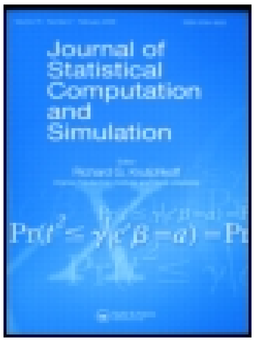

Taylor & Francis
Taylor & Francis Group

Journal of Statistical Computation and Simulation

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/gscs20

Iteratively reweighted least squares with random effects for maximum likelihood in generalized linear mixed effects models

Tonglin Zhang

To cite this article: Tonglin Zhang (2021): Iteratively reweighted least squares with random effects for maximum likelihood in generalized linear mixed effects models, Journal of Statistical Computation and Simulation, DOI: 10.1080/00949655.2021.1928127

To link to this article: https://doi.org/10.1080/00949655.2021.1928127

Published online: 16 May 2021.

JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION
https://doi.org/10.1080/00949655.2021.1928127

Iteratively reweighted least squares with random effects for maximum likelihood in generalized linear mixed effects models

Tonglin Zhang

Department of Statistics, Purdue University, West Lafayette, IN, USA

ABSTRACT

This article develops a new method called iteratively reweighted least squares with random effects (IRWLSR) for maximum likelihood in generalized linear mixed effects models (GLMMs). As normal distributions are used for random effects, the likelihood functions contain intractable integrals except when the responses are normal. This often induces computational difficulties in fitting GLMMs for non-normal responses. The proposed IRWLSR successfully overcomes the difficulties as it only needs computational methods for linear mixed effects models and can be applied to any GLMMs with arbitrary link functions. It can be used even when high-dimensional intractable integrals appear in the likelihood function. The simulation study shows that the results are comparable to and sometimes are more precise than those from the Laplace approximation in the case when the Laplace approximation can be applied. It can also be applied to the case when the Laplace approximation cannot be applied.

ARTICLE HISTORY

Received 21 September 2019
Accepted 6 May 2021

KEYWORDS

Generalized linear mixed effects models; iteratively reweighted least squares with random effects; Laplace approximation; Maximum-likelihood estimation; profile likelihood; variance components

1. Introduction

The main contribution of the article is the development of a new method called iteratively reweighted least squares with random effects (IRWLSR) for maximum-likelihood estimation (MLE) of generalized linear mixed effects models (GLLMs), such that all of the computations can be carried out under the framework of linear mixed effects (LMMs). It is well known that the likelihood function of a GLMM for non-normal responses contains intractable integrals, leading to difficulties in computing MLEs of model parameters. These difficulties can be completely avoided by IRWLSR, because it only needs computational methods for MLEs of LMMs. It is not necessary to consider any numerical methods for intractable integrals in the computation. This means that the derivation of MLEs of GLMMs for non-normal responses can be as easy as that for normal responses.

As normal distributions are used to model random effects, the likelihood functions of GLMMs contain intractable integrals except the case when the responses are normally distributed. This induces difficulties to implement the maximum-likelihood approach to

CONTACT Tonglin Zhang ✉ tlzhang@purdue.edu Department of Statistics, Purdue University, 250 North University Street, West Lafayette, IN 47907-2066, USA

GLMMs when the responses are not normal. To overcome the difficulties, many methods have been proposed. Examples include integrated nested Laplace approximation (INLA) [1], penalized quasi-likelihood (PQL) [2,3], Gauss-Hermite quadrature [4], Monte Carlo EM gradient (MCEMG) [5] and Gibbs sampler [6]. These methods may be efficient if the dimension of intractable integrals is low. However, they have difficulties when high-dimensional and irreducible intractable integrals are present, meaning that the dimension of an intractable integral increases with the sample size [7,8]. Our proposed IRWLSR can overcome these difficulties.

IRWLSR is motivated from iteratively reweighted least squares (IRWLS) [9]. IRWLS is the standard computational method for MLEs of GLMs for non-normal responses. It has been widely used by many software packages, such as R, SAS, Python and MatLab. In each iteration, IRWLS uses the weighted least-squares (WLS) method to solve a working weighted linear model (WWLM) for normal responses. The working weights and responses in the current iteration are determined by solutions of the previous iteration. The exact MLEs are derived if the algorithm converges. This means that IRWLS successfully changes the computation for non-normal responses to that for normal responses. Because solutions in each iteration can be analytically solved, IRWLS is extremely efficient in computing MLEs of GLMs for non-normal responses. As the link function is only used in updating the working weights and responses, IRWLS can be implemented to any GLMs with arbitrary link functions.

We find that the idea of IRWLS can be extended from GLMs to GLMMs, leading to the derivation of IRWLSR. Since a GLMM can be treated as a GLM conditioning on random effects, IRWLS can be used to compute the conditional MLE given the random effects. The link function is only used in updating the working weights and responses. This can provide the exact conditional MLEs given the random effects. To incorporate the idea to GLMMs, we assume that the random effects are not observed. By treating random effects as random variables, we obtain a working weighted linear mixed effects model (WWLMM). We study the likelihood function of the WWLMM in each iteration. We find that the likelihood function does not contain any intractable integrals, and the computation of the MLEs of fixed and random effects parameters is efficient. After they have been obtained, we calculate predicted values of the random effects in the WWLMM. Combine those with the MLEs of fixed effects parameters. We obtain the predicted values of the linear components, which are used in the next iteration. In the end, we obtain the MLEs of the GLMM.

Several advantages are quickly identified. The first is that the implementation of IRWLSR does not need any numerical evaluations of intractable integrals. This means that IRWLSR can be used to any kinds of GLMMs even if high-dimensional intractable integrals are present in the likelihood functions. The second is that any algorithm for LMMs can be modified to an algorithm for GLMMs. In each iteration, the predicted values of the random effects are obtained by their conditional expected values given the working responses. This can be exactly carried out by matrix operations, leading to the next iteration. The third is that IRWLSR can be easily applied to any link functions. The link function is only used in updating the working weights and responses. It is not involved in the computation of the MLEs in the iterations. Thus, the implementation of IRWLSR only needs an algorithm for the MLEs of LMMs.

Although IRWLSR can be used to any kinds of GLMMs with any kinds of reasonable link functions, we focus our presentation on binomial and Poisson data because of their popularity. We evaluate the performance of the approach in two scenarios. In the first scenario, we assume that the dimension of the intractable integrals is low. A typical example is the longitudinal study for count. This usually involves a number of clusters with dependence between clusters ignored. Many software packages can be used. We pick up the lme4 package in R in our comparison. We find that our results are comparable to and sometimes are more precise than those given by the package. To demonstrate flexibility of our approach, we pick up a link function which has not been adopted by the package yet. We want to show that IRWLSR can be implemented with arbitrary link functions. In the second scenario, we assume that the dimension of the intractable integral is high. We study the case when the dimension of the intractable integrals is identical to the sample size. It has been previously pointed out that neither the Laplace approximation nor the MCMC approach can be applied [7,8]. Our method can be easily applied because it does not need any numerical evaluations of high-dimensional intractable integrals.

The article is organized as follows. In Section 2, we provide a review of GLMMs under the framework of exponential family distributions. In Section 3, we present our method. In Section 4, we specify our method to two kinds of GLMMs. In Section 5, we evaluate the performance of our method via simulation studies. In Section 6, we apply our method to a real-world data set. In Section 7, we provide a discussion.

2. Review

GLMs are proposed for exponential family distributions. Their purpose is to model expected values of the response variables via the explanatory variables. Three components are needed to define a GLM. The random component consists of a response vector $\boldsymbol{y} = (y_1, \dots, y_n)^\top$ with $y_1, \dots, y_n$ independently obtained from an exponential family distribution, where n is the sample size. An exponential family distribution has a probability mass function (PMF) or a probability density function (PDF) as

$$f(y_i) = \exp\left[\frac{y_i\omega_i - b(\omega_i)}{a(\phi)} + c(y_i, \phi)\right], \tag{1}$$

where ω_i is a canonical parameter representing the location and $\phi \in \mathbb{R}$ is a dispersion parameter representing the scale. GLMs can be specified to normal, Bernoulli, binomial or Poisson distributions. Under (1), one has $\mathrm{E}(y_i) = b'(\omega_i) = \mu_i$ and $\mathrm{V}(y_i) = a(\phi)b''(\omega_i)$. The linear component $\boldsymbol{\eta} = (\eta_1, \dots, \eta_n)^\top$ is a vector related to explanatory variables by $\eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$ for all $i \in \{1, \dots, n\}$, where $\boldsymbol{x}_i = (1, x_{i1}, \dots, x_{i(p-1)})^\top$ is the ith observed vector of explanatory variables and $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_{p-1})^\top$ represents a vector of regression coefficients. The link function $g(\cdot)$ connects μ_i and η_i through $\eta_i = g(\mu_i) = g[b'(\omega_i)] = \boldsymbol{x}_i^\top \boldsymbol{\beta}$ for all $i \in \{1, \dots, n\}$. The unknown parameters $\boldsymbol{\beta}$ and ϕ are estimated by the maximum-likelihood approach.

Suppose that the conditional distribution of y_i given random effects $\boldsymbol{\gamma}$ is given by (1). A GLMM is proposed as

$$g(\mu_i) = \eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \boldsymbol{z}_i^\top \boldsymbol{\gamma}, \quad i = 1, \dots, n, \tag{2}$$

where $\boldsymbol{\beta}$ is a p-dimensional parameter vector for fixed-effects, $\boldsymbol{\gamma}$ is a q-dimensional vector for random-effects, and $\mu_i = b'(\omega_i) = \mathrm{E}(y_i|\boldsymbol{\gamma})$ is the conditional mean of y_i given $\boldsymbol{\gamma}$. The conditional variance of y_i is $\mathrm{V}(y_i|\boldsymbol{\gamma}) = a(\phi)b''(\omega_i)$. A common way to model $\boldsymbol{\gamma}$ is to use a multivariate normal distribution as

$$\boldsymbol{\gamma} \sim \mathcal{N}(\mathbf{0}, \mathbf{V}_{\boldsymbol{\delta}}), \tag{3}$$

where $\mathbf{V}_{\boldsymbol{\delta}}$ is a $q \times q$-dimensional variance-covariance component matrix for $\boldsymbol{\gamma}$ and $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_r)^\top$ is an r-dimensional parameter vector contained by $\mathbf{V}_{\boldsymbol{\delta}}$.

Let $\ell(\boldsymbol{\beta}, \phi|\boldsymbol{\gamma})$ be the log-likelihood function of (2) under (1) for a given $\boldsymbol{\gamma}$ and $\pi_{\boldsymbol{\delta}}(\boldsymbol{\gamma})$ be the prior density of $\boldsymbol{\gamma}$ given by (3). Since $\boldsymbol{\gamma}$ is not observed, the likelihood function of the GLMM is obtained by integrating out $\boldsymbol{\gamma}$ in the joint distribution of $\boldsymbol{y}$ and $\boldsymbol{\gamma}$ as

$$L(\boldsymbol{\theta}) = \int_{\mathbb{R}^q} \mathrm{e}^{\ell(\boldsymbol{\beta},\phi|\boldsymbol{\gamma})} \pi_{\boldsymbol{\delta}}(\boldsymbol{\gamma})\, \mathrm{d}\boldsymbol{\gamma}, \tag{4}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \phi, \boldsymbol{\delta}^\top)^\top$ is the vector of all parameters contained by the model. The MLE of $\boldsymbol{\theta}$ satisfies

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}), \tag{5}$$

where $\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta})$ is the log-likelihood function of the model.

It is well-known that the right-hand size of (4) is intractable if the distribution given by (1) is not normal. In this case, the computation of $\hat{\boldsymbol{\theta}}$ is challenging as one needs to numerically evaluate the intractable integral over the entire $\mathbb{R}^q$. This is extremely difficult when q is large. A couple of the most popular methods are the penalized quasi-likelihood (PQL) [2] and the integrated nested Laplace approximation (INLA) [1]. They approximate the right-hand side of (4) by the Laplace approximation, which relies on

$$\int_{\mathbb{R}^d} \mathrm{e}^{\varphi_{\boldsymbol{\theta}}(\boldsymbol{b})} d\boldsymbol{b} \approx \frac{(2\pi)^{a/2} \mathrm{e}^{\varphi_{\boldsymbol{\theta}}(\hat{\boldsymbol{b}}_{\boldsymbol{\theta}})}}{|-\det\{\Delta\varphi_{\boldsymbol{\theta}}(\hat{\boldsymbol{b}}_{\boldsymbol{\theta}})\}|^{1/2}}, \tag{6}$$

where $\varphi_{\boldsymbol{\theta}}(\boldsymbol{b})$ is a smooth function of $\boldsymbol{b}$ which may also depend on $\boldsymbol{\theta}$, $\hat{\boldsymbol{b}}_{\boldsymbol{\theta}} = \mathrm{argmax}_{\boldsymbol{b}}\, \varphi_{\boldsymbol{\theta}}(\boldsymbol{b})$, and $\Delta\varphi_{\boldsymbol{\theta}}(\boldsymbol{b})$ is the Hessian matrix of $\varphi_{\boldsymbol{\theta}}(\boldsymbol{b})$ with respect to $\boldsymbol{b}$. After applying the Laplace approximation to (4) for a given $\boldsymbol{\theta}$ with $q = d$ in (6), one also needs to calculate its gradient vector and Hessian matrix. After that, another optimization procedure is needed for $\hat{\boldsymbol{\theta}}$. This often involves an optimization problem of a high-dimensional parameter if q is large.

In addition, MCMC algorithms developed under the Bayesian approach can be used to approximate the right-hand side of (4). Computation in MCMC is also an issue since the usual implementation needs a large amount of computational overhead [10,11]. To apply the MCMC, one needs to specify a proposal distribution on $\mathbb{R}^q$. The acceptance of the new sample is jointly determined by the old sample, the new sample and the proposal distribution. The acceptance probability could be extremely low if q is only moderately large. To solve the problem, Gibbs sampler is proposed [6]. Instead of updating the entire parameter values, Gibbs sampler updates individual parameter values conditioning on the remaining parameter values. Its implementation 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 computation must be conducted in each stage of the updating procedure. Thus, the implementation of the MCMC algorithm is often time-consuming and the result is unstable, especially when q is large.

POL does not work well because it can lead to an asymptotically biased estimator and hence inconsistent [12]. Therefore, it is recommended not used in practice [13, P. 198]. Properties of Laplace approximation and Bayesian approaches have been previously studied for irreducible high-dimensional integrals by Shun and McCullagh [8]. Their article points out that the performance of the Laplace approximation depends on the relationship between d and n. The approximation given by (6) is not valid for the right-hand side of (4) if d does not vanish at rate $n^{1/3}$. This also occurs in the computation of posterior expectations in the Bayesian approach when the parameter is high dimensional.

3. Method

Our method includes the construction of the WWLMM in Section 3.1, the development of the entire IRWLSR procedure in Section 3.2, and the derivation of Fisher Information in Section 3.3. Since $\boldsymbol{\gamma}$ is used in the construction of the WWLMM, we introduce the conditional MLE problem at the beginning of this section. We then assume that $\boldsymbol{\gamma}$ is unknown and propose our IRWLSR. The method implies that any numerical algorithm for LMMs can be extended to a numerical algorithm for GLMMs.

3.1. Working model

We investigate IRWLS for (2) with a given $\boldsymbol{\gamma}$. If $\boldsymbol{\gamma}$ is assumed known, then the GLMM becomes a GLM. We can use IRWLS to compute the exact values of $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ and $\hat{\phi}_{\boldsymbol{\gamma}}$, the conditional MLEs of $\boldsymbol{\beta}$ and ϕ given $\boldsymbol{\gamma}$, respectively. Because the MLE of $\boldsymbol{\delta}$ can be directly computed by (3), it is enough for us to focus on the derivation of $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ and $\hat{\phi}_{\boldsymbol{\gamma}}$ only. In particular, we obtain the log-likelihood function of model parameters given by (1), (2) and (3) with a given $\boldsymbol{\gamma}$ as

$$\ell(\boldsymbol{\theta}|\boldsymbol{\gamma}) = \ell(\boldsymbol{\beta}, \phi, \boldsymbol{\delta}|\boldsymbol{\gamma}) = \ell(\boldsymbol{\beta}, \phi|\boldsymbol{\gamma}) + \ell(\boldsymbol{\delta}|\boldsymbol{\gamma}), \tag{7}$$

where

$$\ell(\boldsymbol{\delta}|\boldsymbol{\gamma}) = \log \pi_{\boldsymbol{\delta}}(\boldsymbol{\gamma}) = -\frac{q}{2}\log(2\pi) - \frac{1}{2}\log|\det(\mathbf{V}_{\boldsymbol{\delta}})| - \frac{1}{2}\boldsymbol{\gamma}^{\top}\mathbf{V}_{\boldsymbol{\delta}}^{-1}\boldsymbol{\gamma} \tag{8}$$

is the log-likelihood function of $\boldsymbol{\delta}$ given by (3).

The conditional MLE of $\boldsymbol{\theta}$ given $\boldsymbol{\gamma}$, denoted by $\hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}} = (\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}^{\top}, \hat{\phi}_{\boldsymbol{\gamma}}, \hat{\delta}_{\boldsymbol{\gamma}}^{\top})^{\top}$, where $\hat{\boldsymbol{\delta}}_{\boldsymbol{\gamma}}$ is the conditional MLE of $\boldsymbol{\delta}$ given $\boldsymbol{\gamma}$, is solved by

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \ell(\boldsymbol{\beta}, \phi, \boldsymbol{\delta}|\boldsymbol{\gamma}). \tag{9}$$

As $(\boldsymbol{\beta}^{\top}, \phi)^{\top}$ and $\boldsymbol{\delta}$ are well separated by the first and second terms on the right-hand side of (7), we estimate them separately by $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} = \operatorname{argmax}_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta}, \phi|\boldsymbol{\gamma})$ and $\hat{\boldsymbol{\delta}}_{\boldsymbol{\gamma}} = \operatorname{argmax}_{\boldsymbol{\delta}} \ell(\boldsymbol{\delta}|\boldsymbol{\gamma})$

with $\hat{\phi}_{\boldsymbol{\gamma}}$ given by a moment estimator [14] as

$$a(\hat{\phi}_{\boldsymbol{\gamma}}) = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_{i,\boldsymbol{\gamma}})^2}{b''[h(\hat{\eta}_{i,\boldsymbol{\gamma}}))]}, \tag{10}$$

where $\hat{\mu}_{i,\boldsymbol{\gamma}} = b'[h(\hat{\eta}_{i,\boldsymbol{\gamma}})]$ and $\hat{\eta}_{i,\boldsymbol{\gamma}} = \boldsymbol{x}_i^\top\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} + \boldsymbol{z}_i^\top\boldsymbol{\gamma}$ are predicted values of the ith response and linear component.

We use IRWLS to compute the exact value of $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$. In particular, let $\boldsymbol{\beta}_{\boldsymbol{\gamma}}^{(t)}$ be the tth iterative value of $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$. Then, the tth iterative value of the linear component is

$$\eta_{i,\boldsymbol{\gamma}}^{(t)} = \boldsymbol{x}_i^\top\boldsymbol{\beta}_{\boldsymbol{\gamma}}^{(t)} + \boldsymbol{z}_i^\top\boldsymbol{\gamma}. \tag{11}$$

Let $\mu_{i,\boldsymbol{\gamma}}^{(t)} = g^{-1}(\eta_{i,\boldsymbol{\gamma}}^{(t)})$ and $\boldsymbol{\mu}_{\boldsymbol{\gamma}}^{(t)} = (\mu_{1,\boldsymbol{\gamma}}^{(t)}, \ldots, \mu_{n,\boldsymbol{\gamma}}^{(t)})^\top$. In the tth iteration, the ith working weights is

$$w_{i,\boldsymbol{\gamma}}^{(t)} = \frac{1}{b''[h(\eta_{i,\boldsymbol{\gamma}}^{(t)})]}\left(\frac{\partial\mu_{i,\boldsymbol{\gamma}}^{(t)}}{\partial\eta_{i,\boldsymbol{\gamma}}^{(t)}}\right) \tag{12}$$

and the ith working responses is

$$u_{i,\boldsymbol{\gamma}}^{(t)} = \eta_{i,\boldsymbol{\gamma}}^{(t)} + (y_i - \mu_{i,\boldsymbol{\gamma}}^{(t)})\frac{\partial\eta_{i,\boldsymbol{\gamma}}^{(t)}}{\partial\mu_{i,\boldsymbol{\gamma}}^{(t)}}. \tag{13}$$

Then, we obtain the tth working model as

$$\boldsymbol{u}_{\boldsymbol{\gamma}}^{(t)} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \tag{14}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\{\mathbf{W}_{\boldsymbol{\gamma}}^{(t)}\}^{-1})$, $\sigma^2 = a(\phi)$, $\mathbf{W}_{\boldsymbol{\gamma}}^{(t)} = \text{diag}(w_{1,\boldsymbol{\gamma}}^{(t)}, \ldots, w_{n,\boldsymbol{\gamma}}^{(t)})$, $\mathbf{X} = (\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_n^\top)^\top$, and $\mathbf{Z} = (\boldsymbol{z}_1^\top, \ldots, \boldsymbol{z}_n^\top)^\top$. By maximizing the log-likelihood function of (14) given $\boldsymbol{\gamma}$, we obtain $\boldsymbol{\beta}_{\boldsymbol{\gamma}}^{(t+1)}$, the next iterated value of $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$. The derivation of $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ does not need σ^2. We only use (10) after $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ is available. Thus, the implementation of IRWLS for $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ does not involve the computation of $\hat{\phi}_{\boldsymbol{\gamma}}$.

IRWLS can only be used when $\boldsymbol{\gamma}$ is known. Thus, it cannot be used to fit GLMMs. Note that the working model given by (14) is completely a normal model. We investigate whether it can be modified for GLMMs. This motivates the development of our IRWLSR.

3.2. IRWLSR

Similar to IRWLS, the major issue in IRWLSR is the construction of the working weights and responses for (14) such that the next iteration can be carried out. This needs expressions similar to (11) –(13). Because $\boldsymbol{\gamma}$ is unknown, we predict $\boldsymbol{\gamma}$ in the iterations. Therefore, IRWLSR needs to update both $\hat{\boldsymbol{\theta}}_{\boldsymbol{\gamma}}$ and $\boldsymbol{\gamma}$. We use $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^\top, \hat{\phi}, \hat{\boldsymbol{\delta}}^\top)^\top$ to represent $\hat{\boldsymbol{\theta}}$ with unobserved $\boldsymbol{\gamma}$ and $\hat{\boldsymbol{\gamma}}$ to represent the predicted value of $\boldsymbol{\gamma}$ under $\hat{\boldsymbol{\theta}}$. A goal of IRWLSR is to provide $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\gamma}}$. Another goal is to provide the Fisher information, which is discussed in Section 3.3.

Let $\boldsymbol{\theta}^{(t)} = (\{\boldsymbol{\beta}^{(t)}\}^\top, \phi^{(t)}, \{\boldsymbol{\delta}^{(t)}\}^\top)^\top$ and $\boldsymbol{\gamma}^{(t)}$ be the tth iterative values of $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\gamma}}$, respectively. Assume that they have been derived in the previous iteration. In the current iteration, we construct the ith linear component as

$$\eta_i^{(t)} = \boldsymbol{x}_i^\top \boldsymbol{\beta}^{(t)} + \boldsymbol{z}_i^\top \boldsymbol{\gamma}^{(t)}, \tag{15}$$

the ith predicted value of μ_i as $\mu_i^{(t)} = g^{-1}(\eta_i^{(t)})$. Then, $\boldsymbol{\mu}^{(t)} = (\mu_1^{(t)}, \dots, \mu_n^{(t)})^\top$ is the predicted conditional mean vector of the entire response. The ith working weight is

$$w_i^{(t)} = \frac{1}{b''[h(\eta_i^{(t)})]} \left(\frac{\partial \mu_i^{(t)}}{\partial \eta_i^{(t)}} \right) \tag{16}$$

and the ith working response is

$$u_i^{(t)} = \eta_i^{(t)} + (y_i - \mu_i^{(t)}) \frac{\partial \eta_i^{(t)}}{\partial \mu_i^{(t)}}. \tag{17}$$

Then, (14) becomes

$$\boldsymbol{u}^{(t)} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \tag{18}$$

where $\boldsymbol{u}^{(t)} = (u_1^{(t)}, \dots, u_n^{(t)})^\top$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \{\mathbf{W}^{(t)}\}^{-1})$, $\sigma^2 = a(\phi)$, $\mathbf{W}^{(t)} = \operatorname{diag}(w_1^{(t)}, \dots, w_n^{(t)})$, the prior distribution of $\boldsymbol{\gamma}$ is given by (3), and $\boldsymbol{\epsilon}$ and $\boldsymbol{\gamma}$ are independent. Similar to IRWLS, we do not need $\phi^{(t)}$ in the derivation of (18). We use (18) to derive $\boldsymbol{\theta}^{(t+1)}$ and $\boldsymbol{\gamma}^{(t+1)}$ for the next iteration.

Since $\boldsymbol{\gamma}$ is unobserved, we cannot use the method for the conditional MLE introduced in Section 3.1. We develop a method for the unconditional MLE, where we need to provide both $\boldsymbol{\theta}^{(t+1)}$ and $\boldsymbol{\gamma}^{(t+1)}$ in the tth iteration. We find that the entire computation does not need any numerical evaluations of intractable integrals. We put the detail of the derivation in Appendix A and only introduce the main steps below.

Integrating $\boldsymbol{\gamma}$ out in the joint distribution of $\boldsymbol{u}^{(t)}$ and $\boldsymbol{\gamma}$ given by (18), we obtain the marginal distribution of $\boldsymbol{u}^{(t)}$ as

$$\boldsymbol{u}^{(t)} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{R}_{\boldsymbol{\delta}}), \tag{19}$$

where $\mathbf{R}_{\boldsymbol{\delta}} = \mathbf{Z}\mathbf{B}_{\boldsymbol{\delta}}\mathbf{Z}^\top + \{\mathbf{W}^{(t)}\}^{-1}$ and $\mathbf{B}_{\boldsymbol{\delta}} = \sigma^{-2}\mathbf{V}_{\boldsymbol{\delta}}$. The marginal log-likelihood function of $\boldsymbol{\theta}$ in the tth iteration is

$$\ell^{(t)}(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2}|\det(\mathbf{R}_{\boldsymbol{\delta}})| - \frac{1}{2\sigma^2}(\boldsymbol{u}^{(t)} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{R}_{\boldsymbol{\delta}}^{-1}(\boldsymbol{u}^{(t)} - \mathbf{X}\boldsymbol{\beta}). \tag{20}$$

We obtain $\boldsymbol{\theta}^{(t+1)}$ by

$$\boldsymbol{\theta}^{(t+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \ell^{(t)}(\boldsymbol{\theta}). \tag{21}$$

We use the profile maximum-likelihood approach to compute $\boldsymbol{\theta}^{(t+1)}$, because it can reduce the dimensionality in the optimization problem given by (21). In particular, for given $\boldsymbol{\delta}$, the conditional MLE of $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta}_{\boldsymbol{\delta}}^{(t+1)} = (\mathbf{X}^\top \mathbf{R}_{\boldsymbol{\delta}}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top \mathbf{R}_{\boldsymbol{\delta}}^{-1}\boldsymbol{u}^{(t)} \tag{22}$$

and the conditional MLE of ϕ is

$$\phi_{\boldsymbol{\delta}}^{(t+1)} = a^{-1}\left(\frac{1}{n}\{\boldsymbol{u}^{(t)}\}^\top \mathbf{M}_{\boldsymbol{\delta}}\boldsymbol{u}^{(t)}\right), \tag{23}$$

where $\mathbf{M}_{\boldsymbol{\delta}} = \mathbf{R}_{\boldsymbol{\delta}}^{-1} - \mathbf{R}_{\boldsymbol{\delta}}^{-1}\mathbf{X}(\mathbf{X}^\top \mathbf{R}_{\boldsymbol{\delta}}^{-1}\mathbf{X})^{-1}\mathbf{X}^\top \mathbf{R}_{\boldsymbol{\delta}}^{-1}$. Putting these into (20), we obtain the profile log-likelihood function of $\boldsymbol{\delta}$ as

$$\ell_P^{(t)}(\boldsymbol{\delta}) = -\frac{n}{2}\left[1 + \log\left(\frac{2\pi}{n}\right)\right] - \frac{1}{2}\log|\det(\mathbf{M}_{\boldsymbol{\delta}})| - \frac{n}{2}\log(\{\boldsymbol{u}^{(t)}\}^\top \mathbf{M}_{\boldsymbol{\delta}}\boldsymbol{u}^{(t)}). \tag{24}$$

We calculate the MLE of $\boldsymbol{\delta}$ by

$$\boldsymbol{\delta}^{(t+1)} = \operatorname*{argmax}_{\boldsymbol{\delta}} \ell_P(\boldsymbol{\delta}). \tag{25}$$

After $\boldsymbol{\delta}^{(t+1)}$ is obtained, we compute $\boldsymbol{\beta}^{(t+1)}$ and $\phi^{(t+1)}$ by $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}_{\boldsymbol{\delta}^{(t+1)}}^{(t+1)}$ and $\phi^{(t+1)} = \phi_{\boldsymbol{\delta}^{(t+1)}}^{(t+1)}$, respectively, which provides the final solution of $\boldsymbol{\theta}^{(t+1)}$.

We use the conditional distribution of $\boldsymbol{\gamma}$ given $\boldsymbol{u}^{(t+1)}$ in the derivation of $\boldsymbol{\gamma}^{(t+1)}$. Note that they are jointly normal. We calculate the conditional mean vector of $\boldsymbol{\gamma}$ given $\boldsymbol{u}^{(t)}$ and obtain $\mathrm{E}(\boldsymbol{\gamma}|\boldsymbol{u}^{(t)}) = \mathbf{B}_{\boldsymbol{\delta}}\mathbf{Z}^\top \mathbf{M}_{\boldsymbol{\delta}}\boldsymbol{u}^{(t)}$, leading to

$$\boldsymbol{\gamma}^{(t+1)} = \mathbf{B}_{\boldsymbol{\delta}^{(t+1)}}\mathbf{Z}^\top \mathbf{M}_{\boldsymbol{\delta}^{(t+1)}}\boldsymbol{u}^{(t)}. \tag{26}$$

After both $\boldsymbol{\theta}^{(t+1)}$ and $\boldsymbol{\gamma}^{(t+1)}$ are obtained, we can carry out the next iteration. To start our method, we need to provide $\boldsymbol{u}^{(0)}$ and $\mathbf{W}^{(0)}$, the initial guesses of the working response and weight. We choose the same as those used by IRWLS. Then, we can conduct the entire IRWLSR procedure. We summarize our algorithm below.

Algorithm of IRWLSR for MLEs of GLMMs.

(i) Obtain $\boldsymbol{u}^{(0)}$ and $\mathbf{W}^{(0)}$ by the traditional IRWLS method.
(ii) Based on $\boldsymbol{u}^{(t)}$ and $\mathbf{W}^{(t)}$ in the previous iteration, calculate $\boldsymbol{\theta}^{(t+1)}$ by (21) and $\boldsymbol{\gamma}^{(t+1)}$ by (26), respectively.
(iii) Update $\boldsymbol{u}^{(t)}$ and $\mathbf{W}^{(t)}$ by $\mathbf{u}^{(t+1)}$ and $\mathbf{W}^{(t+1)}$ using (16) and (17), respectively.
(iv) Iterate (ii) and (iii) until convergence.

We compare $\boldsymbol{\theta}^{(t)}$ given by the IRWLSR with $\boldsymbol{\theta}_{\boldsymbol{\gamma}}^{(t)}$ given by (9). We treat $\boldsymbol{\theta}^{(t)}$ as an estimator of $\boldsymbol{\theta}$ by applying the multiple imputing approach to $\boldsymbol{\theta}_{\boldsymbol{\gamma}}^{(t)}$ with a missing $\boldsymbol{\gamma}$. Following traditional approach for asymptotics of multiple imputation [15,16], we investigate asymptotic properties of $\boldsymbol{\theta}^{(t)}$ by studying the difference between $\boldsymbol{\theta}^{(t)}$ and $\boldsymbol{\theta}_{\boldsymbol{\gamma}}^{(t)}$ for sufficiently large t. Note that t is the number of iterations. It is usually not large in the computation. Therefore, it is enough for us to study the problem for a bounded t (e.g. $t \le 50$). We modify conclusions in the two articles. We summarize our findings by the following proposition.

Proposition 3.1: *If $\sqrt{n}(\boldsymbol{\theta}^{(t)}_{\hat{\boldsymbol{\gamma}}^{(t)}} - \boldsymbol{\theta}^{(t)})$ weakly converges to a multivariate distribution with finite second-order moments for a fixed t as $n \to \infty$, then $\sqrt{n}(\boldsymbol{\theta}^{(t)} - \hat{\boldsymbol{\theta}})$ converges to a multivariate normal distribution with the Fisher Information given by the asymptotic distribution of $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ for a bounded varied t when $n \to \infty$, where $\boldsymbol{\theta}_0$ is the true parameter.*

Proof: By Theorem 1 of [16], we conclude that the asymptotic distributions of $\sqrt{n}(\hat{\boldsymbol{\theta}}_{\hat{\boldsymbol{\gamma}}} - \boldsymbol{\theta}_0)$ and $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ are identical and $\sqrt{n}(\hat{\boldsymbol{\theta}}_{\hat{\boldsymbol{\gamma}}} - \hat{\boldsymbol{\theta}})$ goes to zero in probability. As $\boldsymbol{\theta}^{(t)}_{\hat{\boldsymbol{\gamma}}}$ converges to the global maximum, we can replace $\hat{\boldsymbol{\theta}}_{\hat{\boldsymbol{\gamma}}}$ by $\boldsymbol{\theta}^{(t)}_{\hat{\boldsymbol{\gamma}}}$ in the expression of the asymptotic distribution. Combined with the assumptions, we draw the conclusion. ■

The conditions of Proposition 3.1 are weak as the limiting distribution of $\sqrt{n}(\boldsymbol{\theta}^{(t)}_{\hat{\boldsymbol{\gamma}}^{(t)}} - \boldsymbol{\theta}^{(t)})$ is studied under the WWLMM for normal data. The fact is that for a given t, $\boldsymbol{\theta}^{(t)}_{\boldsymbol{\gamma}}$ is the conditional MLE and $\boldsymbol{\theta}^{(t)}$ is the unconditional MLE of $\boldsymbol{\theta}$ in the working model. Therefore, the relationship between the conditional and unconditional MLEs in LMMs can be migrated to that in GLMMs with predicted $\hat{\boldsymbol{\gamma}}^{(t)}$ in the iterations, which induces the limiting distribution of $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$.

It is well known that the integral in the likelihood function given by the right-hand side of (4) is intractable in binomial or Poisson data. According to the conclusion given by [8], it is unlikely to use the Laplace approximation to evaluate the integral if its dimension is large, but our method can overcome the difficulty.

In binomial GLMMs, we assume that $y_i \sim Bin(m_i, \pi_i)$ independently given $\boldsymbol{\gamma}$, where $g(\cdot)$ in (2) may be specified as the logistic, the cloglog, or the general inverse CDF link. The general inverse CDF link includes the probit and the Cauchy links. To implement IRWLSR, we need to choose the initial $u_i^{(0)}$ and $w_i^{(0)}$ for all $i \in \{1, \dots, n\}$. We use those given by the traditional IRWLS.

If the logistic link is used, then (2) becomes

$$\log \frac{\pi_i}{1-\pi_i} = \eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \boldsymbol{z}_i^\top \boldsymbol{\gamma}, \quad i = 1, \dots, n. \tag{27}$$

Following the traditional IRWLS, we have $u_i^{(0)} = \log[(y_i + 0.5)/(m_i - y_i + 0.5)]$ and $w_i^{(0)} = m_i(y_i + 0.5)(m_i - y_i + 0.5)/(m_i + 1)^2$. Then, we obtain $\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\gamma}^{(1)}$ by (21) and (26), respectively. By (16) and (17), if $t \geq 1$, then we have $w_i^{(t)} = m_i \pi_i^{(t)}(1 - \pi_i^{(t)})$ and $u_i^{(t)} = \eta_i^{(t)} + (y_i - m_i \pi_i^{(t)})/[m_i \pi_i^{(t)}(1 - \pi_i^{(t)})]$, where $\pi_i^{(t)} = e^{\eta_i^{(t)}}/(1 + e^{\eta_i^{(t)}})$ and $\eta_i^{(t)}$ is given by (15). Thus, we can carry out our IRWLSR.

If the general inverse CDF link is used, then (2) becomes

$$\Psi^{-1}(\pi_i) = \eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \boldsymbol{z}_i^\top \boldsymbol{\gamma}, \quad 1 = 1, \dots, n. \tag{28}$$

where $\Psi(\cdot)$ is a CDF of a continuous random variable taking values in the entire $\mathbb{R}$. Following the traditional IRWLS, we have $u_i^{(0)} = \Psi^{-1}[(y_i + 0.5)/(m_i + 1)]$ and $w_i^{(0)} = m_i \psi^2(u_i^{(0)})/\{\Psi(u_i^{(0)})[1 - \Psi(u_i^{(0)})]\}$, where $\psi(\cdot) = \Psi'(\cdot)$ is the PDF of the random variable. We obtain $\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\gamma}^{(1)}$ by (21) and (26), respectively. By (16) and (17), if $t \geq 1$, then $w_i^{(t)} = m_i \psi^2(\eta_i^{(t)})/\{\Psi(\eta_i^{(t)})[1-\Psi(\eta_i^{(t)})]\}$ and $u_i^{(t)} = \eta_i^{(t)} + [y_i - m_i\Psi(\eta_i^{(t)})]/[m_i\psi(\eta_i^{(t)})]$, where $\eta_i^{(t)}$ is given by (15).

If the cloglog link is used, then (2) becomes

$$\log[-\log(1-\pi_i)] = \eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \boldsymbol{z}_i^\top \boldsymbol{\gamma}, \quad 1 = 1, \ldots, n. \tag{29}$$

Following the traditional IRWLS, we have $u_i^{(0)} = \log\{-\log[(m_i - y_i + 0.5)/(m_i + 1)]\}$ and $w_i^{(0)} = m_i[(m_i - y_i + 0.5)/(y_i + 0.5)]\log^2[(m_i - y_i + 0.5)/(m_i + 1)]$. If $t \geq 1$, then $w_i^{(t)} = m_i e^{2\eta_i^{(t)} - e^{\eta_i^{(t)}}}/(1 - e^{-e^{\eta_i^{(t)}}})$ and $u_i^{(t)} = \eta_i^{(t)} + (y_i - n_i + n_i e^{-e^{\eta_i^{(t)}}})/(m_i e^{\eta_i^{(t)}} e^{-e^{\eta_i^{(t)}}})$, where $\eta_i^{(t)}$ is given by (15).

For Poisson data, we assume that $y_i \sim \mathcal{P}(\mu_i)$ independently given $\boldsymbol{\gamma}$, where $g(\cdot)$ in (2) is the log link. Then, (2) becomes

$$\log(\mu_i) = \eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \boldsymbol{z}_i^\top \boldsymbol{\gamma}, \quad 1 = \ldots, n. \tag{30}$$

Following the traditional IRWLS, we have $u_i^{(0)} = \log(y_i + 0.5)$ and $w_i^{(0)} = y_i + 0.5$. Then, we can obtain $\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\gamma}^{(1)}$. If $t \geq 1$, then $w_i = e^{\eta_i^{(t)}}$ and $u_i^{(t)} = \eta_i^{(t)} + (y_i - e^{\eta_i^{(t)}})/e^{\eta_i^{(t)}}$.

We have demonstrated that IRWLSR can be used to GLMMs for binomial and Poisson data, where the entire computation does not need any numerical evaluations of intractable integrals. Therefore, our IRWLSR can be used to fit GLMMs for binomial or Poisson data even if q is large in (4). Since the general implementation of the IRWLSR does not rely on the distribution of the response and the link function, this conclusion holds for any GLMMs with arbitrary reasonable link functions. Although we have provided the method to compute $\hat{\boldsymbol{\theta}}$, we have not provided a method to compute its variance–covariance matrix yet. This is related to the derivation of the Fisher information.

3.3. Fisher information

We use the Fisher information of $\boldsymbol{\theta}^{(t+1)}$ to approximate the variance-covariance matrix of $\hat{\boldsymbol{\theta}}$, where $\boldsymbol{\theta}^{(t+1)}$ is the MLE of $\boldsymbol{\theta}$ for (19) for a given t. We treat it as the working Fisher information given by IRWLSR. We study properties of the working Fisher information in the case when random effects are absent. We find that the working Fisher information approaches the true Fisher information if the algorithm for IRWLSR converges. We provide our results below.

Proposition 3.2: *Let $\mathbf{I}^{(t)}(\boldsymbol{\theta})$ be the Fisher information of $\boldsymbol{\theta}^{(t+1)}$ given by* (19) *and $\mathbf{I}(\boldsymbol{\theta})$ be the Fisher information of $\hat{\boldsymbol{\theta}}$ in the model given by* (1), (2) *and* (3). *If all assumptions of Proportion 3.1 holds, then $\mathbf{I}^{(t)}(\boldsymbol{\theta}_0)$ converges to $\mathbf{I}(\boldsymbol{\theta}_0)$ in probability as $n \to \infty$ if the algorithm for IRWLSR converges.*

Proof: The conclusion can be proven by the same method in the proof of Proposition 3.1. ■

The Fisher information $\mathbf{I}^{(t)}(\boldsymbol{\theta})$ for (19) can be easily derived. In particular, we calculate the second-order partial derivatives of $\ell^{(t)}(\boldsymbol{\theta})$ given by (20) with respect to $\boldsymbol{\beta}$, σ^2 and $\boldsymbol{\delta}$.

We then take negative expected values of the second-order partial derivatives. In the end, we obtain

$$\mathbf{I}^{(t)}(\boldsymbol{\theta}) = \begin{pmatrix} \mathbf{I}_{\boldsymbol{\beta}}(\boldsymbol{\theta}) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{\sigma^2}(\boldsymbol{\theta}) & \mathbf{I}_{\sigma^2\boldsymbol{\delta}}(\boldsymbol{\theta}) \\ \mathbf{0} & \mathbf{I}_{\boldsymbol{\delta}\sigma^2}(\boldsymbol{\theta}) & \mathbf{I}_{\boldsymbol{\delta}}(\boldsymbol{\theta}) \end{pmatrix}, \tag{31}$$

where $\mathbf{I}_{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \mathbf{X}^\top \mathbf{R}_{\boldsymbol{\delta}}^{-1}\mathbf{X}/n$, $\mathbf{I}_{\sigma^2}(\boldsymbol{\theta}) = 1/(2\sigma^4)$, the jth component of $\mathbf{I}_{\sigma^2\boldsymbol{\delta}}(\boldsymbol{\theta})$ is $\mathrm{tr}[\mathbf{R}_{\boldsymbol{\delta}}^{-1}(\partial \mathbf{R}_{\boldsymbol{\delta}}/\partial\delta_j)]/(2n\sigma^2)$, and the (j_1,j_2)th entry of $\mathbf{I}_{\boldsymbol{\delta}}(\boldsymbol{\theta})$ is $\mathrm{tr}[\mathbf{R}_{\boldsymbol{\delta}}^{-1}(\partial \mathbf{R}_{\boldsymbol{\delta}}^{-1}/\partial\delta_{j_1})\mathbf{R}_{\boldsymbol{\delta}}^{-1}(\partial \mathbf{R}_{\boldsymbol{\delta}}^{-1}/\partial\delta_{j_2})]/(2n)$ for all $j, j_1, j_2 \in \{1,\ldots,r\}$. By Proposition 3.2, we approximate $\mathbf{I}(\boldsymbol{\theta})$ by $\mathbf{I}^{(t)}(\boldsymbol{\theta})$ if the algorithm converges, implying that we can use $\{\mathbf{I}^{(t)}(\boldsymbol{\theta})\}^{-1}$ for a sufficiently large t to approximate the variance-covariance matrix of $\hat{\boldsymbol{\theta}}$. Because the Fisher-scoring algorithm converges fast, t is usually small (e.g. $t \le 50$).

Corollary 3.1: *Suppose that all assumptions of Proposition 3.1 hold. Let* $\mathbf{I}(\boldsymbol{\theta})$ *be the matrix converged by* $\mathbf{I}^{(t)}(\boldsymbol{\theta})$ *for sufficiently large t. If* $\mathbf{I}(\boldsymbol{\theta}_0)$ *is positive definite, then* $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \rightsquigarrow \mathcal{N}[\mathbf{0}, \mathbf{I}^{-1}(\boldsymbol{\theta}_0)]$ *as* $n \to \infty$. *If* $\mathbf{I}(\boldsymbol{\theta}_0)$ *is not positive definite, a similar result can be derived if the generalized inverse of* $\mathbf{I}(\boldsymbol{\theta}_0)$ *is used.*

Proof: The conclusion can be directly implied by Propositions 3.1 and 3.2. ■

4. Specification

We specify our method to two kinds of GLMMs. The first is longitudinal data for count. This problem has been well addressed by the Laplace approximation and the MCMC algorithm. The reason is that the dimension of the intractable integral given by (6) is usually low. The second is spatial data for count. This is a difficult problem since the dimension of the intractable integral given by (6) is often equal to the sample size. Since our method does not need to evaluate the intractable integral, it is not affected by the dimension.

4.1. Longitudinal data

We specify our method to GLMMs under the framework of repeated measurements. Longitudinal data sets are derived if repeated measures are recorded over a period of time. Repeated measurement data sets consist of repeated observations of a response variable and a set of explanatory variables for individual subjects. Individual subjects are called clusters. Dependence between clusters is ignored. Thus, only the covariance within clusters is needed. The covariance structure is accounted for by a statistical model with random effects. A common method is to assume that the random effects are independent between clusters.

Suppose that a longitudinal data set for count has K clusters. Let $\boldsymbol{y}_k = (y_{k1},\ldots,y_{kn_k})^\top$ for $k \in \{1,\ldots,K\}$ be independent random vectors with the distribution of y_{ki} given by $f(y_{ki}) = \exp[\{y_{ki}\omega_{ki} - b(\omega_{ki})\}/a(\phi) + c(y_{ki},\phi)]$. The total sample size of the data is $n = \sum_{k=1}^K n_k$. The response vector of the entire data is $\boldsymbol{y} = (\boldsymbol{y}_1^\top,\ldots,\boldsymbol{y}_K^\top)^\top$. Suppose that the conditional mean of y_{ki} is modelled by $g(\mu_{ki}) = \mathbf{x}_{ki}^\top\boldsymbol{\beta} + \mathbf{z}_{ki}^\top\boldsymbol{\gamma}_k$, where $\mu_{ki} = b'(\omega_{ki})$ and the $\boldsymbol{\gamma}_k$s are iid $\mathcal{N}(\mathbf{0},\mathbf{V}_{\boldsymbol{\delta}})$. Let $\boldsymbol{\mu}_k = (\mu_1,\ldots,\mu_{kn_k})^\top$, $\mathbf{X}_k = (\boldsymbol{x}_{k1},\ldots,\boldsymbol{x}_{kn_k})^\top$, $\mathbf{Z}_k = (\boldsymbol{z}_{k1},\ldots,\boldsymbol{z}_{kn_k})^\top$, $\mathbf{X} = (\mathbf{X}_1^\top,\ldots,\mathbf{X}_K^\top)^\top$, $\mathbf{Z} = \mathrm{diag}(\mathbf{Z}_1,\ldots,\mathbf{Z}_K)$, $\boldsymbol{\mu} = (\boldsymbol{\mu}_1^\top,\ldots,\boldsymbol{\mu}_K^\top)^\top$, and

$\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^\top, \ldots, \boldsymbol{\gamma}_K^\top)^\top$. Then, the GLMM can be expressed as $g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}$ and $\boldsymbol{\gamma} \sim \mathcal{N}\{\mathbf{0}, \mathrm{diag}_K(\mathbf{V}_\delta)\}$, where $\mathrm{diag}_K(\mathbf{V}_\delta)$ is a diagonal matrix obtained by repeating $\mathbf{V}_\delta$ K times. For each specific k, the model is

$$g(\boldsymbol{\mu}_k) = \mathbf{X}_k^\top \boldsymbol{\beta} + \mathbf{Z}_k \boldsymbol{\gamma}_k, \tag{32}$$

where $\boldsymbol{\gamma}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{V}_\delta)$ independently.

Assume that $\boldsymbol{\theta}^{(t)}$ and $\boldsymbol{\gamma}^{(t)}$ have been obtained in the tth iteration. Then, (15) becomes $\eta_{ki}^{(t)} = \boldsymbol{x}_{ki}^\top \boldsymbol{\beta}^{(t)} + \boldsymbol{z}_{ki}^\top \boldsymbol{\gamma}_k^{(t)}$, where $\boldsymbol{\gamma}_k^{(t)}$ is the predicted value of $\boldsymbol{\gamma}_k$ given by the tth iteration, implying that we can carry out (16) and (17) to compute $w_{ki}^{(t)}$ and $u_{ki}^{(t)}$, the working weight and response of the ith record in kth cluster, respectively. The working model becomes $\boldsymbol{u}_k^{(t)} = \mathbf{X}_k \boldsymbol{\beta} + \mathbf{Z}_k \boldsymbol{\gamma}_k + \boldsymbol{\epsilon}_k$ for $k \in \{1, \ldots, K\}$, where $\boldsymbol{u}_k^{(t)} = (u_{k1}^{(t)}, \ldots, u_{kn_K}^{(t)})^\top$, and $\boldsymbol{\epsilon}_k \sim \mathcal{N}(\mathbf{0}, \sigma^2\{\mathbf{W}_k^{(t)}\}^{-1})$ independently, where $\mathbf{W}_k^{(t)} = \mathrm{diag}(w_{k1}^{(t)}, \ldots, w_{kn_k}^{(t)})$.

To carry out the next iteration, we need $\boldsymbol{\theta}^{(t+1)}$ and $\boldsymbol{\gamma}_k^{(t+1)}$ for all $k \in \{1, \ldots, K\}$. It relies on the maximum-likelihood approach to

$$\boldsymbol{u}_k^{(t)} = \mathbf{X}_k \boldsymbol{\beta} + \mathbf{Z}_k \boldsymbol{\gamma}_k + \boldsymbol{\epsilon}_k, \quad k = 1, \ldots, K, \tag{33}$$

where $\boldsymbol{\epsilon}_k \sim \mathcal{N}(\mathbf{0}, \sigma^2\{\mathbf{W}_k^{(t)}\}^{-1})$ independently. Note that (33) is a weighted normal model for longitudinal data. The computation has been well addressed. Numerical algorithms for the MLE of (33) can be found in many software packages (e.g. lme4 in R or the proc glimmix in SAS). By migrating those to IRWLSR, we can derive the MLEs for any GLMMs with any reasonable link functions, including the GLMMs for binomial or Poisson longitudinal data.

4.2. Spatial data

A spatial GLMM for count is developed by hierarchical GLMs for count with spatially correlated or autocorrelated random effects. It contains at least two hierarchies. The first hierarchy specifies a spatial GLM for count given the random effects. The second hierarchy specifies the distribution of the random effects. To incorporate spatial correlation or autocorrelation, we specify $\mathbf{V}_\delta$ in (3) by spatial models, including the geostatistical [5], conditional autoregressive (CAR) [17,18] and spatial autoregressive (SAR) [19] models.

Suppose that a study region has been partitioned into n spatial units. Let y_i be the response for count and $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{i(p-1)})^\top$ be the vector of explanatory variables collected from the ith unit, for all $i \in \{1, \ldots, n\}$. A spatial GLMM is

$$g(\mu_i) = \xi_i + \boldsymbol{x}_i^\top \boldsymbol{\beta} + \gamma_i, \quad i = 1, \ldots, n, \tag{34}$$

where γ_i is the ith component of $\boldsymbol{\gamma}$ and ξ_i is an offset term, which is related to at-risk population size if the Poisson spatial GLMM is used [20,21].

The variance–covariance matrix $\mathbf{V}_\delta$ in (3) is defined by spatial models. If the geostatistical model is used, then $\mathbf{V}_\delta$ is constructed by a stationary covariance function, leading to the (i,j)th entry of $\mathbf{V}_\delta$ as $v_{ij,\delta} = c_\delta(\boldsymbol{d}_{ij})$ for all $i,j \in \{1, \ldots, n\}$, where $\boldsymbol{d}_{ij}$ is the difference between the locations of units i and j and $c_\delta(\cdot)$ is a stationary covariance function.

To ensure $\mathbf{V}_{\boldsymbol{\delta}}$ to be positive definite, a parametric model of $c_{\boldsymbol{\delta}}(\cdot)$ is used. One of the most popular models is the Matérn defined by

$$c_{\boldsymbol{\delta}}(\boldsymbol{d}) = \delta_1 \frac{(\delta_2\|\boldsymbol{d}\|)^{\delta_3}}{2^{\delta_3-1}\Gamma(\delta_3)} K_{\delta_3}(\delta_2\|\boldsymbol{d}\|), \tag{35}$$

where $\boldsymbol{d}$ is the difference between locations, $\boldsymbol{\delta} = (\delta_1, \delta_2, \delta_3)^\top$ is a three-dimensional parameter vector, $K_{\delta_3}(\cdot)$ is a modified Bessel function of the second kind, δ_1, δ_2 and δ_3 are variance, scale and smoothness parameters, respectively. The Matérn family is isotropic in space. It contains the exponential covariance function as a special case given by $\delta_3 = 0.5$. The model was first proposed by Matérn [22] and has received more attention since some theoretical work by Handcock and Stein [23] and Stein [24]. A nice review and discussion on Matérn family is given by Guttorp and Gneiting [25]. The Matérn family has been used [26,27]. In addition to the geostatistical model, one can use the CAR or SAR models. Both of them use neighbouring information to define $\mathbf{V}_{\boldsymbol{\delta}}$. Both has $\boldsymbol{\delta} = (\delta_1, \delta_2)^\top$, where $\boldsymbol{\delta}$ is a two-dimensional parameter vector.

In all of the three spatial GLMMs for count, based on $\boldsymbol{\theta}^{(t)}$ and $\boldsymbol{\gamma}^{(t)}$ given by the previous iteration, we can express (15) as $\eta_i^{(t)} = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \gamma_i^{(t)}$ for all $i \in \{1, \ldots, n\}$, where $\gamma_i^{(t)}$ is the ith component of $\boldsymbol{\gamma}^{(t)}$. We use (16) and (17) to compute $w_i^{(t)}$ and $u_i^{(t)}$ for all $i \in \{1, \ldots, n\}$. The working model becomes $\boldsymbol{u}^{(t)} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\gamma} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\{\mathbf{W}^{(t)}\}^{-1})$. To carry out the next iteration, we need to numerically solve the maximum likelihood for the spatial linear model for normal data as

$$\boldsymbol{u}^{(t)} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\gamma} + \boldsymbol{\epsilon}, \tag{36}$$

where the variance matrix of $\boldsymbol{\gamma}$ is given by a spatial model, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\{\mathbf{W}^{(t)}\}^{-1})$ is the error vector, and $\boldsymbol{\gamma}$ and $\boldsymbol{\epsilon}$ are independent. The MLE of (36) can be solved by the profile maximum-likelihood approach given by Appendix A.

5. Simulation

We carried out simulation studies to evaluate the performance of the IRWLSR for binomial and Poisson data. We evaluated the performance for binomial data under the framework of longitudinal studies, where the dimension of the intractable integral given by (4) was low. We considered two link functions: the logistic and inverse CDF links. We selected the logistic link because of its popularity. As logistic models in longitudinal studies can be fitted by many software packages, we compared our results with those given by the glmer function in the lme4 package of R. We selected the inverse CDF link because we wanted to demonstrate the flexibility of our method. We examined the existing inverse CDF links used by the lme4 and found that it did not contain the CDF of the t-distribution. Then, we decided to use this link to demonstrate the flexibility of IRWLSR. We studied Poisson data because we wanted to demonstrate the feasibility of our method when high-dimensional intractable integral was present in the likelihood function. We chose the spatial Poisson model with the covariance matrix of $\boldsymbol{\gamma}$ given by the exponential covariance function in (34), which was derived by taking $\delta_3 = 0.5$ in (35). The dimension of the intractable integral in the likelihood function was equal to the sample size. It was hard to implement the Laplace approximation or the MCMC algorithm, but our IRWLSR could still be applied.

5.1. Logistic model

We assumed that the logistic GLMM for binomial data was applied to a longitudinal study with identical cluster sizes. Let K be the number of clusters and m be their sizes. For each selected K and m, we assumed that y_{ki} for all $k \in \{1, \ldots, K\}$ and $i \in \{1, \ldots, m\}$ were conditionally independent Bernoulli random variables with π_{ki} given by

$$\log \frac{\pi_{ki}}{1-\pi_{ki}} = \beta_0 + \beta_1 x_{ki1} + \beta_2 x_{ki2} + \beta_3 x_{ki3} + \gamma_{k0} + \gamma_{k1} x_{ki} = \boldsymbol{x}_{ki}^\top \boldsymbol{\beta} + \boldsymbol{z}_{ki}^\top \boldsymbol{\gamma}_k, \tag{37}$$

where $\boldsymbol{x}_{ki} = (1, x_{ki1}, x_{ki2}, x_{ki3})^\top$ represented explanatory vectors for fixed effects, $\boldsymbol{z}_{ki} = (1, x_{ki1})^\top$ represented explanatory vectors for random effects, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^\top$ represented the parameter vector for fixed effects, $\boldsymbol{\gamma}_k = (\gamma_{k0}, \gamma_{k2})^\top$ represented vectors for random effects. We generated x_{ki1}, x_{ki2} and x_{ki3} independently from $\mathcal{N}(0, 0.5^2)$, and $\boldsymbol{\gamma}_k$ independently from $\mathcal{N}(\mathbf{0}, \mathbf{V})$, where $\mathbf{V}$ was 2×2 matrix for variance components. We generated y_{ki} conditionally independently from (37) with fixed $\beta_0 = 0$, $\beta_1 = 0.5$, $\beta_2 = 0.4$, $\beta_3 = 0.3$, $v_{00} = v_{11} = 0.5^2$ and $v_{01} = v_{10} = 0.1$, where $v_{j_1 j_2}$ was the $(j_1 + 1, j_2 + 1)$th entry of $\mathbf{V}$.

We implemented IRWLSR to the logistic linear mixed effects model given by (27) for the data. Following the traditional IRWLS, we chose the initial working response and weight values as $u_{ki}^{(0)} = \log[(y_{ki} + 0.5)/(1.5 - y_{ki})]$ and $w_{ki}^{(0)} = (y_{ki} + 0.5)(1.5 - y_{ki})/4$ for all $k \in \{1, \ldots, K\}$ and $i \in \{1, \ldots, m\}$, respectively. We modified (18) as

$$\mathbf{u}_k^{(t)} = \mathbf{X}_k \boldsymbol{\beta} + \mathbf{Z}_k \boldsymbol{\gamma}_k + \boldsymbol{\epsilon}_k \tag{38}$$

where $\mathbf{X}_k = (\boldsymbol{x}_{k1}^\top, \ldots, \boldsymbol{x}_{km}^\top)^\top$, $\mathbf{Z}_k = (\boldsymbol{z}_{k1}^\top, \ldots, \boldsymbol{z}_{km}^\top)^\top$, $\boldsymbol{\gamma}_k \sim^{iid} \mathcal{N}(\mathbf{0}, \mathbf{V})$, and $\boldsymbol{\epsilon}_k \sim^{ind} \mathcal{N}(\mathbf{0}, \{\mathbf{W}_k^{(t)}\}^{-1})$ with $\mathbf{W}_k^{(t)} = \mathrm{diag}(w_{k1}^{(t)}, \ldots, w_{km}^{(t)})$ for all $k \in \{1, \ldots, K\}$.

We used the lmer function in the lmer4 package of R to fit (38). It provided $\boldsymbol{\theta}^{(t+1)} = (\boldsymbol{\beta}^{(t+1)}\}^\top, \mathbf{V}^{(t+1)})$, where $\boldsymbol{\beta}^{(t+1)}$ and $\mathbf{V}^{(t+1)}$ were the MLEs of $\boldsymbol{\beta}$ and $\mathbf{V}$ under (38). We then modified (26) as

$$\boldsymbol{\gamma}_k^{(t+1)} = \mathbf{V}^{(t+1)} \mathbf{Z}_k^\top \mathbf{M}_k^{(t+1)} \mathbf{u}^{(t)} \tag{39}$$

for all $k \in \{1, \ldots, K\}$, where $\mathbf{M}_k^{(t+1)} = \{\mathbf{R}_k^{(t+1)}\}^{-1} - \{\mathbf{R}_k^{(t+1)}\}^{-1} \mathbf{X}_k (\mathbf{X}\{\mathbf{R}_k^{(t+1)}\}^{-1} \mathbf{X}^\top)^{-1}$ $\mathbf{X}_k \{\mathbf{R}_k^{(t+1)}\}^{-1}$ and $\mathbf{R}_k^{(t+1)} = \mathbf{Z}_k \mathbf{V}^{(t+1)} \mathbf{Z}_k^\top + \{\mathbf{W}_k^{(t)}\}^{-1}$.

By taking $t = 0$ in (38) and (39), we obtained $\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\gamma}_k^{(1)}$. If $t \geq 1$, then we defined $\eta_{ki}^{(t)} = \boldsymbol{x}_{ki}^\top \boldsymbol{\beta}^{(t)} + \boldsymbol{z}_{ki}^\top \boldsymbol{\gamma}_k^{(t)}$, $\pi_{ki}^{(t)} = e^{\eta_{ki}^{(t)}}/(1 + e^{\eta_{ki}^{(t)}})$, $w_{ki}^{(t)} = \pi_{ki}^{(t)}(1 - \pi_{ki}^{(t)})$, and $u_{ki}^{(t)} = \eta_{ki}^{(t)} + (y_{ki} - \pi_{ki}^{(t)})/[\pi_{ki}^{(t)}(1 - \pi_{ki}^{(t)})]$, for all $k \in \{1, \ldots, K\}$ and $i \in \{1, \ldots, m\}$. We then carried out the next integration. This meant that our entire method could be employed. In the end, we obtained $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{V}}$, the MLEs of $\boldsymbol{\beta}$ and $\mathbf{V}$. The result was denoted by $\hat{\boldsymbol{\theta}} = (\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \hat{\theta}_4, \hat{\theta}_5, \hat{\theta}_6)^\top = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{v}_{00}, \hat{v}_{01}, \hat{v}_{11})^\top$.

We simulated 1000 data sets from (37) for each selected m and K. To evaluate the performance of the IRWLSR, we computed the MSE values for each component of $\hat{\boldsymbol{\theta}}$ by

$$\mathrm{MSE}(\hat{\theta}_j) = \frac{1}{1000} \sum_{\ell=1}^{1000} (\hat{\theta}_{j,\ell} - \theta_{j0})^2, \tag{40}$$

Table 1. Simulations (with 1000 replicates) for root MSEs of the MLEs given by our IRWLSR for selected *m* and *K* in the logistic mixed effects model.

		$\hat{\boldsymbol{\theta}}$						
m	K	β_0	β_1	β_2	β_3	v_{00}	v_{01}	v_{11}
10	100	0.077	0.141	0.136	0.137	0.099	0.107	0.218
	200	0.057	0.104	0.094	0.097	0.070	0.078	0.162
	500	0.036	0.068	0.062	0.060	0.045	0.052	0.116
	1000	0.025	0.053	0.046	0.044	0.035	0.040	0.093
20	100	0.067	0.107	0.093	0.092	0.068	0.071	0.146
	200	0.048	0.077	0.068	0.069	0.050	0.051	0.106
	500	0.030	0.054	0.043	0.043	0.034	0.036	0.076
	1000	0.021	0.042	0.032	0.032	0.028	0.027	0.060

Table 2. Simulations (with 1000 replicates) for the ratio of the root MSEs of the MLEs ($\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_{LA}$) given by our IRWLSR and Laplace approximation methods, respectively, where the Laplace approximation is carried out by the glmer function in the lme4 package of R.

		$[\mathrm{MSE}(\hat{\boldsymbol{\theta}})/\mathrm{MSE}(\hat{\boldsymbol{\theta}}_{LA})]^{1/2}$						
m	K	β_0	β_1	β_2	β_3	v_{00}	v_{01}	v_{11}
10	100	0.956	0.958	0.964	0.960	0.932	0.883	0.943
	200	0.959	0.967	0.975	0.967	0.932	0.890	0.943
	500	0.960	1.038	1.015	0.974	0.934	0.938	0.844
	1000	0.963	1.039	1.046	1.015	0.932	1.021	0.739
20	100	0.968	0.978	0.970	0.969	0.954	0.898	0.922
	200	0.968	1.008	0.984	0.977	0.968	0.910	0.917
	500	0.969	1.066	1.029	0.981	1.028	0.970	0.879
	1000	0.972	1.095	1.051	1.028	1.101	1.046	0.854

where $\hat{\theta}_{j,\ell}$ was the estimate of θ_j from the ℓth data set and θ_{j0} was the true value of the jth component of $\boldsymbol{\theta}$. The results are given in Table 1. We found that the root MSE values were small comparing to their true values. The MSE values decreased as either m or K increased. This was expected because n increased if either m or K increased.

We next compare our method and the traditional Laplace approximation method. We compared MSEs of $\hat{\boldsymbol{\theta}}$ given by IRWLSR with those of $\hat{\boldsymbol{\theta}}_{LA}$, where $\hat{\boldsymbol{\theta}}_{LA}$ was given by the Laplace approximation based on the glmer function in the lme4 package of R. The MSEs of $\hat{\boldsymbol{\theta}}_{LA}$ were derived similarly as those of $\hat{\boldsymbol{\theta}}$ given by (40). We computed ratios of root MSEs of each component of $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_{LA}$ (Table 2). Since most of the ratios were less than one, we conclude that IRWLSR is more efficient than Laplace approximation.

5.2. Inverse CDF model

We assumed that the GLMM for binomial data with the inverse CDF link was applied to a longitudinal study with identical cluster sizes. We assumed that the response variable y_{ki} for all $k \in \{1, \dots, K\}$ and $i \in \{1, \dots, m\}$ were conditionally independent Bernoulli random variables with π_{ki} given by

$$\Psi^{-1}(\pi_{ki}) = \beta_0 + \beta_1 x_{ki1} + \beta_2 x_{ki2} + \beta_3 x_{ki3} + \gamma_{k0} + \gamma_{k1} x_{ki} = \boldsymbol{x}_{ki}^{\top}\boldsymbol{\beta} + \boldsymbol{z}_{ki}^{\top}\boldsymbol{\gamma}_k, \tag{41}$$

where $\boldsymbol{x}_{ki}$, $\boldsymbol{z}_{ki}$, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}_k$ represented those given by (37), respectively. We used the same $\boldsymbol{\beta}$ and $\mathbf{V}$ to generate $\boldsymbol{x}_{ki}$ and $\boldsymbol{\gamma}_k$. After that, we generated y_{ki} conditionally independent

Table 3. Simulations (with 1000 replicates) for root MSEs of the MLEs given by the IRWLSR for selected m and K in the inverse CDF mixed effects model, where the link is the inverse CDF of the t_{10} distribution.

		$\hat{\boldsymbol{\theta}}$						
m	K	β_0	β_1	β_2	β_3	v_{00}	v_{01}	v_{11}
10	100	0.063	0.114	0.093	0.089	0.099	0.070	0.141
	200	0.045	0.090	0.070	0.065	0.071	0.053	0.096
	500	0.027	0.072	0.055	0.047	0.054	0.037	0.057
	1000	0.020	0.066	0.046	0.040	0.044	0.031	0.039
20	100	0.055	0.090	0.067	0.062	0.056	0.051	0.099
	200	0.038	0.071	0.053	0.048	0.042	0.036	0.067
	500	0.023	0.058	0.039	0.034	0.031	0.026	0.047
	1000	0.017	0.054	0.034	0.028	0.028	0.023	0.037

from (41), where we chose $\Psi(\cdot)$ as the CDF of the t_{10}-distribution. We used this link because it was not included in the lme4 package. The aim is to demonstrate that our method can be easily implied to any GLMMs with arbitrary links.

Following the traditional IRWLS, we chose initial working response and weight values as $u_{ki}^{(0)} = \Psi^{-1}[(y_{ki}+0.5)/2]$ and $w_{ki}^{(0)} = \psi^2(u_{ki}^{(0)})/\{\Psi(u_{ki}^{(0)})[1-\Psi(u_{ki}^{(0)})]\}$, for all $k \in \{1,\dots,K\}$ and $i \in \{1,\dots,m\}$, where $\psi(\cdot)$ was the PDF of the t_{10}-distribution. We obtained (38) and (39). By taking $t = 0$, we obtained $\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\gamma}_k^{(1)}$ for all $k \in \{1,\dots,K\}$. If $t \geq 1$, then we defined $\eta_{ki}^{(t)} = \boldsymbol{x}_{ki}^\top \boldsymbol{\beta}^{(t)} + \boldsymbol{z}_{ki}^\top \boldsymbol{\gamma}_k^{(t)}$, $w_{ki}^{(t)} = \psi^2(\eta_{ki}^{(t)})/\{\Psi(\eta_{ki}^{(t)})[1-\Psi(\eta_{ki}^{(t)})]$, and $u_{ki}^{(t)} = \eta_{ki}^{(t)} + [y_{ki} - \Psi(\eta_{ki}^{(t)})]/\psi(\eta_{ki}^{(t)})$. We obtained the next iteration. This meant that our method was applied. In the end, we obtained $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{V}}$, the MLEs of $\boldsymbol{\beta}$ and $\mathbf{V}$ by IRWLSR. We denoted them by $\hat{\boldsymbol{\theta}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{v}_{00}, \hat{v}_{01}, \hat{v}_{11})^\top$.

We simulated 1000 data sets from (41) for selected m and K. To evaluate the performance of the IRWLSR, we also computed the MSE values for each component of $\hat{\boldsymbol{\theta}}$ by a formula similar to (40). The results are given in Table 3. We still found that the root MSE values were small comparing to their true values, and they decreased as either m or K increased.

5.3. Spatial poisson model

We assumed that the GLMM for Poisson data with the log link was applied to spatial statistical data set with n spatial units. Each unit was represented by a point in the study region. We used a $K \times K$ lattice to represent these points. The $K \times K$ lattice contained $n = K^2$ points. The coordinates of these points were (i_1, i_2) for all $i_1, i_2 \in \{1,\dots,K\}$, respectively. Thus, the difference and the distance between the ith and jth units were $\boldsymbol{d}_{ij} = (j_1 - i_1, j_2 - i_2)$ and $d_{ij} = \|\boldsymbol{d}_{ij}\| = [(j_1-i_1)^2 + (j_2-i_2)^2]^{1/2}$, respectively. We used $\boldsymbol{d}_{ij}$ to define the Matérn covariance function given by (35). Following [28], we fixed $\delta_3 = 0.5$ such that the covariance function became $c_{\boldsymbol{\delta}}(\boldsymbol{d}) = \delta_1 e^{-\delta_2 d}$, where $d = \|\boldsymbol{d}\|$ was the distance between point locations, and $\boldsymbol{\delta} = (\delta_1, \delta_2)^\top$ is a two-dimensional parameter vector to be estimated by our approach.

For each selected K, we assumed that y_i for all $i \in \{1,\dots,n\}$ were conditionally independent Poisson random variables with μ_i given by

$$\log \mu_i = \log m_i + \beta_0 + \beta_1 x_i + \gamma_i = \log m_i + \boldsymbol{x}_i^\top \boldsymbol{\beta} + \gamma_i, \tag{42}$$

where m_i was the at-risk population size, leading to $\xi_i = \log m_i$ in (34), β_0 and β_1 were the parameters for fixed effects, γ_i were the normally distributed random effects, $\boldsymbol{x}_i = (1, x_i)^\top$, and $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$. The covariance between the random effects was given by $\mathrm{Cov}(\gamma_i, \gamma_j) = \delta_1 e^{-\delta_2 d_{ij}}$ for all $i, j \in \{1, \ldots, n\}$.

We generated m_i independently from $\mathcal{P}(\nu_i)$ with $\nu_i = 2 \times 10^4$ in the previous $[K/2]$ rows and $\nu_i = 10^4$ in the remaining rows, where $[\cdot]$ is the function of the integer part. For each selected K, we generated x_i independently from $\mathcal{N}(0, 0.5^2)$ and $\boldsymbol{\gamma}$ from $\mathcal{N}(\mathbf{0}, \mathbf{V}_{\boldsymbol{\delta}})$, where $\mathbf{V}_{\boldsymbol{\delta}}$ was an $n \times n$ matrix for variance components, and the (i,j)th entry of $\mathbf{V}_{\boldsymbol{\delta}}$ was $\mathrm{Cov}(\gamma_i, \gamma_j)$. We then generated y_i conditionally independently from (42) with fixed $\beta_0 = -5.0$ and $\beta_1 = 0.5$.

We implemented IRWLSR to (42). According to traditional IRWLS, we chose $u_i^{(0)} = \log(y_i + 0.5) - \log m_i$ and $w_i^{(0)} = y_i + 0.5$. We obtained a working model as

$$\mathbf{u}^{(t)} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\gamma} + \boldsymbol{\epsilon}, \tag{43}$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)^\top$ was the vector of spatial random effects, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\{\mathbf{W}^{(t)}\}^{-1})$ was the error vector, and $\mathbf{W}^{(t)} = \mathrm{diag}(w_1^{(t)}, \ldots, w_n^{(t)})$ was the working weight matrix. For the next iteration, we numerically calculated the MLEs of $\boldsymbol{\theta} = (\beta_0, \beta_1, \delta_1, \delta_2)^\top$ and σ^2 given by (43), where we treated σ^2 as a nuisance parameter.

We used the profile maximum-likelihood approach to fit (43). Since the optimization problem given by (25) only involved two parameters, the computation was fast. By taking $t = 0$ in (43), we obtained $\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\gamma}^{(1)}$. If $t \geq 1$, then we defined $\eta_i^{(t)} = \mathbf{x}_i^\top \boldsymbol{\beta}^{(t)} + \gamma_i^{(t)} - \log m_i$, $w_i^{(t)} = m_i e^{\eta_i^{(t)}}$, and $u_i^{(t)} = \eta_i^{(t)} - (y_i - w_i^{(t)})/w_i^{(t)}$. In the end, we obtained $\hat{\boldsymbol{\theta}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\delta}_1, \hat{\delta}_2)^\top$ by IRWLSR.

We simulated 1000 data sets from (42) for selected δ_1, δ_2 and K. To evaluate the performance of IRWLSR, we computed the MSE values for each component of $\hat{\boldsymbol{\theta}}$. The results are given in Table 4. We found that the root MSE values were significantly affected by K, δ_1, and δ_2 values. The root MSE values decreased as K increased. This was expected as the sample size increased with K.

Table 4. Simulations (with 1000 replicates) for root MSEs of the MLEs given by the IRWLSR for selected K, δ_1 and δ_2 in the spatial Poisson mixed effects model, where $\beta_0 = -5.0$ and $\beta_1 = 0.5$.

			$\hat{\boldsymbol{\theta}}$			
K	δ_1	δ_2	β_0	β_1	δ_1	δ_2
15	0.1	0.5	0.077	0.032	0.072	0.144
		1.0	0.052	0.036	0.060	0.169
	0.2	0.5	0.134	0.037	0.156	0.140
		1.0	0.068	0.055	0.112	0.184
20	0.1	0.5	0.063	0.023	0.061	0.111
		1.0	0.035	0.032	0.048	0.140
	0.2	0.5	0.099	0.029	0.098	0.097
		1.0	0.055	0.042	0.106	0.150
25	0.1	0.5	0.054	0.018	0.038	0.095
		1.0	0.031	0.023	0.040	0.123
	0.2	0.5	0.082	0.021	0.073	0.078
		1.0	0.046	0.032	0.098	0.124

6. Application

We applied our method to the Guangxi infant mortality data set, which was previously analyzed by Zhang and Lin [29]. Guangxi is one of five autonomous ethnic provinces regions in China. The province contains 110 counties. The total area is about 236,000km^2. The total population size was about 49.6 million in 2020.

We obtained the county-level infant birth and death counts from the 2000 Census in China. The data set contained 14,508 total infant mortality and 603,910 total infant birth counts. The province-level infant mortality rate was 2402 per 100,000. The county-level infant mortality rates varied substantially between counties. The lowest rate was 248 per 100,000 at Jiangnan Qu, which was close to the capital city of the province. The highest rate was 7260 per 100,000 at Shangsi Xian, which was close to the border between China and Vietnam. The county-level elevation varies from 20 to 1140 metres above sea level. The value was low in the southeastern area but high in the western and northwestern area. As it was highly correlated with medical and socioeconomic resources, we treated elevation as an important explanatory variable. To confirm this, we carried out an initial study via a loglinear model for Poisson data as

$$\log \mu_i = \log m_i + \beta_0 + \beta_1 x_i, \tag{44}$$

where μ_i was the expected value of the county-level infant morality count, y_i was the observed value of the county-level infant morality count, m_i was the county-level infant birth, and x_i was the county-level elevation value. We fitted (44) by a standard R package. We had $\hat{\beta}_0 = -3.9774$ and $\hat{\beta}_1 = 0.8498$ with standard errors $s(\hat{\beta}_0) = 0.0133$ and $s(\hat{\beta}_1) = 0.0329$, respectively, indicating that elevation was important (Figure 1).

A significant defect of the previous analysis [29] was the ignorance of spatial dependency. The main reason was the difficulty in the computation of the MLEs in spatial Poisson models for count, where the dimension of the intractable integral in the likelihood function given by (4) was equal to the sample size. This difficulty was completely overcome by IRWLSR. In particular, we added a spatial random effect in (44), such that it became (42). We used the Matérn covariance function given by (35) to model the dependence between the random effects, where the distance was given by kilometres. We used the methods given by Sections 3.2 and 3.3 to compute the MLEs of model parameters and their standard errors, respectively.

We considered two cases in the implementation of the model. In the first case, we fixed $\delta_3 = 0.5$. We had $\boldsymbol{\delta} = (\delta_1, \delta_2)^\top$, and $\boldsymbol{\theta} = (\beta_0, \beta_1, \delta_1, \delta_2)^\top$. We obtained $\hat{\beta}_0 = -4.017$, $\hat{\beta}_1 = 0.966$, $\hat{\delta}_1 = 0.0190$ and $\hat{\delta}_2 = 0.0196$. We calculated their standard errors by the Fisher information given by (31). We obtained $s(\hat{\beta}_0) = 0.142$, $s(\hat{\beta}_1) = 0.262$, $s(\hat{\delta}_1) = 0.0096$ and $s(\hat{\delta}_2) = 0.0074$, respectively. All of them were significant based on their p-values. In the second case, we treated δ_3 as a parameter. We estimated it. We had $\boldsymbol{\delta} = (\delta_1, \delta_2, \delta_3)^\top$ and $\boldsymbol{\theta} = (\beta_0, \beta_1, \delta_1, \delta_2, \delta_3)^\top$. As we found that $\hat{\delta}_3$ was extremely large (≥ 100), we roughly treated it as infinity, leading to the Gaussian covariance function in (4). Then, we had $\hat{\beta}_0 = -4.016$, $\hat{\beta}_1 = 0.979$, $\hat{\delta}_1 = 0.0136$ and $\hat{\delta}_2 = 0.359$. We calculated their standard errors and obtained $s(\hat{\beta}_0) = 0.122$, $s(\hat{\beta}_1) = 0.245$, $s(\hat{\delta}_1) = 0.0062$ and $s(\hat{\delta}_2) = 0.0636$, respectively. We studied the difference between the two cases by the likelihood ratio test. The value of the likelihood ratio statistic was 2.71. It was insignificant based on the χ_1^2-approximation. Therefore, we could use $\delta_3 = 0.5$, indicating that the exponential covariance function was acceptable.

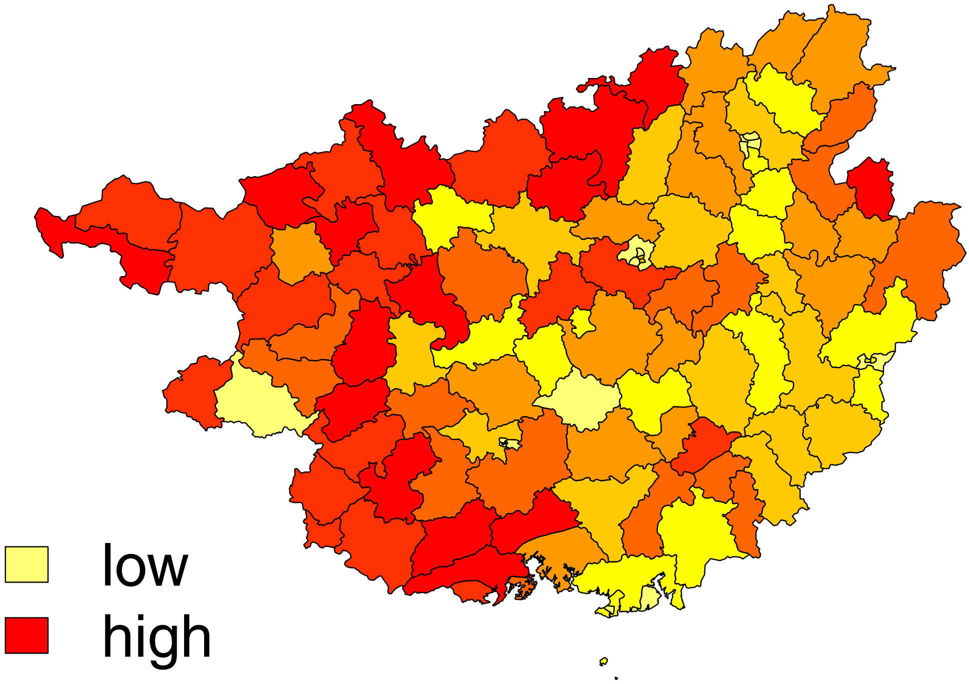

Figure 1. County-level infant mortality rates in Guangxi, China, 2000.

We also investigated two well known previous methods. To implement the Laplace approximation, we studied numerical issues for $\hat{\boldsymbol{b}}_{\boldsymbol{\theta}}$ given by (6) for each candidate $\boldsymbol{\theta}$. We needed to solve the optimization problem for a 110-dimensional parameter. We checked the eigenvalues of the Hessian matrix of the objective function used in the Laplace approximation at different values of $\boldsymbol{\theta}$. We found the Hessian matrix was not always negative definite. We concluded that the objective function might contain many local optimizers. This made it difficult to compute $\hat{\boldsymbol{b}}_{\boldsymbol{\theta}}$ in (6) since $\hat{\boldsymbol{b}}_{\boldsymbol{\theta}}$ must be the global maximizer. In addition, even if $\hat{\boldsymbol{b}}_{\boldsymbol{\theta}}$ could be successfully derived, according to [8], the Laplace approximation approach might contain serious bias because the dimension of the intractable integral was not $o(n^{1/3})$. In the MCMC, we wanted to update samples of $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ based on probability values derived by the combination of a proposal function, and the previous and current samples. We tried a number of proposal functions and all of them had high probabilities to reject the new samples. In many cases, the MCMC provided a sample value with over 99% and the rest sample values with lower than 1%. We investigated the reason and found that the problem was caused by the dimension of $\boldsymbol{\gamma}$. Then, we decided to try the Gibbs sampler approach. To update individual parameter values, we needed to compute the inverse and determinant of a 109-dimensional variance-covariance matrix in the conditional distribution of the current component of $\boldsymbol{\gamma}$ given the remaining components of $\boldsymbol{\gamma}$. This was extremely time-consuming.

7. Discussion

We propose iteratively reweighted least squares with random effects (IRWLSR) for MLEs of generalized linear mixed effects models (GLMMs), which successfully overcomes the difficulty caused by high dimensional intractable integrals in fitting GLMMs for non-normal data. An advantage is that the major computation and prediction are carried out by a working linear mixed effects model (WWLMM) for normal data. The distribution of the GLMM and the link function are only used in updating the working responses and weights in the WWLMM. Since the entire method does not need any numerical evaluations of intractable integrals, it can be easily implemented even if high-dimensional intractable integrals appear in the likelihood function.

IRWLSR successfully reduces the computational burden of MLE of GLMMs for non-normal data. The computational complexity is equal to that of LMMs for normal data. Given that an algorithm for a specific LMM exists, we can modify it to MLEs of a corresponding GLMM. Therefore, the important issue is to develop an efficient algorithm for MLEs of LMMs for normal data but not GLMMs for non-normal data. This kind of problems has not been completely solved yet. An example is the geostatistical model for normal data, displayed by (36). Although the likelihood function can be expressed by closed-forms, the size of the variance–covariance matrix may be extremely large if the number of observations is only moderately large, leading to a difficulty in the computation of the MLEs of a spatial statistical model for normal data. This problem has been previously studied by many articles [28,30–33].

We believe that IRWLSR can be combined with hierarchical generalized linear model (HGLM) for count data. Assume that many hierarchical levels are involved. The first level is given by a GLM. The second level is given by normally distributed random effects with parameters to be modelled by other levels. Note that the HGLM becomes a GLMM for count data given parameters of the second level. We can use IRWLSR to estimate the conditional MLEs of model parameters given parameters of the second level. Then, we can study the HGLM based on the working normal model given by IRWLSR, indicating that it is possible to use tools for normal data to study models for non-normal data. This is left to future research.

Acknowledgments

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

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

[1] Rue H, Martino S, Chopin N. Approximate Bayesian inference for laten Gaussian models by using integrated nested Laplace approximation. J R Stat Soc Ser B. 2009;71:319–392.

[2] Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. J Am Stat Assoc. 1993;88:9–25.

[3] Breslow NE, Lin X. Bias correlation in generalised linear mixed models with a single component of dispersion. Biometrika. 1995;82:81–91.

[4] Liu Q, Pierce DA. A note on Gauss-Hermite quadrature. Biometrika. 1994;81:624–629.

[5] Zhang H. On estimation and prediction for spatial generalized linear mixed models. Biometrics. 2002;58:129–136.

[6] Zeger SL, Karim MR. Generalized linear models with random effects: a Gibbs sampling approach. J Am Stat Assoc. 1991;86:79–86.

[7] Best NG, Wakefield JC. Accounting for inaccuracies in population counts can case registration in cancer mapping studies. J R Stat Soc Ser A. 1999;162:363–382.

[8] Shun Z, McCullagh P. Laplace approximation for high dimensional integrals. J R Stat Soc Ser B. 1995;4:749–760.

[9] Green PJ. Iteratively weighted least squares for maximum likelihood estimation and some robust and resistant alternatives. J R Stat Soc Ser B. 1984;46:149–192.

[10] Fong Y, Rue H, Wakefield J. Bayesian inference for generalized linear mixed models. Biostatistics. 2010;11:397–412.

[11] Zhang T. General Gaussian estimation. J Multivar Anal. 2019;169:234–247.

[12] Lin X, Breslow NE. Bias correction in generalized linear mixed models with multiple components of dispersion. J Am Stat Assoc. 1996;91:1007–1016.

[13] McCulloch CE, Searle SR, Neuhaus JR. Generalized, linear, and mixed models. Hoboken, NJ: John Wiley & Sons; 2008.

[14] McCullagh P. Quasi-likelihood functions. Ann Stat. 1983;11:59–67.

[15] Schenker N, Welsh AH. Asymptotic results for multiple imputation. Ann Stat. 1988;16: 1550–1566.

[16] Wang N, Robins JM. Large-sample theory for parametric multiple imputation procedures. Biometrika. 1998;85:935–948.

[17] Besag J. Spatial interaction and the statistical analysis of lattice systems. J R Stat Soc, B. 1974;36:192–236.

[18] Cressie N, Chan N. Spatial modeling of regional variables. J Am Stat Assoc. 1989;84:393–401.

[19] Ord K. Estimation methods for models of spatial interaction. J Am Stat Assoc. 1975;70: 120–126.

[20] Liu Y, Liu Y, Zhang T. Wald-based spatial scan statistics for cluster detection. Comput Stat Data Anal. 2018;127:298–310.

[21] Zhang T, Lin G. Cluster detection based on spatial association and iterated residuals in generalized linear mixed models. Biometrics. 2009b;65:353–360.

[22] Matérn B. Spatial variation. 2nd ed. Berlin: Springer-Verlag; 1986.

[23] Handcock MS, Stein ML. A Bayesian analysis of Kriging. Technometrics. 1993;35:403–410.

[24] Stein ML. Interpolation of spatial data. some theory for Kriging. New York: Springer-Verlag; 1999.

[25] Guttorp P, Gneiting T. Studies in the history of probability and statistics XLIX: on the Matérn correlation family. Biometrika. 2006;93:989–995.

[26] Lee D, Shaddick G. Spatial modeling of air pollution in studies of its short-term health effects. Biometrics. 2010;66:1238–1246.

[27] North G, Wang J, Genton M. Correlation models for temperature fields. J Clim. 2011;24:5850–5862.

[28] Liang F, Cheng Y, Song Q, et al. A resampling-based stochastic approximation method for analysis of large geostatistical data. J Am Stat Assoc. 2013;108:325–339.

[29] Zhang T, Lin G. Spatial scan statistics in loglinear models. Comput Stat Data Anal. 2009a;53:2851–2858.

[30] Cressie N, Johannesson G. Fixed rank kriging for very large spatial data sets. J R Stat Soc B. 2008;70:209–226.

[31] Eidsvik J, Shaby BA, Reich BJ, et al. Estimation and prediction in spatial models with block composite likelihoods. J Comput Graph Stat. 2014;23:295–315.

[32] Fuentes M. Approximate likelihood for large irregularly spaced spatial data. J Am Stat Assoc. 2007;102:321–331.

[33] Kaufman CG, Schervish MJ, Nychka DW. Covariance tapering for likelihood-based estimation in large spatial data sets. J Am Stat Assoc. 2008;103:1545–1555.

Appendix 1. Maximum likelihood for LMMs

Let $\boldsymbol{u}$ be the n-dimensional response, $\mathbf{X}$ be the $n \times p$-dimensional design matrix for fixed effects, $\mathbf{Z}$ be the $n \times q$-dimensional design matrix for random effects. A general LMM can be expressed as

$$\boldsymbol{u} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\gamma} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{B}_\delta)$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{W})$ are independent, $\mathbf{W}$ is a known weight matrix, and $\boldsymbol{\delta} = (\delta_1, \dots, \delta_r)^\top$ is an r-dimensional parameter vector for the variance-components. The log-likelihood function of the model is

$$\ell(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\delta}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2}|\det(\mathbf{R}_\delta)| - \frac{1}{2\sigma^2}(\boldsymbol{u} - \mathbf{X}\boldsymbol{\beta})^\top\mathbf{R}_\delta^{-1}(\boldsymbol{u} - \mathbf{X}\boldsymbol{\beta}),$$

where $\mathbf{R}_\delta = \mathbf{Z}\mathbf{B}_\delta\mathbf{Z}^\top + \mathbf{W}^{-1}$. Given $\boldsymbol{\delta}$, the conditional MLE of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}}_\delta = (\mathbf{X}^\top\mathbf{R}_\delta^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{R}_\delta^{-1}\boldsymbol{u}$ and the conditional MLE of σ^2 is $\hat{\sigma}_\delta^2 = \boldsymbol{u}^\top\mathbf{M}_\delta\mathbf{u}/n$, where $\mathbf{M}_\delta = \mathbf{R}_\delta^{-1} - \mathbf{R}_\delta^{-1}\mathbf{X}(\mathbf{X}^\top\mathbf{R}_\delta^{-1}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{R}_\delta^{-1}$. Putting these into the expression of the log-likelihood function, we obtain the profile log-likelihood function of $\boldsymbol{\delta}$ as

$$\ell_P(\boldsymbol{\delta}) = -\frac{n}{2}\left[1 + \log\left(\frac{2\pi}{n}\right)\right] - \frac{1}{2}\log|\det(\mathbf{M}_\delta)| - \frac{n}{2}\log(\mathbf{u}^\top\mathbf{M}_\delta\mathbf{u}).$$

For any $k \in \{1, \dots, r\}$, we have

$$\frac{\partial\ell_P(\boldsymbol{\delta})}{\partial\delta_k} = -12\mathrm{tr}\left(\mathbf{R}_\delta^{-1}\partial\mathbf{R}_\delta\partial\delta_k\right) + n2\boldsymbol{u}^\top\mathbf{M}_\delta\partial\mathbf{R}_\delta\partial\delta_k\mathbf{M}_\delta\boldsymbol{u}\boldsymbol{u}^\top\mathbf{M}_\delta\boldsymbol{u}.$$

For any $k_1, k_2 \in \{1, \dots, r\}$, we have

$$\begin{aligned}\frac{\partial^2\ell_P(\boldsymbol{\delta})}{\partial\delta_{k_1}\partial\delta_{k_2}} &= -\frac{1}{2}\mathrm{tr}\left(\mathbf{R}_\delta^{-1}\frac{\partial^2\mathbf{R}_\delta}{\partial\delta_{k_1}\partial\delta_{k_2}}\right) + \frac{1}{2}\mathrm{tr}\left(\mathbf{R}_\delta^{-1}\frac{\partial\mathbf{R}_\delta}{\partial\delta_{k_1}}\mathbf{R}_\delta^{-1}\frac{\partial\mathbf{R}_\delta}{\partial\delta_{k_2}}\right)\\ &\quad + \frac{n}{2}\frac{\boldsymbol{u}^\top\mathbf{M}_\delta\frac{\partial^2\mathbf{R}_\delta}{\partial\delta_{k_1}\partial\delta_{k_2}}\mathbf{M}_\delta\boldsymbol{u}}{\boldsymbol{u}^\top\mathbf{M}_\delta\boldsymbol{u}} + \frac{n}{2}\frac{\left(\boldsymbol{u}^\top\mathbf{M}_\delta\frac{\partial\mathbf{R}_\delta}{\partial\delta_{k_1}}\mathbf{M}_\delta\boldsymbol{u}\right)\left(\boldsymbol{u}^\top\mathbf{M}_\delta\frac{\partial\mathbf{R}_\delta}{\partial\delta_{k_2}}\mathbf{M}_\delta\boldsymbol{u}\right)}{(\boldsymbol{u}^\top\mathbf{M}_\delta\boldsymbol{u})^2}\\ &\quad - \frac{n}{2}\frac{\boldsymbol{u}^\top\mathbf{M}_\delta\left(\frac{\partial\mathbf{R}_\delta}{\partial\delta_{k_1}}\mathbf{M}_\delta\frac{\partial\mathbf{R}_\delta}{\partial\delta_{k_2}} + \frac{\partial\mathbf{R}_\delta}{\partial\delta_{k_1}}\mathbf{M}_\delta\frac{\partial\mathbf{R}_\delta}{\partial\delta_{k_2}}\right)\mathbf{M}_\delta\boldsymbol{u}}{\boldsymbol{u}^\top\mathbf{M}_\delta\boldsymbol{u}}.\end{aligned}$$

Therefore, we can implement the Newton–Raphson algorithm to compute the maximizer of $\ell_P(\boldsymbol{\delta})$, which provides the MLE of $\boldsymbol{\delta}$, denoted by $\hat{\boldsymbol{\delta}}$. The MLEs of $\boldsymbol{\beta}$ and σ^2 are derived by $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{\hat{\delta}}$ and $\hat{\sigma}^2 = \hat{\sigma}_{\hat{\delta}}^2$, respectively.

After $\hat{\boldsymbol{\theta}}$ is derived, we predict $\boldsymbol{\gamma}$ by its conditional mean given $\boldsymbol{u}$. By $\mathrm{V}(\mathbf{u}|\boldsymbol{\gamma}) = \sigma^2\mathbf{W}^{-1}$ and $\mathrm{V}(\boldsymbol{\gamma}) = \sigma^2\mathbf{B}_\delta$, we have

$$\begin{pmatrix}\boldsymbol{\gamma}\\ \boldsymbol{u}\end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix}\mathbf{0}\\ \mathbf{X}\boldsymbol{\beta}\end{pmatrix},\quad \sigma^2\begin{pmatrix}\mathbf{B}_\delta & \mathbf{B}_\delta\mathbf{Z}^\top\\ \mathbf{Z}\mathbf{B}_\delta & \mathbf{R}_\delta\end{pmatrix}\right].$$

Because $\mathrm{Cov}[\boldsymbol{\gamma} - \mathbf{B}_\delta\mathbf{Z}^\top\mathbf{R}_\delta^{-1}(\boldsymbol{u} - \mathbf{X}\boldsymbol{\beta}), \boldsymbol{u}] = 0$, we conclude that $\boldsymbol{\gamma} - \mathbf{B}_\delta\mathbf{Z}^\top\mathbf{R}_\delta^{-1}(\boldsymbol{u} - \mathbf{X}\boldsymbol{\beta})$ and $\boldsymbol{u}$ are independent. By $\mathrm{E}[\boldsymbol{\gamma} - \mathbf{B}_\delta\mathbf{Z}^\top\mathbf{R}_\delta^{-1}(\boldsymbol{u} - \mathbf{X}\boldsymbol{\beta})] = 0$ and $\mathrm{Cov}[\boldsymbol{\gamma} - \mathbf{B}_\delta\mathbf{Z}^\top\mathbf{R}_\delta^{-1}(\boldsymbol{u} - \mathbf{X}\boldsymbol{\beta})|\boldsymbol{u}] = \sigma^2(\mathbf{B}_\delta - \mathbf{B}_\delta\mathbf{Z}^\top\mathbf{R}_\delta^{-1}\mathbf{Z}\mathbf{B}_\delta)$, we obtain $\boldsymbol{\gamma}|\boldsymbol{u} \sim \mathcal{N}[\mathbf{B}_\delta\mathbf{Z}^\top\mathbf{R}_\delta^{-1}(\boldsymbol{u} - \mathbf{X}\boldsymbol{\beta}), \sigma^2(\mathbf{B}_\delta - \mathbf{B}_\delta\mathbf{Z}^\top\mathbf{R}_\delta^{-1}\mathbf{Z}\mathbf{B}_\delta)]$. Thus, we predict $\boldsymbol{\gamma}$ by

$$\hat{\boldsymbol{\gamma}} = \mathbf{B}_\delta\mathbf{Z}^\top\mathbf{M}_\delta\mathbf{u}.$$