Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Computational Statistics & Data Analysis 49 (2005) 1173–1191
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Pairwise likelihood inference in spatial generalized linear mixed models

Cristiano Varin[a,*], Gudmund Høst[b], Øivind Skare[c]

[a] *Department of Statistics, University of Padova, via Cesare Battisti, 241, 35121 Padova, Italy*
[b] *The Research Council of Norway, Oslo, Norway*
[c] *Institute of Biology, University of Oslo, Norway*

Received 23 January 2004; received in revised form 2 July 2004; accepted 26 July 2004
Available online 21 August 2004

Abstract

Spatial generalized linear mixed models are flexible models for a variety of applications, where spatially dependent and non-Gaussian random variables are observed. The 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. In particular, a composite likelihood based on pairs of observations is adopted. In order to maximize the pairwise likelihood, a new expectation–maximization-type algorithm which uses numerical quadrature is introduced. The method is illustrated on simulated data and on data from air pollution effects for fish populations in Norwegian lakes. A comparison with alternative methods is given. The proposed algorithm is found to give reasonable parameter estimates and to be computationally efficient.

Keywords: Composite likelihood; Expectation–maximization algorithm; Gauss–Hermite quadrature; Model-based geostatistics; Pairwise likelihood

1. Introduction

We present a methodology for computationally efficient parameter estimation in generalized linear mixed models (GLMMs) for spatial data. This class of models has found

* Corresponding author. Tel.: +39-049-8274192; fax: +39-049-8274140.
E-mail addresses: sammy@stat.unipd.it (C. Varin), gho@rcn.no (G. Høst), skare@bio.uio.no (Ø. Skare).

0167-9473/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2004.07.021

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 et al. (1998). Here, the underlying random effects were modeled by a Gaussian random field (GRF). As in standard GLMM (Breslow and 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 and Waagepetersen (2002).

Maximum likelihood (ML) 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 ML estimation within GLMMs. Zhang (2002) used a ML approach together with an Monte Carlo EM (MCEM) algorithm to estimate parameters of a general spatial GLMM. Alternatively, the integral may be computed by the Laplace's method, which uses a Gaussian approximation of the integrand. In GLMM inference, this method has been used by Breslow and Clayton (1993) and Skaug (2002).

Both Bayesian MCMC and ML MCEM inference involve 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 and 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 et al. (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). A general discussion on its merits is given in Cox and Reid (2004). Applications to correlated data include random set models in image analysis (Nott and Rydén, 1999), correlated binary data (Kuk and Nott, 2000), multivariate survival data analysis (Parner, 2001), multilevel models (Renard et al., 2004) and frailty models for longitudinal data (Henderson and Shimakura, 2003). Applications to Gaussian spatial data have been described by Hjort and Omre (1994). Heagerty and 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 implementation 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.

(1) Denote by $\{u(\boldsymbol{s}) : \boldsymbol{s} \in S\}$ a stationary GRF with zero mean and spatial covariance function $\mathrm{cov}(u(\boldsymbol{s}), u(\boldsymbol{s}')) = \sigma^2\rho(\boldsymbol{s}-\boldsymbol{s}';\boldsymbol{\alpha})$. Here, $\rho(\cdot;\boldsymbol{\alpha})$ is a positive definite function and $\boldsymbol{\alpha}$ is a vector of correlation parameters.
(2) Given $\boldsymbol{u} = (u(\boldsymbol{s}_1), \ldots, u(\boldsymbol{s}_n))^{\mathrm{T}}$; $\boldsymbol{s}_i \in S$; $i = 1, \ldots, n$, the observations $\boldsymbol{y} = (y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_n))^{\mathrm{T}}$ are mutually independent.
(3) The conditional mean of an observation at $\boldsymbol{s}$ is $\mathrm{E}(y(\boldsymbol{s})|u(\boldsymbol{s})) = g^{-1}(\eta(\boldsymbol{s}))$, 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.
(4) Given a dispersion 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

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

Here, $\mu_i = \mathrm{E}(y_i|u_i)$ while $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are specific functions, see McCullagh and 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 nested GLMM (McCulloch and 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 the pairwise likelihood (Lindsay, 1988), which is the product of the bivariate likelihoods

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

where $\mathscr{R}$ is a subset of all possible pairwise neighbors. The element in $\boldsymbol{\Psi}$ which maximizes the pairwise likelihood is $\widehat{\boldsymbol{\psi}}_{\mathrm{MPL}}$, 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

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

Conditions for consistency and asymptotic normality of the MPL estimator under increasing-domain asymptotics were given by Heagerty and Lele (1998) and Heagerty and Lumley (2000). Zhang (2004) showed that the same results are not valid under fixed-domain aymptotics. The Author pointed out that, if the correlation function of the hidden GRF belongs to the Matern class, it is not possible to consistently estimate both the variance and the correlation range of the hidden GRF, regardless of the estimation method. Instead, the inferential focus should be on the ratio of these two parameters which is also more relevant for interpolations. We will turn to this point in greater detail in Section 5.

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.

Algorithm 1. *Choose a starting value* $\boldsymbol{\psi}^{(0)}$ *such that* $PL(\boldsymbol{\psi}^{(0)};\boldsymbol{y}) > 0$ and set $d = 0$. *The pairwise EM* (*PEM*) *algorithm iterates the following steps until convergence*.

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

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

(2) *Maximization step*: *choose* $\boldsymbol{\psi}^{(d+1)}$ *such that* $\boldsymbol{\psi}^{(d+1)} = \text{argmax}_{\boldsymbol{\psi}} Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(d)})$.
(3) *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* $\boldsymbol{\psi}^{(0)}, \boldsymbol{\psi}^{(1)}, \boldsymbol{\psi}^{(2)}, \ldots$ *be the sequence of iterates of PEM, then the pairwise likelihood does not decrease at each iteration of the PEM algorithm*, i.e., $PL(\boldsymbol{\psi}^{(d)}; \boldsymbol{y}) \leqslant PL(\boldsymbol{\psi}^{(d+1)}; \boldsymbol{y})$.

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 $\boldsymbol{\psi}^{(d)}$ maximizes $Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(d)})$ for the convergence of PEM. Indeed, the ascent property is still satisfied if $\boldsymbol{\psi}^{(d+1)}$ is chosen such that

$$Q(\boldsymbol{\psi}^{(d+1)}|\boldsymbol{\psi}^{(d)}) \geqslant Q(\boldsymbol{\psi}^{(d)}|\boldsymbol{\psi}^{(d)}). \tag{4}$$

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

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

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

(1) *Approximate expectation step*: *approximate the expectation step* (3) *in PEM by* $\widehat{Q}(\boldsymbol{\psi}; \boldsymbol{\psi}^{(d)})$.
(2) *Generalized maximization step*: *choose* $\boldsymbol{\psi}^{(d+1)}$ *such that* $\widehat{Q}(\boldsymbol{\psi}^{(d+1)}|\boldsymbol{\psi}^{(d)}) \geqslant \widehat{Q}(\boldsymbol{\psi}^{(d)}|\boldsymbol{\psi}^{(d)})$.
(3) *Set* $d = d + 1$.

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

Booth and 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.

Observe that the above algorithm could be generalized to any form of composite likelihood.

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 $Q(\boldsymbol{\psi}; \boldsymbol{\psi}^{(d)})$ is

$$\widehat{Q}(\boldsymbol{\psi}; \boldsymbol{\psi}^{(d)}) = \sum_{(i,j)\in\mathscr{R}} \sum_{m=1}^{M} \log\, f(\boldsymbol{u}_{(i,j)}(m), y_i, y_j; \boldsymbol{\psi}) w\left(\boldsymbol{u}_{(i,j)}(m); \boldsymbol{\psi}^{(d)}\right). \tag{5}$$

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

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

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

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

and

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

The advantage of the decomposition above is that the two terms may be maximized separately. The first term, $\widehat{Q}(\boldsymbol{\beta}, \phi; \boldsymbol{\psi}^{(d)})$, is the usual GLM term, involving only the fixed effects. The second term, $\widehat{Q}(\sigma^2, \boldsymbol{\alpha}; \boldsymbol{\psi}^{(d)})$, involves only the random effects parameters. The fixed effects term may be maximized by iterative weighted least squares as described in McCullagh and Nelder (1989).

The random effects term $\widehat{Q}(\sigma^2, \boldsymbol{\alpha}; \boldsymbol{\psi}^{(d)})$ 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 and 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., $\widehat{r}(\boldsymbol{s}_i) = g(y_i) - \boldsymbol{x}_i \widehat{\boldsymbol{\beta}}^{(0)}$, $i = 1, \ldots, n$. 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 σ^2 and $\boldsymbol{\alpha}$ 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 and 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 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 ML parameter estimates are often based on the information matrix. For pairwise likelihood, we suggest corresponding variance estimates based on *pairwise information*. Write $pl(\boldsymbol{\psi}; \boldsymbol{y}) = \log PL(\boldsymbol{\psi}; \boldsymbol{y})$, then a Taylor series argument, (Heagerty and Lele, 1998; Nott and Rydén, 1999), shows that the asymptotic variance of the MPL estimator is the inverse of the pairwise information

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

To estimate $\boldsymbol{I}(\boldsymbol{\psi})$, we need the gradient and the Hessian of each bivariate log likelihood $\log f(y_i, y_j; \boldsymbol{\psi})$. These are obtained by straightforward differentiation using the approach in Louis (1982) applied to pairs of indices (i, j). Thus, the gradient and the Hessian matrix of $\log f(y_i, y_j; \boldsymbol{\psi})$ are given by

$$\nabla \log f(y_i, y_j; \boldsymbol{\psi}) = \mathrm{E} \left\{ \nabla \log f(u_i, u_j, y_i, y_j; \boldsymbol{\psi}) \middle| y_i, y_j; \boldsymbol{\psi}^{(d)} \right\},$$

and

$$\nabla^2 \log\, f(y_i, y_j; \boldsymbol{\psi}) = \mathrm{E}\left\{ \nabla^2 \log\, f(u_i, u_j, y_i, y_j; \boldsymbol{\psi}) \,\middle|\, y_i, y_j; \boldsymbol{\psi}^{(d)} \right\} + \mathrm{var}\left\{ \nabla \log\, f(u_i, u_j, y_i, y_j; \boldsymbol{\psi}) | y_i, y_j; \boldsymbol{\psi}^{(d)} \right\}.$$

From the above formulae, we see that the gradient and Hessian of the log-pairwise likelihood are functions of the derivatives of the terms forming $Q(\boldsymbol{\psi}; \boldsymbol{\psi}^{(d)})$.

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 $\boldsymbol{y}$. The term $\mathrm{E}_{\boldsymbol{y}}\{\nabla^2 pl(\boldsymbol{\psi}; \boldsymbol{y})\}$ can be consistently estimated by the Hessian of the log-pairwise likelihood evaluated at the MPL estimator, i.e., $\nabla^2 pl(\widehat{\boldsymbol{\psi}}_{\mathrm{MPL}}; \boldsymbol{y})$. It is more difficult to estimate $J(\boldsymbol{\psi}) = \mathrm{var}_{\boldsymbol{y}}\{\nabla pl(\boldsymbol{\psi}; \boldsymbol{y})\}$. In the real data applications we consider a Monte Carlo estimate of this quantity. We generate K replicates of the data, $\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(K)}$, fixing $\boldsymbol{\psi} \equiv \widehat{\boldsymbol{\psi}}_{\mathrm{MPL}}$, and then estimate $J(\boldsymbol{\psi})$ by

$$\widehat{J}^{(K)}(\widehat{\boldsymbol{\psi}}_{\mathrm{MPL}}) = \frac{1}{K} \sum_{k=1}^{K} \nabla pl\left(\widehat{\boldsymbol{\psi}}_{\mathrm{MPL}}; \boldsymbol{y}^{(k)}\right) \nabla pl(\widehat{\boldsymbol{\psi}}_{\mathrm{MPL}}; \boldsymbol{y}^{(k)})^{\mathrm{T}}.$$

5. Simulated data examples

We illustrate the use of our method on a spatial GLMM with Poisson responses, 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

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

Here, the constant 3 is introduced in accordance with the 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 Fig. 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 $(0.1, 10)$ and α to the interval $(0.1, 15)$. We used relative difference $\max_i |\boldsymbol{\psi}_i^{(d+1)} - \boldsymbol{\psi}_i^{(d)}| / |\boldsymbol{\psi}_i^{(d)}| < 0.0005$ as convergence criterion. Alternatively, the algorithm was stopped when $Q(\boldsymbol{\psi}; \boldsymbol{\psi}^{(d)})$ 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. Fig. 2 shows parameter values as function of iteration number for one simulated data set. In one case the algorithm did not converge since $\widehat{\sigma}^2$ tended to 10.0, the maximal value admissible in our constrains. In the remaining cases, the algorithm converged in less than 40 iterations for 81 of the data sets. For 14 data sets 40–100 iterations

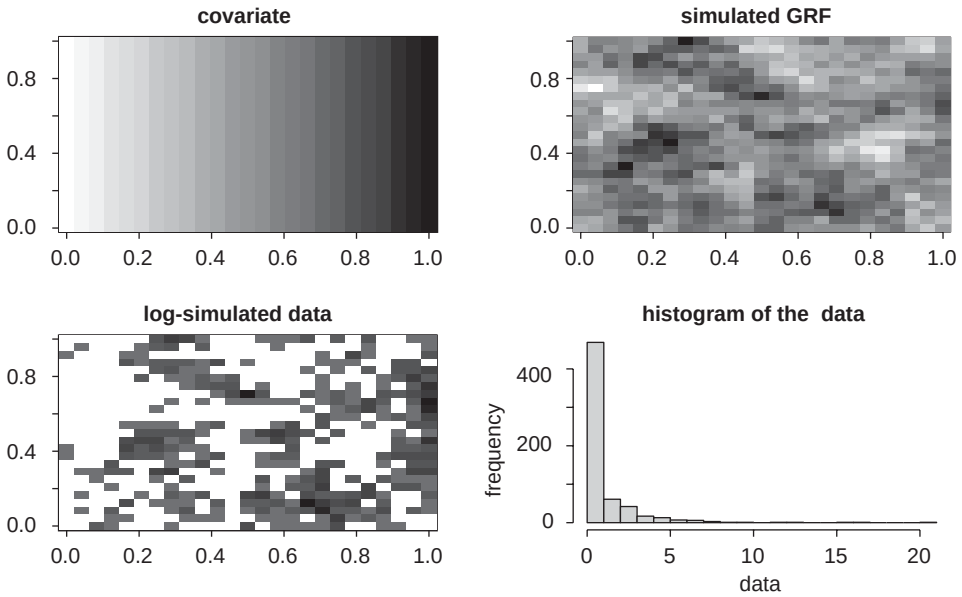

Fig. 1. Realization from the spatial Poisson model. From top-left to bottom-right: covariate $x(\boldsymbol{s}) = \boldsymbol{s}_1$, simulated mixed effects field $u(\boldsymbol{s})$, (log) simulated data, histogram of the data.

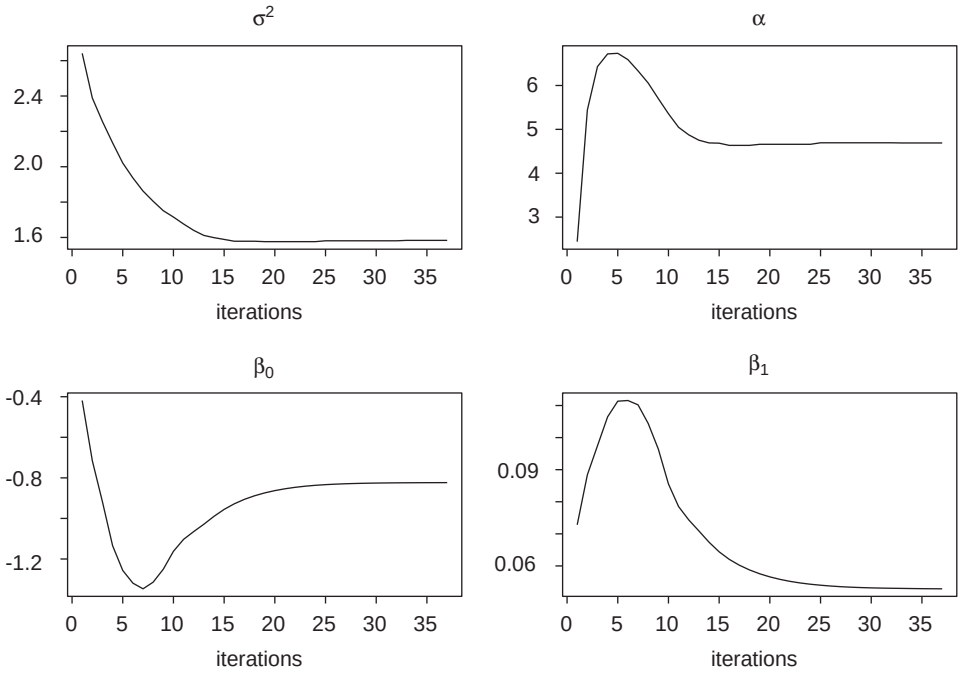

Fig. 2. QPEM iterates for a typical data set. From top-left to bottom-right: σ^2, α, β_0 and β_1.

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

Parameter	True	Estimate (mean)	SD	MSE
σ^2	1.5	1.40	0.40	0.1672
α	6.0	5.33	2.04	4.6027
α/σ^2	4.0	3.95	1.40	1.9759
β_0	−2.0	−2.008	0.68	0.4615
β_1	0.1	0.105	0.05	0.0024

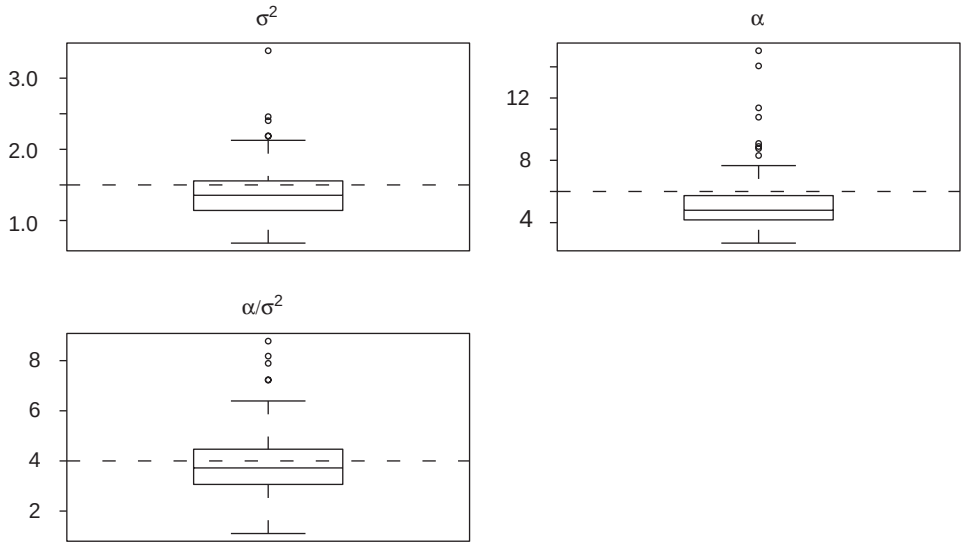

Fig. 3. Box-plots for σ^2, α and α/σ^2. The horizontal lines represents the true values.

were needed, while the remaining 4 converging 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 the estimated regression parameters have negligible bias and variance, while some bias occured for the random effects parameters, in particular for the range parameter α. Since the exponential correlation function belongs to the Matern class α and σ^2 cannot be consistently estimated under fixed-domain asympototics (Zhang, 2004) and the ML inference should instead focus on the ratio between these two parameters. The box-plots of MPL estimators (Fig. 3) seem leading to the same conclusions. In fact, the ratio α/σ^2 is more accurately estimated by QPEM.

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 within a neighborhood of radius 4, as shown

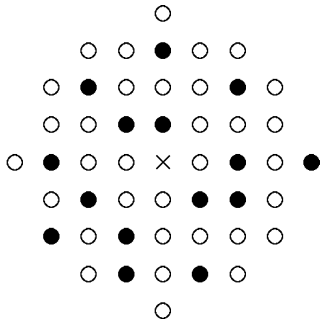

Fig. 4. Sampling pairs within a neighborhood of radius 4. Here, × is the observation location and the filled circles are 15 neighbors sampled at random without replacement. The contributing pairs consist of × and each of the 15 sampled neighbors.

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

Parameter	True	Estimate (mean)	SD	MSE
σ^2	1.5	1.44	0.44	0.1997
α	6.0	5.50	2.21	5.1651
α/σ^2	4.0	3.97	1.39	1.9319
β_0	−2.0	−2.026	0.66	0.4314
β_1	0.1	0.107	0.05	0.0025

in Fig. 4. Running through all data locations, we obtained about $15n = 9375$ 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 . The algorithm converged in all the 100 data sets, that is also in the case where failed considering all the pairs. Again, there is good correspondence between true and estimated values for β_0, β_1 and α/σ^2. We see that negligible 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 ML estimation. The results from this study comparison will depend 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 and McCullogh, 1995). A general

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

Pram	ADMB			QPEM		
	True	Mean	MSE	Mean	MSE	Corr
σ^2	1.0	0.93	0.060	1.09	0.346	0.524
α	3.0	2.48	1.814	2.54	1.471	0.543
α/σ^2	3.0	2.73	2.092	2.60	1.621	0.525
β_0	−2.0	−1.97	0.257	−1.99	0.476	0.757
β_1	0.3	0.29	0.003	0.29	0.007	0.661
σ^2	1.0	0.77	0.112	0.90	0.362	0.601
α	6.0	3.90	6.817	4.08	6.790	0.458
α/σ^2	6.0	5.46	8.763	5.17	6.377	0.473
β_0	−2.0	−-1.96	0.526	−1.99	0.597	0.927
β_1	0.3	0.29	0.007	0.29	0.008	0.868

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 $\alpha = 3.0$. 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. In 3 data sets QPEM did not converge. In two cases, $\widehat{\sigma}^2$ tended to the upper boundary of our constrains (10), while in another case $\widehat{\alpha}$ tended to its lower boundary (0.1). Even more convergence problems occur with ADMB. Indeed, the algorithm did not converge in 11 data sets always for problems with α. We note that the present choice of β_0 and β_1 typically gives simulated data with very skewed likelihood. Laplace approximation used in ADMB is based on a Gaussian approximation to the likelihood which may not work well in this situation. We see that the mean squared errors for the estimated regression parameters β_0 and β_1 are sligthly smaller for ML than for QPEM. Viceversa, the ratio α/σ^2 is better estimated by QPEM.

We also repeat the exercise doubling the range parameter, Table 3. In this case QPEM converged in all the data set but one where $\widehat{\alpha}$ tended to the upper boundary in our constrains. Instead, ADMB did not converge in 10 data sets. As for the case before, ADMB convergence problems were caused by the estimation of the range. Here, the performance of ML and MPL estimators of the regression parameters is very similar, while the ratio α/σ^2 is again better estimated by QPEM.

The computing time for QPEM on a typical data set using a 550 MHz Pentium III with 4 GB RAM was 69 s. The ML estimates for the same data set was computed in 756 s. In this case, QPEM used 21 MB 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.

Fig. 5 shows the observed population status. Here, light gray circles mark lakes unaffected by acidification, while dark gray circles mark affected lakes. We see that trout in the southern and western parts of Norway are most affected by acidification.

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, logit-link and two regression parameters, i.e., $\eta(\boldsymbol{s}) = \beta_0 + \beta_1 ANC(\boldsymbol{s}) + u(\boldsymbol{s})$. 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 function $glm()$ implemented in the R language (Ihaka and Gentleman, 1996). The parameter estimates obtained by R $glm()$ were $\beta_0^{(0)} = 0.222$ and $\beta_1^{(1)} = -0.120$. As suggested in Section 4, we calculated an empirical variogram of transformed residuals. The shape of the variogram suggested an exponential

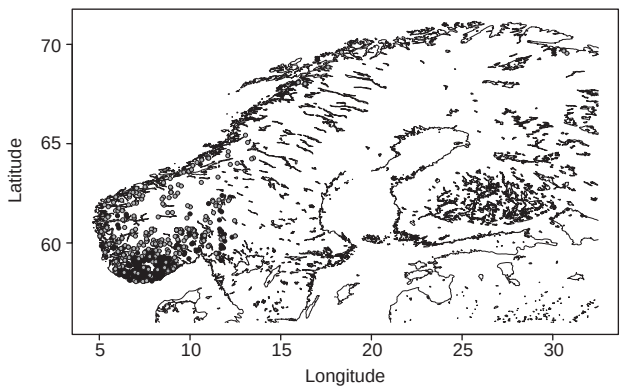

Fig. 5. Trout data. light gray denotes lakes where trout are not affected by acidification, dark gray circles lakes where the trout population has decreased or is extinct.

covariance function, and a least squares fit provided starting values $\sigma^{2(0)} = 2.248$ and $\alpha^{(0)} = 171.8\,\text{km}$.

Data pairs for the QPEM algorithm were obtained by using $r = 30$ locations sampled from a neighborhood of radius 250 km around each of the 400 data locations. We used 5×5 points Gauss–Hermite quadrature and a tolerance of 10^{-5} for convergence (one order less than in the simulated data examples). The algorithm converged after 77 iterations to the values $\widehat{\alpha}/\widehat{\sigma}^2 = 78.3$, ($\widehat{\sigma}^2 = 2.703$, $\widehat{\alpha} = 211.5\,\text{km}$), $\widehat{\beta}_0 = 0.276$ and $\widehat{\beta}_1 = -0.169$. Convergence to a local maximum was checked by re-running the algorithm with perturbed starting values. The matrix $J(\boldsymbol{\psi})$ was estimated by Monte Carlo using 500 simulated replicates of the data. The standard deviations obtained for β_0, β_1 and α/σ^2 were 0.457, 0.0344 and 10.159, respectively.

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 means 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 134 of the 142 locations, i.e., about 94% 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 such situations, pairwise likelihood can benefit from a parallel implementation where different processors are used to evaluate each pairwise component. In contrast, QPEM may also be a practical tool for finding starting values for ML 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. Adaptive quadrature methods could be appealing in various situations. Advantages of adaptive quadrature include robustnes, see Lesaffre and Speissens (2001), and the possibility to fix the precision.

The computation of the standard deviations of the MPL estimators requires to estimate $J(\psi)$, the variance of the log-pairwise likelihood. In the paper, we consider a Monte Carlo estimator of this quantity. Alternatively, when the data are disposed on a spatial grid, one could consider a window resampling strategy as described by Heagerty and Lele (1998). In window resampling, we subdivide the study region into overlapping spatial windows and compute empirical estimates of $J(\psi)$ 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 for general estimating functions in spatial models are given in Heagerty and Lumley (2000). Window-subsampling could be promising for very large applications where simulations should be avoided.

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 and non-stationary mixed effects models.

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. They also acknowledge Guido Masarotto and two anonymous referees whose suggestions led to greatly improve the manuscript. Norwegian Institute of Water research kindly provided the data for the fish population example. The first author was partially supported by MIUR (Italy) grant 2002134337 “Statistics as an aid for environmental decisions: identification, monitoring and evaluation” and by a three months scholarship from the Research Council of Norway.

Appendix A. Proof of Proposition 1

Choose a starting value $\boldsymbol{\psi}^{(0)}$ and write $pl(\boldsymbol{\psi};\boldsymbol{y}) = \log\, PL(\boldsymbol{\psi};\boldsymbol{y})$. We have

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

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

$$\begin{aligned}
pl(\boldsymbol{\psi}^{(d+1)};\boldsymbol{y}) - pl(\boldsymbol{\psi}^{(d)};\boldsymbol{y}) &= Q(\boldsymbol{\psi}^{(d+1)}|\boldsymbol{\psi}^{(d)}) - Q(\boldsymbol{\psi}^{(d)}|\boldsymbol{\psi}^{(d)}) \\
&\quad + \sum_{(i,j)\in\mathscr{R}} D_{(i,j)}(\boldsymbol{\psi}^{(d+1)}|\boldsymbol{\psi}^{(d)}),
\end{aligned}$$

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

$$\begin{aligned}
D_{(i,j)}(\boldsymbol{\psi}^{(d+1)}|\boldsymbol{\psi}^{(d)}) = &-\int\int \log\left\{\frac{f(u_i, u_j|y_i, y_j;\boldsymbol{\psi}^{(d+1)})}{f(u_i, u_j|y_i, y_j;\boldsymbol{\psi}^{(d)})}\right\} \\
&\times f(u_i, u_j|y_i, y_j;\boldsymbol{\psi}^{(d)})\,\mathrm{d}u_i\,\mathrm{d}u_j.
\end{aligned}$$

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

Appendix 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 $\tilde{f}(\mathbf{t}) = \mathrm{e}^{\|\mathbf{t}\|^2/2} f(\mathbf{t})$. Gauss–Hermite quadrature reduces each 1D integral to a weighted sum of $\tilde{f}(\mathbf{t})$ evaluated at M specific nodes. If $\tilde{f}(\mathbf{t})$ is well approximated by a polynomial of low order, then the integral may be accurately computed using a small M. Adaptive Gauss–Hermite quadrature, using a Gaussian approximation of $f(u_i, u_j|y_i, y_j;\boldsymbol{\psi}^{(d)})$, could give acceptable

accuracy with a low M. However, to compute the approximation, we need to compute the mode and the second derivatives of $f(u_i, u_j|y_i, y_j; \boldsymbol{\psi}^{(d)})$. 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 $Q(\boldsymbol{\psi}; \boldsymbol{\psi}^{(d)}) = \sum_{(i,j)} Q_{(i,j)}(\boldsymbol{\psi}; \boldsymbol{\psi}^{(d)})$. Each $Q_{(i,j)}(\boldsymbol{\psi}; \boldsymbol{\psi}^{(d)})$ is a ratio of two double integrals. The numerator of $Q_{(i,j)}(\boldsymbol{\psi}; \boldsymbol{\psi}^{(d)})$ is given by

$$\int\int \log\{f(u_i, u_j, y_i, y_j; \boldsymbol{\psi})\} f(y_i|u_i; \boldsymbol{\psi}^{(d)}) f(y_j|u_j; \boldsymbol{\psi}^{(d)}) \times f(u_i, u_j; \boldsymbol{\psi}^{(d)})\, \mathrm{d}u_i\, \mathrm{d}u_j, \tag{B.1}$$

while the denominator is

$$\int\int f(y_i|u_i; \boldsymbol{\psi}^{(d)}) f(y_j|u_j; \boldsymbol{\psi}^{(d)}) f(u_i, u_j; \boldsymbol{\psi}^{(d)})\, \mathrm{d}u_i\, \mathrm{d}u_j. \tag{B.2}$$

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

$$\begin{aligned} v_i &= \frac{u_i}{\sigma}, \\ v_j &= \frac{u_j - \rho_{(i,j)} u_i}{\sigma(1-\rho_{(i,j)}^2)^{1/2}}, \end{aligned} \tag{B.3}$$

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 (B.3), the denominator (B.2) becomes

$$\frac{1}{2\pi}\int\int f(y_i|u_i(v_i); \boldsymbol{\psi}^{(d)}) f(y_j|u_j(v_i, v_j); \boldsymbol{\psi}^{(d)}) \mathrm{e}^{-v_i^2/2} \mathrm{e}^{-v_j^2/2}\, \mathrm{d}v_i\, \mathrm{d}v_j. \tag{B.4}$$

Now, (B.4) can be approximated by Gauss–Hermite quadrature

$$\frac{1}{2\pi}\sum_{m_1=1}^{M_1}\sum_{m_2=1}^{M_2} f\left\{y_i|u_i\,(h(m_1)); \boldsymbol{\psi}^{(d)}\right\} f\left\{y_j|u_j\,(h(m_1), h(m_2)); \boldsymbol{\psi}^{(d)}\right\} k(m_1)k(m_2). \tag{B.5}$$

Here, $h(m)$ are the quadrature nodes and $k(m)$ are the quadrature weights. The quadrature formula for the numerator (B.1) is derived by a similar procedure. The denominator depends on $\boldsymbol{\psi}^{(d)}$ 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 $Q_{(i,j)}(\boldsymbol{\psi}; \boldsymbol{\psi}^{(d)})$ becomes

$$\sum_{m_1,m_2} \log\, f\left\{u_i\,(h(m_1))\,, u_j\,(h(m_1), h(m_2))\,, y_i, y_j; \boldsymbol{\psi}\right\} w_{(i,j)}\left(m_1, m_2; \boldsymbol{\psi}^{(d)}\right),$$

where the new weights are

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

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

$$w\left(\boldsymbol{u}_{(i,j)}(m);\boldsymbol{\psi}^{(d)}\right)=w_{(i,j)}\left(m_1,m_2;\boldsymbol{\psi}^{(d)}\right),$$

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. J. Roy. Statist. Soc. Ser. B 61, 265–285.

Breslow, N.E., Clayton, D.G., 1993. Approximate inference in generalized linear mixed models. J. Amer. Statist. Assoc. 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.

Cox, D.R., Reid, N., 2004. A note on pseudolikelihood constructed from marginal densities. Biometrika, to appear.

Cressie, N., 1993. Statistics for Spatial Data, second ed. Wiley, New York.

Diggle, P.J., Tawn, J.A., Moyeed, R.A., 1998. Model-based Geostatistics (with discussion). J. Roy. Statist. Soc. Ser. B 47 (2), 299–350.

Heagerty, P.J., Lele, S.R., 1998. A composite likelihood approach to binary spatial data. J. Amer. Statist. Assoc. 93, 1099–1111.

Heagerty, P.J., Lumley, T., 2000. Window subsampling of estimating functions with application to regression models. J. Amer. Statist. Assoc. 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. Scand. J. Statist. 21, 289–357.

Huang, H.-C., Cressie, N., Gabrosek, J., 2002. Fast resolution consistent spatial prediction of global processes from satellite data. J. Comput. Graphical Statist. 11 (1), 63–88.

Ihaka, R., Gentleman, R., 1996. R: A language for data analysis and graphics. J. Comput. Graphical Statist. 5 (3), 299–314.

Kuk, A.Y., Nott, D., 2000. A pairwise likelihood approach to analyzing correlated binary data. Statist. Probab. Lett. 47, 329–335.

Lesaffre, E., Speissens, B., 2001. On the effect of the number of quadrature points in a logistic random-effects model: an example. Appl. Statist. 50 (3), 325–335.

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

Louis, T.A., 1982. Finding the observed information matrix when using the EM algorithm. J. Roy. Statist. Soc. Ser. B 44 (2), 226–233.

McCulloch, C.E., 1997. Maximum likelihood algorithms for generalized linear mixed models. J. Amer. Statist. Assoc. 92 (437), 162–170.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, second ed. Chapman & Hall, London.

McCulloch, C.E., Searle, S.R., 2001. Generalized, Linear, and Mixed Models, Wiley, New York.

McLachlan, G.J., Krishnan, T., 1997. The EM Algorithm and Extensions, Wiley, New York.

Nelder, J.A., Mead, R., 1965. A simplex method for function minimization. Comput. J. 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. Scand. J. Statist. 28, 295–302.

Renard, D., Molenberghs, G., Geys, H., 2004. A pairwise likelihood approach to estimation in multilevel probit models. Comput. Statist. Data Anal. 44, 649–667.

Rue, H., 2001. Fast sampling of Gaussian Markov random fields. J. Roy. Soc. Ser. B 63 (2),

Rue, H., Tjelmeland, H., 2002. Fitting Gaussian Markov random fields to Gaussian fields. Scand. J. Statist. 29 (1), 31–49.

Shun, Z., McCullogh, C.E., 1995. Laplace approximation of high-dimensional integrals. J. Roy. Statist. Soc. Ser. B 57, 749–760.

Skaug, H.J., 2002. Automatic differentiation to facilitate maximum likelihood estimation in nonlinear random effects models. J. Comput. Graphical Statist. 11 (2), 458–470.

Wu, C.F.J., 1983. On the convergence properties of the EM algorithm. Anna. Statist. 11 (1), 95–103.

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

Zhang, H., 2004. Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. J. Amer. Statist. Assoc. 99, 250–261.