, y_j; {}_{d+1}\boldsymbol{\psi})$ and $f(u_i, u_j|y_i, y_j; {}_{d}\boldsymbol{\psi})$

$$D_{(i,j)}({}_{d+1}\boldsymbol{\psi}|{}_{d}\boldsymbol{\psi}) = -\iint \log\left\{\frac{f(u_i, u_j|y_i, y_j; {}_{d+1}\boldsymbol{\psi})}{f(u_i, u_j|y_i, y_j; {}_{d}\boldsymbol{\psi})}\right\} f(u_i, u_j|y_i, y_j; {}_{d}\boldsymbol{\psi})\mathrm{d}u_i\mathrm{d}u_j.$$

Since the Kullback-Leibler distance is non-negative, then $pl({}_{d+1}\boldsymbol{\psi}; \boldsymbol{y}) - pl({}_{d}\boldsymbol{\psi}; \boldsymbol{y})$ is non-negative and the map induced by PEM into the parametric space is non-decreasing. □

B E step: Gauss-Hermite quadrature

Gauss-Hermite quadrature is designed to approximate integrals involving distributions close to the normal distribution. The integrand $f(\mathbf{t})$ is split in a Gaussian part (the envelope) and a remaining part, which after transformation of variables will be of the form $e^{-\|\mathbf{t}\|^2/2} f(\mathbf{t})$. Gauss-Hermite quadrature reduces each 1D integral to a weighted sum of $f(\mathbf{t})$ evaluated at M specific nodes. If $f(\mathbf{t})$ is well approximated by a polynomial of low order, then the integral may be accurately computed using a small M. A Gaussian approximation of $f(u_i, u_j|y_i, y_j; {}_{d}\boldsymbol{\psi})$ could give acceptable accuracy with $M = 1$. However, to compute the approximation, we need to compute the mode and the second derivatives of $f(u_i, u_j|y_i, y_j; {}_{d}\boldsymbol{\psi})$. Therefore, we chose instead the distribution $f(u_i, u_j)$ as envelope. The remaining part involves the likelihoods $f(y_i|u_i)$, and the choice of M has to be tuned to the actual application.

Write $Q(\boldsymbol{\psi}; {}_{d}\boldsymbol{\psi}) = \sum_{(i,j)} Q_{(i,j)}(\boldsymbol{\psi}; {}_{d}\boldsymbol{\psi})$. Each $Q_{(i,j)}(\boldsymbol{\psi}; {}_{d}\boldsymbol{\psi})$ is a ratio of two double integrals. The numerator of $Q_{(i,j)}(\boldsymbol{\psi}; {}_{d}\boldsymbol{\psi})$ is given by

$$\iint \log\{f(u_i, u_j, y_i, y_j; \boldsymbol{\psi})\} f(y_i|u_i; {}_{d}\boldsymbol{\psi}) f(y_j|u_j; {}_{d}\boldsymbol{\psi}) f(u_i, u_j; {}_{d}\boldsymbol{\psi})\mathrm{d}u_i\mathrm{d}u_j, \quad (7)$$

while the denominator is

$$\iint f(y_i|u_i; {}_{d}\boldsymbol{\psi}) f(y_j|u_j; {}_{d}\boldsymbol{\psi}) f(u_i, u_j; {}_{d}\boldsymbol{\psi})\mathrm{d}u_i\mathrm{d}u_j. \quad (8)$$

In order to approximate these integrals by Gauss-Hermite quadrature, we transform the normal vector $(u_i, u_j)^{\mathrm{T}}$ into independent standardized components $(v_i, v_j)^{\mathrm{T}}$. This gives

$$\begin{cases} v_i = \frac{u_i}{\sigma}, \\ v_j = \frac{u_j - \rho_{(i,j)} u_i}{\sigma(1 - \rho^2_{(i,j)})^{1/2}}, \end{cases} \quad (9)$$

where $\rho_{(i,j)} = \rho(\boldsymbol{s}_i - \boldsymbol{s}_j; \boldsymbol{\alpha})$. By solving for $u_i(v_i)$ and $u_j(v_i, v_j)$ in (9), the denominator (8) becomes

$$\frac{1}{2\pi} \iint f(y_i|u_i(v_i); {}_d\boldsymbol{\psi}) f(y_j|u_j(v_i, v_j); {}_d\boldsymbol{\psi}) e^{-v_i^2/2} e^{-v_j^2/2} \mathrm{d}v_i \mathrm{d}v_j. \qquad (10)$$

Now, (10) can be approximated by Gauss-Hermite quadrature

$$\frac{1}{2\pi} \sum_{m_1=1}^{M_1} \sum_{m_2=1}^{M_2} f\big\{y_i|u_i\big(h(m_1)\big); {}_d\boldsymbol{\psi}\big\} f\big\{y_j|u_j\big(h(m_1), h(m_2)\big); {}_d\boldsymbol{\psi}\big\} k(m_1)k(m_2). \qquad (11)$$

Here, $h(m)$ are the quadrature nodes and $k(m)$ are the quadrature weights.

The quadrature formula for the numerator (7) is derived by a similar procedure. The denominator depends on ${}_d\boldsymbol{\psi}$ and on the data, but not on $\boldsymbol{\psi}$. Therefore, its contribution is only to change the weights of the quadrature formula for the numerator. The final Gauss-Hermite quadrature formula for $Q_{(i,j)}(\boldsymbol{\psi}; {}_d\boldsymbol{\psi})$ becomes

$$\sum_{m_1, m_2} \log f\big\{u_i\big(h(m_1)\big), u_j\big(h(m_1), h(m_2)\big), y_i, y_j; \boldsymbol{\psi}\big\} w_{(i,j)}(m_1, m_2; {}_d\boldsymbol{\psi})$$

where the new weights are

$$w_{(i,j)}(m_1, m_2; {}_d\boldsymbol{\psi}) = \frac{f\big\{y_i|u_i\big(h(m_1)\big); {}_d\boldsymbol{\psi}\big\} f\big\{y_j|u_j\big(h(m_1), h(m_2)\big); {}_d\boldsymbol{\psi}\big\} k(m_1)k(m_2)}{\sum_{m_1, m_2} f\big\{y_i|u_i\big(h(m_1)\big); {}_d\boldsymbol{\psi}\big\} f\big\{y_j|u_j\big(h(m_1), h(m_2)\big); {}_d\boldsymbol{\psi}\big\} k(m_1)k(m_2)}.$$

Note that the above weights corresponds to that used in the (5) with

$$w\big(\boldsymbol{u}_{(i,j)}(m); {}_d\boldsymbol{\psi}\big) = w_{(i,j)}(m_1, m_2; {}_d\boldsymbol{\psi}),$$

where the double indices (m_1, m_2) have been aggregated to simplify the notation.

References

Booth, J. G. & 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* **61**, 265 – 285.

Breslow, N. E. & Clayton, D. G. (1993), 'Approximate inference in generalized linear mixed models', *Journal of the American Statistical Association* **88**(421), 9 – 25.

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

Cressie, N. (1993), *Statistics for Spatial Data*, 2nd edn, John Wiley and Sons, inc.

Diggle, P. J., Tawn, J. A. & Moyeed, R. A. (1998), 'Model-based Geostatistics (with discussion)', *Journal of the Royal Statistical Society, Series B* **47**(2), 299–350.

Heagerty, P. J. & Lele, S. R. (1998), 'A composite likelihood approach to binary spatial data', *Journal of the American Statistical Association* **93**, 1099–1111.

Heagerty, P. J. & Lumley, T. (2000), 'Window subsampling of estimating functions with application to regression models', *Journal of the American Statistical Association* **95**, 197–211.

Henderson, R. & Shimakura, S. (2003), 'A serially correlated gamma frailty model for longitudinal count data', *Biometrika* **90**(2), 355–366.

Hjort, N. L. & Omre, H. (1994), 'Topics in spatial statistics', *The Scandinavian Journal of Statistics* **21**, 289–357.

Huang, H.-C., Cressie, N. & Gabrosek, J. (2002), 'Fast resolution consistent spatial prediction of global processes from satellite data', *Journal of Computational and Graphical Statistics* **11**(1), 63–88.

Ihaka, R. & Gentleman, R. (1996), 'R: A language for data analysis and graphics', *Journal of Computational and Graphical Statistics* **5**(3), 299–314.

Kuk, A. Y. & Nott, D. (2000), 'A pairwise likelihood approach to analyzing correlated binary data', *Statistics & Probability Letters* **47**, 329–335.

Lindsay, B. (1988), Composite likelihood methods, *in* N. U. Prabhu, ed., 'Statistical Inference from Stochastic Processes', Providence RI: American Mathematical Society.

Louis, T. A. (1982), 'Finding the observed information matrix when using the EM algorithm', *Journal of the Royal Statistical Society, Series B* **44**(2), 226–233.

McCullagh, P. & Nelder, J. A. (1989), *Generalized Linear Models*, 2nd edn, Chapman and Hall, London.

McCulloch, C. E. (1997), 'Maximum likelihood algorithms for generalized linear mixed models', *Journal of the American Statistical Association* **92**(437), 162–170.

McCulloch, C. E. & Searle, S. R. (2001), *Generalized, Linear, and Mixed Models*, John Wiley & Sons, Inc.

McLachlan, G. J. & Krishnan, T. (1997), *The EM Algorithm and Extensions*, Wiley.

Nelder, J. A. & Mead, R. (1965), 'A simplex method for function minimization', *Computational Journal* **7**, 308–313.

Nott, D. J. & Rydén, T. (1999), 'Pairwise likelihood methods for inference in image models', *Biometrika* **86**(3), 661–676.

Parner, E. T. (2001), 'A composite likelihood approach to multivariate survival data', *The Scandinavian Journal of Statistics* **28**, 295–302.

Renard, D., Molenberghs, G. & Geys, H. (2003), 'A pairwise likelihood approach to estimation in multilevel probit models', *Computational Statistics and Data Analysis* **42**, 000 – 000.

Rue, H. (2001), 'Fast sampling of Gaussian Markov random fields', *Journal of the Royal Society, Series B.*

Rue, H. & Tjelmeland, H. (2002), 'Fitting Gaussian Markov random fields to Gaussian fields', *Scandinavian Journal of Statistics.*

Shun, Z. & McCullogh, C. E. (1995), 'Laplace approximation of high-dimensional integrals', *Journal of the Royal Statistical Society, Series B* **57**, 749 – 760.

Skaug, H. J. (2002), 'Automatic differentiation to facilitate maximum likelihood estimation in nonlinear random effects models', *Journal of Computational and Graphical Statistics* **11**(2), 458–470.

Varin, C. & Vidoni, P. (2003), Composite likelihood model selection, Technical report, Department of Statistical Sciences, Udine (Italy). submitted.

Wu, C. F. J. (1983), 'On the convergence properties of the EM algorithm', *The Annals of Statistics* **11**(1), 95–103.

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