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
© 2012 Taylor & Francis
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

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

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

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

where $\nu(\cdot)$ is a known variance function and $\phi$ 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

$$
\mu_{ij}(\psi) = E(y_{ij}|X_i, Z_i) = \int g(x_{ij}' \beta + z_{ij}' u) f_b(u; \theta) \, du \quad (3)
$$

and

$$
\eta_{ijk}(\psi) = E(y_{ij}y_{ik}|X_i, Z_i) = \int g(x_{ij}' \beta + z_{ij}' u)g(x_{ik}' \beta + z_{ik}' u) f_b(u; \theta) \, du 
+ \delta_{jk}\phi \int \nu(g(x_{ij}' \beta + z_{ij}' u)) f_b(u; \theta) \, du, \quad (4)
$$

where $X_i = (x_{i1}', x_{i2}', \ldots, x_{ini}')'$, $Z_i = (z_{i1}', z_{i2}', \ldots, z_{ini}')'$, $\delta_{jk} = 1$ if $j = k$ and 0 otherwise.
In the following, we motivate our approach using two most popular GLMMs as examples to
demonstrate that $\psi$ 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

$$
\mu_{ij}(\psi) = \exp \left( x_{ij}' \beta + \frac{z_{ij}' D(\theta) z_{ij}}{2} \right),
$$

and

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

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
logit $\{Pr(y_{ij} = 1|b_i)\} = x_{ij}' \beta + z_{ij}' b_i$. For this model, we find

$$
\mu_{ij}(\psi) = E(y_{ij}^2|X_i, Z_i) = \int \left( \frac{e^{x_{ij}' \beta + z_{ij}' u}}{1 + e^{x_{ij}' \beta + z_{ij}' u}} \right) f_b(u; \theta) h(u) du,
$$

and

$$
\eta_{ijk}(\psi) = \int \left( \frac{e^{x_{ij}' \beta + z_{ij}' u}}{1 + e^{x_{ij}' \beta + z_{ij}' u}} \right) \left( \frac{e^{x_{ik}' \beta + z_{ik}' u}}{1 + e^{x_{ik}' \beta + z_{ik}' u}} \right) f_b(u; \theta) h(u) du, \quad \text{for } j < k.
$$

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

$$
\mu_{ij,1}(\psi) = \frac{1}{S} \sum_{s=1}^{S} g(x_{ij}' \beta + z_{ij}' u_{is}) f_b(u_{is}; \theta) h(u_{is}),
$$

and

$$
\eta_{ijk,1}(\psi) = \frac{1}{S} \sum_{s=1}^{S} g(x_{ij}' \beta + z_{ij}' u_{is}) g(x_{ik}' \beta + z_{ik}' u_{is}) f_b(u_{is}; \theta) h(u_{is})
\quad + \frac{\delta_{jk} \phi}{S} \sum_{s=1}^{S} \nu(g(x_{ij}' \beta + z_{ij}' u_{is})) f_b(u_{is}; \theta) h(u_{is})
$$

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 \( \psi \) is defined as

\[
\hat{\psi}_{m,S} = \arg\min_{\psi \in \Gamma} Q_{m,S}(\psi) = \arg\min_{\psi \in \Gamma} \sum_{i=1}^{m} \rho_{i,1}(\psi) W_i \rho_{i,2}(\psi),
\]

where \( \rho_{i,1}(\psi) = (y_{ij} - \mu_{ij}(\psi), 1 \leq j \leq n_i, y_{ij} y_{ik} - \eta_{ijk}(\psi), 1 \leq j \leq k \leq n_i)' \) 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 \( Q(\psi) = \sum_{i=1}^{m} \rho_i'(\psi) W_i \rho_i(\psi) \) because \( \rho_{i,1}(\psi) \) and \( \rho_{i,2}(\psi) \) are conditionally independent given \( (Y_i, X_i, Z_i), \) and hence,

\[
E[\rho_{i,1}(\psi) W_i \rho_{i,2}(\psi)] = E[E(\rho_{i,1}(\psi)|Y_i, X_i, Z_i) W_i E(\rho_{i,2}(\psi)|Y_i, X_i, Z_i)] = E(\rho_i(\psi) W_i \rho_i(\psi)),
\]

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

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 \( Z_i, \) 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 \( b_i \) and \( X_i \) 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 \( E[||W_i||(\int G(X_i, Z_i, u) du)^2] < \infty. \)

A3. The parameter space \( \Gamma \subseteq \mathbb{R}^{p+r+1} \) is compact.

A4. \( E[\rho_1(\psi) - \rho_i(\psi_0)] W_i [\rho_1(\psi) - \rho_i(\psi_0)] = 0 \) 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. \( \theta \) in an open subset \( \Theta_0 \subseteq \Theta. \) 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

\[
B = E \left[ \frac{\partial \rho_1'(\psi_0)}{\partial \psi} W_i \frac{\partial \rho_1(\psi_0)}{\partial \psi'} \right]
\]

is nonsingular.

**Theorem 3.1** Suppose that \( \text{Supp}(h) \supseteq \text{Supp}(f_b(\cdot; \theta)) \) for all \( \theta \in \Theta_0. \) Then, \textit{for any fixed} \( S > 0, \) \( m \to \infty, \)

(1) \textit{under A1–A4, } \( \hat{\psi}_{m,S} \overset{a.s.}{\longrightarrow} \psi_0; \)
(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, \( \sqrt{m}(\hat{\psi}_m - \psi_0) \xrightarrow{L} N(0, B^{-1}C_B B^{-1}) \), where \( B \) is given in Equation (12) and

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

**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 \( \xi \) has mean zero and covariance matrix \( I_q \). Hence alternatively, we can rewrite Equations (9) and (10) as

\[
\mu_{ij,1}(\psi) = \frac{1}{S} \sum_{s=1}^{S} \frac{g(x_{ij}' \beta + z_{ij}' u)f_{\xi}(u_{is})}{h(u_{is})},
\]

\[
\eta_{ijk,1}(\psi) = \frac{1}{S} \sum_{s=1}^{S} \frac{g(x_{ij}' \beta + z_{ij}' D(\theta)^{1/2} u_{is}) g(x_{ik}' \beta + z_{ik}' D(\theta)^{1/2} u_{is}) f_{\xi}(u_{is})}{h(u_{is})} \]

\[
+ \frac{\delta_{jk} \phi}{S} \sum_{s=1}^{S} \frac{v(g(x_{ij}' \beta + z_{ij}' D(\theta)^{1/2} u_{is})) f_{\xi}(u_{is})}{h(u_{is})}.
\]

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 \( \psi \) at the \( \tau \)th iteration, and

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

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

The terms \((\rho_{i,1} W_i \otimes I)(\partial \text{vec}(\rho_{i,2}/\partial \psi))/\partial \psi'\) and \((\rho_{i,2} W_i \otimes I)(\partial \text{vec}(\rho_{i,1}/\partial \psi))/\partial \psi'\) 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 \( \text{AV}(\hat{\psi}_{m}(W_i)) - \text{AV}(\hat{\psi}_{m}(W_{i}^{\text{opt}})) \) is nonnegative definite for all possible \( W_i \). It can be shown that \( W_{i}^{\text{opt}} \) is equal to

\[
A_i^{-1} = E[\rho_{i,1}(\psi_0)\rho_{i,2}(\psi_0)|X_i, Z_i]^{-1}.
\]

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

\[
A(\hat{\psi}) = \frac{1}{m} \sum_{i=1}^{m} \rho_{i,1}(\hat{\psi}_{m1})\rho_{i,2}(\hat{\psi}_{m1}).
\]

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]

\[
\text{IF}(v; \hat{\psi}_{m,S}, F) = -B(\hat{\psi}_{m}(F))^{-1} \frac{\partial\rho'_{l,1}(v; \hat{\psi}_{m,S}(F))}{\partial\psi} \hat{A}^{-1} \rho_{l,2}(v; \hat{\psi}_{m,S}(F)),
\]

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 \( \|\text{IF}(v; \hat{\psi}_{m,S}, F)\| \to 0 \) 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 \( \rho_{i,1}(\psi) \) 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) \( \hat{\psi}_{m,S}^{\text{IW}} \) for \( \psi \) is defined as the measurable function that minimizes

\[
Q_{m,S}(\psi) = \frac{1}{K} \sum_{k=1}^{K} \sum_{i=1}^{m_k} (\rho_{i,1}^k(\psi))^\prime A_{i,k}^{-1}(\hat{\psi}) \rho_{i,2}^k(\psi),
\]

where \( \rho_{i,1}^k(\psi) \) is constructed for the \( k \)th group and \( A_{i,k}^{-1}(\hat{\psi}) \) is constructed using all but the \( k \)th 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 \( \text{cov}(\hat{\psi}_{m,S}^k, \hat{\psi}_{m,S}^{k+1}) = 0 \) for \( k \neq 0 \) by design, it can
be easily shown that
\[
\text{cov}(\hat{\psi}_{m,S}^{\text{IW}}) = \frac{1}{K^2} \sum_{k=1}^{K} \text{cov}(\hat{\psi}_{m,S}^{k}),
\]
where \(\hat{\psi}_{m,S}^{k}\) is obtained by minimizing \(\sum_{i=1}^{m_i} (\rho_{i,1}^{k}(\psi))^{-1} A_{i,k}^{-1}(\hat{\psi}) \rho_{i,2}^{k}(\psi)\). 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 \texttt{glmmPQL} package.

The first simulation study was designed based on Example 2.1. In particular, we simulated the model \(E(y_{ij}|b_i) = \beta_0 \gamma_0 + \beta_1 x_{ij} + b_i\), \(j = 1, \ldots, 4\), where \(x_{ij} = 0.1\), \(\beta = (3, -1)\) 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 \(\psi_{m}\) 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 \(\psi_{m}^{\text{IW}}\) and \(\psi_{m,S}^{\text{IW}}\) have slightly higher RMSEs.

<table>
<thead>
<tr>
<th>(m)</th>
<th>PQLE</th>
<th>SLSE</th>
<th>SLSIW</th>
<th>SBE</th>
<th>SBEIW</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\beta_0 = 3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>0.006 (0.082)</td>
<td>−0.086 (0.115)</td>
<td>0.001 (0.162)</td>
<td>−0.069 (0.109)</td>
<td>0.012 (0.168)</td>
</tr>
<tr>
<td>100</td>
<td>0.012 (0.060)</td>
<td>−0.053 (0.077)</td>
<td>−0.009 (0.090)</td>
<td>−0.039 (0.075)</td>
<td>0.007 (0.103)</td>
</tr>
<tr>
<td>200</td>
<td>0.010 (0.040)</td>
<td>−0.029 (0.052)</td>
<td>−0.009 (0.058)</td>
<td>−0.022 (0.055)</td>
<td>0.005 (0.061)</td>
</tr>
<tr>
<td>300</td>
<td>0.006 (0.033)</td>
<td>−0.021 (0.040)</td>
<td>−0.005 (0.040)</td>
<td>−0.016 (0.047)</td>
<td>−0.003 (0.052)</td>
</tr>
<tr>
<td>400</td>
<td>0.009 (0.031)</td>
<td>−0.017 (0.035)</td>
<td>−0.005 (0.034)</td>
<td>−0.010 (0.044)</td>
<td>−0.003 (0.043)</td>
</tr>
<tr>
<td>(\beta_1 = -1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>−0.007 (0.152)</td>
<td>0.007 (0.143)</td>
<td>0.020 (0.341)</td>
<td>0.009 (0.130)</td>
<td>0.005 (0.329)</td>
</tr>
<tr>
<td>100</td>
<td>−0.004 (0.109)</td>
<td>0.006 (0.106)</td>
<td>0.013 (0.180)</td>
<td>0.008 (0.107)</td>
<td>0.007 (0.195)</td>
</tr>
<tr>
<td>200</td>
<td>0.000 (0.073)</td>
<td>0.002 (0.077)</td>
<td>0.015 (0.109)</td>
<td>0.004 (0.074)</td>
<td>0.013 (0.115)</td>
</tr>
<tr>
<td>300</td>
<td>−0.001 (0.064)</td>
<td>0.003 (0.061)</td>
<td>0.007 (0.081)</td>
<td>0.000 (0.058)</td>
<td>0.003 (0.081)</td>
</tr>
<tr>
<td>400</td>
<td>−0.001 (0.056)</td>
<td>0.001 (0.054)</td>
<td>0.003 (0.067)</td>
<td>0.003 (0.054)</td>
<td>0.006 (0.065)</td>
</tr>
<tr>
<td>(\theta = 0.25)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>−0.010 (0.053)</td>
<td>−0.043 (0.060)</td>
<td>0.011 (0.105)</td>
<td>−0.054 (0.076)</td>
<td>0.012 (0.122)</td>
</tr>
<tr>
<td>100</td>
<td>−0.007 (0.040)</td>
<td>−0.043 (0.056)</td>
<td>0.004 (0.066)</td>
<td>−0.045 (0.069)</td>
<td>0.001 (0.081)</td>
</tr>
<tr>
<td>200</td>
<td>−0.004 (0.026)</td>
<td>−0.030 (0.042)</td>
<td>0.012 (0.059)</td>
<td>−0.036 (0.059)</td>
<td>0.000 (0.060)</td>
</tr>
<tr>
<td>300</td>
<td>−0.003 (0.023)</td>
<td>−0.024 (0.035)</td>
<td>0.006 (0.048)</td>
<td>−0.027 (0.051)</td>
<td>0.002 (0.055)</td>
</tr>
<tr>
<td>400</td>
<td>−0.004 (0.019)</td>
<td>−0.022 (0.032)</td>
<td>0.002 (0.033)</td>
<td>−0.025 (0.048)</td>
<td>0.005 (0.048)</td>
</tr>
</tbody>
</table>
for $\beta_1$. For the random effect parameter $\theta$, all estimators present similar estimated RMSEs and PQLE; $\psi_m$ and $\psi_{m,S}$ show some downward bias, while $\psi_{m}^{IW}$ and $\psi_{m,S}^{IW}$ show some upward bias. In Figure 1, a significant higher percent (10–20%) bias is observed in $\psi_m$ 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, $\psi_{m}^{IW}$ and $\psi_{m,S}^{IW}$ 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: $\text{logit}(\Pr(y_{ij} = 1|b_i)) = \beta_0 + \beta_1 \times \text{trt}_i + \beta_2 x_{ij} + b_{i0} + b_{i1}x_{ij}$, where $b_i \sim N((0, 0)', \text{diag}(\theta_0, \theta_1))$. In the present simulation, we selected $m = 200, 300$ and $n = 5$; covariates $\text{trt}_i = 1$ 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 $N((0, 0)', \text{diag}(2, 2))$ 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.

<table>
<thead>
<tr>
<th></th>
<th>PQLE</th>
<th>SBE</th>
<th>PQLE</th>
<th>SBE</th>
<th>PQLE</th>
<th>SBE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\chi^2(3)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_0 = 3$</td>
<td>0.006 (0.011)</td>
<td>-0.031 (0.055)</td>
<td>0.010 (0.101)</td>
<td>-0.028 (0.053)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_1 = -1$</td>
<td>0.002 (0.073)</td>
<td>0.005 (0.079)</td>
<td>0.002 (0.056)</td>
<td>0.007 (0.072)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\theta = 0.25$</td>
<td>0.093 (0.394)</td>
<td>-0.023 (0.039)</td>
<td>0.116 (1.106)</td>
<td>-0.027 (0.041)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$t(4)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_0 = 3$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_1 = -1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\theta = 0.25$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

the random effect $\theta_0$, while the SBE has slightly bigger bias only for the random effect $\theta_1$. 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
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 $\beta_1$ and $\theta$ in the presence of outliers. For the estimation of fixed effects $\beta_0$ and $\beta_1$, 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 $\theta$ 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 4. Biases (RMSE) for the parameter estimates with and without outliers. |
|---------------------------------|----------|----------|----------|
|                                | PQLE     | GEE      | SLSE/SBE |
| **No outliers**                |          |          |          |
| Poisson model                  |          |          |          |
| $\beta_0 = 1$                  | 0.021 (0.060) | 0.1232 (0.1369) | -0.082 (0.103) |
| $\beta_1 = 1$                  | -0.001 (0.038) | -0.0019 (0.0373) | 0.017 (0.043) |
| $\theta = 0.25$                | -0.013 (0.043) | -0.047 (0.062) | 0.097 (1.029) |
| Logistic model                 |          |          |          |
| $\beta_0 = 1$                  | 0.020 (0.212) | -0.0440 (0.0699) | 0.066 (0.306) |
| $\beta_1 = 1$                  | 0.051 (0.229) | -0.0435 (0.0744) | 0.117 (0.317) |
| $\theta = 0.25$                | 0.017 (0.320) | -0.021 (0.571) | -0.013 (0.295) |
| **With outliers**              |          |          |          |
| Poisson model                  |          |          |          |
| $\beta_0 = 1$                  | 0.109 (0.180) | -0.071 (0.188) | -0.070 (0.200) |
| $\beta_1 = 0.5$                | -0.054 (0.189) | 0.029 (0.217) | 0.040 (0.218) |
| $\theta = 1$                  | -0.057 (0.124) | 0.030 (0.141) | 0.030 (0.139) |
| $\theta_0 = 1$                 | -0.108 (0.258) | 0.103 (0.332) | 0.112 (0.375) |
| ($\theta_1 = 0.5$           | 0.074 (0.279) | 0.082 (0.402) | 0.107 (0.392) |
| |                                  |          |          |          |
| Logistic model                 |          |          |          |
| $\beta_0 = 1$                  | 0.113 (0.164) | -0.030 (0.135) | -0.045 (0.154) |
| $\beta_1 = 0.5$                | -0.067 (0.176) | 0.021 (0.170) | 0.024 (0.169) |
| $\theta_0 = 1$                 | -0.116 (0.210) | 0.055 (0.255) | 0.071 (0.298) |
| $\theta_1 = 0.5$               | 0.088 (0.241) | 0.074 (0.319) | 0.073 (0.324) |
Table 5. Comparison of parameter estimates and their SEs for the seizure count data.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>SLSE estimates (SE)</th>
<th>RQLE estimates (SE)</th>
<th>MQLE estimates (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTERCEPT</td>
<td>-1.324 (1.672)</td>
<td>-1.330 (0.928)</td>
<td>-1.388 (1.248)</td>
</tr>
<tr>
<td>BASE</td>
<td>0.915 (0.117)</td>
<td>0.895 (0.083)</td>
<td>0.890 (0.141)</td>
</tr>
<tr>
<td>TRT</td>
<td>-0.758 (0.627)</td>
<td>-0.795 (0.446)</td>
<td>-0.849 (0.424)</td>
</tr>
<tr>
<td>TRT × Base</td>
<td>0.397 (0.205)</td>
<td>0.260 (0.238)</td>
<td>0.324 (0.216)</td>
</tr>
<tr>
<td>AGE</td>
<td>0.453 (0.485)</td>
<td>0.462 (0.277)</td>
<td>0.463 (0.365)</td>
</tr>
<tr>
<td>VISIT/10</td>
<td>-0.230 (0.268)</td>
<td>-0.230 (0.156)</td>
<td>-0.253 (0.241)</td>
</tr>
<tr>
<td>(\theta_0)</td>
<td>0.135 (0.093)</td>
<td>0.130 (0.050)</td>
<td>0.257 (0.083)</td>
</tr>
<tr>
<td>(\theta_1)</td>
<td>0.117 (0.709)</td>
<td>0.116 (0.357)</td>
<td>1.904 (1.386)</td>
</tr>
</tbody>
</table>

*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 \(\theta\) 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]:

\[
\log E(y_{ij}|b_i) = x_{ij}'\beta + b_{i0} + b_{i1}\text{VISIT}_{ij},
\]

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 MQLE method. The standard errors (SEs) of \(\theta_0^2\) from all approaches are relatively close, but the SBE results in a SE reduction of 50% for \(\theta_1^2\) 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

Appendix: Technical proofs

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

Proof of Corollary 3.2(1)

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

$$
\|\rho_i(\psi)\|^2 \leq 2 \sum_j y^2_j \nu^{(0)} + 2 \sum_{j \leq k} y^2_j \nu^{(2)} + 2 \sum_j \left( \int g(x'_{ij} \beta + z_{ij} u) f_b(u; \theta) \, du \right)^2
$$

$$
+ 4 \sum_{j \leq k} \left( \int v(g(x'_{ij} \beta + z_{ij} u)) f_b(u; \theta) \, du \right)^2
$$

$$
+ 4 \phi^2 \sum_j \left( \int v(g(x'_{ij} \beta + z_{ij} u)) f_b(u; \theta) \, du \right)^2
$$

$$
\leq 2 \sum_j y^2_j \nu^{(0)} + 2 \sum_{j \leq k} y^2_j \nu^{(2)} + 2 \int g^2(x'_{ij} \beta + z_{ij} u) f_b(u; \theta) \, du
$$

$$
+ 4 \sum_{j \leq k} \int g^2(x'_{ij} \beta + z_{ij} u) f_b(u; \theta) \, du \int g^2(x'_{ik} \beta + z_{ik} u) f_b(u; \theta) \, du
$$

$$
+ 4 \phi^2 \sum_j \left( \int v(g(x'_{ij} \beta + z_{ij} u)) f_b(u; \theta) \, du \right)^2,
$$

and therefore,

$$
E \sup_{\psi \in \Gamma} \rho_i(\psi)^{W_i(\psi)} \leq E\|W_i\| \sup_{\Gamma} \|\rho_i(\psi)\|^2
$$

$$
\leq 2n_i E\|W_i\| y^2_j \nu^{(0)} + n_i(n_i + 1)E\|W_i\| y^2_j \nu^{(2)} + 2n_i E\|W_i\| \int G(X_i, Z_i, u) \, du
$$

$$
+ 2n_i(n_i + 1 + 2 \sup_{\sum} \phi^2)E\|W_i\| \left( \int G(X_i, Z_i, u) \, du \right)^2
$$

$$
< \infty.
$$

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 $Q(\psi) = E[\rho_i(\psi)^{W_i(\psi)}]$. Furthermore, since $\rho_i(\psi) - \rho_i(\psi_0)$ does not depend on $Y_i$,

$$
Q(\psi) = E(\rho_i(\psi) - \rho_i(\psi_0)) W_i(\rho_i(\psi) - \rho_i(\psi_0) + \rho_i(\psi_0))
$$

$$
= Q(\psi_0) + E(\rho_i(\psi) - \rho_i(\psi_0)) W_i(\rho_i(\psi) - \rho_i(\psi_0)),
$$

It follows from A4 that $Q(\psi) \geq Q(\psi_0)$ 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 \( \frac{\partial Q_m(\psi)}{\partial \psi} \) exists and has the first-order Taylor expansion in \( \Gamma_1 \). Since \( \hat{\psi}_m \xrightarrow{a.s.} \psi_0 \), for sufficiently large \( m \), we have

\[
\frac{\partial Q_m(\hat{\psi}_m)}{\partial \psi} = \frac{\partial Q_m(\psi_0)}{\partial \psi} + \frac{\partial^2 Q_m(\hat{\psi}_m)}{\partial \psi \partial \psi'} (\hat{\psi}_m - \psi_0) = 0,
\]

where \( \| \hat{\psi}_m - \psi_0 \| \leq \| \hat{\psi}_m - \psi_0 \| \). 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

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

with nonzero first derivatives

\[
\frac{\partial \mu_{ij}(\psi)}{\partial \beta} = x_{ij} \int g^{(1)}(x'_{ij} \beta + z'_{ij} u) f_b(u; \theta) du,
\]

\[
\frac{\partial \mu_{ij}(\psi)}{\partial \theta} = \int g(x'_{ij} \beta + z'_{ij} u) f^{(1)}(u; \theta) du,
\]

\[
\frac{\partial \eta_{ijk}(\psi)}{\partial \beta} = x_{ij} \int g^{(1)}(x'_{ij} \beta + z'_{ij} u) g(x'_{ik} \beta + z'_{ik} u) f_b(u; \theta) du
\]
\[
+ x_{ik} \int g(x'_{ij} \beta + z'_{ij} u) g^{(1)}(x'_{ik} \beta + z'_{ik} u) f_b(u; \theta) du
\]
\[
+ \delta_{jk} \phi x_{ij} \int v^{(1)}(g(x'_{ij} \beta + z'_{ij} u)) g^{(1)}(x'_{ij} \beta + z'_{ij} u) f_b(u; \theta) du,
\]

\[
\frac{\partial \eta_{ijk}(\psi)}{\partial \theta} = \int g(x'_{ij} \beta + z'_{ij} u) g(x'_{ik} \beta + z'_{ik} u) f^{(1)}(u; \theta) du
\]
\[
+ \delta_{jk} \phi \int v(g(x'_{ij} \beta + z'_{ij} u)) f^{(1)}(u; \theta) du,
\]

\[
\frac{\partial \eta_{ijk}(\psi)}{\partial \phi} = \delta_{jk} \int v(g(x'_{ij} \beta + z'_{ij} u)) f_b(u; \theta) du.
\]

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 \( m \to \infty \),

\[
\frac{1}{\sqrt{m}} \frac{\partial Q_m(\psi_0)}{\partial \psi} \xrightarrow{L} \mathcal{N}(0, 4C).
\]

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

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

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

\[
\frac{\partial \text{vec}(\partial \rho_i'(\psi)/\partial \psi')}{\partial \psi'} = -\left( \frac{\partial^2 \mu_{ij}(\psi)}{\partial \psi \partial \psi'}, 1 \leq j \leq n_i, \frac{\partial^2 \eta_{ijk}(\psi)}{\partial \psi \partial \psi'}, 1 \leq j \leq k \leq n_i \right).
\]
with nonzero partial derivatives
\[
\frac{\partial^2 \mu_{ij}(\psi)}{\partial \beta \partial \beta'} = x_{ij} x'_{ik} \int g^{(2)}(x'_{jk} \beta + z'_{jk} u) f_b(u; \theta) \, du,
\]
\[
\frac{\partial^2 \mu_{ij}(\psi)}{\partial \theta \partial \beta'} = \int g(x'_{jk} \beta + z'_{jk} u) f_b^{(2)}(u; \theta) \, du,
\]
\[
\frac{\partial^2 \theta_{ij}(\psi)}{\partial \beta \partial \beta'} = x_{ij} \int g^{(1)}(x'_{jk} \beta + z'_{jk} u) f_b^{(1)}(u; \theta) \, du,
\]
\[
\frac{\partial^2 \theta_{ij}(\psi)}{\partial \theta \partial \beta'} = \int g(x'_{jk} \beta + z'_{jk} u) g^{(2)}(x'_{jk} \beta + z'_{jk} u) f_b(u; \theta) \, du,
\]
\[
\frac{\partial^2 \theta_{ij}(\psi)}{\partial \theta \partial \theta'} = x_{ij} \int g^{(1)}(x'_{jk} \beta + z'_{jk} u) g^{(1)}(x'_{jk} \beta + z'_{jk} u) f_b(u; \theta) \, du.
\]

Analogous to the proof of Corollary 3.2(1), by A1–A5 and the Cauchy–Schwartz inequality, we can verify that
\[
E \sup_{r} \left\| \frac{\partial \rho_{i}^{(j)}(\psi)}{\partial \psi} W_{i} \frac{\partial \rho_{i}^{(j)}(\psi)}{\partial \psi'} \right\| \leq E \| W_{i} \| \sup_{r} \left\| \frac{\partial \rho_{i}^{(j)}(\psi)}{\partial \psi} \right\|^{2} < \infty
\]
and
\[
E \sup_{r} \left( \rho_{i}^{(j)}(\psi) W_{i} \otimes I \right) \frac{\partial \text{vec}(\rho_{i}^{(j)}(\psi) / \partial \psi)}{\partial \psi'} \leq \sqrt{2m(p + r + 1)E \| W_{i} \| \sup_{r} \| \rho_{i}^{(j)}(\psi) \|} \left\| \frac{\partial \text{vec}(\rho_{i}^{(j)}(\psi) / \partial \psi)}{\partial \psi'} \right\|^{2}^{1/2} < \infty.
\]
Therefore, by the ULLN and Lemma 4 reported by Amemiya [27], we have
\[
\frac{1}{2m} \frac{\partial^2 Q_{m}(\psi)}{\partial \psi \partial \psi'} \xrightarrow{a.s.} E \left[ \frac{\partial \rho_{i}^{(j)}(\psi)}{\partial \psi} W_{i} \frac{\partial \rho_{i}^{(j)}(\psi)}{\partial \psi'} + (\rho_{i}^{(j)}(\psi) W_{i} \otimes I) \frac{\partial \text{vec}(\rho_{i}^{(j)}(\psi) / \partial \psi)}{\partial \psi'} \right] = B,
\]
where the second equality holds because
\[
E \left[ (\rho_{i}^{(j)}(\psi_{0}) W_{i} \otimes I) \frac{\partial \text{vec}(\rho_{i}^{(j)}(\psi_{0}) / \partial \psi)}{\partial \psi'} \right] = 0.
\]
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
\[
E \left( \sup_{\Gamma} \| \rho_{1.1}(\psi) \|_{Y_i, X_i, Z_i} \right) \leq \sum_j |y_{ij}| + \sum_{j \leq k} |y_{ij} y_{ik}| + \frac{1}{S} \sum_{j} \sum_{s=1}^{S} E \left( \sup_{\rho} \left| g(x_{ij} \beta + z_{ij} u_{is}) \right| f_b(u_{is}; \theta) \right)_{X_i, Z_i}
\]
\[
+ \frac{1}{S} \sum_{j} \sum_{s=1}^{S} E \left( \sup_{\rho} \left| g(x_{ij} \beta + z_{ij} u_{is}) \right| f_b(u_{is}; \theta) \right)_{X_i, Z_i}
\]
\[
+ \sup_{\rho} \sum_{j} \left( \int \sup_{\rho} \left| g(x_{ij} \beta + z_{ij} u) \right| f_b(u; \theta) \right)_{X_i, Z_i}
\]
\[
+ \sup_{\rho} \sum_{j} \left( \int \sup_{\rho} \left| g(x_{ij} \beta + z_{ij} u) \right| f_b(u; \theta) \right)_{X_i, Z_i}.
\]

Similarly, the above upper bound applies to \( E(\sup_{\Gamma} \| \rho_{1.2}(\psi) \|_{Y_i, X_i, Z_i}) \) as well. Furthermore, since \( \rho_{1.1} \) and \( \rho_{1.2} \) are conditionally independent given \((Y_i, X_i, Z_i)\), we have
\[
E \left( \sup_{\Gamma} \| \rho_{1.1}(\psi) \|_{Y_i, X_i, Z_i} \right) \leq E \left[ \| W_i \| E \left( \sup_{\Gamma} \| \rho_{1.1}(\psi) \|_{Y_i, X_i, Z_i} \right) \right]
\]
\[
\leq E \| W_i \| \left( \sum_j |y_{ij}| + \sum_{j \leq k} |y_{ij} y_{ik}| + \sum_{j} \int \sup_{\rho} \left| g(x_{ij} \beta + z_{ij} u) \right| f_b(u; \theta) \right. \right.
\]
\[
+ \left. \sum_{j \leq k} \int \sup_{\rho} \left| g(x_{ij} \beta + z_{ij} u) \right| f_b(u; \theta) \right)
\]
\[
+ \left. \sup_{\rho} \sum_{j} \left( \int \sup_{\rho} \left| g(x_{ij} \beta + z_{ij} u) \right| f_b(u; \theta) \right) \right)^2.
\]

Analogous to the proof of Corollary 3.2(1), we have \( E(\sup_{\Gamma} \| \rho_{1.1}(\psi) W_i \rho_{1.2}(\psi) \|) < \infty \), and therefore, by the ULLN,
\[
\frac{1}{m} Q_{m.5}(\psi) \xrightarrow{a.s.} E \rho_{1.1}'(\psi) W_i \rho_{1.2}(\psi)
\]
uniformly in \( \psi \in \Gamma \), where
\[
E \rho_{1.1}'(\psi) W_i \rho_{1.2}(\psi) = E[\rho_{1.1}'(\psi)|X_i, Z_i] W_i E[\rho_{1.2}(\psi)|X_i, Z_i] = Q(\psi).
\]

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], \( \hat{\psi}_{m.5} \xrightarrow{a.s.} \psi_0 \), as \( m \xrightarrow{a.s.} \infty \).

Proof of Theorem 3.1(2)

For sufficiently large \( m \), we have
\[
\frac{\partial Q_{m.5}(\psi_0)}{\partial \psi} + \frac{\partial^2 Q_{m.5}(\hat{\psi}_{m.5})}{\partial \psi \partial \psi'} (\hat{\psi}_{m.5} - \psi_0) = 0,
\]
where \(\|\hat{\psi}_{m,S} - \psi_0\| \leq \|\tilde{\psi}_{m,S} - \psi_0\|\) 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
\[
\frac{\partial^2 Q_{m,S}(\psi)}{\partial \psi \partial \psi'} = \sum_{i=1}^{m} \left[ \frac{\partial \rho_{i,1}'(\psi)}{\partial \psi} W_i \frac{\partial \rho_{i,2}(\psi)}{\partial \psi'} + (\rho_{i,1}'(\psi) W_i \otimes I) \frac{\partial \text{vec}(\rho_{i,1}'(\psi)/\partial \psi)}{\partial \psi'} \right]
+ \sum_{i=1}^{m} \left[ \frac{\partial \rho_{i,2}'(\psi)}{\partial \psi} W_i \frac{\partial \rho_{i,1}(\psi)}{\partial \psi'} + (\rho_{i,1}'(\psi) W_i \otimes I) \frac{\partial \text{vec}(\rho_{i,2}'(\psi)/\partial \psi)}{\partial \psi'} \right],
\]
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
\[
E \left[ \frac{\partial \rho_{i,1}'(\psi_0)}{\partial \psi} W_i \frac{\partial \rho_{i,2}(\psi_0)}{\partial \psi'} + (\rho_{i,1}'(\psi_0) W_i \otimes I) \frac{\partial \text{vec}(\rho_{i,1}'(\psi_0)/\partial \psi)}{\partial \psi'} \right]
+ E \left[ \frac{\partial \rho_{i,2}'(\psi_0)}{\partial \psi} W_i \frac{\partial \rho_{i,1}(\psi_0)}{\partial \psi'} + (\rho_{i,1}'(\psi_0) W_i \otimes I) \frac{\partial \text{vec}(\rho_{i,2}'(\psi_0)/\partial \psi)}{\partial \psi'} \right]
\]
uniformly for all \(\psi \in \Gamma\). Since
\[
E \left[ \frac{\partial \rho_{i,1}'(\psi_0)}{\partial \psi} W_i \frac{\partial \rho_{i,2}(\psi_0)}{\partial \psi'} \right] = E \left[ \frac{\partial \rho_{i,2}'(\psi_0)}{\partial \psi} W_i \frac{\partial \rho_{i,1}(\psi_0)}{\partial \psi'} \right] = B
\]
and
\[
E \left[ (\rho_{i,1}'(\psi_0) W_i \otimes I) \frac{\partial \text{vec}(\rho_{i,2}'(\psi_0)/\partial \psi)}{\partial \psi'} \right] = 0,
\]
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
\[
\frac{\partial \rho_{i,1}'(\psi; \hat{\psi}_{m,S}(F))}{\partial \psi} A^{-1} \rho_{i,2}(v; \hat{\psi}_{m,S}(F)) \tag{A7}
\]
is bounded. We can express \(\hat{A}\) as
\[
\hat{A} = \frac{1}{m} \sum_{i=1}^{m} \rho_{i,2} \rho_{i,1}' = \frac{1}{m} (V_i + \rho_{i,2} \rho_{i,1}).
\]
where $V_l = \sum_{i \neq l} \rho_{l,2} \rho'_{l,1}$. Then, by the Sherman–Morrison–Woodbury formula, we have

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

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

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

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