Journal of Statistical Computation and Simulation
Vol. 82, No. 8, August 2012, 1085–1103

A consistent simulation-based estimator in generalized linear mixed models

H. Li and L. Wang*

Department of Statistics, University of Manitoba, Winnipeg, Manitoba, Canada R3T 2N2

(*Received 3 September 2010; final version received 4 March 2011*)

We propose a strongly root-*n* consistent simulation-based estimator for the generalized linear mixed models. This estimator is constructed based on the first two marginal moments of the response variables, and it allows the random effects to have any parametric distribution (not necessarily normal). Consistency and asymptotic normality for the proposed estimator are derived under fairly general regularity conditions. We also demonstrate that this estimator has a bounded influence function and that it is robust against data outliers. A bias correction technique is proposed to reduce the finite sample bias in the estimation of variance components. The methodology is illustrated through an application to the famed seizure count data and some simulation studies.

Keywords: bias reduction; influence function; M-estimator; mixed models; robustness; simulation-based estimator

1. Introduction

Generalized linear mixed models (GLMMs) have been widely used in the modelling of longitudinal data where the response is discrete. They can be viewed as a natural combination of linear mixed models [1] and generalized linear models. In contrast to marginal or generalized estimating equation (GEE) models [2], GLMMs emphasize on the regression coefficients as well as the variance components of random effects.

For estimation and inference in GLMMs, the most frequently employed approach is likelihood based. However, the likelihood function of a GLMM involves integrals with respect to the distribution of the random effects and is generally intractable analytically. The analysis is even more difficult when the dimension of random effects is high or there are crossed random effects. To overcome this numerical difficulty, several methods have been proposed to approximate the integrals in the likelihood function, for example, marginal quasi-likelihood and penalized quasi-likelihood (PQL) estimation [3], adaptive quadrature [4] and maximum simulated likelihood [5]. A comprehensive evaluation and comparison of these approximate methods are unavailable in the statistical literature. However, some limited studies have shown that the analytical simplification

*Corresponding author. Email: liqun_wang@umanitoba.ca

ISSN 0094-9655 print/ISSN 1563-5163 online

http://dx.doi.org/10.1080/00949655.2011.569721
http://www.tandfonline.com

may not be always satisfactory and may produce biased and highly inefficient estimates [6,7]. Furthermore, the likelihood methods rely on normal assumption for random effects. Since the random effects are unobservable, it is not feasible to verify their distributional assumptions. It is thus natural to be concerned whether these methods yield reliable results when the normality assumption is violated. In addition, it is also known that likelihood-based methods are sensitive to data outliers. On the other hand, there are many works extending the GEE type or quasi-likelihood to the estimation of GLMMs [2,8,9]. However, these methods are usually inefficient and require the simulation size S to go to infinity to obtain consistent estimators. In practice, since S has to be fixed, these methods only produce approximate consistent estimates.

In this paper, we propose an exact (not approximate) consistent simulation-based estimator (SBE) using fixed S in the framework of GLMMs. This estimator is constructed based on the first two marginal moments of the response variables, and it allows random effects follow a flexible distribution. This approach was originally studied by Wang [10] for nonlinear mixed effects models with homoscedastic errors. This paper extends this methodology to a GLMM which allows very general heteroscedastic errors, and we further investigate its robustness against data outliers using its influence function (IF). In addition, this paper proposes a bias reduction technique to reduce the finite sample bias for the estimation of variance components.

The structure of the paper is as follows. In Section 2, we introduce the model and give some examples to illustrate model identifiability. In Section 3, we introduce the SBE and its properties. In Section 4, we present simulation studies to examine the finite sample performances of the proposed estimators. In Section 5, a real data application is given, and in Section 6, a discussion is given. Proofs of the theorems are provided in the appendix.

2. The model

Suppose a subject i is measured repeatedly on n_i occasions and it is assumed as the conditional distribution of the response variable $y_{ij} \in \mathbb{R}$, given that the random effects $b_i \in \mathbb{R}^q$ are independent and belong to an exponential family. The random effects are assumed to have mean zero and distribution $f_b(u;\theta)$ with unknown parameters $\theta \in \mathbb{R}^r$. The conditional mean of y_{ij} is assumed to depend upon fixed and random effects via a linear predictor and can be written as

$$g^{-1}\{E(y_{ij}|b_i, x_{ij}, z_{ij})\} = x'_{ij}\beta + z'_{ij}b_i, \quad i = 1,\dots,m, \quad j = 1,\dots,n_i, \tag{1}$$

where $x_{ij} \in \mathbb{R}^p$ and $z_{ij} \in \mathbb{R}^q$ are the predictors, $\beta \in \mathbb{R}^p$ is a vector of the fixed effects and $g^{-1}(\cdot)$ is a link function. The conditional variance is given by

$$V(y_{ij}|b_i, x_{ij}, z_{ij}) = \phi\nu(g(x'_{ij}\beta + z'_{ij}b_i)), \tag{2}$$

where $\nu(\cdot)$ is a known variance function and ϕ is a scale parameter that may be known or unknown. In this model, the parameter of interest is $\psi = (\beta', \theta', \phi)'$. Based on the model assumptions, the first and second marginal moments can be expressed as

$$\mu_{ij}(\psi) = E(y_{ij}|X_i, Z_i) = \int g(x'_{ij}\beta + z'_{ij}u) f_b(u;\theta)\,\mathrm{d}u \tag{3}$$

and

$$\eta_{ijk}(\psi) = E(y_{ij}y_{ik}|X_i, Z_i) = \int g(x'_{ij}\beta + z'_{ij}u) g(x'_{ik}\beta + z'_{ik}u) f_b(u;\theta)\,\mathrm{d}u + \delta_{jk}\phi \int \nu(g(x'_{ij}\beta + z'_{ij}u)) f_b(u;\theta)\,\mathrm{d}u, \tag{4}$$

where $X_i = (x'_{i1}, x'_{i2}, \dots, x'_{in_i})'$, $Z_i = (z'_{i1}, z'_{i2}, \dots, z'_{in_i})'$, $\delta_{jk} = 1$ if $j = k$ and 0 otherwise.

In the following, we motivate our approach using two most popular GLMMs as examples to demonstrate that ψ can indeed be identified and consistently estimated using the first two marginal moments (3) and (4).

Example 2.1 Consider a mixed Poisson model for counts, where $V(y_{ij}|b_i) = E(y_{ij}|b_i)$ and $\log E(y_{ij}|b_i) = x'_{ij}\beta + z'_{ij}b_i$. Assuming $b_i \sim N(0, D(\theta))$, we have

$$\mu_{ij}(\psi) = \exp\left(x'_{ij}\beta + \frac{z'_{ij}D(\theta)z_{ij}}{2}\right), \tag{5}$$

and

$$\eta_{ijk}(\psi) = \mu_{ij}(\psi)\mu_{ik}(\psi)\exp[z'_{ij}D(\theta)z_{ik})] + \delta_{jk}\phi\mu_{ij}(\psi). \tag{6}$$

All unknown parameters in this model can be consistently estimated by Equations (5) and (6).

Example 2.2 Consider a mixed logistic model for a binary response y_{ij}, where $\phi = 1$ and $\text{logit}\{Pr(y_{ij} = 1|b_i)\} = x'_{ij}\beta + z'_{ij}b_i$. For this model, we find

$$\mu_{ij}(\psi) = E(y_{ij}^2|X_i, Z_i) = \int \left(\frac{e^{x'_{ij}\beta + z'_{ij}u}}{1 + e^{x'_{ij}\beta + z'_{ij}u}}\right) f_b(u;\theta)\,\mathrm{d}u, \tag{7}$$

and

$$\eta_{ijk}(\psi) = \int \left(\frac{e^{x'_{ij}\beta + z'_{ij}u}}{1 + e^{x'_{ij}\beta + z'_{ij}u}}\right)\left(\frac{e^{x'_{ik}\beta + z'_{ik}u}}{1 + e^{x'_{ik}\beta + z'_{ik}u}}\right) f_b(u;\theta)\,\mathrm{d}u, \quad \text{for } j < k. \tag{8}$$

The integrals in Equations (7) and (8) are intractable but can be approximated using Monte Carlo simulation techniques.

3. Simulation-based estimator

3.1. *The estimator and its asymptotic properties*

The first two marginal moments usually do not have closed forms in GLMMs, and the density $f_b(u;\theta)$ is typically unknown. Here, we propose a simulation-based approach to overcome these two difficulties simultaneously. As it is well known, SBE is computationally convenient when moment functions cannot be evaluated directly [11–13]. The basic idea is to form unbiased estimators of integrals in moment equations with their Monte Carlo simulators. In particular, we propose a simulation-by-parts [14] technique to construct two sets of moments. First, generate random points u_{is}, $s = 1, 2, \ldots, 2S$, from a known density $h(u)$, and construct

$$\mu_{ij,1}(\psi) = \frac{1}{S}\sum_{s=1}^{S} \frac{g(x'_{ij}\beta + z'_{ij}u_{is})f_b(u_{is};\theta)}{h(u_{is})}, \tag{9}$$

$$\begin{aligned}\eta_{ijk,1}(\psi) = {} & \frac{1}{S}\sum_{s=1}^{S} \frac{g(x'_{ij}\beta + z'_{ij}u_{is})g(x'_{ik}\beta + z'_{ik}u_{is})f_b(u_{is};\theta)}{h(u_{is})} \\ & + \frac{\delta_{jk}\phi}{S}\sum_{s=1}^{S} \frac{\nu(g(x'_{ij}\beta + z'_{ij}u_{is}))f_b(u_{is};\theta)}{h(u_{is})}\end{aligned} \tag{10}$$

using the first half of the points u_{is}, $s = 1, 2, \ldots, S$. Then, construct $\mu_{ij,2}(\psi)$ and $\eta_{ijk,2}(\psi)$ similarly using the second half of the points u_{is}, $s = S+1, S+2, \ldots, 2S$. It is obvious that the simulated moments are unbiased estimates of the true moments, since $E(\mu_{ij,t}(\psi)|X_i, Z_i) = \mu_{ij}(\psi)$ and $E(\eta_{ijk,t}(\psi)|X_i, Z_i) = \eta_{ijk}(\psi)$, $t = 1, 2$. We denote the parameter space by $\Gamma = \Omega \times \Theta \times \Sigma \in \mathbb{R}^{p+r+1}$, and the true parameter value by $\psi_0 = (\beta_0', \theta_0', \phi_0)' \in \Gamma$. Finally, the SBE $\hat{\psi}_{m,S}$ for ψ is defined as

$$\hat{\psi}_{m,S} = \operatorname*{argmin}_{\psi\in\Gamma} Q_{m,S}(\psi) = \operatorname*{argmin}_{\psi\in\Gamma} \sum_{i=1}^{m} \rho_{i,1}'(\psi) W_i \rho_{i,2}(\psi),$$

where $\rho_{i,t}(\psi) = (y_{ij} - \mu_{ij,t}(\psi), 1 \le j \le n_i, y_{ij}y_{ik} - \eta_{ijk,t}(\psi), 1 \le j \le k \le n_i)'$ and $W_i = W(X_i, Z_i)$ is a nonnegative definite matrix which may depend on X_i and Z_i. By using two different sets of independent simulated points, $Q_{m,S}(\psi)$ is an unbiased estimator of $Q_m(\psi) = \sum_{i=1}^{m} \rho_i'(\psi) W_i \rho_i(\psi)$ because $\rho_{i,1}(\psi)$ and $\rho_{i,2}(\psi)$ are conditionally independent given (Y_i, X_i, Z_i), and hence,

$$\begin{aligned} E[\rho_{i,1}(\psi) W_i \rho_{i,2}(\psi)] &= E[E(\rho_{i,1}(\psi)|Y_i, X_i, Z_i) W_i E(\rho_{i,2}(\psi)|Y_i, X_i, Z_i)] \\ &= E(\rho_i(\psi) W_i \rho_i(\psi)), \end{aligned} \tag{11}$$

where $\rho_i(\psi) = (y_{ij} - \mu_{ij}(\psi), 1 \le j \le n_i, y_{ij}y_{ik} - \eta_{ijk}(\psi), 1 \le j \le k \le n_i)'$.

To construct simulated moments in Equations (9) and (10), the random effect distribution is only required to have a known parametric form. Hence, instead of relying on normality assumption on b_i, we can use more flexible distributions. For example, one can follow Davidian and Gallant [15] and Zhang and Davidian [16] to represent the density of b_i by the standard semi-nonparametric densities, which include normal, skewed, multi-modal and fat- or thin-tailed densities. One can use the Tukey(g, h) family distribution [17] for b_i as well, which is generated by a single transformation of the standard normal and covers a variety of distributions.

To establish the consistency and asymptotic normality of $\hat{\psi}_{m,S}$, we make the following assumptions:

A1. $g(\cdot)$ and $\nu(\cdot)$ are continuous functions; $f_b(u;\theta)$ is continuous in $\theta \in \Theta$ for all u.

A2. $E[\|W_i\|(y_{ij}^4 + 1)] < \infty$; $g^2(x'\beta + z'u) f_b(u;\theta)$ and $|\nu(g(x'\beta + z'u))| f_b(u;\theta)$ are bounded by a positive function $G(x, z, u)$ satisfying $E[\|W_i\|(\int G(X_i, Z_i, u)\,\mathrm{d}u)^2] < \infty$.

A3. The parameter space $\Gamma \subset \mathbb{R}^{p+r+1}$ is compact.

A4. $E[\rho_i(\psi) - \rho_i(\psi_0)]' W_i [\rho_i(\psi) - \rho_i(\psi_0)] = 0$ if and only if $\psi = \psi_0$.

A5. $g(\cdot)$ and $\nu(\cdot)$ are twice continuously differentiable and $f_b(u;\theta)$ is twice continuously differentiable w.r.t. to θ in an open subset $\theta_0 \in \Theta_0 \subset \Theta$. Furthermore, all first- and second-order partial derivatives of $g(x'\beta + z'u) f_b(u;\theta)$ and $\nu(g(x'\beta + z'u)) f_b(u;\theta)$ w.r.t. $(\beta', \theta')'$ are bounded absolutely by the positive function $G(x, z, u)$ given in A2.

A6. The matrix

$$B = E\left[\frac{\partial \rho_i'(\psi_0)}{\partial \psi} W_i \frac{\partial \rho_i(\psi_0)}{\partial \psi'}\right] \tag{12}$$

is nonsingular.

THEOREM 3.1 *Suppose that* $Supp(h) \supseteq Supp(f_b(\cdot;\theta))$ *for all* $\theta \in \Theta_0$. *Then, for any fixed* $S > 0$, *as* $m \to \infty$,

(1) *under* A1–A4, $\hat{\psi}_{m,S} \xrightarrow{a.s.} \psi_0$;

(2) *under* A1–A6, $\sqrt{m}(\hat{\psi}_{m,S} - \psi_0) \xrightarrow{L} N(0, B^{-1}C_S B^{-1})$, *where*

$$2C_S = E\left[\frac{\partial \rho'_{i,1}(\psi_0)}{\partial \psi} W_i \rho_{i,2}(\psi_0)\rho'_{i,2}(\psi_0) W_i \frac{\partial \rho_{i,1}(\psi_0)}{\partial \psi'}\right] + E\left[\frac{\partial \rho'_{i,1}(\psi_0)}{\partial \psi} W_i \rho_{i,2}(\psi_0)\rho'_{i,1}(\psi_0) W_i \frac{\partial \rho_{i,2}(\psi_0)}{\partial \psi'}\right]. \tag{13}$$

Note that the above asymptotic results do not require that the simulation size S tends to infinity, because we use the simulation-by-parts technique to approximate moments. This is fundamentally different from other simulation-based methods, which require that S goes to infinity to obtain consistent estimators [2,8,9]. In general, the simulation approximation of the integrals will result in certain efficiency loss, but this loss decreases at the rate $O(1/S)$ [14]. Therefore, the efficiency loss due to the simulations can be made small by increasing S. In general, a simulation size of 1000–3000 is sufficient to obtain satisfactory estimates. For the choice of $h(u)$, in theory, it has no impact on the asymptotic efficiency of the estimator, as long as it has sufficiently large support. However, the choice of $h(u)$ will affect the finite sample variances of the simulated moments. It is well known that the finite sample variances will be minimized when $h(u) \propto |g(x'_{ij}\beta + z'_{ij}u) f_b(u;\theta)|$ and $h(u) \propto |g(x'_{ij}\beta + z'_{ij}u) g(x'_{ik}\beta + z'_{ik}u) f_b(u;\theta)|$.

When closed forms of moments exist such as in Example 2.1, the SBE becomes M-estimator [18] $\hat{\psi}_m$ or the second-order least squares estimator (SLSE) of Wang [10]. We can show that $\hat{\psi}_m$ is consistent and asymptotically normally distributed. In particular, we have the following corollary.

Corollary 3.2 *As* $m \to \infty$, $\hat{\psi}_m = \arg\min Q_m(\psi)$ *has properties*

(1) *under* A1–A4, $\hat{\psi}_m \xrightarrow{a.s.} \psi_0$;
(2) *under* A1–A6, $\sqrt{m}(\hat{\psi}_m - \psi_0) \xrightarrow{L} N(0, B^{-1}CB^{-1})$, *where B is given in Equation* (12) *and*

$$C = E\left[\frac{\partial \rho'_i(\psi_0)}{\partial \psi} W_i \rho_i(\psi_0)\rho'_i(\psi_0) W_i \frac{\partial \rho_i(\psi_0)}{\partial \psi'}\right].$$

Remark 3.3 Since random effects are usually assumed to have zero mean, it is more convenient to define $b_i = D(\theta)^{1/2}\xi_i$, where the random variable ξ has mean zero and covariance matrix I_q. Hence alternatively, we can rewrite Equations (9) and (10) as

$$\mu_{ij,1}(\psi) = \frac{1}{S}\sum_{s=1}^{S} \frac{g(x'_{ij}\beta + z'_{ij}D(\theta)^{1/2}u_{is}) f_\xi(u_{is})}{h(u_{is})},$$

$$\eta_{ijk,1}(\psi) = \frac{1}{S}\sum_{s=1}^{S} \frac{g(x'_{ij}\beta + z'_{ij}D(\theta)^{1/2}u_{is}) g(x'_{ik}\beta + z'_{ik}D(\theta)^{1/2}u_{is}) f_\xi(u_{is})}{h(u_{is})} + \frac{\delta_{jk}\phi}{S}\sum_{s=1}^{S} \frac{\nu(g(x'_{ij}\beta + z'_{ij}D(\theta)^{1/2}u_{is})) f_\xi(u_{is})}{h(u_{is})}.$$

In this case, there is no parameter of interest in $f_\xi(u_{is})$.

Remark 3.4 For binary responses y_{ij}, $E(y_{ij}|X_i, Z_i) = E(y_{ij}^2|X_i, Z_i)$ with probability one. Therefore, the terms $y_{ij}^2 - E(y_{ij}^2|X_i, Z_i)$ in $\rho_{i,1}(\psi)$ and $\rho_{i,2}(\psi)$ are redundant and do not need to be included.

Remark 3.5 For certain GLMMs such as a probit model with normal distributed random effects, the first marginal moment admits an analytical form but not the second marginal moments. In this case, only the second moments need to be simulated.

3.2. *Computation of the SBE*

In general, the SBE does not admit an explicit solution and can be computed using the Newton–Raphson algorithm as

$$\hat{\psi}^{(\tau+1)} = \hat{\psi}^{(\tau)} - \left(\frac{\partial^2 Q_{m,S}(\hat{\psi}^{(\tau)})}{\partial\psi\partial\psi'}\right)^{-1} \frac{\partial Q_{m,S}(\hat{\psi}^{(\tau)})}{\partial\psi},$$

where $\hat{\psi}^{(\tau)}$ denotes the estimate of ψ at the τth iteration, and

$$\frac{\partial Q_{m,S}(\hat{\psi}^{(\tau)})}{\partial\psi} = \sum_{i=1}^{m}\left[\frac{\partial\rho'_{i,1}(\hat{\psi}^{(\tau)})}{\partial\psi}W_i\rho_{i,2}(\hat{\psi}^{(\tau)}) + \frac{\partial\rho'_{i,2}(\hat{\psi}^{(\tau)})}{\partial\psi}W_i\rho_{i,1}(\hat{\psi}^{(\tau)})\right], \tag{14}$$

$$\begin{aligned}\frac{\partial^2 Q_{m,S}(\hat{\psi}^{(\tau)})}{\partial\psi\partial\psi'} &= \sum_{i=1}^{m}\left[\frac{\partial\rho'_{i,1}(\hat{\psi}^{(\tau)})}{\partial\psi}W_i\frac{\partial\rho_{i,2}(\hat{\psi}^{(\tau)})}{\partial\psi'} + (\rho'_{i,2}(\hat{\psi}^{(\tau)})W_i\otimes I)\frac{\partial\mathrm{vec}(\partial\rho'_{i,1}(\hat{\psi}^{(\tau)})/\partial\psi)}{\partial\psi'}\right]\\ &\quad+\sum_{i=1}^{m}\left[\frac{\partial\rho'_{i,2}(\hat{\psi}^{(\tau)})}{\partial\psi}W_i\frac{\partial\rho_{i,1}(\hat{\psi}^{(\tau)})}{\partial\psi'}\right.\\ &\quad\left.+(\rho'_{i,1}(\hat{\psi}^{(\tau)})W_i\otimes I)\frac{\partial\mathrm{vec}(\partial\rho'_{i,2}(\hat{\psi}^{(\tau)})/\partial\psi)}{\partial\psi'}\right].\end{aligned} \tag{15}$$

The terms $(\rho'_{i,1}W_i\otimes I)(\partial\mathrm{vec}(\partial\rho'_{i,2}/\partial\psi))/\partial\psi'$ and $(\rho'_{i,2}W_i\otimes I)(\partial\mathrm{vec}(\partial\rho'_{i,1}/\partial\psi))/\partial\psi'$ are $o_p(1)$, so they can be omitted from the second derivative for computational convenience.

Another important question is how to specify the form of weight W_i to compute $\hat{\psi}_{m,S}$ in an optimal way, such that $AV(\hat{\psi}_m(W_i)) - AV(\hat{\psi}_m(W_i^{\mathrm{opt}}))$ is nonnegative definite for all possible W_i. It can be shown that W_i^{opt} is equal to

$$A_i^{-1} = E[\rho_{i,1}(\psi_0)\rho'_{i,2}(\psi_0)|X_i, Z_i]^{-1}. \tag{16}$$

The proof is analogous to that reported by Abarin and Wang [19] and is, therefore, omitted. In practice, A_i is not feasible, since it involves unknown parameters to be estimated. One possible solution is using a two-stage procedure. First, minimize $Q_{m,S}(\psi)$ using a sub-optimal choice of W_i, such as an identity weight matrix, to obtain the first-stage estimator $\hat{\psi}_{m1,S}$. Second, estimate $W_i = \hat{A}_i^{-1}$ using $\hat{\psi}_{m1,S}$ and then minimize $Q_{m,S}(\psi)$ again with $\hat{A}_i^{-1}$ to obtain the second-stage estimator $\hat{\psi}_{m2,S}$. In general, the computation of A_i in Equation (16) is difficult, since it requires the specification of the third- and fourth-order moments of y_{ij}. However, these high-order moments can be easily approximated using the Monte Carlo simulation method introduced in this section. Alternatively, A_i can be estimated using any nonparametric method such as kernel or spline estimation. A simple estimator of A_i would be

$$A(\hat{\psi}) = \frac{1}{m}\sum_{i=1}^{m}\rho_{i,1}(\hat{\psi}_{m1})\rho'_{i,2}(\hat{\psi}_{m1}). \tag{17}$$

In many real data applications, the subjects are clustered so that the values of X_i, Z_i are equal or close for all subjects within one cluster. In such cases, each A_i can be estimated similar to Equation (17) using all the subjects within the same cluster.

3.3. *Robustness*

Many simulation studies that we have done show that the estimated optimal weight (17) provides not only efficient estimates but also protection against influential measurements. This motivated us to investigate the robustness property of the proposed estimator theoretically. In particular, we study the robustness property of the SBE by means of the IF, which was introduced by Hampel *et al.* [20]. Let v be the subset of observations (X_l, Y_l) under investigation, and the IF of SBE at point v takes the form [20]

$$\mathrm{IF}(v; \hat{\psi}_{m,S}, F) = -B(\hat{\psi}_m(F))^{-1} \frac{\partial \rho'_{l,1}(v; \hat{\psi}_{m,S}(F))}{\partial \psi} \hat{A}^{-1} \rho_{l,2}(v; \hat{\psi}_{m,S}(F)), \tag{18}$$

where F is the underlying distribution and B is given in Equation (12).

COROLLARY 3.6 *If the SBE $\hat{\psi}_{m,S}$ is computed using the estimated optimal weight* (17), *then* $\|IF(v; \hat{\psi}_{m,S}, F)\| \to 0$ *as* $\|v\| \to \infty$.

The implication of the above corollary is that the IF of $\hat{\psi}_m$ is bounded and $\hat{\psi}_m$ has a redescending property [18]. It is expected that data outliers in either x or y direction will be automatically downweighted by the inverse of the estimated optimal weight matrix. It does not require detection for outliers beforehand to implement downweighting strategy.

3.4. *Bias reduction*

It is noticed in the simulation studies done by Wang [10] and our preliminary simulation studies that there are some finite sample biases for the estimation of variance components by the SBE. These biases are downward oriented and diminish with the increase in sample sizes. The source of this bias lies in the fact that the optimal weight in Equation (16) is replaced by a root-m estimate given in Equation (17) for the second-stage minimization. Asymptotically, this replacement has no impact on the properties of SBE. However, it does make a difference in finite samples because $A_i(\hat{\psi})$ depends on y_i and causes the correlation with $\rho_{i,1}(\psi)$ and $\rho_{i,2}(\psi)$. Note in the setup of the SBE, we require W_i that may only depend on X_i and Z_i. Evaluating this bias analytically is not easy. Instead, we extend the independently weighted method proposed by Altonji and Segal [21] for the bias reduction. The basic idea is to break the correlation between $A_i(\hat{\psi})$ and $\rho_{i,t}(\psi)$ by designing the weighting matrix using observations other than those used to construct the sample moments. We randomly split the sample into K groups with m_k subjects in each group, and the independently weighted SBE (SBEIW) $\hat{\psi}^{\mathrm{IW}}_{m,S}$ for ψ is defined as the measurable function that minimizes

$$Q_{m,S}(\psi) = \frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{m_k} (\rho^k_{i,1}(\psi))' A^{-1}_{i,k}(\hat{\psi}) \rho^k_{i,2}(\psi), \tag{19}$$

where $\rho^k_{i,t}(\psi)$ is constructed for the kth group and $A^{-1}_{i,k}(\hat{\psi})$ is constructed using all but the kth group. Intuitively, this estimator is less biased because the statistical dependence between the weight matrix and sample moments is broken. However, splitting the sample causes efficiency loss due to the loss in degrees of freedom. Since $\mathrm{cov}(\hat{\psi}^k_{m,S}, \hat{\psi}^{k+l}_{m,S}) = 0$ for $l \neq 0$ by design, it can be easily shown that

$$\mathrm{cov}(\hat{\psi}^{\mathrm{IW}}_{m,S}) = \frac{1}{K^2}\sum_{k=1}^{K}\mathrm{cov}(\hat{\psi}^{k}_{m,S}),$$

where $\hat{\psi}^k_{m,S}$ is obtained by minimizing $\sum_{i=1}^{m_k}(\rho^k_{i,1}(\psi))' A^{-1}_{i,k}(\hat{\psi})\rho^k_{i,2}(\psi)$. In the simulation studies presented in Section 4, we select $K = 2$ and observe significant improvement in estimation bias over SBE with negligible efficiency loss.

4. Monte Carlo simulation studies

In this section, we evaluate the finite sample behaviours of the proposed estimator and compare them with that of the penalized quasi-likelihood estimator (PQLE) reported by Breslow and Clayton [3]. We conducted substantial numerical studies by using different GLMMs and parameter configurations. We carried out 500 Monte Carlo replications in each simulation study and reported the biases and the root mean square errors (RMSEs). All computations were done in R and PQL estimates were obtained from the `glmmPQL` package.

The first simulation study was designed based on Example 2.1. In particular, we simulated the model $\log E(y_{ij}|b_i) = \beta_0 + \beta_1 x_{ij} + b_i$, $j = 1, \ldots, 4$, where $x_{ij} = 0.1j$, $\beta = (3, -1)'$ and $b_i \sim N(0, 0.25)$. In the present simulation, we set $m = 50, 100, 200, 300, 400$ and chose the density $N(0, 1)$ to be $h(u)$ and generated $S = 1000$ independent u_{is} for the SBE. For comparison purpose, we also computed the ψ_m by using the two marginal moments from Equations (5) and (6).

Table 1 reports the biases and the RMSEs. Figure 1 visually summarizes the performance of all estimators at various sample sizes in terms of RMSEs and percentage of bias. From Table 1 and Figure 1, we can see that all estimators perform satisfactorily and show clearly their asymptotic proprieties, that is, the estimated RMSEs decrease with the increase in sample size. For fixed effects, both estimated RMSEs and biases from the proposed estimators are very close to each other and are comparable to the PQLE, although ψ^{IW}_m and $\psi^{\mathrm{IW}}_{m,S}$ have slightly higher RMSEs

Table 1. Biases (RMSE) of the parameter estimates.

m	PQLE	SLSE	SLSIW	SBE	SBEIW
$\beta_0 = 3$					
50	0.006 (0.082)	−0.086 (0.115)	0.001 (0.162)	−0.069 (0.109)	0.012 (0.168)
100	0.012 (0.060)	−0.053 (0.077)	−0.009 (0.090)	−0.039 (0.075)	0.007 (0.103)
200	0.010 (0.040)	−0.029 (0.052)	−0.009 (0.058)	−0.022 (0.055)	0.005 (0.061)
300	0.006 (0.033)	−0.021 (0.040)	−0.005 (0.040)	−0.016 (0.047)	−0.003 (0.052)
400	0.009 (0.031)	−0.017 (0.035)	−0.005 (0.034)	−0.010 (0.044)	−0.003 (0.043)
$\beta_1 = -1$					
50	−0.007 (0.152)	0.007 (0.143)	0.020 (0.341)	0.009 (0.130)	−0.005 (0.329)
100	−0.004 (0.109)	0.006 (0.106)	0.013 (0.180)	0.008 (0.107)	0.007 (0.195)
200	0.000 (0.073)	0.002 (0.077)	0.015 (0.109)	0.004 (0.074)	0.013 (0.115)
300	−0.001 (0.064)	0.003 (0.061)	0.007 (0.081)	0.000 (0.058)	0.003 (0.081)
400	−0.001 (0.056)	0.001 (0.054)	0.003 (0.067)	0.003 (0.054)	0.006 (0.065)
$\theta = 0.25$					
50	−0.010 (0.053)	−0.043 (0.060)	0.011 (0.105)	−0.054 (0.076)	0.012 (0.122)
100	−0.007 (0.040)	−0.043 (0.056)	0.004 (0.066)	−0.045 (0.069)	0.001 (0.081)
200	−0.004 (0.026)	−0.030 (0.042)	0.012 (0.059)	−0.036 (0.059)	0.000 (0.060)
300	−0.003 (0.023)	−0.024 (0.035)	0.006 (0.048)	−0.027 (0.051)	0.002 (0.055)
400	−0.004 (0.019)	−0.022 (0.032)	0.002 (0.033)	−0.025 (0.048)	0.005 (0.048)

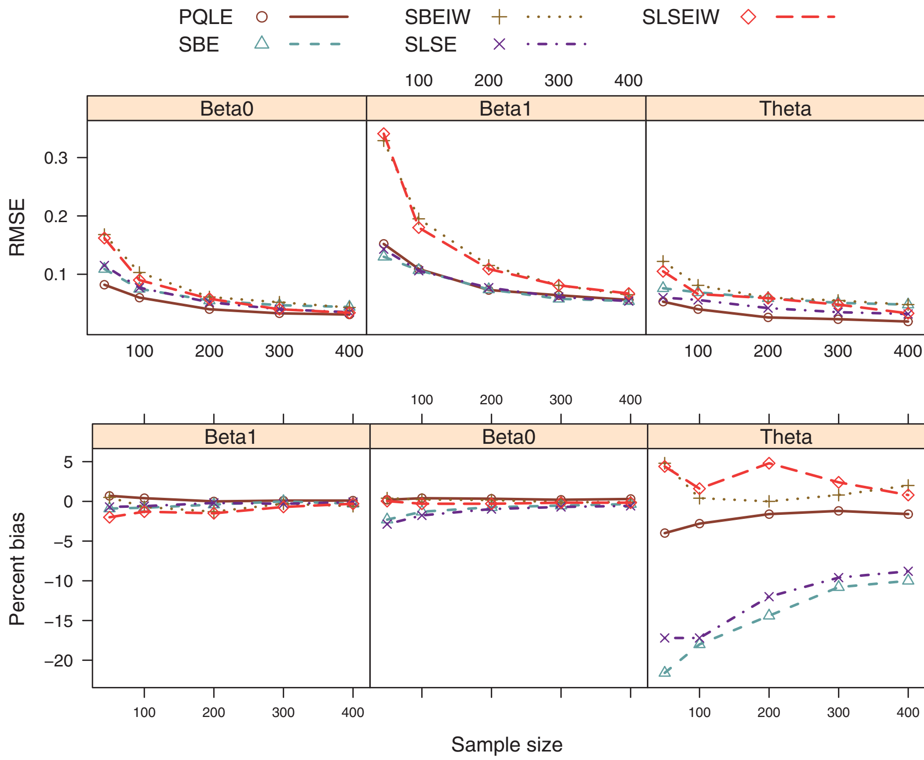

Figure 1. RMSE and percentage of bias of parameter estimates for a model at various sample sizes.

for β_1. For the random effect parameter θ, all estimators present similar estimated RMSEs and PQLE; ψ_m and $\psi_{m,S}$ show some downward bias, while ψ_m^{IW} and $\psi_{m,S}^{\mathrm{IW}}$ show some upward bias. In Figure 1, a significant higher percent (10–20%) bias is observed in ψ_m as well as in $\psi_{m,S}$; however, it is worth noting that this bias gradually reduces with the increase in sample size. In contrast, ψ_m^{IW} and $\psi_{m,S}^{\mathrm{IW}}$ have less than 5% bias, which demonstrates bias reduction by using the proposed independent weight methodology. In addition, we use histograms to show how close the distributions of the SBE estimates are to the normal distributions and compare them with those of the PQL estimates. In Figure 2, we can find that when $m = 200$, the distribution is already fairly close to normal for all estimators; thus, the asymptotic normality properties of the proposed estimates are justified.

A second simulation study was conducted based on a model setup that was the same as the one in the previous simulation study, except the random effect was generated from either a $t(4)$ or a $\chi^2(3)$ distribution. $h(u)$ was set as the same distribution as the random effect for SBE. Table 2 summarizes the simulation results. For fixed effects, Monte Carlo mean estimates from both PQLE and SBE are close to the true parameter values and no apparent biases are observed. For the random effect, PQLE results in a larger bias and RMSEs in comparison with the SBE.

In the third simulation study, we considered a logistic model: $\mathrm{logit}(\Pr(y_{ij} = 1|b_i)) = \beta_0 + \beta_1 \times \mathrm{trt}_i + \beta_2 x_{ij} + b_{i0} + b_{i1}x_{ij}$, where $b_i \sim N[(0,0)', \mathrm{diag}(\theta_0, \theta_1)]$. In the present simulation, we selected $m = 200, 300$ and $n = 5$; covariates $\mathrm{trt}_i = 1$ for half the sample and 0 for the remainder, $x_{ij} = (j-3)/2$; $\beta = (-1.0,\ 0.5,\ 0.5)'$; $\theta_0 = 1$ and $\theta_1 = 0.5$. To compute the SBE, we chose the density of $N[(0,0)', \mathrm{diag}(2,2)]$ to be $h(u)$ and generated independent points u_{is}, $s = 1, \ldots, 2S$, using $S = 500$, 1000 and 2000, respectively. Table 3 reports the simulation results. Overall, it is clear that the SBE results in smaller bias than the PQLE for fixed effects as well as

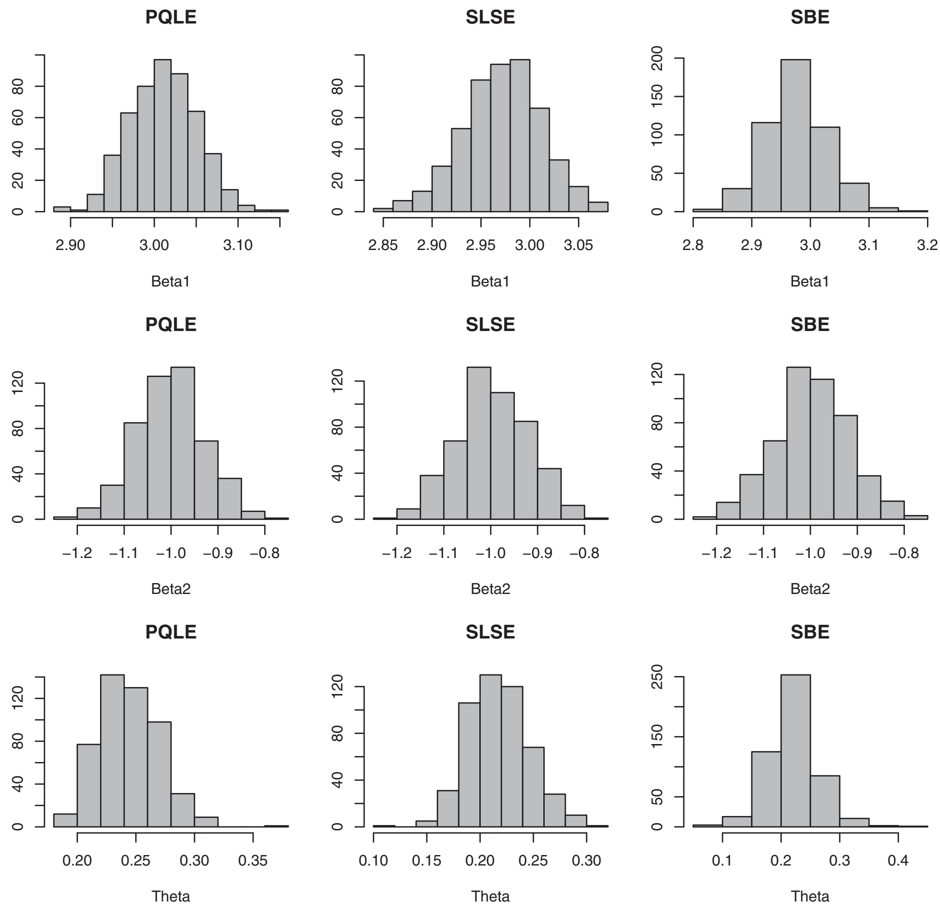

Figure 2. Histograms of PQLE, SLSE and SBE for a model with $m = 200$.

Table 2. Biases (RMSE) of the parameter estimates at $m = 200$ and non-normal random effect distribution.

	$\chi^2(3)$ distribution		$t(4)$ distribution	
	PQLE	SBE	PQLE	SBE
$\beta_0 = 3$	0.006 (0.011)	−0.031 (0.055)	0.010 (0.101)	−0.028 (0.053)
$\beta_1 = -1$	0.002 (0.073)	0.005 (0.079)	0.002 (0.056)	0.007 (0.072)
$\theta = 0.25$	0.093 (0.394)	−0.023 (0.039)	0.116 (1.106)	−0.027 (0.041)

the random effect θ_0, while the SBE has slightly bigger bias only for the random effect θ_1. The finding is not surprising, as it is known that the PQLE may have severe bias in the estimates of the fixed effects and variance components of random effects, when repeated measures data are binary. As the sample size m increases from 200 to 300, the RMSEs for all parameters from all methods decrease. For the SBE, as the number of simulated values S decreases from 2000 to 500, RMSEs become slightly bigger, but the estimates stay relatively stable. This implies that even at a relative small sample size of simulated values $S = 500$, the SBE still produces reasonable estimates. On comparing the PQLE with the SBE computed using $S = 2000$, the PQLE seems to have smaller RMSEs, especially for the random effect estimates. The SBEIW has also been computed and it

Table 3. Biases (RMSE) of the parameter estimates with a different number of the simulated points S for SBE.

		SBE		
	PQLE	$S = 2000$	$S = 1000$	$S = 500$
$m = 200$				
$\beta_0 = -1$	0.109 (0.180)	−0.071 (0.188)	−0.070 (0.200)	−0.049 (0.191)
$\beta_1 = 0.5$	−0.054 (0.189)	0.029 (0.217)	0.040 (0.218)	0.032 (0.174)
$\beta_2 = 0.5$	−0.057 (0.124)	0.030 (0.141)	0.030 (0.139)	0.024 (0.109)
$\theta_0 = 1$	−0.108 (0.258)	0.103 (0.332)	0.112 (0.375)	0.063 (0.358)
$\theta_1 = 0.5$	0.074 (0.279)	0.082 (0.402)	0.107 (0.392)	0.061 (0.366)
$m = 300$				
$\beta_0 = -1$	0.113 (0.164)	−0.030 (0.135)	−0.045 (0.154)	−0.033 (0.178)
$\beta_1 = 0.5$	−0.067 (0.176)	0.021 (0.170)	0.024 (0.169)	0.027 (0.183)
$\beta_2 = 0.5$	−0.058 (0.109)	0.022 (0.109)	0.022 (0.107)	0.013 (0.108)
$\theta_0 = 1$	−0.116 (0.210)	0.055 (0.255)	0.071 (0.298)	0.051 (0.345)
$\theta_1 = 0.5$	0.088 (0.241)	0.074 (0.319)	0.073 (0.324)	0.045 (0.334)

Table 4. Biases (RMSE) for the parameter estimates with and without outliers.

	No outliers			With outliers		
	PQLE	GEE	SLSE/SBE	PQLE	GEE	SLSE/SBE
Poisson model						
$\beta_0 = 1$	0.021 (0.060)	0.1232 (0.1369)	−0.082 (0.103)	0.162 (0.205)	0.2907 (0.3080)	−0.057 (0.082)
$\beta_1 = 1$	−0.001 (0.038)	−0.0019 (0.0373)	0.017 (0.043)	−0.004 (0.163)	0.0081 (0.1716)	0.011 (0.041)
$\theta = 0.25$	−0.013 (0.043)	–	−0.047 (0.062)	0.097 (1.029)	–	−0.040 (0.059)
Logistic model						
$\beta_0 = 1$	0.020 (0.212)	−0.0440 (0.0699)	0.066 (0.306)	−0.059 (0.412)	−0.0570 (0.0807)	0.074 (0.365)
$\beta_1 = 1$	0.051 (0.229)	−0.0435 (0.0744)	0.117 (0.317)	−0.108 (0.433)	−0.0551 (0.0842)	0.073 (0.301)
$\theta = 0.25$	0.017 (0.320)	–	−0.021 (0.571)	−0.013 (0.295)	–	0.028 (0.643)

showed smaller biases than SBE. The simulation results from SBEIW are not provided here for the sake of saving space, since SBE has already demonstrated smaller biases than PQLE.

The last simulation study here is to demonstrate the robustness of the proposed estimator in the presence of outliers; we conducted simulation studies on random intercept Poisson and logistic models with one covariate and the parameter values $\beta = (1, 1)'$ and $\theta = 0.25$. We generated $m = 100$ subjects with $n = 5$ measurements per subject. The values of the covariate $x_{ij} = (j - 3)/2$ in the Poisson mixed model and one random measurement within five different subjects were contaminated by using $100y_{ij}$ (i.e. 5% subjects with one outlier). For the logistic model, x_{ij} was generated from $N(0, 1)$. Since the response variable y_{ij} is binary in the logistic model, outliers arise in x. To create outliers, we followed Sinha [22,23] to replace one randomly chosen x value within five different subjects by $x + 3$ (i.e. 5% subjects with one outlier). In this simulation study, we also included the GEE estimates based on an independent working correlation. For comparison, we also present the simulation results without outliers. Table 4 summarizes the simulation results. In the case of the Poisson mixed model, the SBE stays almost the same as outliers increase from 0% to 5%, while a significant increase from PQLE and GEE is observed. For the logistic model, the SBE shows smaller biases for the estimation of β_1 and θ in the presence of outliers. For the estimation of fixed effects β_0 and β_1, the SBE provides smaller RMSEs than the PQLE and GEE. It is known that GEE is unbounded and sensitive to data outliers [24]. However, the PQLE of θ appears to have smaller RMSEs. This interesting and counterintuitive phenomenon was also found in a similar simulation study conducted by Sinha [22] and Noh and Lee [25] when they

Table 5. Comparison of parameter estimates and their SEs for the seizure count data.

Parameter	SLSE estimates (SE)	RQLE[a] estimates (SE)	MQLE[a] estimates (SE)
INTERCEPT	−1.324 (1.672)	−1.330 (0.928)	−1.388 (1.248)
BASE	0.915 (0.117)	0.895 (0.083)	0.890 (0.141)
TRT	−0.758 (0.627)	−0.795 (0.446)	−0.849 (0.424)
TRT × Base	0.397 (0.205)	0.260 (0.238)	0.324 (0.216)
AGE	0.453 (0.485)	0.462 (0.277)	0.463 (0.365)
VISIT/10	−0.230 (0.268)	−0.230 (0.156)	−0.253 (0.241)
θ_0	0.135 (0.093)	0.130 (0.050)	0.257 (0.083)
θ_1	0.117 (0.709)	0.116 (0.357)	1.904 (1.386)

[a]Obtained from Sinha [23].

compared their proposed robust estimation methods with the classical likelihood-based method. Similarly, we can argue that the RMSE of the PQLE of θ underestimates because of the relatively larger biases observed in the PQLE of the fixed effects.

5. Application to the seizure count data

In this section, we apply the proposed methods to analyse the popular epilepsy seizure count data presented by Thall and Vail [26]. The data come from a clinical trial of 59 epileptics who were randomized to receive either the antiepileptic drug progabide (TRT $= 1$) or a placebo (TRT $= 0$), as an adjuvant to standard chemotherapy. The logarithm of a quarter of the number of epileptic seizures in the 8-week period preceding the trial (BASE) and the logarithm of age (Age) were included as covariates in the analysis. For each individual, a multivariate response variable consisting of the seizure counts during 2-week periods before each of four clinical visits (VISIT, coded -0.3, -0.1, 0.1 and 0.3) was collected. By a thorough investigation, Thall and Vail [26] identified a number of patients as outliers, who have irregular large counts. Recently, the data were further analysed by Sinha [23] using the robust quasi-likelihood estimator (RQLE) proposed by him. Here, we consider the following model used by Sinha [23]:

$$\log E(y_{ij}|b_i) = x_{ij}'\beta + b_{i0} + b_{i1}\mathrm{VISIT}_{ij}, \tag{20}$$

where $b_{i0} \sim N(0, \theta_0)$ and $b_{i1} \sim N(0, \theta_1)$ are the independent random effects, and x_{ij} represents the vector of the predictors BASE, TRT, AGE and VISIT, and the interaction between BASE and TRT.

Table 5 reports the fixed and random effect estimates by the SBE, the RQLE and the classical marginal quasi-likelihood estimator (MQLE). The estimates of the fixed effects are very similar and the covariate BASE is highly significant by all the three approaches. However, we observed a significant difference in the estimates of the random effects. In particular, the SBE estimates highly agree with the RQL estimates, but are quite different from those obtained by the MQL method. The standard errors (SEs) of θ_0^2 from all approaches are relatively close, but the SBE results in a SE reduction of 50% for θ_1^2 in comparison with the MQLE. Since Sinha [23] concludes that the RQL method appears to be successful in handling outliers in the epilepsy data, we confirm that the SBE has the same property.

6. Concluding remarks

This paper proposes an exact consistent SBE for GLMMs with flexible distributions of random effects. We have established the asymptotic properties of the proposed estimators under mild regularity conditions, and we have demonstrated that the proposed estimator has desirable finite sample properties by simulation studies. In comparison with the likelihood-based method, the proposed approach requires less distributional assumptions and leads to exact consistent (not approximately) estimation. In comparison to GEE and associated simulation-based methods, it is computationally more attractive and does not require any 'working' specification of the weight matrix. Furthermore, the proposed estimator is robust against data outliers. Since the main purpose of this paper is to introduce a new consistent estimator for a GLMM, we did not fully explore its robustness property, except some limited simulation studies. Some future research may be done to investigate its breakdown points and compare it with some popular robust estimation methods in the literature.

Acknowledgements

This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC) and the National Institute for Complex Data Structures (NICDS).

References

[1] N.M. Laird and J.H. Wair, *Random-effects models for longitudinal data*, Biometrics 38 (1982), pp. 963–974.
[2] S.L. Zeger, K.Y. Liang, and P.S. Albert, *Models for longitudinal data: A generalized estimating equation approach*, Biometrics 44 (1988), pp. 1049–1060.
[3] N.E. Breslow and D.G. Clayton, *Approximate inference in generalized linear mixed models*, J. Amer. Statist. Assoc. 88 (1993), pp. 9–25.
[4] S. Rabe-Hesketh, A. Skrondal, and A. Pickles, *Reliable estimation of generalized linear mixed models using adaptive quadrature*, Stata J. 2 (2002), pp. 1–21.
[5] J. Durbin and S.J. Koopman, *Monte Carlo maximum likelihood estimation for non-Gaussian state space models*, Biometrika 84 (1997), pp. 669–684.
[6] H. Joe, *Accuracy of Laplace approximation for discrete response mixed models*, Comput. Stat. Data Anal. 52 (2008), pp. 5066–5074.
[7] X. Lin and N.E. Breslow, *Bias correction in generalized linear mixed models with multiple components of dispersion*, J. Amer. Statist. Assoc. 91 (1996), pp. 1007–1016
[8] J.M. Jiang, *Consistent estimators in generalized linear mixed models*, J. Amer. Statist. Assoc. 93 (1998), pp. 720–729.
[9] B.C. Sutradhar, *On exact quasilikelihood inference in generalized linear mixed models*, Sankhya Indian J. Stat. 66 (2004), pp. 261–289.
[10] L. Wang, *A unified approach to estimation of nonlinear mixed effects and Berkson measurement error models*, Can. J. Stat. 35 (2007), pp. 233–248.
[11] C. Gouriéroux and A. Monfort, *Simulation-Based Econometric Methods*, Oxford University Press, New York, 1996.
[12] D. McFadden, *A method of simulated moments for estimation of discrete response models without numerical integration*, Econometrica 57 (1989), pp. 995–1026.
[13] A. Pakes and D. Pollard, *Simulation and the asymptotics of optimization estimators*, Econometrica 57 (1989), pp. 1027–1057.
[14] L. Wang, *Estimation of nonlinear models with Berkson measurement errors*, Ann. Stat. 32 (2004), pp. 2559–2579.
[15] M. Davidian and A.R. Gallant, *The nonlinear mixed effects model with a smooth random effects density*, Biometrika 80 (1993), pp. 475–488.
[16] D. Zhang and M. Davidian, *Linear mixed models with flexible distributions of random effects for longitudinal data*, Biometrics 57 (2001), pp. 795–802.
[17] C.A. Field and M.G. Genton, *The multivariate g-and-h distribution*, Technometrics 48 (2006), pp. 104–111.
[18] P.J. Huber, *Robust Statistics*, Wiley, New York, 2004.
[19] T. Abarin and L. Wang, *Comparison of GMM with second-order least squares estimator in nonlinear models*, Far East J. Theoret. Stat. 20 (2006), pp. 179–196.
[20] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel, *Robust Statistics: The Approach Based on Influence Functions*, Wiley, New York, 1986.
[21] J.G. Altonji and L.M. Segal, *Small sample bias in GMM estimation of covariance structures*, J. Bus. Econ. Stat. 14 (1996), pp. 353–366.
[22] S.K. Sinha, *Robust analysis of generalized linear mixed models*, J. Amer. Statist. Assoc. 99 (2004), pp. 451–460.
[23] S.K. Sinha, *Robust inference in generalized linear mixed models for longitudinal data*, Can. J. Stat. 34(2) (2006), pp. 1–18.
[24] A. Qu and X.K. Song, *Assessing robustness of generalised estimating equations and quadratic inference functions*, Biometrika 91 (2004), pp. 447–459.

[25] M. Noh and Y. Lee, *Robust modeling for inference from generalized linear model classes*, J. Amer. Statist. Assoc. 102 (2007), pp. 1059–1072.
[26] P.F. Thall and S.C. Vail, *Some covariance models for longitudinal count data with overdispersion*, Biometrics 46 (1990), pp. 657–671.
[27] T. Amemiya, *Regression analysis when the dependent variable is truncated normal*, Econometrica 41 (1973), pp. 997–1016.

Appendix: Technical proofs

Throughout the proofs, we use $g^{(d)}(\cdot)$ and $\nu^{(d)}(\cdot)$, $d = 0, 1, 2$, to denote their dth-order derivatives, and use $f_b^{(d)}(u;\theta)$ to denote its dth-order partial derivative w.r.t. θ.

Proof of Corollary 3.2(1)

For any $1 \le i \le m$, by A1–A3 and the Cauchy–Schwartz inequality, we have

$$\begin{aligned}
\|\rho_i(\psi)\|^2 &\le 2\sum_j y_{ij}^2 + 2\sum_{j\le k} y_{ij}^2 y_{ik}^2 + 2\sum_j \left(\int g(x_{ij}'\beta + z_{ij}'u) f_b(u;\theta)\,\mathrm{d}u\right)^2 \\
&\quad + 4\sum_{j\le k}\left(\int g(x_{ij}'\beta + z_{ij}'u) g(x_{ik}'\beta + z_{ik}'u) f_b(u;\theta)\,\mathrm{d}u\right)^2 \\
&\quad + 4\phi^2 \sum_j \left(\int \nu(g(x_{ij}'\beta + z_{ij}'u)) f_b(u;\theta)\,\mathrm{d}u\right)^2 \\
&\le 2\sum_j y_{ij}^2 + 2\sum_{j\le k} y_{ij}^2 y_{ik}^2 + 2\sum_j \int g^2(x_{ij}'\beta + z_{ij}'u) f_b(u;\theta)\,\mathrm{d}u \\
&\quad + 4\sum_{j\le k}\int g^2(x_{ij}'\beta + z_{ij}'u) f_b(u;\theta) du \int g^2(x_{ik}'\beta + z_{ik}'u) f_b(u;\theta)\,\mathrm{d}u \\
&\quad + 4\phi^2 \sum_j \left(\int \nu(g(x_{ij}'\beta + z_{ij}'u)) f_b(u;\theta)\,\mathrm{d}u\right)^2,
\end{aligned}$$

and therefore,

$$\begin{aligned}
E\sup_\Gamma \rho_i'(\psi) W_i \rho_i(\psi) &\le E\|W_i\| \sup_\Gamma \|\rho_i'(\psi)\|^2 \\
&\le 2n_i E\|W_i\| y_{ij}^2 + n_i(n_i+1) E\|W_i\| y_{ij}^2 y_{ik}^2 + 2n_i E\|W_i\| \int G(X_i, Z_i, u)\,\mathrm{d}u \\
&\quad + 2n_i(n_i + 1 + 2\sup_\Sigma \phi^2) E\|W_i\| \left(\int G(X_i, Z_i, u)\,\mathrm{d}u\right)^2 \\
&< \infty.
\end{aligned}$$

Hence, by the uniform law of large numbers (ULLN), $\sup_{\psi\in\Gamma} |(1/m)Q_m(\psi) - Q(\psi)| \xrightarrow{a.s.} 0$, where $Q(\psi) = E[\rho_i'(\psi) W_i \rho_i(\psi)]$. Furthermore, since $\rho_i(\psi) - \rho_i(\psi_0)$ does not depend on Y_i,

$$\begin{aligned}
Q(\psi) &= E(\rho_i'(\psi) - \rho_i'(\psi_0) + \rho_i'(\psi_0)) W_i (\rho_i(\psi) - \rho_i(\psi_0) + \rho_i(\psi_0)) \\
&= Q(\psi_0) + E(\rho_i(\psi) - \rho_i(\psi_0))' W_i (\rho_i(\psi) - \rho_i(\psi_0)).
\end{aligned}$$

It follows from A4 that $Q(\psi) \ge Q(\psi_0)$ and the equality holds if and only if $\psi = \psi_0$. Thus, all conditions reported by Amemiya [27, Lemma 3] are satisfied, and therefore, $\hat{\psi}_m \xrightarrow{a.s.} \psi_0$, as $m \to \infty$.

Proof of Corollary 3.2(2)

By A5 and the dominated convergence theorem, the first derivative $\partial Q_m(\psi)/\partial\psi$ exists and has the first-order Taylor expansion in Γ. Since $\hat{\psi}_m \xrightarrow{a.s.} \psi_0$, for sufficiently large m, we have

$$\frac{\partial Q_m(\hat{\psi}_m)}{\partial\psi} = \frac{\partial Q_m(\psi_0)}{\partial\psi} + \frac{\partial^2 Q_m(\tilde{\psi}_m)}{\partial\psi\,\partial\psi'}(\hat{\psi}_m - \psi_0) = 0, \tag{A1}$$

where $\|\tilde{\psi}_m - \psi_0\| \le \|\hat{\psi}_m - \psi_0\|$. The first derivative of $Q_m(\psi)$ in Equation (A1) is given by

$$\frac{\partial Q_m(\psi)}{\partial\psi} = 2\sum_{i=1}^{m} \frac{\partial\rho_i'(\psi)}{\partial\psi} W_i \rho_i(\psi),$$

where

$$\frac{\partial\rho_i'(\psi)}{\partial\psi} = -\left(\frac{\partial\mu_{ij}(\psi)}{\partial\psi}, 1 \le j \le n_i, \frac{\partial\eta_{ijk}(\psi)}{\partial\psi}, 1 \le j \le k \le n_i\right)$$

with nonzero first derivatives

$$\begin{aligned}
\frac{\partial\mu_{ij}(\psi)}{\partial\beta} &= x_{ij}\int g^{(1)}(x_{ij}'\beta + z_{ij}'u) f_b(u;\theta)\,\mathrm{d}u,\\
\frac{\partial\mu_{ij}(\psi)}{\partial\theta} &= \int g(x_{ij}'\beta + z_{ij}'u) f_b^{(1)}(u;\theta)\,\mathrm{d}u,\\
\frac{\partial\eta_{ijk}(\psi)}{\partial\beta} &= x_{ij}\int g^{(1)}(x_{ij}'\beta + z_{ij}'u) g(x_{ik}'\beta + z_{ik}'u) f_b(u;\theta)\,\mathrm{d}u\\
&\quad + x_{ik}\int g(x_{ij}'\beta + z_{ij}'u) g^{(1)}(x_{ik}'\beta + z_{ik}'u) f_b(u;\theta)\,\mathrm{d}u\\
&\quad + \delta_{jk}\phi x_{ij}\int \nu^{(1)}(g(x_{ij}'\beta + z_{ij}'u)) g^{(1)}(x_{ij}'\beta + z_{ij}'u) f_b(u;\theta)\,\mathrm{d}u,\\
\frac{\partial\eta_{ijk}(\psi)}{\partial\theta} &= \int g(x_{ij}'\beta + z_{ij}'u) g(x_{ik}'\beta + z_{ik}'u) f_b^{(1)}(u;\theta)\,\mathrm{d}u\\
&\quad + \delta_{jk}\phi\int \nu(g(x_{ij}'\beta + z_{ij}'u)) f_b^{(1)}(u;\theta)\,\mathrm{d}u,\\
\frac{\partial\eta_{ijk}(\psi)}{\partial\phi} &= \delta_{jk}\int \nu(g(x_{ij}'\beta + z_{ij}'u)) f_b(u;\theta)\,\mathrm{d}u.
\end{aligned}$$

Since $(\partial\rho_i'(\psi)/\partial\psi)W_i\rho_i(\psi)$ are i.i.d. with zero mean, it follows from the central limit theorem that, as $m \to \infty$,

$$\frac{1}{\sqrt{m}}\frac{\partial Q_m(\psi_0)}{\partial\psi} \xrightarrow{L} N(0, 4C). \tag{A2}$$

The second derivative of $Q_m(\psi)$ in Equation (A1) is given by

$$\frac{\partial^2 Q_m(\psi)}{\partial\psi\,\partial\psi'} = 2\sum_{i=1}^{m}\left[\frac{\partial\rho_i'(\psi)}{\partial\psi} W_i \frac{\partial\rho_i(\psi)}{\partial\psi'} + (\rho_i'(\psi)W_i \otimes I)\frac{\partial\mathrm{vec}(\partial\rho_i'(\psi)/\partial\psi)}{\partial\psi'}\right],$$

where I is the $2m(p + r + 1)$-dimensional identity matrix and

$$\frac{\partial\mathrm{vec}(\partial\rho_i'(\psi)/\partial\psi)}{\partial\psi'} = -\left(\frac{\partial^2\mu_{ij}(\psi)}{\partial\psi\,\partial\psi'}, 1 \le j \le n_i, \frac{\partial^2\nu_{ijk}(\psi)}{\partial\psi\,\partial\psi'}, 1 \le j \le k \le n_i\right)'$$

with nonzero partial derivatives

$$\frac{\partial^2 \mu_{ij}(\psi)}{\partial\beta\partial\beta'} = x_{ij}x'_{ij}\int g^{(2)}(x'_{ij}\beta + z'_{ij}u) f_b(u;\theta)\,\mathrm{d}u,$$

$$\frac{\partial^2 \mu_{ij}(\psi)}{\partial\theta\partial\theta'} = \int g(x'_{ij}\beta + z'_{ij}u) f_b^{(2)}(u;\theta)\,\mathrm{d}u,$$

$$\frac{\partial^2 \mu_{ij}(\psi)}{\partial\beta\partial\theta'} = x_{ij}\int g^{(1)}(x'_{ij}\beta + z'_{ij}u) f_b^{(1)}(u;\theta)\,\mathrm{d}u,$$

$$\begin{aligned}\frac{\partial^2 \eta_{ijk}(\psi)}{\partial\beta\partial\beta'} = {} & x_{ij}x'_{ij}\int g^{(2)}(x'_{ij}\beta + z'_{ij}u) g(x'_{ik}\beta + z'_{ik}u) f_b(u;\theta)\,\mathrm{d}u \\ & + 2x_{ij}x'_{ik}\int g^{(1)}(x'_{ij}\beta + z'_{ij}u) g^{(1)}(x'_{ik}\beta + z'_{ik}u) f_b(u;\theta)\,\mathrm{d}u \\ & + x_{ik}x'_{ik}\int g(x'_{ij}\beta + z'_{ij}u) g^{(2)}(x'_{ik}\beta + z'_{ik}u) f_b(u;\theta)\,\mathrm{d}u \\ & + \delta_{jk}\phi x_{ij}x'_{ij}\int \nu^{(2)}(g(x'_{ij}\beta + z'_{ij}u))\left(g^{(1)}(x'_{ij}\beta + z'_{ij}u)\right)^2 f_b(u;\theta)\,\mathrm{d}u, \\ & + \delta_{jk}\phi x_{ij}x'_{ij}\int \nu^{(1)}(g(x'_{ij}\beta + z'_{ij}u)) g^{(2)}(x'_{ij}\beta + z'_{ij}u) f_b(u;\theta)\,\mathrm{d}u,\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 \eta_{ijk}(\psi)}{\partial\theta\partial\theta'} = {} & \int g(x'_{ij}\beta + z'_{ij}u) g(x'_{ik}\beta + z'_{ik}u) f_b^{(2)}(u;\theta)\,\mathrm{d}u \\ & + \delta_{jk}\phi\int \nu(g(x'_{ij}\beta + z'_{ij}u)) f_b^{(2)}(u;\theta)\,\mathrm{d}u,\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 \eta_{ijk}(\psi)}{\partial\beta\partial\theta'} = {} & x_{ij}\int g^{(1)}(x'_{ij}\beta + z'_{ij}u) g(x'_{ik}\beta + z'_{ik}u) f_b^{(1)}(u;\theta)\,\mathrm{d}u \\ & + x_{ik}\int g(x'_{ij}\beta + z'_{ij}u) g^{(1)}(x'_{ik}\beta + z'_{ik}u) f_b^{(1)}(u;\theta)\,\mathrm{d}u \\ & + \delta_{jk}\phi\int \nu^{(1)}(g(x'_{ij}\beta + z'_{ij}u)) g^{(1)}(x'_{ij}\beta + z'_{ij}u) f_b^{(1)}(u;\theta)\,\mathrm{d}u.\end{aligned}$$

Analogous to the proof of Corollary 3.2(1), by A1–A5 and the Cauchy–Schwartz inequality, we can verify that

$$E\sup_{\Gamma}\left\|\frac{\partial\rho'_i(\psi)}{\partial\psi}W_i\frac{\partial\rho_i(\psi)}{\partial\psi'}\right\| \le E\|W_i\|\sup_{\Gamma}\left\|\frac{\partial\rho'_i(\psi)}{\partial\psi}\right\|^2 < \infty$$

and

$$\begin{aligned}E\sup_{\Gamma}\left\|(\rho'_i(\psi)W_i\otimes I)\frac{\partial\mathrm{vec}(\partial\rho'_i(\psi)/\partial\psi)}{\partial\psi'}\right\| & \le \sqrt{2m(p+r+1)}E\|W_i\|\sup_{\Gamma}\|\rho_i(\psi)\|\left\|\frac{\partial\mathrm{vec}(\partial\rho'_i(\psi)/\partial\psi)}{\partial\psi'}\right\| \\ & \le \sqrt{2m(p+r+1)}(E\|W_i\|\sup_{\Gamma}\|\rho_i(\psi)\|^2)^{1/2} \\ & \quad \times\left(E\|W_i\|\sup_{\Gamma}\left\|\frac{\partial\mathrm{vec}(\partial\rho'_i(\psi)/\partial\psi)}{\partial\psi'}\right\|^2\right)^{1/2}. \\ & < \infty.\end{aligned}$$

Therefore, by the ULLN and Lemma 4 reported by Amemiya [27], we have

$$\frac{1}{2m}\frac{\partial^2 Q_m(\psi)}{\partial\psi\partial\psi'} \xrightarrow{a.s.} E\left[\frac{\partial\rho'_i(\psi)}{\partial\psi}W_i\frac{\partial\rho_i(\psi)}{\partial\psi'} + (\rho'_i(\psi)W_i\otimes I)\frac{\partial\mathrm{vec}(\partial\rho'_i(\psi)/\partial\psi)}{\partial\psi'}\right] = B, \tag{A3}$$

where the second equality holds because

$$E\left[(\rho'_i(\psi_0)W_i\otimes I)\frac{\partial\mathrm{vec}(\partial\rho'_i(\psi_0)/\partial\psi)}{\partial\psi'}\right] = 0.$$

The result then follows from Equations (A1)–(A3), assumption (A6) and Slutsky's theorem.

Proof of Theorem 3.1(1)

First, the conditional expectation satisfies

$$
\begin{aligned}
E\left(\sup_{\Gamma}\|\rho_{i,1}(\psi)\||Y_i, X_i, Z_i\right) &\leq \sum_j |y_{ij}| + \sum_{j\leq k}|y_{ij}y_{ik}| + \frac{1}{S}\sum_j\sum_{s=1}^{S} E\left(\left.\frac{\sup_\Psi |g(x'_{ij}\beta + z'_{ij}u_{is})| f_b(u_{is};\theta)}{h(u_{is})}\right| X_i, Z_i\right)\\
&\quad + \frac{1}{S}\sum_{j\leq k}\sum_{s=1}^{S} E\left(\left.\frac{\sup_\Psi |g(x'_{ij}\beta + z'_{ij}u_{is}) g(x'_{ik}\beta + z'_{ik}u_{is})| f_b(u_{is};\theta)}{h(u_{is})}\right| X_i, Z_i\right)\\
&\quad + \frac{\sup_\Sigma \phi}{S}\sum_j\sum_{s=1}^{S} E\left(\left.\frac{\sup_\Psi |\nu(g(x'_{ij}\beta + z'_{ij}u_{is}))| f_b(u_{is};\theta)}{h(u_{is})}\right| X_i, Z_i\right)\\
&\leq \sum_j |y_{ij}| + \sum_{j\leq k}|y_{ij}y_{ik}| + \sum_j\left(\int \sup_\Psi |g(x'_{ij}\beta + z'_{ij}u)| f_b(u;\theta)\mathrm{d}u\right)\\
&\quad + \sum_{j\leq k}\left(\int \sup_\Psi |g(x'_{ij}\beta + z'_{ij}u) g(x'_{ik}\beta + z'_{ik}u)| f_b(u;\theta)\mathrm{d}u\right)\\
&\quad + \sup_\Sigma \phi \sum_j\left(\int \sup_\Psi |\nu(g(x'_{ij}\beta + z'_{ij}u))| f_b(u;\theta)\mathrm{d}u\right).
\end{aligned}
$$

Similarly, the above upper bound applies to $E(\sup_\Gamma \|\rho_{i,2}(\psi)\||Y_i, X_i, Z_i)$ as well. Furthermore, since $\rho_{i,1}$ and $\rho_{i,2}$ are conditionally independent given (Y_i, X_i, Z_i), we have

$$
\begin{aligned}
E\left(\sup_\Gamma |\rho_{i,1}(\psi) W_i \rho_{i,2}(\psi)|\right) &\leq E\left[\|W_i\| E\left(\sup_\Gamma \|\rho_{i,1}(\psi)\||Y_i, X_i, Z_i\right) E\left(\sup_\Gamma \|\rho_{i,2}(\psi)\||Y_i, X_i, Z_i\right)\right]\\
&\leq E\|W_i\|\left(\sum_j |y_{ij}| + \sum_{j\leq k}|y_{ij}y_{ik}| + \sum_j \int \sup_\Psi |g(x'_{ij}\beta + z'_{ij}u)| f_b(u;\theta)\,\mathrm{d}u\right.\\
&\quad + \sum_{j\leq k}\int \sup_\Psi |g(x'_{ij}\beta + z'_{ij}u) g(x'_{ik}\beta + z'_{ik}u)| f_b(u;\theta)\,\mathrm{d}u\\
&\quad \left. + \sup_\Sigma \phi \sum_j \int \sup_\Psi |\nu(g(x'_{ij}\beta + z'_{ij}u))| f_b(u;\theta)\,\mathrm{d}u\right)^2.
\end{aligned}
$$

Analogous to the proof of Corollary 3.2(1), we have $E(\sup_\Gamma |\rho_{i,1}(\psi) W_i \rho_{i,2}(\psi)|) < \infty$, and therefore, by the ULLN,

$$
\frac{1}{m} Q_{m,S}(\psi) \xrightarrow{a.s.} E\rho'_{i,1}(\psi) W_i \rho_{i,2}(\psi)
$$

uniformly in $\psi \in \Gamma$, where

$$
E\rho'_{i,1}(\psi) W_i \rho_{i,2}(\psi) = E[E(\rho'_{i,1}(\psi)|X_i, Z_i) W_i E(\rho'_{i,2}(\psi)|X_i, Z_i)] = Q(\psi).
$$

It has been proved previously that $Q(\psi)$ attains a unique minimum at $\psi_0 \in \Gamma$. Therefore, by Lemma 3 reported by Amemiya [27], $\hat{\psi}_{m,S} \xrightarrow{a.s.} \psi_0$, as $m \xrightarrow{a.s.} \infty$.

Proof of Theorem 3.1(2)

For sufficiently large m, we have

$$
\frac{\partial Q_{m,S}(\psi_0)}{\partial \psi} + \frac{\partial^2 Q_{m,S}(\tilde{\psi}_{m,S})}{\partial \psi \partial \psi'}(\hat{\psi}_{m,S} - \psi_0) = 0, \tag{A4}
$$

where $\|\tilde{\psi}_{m,S} - \psi_0\| \leq \|\hat{\psi}_{m,S} - \psi_0\|$ and the first derivative

$$\frac{\partial Q_{m,S}(\psi)}{\partial \psi} = \sum_{i=1}^{m} \left(\frac{\partial \rho_{i,1}'(\psi)}{\partial \psi} W_i \rho_{i,2}(\psi) + \frac{\partial \rho_{i,2}'(\psi)}{\partial \psi} W_i \rho_{i,1}(\psi) \right)$$

is a summation, which are i.i.d. terms with mean zero and common covariance matrix $4C_S$. Hence, by the central limit theorem, we have

$$\frac{1}{\sqrt{m}} \frac{\partial Q_{m,S}(\psi)}{\partial \psi} \xrightarrow{a.s.} N(0, 4C_S). \tag{A5}$$

Next, the second derivative is given by

$$\frac{\partial^2 Q_{m,S}(\psi)}{\partial \psi \partial \psi'} = \sum_{i=1}^{m} \left[\frac{\partial \rho_{i,1}'(\psi)}{\partial \psi} W_i \frac{\partial \rho_{i,2}(\psi)}{\partial \psi'} + (\rho_{i,2}'(\psi) W_i \otimes I) \frac{\partial \mathrm{vec}(\partial \rho_{i,1}'(\psi)/\partial \psi)}{\partial \psi'} \right] + \sum_{i=1}^{m} \left[\frac{\partial \rho_{i,2}'(\psi)}{\partial \psi} W_i \frac{\partial \rho_{i,1}(\psi)}{\partial \psi'} + (\rho_{i,1}'(\psi) W_i \otimes I) \frac{\partial \mathrm{vec}(\partial \rho_{i,2}'(\psi)/\partial \psi)}{\partial \psi'} \right],$$

where I is the $2m(p+r+1)$-dimensional identity matrix. Similar to previous proofs, it can be shown that $(1/m)(\partial^2 Q_{m,S}(\psi)/\partial \psi \partial \psi')$ converges to

$$E\left[\frac{\partial \rho_{i,1}'(\psi_0)}{\partial \psi} W_i \frac{\partial \rho_{i,2}(\psi_0)}{\partial \psi'} + (\rho_{i,2}'(\psi_0) W_i \otimes I) \frac{\partial \mathrm{vec}(\partial \rho_{i,1}'(\psi_0)/\partial \psi)}{\partial \psi'} \right] + E\left[\frac{\partial \rho_{i,2}'(\psi_0)}{\partial \psi} W_i \frac{\partial \rho_{i,1}(\psi_0)}{\partial \psi'} + (\rho_{i,1}'(\psi_0) W_i \otimes I) \frac{\partial \mathrm{vec}(\partial \rho_{i,2}'(\psi_0)/\partial \psi)}{\partial \psi'} \right]$$

uniformly for all $\psi \in \Gamma$. Since

$$E\left[\frac{\partial \rho_{i,1}'(\psi_0)}{\partial \psi} W_i \frac{\partial \rho_{i,2}(\psi_0)}{\partial \psi'} \right] = E\left[\frac{\partial \rho_i'(\psi_0)}{\partial \psi} W_i \frac{\partial \rho_i(\psi_0)}{\partial \psi'} \right] = B$$

and

$$E\left[(\rho_{i,1}'(\psi_0) W_i \otimes I) \frac{\partial \mathrm{vec}(\partial \rho_{i,2}'(\psi_0)/\partial \psi)}{\partial \psi'} \right] = 0,$$

we have

$$\frac{1}{m} \frac{\partial^2 Q_{m,S}(\psi)}{\partial \psi \partial \psi'} \xrightarrow{a.s.} 2B. \tag{A6}$$

Finally, the result follows from Equations (A4)–(A5) and Slutsky's theorem.

Proof of Corollary 3.6

The IF of SBE is bounded if and only if

$$\frac{\partial \rho_{l,1}'(v; \hat{\psi}_{m,S}(F))}{\partial \psi} \hat{A}^{-1} \rho_{l,2}(v; \hat{\psi}_{m,S}(F)) \tag{A7}$$

is bounded. We can express $\hat{A}$ as

$$\hat{A} = \frac{1}{m} \sum_{i=1}^{m} \rho_{i,2} \rho_{i,1}' = \frac{1}{m} (V_l + \rho_{l,2} \rho_{l,1}'),$$

where $V_l = \sum_{i \neq l} \rho_{i,2}\rho'_{i,1}$. Then, by the Sherman–Morrison–Woodbury formula, we have

$$\hat{A}^{-1} = m(V_l + \rho_{i,2}\rho'_{i,1})^{-1} = m\left(V_l^{-1} - \frac{V_l^{-1}\rho_{l,2}\rho'_{l,1}V_l^{-1}}{1+\rho'_{l,1}V_l^{-1}\rho_{l,2}}\right)$$

if V_l is nonsingular, V_l^{-1} and $\hat{A}^{-1}$ exist. Therefore,

$$\hat{A}^{-1}\rho_{l,2} = m\left(V_l^{-1}\rho_{l,2} - \frac{V_l^{-1}\rho_{l,2}\rho'_{l,1}V_l^{-1}\rho_{l,2}}{1+\rho'_{l,1}V_l^{-1}\rho_{l,2}}\right) = m\left(\frac{V_l^{-1}\rho_{l,2}}{1+\rho'_{l,1}V_l^{-1}\rho_{l,2}}\right),$$

and accordingly,

$$\left\|\frac{\partial\rho'_{l,1}}{\partial\psi}\hat{A}^{-1}\rho_{l,2}(v)\right\|^2 = m^2\left(\frac{\rho'_{l,2}(v)V_l^{-1}(\partial\rho_{l,1}/\partial\psi)(\partial\rho'_{l,1}/\partial\psi)V_l^{-1}\rho_{l,2}(v)}{1+\rho'_{l,1}(v)V_l^{-1}\rho_{l,2}(v)}\frac{1}{1+\rho'_{l,1}(v)V_l^{-1}\rho_{l,2}(v)}\right) \to 0$$

as $\|v\| \to \infty$.