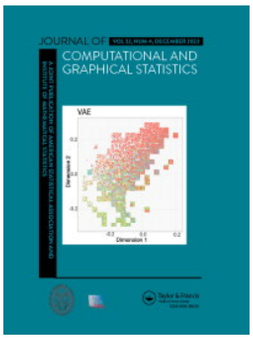

Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/ucgs20

Maximum Likelihood Algorithm for Spatial Generalized Linear Mixed Models without Numerical Evaluations of Intractable Integrals

Tonglin Zhang

To cite this article: Tonglin Zhang (2023) Maximum Likelihood Algorithm for Spatial Generalized Linear Mixed Models without Numerical Evaluations of Intractable Integrals, Journal of Computational and Graphical Statistics, 32:4, 1636-1648, DOI: 10.1080/10618600.2022.2163651

To link to this article: https://doi.org/10.1080/10618600.2022.2163651

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2023, VOL. 32, NO. 4, 1636–1648
https://doi.org/10.1080/10618600.2023.2163651

Maximum Likelihood Algorithm for Spatial Generalized Linear Mixed Models without Numerical Evaluations of Intractable Integrals

Tonglin Zhang

Department of Statistics, Purdue University, West Lafayette, IN

ABSTRACT

Spatial generalized linear mixed effects models are popular in spatial or spatiotemporal data analysis when the responses are counts and the random effects are modeled by multivariate normal distributions. Direct computation of the MLEs of model parameters is impossible because the likelihood functions contain high-dimensional intractable integrals. To overcome the difficulty, a new method called the prediction-maximization algorithm is proposed. The method has a maximization step for the MLEs of spatial linear mixed effects models for normal responses and a prediction step for the prediction of the random effects. None of them involves high-dimensional intractable integrals. Because only algorithms for the normal responses are needed, the derivation of the MLEs of a spatial generalized linear mixed effects model for count responses by the proposed method is not computationally harder than a model for normal responses. The simulation study shows that the performance of the proposed method is comparable to that of the previous maximum likelihood algorithms formulated by high-order Laplace approximations and is better than that of Bayesian methods formulated by MCMC algorithms. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received July 2021
Accepted December 2022

KEYWORDS
Exponential family distribution; High-dimensional intractable integrals; Laplace approximation; Penalized quasi-likelihood; Prediction-maximization algorithm; Profile likelihood

1. Introduction

Spatial generalized linear mixed effect models (SGLMMs), introduced by Diggle, Tawn, and Moyeed (1998) named as model-based geostatistics, are popular in the analysis of non-Gaussian spatial data collected from a geological or geographical region. Theoretically, an SGLMM can be treated as a special case of a generalized linear mixed effects model (GLMM) when the random effects are modeled by a spatial Gaussian process. Conditional on the spatial Gaussian process, the response variables are assumed to follow an exponential family distribution with site-specific conditional expected values. A link function connects the site-specific conditional expected values with the underlying spatial Gaussian process. Except the case when the response follows a normal distribution given the process, the likelihood function of the SGLMM in the remaining cases contains high-dimensional intractable integrals (HDIIs), leading to difficulties in the derivation of the exact MLEs of model parameters. To overcome the difficulties, numerical approximations of HDIIs are often used. Examples include the well-known Laplace Approximation (LA) (Barndorff-Neilsen and Cox 1989) and the Penalized Quasi-Likelihood (PQL) (Breslow and Clayton 1993) methods. It is believed that numerical approximations have to be used in the derivation of the approximate MLEs. The main contribution of the article is the development of the PM algorithm for the MLE of SGLMMs for count responses without the numerical approximations, which resolves the long-standing difficulties caused by the HDII problems in fitting SGLMMs.

The proposed PM method has a prediction step for the prediction of the random effects and a maximization step for the MLE of spatial linear mixed effects models (SLMMs) for normal responses. None of them needs numerical evaluations of HDIIs. Thus, HDIIs are not an issue in the proposed method. In particular, suppose that count responses are collected at sites in a spatial (or spatiotemporal) domain. Assume that an exponential family distribution is used to model the response given the random effects and a spatial Gaussian process is used to model the underlying distribution of the random effects. Then, the likelihood function of the SGLMM contains HDIIs, which is the major difficulty in the computation of the MLEs of the model parameters. This is not an issue in SLMMs for normal responses because the likelihood functions have closed-form expressions. Our work shows that the MLE of the SGLMM for count responses can be derived by iteratively using ML algorithms for SLMMs. It is not necessary to treat any HDIIs in the entire computation. The derivation of the MLEs of SGLMMs for count responses is not computationally harder than that of SLMMs for normal responses. Theoretically, our method is asymptotically equivalent to the exact MLE, meaning that $\sqrt{n}$ times the difference between the solution provided the PM and the exact MLE is asymptotically negligible (i.e., Corollary 2).

The idea of the PM is motivated by the relationship between the hierarchical-likelihood (i.e., h-likelihood) (Lee and Nelder 1996) and the likelihood approaches. The h-likelihood function is the conditional likelihood of the responses given the random effects. The likelihood function is the marginal

CONTACT Tonglin Zhang *tlzhang@purdue.edu* Department of Statistics, Purdue University, 250 North University Street, West Lafayette, IN 47907-2066.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JCGS*.

likelihood obtained by integrating out the random effects. In the h-likelihood approach, the random effects are assumed to be observed. They can be used in the derivation of the MLEs. The computation can be carried based on the h-likelihood function, but this is not a usable approach in practice. The likelihood approach, which is the usable approach, assumes that the random effects are unobserved. The computation has to be carried out based on the likelihood function. HDIIs appear in the likelihood function for counts responses when the random effects are modeled by normal distributions. To overcome the difficulty, a straightforward method is to impute the random effects by the EM-algorithm (Dempster, Laird, and Rubin 1977). In the E-step, the EM-algorithm calculates the expected value of the logarithm of the h-likelihood conditioning on the logarithm of the likelihood, implying that HDIIs are still present. Therefore, the EM is often carried out by a Monte Carlo EM (MCEM) algorithm (Zhang 2002; Guan 2021).

The HDII problem can be completely avoided in the proposed PM. To formulate the PM, we examine the algorithm for the h-likelihood approach. Because it assumes that the random effects are observed, the MLEs of fixed effects parameters can be computed by the standard iteratively reweighted least squares (IRWLS) algorithm, and the MLEs of the variance components parameters can be computed by maximizing the likelihood function of the random effects specified by the prior distribution. None of them involve HDIIs. In the working models considered by the iterations of the h-likelihood approach, the random effects appear in three places: the linear components, the working weights, and the working responses. The working weights and the working responses are constructed by the previous iteration. Only the random effects contained by the linear components are used in the current iteration. The roles of the random effects can be classified into two groups. The first group contains the linear components only. The second group contains the working weights and responses. We migrate this idea to SGLMMs, leading to the proposed PM. The PM is analogous to the EM. Both the P-step of the PM and the E-step of the EM predict the random effects. The difference is that the PM does not use the conditional expected value of the logarithm of the h-likelihood given the logarithm of the likelihood. The PM is identical to the EM only when the responses are normal.

The PM has connections with the recently developed iteratively reweighted least squares with random effects (IRWLSR) for the MLE of generalized linear mixed models (GLMMs) (Zhang 2021). Our research shows that the usage of the redundant dispersion parameter (for the variance of the errors) used by the iterations of the IRWLSR can cause the resulting estimators inconsistent. The redundant dispersion parameter should be excluded in the iterations to stabilize the computation. As all of linear mixed models for normal responses contain the dispersion parameter, it is impossible to directly use existing software packages to implement the PM. We need to write new computer code to remove impact of the redundant dispersion parameter used by the iterations. This can lead to a more reliable approach, leading to consistency of the proposed PM.

SGLMMs are special cases of GLMMs. The random effects of SLMMs are specified for spatial or spatiotemporal dependencies. GLMMs contain all scenarios of mixed effects models when the responses are modeled by exponential family distributions. It is well-known that the main difficulty in fitting GLMMs for count responses is computational because the likelihood function contains intractable integrals (IIs). To overcome the difficulty, numerical evaluations of the IIs are often used in previous methods. Examples include the penalized quasi-likelihood (PQL) (Breslow and Clayton 1993; Breslow and Lin 1996), the Gauss–Hermite quadrature (Liu and Pierce 1994), the Laplace approximation (LA) (Barndorff-Neilsen and Cox 1989; Evangelou, Zhu, and Smith 2011), and the integrated nested Laplace approximation (INLA) (Rue, Martino, and Chopin 2009). Bayesian methods developed under the Markov chain Monte Carlo (MCMC) or the Metropolis-Hasting algorithms are also used. Examples include the MCEM gradient (McCulloch 1997) and the Gibbs sampler (Zeger and Karim 1991). To handle large matrices in MCMC, an effort has been put into the combination C++ and R packages. Examples include the `mcemGLM`, `geoCount` (Jing and Oliveira 2015), `gcKrig` (Han and Oliveria 2018), and `spaMM` (Rousset 2021) packages of R.

SGLMMs is one of the hardest cases in GLMMs because the dimension of intractable integrals increases with the sample size. This is different from longitudinal data studies, because they assume that dependencies are only present for observations within clusters. In SGLMMs, dependencies have to be considered between all observations in the data, leading to difficulties in MCMC and many simulation-based methods for the computation of the MLEs (Best and Wakefield 1999; Shun and McCullagh 1995). The proposed PM approach successfully overcomes the difficulties.

The article is organized as follows. In Section 2, we introduce our method. In Section 3, we compare our method with our competitors via Monte Carlo simulations. In Section 4, we apply our method to a real world dataset. In Section 5, we provide a discussion. We put all proofs in the Appendix, supplementary materials.

2. Method

We review the definition of SGLMMs in Section 2.1. We introduce our method in Section 2.2. We compare our method to a few well-known SGLMMs in Section 2.3.

2.1. SGLMM

Let $\boldsymbol{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$ be an n-dimensional count response vector collected from n distinct sites in a spatial or spatiotemporal domain. Suppose that $\boldsymbol{y}$ follows an exponential family distribution without the dispersion parameter (e.g., for binomial or Poisson data), such that its probability mass function (PMF) can be expressed as

$$f(\boldsymbol{y}|\boldsymbol{\gamma}) = e^{\boldsymbol{y}*\boldsymbol{\theta} - b(\boldsymbol{\theta}) + c(\boldsymbol{y})}, \tag{1}$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)^\top \in \mathbb{R}^n$ is an n-dimensional vector for spatial (including spatiotemporal) random effects, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^\top$ is an n-dimensional vector, $b(\boldsymbol{\theta}) = (b(\theta_1), \ldots, b(\theta_n))^\top$ is derived by a real transformation $b(\cdot)$ on $\boldsymbol{\theta}$, $\boldsymbol{y} * \boldsymbol{\theta}$ is the Hadamard product (i.e., element-wise product) of vectors, and $c(\boldsymbol{y})$ is the normalizing constant. The SGLMM is treated as a realization of a spatial count process as

$$g[\mu(s)] = \boldsymbol{x}^\top(s)\boldsymbol{\beta} + \gamma(s), \tag{2}$$

where $\mu(s) = \mathrm{E}[y(s)]$, $y(s)$ is a count response at site s, $\boldsymbol{x}(s)$ is a p-dimensional vector of explanatory variables at site s, $\boldsymbol{\beta}$ is a p-dimensional parameter for fixed effects, and $\gamma(s)$ is an observed (i.e., latent) spatial Gaussian process. Let $y_i = y(s_i)$ and $\boldsymbol{x}_i = \boldsymbol{x}(s_i)$ be the response and explanatory variable values observed at site s_i for $i = 1, \ldots, n$, respectively. The SGLMM can be equivalently expressed as

$$g(\boldsymbol{\mu}) = g[b'(\boldsymbol{\theta})] = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\gamma}, \tag{3}$$

where $\mathbf{X} = (\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_n^\top)^\top$ is an $n \times p$ design matrix for explanatory variables given by the realization and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)^\top$ with $\gamma_i = \gamma(s_i)$ is the vector of the spatial random effects. By (1), we have $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top = b'(\boldsymbol{\theta}) = \mathrm{E}(\boldsymbol{y}|\boldsymbol{\gamma})$ and $\mathrm{cov}(\boldsymbol{y}|\boldsymbol{\gamma}) = \mathrm{diag}\{b''(\boldsymbol{\theta})\}$ with $\mu_i = b'(\theta_i) = \mathrm{E}(y_i|\boldsymbol{\gamma})$ and $\mathrm{V}(y_i|\boldsymbol{\gamma}) = b''(\theta_i)$. The prior distribution for $\boldsymbol{\gamma}$ in (3) is

$$\boldsymbol{\gamma} \sim \mathcal{N}(\mathbf{0}, \Sigma_{\boldsymbol{\omega}}), \tag{4}$$

where $\boldsymbol{\omega} \in \mathbb{R}^q$ is the parameter vector in variance components provided by the spatial Gaussian process $\gamma(s)$ of (2). In (4), $\Sigma_{\boldsymbol{\omega}}$ can be specified by the geostatistical (GEO) (Diggle, Tawn, and Moyeed 1998), the spatial autoregressive (SAR) (Ork 1975), or the conditional autoregressive (CAR) (Besag 1974; Cressie and Chan 1989) models. This is discussed in Section 2.3.

The SGLMM jointly defined by (1), (3), and (4) can be interpreted by spatial hierarchical models with two hierarchical levels. The first level, given by (1) and (3), specifies the conditional distribution of the response given the random effects. The second level, given by (4), specifies the distribution of the random effects. The choice of the multivariate normal distribution for $\boldsymbol{\gamma}$ is convenient in modeling spatial or spatiotemporal dependencies.

The h-likelihood (Lee and Nelder 1996) is given by the joint distribution of $\boldsymbol{y}$ and $\boldsymbol{\gamma}$ as

$$\begin{aligned} L_h(\boldsymbol{\beta}, \boldsymbol{\omega}) &= f_{h,\boldsymbol{\beta}\boldsymbol{\omega}}(\boldsymbol{y}, \boldsymbol{\gamma}) \\ &= e^{\boldsymbol{y}^*\boldsymbol{\theta} - b(\boldsymbol{\theta}) + c(\boldsymbol{y})}(2\pi)^{-\frac{n}{2}}[\det(\Sigma_{\boldsymbol{\omega}})]^{-\frac{1}{2}} e^{-\frac{1}{2}\boldsymbol{\gamma}^\top \Sigma_{\boldsymbol{\omega}}^{-1}\boldsymbol{\gamma}}, \end{aligned} \tag{5}$$

where $f_{h,\boldsymbol{\beta}\boldsymbol{\omega}}(\boldsymbol{y}, \boldsymbol{\gamma})$ is the joint probability mass function (PMF)-probability density function(PDF) of $\boldsymbol{y}$ and $\boldsymbol{\gamma}$. This is used to compute the conditional MLEs of $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ given $\boldsymbol{\gamma}$, leading to

$$(\hat{\boldsymbol{\beta}}_{h,\gamma}^\top, \hat{\boldsymbol{\omega}}_{h,\gamma}^\top)^\top = \underset{\boldsymbol{\beta},\boldsymbol{\omega}}{\mathrm{argmin}}\, \ell_h(\boldsymbol{\beta}, \boldsymbol{\omega}), \tag{6}$$

where $\ell_h(\boldsymbol{\beta}, \boldsymbol{\omega}) = \log L_h(\boldsymbol{\beta}, \boldsymbol{\gamma})$ is the h-loglikelihood function. If $\boldsymbol{\gamma}$ is unobserved, then the computation of the MLEs has to be carried out based on the likelihood as

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{\omega}) &= f_{\boldsymbol{\beta}\boldsymbol{\omega}}(\boldsymbol{y}) = \int_{\mathbb{R}^n} L_h(\boldsymbol{\beta}, \boldsymbol{\omega}) d\boldsymbol{\gamma} = \int_{\mathbb{R}^n} f_{h,\boldsymbol{\beta}\boldsymbol{\omega}}(\boldsymbol{y}, \boldsymbol{\gamma}) d\boldsymbol{\gamma} \\ &= \int_{\mathbb{R}^n} e^{\ell_h(\boldsymbol{\beta},\boldsymbol{\omega})} d\boldsymbol{\gamma}, \end{aligned} \tag{7}$$

where $f_{\boldsymbol{\beta}\boldsymbol{\omega}}(\boldsymbol{y})$ is the marginal PMF of $\boldsymbol{y}$. This is used to compute the MLEs of $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ by

$$(\hat{\boldsymbol{\beta}}^\top, \hat{\boldsymbol{\omega}}^\top)^\top = \underset{\boldsymbol{\beta},\boldsymbol{\omega}}{\mathrm{argmax}}\, \ell(\boldsymbol{\beta}, \boldsymbol{\omega}), \tag{8}$$

where $\ell(\boldsymbol{\beta}, \boldsymbol{\omega}) = \log L(\boldsymbol{\beta}, \boldsymbol{\omega})$ is the loglikelihood function. The conditional (i.e., the posterior) PDF of $\boldsymbol{\gamma}$ given $\boldsymbol{y}$ is $f_c(\boldsymbol{\gamma}|\boldsymbol{y}) = f_{h,\boldsymbol{\beta}\boldsymbol{\omega}}(\boldsymbol{y}, \boldsymbol{\gamma})/f_{\boldsymbol{\beta}\boldsymbol{\omega}}(\boldsymbol{y})$.

The high-dimensional integral on the right-hand side of (7) is intractable if (1) is used to model the count responses and (4) is used to model the random effects. In this case, direct implementation of (8) is hard. Numerical evaluations of the HDIIs are often used. One of the most popular methods is the Laplace Approximation (LA). The idea of LA is to replace the integral by a Taylor expansion around the maximizer (i.e., mode) of the integrand. LA has also been used in the Bayesian approach for approximating posterior expectations. Two well-known methods under LA are the PQL (Breslow and Clayton 1993) and the INLA (Rue, Martino, and Chopin 2009). Because the simple version of the PQL, which only uses the first-order Taylor expansion to approximate the HDII, contains asymptotic biases (Breslow and Lin 1996; Lin and Breslow 1996), it is not recommended in practice (McCulloch, Searle, and Neuhaus 2008, p. 341). To reduce the asymptotic bias, higher-order LA is often used (Evangelou, Zhu, and Smith 2011). This has been incorporated in the recent developed `spaMM` package of R.

In addition, MCMC algorithms developed under the Bayesian framework can be used. These methods contain concerns because they have a large amount of computational overhead in their implementations (Fong, Rue, and Wakefield 2010; Zhang 2019). To apply MCMC, one needs to specify a proposal distribution on $\mathbb{R}^n$ with the new sample to be accepted by a probability depending on the old sample, the new sample, and the proposal distribution. The acceptance probability is usually low if n is not very small. To solve the problem, the Gibbs sampler is proposed (Zeger and Karim 1991). Instead of updating the entire parameter vector simultaneously, the Gibbs sampler updates individual components conditioning on the remaining components. This means that it needs to compute the inverse of a sub-matrix of the variance-covariance matrix of the random effects and its determinant. As the inverse and the determinant depend on the parameters, the implementation of MCMC to SGLMMs is time-consuming and the result is usually unstable.

Our research indicates that numerical evaluations of the HDII on the right-hand side of (7) are not necessary. Our idea is motivated by a recent article for MLE of GLMMs in general exponential family distributions (Zhang 2021). The previous method contains a dispersion parameter in its iterations. This does not occur in the proposed PM. Our study shows that the PM is more reliable than the IRWLSR in fitting SGLMMs. We introduce our method in the following section.

2.2. Maximum Likelihood Algorithm

We calculate the first-order Taylor expansion of y_i at μ_i and obtain $y_i \approx \mu_i + (\partial\mu_i/\partial\eta_i)[g^{-1}(y_i) - \eta_i]$, leading to the definition of the ith component of the working response vector $\boldsymbol{z}_{\boldsymbol{\beta}\boldsymbol{\gamma}}$ as

$$z_{i\boldsymbol{\beta}\boldsymbol{\gamma}} = \eta_i + (y_i - \mu_i)(\partial\eta_i/\partial\mu_i), \tag{9}$$

where η_i is the ith component of $\boldsymbol{\eta}$. By (1), we have $\mathrm{E}(z_{i\boldsymbol{\beta}\boldsymbol{\gamma}}|\boldsymbol{\gamma}) = \eta_i$ and $\mathrm{V}(z_{i\boldsymbol{\beta}\boldsymbol{\gamma}}|\boldsymbol{\gamma}) = (\partial\eta_i/\partial\mu_i)^2 b''(\theta_i)$, leading to the definition of the working weight matrix as

$$\mathbf{W}_{\boldsymbol{\beta}\boldsymbol{\gamma}} = \mathrm{diag}(w_{1\boldsymbol{\beta}\boldsymbol{\gamma}}, \ldots, w_{n\boldsymbol{\beta}\boldsymbol{\gamma}}), \tag{10}$$

where $w_{i\boldsymbol{\beta\gamma}}^{-1} = (\partial\eta_i/\partial\mu_i)^2 b''(\theta_i)$. Using $\boldsymbol{\eta} = \mathrm{E}(\boldsymbol{z}_{\boldsymbol{\beta\gamma}}|\gamma) = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\gamma}$ and $\mathrm{cov}(\boldsymbol{z}_{\boldsymbol{\beta\gamma}}|\boldsymbol{\gamma}) = \mathbf{W}_{\boldsymbol{\beta\gamma}}^{-1}$, we obtain the initial working model by treating $\boldsymbol{z}_{\boldsymbol{\beta\gamma}}$ as a normal random vector. Although the true distribution of $\boldsymbol{z}_{\boldsymbol{\beta\gamma}}$ is not normal, we can still use the normal likelihood to estimate the parameters because this is a consistent method (Zhang 2019). A difficulty is that both $\boldsymbol{z}_{\boldsymbol{\beta\gamma}}$ and $\mathbf{W}_{\boldsymbol{\beta\gamma}}$ depend on $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. We address this issue by a modification. Note that $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ appear in three places of the initial working model: the linear component $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\gamma}$, the working response vector $\boldsymbol{z}_{\boldsymbol{\beta\gamma}}$, and the working weight matrix $\mathbf{W}_{\boldsymbol{\beta\gamma}}$. We propose our modification by treating $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ as fixed in $\boldsymbol{z}_{\boldsymbol{\beta\gamma}}$ and $\mathbf{W}_{\boldsymbol{\beta\gamma}}$. We replace $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ contained by $\boldsymbol{z}_{\boldsymbol{\beta\gamma}}$ and $\mathbf{W}_{\boldsymbol{\beta\gamma}}$ with $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\gamma}}$, respectively, in our notations. We obtain the working model in our PM as

$$\boldsymbol{z}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\gamma} + \boldsymbol{\epsilon}, \tag{11}$$

where the distribution of $\boldsymbol{\gamma}$ is given by (4), $\mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0}$, and $\mathrm{cov}(\boldsymbol{\epsilon}) = \mathbf{W}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}}^{-1}$. We assume that $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\gamma}}$ have been derived by the previous step of the iterations, meaning that both $\boldsymbol{z}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}}$ and $\mathbf{W}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}}$ are assumed to be known in fitting (11) in the current iteration. We show that this is appropriate.

Our method is identical to the h-likelihood approach when $\boldsymbol{\gamma}$ is observed. We use it to compute the MLEs of $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ given $\boldsymbol{\gamma}$ of (11) defined as

$$(\hat{\boldsymbol{\beta}}_{h,\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}}^{\top}, \hat{\boldsymbol{\omega}}_{h,\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}}^{\top})^{\top} = \underset{\boldsymbol{\beta},\boldsymbol{\omega}}{\mathrm{argmax}}\, \ell_{h,\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}}(\boldsymbol{\beta},\boldsymbol{\omega}), \tag{12}$$

where $\ell_{h,\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}}(\boldsymbol{\beta},\boldsymbol{\omega})$ is the h-loglikelihood function of (11) derived by assuming $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{W}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}}^{-1})$. Although $\hat{\boldsymbol{\beta}}_{h,\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}}$ and $\hat{\boldsymbol{\omega}}_{h,\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}}$ are different from $\hat{\boldsymbol{\beta}}_{h,\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\omega}}_{h,\boldsymbol{\gamma}}$ given by (6) in general, respectively, they can be the identical in special cases.

Theorem 1. If the root of the score function of $\tilde{\ell}_{h,\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}}(\boldsymbol{\beta},\boldsymbol{\omega})$ as a function of $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ is unique for any fixed $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\gamma}}$, then $\hat{\boldsymbol{\beta}}_{h,\gamma} = \hat{\boldsymbol{\beta}}_{h,\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}}$ and $\hat{\boldsymbol{\omega}}_{h,\boldsymbol{\gamma}} = \hat{\boldsymbol{\omega}}_{h,\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}}$ when $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{h,\boldsymbol{\gamma}}$ and $\tilde{\boldsymbol{\gamma}} = \boldsymbol{\gamma}$.

Theorem 1 points out that $(\hat{\boldsymbol{\beta}}_{h,\boldsymbol{\gamma}}^{\top}, \hat{\boldsymbol{\omega}}_{h,\boldsymbol{\gamma}}^{\top})^{\top}$ given by (6) is a stationary point of the numerical algorithm formulated by (12). This provides a method to compute the MLEs of $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ under the h-likelihood approach. Based on initial guesses of $\hat{\boldsymbol{\beta}}_{h,\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\omega}}_{h,\boldsymbol{\gamma}}$, we calculate $\hat{\boldsymbol{\beta}}_{h,\tilde{\boldsymbol{\beta}}\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\omega}}_{h,\tilde{\boldsymbol{\beta}}\boldsymbol{\gamma}}$ by (12). We treat them as the next guesses of $\hat{\boldsymbol{\beta}}_{h,\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\omega}}_{h,\boldsymbol{\gamma}}$. By iterating the procedure, we obtain the exact values of $\hat{\boldsymbol{\beta}}_{h,\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\omega}}_{h,\boldsymbol{\gamma}}$.

Our interest is the likelihood approach but not the h-likelihood approach. We modify the algorithm such that it can be used to compute $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\omega}}$ given by (8) under the likelihood approach. The idea is to predict $\boldsymbol{\gamma}$. In particular, we use the likelihood function of (11) as

$$L_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}}(\boldsymbol{\beta},\boldsymbol{\omega}) = (2\pi)^{-\frac{n}{2}}[\det(\mathbf{V}_{\boldsymbol{\omega},\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}})]^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{z}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}}^{\top} - \mathbf{X}\boldsymbol{\beta})^{\top}\mathbf{V}_{\boldsymbol{\omega},\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}}^{-1}(\boldsymbol{z}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}} - \mathbf{X}\boldsymbol{\beta})\right\}, \tag{13}$$

where $\mathbf{V}_{\boldsymbol{\omega},\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}} = \Sigma_{\boldsymbol{\omega}} + \mathbf{W}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}}^{-1}$, which is derived by integrating out $\boldsymbol{\gamma}$ from the h-loglikelihood. Then, we obtain a modification of (12) as

$$(\hat{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}}^{\top}, \hat{\boldsymbol{\omega}}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}}^{\top})^{\top} = \underset{\boldsymbol{\beta},\boldsymbol{\omega}}{\mathrm{argmax}}\, \ell_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}}(\boldsymbol{\beta},\boldsymbol{\omega}), \tag{14}$$

where $\ell_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}}(\boldsymbol{\beta},\boldsymbol{\omega}) = \log L_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}}(\boldsymbol{\beta},\boldsymbol{\omega})$. Although $\hat{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}}$ and $\hat{\boldsymbol{\omega}}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}}$ given by (14) may be different from the true MLEs $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\omega}}$ given by (8), the differences can be ignored asymptotically. To show that this is appropriate, we connect our idea based on (14) to the EM algorithm at the beginning.

Theorem 2. (Connection to the EM). If $\tilde{\boldsymbol{\gamma}}$ satisfies $\log f_{h,\boldsymbol{\beta\omega}}(\boldsymbol{y},\tilde{\boldsymbol{\gamma}}) = \mathrm{E}_{\boldsymbol{\beta\omega}}[\log f_{h,\boldsymbol{\beta\omega}}(\boldsymbol{y},\boldsymbol{\gamma})|\boldsymbol{y}]$, meaning that it is provided by the E-step of the EM algorithm in the iterations, then the final answers of $\hat{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}}$ and $\hat{\boldsymbol{\omega}}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}}$ given by (14) are identical to the true MLE given by (8).

Theorem 2 points out that the EM-algorithm can be obtained by an appropriate formulation of $\tilde{\boldsymbol{\gamma}}$. Because $f_{c,\boldsymbol{\gamma\omega}}(\boldsymbol{\gamma}|\boldsymbol{y})$ contains HDIIs, it is hard to be used in predicting $\boldsymbol{\gamma}$. To overcome the difficulty, we predict γ based on (11) because it is a normal model and the prediction does not involve any HDIIs. We obtain a matrix expression for the prediction, leading to the PM algorithm. Thus, we can treat the proposed PM approach as a modification of the EM approach. Using this idea, we obtain Algorithm 1.

Algorithm 1 The Prediction-Maximization (PM) algorithm for MLEs of SGLMMs

Input: Data from the SGLMM defined by (1), (3), and (4)
Output: MLEs and their variance-covariance matrix
Initialization
1: Let $\boldsymbol{z}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}}$ and $\mathbf{W}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}}$ be the same as those adopted by the traditional IRWLS
Begin Iteration
2: M-Step: Compute $\hat{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}}$ and $\hat{\boldsymbol{\omega}}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}}$ defined by (14) using Appendix A
3: P-Step: Predict $\boldsymbol{\gamma}$ by $\hat{\boldsymbol{\gamma}}_{\hat{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}}\hat{\boldsymbol{\omega}}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}}}$ using Appendix B
4: $\tilde{\boldsymbol{\beta}} \leftarrow \hat{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}}$ and $\tilde{\boldsymbol{\gamma}} \leftarrow \hat{\boldsymbol{\gamma}}_{\hat{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}}\hat{\boldsymbol{\omega}}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}}}$
5: Compute $\boldsymbol{z}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}}$ and $\mathbf{W}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}}$ by (9) and (10), respectively
End Iteration
6: Compute the variance-covariance matrix of the MLEs using Appendix C
7: Output

Algorithm 1 has two stages. The first stage is given by Step 1. It provides the initial working responses and weights of (11) without the need of initial guesses of $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\gamma}}$. This is emphasized by a few examples to be introduced in Section 2.3. Although we use the same initialization of the traditional IRWLS, the final solution is not sensitive to that. The second stage is the iterations. It starts from the initial $\boldsymbol{z}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}}$ and $\mathbf{W}_{\tilde{\boldsymbol{\beta}}\boldsymbol{\gamma}}$ given by the first stage. The final output is the approximation of the MLEs of $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$. To implement Algorithm 1, we need the prediction and maximization step. Therefore, we call it the PM algorithm. The

maximization step is given by Step 2. It is completely formulated by MLE of SLMMs for normal responses. The prediction step is given by Step 3. It is carried out by matrix operations. None of them involve any numerical evaluations of the HDIIs. Therefore, HDIIs are not an issue in the PM.

The difference between Algorithm 1 and the algorithm proposed by Zhang (2021) is the mathematical formulations for the M-step and the P-step. The dispersion parameter is not contained in the iterations of the proposed PM, but it is contained in the iterations of the IRWLSR. This is not critical in GLMMs for longitudinal count data, but it is critical in SGLMMs for spatial count data. Thus, the dispersion should be removed from the iterations, leading to Algorithm 1. We next show that the estimators provided by our approach and the EM approach are asymptotically negligible, implying that the PM approach is asymptotically equivalent to the true MLE approach.

Theorem 3. If $\alpha < 1/2$, then $n^{\alpha}(\hat{\boldsymbol{\beta}}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}} - \hat{\boldsymbol{\beta}}_{h,\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}}) \xrightarrow{P} \mathbf{0}$ and $n^{\alpha}(\hat{\boldsymbol{\omega}}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}} - \hat{\boldsymbol{\omega}}_{h,\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}}) \xrightarrow{P} \mathbf{0}$ as $n \to \infty$ under (11).

Theorem 3 shows that the difference between the estimators given by (12) and (14) is asymptotically negligible, meaning that we can use (14) to replace (12). This removes numerical issues for HDIIs given by the right-hand side of (7), meaning that we only need to provide a formulation to $\boldsymbol{\gamma}$ by (11), which can be easily handled. We put the details in Appendices A and B.

We still have two theoretical issues to be addressed. The first is the difference between the final answers given by Algorithm 1 and $\hat{\boldsymbol{\beta}}_{h,\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\omega}}_{h,\boldsymbol{\gamma}}$ given by the h-likelihood approach. The second is the derivation of variance-covariance matrix for the estimator provided by Algorithm 1. We investigate both issues under the framework of increasing domain asymptotic theory. We do not expect that the estimators would be consistent under the fixed (i.e., infill) domain asymptotics due to the work of Zhang (2004), as the article points out that the exact MLE is not consistent either. For the first issue, we compare Algorithm 1 with the traditional IRWLS. In particular, let $\boldsymbol{\beta}^{(t)}$ and $\boldsymbol{\omega}^{(t)}$ be the tth iterated values of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\omega}}$ given by Algorithm 1. Let $\boldsymbol{\beta}_h^{(t)}$ and $\boldsymbol{\omega}_h^{(t)}$ be the tth iterated values of $\hat{\boldsymbol{\beta}}_{h,\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\omega}}_{h,\boldsymbol{\gamma}}$ given by the traditional IRWLS in the h-likelihood approach. The initial working responses and weights in (11) are identical. We immediately conclude that $n^{\alpha}(\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}_h^{(0)}) \xrightarrow{P} 0$ and $n^{\alpha}(\boldsymbol{\omega}^{(0)} - \boldsymbol{\omega}_h^{(0)}) \xrightarrow{P} 0$ for any $\alpha < 1/2$ by Theorem 3. The working responses and weights are difference in Algorithm 1 for the likelihood approach and the traditional IRWLS for the h-likelihood approach. We evaluate this and summarize our finding in the following theorem.

Theorem 4. Assume that

(i) the domain of $(\boldsymbol{\beta}^{\top}, \boldsymbol{\omega}^{\top})^{\top}$ is compact, and the true $(\boldsymbol{\beta}^{\top}, \boldsymbol{\omega}^{\top})^{\top}$, denoted by $(\boldsymbol{\beta}_0^{\top}, \boldsymbol{\omega}_0^{\top})^{\top}$, belongs to the interior of the domain;

(ii) there exists a function $K(\boldsymbol{y})$ such that $\mathrm{E}|K(\boldsymbol{y})| < \infty$ and

$$n^{-1}\{\log \ell_{h,\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\omega}}}(\boldsymbol{\beta}, \boldsymbol{\omega}) - \log \ell_{h,\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\omega}}}(\boldsymbol{\beta}_0, \boldsymbol{\omega}_0)\} \le K(\boldsymbol{y})$$

for all $\boldsymbol{y}$, $\mathbf{X}$, $\tilde{\boldsymbol{\beta}}$, $\boldsymbol{\beta}$, $\tilde{\boldsymbol{\omega}}$, and $\boldsymbol{\omega}$ in their domains;

(iii) there exists a function $\psi_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\omega}}}(\boldsymbol{\beta}, \boldsymbol{\omega})$ not dependent on $\boldsymbol{y}$, $\mathbf{X}$, and $\boldsymbol{\gamma}$, such that

$$P\left\{\lim_{n\to\infty} \sup_{\boldsymbol{\theta},\boldsymbol{\omega}} \left|\frac{1}{n}\ell_{h,\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\omega}}}(\boldsymbol{\beta}, \boldsymbol{\omega}) - \psi_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\omega}}}(\boldsymbol{\beta}, \boldsymbol{\omega})\right| = 0\right\} = 1.$$

Then, $n^{\alpha}(\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}_h^{(t)}) \xrightarrow{P} 0$ and $n^{\alpha}(\boldsymbol{\omega}^{(t)} - \boldsymbol{\omega}_h^{(t)}) \xrightarrow{P} 0$ for any $\alpha < 1/2$.

Conditions (i) and (ii) are the usual conditions and for consistency and asymptotic normality of the MLE (Ferguson 1996, theorems 17, 18). Condition (iii) is modified from the usual conditions for the same problem (Ferguson 1996, theorem 16). Condition (iii) is usually satisfied when strong mixing conditions for weak dependency are satisfies (Rosenblatt 1984). It is an appropriate assumption under the increasing domain asymptotics but not under the fixed domain asymptotics. Theorem 4 can only be applied to asymptotic normality under the framework of the increasing domain asymptotics.

The number of iterations in Algorithm 1 is usually small, because it is similar to the Newton-Raphson algorithm. For instance, the `glm` function of R chooses 25 as the maximum number of iterations. It stops the computation if the algorithm does not converge within 25 iterations. Therefore, we only need to study the properties of our method with a bounded number of iterations. In this case, the difference between the results given by our method and the results given by the h-likelihood method vanishes as $n \to \infty$, leading to consistency of our method. This is Theorem 4. By the connection between the two methods, we derive the variance-covariance matrix of our estimator. We only state the theorem. The detailed formulation is displayed in Appendix C, supplementary materials.

Corollary 1. Assume that Conditions (i), (ii), and (iii) of Theorem 4 hold. If (iv) there exists $\alpha > 0$ such that $\sum_{i=1}^{n} E\{[y_i - \mathrm{E}(y_i)]^{2+\alpha}\}/[\sum_{i=1}^{n} \mathrm{V}(y_i)]^{2+\alpha} \to 0$ as $n \to \infty$, then the estimator of $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ provided by the PM, denoted by $\hat{\boldsymbol{\beta}}_{\mathrm{PM}}$ and $\hat{\boldsymbol{\omega}}_{\mathrm{PM}}$, respectively, are $\sqrt{n}$-consistent, and their variance-covariance matrix is the inverse of the Fisher information provided by the likelihood function of (11).

Condition (iv) is the Lyapunov condition, which implies the Linderberg-Feller condition for derivation of asymptotic normality. The Lyapunov condition is the usual assumption for asymptotic normality for dependent data. It generally holds under the strong mixing condition. Therefore, Corollary 1 provides the asymptotic normality of the estimators under the framework of the increasing domain asymptotics. This can induce that the estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ provided by the PM is asymptotically equivalent to the true MLEs, which means that $\sqrt{n}$ times the difference between the estimator provided by the PM and the true MLE goes to 0 in probability as $n \to \infty$. We summarize this by the following Corollary.

Corollary 2. If Conditions (i), (ii), and (iii) of Theorem 4 and Condition (iv) of Corollary 1 hold, then $\sqrt{n}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\mathrm{PM}}) \xrightarrow{P} 0$ and $\sqrt{n}(\hat{\boldsymbol{\omega}} - \hat{\boldsymbol{\omega}}_{\mathrm{PM}}) \xrightarrow{P} 0$.

2.3. Examples

We compare our method to three well-known SGLMMs: the generalized geostatistical (GGEO) (Diggle, Tawn, and Moyeed 1998), the generalized conditional autoregressive (GCAR) (Cressie and Chan 1989), and the generalized spatial autoregression (GSAR) (Ork 1975) models for count responses. All of them can be applied to spatiotemproal data. It is enough to display our method for spatial data only. We study the cases when the count responses are binomial or Poisson due to their popularity. We provide the initialization of $\boldsymbol{z}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}}$ and $\mathbf{W}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\gamma}}}$ for Step 1 of Algorithm 1.

Example 1 (GGEO). Let y_i and $\boldsymbol{x}_i$ be response and explanatory variables collected at the ith site in a spatial or spatiotemporal domain. The GGEO model is jointly defined by (3) and (4) with

$$\Sigma_{\boldsymbol{\omega}} = (c_{\boldsymbol{\omega}}(d_{ij})), i, j = 1, \ldots, n, \tag{15}$$

where d_{ij} is the distance between sites i and j. For purely spatial data, d_{ij} is chosen as the Euclidean distance. For spatiotemporal data, d_{ij} is defined by a bivariate vector with the first and the second components to be the spatial and temporal distances, respectively. Because $\Sigma_{\boldsymbol{\omega}}$ must be positive definite, parametric methods are used to specify it. Examples include the Matérn (Matérn 1986) and the generalized Cauchy (Gneiting and Schlather 2004) families in purely spatial case, and a number of nonseparable families in the spatiotemporal case (Cressie and Huang 1998; Gneiting 2002; Stein 2005). Because the PM for SGLMMs and fitting methods for SLMMs are developed independently, it is enough to display it based on the well-known Matérn covariance family for spatial data only.

The Matérn covariance family specifies the variance-covariance matrix between $\boldsymbol{\gamma}$ as

$$c_{\boldsymbol{\omega}}(d_{ij}) = \omega_1 \frac{(\omega_2 d_{ij})^{\omega_3}}{2^{\omega_3 - 1}\Gamma(\omega_3)} K_{\omega_3}(\omega_2 d_{ij}) = \omega_1 M_{\omega_3}(\omega_2 d_{ij}) \tag{16}$$

with $\boldsymbol{\omega} = (\omega_1, \omega_2, \omega_3)^\top$, where $K_{\delta_3}(\cdot)$ is the modified Bessel function of the second kind, and ω_1, $1/\omega_2$, and ω_3 are variance, scale, and smoothness parameters, respectively. The Matérn family is isotropic in space. It becomes the exponential covariance function when $\delta_3 = 0.5$. The family was first proposed by Matérn (1986) and has received more attention since some theoretical work of Handcock and Stein (1993) and Stein (1999).

To fit the GGEO model, we only need an algorithm for the GEO model for normal responses. The formulation of GEO is identical to (19) in Appendix A, supplementary materials with the variance-covariance matrix for $\boldsymbol{\gamma}$ specified by (16). We need the first and second-order partial derivatives of $\Sigma_{\boldsymbol{\omega}}$ with respect to $\boldsymbol{\omega}$. They have been worked out in the literature. For instance, for ω_2, we have

$$\frac{\partial c_{\boldsymbol{\omega}}(d_{ij})}{\partial \omega_2} = \frac{2\omega_1\omega_3}{\omega_2}[M_{\omega_3}(\omega_2 d_{ij}) - M_{\omega_3+1}(\omega_2 d_{ij})]$$

and

$$\frac{\partial^2 c_{\boldsymbol{\omega}}(d_{ij})}{\partial \omega_2^2} = \omega_1 \left(\frac{4\omega_3^2 - 2\omega_3}{\omega_2^2} + d_{ij}^2 \right) M_{\omega_3}(\omega_2 d_{ij}) + \frac{\omega_2(2\omega_3 - 4\omega_3^2)}{\omega_2^2} M_{\omega_3+1}(\omega_2 d_{ij}).$$

The derivation of the first and second-order partial derivatives of $c_{\boldsymbol{\omega}}(d_{ij})$ with respect to ω_1 is trivial. It is omitted. The partial derivatives of $c_{\boldsymbol{\omega}}(d_{ij})$ with respect to ω_3 are rarely used in the computation because the MLE of ω_3 is often obtained by a golden section search. Therefore, we can use Appendix A, supplementary materials to fit the GEO model with Matérn covariance functions. Combining it with Appendix B, supplementary materials, we obtain the PM algorithm for the GGEO model with the Matérn covariance function for count responses.

We then implement our method to GGEO for binomial or Poisson data. For binomial data, we assume that $y_i \sim Bin(m_i, \pi_i)$ with $\log[\pi_i/(1-\pi_i)]$ modeled by (3). At the beginning (i.e., Step 1) of Algorithm 1, we choose the ith components of $\boldsymbol{z}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\omega}}}$ as $z_i^{(0)} = \log[(y_i+0.5)/(m_i-y_i+0.5)]$ and the ith diagonal entry of $\mathbf{W}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\omega}}}$ as $w_i^{(0)} = m_i(y_i+0.5)(m_i-y_i+0.5)/(m_i+1)^2$. Let $\boldsymbol{\beta}^{(t)}$ be the estimated value of $\boldsymbol{\beta}$ and $\gamma_i^{(t)}$ be the predicted value of γ_i in the tth iteration. We predict η_i by $\eta_i^{(t)} = \boldsymbol{x}_i^\top \boldsymbol{\beta}^{(t)} + \gamma_i^{(t)}$, leading to the $(t+1)$th iterated value of the ith component of $\boldsymbol{z}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\omega}}}$ as $z_i^{(t)} = \eta_i^{(t)} + (y_i - m_i\pi_i^{(t)})/[m_i\pi_i^{(t)}(1-\pi_i^{(t)})]$ and the $(t+1)$th iterated value of the ith diagonal entry of $\mathbf{W}_{\tilde{\boldsymbol{\beta}}\tilde{\boldsymbol{\omega}}}$ as $w_i^{(t)} = m_i\pi_i^{(t)}(1-\pi_i^{(t)})$, where $\pi_i^{(t)} = e^{\eta_i^{(t)}}/(1+e^{\eta_i^{(t)}})$. We can carry out the next iteration. The entire computation does not need initial guesses of $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\omega}}$.

For Poisson data, we assume that $y_i \sim \mathcal{P}(\lambda_i)$ with $\log \lambda_i$ modeled by (3). At the beginning (i.e., Step 1) of Algorithm 1, we choose $z_i^{(0)} = \log(y_i + 0.5)$ and $w_i^{(0)} = y_i + 0.5$. Let $\boldsymbol{\beta}^{(t)}$ and $\gamma_i^{(t)}$ be the tth iterated values of $\hat{\boldsymbol{\beta}}$ and γ_i, respectively. We have $z_i^{(t)} = \eta_i^{(t)} + (y_i - e^{\eta_i^{(t)}})/e^{\eta_i^{(t)}}$ and $w_i = e^{\eta_i^{(t)}}$ with $\eta_i^{(t)} = \boldsymbol{x}_i^\top \boldsymbol{\beta}^{(t)} + \gamma_i^{(t)}$, leading to the next iteration. The entire computation does not need initial guesses of $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\omega}}$ either.

Example 2 (GCAR). The GCAR model is proposed for aggregated data (also called lattice data) with at-risk population in spatial statistics. Suppose that a spatial or spatiotemporal domain is partitioned into K units. Let y_i and $\boldsymbol{x}_i$ be the response and explanatory variables associated with at-risk population size n_i from the ith units. The GCAR model is jointly defined by (3) and (4) with $\Sigma_{\boldsymbol{\omega}} = \omega_1(\mathbf{I} - \omega_2\mathbf{H})^{-1}\mathbf{D}$, where $\boldsymbol{\omega} = (\omega_1, \omega_2)^\top$, ω_1 is the variance parameter, and ω_2 is the spatial autocorrelation parameter. To ensure $\Sigma_{\boldsymbol{\omega}}$ symmetric, we need to choose $\mathbf{H} = (h_{ij})$ with $h_{ij} = a_{ij}(n_j/n_i)^{1/2}$ and $\mathbf{D} = \text{diag}(n_1^{-1}, \ldots, n_K^{-1})$, where a_{ij} with $a_{ii} = 0$ is the (i,j)th entry of a symmetric matrix (Cressie and Chan 1989).

To fit the GCAR model, we only need a method for the CAR model for normal responses. The format of the model is identical to (19) in Appendix A, supplementary materials with the variance-covariance matrix for $\boldsymbol{\gamma}$ specified by $\Sigma_{\boldsymbol{\omega}}$. We need the first and second-order partial derivatives of $\Sigma_{\boldsymbol{\omega}}$ with respect to $\boldsymbol{\omega}$. The first-order partial derivatives are $\partial\Sigma_{\boldsymbol{\omega}}/\partial\omega_1 =$

$(\mathbf{I}-\omega_2\mathbf{H})^{-1}\mathbf{D}$ and $\partial\Sigma_{\boldsymbol{\omega}}/\partial\omega_2 = \omega_1(\mathbf{I}-\omega_2\mathbf{H})^{-1}\mathbf{H}(\mathbf{I}-\omega_2\mathbf{H})^{-1}\mathbf{D}$. The second-order partial derivatives are $\partial^2\Sigma_{\boldsymbol{\omega}}/\partial\omega_1^2 = \mathbf{0}$, $\partial^2\Sigma_{\boldsymbol{\omega}}/\partial\omega_2^2 = 2\omega_1(\mathbf{I}-\omega_2\mathbf{H})^{-1}\mathbf{H}(\mathbf{I}-\omega_2\mathbf{H})^{-1}\mathbf{H}(\mathbf{I}-\omega_2\mathbf{H})^{-1}\mathbf{D}$, and $\partial^2\Sigma_{\boldsymbol{\omega}}/\partial\omega_1\partial\omega_2 = (\mathbf{I}-\omega_2\mathbf{H})^{-1}\mathbf{H}(\mathbf{I}-\omega_2\mathbf{H})^{-1}\mathbf{D}$. Therefore, we can use Appendix A, supplementary materials to compute the MLEs of $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ in the CAR model. Combine it with Appendix B, supplementary materials. We obtain the PM algorithm for the GCAR model. The specification of the PM to the GCAR model for binomial or Poisson responses is similar to that we have just displayed for the GGEO model. We can use the same $z_i^{(0)}$ and $w_i^{(0)}$ of Example 1 for the Poisson or the binomial data at the beginning (i.e., Step 1) of Algorithm 1. We omit it.

Example 3 (GSAR). The GSAR model is also proposed for aggregated data with at-risk population. It is jointly defined by (3) with the distribution of $\boldsymbol{\gamma}$ given by $\boldsymbol{\gamma} = \omega_2\mathbf{A}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \omega_1\mathbf{I})$, and $\mathbf{A}$ is a spatial weight matrix. Then, (4) becomes $\Sigma_{\boldsymbol{\omega}} = \omega_1(\mathbf{I} - \omega_2\mathbf{A})^{-1}(\mathbf{I} - \omega_2\mathbf{A}^\top)^{-1}$, where $\boldsymbol{\omega} = (\omega_1, \omega_2)^\top$. To fit the GCAR, we only need a method for the CAR model for normal responses. We need the first and the second-order partial derivatives of $\Sigma_{\boldsymbol{\omega}}$ with respect to $\boldsymbol{\omega}$. They are provided by matrix expression. We use Appendices A and B to calculate the MLEs of $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ in the GSAR model. We can use the same $z_i^{(0)}$ and $w_i^{(0)}$ of Example 1 for the Poisson or the binomial data at the beginning (i.e., Step 1) of Algorithm 1. The detailed specification of our method to the GSAR for binomial or Poisson responses is also omitted.

For spatiotemporal data, the total number of observations is usually large. Large matrix approaches are often needed. To implement the PM to spatiotemporal data, we only need an efficient algorithm for normal responses. After it is derived, we can easily modify it for count responses. Therefore, our method provides a bridge to connect mixed effects models for normal responses with those for nonnormal responses.

3. Simulation

We evaluated our proposed method (i.e., the PM) with the comparison to seven previous methods for the MLEs of SGLMMs. The seven previous methods were the IRWLSR (Zhang 2021), the PQL (Breslow and Clayton 1993), the LA (Evangelou, Zhu, and Smith 2011), the PrevMap (Girogi and Diggle 2017), the geoCount (Jing and Oliveira 2015), the geoRglm (Christensen and Ribeiro 2002), and the INLA (Rue, Martino, and Chopin 2009). We classified the eight methods into two groups. The likelihood-based group contained the PM, the IRWLSR, the PQL, the LA, and the PrevMap methods. We treated the PQL as a restricted maximum likelihood (REML) method, because it is identical to the REML method for the SGLMM if the response is normal. We treated the PM, the IRWLSR, and the LA methods as the ML methods, because they are identical to the ML method for the SGLMM if the response is normal. The PrevMap was also treated as one of the ML methods because it uses the Monte Carlo approach to approximate the likelihood function with the parameter estimators derived by maximizing the simulated likelihood function. The Bayesian group contained the geoCount, the geoRglm, and the INLA methods. Both the geoCount and the geoRglm use MCMC algorithms to sample the posterior distributions of $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$, and $\boldsymbol{\omega}$ by the Metropolis-Hasting approach with the Bayesian estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ derived by the posterior distributions. Following Jing and Oliveira (2015), we generated 2200 MCMC samples in the geoCount and 50,000 in the geoRglm. Our simulation showed that they were big enough to obtain the stable posterior distributions from the MCMC. The computation of the geoCount was time-consuming if more samples were used. This was why the authors only used 2200 MCMC samples in the geoCount method. The INLA is an approximate Bayesian method. It incorporates the LA into MCMC for approximating the marginal distributions of parameters of interests. The approximate Bayesian method is carried out based on a mesh structure composed by less than 200 points in the domain. The INLA generated 2048 MCMC samples based on the mesh. This makes the INLA more computationally efficient than the usual MCMC algorithms.

We studied both increasing and fixed (i.e., infill) domain asymptotics problems. We focus on the comparison under the increasing domain asymptotics and simply mention that under the fixed domain asymptotics. In the increasing domain asymptotics problem, we randomly generated n spatial sites from square region $[0, \sqrt{n}]^2$. In the fixed domain asymptotics problem, we randomly generated the locations from the squared region $[0, 10]^2$. After the sites were derived, we calculated the Euclidean distance d_{ij} between the sites for $i, j = 1, \ldots, n$. We then generated data from the Poisson GGEO model with only the intercept in its linear component, leading to the SGLMM in our simulation as

$$\log \mu_i = \beta_0 + \gamma_i,\ i = 1, \ldots, n, \tag{17}$$

where $\Sigma_{\boldsymbol{\omega}}$ was defined by the Matérn family given by (16). We fixed $\omega_3 = 0.5$ in (17), such that the Matérn covariance function reduced to the exponential covariance function as

$$c_{\boldsymbol{\omega}}(d_{ij}) = \omega_1 e^{-\omega_2 d_{ij}}. \tag{18}$$

Therefore, we had $\boldsymbol{\omega} = (\omega_1, \omega_2)^\top$, where ω_1 was the variance parameter and $1/\omega_2$ was the scale parameter. We used fixed $\beta_0 = 5.0$, $\omega_1 = 0.5$, and $\omega_2 = 1.00$ and varied $n = 100, 200, 400$. We reparameterize ω_1 and ω_2 for nugget effect formulations by $\delta_1 = \omega_1/(1 + \omega_1)$ and $\delta_2 = \omega_2$. We evaluated the performance based on $\boldsymbol{\delta} = (\delta_1, \delta_2)$. Therefore, the variance parameter was $\omega_1 = \delta_1/(1 - \delta_1)$ and the scale parameter was $\omega_2 = \delta_2$.

Following Jing and Oliveira (2015), we used the proper half-t prior for ω_1 and the proper uniform prior for $1/\omega_2$ in the geoCount. We selected the bounds of the uniform prior to ensure that it contained the true ω_2. We tried different bounds and found that the results were similar. Then, we fixed the lower bound at 0.1 and the upper bound at 10. In geoRglm, still following Jing and Oliveira (2015), we used the proper scaled inverse Chi-squared prior for the nugget parameter and also the Uniform[0.1, 10] prior for the scale parameter. We also tried other bounds and found that the changes could be ignored. In INLA, we used the R code provided by Jing and Oliveira (2015) to generate the mesh. We tried different options in the generations and did not find any significant differences between the options.

We also examined many other priors for the influence of the prior distributions on the Bayesian methods. For the geoCount,

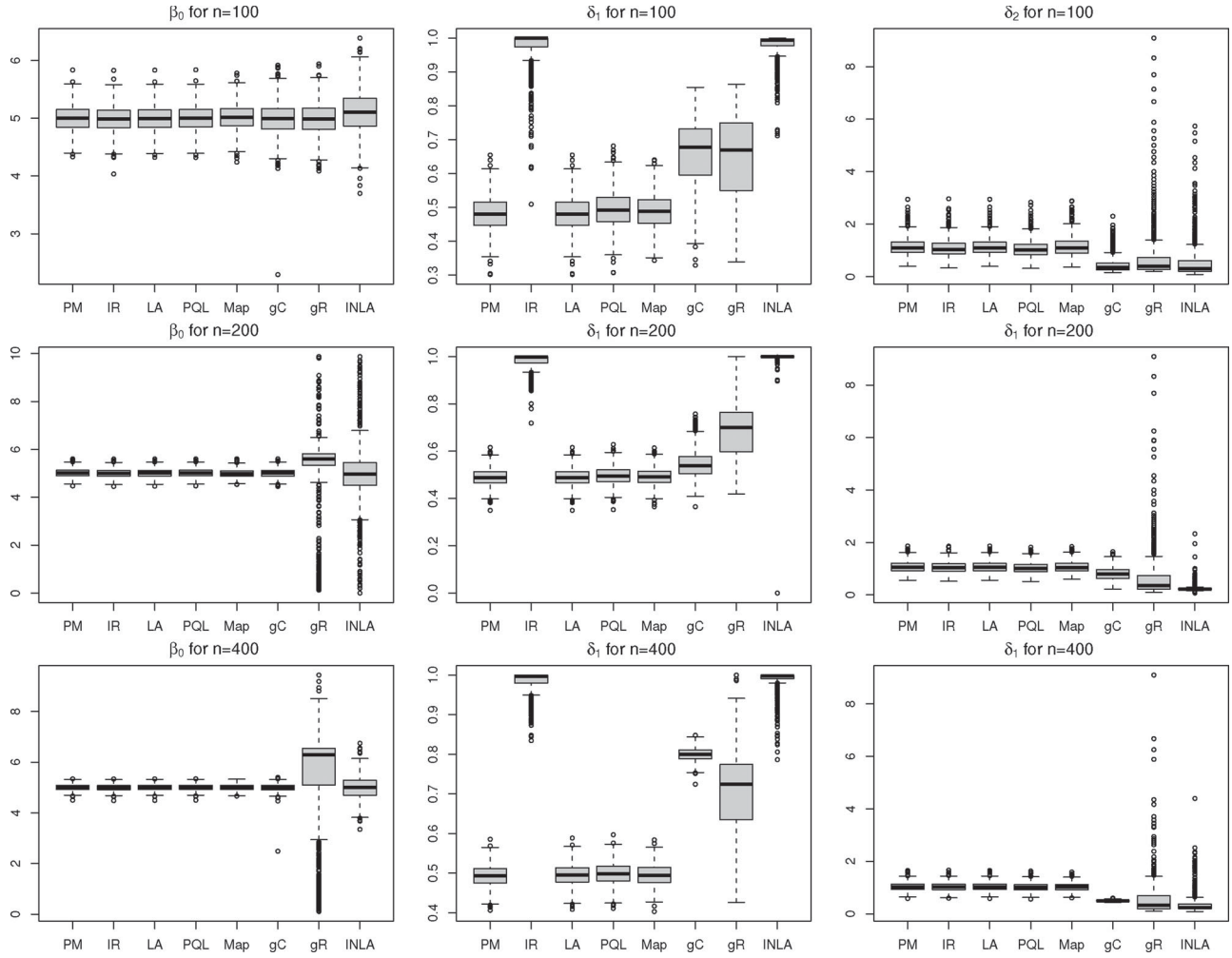

Figure 1. Boxplot for the sample distributions (obtained from 1000 replications) of the 8 estimators (from the left to the right) from the proposed PM approach with the comparison to the previous IRWLSR (IR), LA, PQL, PrevMap (Map), geoCount (gC), geoRglm (gR), and INLA approaches when data are generated from (17) with $\beta_0 = 5.0$, $\delta_1 = 0.5$, and $\delta_2 = 1.0$ in (18) for the increasing domain asymptotics problem.

we investigated the Inverse Gamma and Reciprocal priors for ω_1 and ω_2, respectively. We found that the influences of prior distributions on estimates of β_0 were all less than 1%, of ω_1 were mostly less than 5%, and of ω_2 were mostly less than 20%. For the geoRglm, we investigated the Flat, Normal, and Fixed priors for β_0, the Uniform, Scaled Inverse χ^2_2, Reciprocal, Fixed priors for ω_1, and the Uniform, Exponential, Fixed, Squared Reciprocal, Reciprocal priors for $1/\omega_2$, respectively. We found that the influences on estimates of β_1, ω_1, and ω_2 were all less than 1%. For the INLA, we investigated various options of prior distributions provided by the `INLA` package of R. We found that the influences on estimates of β_1 and ω_1 were all less than 1%, and of ω_2 were mostly less than 15%. The selections of the prior distributions did not significantly affect the performance of the Bayesian methods.

We generated 1000 datasets from (17) for each selected n. In each generated dataset, we implemented our proposed PM by Algorithm 1 and the PQL and the LA methods by the `spaMM` package of R. We implemented the PrevMap method by the `PrevMap` packages of R. We implemented the previous IRWLSR method proposed by Zhang (2021) by a modification of Algorithm 1 to incorporate the dispersion parameter. We implemented the geoCount, the geoRglm, and INLA methods by the `geoCount`, `geoRglm`, and `inla` packages of R.

We first compared the performance of the eight methods by looking at the sample distributions of the estimators. For each of the methods, we calculated the estimates of β_0, δ_1, and δ_2 using the generated datasets. We obtained 1000 estimated values. We plotted these and obtained the simulated sample distributions (Figure 1). We found that the sampling distributions of the IRWLSR, geoCount, geoRglm, and INLA were much wider than those of the PM, LA, PQL, and PrevMap, implying that the likelihood-based methods were more reliable than the MCMC methods. We found significant biases in the MCMC approaches when they were used to estimate the nugget and scale effects. The results of the IRWLSR were not stable either, because it contains the redundant dispersion parameter in its iterations.

We next compared the performance of the eight methods by looking at their root mean squares errors (RMSE) values. The RMSE values were obtained by the squared root of the average of the MSE values from the 1000 replications. The result showed that our proposed method was equally good as the PQL, LA, and PrevMap methods (Table 1). The performance of likelihood-based methods (i.e., the PM, the PA, the LA, and the PrevMap)

Table 1. Comparison of the proposed PM and the previous IRWLSR (using R for both) with the LA, PQL, PrevMap, geoCount, geoRglm, and INLA (using C++ for all the six) based on RMSE values obtained from simulations with 1000 replications generated by (17) when $\beta_0 = 5$, $\delta_1 = 0.5$, and $\delta_2 = 1.0$ in the increasing domain asymptotics problem.

	$n = 100$				$n = 200$				$n = 400$			
Method	Minutes	β_0	δ_1	δ_2	Minutes	β_0	δ_1	δ_2	Minutes	β_0	δ_1	δ_2
PM	0.012	0.222	0.054	0.324	0.037	0.168	0.038	0.210	0.149	0.123	0.027	0.145
IRWLS	0.063	0.227	0.477	0.345	0.186	0.169	0.483	0.225	0.790	0.124	0.486	0.155
LA	0.021	0.223	0.054	0.325	0.067	0.168	0.038	0.210	0.360	0.123	0.027	0.146
PQL	0.011	0.223	0.054	0.324	0.035	0.168	0.038	0.209	0.237	0.123	0.027	0.146
PrevMap	1.446	0.262	0.059	0.377	2.067	0.167	0.037	0.222	5.011	0.121	0.028	0.152
geoCount	0.367	0.281	0.185	0.626	1.190	0.172	0.074	0.317	5.644	0.153	0.300	0.498
geoRglm	0.111	0.269	0.190	1.653	0.464	18.623	0.208	3.118	3.177	149.576	0.226	3.588
INLA	0.084	0.190	0.452	0.520	0.374	0.369	0.496	0.647	2.030	0.773	0.500	0.678

was generally better than that of the Bayesian methods (i.e., the geoCount, the geoRglm, and INLA). We did not expect that the performance of the IRWLSR could be as good as the PM, the LA, the PQL, and the PrevMap because it contained the redundant dispersion parameter in the computation. As n became large, the RMSE of the estimator obtained from PM, the LA, and the PQL decreased. The rates were approximately inversely proportional to $\sqrt{n}$, indicating that the conclusion provided by Corollary 1 for the increasing domain asymptotics was precise. In addition, we investigated the same issue for the fixed domain asymptotics (not shown). We found that the RMSE did not significantly reduce with n, implying that Corollaries 1 and 2 cannot be applied to the fixed domain asymptotics.

We also evaluated the properties of the Fisher information provided by the PM. In particular, we calculated $s_{\mathrm{PM}}(\hat{\delta}_1)$, $s_{\mathrm{PM}}(\hat{\delta}_1)$, and $s_{\mathrm{PM}}(\hat{\delta}_2)$, the standard errors of the estimators of β_0, β_1, and β_2, by the Fisher information of the PM. We computed the 95% Wald confidence intervals for β_0, δ_1, and δ_2 by $\hat{\beta}_0 \pm 1.96 s_{\mathrm{PM}}(\hat{\beta}_0)$, $\hat{\delta}_1 \pm 1.96 s_{\mathrm{PM}}(\hat{\delta}_1)$, and $\hat{\delta}_2 \pm 1.96 s_{\mathrm{PM}}(\hat{\delta}_2)$, respectively. Theoretically, we need to choose $\hat{\beta}_0$, $\hat{\delta}_1$, $\hat{\delta}_2$ as the estimators of β_0, δ_1, and δ_2 provided by the PM only. To understand the impacts of the Fisher information more deeply, we extended the corresponding formulations by choosing those as the estimators provided by all of the eight methods, respectively. For each of the eight methods, we compared the coverage probabilities of the confidence intervals. The coverage probabilities were derived by the proportions for the confidence intervals to contain the true parameters (Table 2). Our result showed that the coverage probabilities of the confidence intervals provided by the PM, the LA, the PQL, and the PrevMap were all close to 95%, indicating that the standard errors provided by the PM can be extended to all of the three methods. Because we did not find this phenomenon in the IRWLS, geoCount, geoRglm, and INLA, we concluded that the standard errors provided by the PM cannot be extended to the four methods. Due to the bias of the MCMC methods, the coverage probabilities of the 95% confidence intervals were also low when the standard errors were calculated by their posterior samples.

We compared the difference between the proposed PM with the previous IRWLSR. In each iteration of the IRWLSR, it uses an SLMM for normal responses to update the estimates of parameters. The SLMM contains a dispersion parameter for the variance of random errors. The proposed PM removes the dispersion parameter in its iterations by assuming that the dispersion parameter is equal to 1. The dispersion parameter is not present in the SGLMM but it is used in the iterations of IRWLSR.

Table 2. Comparison of the coverage probabilities of the 95% confidence intervals provided by the estimators of the eight methods with the standard errors given by the Fisher Information of the PM from simulations with 1000 datasets generated by (17) when $\beta_0 = 5$, $\delta_1 = 0.5$, and $\delta_2 = 1.0$ in the increasing domain asymptotics problem.

	$n = 100$			$n = 200$			$n = 400$		
Method	β_0	δ_1	δ_2	β_0	δ_1	δ_2	β_0	δ_1	δ_2
PM	0.916	0.924	0.936	0.916	0.924	0.936	0.916	0.924	0.936
IRWLSR	0.910	0.001	0.941	0.910	0.001	0.941	0.910	0.001	0.941
LA	0.915	0.924	0.936	0.915	0.924	0.936	0.915	0.924	0.936
PQL	0.915	0.925	0.950	0.915	0.925	0.950	0.915	0.925	0.950
PrevMap	0.912	0.932	0.925	0.924	0.934	0.918	0.924	0.951	0.938
geoCount	0.865	0.245	0.473	0.865	0.245	0.473	0.865	0.245	0.473
geoRglm	0.864	0.316	0.000	0.864	0.316	0.000	0.864	0.316	0.000
INLA	0.842	0.000	0.666	0.842	0.000	0.666	0.842	0.000	0.666

This causes serious biases in the estimates of δ_1, leading to low coverage of the IRWLSR for δ_1. This issue is solved by the proposed PM.

We then studied the prediction performance. As the prediction of $\boldsymbol{\gamma}$ was available in the PM, we quickly obtained the predicted value of the responses by $\hat{y}_i = \exp(\hat{\beta}_0 + \hat{\gamma}_i)$, where $\hat{\gamma}_i$ was the predicted value of γ_i. We looked at two quantities. The first was the relative total variation (RTV) of the random effects not explained by the model defined as $\mathrm{NRTV} = \sum_{i=1}^{n}(\hat{\gamma}_i - \gamma_i)^2 / \sum_{i=1}^{n} \gamma_i^2$. The second was the deviance coefficient of determination for the reduction of deviance goodness-of-fit (GOF) statistic explained by the model. The formulation was $R^2 = (G_0 - G^2)/G_0^2$, where $G^2 = \sum_{i=1}^{2} y_i \log(y_i/\hat{y}_i)$ was the deviance GOF statistic for (17) and G_0^2 was that for the null model with β_0 only. For the PM, we found that all of the values of NRTV were less than 0.001 and all of the values of R^2 were greater than 0.99, indicating that the prediction was precise. To compare, we also used `spaMM` to calculate the prediction of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$. We found all of them were close to that provided by the PM. The differences were less than 0.1%. This also occurred in the PrevMap method. In addition, we implemented the same approach to evaluate the prediction performance of the MCMC methods. For all of the geoCount, the geoRglm, and the INLA, the values of NRTV were only slightly larger than 0.001, which was still small. All of the values of R^2 were still higher than 0.99, indicating that the prediction of the MCMC methods was still precise. Therefore, we conclude that although MCMC was not stable in estimation, it can still provide nice predictions of the random effects and the responses.

In the end, we compared the computational times of the eight methods. We obtained the computational times by the average

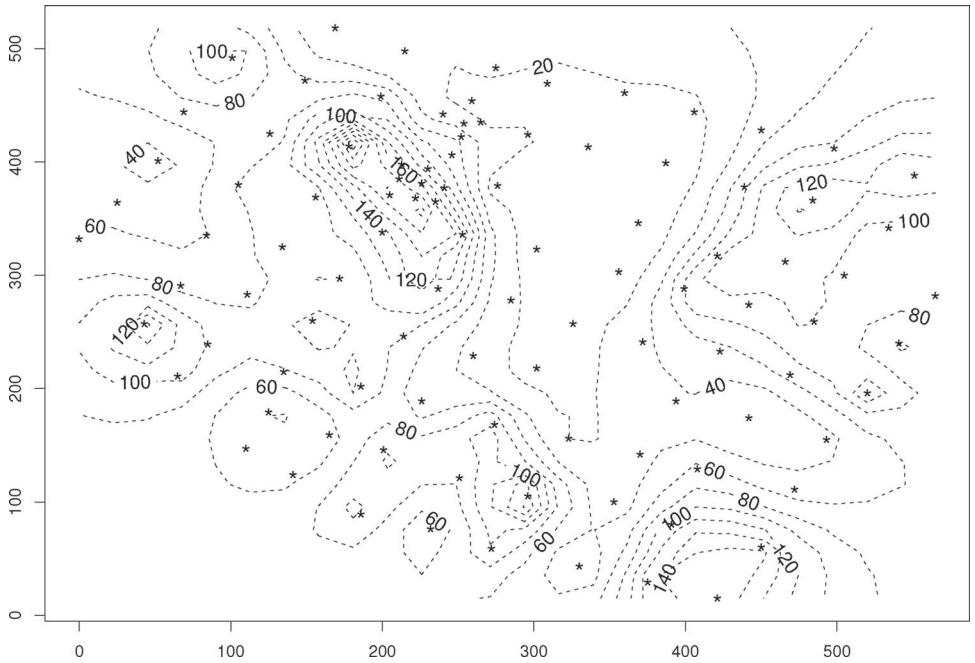

Figure 2. Locations of weed samples (marked by *) and the contour plot for the predicted weed counts given by the PM. We only display those given by the PM because they are almost identical to those given by the PQL and the LA.

minutes in the 1000 replications. Our results showed that the likelihood-based methods were generally more computationally efficient than the MCMC methods, and the geoCount was the slowest. Therefore, we conclude that the likelihood-based methods are more computationally efficient than the MCMC methods. The advantage of the PM is that we do not need any numerical evaluation of HDIIs. Thus, it is more convenient than its competitors. In the comparison, the implementations of the PM and the IRWLSR were based on their R code, respectively. The implementations of all of the other six methods were based on their C++ code, respectively, where R is only a platform. As it has previously pointed out that R can be 500 times slower than C++ (Aruoba and Fernandez-Villaverde 2015), it is enough for us to conclude that the proposed PM and the previous IRWLSR are more computationally efficient than the other six methods. Therefore, the PM has more computational advantages than its competitors.

4. Application

We implemented our method with the comparison of our competitors to the Weed dataset. The Weed dataset was previously analyzed by Guillot, Lorén, and Rudemo (2009) using the spatial Poisson-lognormal model, which was identical to (17). It was also previously analyzed by many other authors using various methods, including the Bayesian method adopted by Christensen and Waagepetersen (2002).

Because collecting exact weed counts in agricultural fields is expensive and time-consuming, statistical analysis methods are recommended for predicting weed counts to reduce the cost at an appropriate level. To fulfill the goal, the Weed dataset was collected at the Bjertorp farm located 58.26°N and 13.13°E in the south-west of Sweden. The size of the farm was about 30 hectares. The data were collected from 100 sites with exact counts of noncrop plants. The size of the sites was given by a 0.5m × 0.75m frame. The locations of the sites are displayed in Figure 2 (marked by *).

Table 3. Estimates of model parameters for (17) in the Weed dataset by the PM, IRWLSR, PQL, PrevMap, LA, geoCount, geoRglm, and INLA methods under (17) with $\Sigma_{\boldsymbol{\omega}}$ is given by (18).

	β_0	ω_1	$1/\omega_2$
PM	4.080	0.894	70.20
PQL	4.080	0.918	70.45
LA	4.069	0.917	70.44
PrevMap	4.068	0.920	70.18
geoCount	4.169	1.424	62.87
geoRglm	4.119	1.374	111.33
INLA	4.072	1.070	45.02

We analyzed the Weed dataset by the PM, the LA, the PQL, the PrevMap, the geoCount, the geoRglm, and the INLA under (17) with $\Sigma_{\boldsymbol{\omega}}$ given by (18) (Table 3). We excluded the IRWLSR method because it contains a redundant dispersion parameter in its iteration but it is not present in the model. We treated the LA as ML methods because it uses the approximate expression of the likelihood function to estimate the parameters. We also treated the PrevMap as an ML method because it uses Monte Carlo method to approximate the likelihood function. We treated the PrevMap as an approximate ML method because it uses the Monte Carlo approach to approximate the likelihood function. Therefore, we had four likelihood-based methods. They were the PM, the LA, the PQL, and the PrevMap. The remaining three, which were the geoCount, the geoRglm, and the INLA, were treated as Bayesian methods, because they used the MCMC.

The results given by the likelihood-based methods were close but the results given by the Bayesian methods were not. The estimates of the intercept parameter (i.e., β_0) in all of the seven

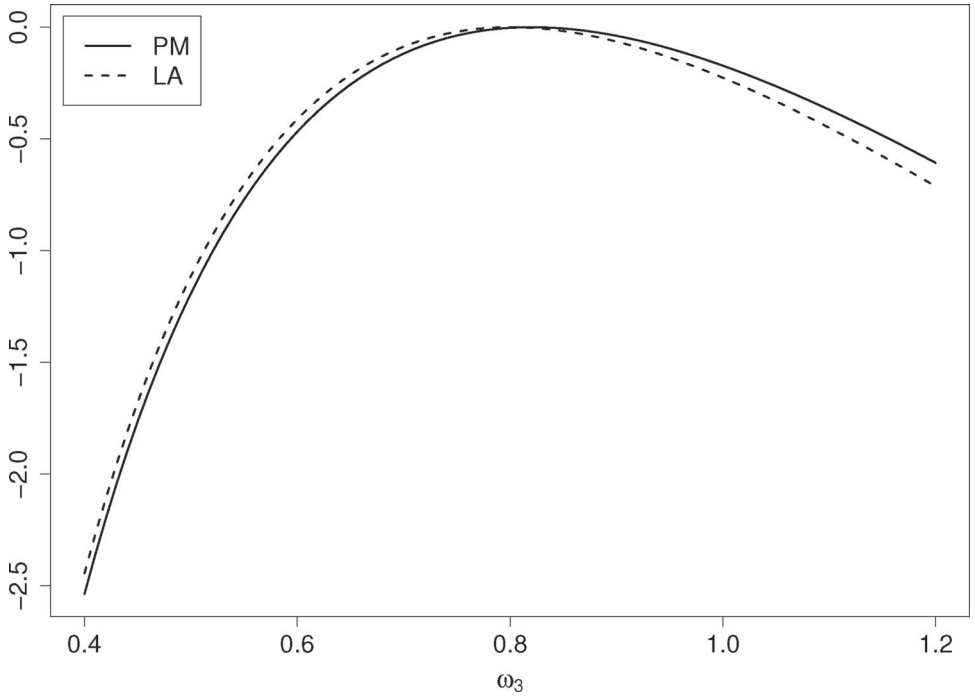

Figure 3. Plots of $2\{\ell_P(\omega_3) - \max[\ell_P(\omega_3)]\}$ in our proposed PM and the LA methods. Because the plot in the PQL is almost identical to that in the LA, it is not contained.

methods were close. The estimates of the variance parameter (i.e., ω_1) were a little bit different. Significant differences were found in the estimates of the scale parameter (i.e., $1/\omega_2$) in the Bayesian methods. The largest one given by geoRglm was higher than those given by the likelihood methods. The lowest one given by the INLA was lower than those. Thus, we concluded that results given by the likelihood-based methods were more reliable than those given by the Bayesian methods or the MCMC algorithms. We also studied the prediction performance of the methods. We then predicted the random effects at unobserved sites. We used those to predict the response at all the sites in the study region. We obtained the contour plot in Figure 2 by the PM. We also computed $\hat{\boldsymbol{y}}$ at the observed sites. We used it to calculate the G^2 value of the model. We compared it with G_0^2, the deviance GOF value of the null model (i.e., the model without γ_i in (17)). We had $G^2 = 9.70$ and $G_0^2 = 4506.9$, leading to the deviance coefficient of determination as $R^2 = 1 - G^2/G_0^2 = 0.998$ from the PM. In addition, we had $G^2 = 9.73$ from the PQL and $G^2 = 9.63$ from the LA. This means that all of the likelihood-based methods well explained the variations of the weed counts. We also studied similar issues for the MCMC methods. The result showed that their G^2 values were still low and their R^2 values also close to 1. This means that the MCMC methods also well explained the variations of the responses.

To assess whether we could assume $\omega_3 = 0.5$ in (16), we studied the likelihood-based methods with a varied ω_3. Theoretically, it was a profile likelihood approach because we fixed ω_3 at each candidate value and estimated all of the rest parameters. We put the estimates of those parameters in the log-likelihood function and obtained the profile log-likelihood function of ω_3, denoted by $\ell_P(\omega_3)$ (Figure 3). We estimated ω_3 by maximizing $\ell_P(\omega_3)$. We obtained $\hat{\omega}_3 = 0.82$ in our method and $\hat{\omega}_3 = 0.80$ in the PQL and the LA methods. Because the differences between $\ell_P(\hat{\omega}_3)$ and $\ell_P(0.5)$ were not higher than $\chi^2_{0.05,1} = 3.84$, we accepted the null hypothesis in the test for $H_0 : \omega_3 = 0.5$ against $H_1 : \omega_3 \neq 0.5$. We concluded that the true ω_3 was not significantly different from 0.5 and it was appropriate to use $\omega_3 = 0.5$ in the computation. Thus, the results given by Table 3 were appropriate.

5. Discussion

In this article, we propose a new method called the PM algorithm for MLEs of SGLMMs for count responses. The implementation of our method does not need any numerical evaluations of HDIIs contained by the likelihood functions of the models. It only needs a numerical algorithm for MLEs of SLMMs for normal responses and prediction of the random effects under the MLEs. If the MLEs are derived, then the prediction of the random effects can be easily carried out by matrix operations. This means that the computation of MLEs in SGLMMs for count responses is not harded than the computation of MLEs in SLMMs for normal responses. It is not necessary to develop specific algorithms for MLEs in SGLMMs for count responses. High-dimensional integrals are not issues in the computation of MLEs of SGLMMs for count responses. As an SLMM for normal responses contains the dispersion parameter but an SGLMM for count responses does not, we cannot directly use established fitting procedures to implement the PM. Therefore, new algorithms without the dispersion parameter are needed.

Mathematically, the PM is simpler than the previous LA and PQL. Although all of the purposes are the computations of the MLEs of SGLMMs, the LA and the PQL use higher-order (e.g., the third and the fourth order) moment expressions. The goal is to reduce the asymptotic bias. The PM does not have an asymptotic bias. It does use higher-order moments in its mathematical formulations either. It is asymptotically equivalent to the true MLE. Thus, the properties of the PM are nice. The PM

can be easily extended to any generalized mixed effects models for counts.

Our method successfully reduces the computational complexity for MLEs in SGLMMs for count responses to that for normal responses. Given that an algorithm for MLEs in a specific SLMM exists, we can modify it to a method for MLEs in the corresponding SGLMMs for count responses. Therefore, it is more important to develop efficient algorithms for SLMMs. This kind of problems has not been completely solved yet. An example is the spatial or spatiotemporal model with a large number of observations. Because the size of the variance-covariance matrix of the random effects is large, it is generally impossible to load the entire matrix to memory of a computing system, leading to a difficulty in computing the inverse and the determinant of the variance-covariance matrix. This kind of problems has been previously studied by many articles (Fuentes 2007; Cressie and Johannesson 2008; Kaufman, Schervish, and Nychka 2008; Liang, et. al 2013; Eidsvik, et. al 2014) for normal responses. Our research indicates that the corresponding methods can be modified to count responses. Therefore, our research provides a pathway to connect SGLMMs for count responses with SLMMs for normal responses. This is left to future research.

Supplementary Materials

The online supplementary materials contain the standard MLEs for spatial linear mixed models (SLMMs) for normal data, and the prediction of random effects based on the SLMMs, the formulations of the Fisher information matrix for the MLEs of the SLMMs, and the proofs of all of the lemmas, theorems, and corollaries.

Acknowledgments

The author appreciates great comments from the Associate Editor and two anonymous referees which significantly improve the quality of the article.

References

Aruoba, S. B., and Fernandez-Villaverde, J. (2015), "A Comparison of Programming Languages in Macroeconomics," *Journal of Economic Dynamics and Control*, 58, 265–273. [1645]

Besag, J. (1974), "Spatial Interaction and the Statistical Analysis of Lattice Systems," *Journal of the Royal Statistical Society*, Series B, 36, 192–236. [1638]

Barndorff-Neilsen, O. E., and Cox, D. R. (1989), *Asymptotic Techniques for Use in Statistics*, London: Chapman & Hall. [1636,1637]

Best, N. G., and Wakefield, J. C. (1999), "Accounting for Inaccuracies in Population Counts can Case Registration in Cancer Mapping Studies," *Journal of the Royal Statistical Society*, Series A, 162, 363–382. [1637]

Breslow, N. E., and Clayton, D. G. (1993), "Approximate Inference in Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 88, 9–25. [1636,1637,1638,1642]

Breslow, N. E., and Lin, X. (1995), "Bias Correlation in Generalised Linear Mixed Models with a Single Component of Dispersion," *Biometrika*, 82, 81–91. [1637,1638]

Christensen, O. F., and Ribeiro, P. J. (2002), "**geoRglm**: A Package for Generalized Linear Spatial Models," *R News*, 2, 26–28. [1642]

Christensen, O. F., and Waagepetersen, R. (2002), "Bayesian Prediction of Spatial Count Data using Generalized Linear Mixed Models," *Biometrics*, 58, 280–286. [1645]

Cressie, N., and Chan, N. (1989), "Spatial Modeling of Regional Variables," *Journal of the American Statistical Association*, 84, 393–401. [1638,1641]

Cressie, N., and Huang, H. C. (1999), "Classes of Nonseparable, Spatiotemporal Stationary Covariance Functions," *Journal of the American Statistical Association*, 94, 1330–1340. [1641]

Cressie, N., and Johannesson, G. (2008), "Fixed Rank Kriging for Very Large Spatial Data Sets," *Journal of the Royal Statistical Society*, Series B, 70, 209–226. [1647]

Dempster, A. R., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm," (with discussion), *Journal of the Royal Statistical Society*, Series B, 39, 1–38. [1637]

Diggle, P., Tawn, J. A., and Moyeed, R. A. (1998), "Model-based Geostatistics," (with discussion), *Applied Statistics*, 47, 299–350. [1636,1638,1641]

Eidsvik, J., Shaby, B. A., Reich, B. J., Wheeler, M., and Niemi, J. (2014), "Estimation and Prediction in Spatial Models with Block Composite Likelihoods," *Journal of Computational and Graphical Statistics*, 23, 295–315. [1647]

Evangelou, E., Zhu, Z., and Smith, R. L. (2011), "Estimation and Prediction for Spatial Generalized Linear Mixed Models using High Order Laplace Approximation," *Journal of Statistical Planning and Inference*, 141, 3564–3577. [1637,1638,1642]

Ferguson, T. S. (1996). *A Course in Large Sample Theory*, Boca Raton, FL: Chapman & Hall/CRC Press. [1640]

Fong, Y., Rue, H., and Wakefield, J. (2010), "Bayesian Inference for Generalized Linear Mixed Models," *Biostatistics*, 11, 397–412. [1638]

Fuentes, M. (2007), "Approximate Likelihood for Large Irregularly Spaced Spatial Data," *Journal of the American Statistical Association*, 102, 321–331. [1647]

Girogi, E., and Diggle, P. J. (2017), "Premap: An R Package for Prevalence Mapping," *Journal of Statistical Software*, 78, 1–29. [1642]

Gneiting, T. (2002), "Nonseparable, Stationary Covariance Functions for Space-Time Data," *Journal of the American Statistical Association*, 97, 590–600. [1641]

Gneiting, T., and Schlather, M. (2004), "Stochastic Models that Separate Fractal Dimension and the Hurst Effect," *SIAM Review*, 46, 269–282. [1641]

Guan, Y., and Haran, M. (2021), "Fast Expectation-Maximization Algorithms for Spatial Generalized Linear Mixed Models," PrePrint, arxiv.org/pdf/1909.05440.pdf. [1637]

Guillot, G., Lorén N., and Rudemo, M. (2009), "Spatial Prediction of Weed Intensities from Exact Count Data and Image-based Estimates," *Applied Statistics*, 58, 525–542. [1645]

Han, Z., and Oliveria, V. D. (2018), "gcKrig: An R Package for the Analysis of Geostatistical Count Data Using Gaussian Copulas," *Journal of Statistical Software*, 87, 1–32. [1637]

Handcock, M. S., and Stein, M. L. (1993), "A Bayesian Analysis of Kriging," *Technometrics*, 35, 403–410. [1641]

Jing, L., and Oliveira, D. V. (2015), "geoCount: An R Package for the Analysis of Geostatistical Count Data," *Journal of Statistical Software*, 63, 1–33. [1637,1642]

Kaufman, C. G., Schervish, M. J., and Nychka, D. W. (2008), "Covariance Tapering for Likelihood-Based Estimation in Large Spatial Data Sets," *Journal of the American Statistical Association*, 103, 1545–1555. [1647]

Lee, Y., and Nelder, J. A. (1996), "Hierarchical Generalized Linear Models," *Journal of Royal Statistical Society*, Series B, 58, 619–678. [1636,1638]

Liang, F., Cheng, Y., Song, Q., Park, J., and Yang, P. (2013), "A Resampling-based Stochastic Approximation Method for Analysis of Large Geostatistical Data," *Journal of the American Statistical Association*, 108, 325–339. [1647]

Lin, X., and Breslow, N. E. (1996), "Bias Correction in Generalized Linear Mixed Models with Multiple Components of Dispersion," *Journal of the American Statistical Association*, 91, 1007–1016. [1638]

Liu, Q., and Pierce, D. A. (1994), "A Note on Gauss-Hermite Quadrature," *Biometrika*, 81, 624–629. [1637]

Matérn, B. (1986). *Spatial Variation* (2nd ed.), Berlin: Springer. [1641]

McCulloch, C. E. (1997), "Maximum Likelihood Algorithms for Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 92, 162–170. [1637]

McCulloch, C. E., Searle, S. R., and Neuhaus, J. M. (2008), *Generalized, Linear, and Mixed Models*, Hoboken, NJ: Wiley. [1638]

Ord, K. (1975), "Estimation Methods for Models of Spatial Interaction," *Journal of the American Statistical Association*, 70, 120–126. [1638,1641]

Rosenblatt, M. (1984), "Asymptotic Normality, Strong Mixing and Spectral Density Estimation," *Annals of Probability*, 12, 1167–1180. [1640]

Rousset, F. (2021), "An Introduction to the spaMM Package for Mixed Models," available at *https://kimura.univ-montp2.fr/~rousset/spaMM/spaMMintro.pdf*. [1637]

Rue, H., Martino, S., and Chopin, N. (2009), "Approximate Bayesian Inference for Laten Gaussian Models by Using Integrated Nested Laplace Approximation," *Journal of the Royal Statistical Society*, Series B, 71, 319–392. [1637,1638,1642]

Stein, M. L. (1999), *Interpolation of Spatial Data. Some Theory for Kriging*, New York: Springer-Verlag. [1641]

——— (2005), "Space-Time Covariance Function," *Journal of the American Statistical Association*, 100, 310–321. [1641]

Shun, Z., and McCullagh, P. (1995), "Laplace Approximation for High Dimensional Integrals," *Journal of Royal Statistical Society*, Series B, 4, 749–760. [1637]

Zeger, S. L., and Karim, M. R. (1991), "Generalized Linear Models with Random Effects: A Gibbs Sampling Approach," *Journal of the American Statistical Association*, 86, 79–86. [1637,1638]

Zhang, H. (2002), "On Estimation and Prediction for Spatial Generalized Linear Mixed Models," *Biometrics*, 58, 129–136. [1637]

——— (2004), "Inconsistent Estimation and Asymptotically Equal Interpolations in Model-Based Geostatistics," *Journal of the American Statistical Association*, 99, 250–261. [1640]

Zhang, T. (2019), "General Gaussian Estimation," *Journal of Multivariate Analysis*, 169, 234–247. [1638,1639]

——— (2021), "Iteratively Reweighted Least Squares with Random Effects for Maximum Likelihood in Generalized Linear Mixed Effects Models," *Journal of Statistical Computation and Simulation*, 91, 3404–3425. DOI: 10.1080/00949655.2021.1928127. [1637,1638,1640,1642,1643]