Communications in Statistics—Theory and Methods, 42: 1982–1997, 2013

ISSN: 0361-0926 print/1532-415X online
DOI: 10.1080/03610926.2011.601837

A Semiparametric Estimation Approach for Linear Mixed Models

DANIEL LI[1] AND LIQUN WANG[1,2]

[1]Department of Statistics, University of Manitoba, Winnipeg, Manitoba, Canada
[2]School of Science, Beijing Jiaotong University, Beijing, China

Maximum likelihood approach is the most frequently employed approach for the inference of linear mixed models. However, it relies on the normal distributional assumption of the random effects and the within-subject errors, and it is lack of robustness against outliers. This article proposes a semiparametric estimation approach for linear mixed models. This approach is based on the first two marginal moments of the response variable, and does not require any parametric distributional assumptions of random effects or error terms. The consistency and asymptotically normality of the estimator are derived under fairly general conditions. In addition, we show that the proposed estimator has a bounded influence function and a redescending property so it is robust to outliers. The methodology is illustrated through an application to the famed Framingham cholesterol data. The finite sample behavior and the robustness properties of the proposed estimator are evaluated through extensive simulation studies.

Keywords Least squares method; Linear mixed models; Misspecification; Outliers; Redescending M-estimator; Robustness.

1. Introduction

Linear mixed models (LMM; Laird and Wair, 1982) are a common framework used to analyze repeatedly measured and clustered data which arise in many areas, such as medical and biological sciences, epidemiology, agriculture, social, and environmental sciences. For the estimation and inference of LMM, the most frequently employed approach is the maximum likelihood (ML) approach. In general, the computation of likelihood function is not simple and relies on Gaussian assumption for both random effects and residual error terms. 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 Gaussian assumption is not appropriate. Several extensions of the LMM have been proposed to relax the Gaussian assumption for the random effects (e.g., Lin

Received August 27, 2010; Accepted June 24, 2011
Address correspondence to Liqun Wang, University of Manitoba, Winnipeg, Manitoba, Canada R3T 2N2; E-mail: liqun.wang@ad.umanitoba.ca

and Lee, 2008; Verbeke and Lesaffre, 1997; Zhang and Davidian, 2001). However, these works still assume the distribution of residual errors to be normal, and impose certain parametric assumptions for random effects distribution, such as Student-t, mixture-normal, or skew-normal. Further, it is well known that MLE is vulnerable to data outliers (Pinheiro et al., 2004).

In this article, we propose the second-order least squares estimator (SLSE) for LMM. This estimator is based on the first two marginal moments of the response variables, which can be computed easily without any further distributional assumptions on random effects or residual error terms. In this sense, the proposed approach can be viewed as a semiparametric approach. The idea of the SLSE was introduced by Wang (2007) for nonlinear mixed models, and Wang and Leblanc (2008) for nonlinear regression models. However, as any other computation intensive methods, SLSE has many computational issues that need to be addressed, and its finite sample performance under misspecified models need to be investigated. Further, the robustness of the SLSE against data outliers remains unknown. The later is a natural and intuitive concern because the effect of outliers may be exaggerated by taking the second absolute moments. In this article, we will address these issues in the framework of LMM. In addition, we relax the high-level regularity conditions in Wang (2007) that are only necessary for nonlinear settings. We derive the asymptotic properties of the SLSE for LMM without these condisiotns.

This article is organized as follows. Section 2 introduces the SLSE, and gives its consistency, asymptotic normality, and redescending properties. Section 3 examines the performance of the SLSE in comparison with the MLE when the distributional assumptions of random effects and error terms are misspecified, and investigates how SLSE behaves by implementing different specifications of the optimal weight matrix. The robustness property of the SLSE against data contamination is also studied via simulation in this section. A real data application is given in Sec. 4, and a discussion is given in Sec. 5. Finally, proofs of the theorems are provided in the Appendix.

2. Second-Order Least Squares Estimator

For a subject i $(i = 1, \ldots, m)$ being observed or measured repeatedly on n_i occasions, the linear mixed model (LMM) can be expressed as

$$y_i = X_i\beta + Z_i b_i + \epsilon_i,$$

where y_i is the $n_i \times 1$ vector of responses, β is a $p \times 1$ vector of the fixed population effects, and b_i is a $q \times 1$ vector of ith subject's random effects and follows a certain distribution with mean 0 and covariance $D(\theta)$. $D(\theta)$ is a $q \times q$ positive-definite covariance matrix depending on a $r \times 1$ vector of parameters θ. X_i and Z_i are the $n_i \times p$ and $n_i \times q$ design matrices to link β and b_i to y_i, respectively. ϵ_i is the $n_i \times 1$ vector of residual error terms following a certain distribution with mean 0 and covariance $\sigma^2 I_{n_i}$. Also, all random vectors $\{b_i, \epsilon_i, i = 1, \ldots, m\}$ are assumed mutually independent. For a subject i at a given occasion j, the LMM can be written as

$$y_{ij} = x'_{ij}\beta + z'_{ij}b_i + \epsilon_{ij}, \tag{2.1}$$

where x'_{ij} and z'_{ij} are the jth rows of the design matrixes X_i and Z_i, respectively. The closed form of the first two marginal moments of the response in model (2.1) are

$$E(y_{ij} \mid X_i, Z_i) = x'_{ij}\beta, \tag{2.2}$$

$$E(y_{ij}y_{ik} \mid X_i, Z_i) = (x'_{ij}\beta)(x'_{ik}\beta) + z'_{ij}D(\theta)z_{ik} + \delta_{jk}\sigma^2, \tag{2.3}$$

where $\delta_{jk} = 1$ if $j = k$ and 0 otherwise. Note that the derivation of Eqs. (2) and (3) dose not require any parametric assumption for the distribution of random effects or error terms.

Let $\Upsilon = (\beta', \theta', \sigma^2)'$ and the parameter space $\Gamma = \Omega \times \Theta \times \Sigma \subset \mathbb{R}^{p+r+1}$. Following Wang (2007), the SLSE $\hat{\Upsilon}_m$ for Υ is defined as the measurable function that minimizes

$$Q_m(\Upsilon) = \sum_{i=1}^{m} \rho'_i(\Upsilon)W_i\rho_i(\Upsilon), \tag{2.4}$$

where $\rho_i(\Upsilon) = (y_{ij} - \mu_{ij}(\Upsilon), 1 \le j \le n_i, y_{ij}y_{ik} - \eta_{ijk}(\Upsilon), 1 \le j \le k \le n_i)'$, $\mu_{ij}(\Upsilon) = E(y_{ij} \mid X_i, Z_i)$, $\eta_{ijk}(\Upsilon) = E(y_{ij}y_{ik} \mid X_i, Z_i)$ and $W_i = W(X_i, Z_i)$ is a non negative definite matrix of dimension $n_i(n_i + 3)/2$.

2.1. *Asymptotic Properties of the SLSE*

To simplify the notation, we present our theoretical results for the case where $n_i = n, i = 1, \ldots, m$ without loss of generality. The following assumptions are used for the proof of the consistency and asymptotic properties of $\hat{\Upsilon}_m$.

Assumption 1. $(y_i, X_i, Z_i, n_i), i = 1, \ldots, m$ are independent and identically distributed and satisfy $E\|W_i\|(y_{ij}^4 + \|x_{ij}\|^4 + \|z_{ij}\|^4 + 1) < \infty$, where $\|\cdot\|$ denotes the Euclidean norm.

Assumption 2. The parameter space $\Gamma \subset \mathbb{R}^{p+r+1}$ is compact.

Assumption 3. $E[(\rho_i(\Upsilon) - \rho_i(\Upsilon_0))'W_i(\rho_i(\Upsilon) - \rho_i(\Upsilon_0))] = 0$ if and only if $\Upsilon = \Upsilon_0$.

Assumption 4. The matrix $B = E[\frac{\partial \rho'_i(\Upsilon_0)}{\partial \Upsilon} W_i \frac{\partial \rho_i(\Upsilon_0)}{\partial \Upsilon'}]$ is non singular.

These are common assumptions in the literature of linear models. In particular, Assumptions 1 and 2 ensure that $Q_m(\Upsilon)$ uniformly converges to $Q(\Upsilon) = E\rho'_i(\Upsilon)W_i\rho_i(\Upsilon)$. Assumption 3 is a high-level identification condition to guarantee that $Q(\Upsilon)$ attains a unique minimum at the true parameter value $\Upsilon_0 \in \Gamma$. A sufficient condition for Assumption 3 is that the matrix $\sum X'_iX_i$ is non singular with $\sum n_i > p$ and at least one matrix Z'_iZ_i is positive definite with $\sum_{i=1}^m (n_i - q) > 0$, provided all random variables in the model are normally distributed (Demidenko, 2004). Finally, Assumption 4 is necessary for the existence of the variance of $\hat{\Upsilon}_m$.

Theorem 2.1. *Under Assumptions* 1–3, *as* $m \to \infty$, $\hat{\Upsilon}_m \xrightarrow{a.s.} \Upsilon_0$.

Theorem 2.2. *Under Assumptions* 1–4, *as* $m \to \infty$, $\sqrt{m}(\widehat{\Upsilon}_m - \Upsilon_0) \xrightarrow{L} N(0, B^{-1}CB^{-1})$, *where*

$$B = E\left[\frac{\partial \rho_i'(\Upsilon_0)}{\partial \Upsilon} W_i \frac{\partial \rho_i(\Upsilon_0)}{\partial \Upsilon'}\right] \tag{2.5}$$

and

$$C = E\left[\frac{\partial \rho_i'(\Upsilon_0)}{\partial \Upsilon} W_i \rho_i(\Upsilon_0)\rho_i'(\Upsilon_0) W_i \frac{\partial \rho_i(\Upsilon_0)}{\partial \Upsilon'}\right]. \tag{2.6}$$

Furthermore, with probability one,

$$B = \lim_{m\to\infty} \frac{1}{m}\sum_{i=1}^{m}\left[\frac{\partial \rho_i'(\widehat{\Upsilon}_m)}{\partial \Upsilon} W_i \frac{\partial \rho_i(\widehat{\Upsilon}_m)}{\partial \Upsilon'}\right]$$

and

$$C = \lim_{m\to\infty} \frac{1}{m}\sum_{i=1}^{m}\frac{\partial \rho_i'(\widehat{\Upsilon}_m)}{\partial \Upsilon} W_i \rho_i(\widehat{\Upsilon}_m)\rho_i'(\widehat{\Upsilon}_m) W_i \frac{\partial \rho_i(\widehat{\Upsilon}_m)}{\partial \Upsilon'}.$$

2.2. *Computation of the SLSE*

In general, there is no explicit solution for the SLSE. The iterative Newton-Raphson algorithm could be used to compute SLSE, that is,

$$\hat{\Upsilon}^{(t+1)} = \hat{\Upsilon}^{(t)} - \left[\frac{\partial^2 Q_m(\hat{\Upsilon}^{(t)})}{\partial\Upsilon\partial\Upsilon'}\right]^{-1}\frac{\partial Q_m(\hat{\Upsilon}^{(t)})}{\partial \Upsilon},$$

where $\hat{\Upsilon}^{(t)}$ denotes the estimate of Υ at the tth iteration,

$$\frac{\partial Q_m(\hat{\Upsilon}^{(t)})}{\partial \Upsilon} = 2\sum_{i=1}^{m}\frac{\partial \rho_i'(\hat{\Upsilon}^{(t)})}{\partial \Upsilon} W_i \rho_i(\hat{\Upsilon}^{(t)}), \quad \text{and} \tag{2.7}$$

$$\frac{\partial^2 Q_m(\hat{\Upsilon}^{(t)})}{\partial\Upsilon\partial\Upsilon'} = 2\sum_{i=1}^{m}\left[\frac{\partial \rho_i'(\hat{\Upsilon}^{(t)})}{\partial \Upsilon} W_i \frac{\partial \rho_i(\hat{\Upsilon}^{(t)})}{\partial \Upsilon'} + (\rho_i'(\hat{\Upsilon}^{(t)})W_i \otimes I)\frac{\partial \mathrm{vec}(\partial\rho_i'(\hat{\Upsilon}^{(t)})/\partial\Upsilon)}{\partial\Upsilon'}\right], \tag{2.8}$$

where $\partial\rho_i'(\Upsilon)/\partial\Upsilon = -(\partial\mu_{ij}(\Upsilon)/\partial\Upsilon, 1 \le j \le n_i, \partial\eta_{ijk}(\Upsilon)/\partial\Upsilon, 1 \le j \le k \le n_i)$ and I is the $2m(p+r+1)$ dimensional identity matrix. In the above equation, since the term $(\rho_i'(\hat{\Upsilon}^{(t)})W_i \otimes I)\frac{\partial \mathrm{vec}(\partial\rho_i'(\hat{\Upsilon}_t)/\partial\Upsilon)}{\partial\Upsilon'}$ has expectation zero, it can be ignored from the second derivative. Therefore, we have the following modified Newton-Raphson algorithm:

$$\hat{\Upsilon}^{(t+1)} = \hat{\Upsilon}^{(t)} - \left[\sum_{i=1}^{m}\frac{\partial \rho_i'(\hat{\Upsilon}^{(t)})}{\partial \Upsilon} W_i \frac{\partial \rho_i(\hat{\Upsilon}^{(t)})}{\partial \Upsilon'}\right]^{-1}\sum_{i=1}^{m}\frac{\partial \rho_i'(\hat{\Upsilon}^{(t)})}{\partial \Upsilon} W_i \rho_i(\hat{\Upsilon}^{(t)}). \tag{2.9}$$

For initial values in (2.9), we can use either the ML estimates or method of moment estimates. To avoid the complexity of finding the derivatives of $Q_N(\psi)$, we can also choose the Nelder-Mead simplex method (Nelder and Mead, 1965) to minimize the quadratic inference function $Q_N(\psi)$ to obtain $\hat{\psi}$.

Another question is how to specify the form of weight W_i to carry out the SLSE. In theory, W_i only depends on X_i and Z_i, and any form of W_i satisfying the regularity conditions is valid for the SLS estimator. However, it would be desirable to make inferences based on the more precise estimator, so the optimal choice of W_i is the one which yields the minimum variance-covariance matrix of $\hat{\Upsilon}_m$. Abarin and Wang (2006) showed this optimal choice is $W_i = U_i^{-1}$ where $U_i = E[\rho_i(\Upsilon_0)\rho_i'(\Upsilon_0) \mid X_i, Z_i]$. In this case, the minimum asymptotic variance-covariance matrix of $\widehat{\Upsilon}_m$ is $E[\frac{\partial \rho_i'(\Upsilon_0)}{\partial \Upsilon} U_i^{-1} \frac{\partial \rho_i(\Upsilon_0)}{\partial \Upsilon'}]$. In practice, the calculation of U_i is not feasible since it involves unknown parameters which need to be estimated first. One of the possible solution is using a two-stage procedure. First, minimize $Q_m(\Upsilon)$ using a sub-optimal choice of W_i, such as an identity matrix, to obtain the first stage estimator $\hat{\Upsilon}_{m1}$. Second, estimate U_i using $\hat{\Upsilon}_{m1}$ and then minimize $Q_m(\Upsilon)$ again with $W_i = \hat{U}_i^{-1}$ to obtain the second stage estimator $\hat{\Upsilon}_{m2}$. In theory, $\hat{\Upsilon}_{m2}$ is asymptotically more efficient than $\hat{\Upsilon}_{m1}$ because $\hat{\Upsilon}_{m2}$ has the minimum asymptotic variance-covariance matrix. In general, U_i can be estimated using any nonparametric method, such as kernel or spline estimators. However, a simpler estimator of U_i would be

$$\hat{U}_i = \frac{1}{m}\sum_{i=1}^{m} \rho_i(\hat{\Upsilon}_{m1})\rho_i'(\hat{\Upsilon}_{m1}). \tag{2.10}$$

In many real data applications, the subjects are clustered so that the values of X_i, Z_i are equal for all subjects within one cluster. In such cases, each U_i can be estimated similarly to (2.10) using all the subjects within the same cluster. Since $\hat{U}_i$ is of dimension $n_i(n_i+3)/2$, numerical inversion of $\hat{U}_i$ may be difficult when n_i is large. In this case, one may consider using diagonal or certain block diagonal sub-matrix of U_i. In Sec. 3, we conduct extensive simulation studies to investigate the sensitivity and efficiency of SLSE by using different specifications of the weight matrix.

2.3. *Robustness of the SLSE*

Outliers are common in experimental research data for reasons such as transcription error or technical equipment malfunction. If no action is implemented, such outliers may distort an analysis completely and lead to inappropriate conclusions. In mixed models, outliers may happen not only at the level of within-subject error but also at the level of within-subject variations.

Here, we study the robustness property of SLSE by means of the influence function (IF), which was introduced by Hampel et al. (1986). The essential concept of IF is that one can use it to assess the asymptotical bias of the estimator caused by a certain degree of data contamination. The estimator is robust if the IF is bounded (Huber, 2004). In principle, the SLSE is an M-estimator (Huber, 2004) and minimizing the quadratic distance function (2.4) with optimal weight matrix in (2.10) is asymptotically equivalent to solving the equation

$$\sum_{i=1}^{m} \frac{\partial \rho_i'(\Upsilon)}{\partial \Upsilon} W_i \rho_i(\Upsilon) = 0. \tag{2.11}$$

It follows from Hampel et al. (1986) that when $m \to \infty$, the IF of the SLSE at point $v = (x_l, z_l)'$ is

$$\mathrm{IF}(v; \hat{\Upsilon}_m, F) = -B(\hat{\Upsilon}_m)^{-1} G(v; \hat{\Upsilon}_m, F) \tag{2.12}$$

where F is the underlying distribution and B is given in (2.5), and

$$G(v; \hat{\Upsilon}_m, F) = \frac{\partial \rho_l'(\hat{\Upsilon}_m)}{\partial \Upsilon} W_l \rho_l(\hat{\Upsilon}_m). \tag{2.13}$$

If $\hat{\Upsilon}_m$ is computed using the estimated optimal weight (2.10), we can show that as $\|v\| \to \infty$

$$\|\mathrm{IF}(v, \hat{\Upsilon}_m)\| \to 0. \tag{2.14}$$

This implies that the $\hat{\Upsilon}_m$ is a redescending M-estimator (Huber, 2004) so it is able to reject extreme outliers completely. Intuitively, it is expected that the outlier will be automatically downweighted by the inverse of the optimal weight matrix. It does not require to screen data for outliers and make a subjective decision to exclude them from the analysis. This is practically meaningful because an outlier may be an indication of a problem with the data generation process but more importantly it may be a true unusual observation about reality.

3. Monte Carlo Simulation Studies

In this section, we carry out simulation studies: (1) to examine finite sample behavior of the SLSE; (2) to evaluate and compare the robustness of SLSE with restricted maximum likelihood (REML) estimator under misspecified random effects and residual error distributions; (3) to investigate the sensitivity and efficiency of SLSE by using different specifications of the weight; and (4) to demonstrate the robustness of SLSE against outliers. We considered the following two linear mixed models commonly used to study the growth curves (Demidenko, 2004; Jacqmin-Gadda et al., 2006):

1. random intercept (RI) model: $y_{ij} = \beta_1 + \beta_2 x_{ij} + b_{i1} + \epsilon_{ij}$;
2. random intercept and slope (RIS) model: $y_{ij} = \beta_1 + \beta_2 x_{ij} + b_{i1} + b_{i2} x_{ij} + \epsilon_{ij}$.

The following configurations are used for simulation:

- $m = 20, 50, 100, 200, 300, 400, 500$; $n = 4$ or 8; and $x_{ij} = j$, $j = 1, \ldots, n$;
- b_{i1}, b_{i2}, and ϵ_{ij} are all generated independently from one of the following distribution: Gaussian, $\chi^2(3)$ and student's $t(4)$ distributions with mean 0 and variance θ_{11}, θ_{22}, and σ^2 respectively;
- $\beta_1 = 8$, $\beta_2 = 2$, $\theta_{11} = 1.96$, $\theta_{22} = 1$ and $\sigma^2 = 1$.

All computations are done in R and the restricted maximum likelihood (REML) estimates are obtained from `lme` package. The SLSEs are computed using three different weight matrices:

1. identity weight (SLS1);
2. diagonal of the estimated optimal weight (2.10) (SLS2);
3. fully estimated optimal weight (2.10) (SLS3).

To determine how well the methods perform, we present the estimation bias and mean squared errors (MSE) of the estimators. For each model, 1,000 Monte Carlo replications were carried out. For fair comparisons, the same dataset was used to obtain both REML estimates and SLS estimates, at each replication. To eliminate potential nonlinear numerical optimization problems on the selection of starting points, the true parameter values were used as starting values for the minimization and the optimal weight calculation for SLS method.

Table 1
Simulation results from RI model with Gaussian and non-Gaussian distributed random effect and residual errors based on 1000 iterations

		Normal			$t(4)$			$\chi^2(3)$		
		$n = 20$	$n = 100$	$n = 500$	$n = 20$	$n = 100$	$n = 500$	$n = 20$	$n = 100$	$n = 500$
					$\beta_1 = 8$					
REML	Bias	0.0122	0.0000	0.0005	−0.0005	0.0020	0.0018	−0.0474	0.0012	0.0032
	MSE	0.1577	0.0299	0.0065	0.1523	0.0297	0.0062	0.1498	0.0333	0.0061
SLS1	Bias	−0.0241	−0.0131	−0.0101	−0.0485	−0.0151	−0.0079	−0.0963	−0.0139	−0.0075
	MSE	0.1590	0.0295	0.0062	0.1571	0.0299	0.0059	0.1771	0.0388	0.0068
SLS2	Bias	−0.0169	−0.0124	−0.0109	−0.0310	−0.0157	−0.0091	−0.0811	−0.0144	−0.0074
	MSE	0.1530	0.0301	0.0065	0.1457	0.0301	0.0062	0.1555	0.0369	0.0068
SLS3	Bias	−0.0206	−0.0016	−0.0002	−0.0011	−0.0027	0.0010	−0.1448	−0.0779	−0.0016
	MSE	0.0789	0.0270	0.0066	0.0624	0.0255	0.0063	0.0872	0.0308	0.0043
					$\beta_2 = 2$					
REML	Bias	−0.0026	0.0013	0.0001	0.0016	−0.0007	−0.0001	0.0064	−0.0003	−0.0004
	MSE	0.0099	0.0010	0.0002	0.0091	0.0010	0.0002	0.0098	0.0010	0.0002
SLS1	Bias	0.0041	0.0039	0.0022	0.0102	0.0021	0.0018	0.0141	0.0025	0.0018
	MSE	0.0097	0.0010	0.0002	0.0089	0.0010	0.0002	0.0103	0.0010	0.0002
SLS2	Bias	0.0048	0.0038	0.0028	0.0090	0.0026	0.0023	0.0125	0.0024	0.0020
	MSE	0.0099	0.0010	0.0002	0.0086	0.0010	0.0002	0.0098	0.0010	0.0002
SLS3	Bias	−0.0031	0.0014	0.0002	0.0011	−0.0008	−0.0001	0.0005	−0.0006	−0.0004
	MSE	0.0049	0.0009	0.0002	0.0034	0.0007	0.0002	0.0028	0.0005	0.0001
					$\theta_{11} = 1.96$					
REML	Bias	−0.0062	0.0081	−0.0050	−0.1472	−0.0745	−0.0012	−0.1634	−0.0155	0.0086
	MSE	0.5001	0.0966	0.0187	0.8474	0.3146	0.1045	0.7556	0.2537	0.0475
SLS1	Bias	0.2445	0.0293	0.0164	0.2836	0.0439	0.0198	0.3385	0.0401	0.0209
	MSE	0.2241	0.0072	0.0019	0.2822	0.0204	0.0079	0.3352	0.0216	0.0052
SLS2	Bias	0.0617	0.0525	0.0135	−0.0145	0.0561	0.0258	−0.0199	0.0447	0.0154
	MSE	0.1729	0.0268	0.0024	0.2336	0.0575	0.0143	0.2233	0.0604	0.0058
SLS3	Bias	−0.2272	−0.1218	−0.0367	−0.3357	−0.3121	−0.1026	−0.2804	−0.2342	−0.0389
	MSE	0.2037	0.0909	0.0186	0.2594	0.2325	0.0725	0.2168	0.2033	0.0233
					$\sigma^2 = 1$					
REML	Bias	0.0058	−0.0006	−0.0014	−0.0419	−0.0029	−0.0013	−0.0158	−0.0038	−0.0011
	MSE	0.0351	0.0049	0.0010	0.0935	0.0271	0.0066	0.0735	0.0130	0.0026
SLS1	Bias	0.0846	0.0089	0.0048	0.0759	0.0151	0.0064	0.0983	0.0119	0.0060
	MSE	0.0324	0.0009	0.0002	0.0466	0.0046	0.0010	0.0507	0.0021	0.0005
SLS2	Bias	−0.0125	0.0162	0.0040	−0.0472	0.0087	0.0083	−0.0657	−0.0026	0.0039
	MSE	0.0337	0.0028	0.0002	0.0553	0.0057	0.0015	0.0433	0.0048	0.0004
SLS3	Bias	−0.1845	−0.0983	−0.0254	−0.2611	−0.2186	−0.0981	−0.2516	−0.1884	−0.0575
	MSE	0.0435	0.0137	0.0016	0.0789	0.0542	0.0125	0.0755	0.0437	0.0060

Since the relative performances of the estimates are similar for RI and RIS model with $n = 4$ or 8, in consideration of space and clarity, we only present the simulation results for the RI model at $n = 4$ in this section. A selection of simulation results for $m = 20$, 100, 500 are provided in Table 1. Overall simulation results in all sample sizes are summarized in Figs. 1–4. These figures contain the percentage of estimation bias and MSE.

Figures 1 and 2 depict the performance of SLS and REML methods for fixed effects. They show all Monte Carlo mean estimates are close to the true parameter values and no apparent biases are observed across all methods. This is not surprising as a few simulations studies (e.g., Jacqmin-Gadda et al., 2006; Verbeke and Lesaffre, 1997) have shown that maximum likelihood inference on fixed effects is robust to misspecified LMM. At relative small sample size ($m = 20$, 50, 100), SLS2 and SLS3 have lower MSE than REML and SLS1. As sample size increases from 200 to 500, all 4 methods behave very closely. Figures 3 and 4 depict the performance of estimators for θ_{11} and σ^2. REML results in very small biases under all circumstances even though the bias increases under misspecified models especially in small sample sizes. In all cases, SLS estimates show similar or much smaller MSE than REML, particulary when the model is misspecified. The MSE reduction in the misspecified model can be as high as 70–80% in some instances. SLS3 suffers some downward bias, although this bias decreases with the increase of sample size. As a result, SLS3 is usually dominated by SLS1 and SLS2 based on the criteria of MSE.

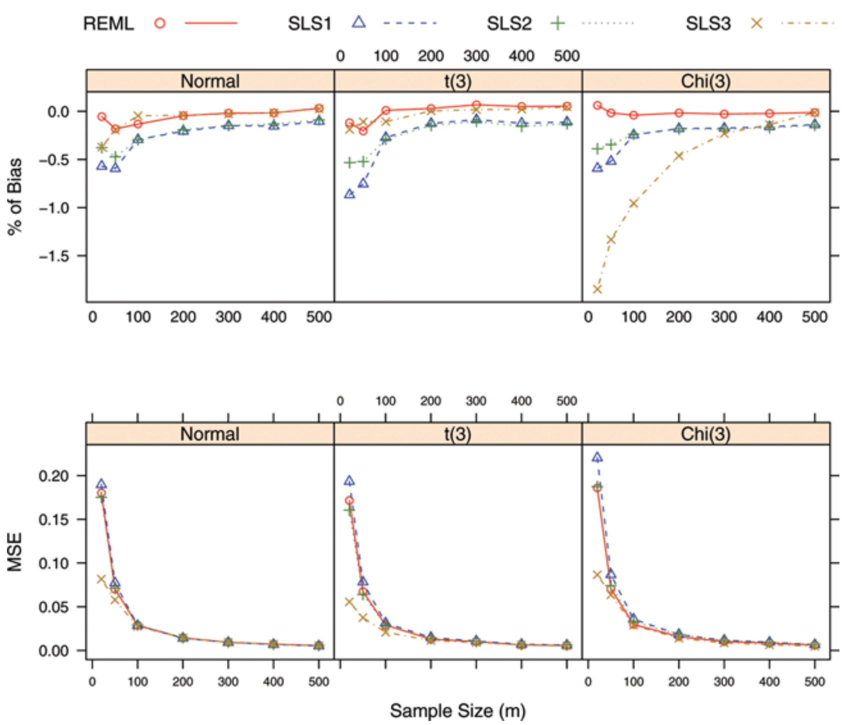

Figure 1. Bias and MSE of β_1 from REML and SLS estimates based on a RI model with Gaussian and non-Gaussian distributed random effect and residual errors. (color figure available online.)

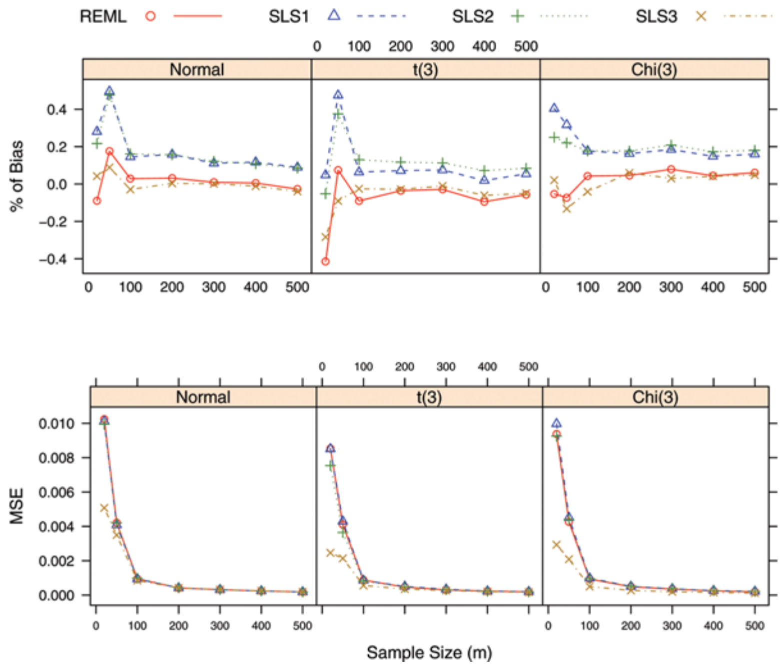

Figure 2. Bias and MSE of β_2 from REML and SLS estimates based on a RI model with Gaussian and non-Gaussian distributed random effect and residual errors. (color figure available online.)

In the second simulation study, we compare the estimates of REML with SLS when outliers exist. We generated 100 subjects with 8 measurements per subject and randomly contaminated one measurement using $100y_{ij}$ within some subjects. The proportions of contaminated subjects were chosen as 0%, 5%, 10%, 15%, 20%, 25%, 30%, and 35%. Table 2 reports the Monte Carlo mean estimates and MSE in simulation study one. For the sake of saving space, we only present the simulation results based on the RI model with 0%, 5%, 15%, and 30%, since similar pattern of results are observed. The influence of the outliers is clearly unbounded for REML estimates because the estimation bias and MSE increase as the percentage of data contamination increases. The magnitude of increase is especially dramatic for the random effect and residual error variances. The same phenomenon is observed in SLS1 estimates. This is not surprising because no downweight is applied in SLSE by using identity weight matrix, and the marginal second moments enlarge the affect of outliers. SLS2 is relatively more robust than SLS1 and REML with a smaller MSE, especially for moderate percentages of outliers. In contrast, SLS3 is clearly bounded and provides consistent mean and MSE estimates regardless of the percentage of data contamination.

4. Application

The proposed estimator is applied to the longitudinal data on cholesterol levels collected as part of the famed Framingham heart study. In the study, 2,634 participants' cholesterol level was measured every 2 years over 10-year period. The

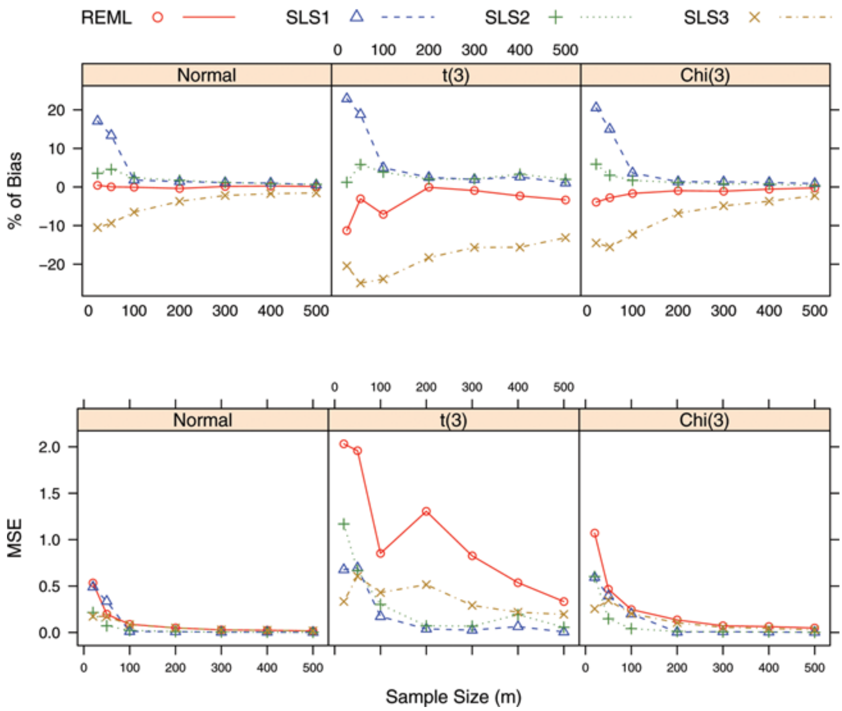

Figure 3. Bias and MSE of θ_{11} from REML and SLS estimates based on a RI model with Gaussian and non-Gaussian distributed random effect and residual errors. (color figure available online.)

objective is to study change in cholesterol over time and examine the association with age at baseline and gender. This dataset is widely used in the linear mixed model literature, partly because many studies conclude that the distribution of subject-specific intercept is non Gaussian; see, e.g., Zhang and Davidian (2001) and Lin and Lee (2008). For illustration, we select a sample of 133 participants (60 men and 73 women) whose cholesterol measurements as well as covariates of interest are completely observed at the duration of follow-up time. In general, the following linear mixed effect model is well accepted to fit the data:

$$y_{ij} = \beta_0 + \beta_1 \,\mathrm{Sex}_i + \beta_2 \,\mathrm{Age}_i + \beta_3 \mathrm{t}_{ij} + b_{0i} + b_{1i} \mathrm{t}_{ij} + \epsilon_{ij}, \quad i = 1, \ldots, 133, \quad j = 1, \ldots, 6,$$

where y_{ij} is the cholesterol level for the ith subject at the jth time point, and y_{ij} was divided by 100 for numerical calculation stability; t_{ij} (in years) was taken as (time – 5)/10 measured from the baseline; Sex_i is a gender indicator (0 = female, 1 = male); and Age_i is age at baseline. $(b_{0i}, b_{1i})'$ is assumed to be normally distributed with mean zero and covariance $D = (\theta_{11}, \theta_{12}, \theta_{22})'$, and ϵ_{ij} is assumed to be normally distributed with mean zero and variance σ^2.

Table 3 includes the estimates and the corresponding 95% confidence interval. For fixed effects, SLS estimates are highly agree with ML, but with slightly tighter confidence intervals. Regarding the random effects and the residual errors, the estimates are quite different between these two methods. This finding is not surprising because the estimates of variance components are usually more difficult

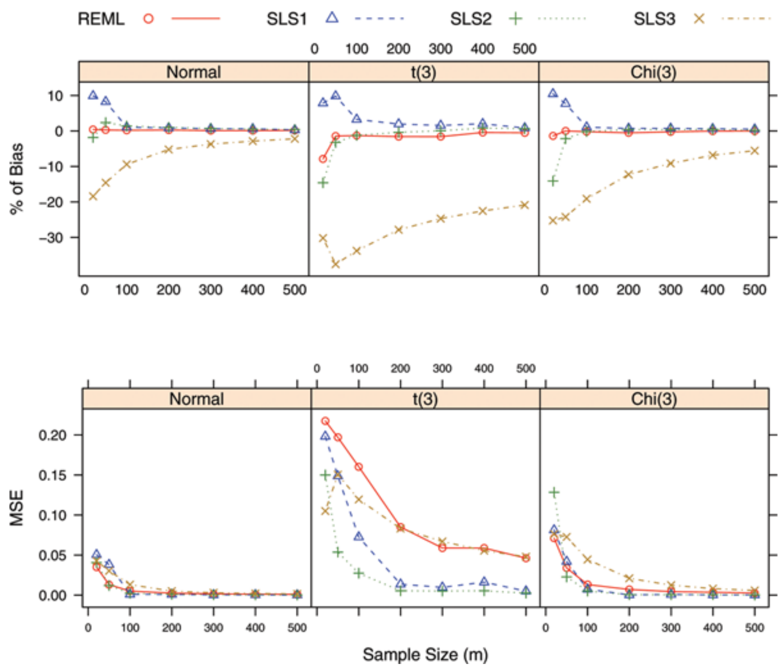

Figure 4. Bias and MSE of σ^2 from REML and SLS estimates based on a RI model with Gaussian and non-Gaussian distributed random effect and residual errors. (color figure available online.)

to estimate and known to have fairly large variabilities. However, the confidence intervals from SLSE are much smaller, which may due to the non-normality distributed random effects. Thus, SLSE provides more precise estimates than ML in this example.

5. Discussion

In statistical literature, the most popular estimation approach for linear mixed effects models is the likelihood method; however, it relies on the normality assumption of the variance components. This article proposes a semiparametric estimation approach, which does not require any distributional assumptions on variance components. The consistency and asymptotic properties of the proposed estimator are derived under fairly mild regularity assumptions. The superiority of the proposed estimator under non normal distributed variance components over maximum likelihood estimator was demonstrated though simulation studies. In addition, we investigated its finite sample properties with diffident choices of the weighting matrices. Although in theory SLS3 should be most efficient, our Monte Carlo simulation results reveal that it is severally biased and usually dominated by SLS1 and SLS2 based on MSE criteria for the estimation of variance components. Based on the extensive simulation studies, we suggest using the diagonal of the optimal weight matrix (SLS2) in practice. This choice not only provides us with the best trade-off between bias and efficiency, but also eases the computation complexity. Furthermore, we demonstrate the robustness property of the proposed

Table 2
Simulation results for different percentage contaminations of a single response in a RI model at $m = 100$ and $n = 8$

		RMEL		SLS1		SLS2		SLS3	
	%	Mean	MSE	Mean	MSE	Mean	MSE	Mean	MSE
$\beta_1 = 8$	0	8.0	0.0	8.0	0.0	8.0	0.0	7.9990	0.02
	5	8.5	1.4	5.5	68	8.1	0.1	7.9897	0.02
	15	9.8	7.8	5.5	377	9.4	3.9	7.9985	0.01
	30	11.9	27.7	1.8	958	10.4	9.7	7.9945	0.02
$\beta_2 = 2$	0	2.0	0.0	2.0	0.0	2.0	0.0	1.9992	0.00
	5	2.1	0.1	2.6	4	2.0	0.0	1.9999	0.00
	15	2.3	0.4	3.2	26	2.0	0.2	1.9997	0.00
	30	2.6	1.2	4.2	66	2.4	0.5	1.9994	0.00
$\theta_{11} = 1.96$	0	1.97	0	1.97	0	2.0	0	1.8473	0.05
	5	2.27	11	35	3396	2.0	0	1.8427	0.05
	15	6.44	260	56	14374	3.4	9	1.8547	0.04
	30	19	6165	125	64188	9.3	140	1.8770	0.04
$\sigma^2 = 1$	0	1	0.0028	1	0.0009	1.0	0.00	0.8984	0.01
	5	253	552545	162	135993	1.0	0.01	0.9038	0.01
	15	1087	3804815	460	849707	85	18598	0.9774	0.12
	30	3061	17463520	1426	5416065	248	101369	1.0763	0.54

%: Percentage of Contaminations

Table 3
SLS and ML estimation of Framingham cholesterol data

	SLS		ML	
Parameter	Estimate	95% Confidence interval	Estimate	95% Confidence interval
β_0	1.5380	(1.3028, 1.7732)	1.5740	(1.2343, 1.9137)
β_1	−0.0369	(−0.1178, 0.0440)	−0.0338	(−0.1564, 0.0889)
β_2	0.0193	(0.0138, 0.0248)	0.0186	(0.0107, 0.0265)
β_3	0.2745	(0.2341, 0.3149)	0.2787	(0.2248, 0.3326)
θ_{11}	0.1033	(0.0731, 0.1335)	0.1259	(0.0934, 0.1584)
θ_{12}	0.0077	(0.0000, 0.0236)	0.0218	(0.0005, 0.0430)
θ_{22}	0.0418	(0.0208, 0.0628)	0.0390	(0.0136, 0.0644)
σ^2	0.0329	(0.0280, 0.0378)	0.0432	(0.0380, 0.0484)

estimator against outliers theoretically and by simulation studies. Although we assume the data are independently and identically distributed in the article, it is a straightforward extension to derive the asymptotic properties of the proposed estimator for independent but not identically distributed data based on the central limit theorem of Lindeberg-Feller, instead of Lindeberg-Lévy. Some future research may be done to explore the robustness property of the proposed estimator by studying its breakdown points and compare it with some popular robust estimation methods in the literature. Some future research can also be done to correct the finite sample bias of SLS3 for the estimation of variance components.

Appendix: Proofs

Proof of Theorem 2.1. First, for any $1 \le i \le m$, by Assumptions 1–2 and Cauchy-Schwartz inequality, we have

$$E\left[\|W_i\|0 \sup_{\Omega} \sum_j (y_{ij} - x'_{ij}\beta)^2\right] \le 2\sum_j E\|W_i\|y_{ij}^2 + 2\sum_j E\|W_i\|\|x_{ij}\|^2 \sup_{\Omega} \|\beta\|^2 < \infty,$$

and

$$\begin{aligned} E&\left[\|W_i\| \sup_{\Gamma} \sum_j \sum_k (y_{ij}y_{ik} - (x'_{ij}\beta x'_{ik}\beta + z'_{ij}Dz_{ik} + \delta_{jk}\sigma^2))^2\right] \\ &\le 2\sum_j\sum_k E\|W_i\|y_{ij}^2y_{ik}^2 + 6\sum_j\sum_k E\|W_i\|\|x_{ij}\|^2\|x_{ik}\|^2 \sup_{\Omega}\|\beta\|^2 \\ &\quad + 6\sum_j\sum_k E\|W_i\|\|z_{ij}\|^2\|z_{ik}\|^2 \sup_{\Theta}\|D\|^2 + 6n\sup_{\Sigma}\sigma^4 E\|W_i\| < \infty, \end{aligned}$$

which imply $E\sup_{\Gamma}\rho'_i(\Upsilon)W_i\rho_i(\Upsilon) \le E\|W_i\|\sup_{\Gamma}\|\rho'_i(\Upsilon)\|^2 < \infty$. Then, it follows from the uniform law of large numbers (ULLN; Jennrich, 1969, Theorem 2), that $\frac{1}{m}Q_m(\Upsilon)$ converges almost surely to $Q(\Upsilon) = E\rho'_i(\Upsilon)W_i\rho_i(\Upsilon)$ uniformly for all Υ in Γ. Furthermore, we have

$$\begin{aligned} Q(\Upsilon) &= Q(\Upsilon_0) + 2E\rho'_i(\Upsilon_0)W_i(\rho_i(\Upsilon) - \rho_i(\Upsilon_0)) + E(\rho_i(\Upsilon) - \rho_i(\Upsilon_0))'W_i(\rho_i(\Upsilon) - \rho_i(\Upsilon_0)) \\ &= Q(\Upsilon_0) + E\left[(\rho_i(\Upsilon) - \rho_i(\Upsilon_0))'W_i(\rho_i(\Upsilon) - \rho_i(\Upsilon_0))\right] \end{aligned}$$

because $\rho_i(\Upsilon) - \rho_i(\Upsilon_0)$ does not depend on Y_i and hence

$$E\left[\rho'_i(\Upsilon_0)W_i(\rho_i(\Upsilon) - \rho_i(\Upsilon_0))\right] = E\left[E(\rho'_i(\Upsilon_0) \mid X_i, Z_i)W_i(\rho_i(\Upsilon) - \rho_i(\Upsilon_0))\right] = 0.$$

Therefore, by Assumption 3, $Q(\Upsilon) \ge Q(\Upsilon_0)$ and the equality holds if and only if $\Upsilon = \Upsilon_0$. Thus, all conditions of Lemma 3 in Amemiya (1973) are satisfied, so we have $\hat{\Upsilon}_m \overset{a.s.}{\to} \Upsilon_0$, as $m \to \infty$.

Proof of Theorem 2.2. The first derivative $\partial Q_m(\Upsilon)/\partial\Upsilon$ exists and has the first-order Taylor expansion in Γ. Since $\partial Q_m(\hat{\Upsilon}_m)/\partial\Upsilon = 0$ and $\hat{\Upsilon}_m \overset{a.s}{\to} \Upsilon_0$, for sufficiently large m we have

$$\frac{\partial Q_m(\widehat{\Upsilon}_m)}{\partial\Upsilon} = \frac{\partial Q_m(\Upsilon_0)}{\partial\Upsilon} + \frac{\partial^2 Q_m(\tilde{\Upsilon}_m)}{\partial\Upsilon\partial\Upsilon'}(\widehat{\Upsilon}_m - \Upsilon_0) = 0, \tag{5.1}$$

where $\|\widetilde{\Upsilon}_m - \Upsilon_0\| \le \|\widehat{\Upsilon}_m - \Upsilon_0\|$. The first derivative of $Q_m(\Upsilon)$ in (5.1) is given in (2.7) with

$$\frac{\partial\mu_{ij}(\Upsilon)}{\partial\Upsilon} = (x_{ij}, 0, 0)', \quad \frac{\partial\eta_{ijk}(\Upsilon)}{\partial\Upsilon} = \left((x_{ij}x'_{ik} + x_{ik}x'_{ij})\beta, \frac{\partial + \operatorname{vec}(D)}{\partial\theta}\operatorname{vec}(z_{ij}z'_{ik}), \delta_{jk}\right)'.$$

Since $\frac{\partial\rho_i'(\Upsilon)}{\partial\Upsilon}W_i\rho_i(\Upsilon)$ are *i.i.d.*, it follows the Central Limit Theorem, as $m \to \infty$,

$$\frac{1}{\sqrt{m}}\frac{\partial Q_m(\Upsilon_0)}{\partial\Upsilon} \xrightarrow{L} N(0, 4C), \tag{5.2}$$

where C is as in (2.6). The second derivative of $Q_m(\Upsilon)$ in (5.1) is given in (2.8) with

$$\frac{\partial^2\mu_{ij}(\Upsilon)}{\partial\Upsilon\partial\Upsilon'} = 0, \quad \frac{\partial^2\eta_{ijk}(\Upsilon)}{\partial\Upsilon\partial\Upsilon'} = \begin{pmatrix} x_{ij}x_{ik}' + x_{ik}x_{ij}' & 0 \\ 0 & 0 \end{pmatrix}.$$

By Assumptions 1–2 and Cauchy-Schwartz inequality,

$$\begin{aligned} E\sup_{\Gamma}\|\frac{\partial\rho_i'(\Upsilon)}{\partial\Upsilon}W_i\frac{\partial\rho_i(\Upsilon)}{\partial\Upsilon'}\| &\le E\|W_i\|\sup_{\Gamma}\|\frac{\partial\rho_i'(\Upsilon)}{\partial\Upsilon}\|^2 \\ &\le \sum_j E\|W_i\|\|x_{ij}\|^2 + 2\sum_j\sum_k E\|W_i\|\|x_{ij}\|^2\|x_{ik}\|^2\sup_{\Omega}\|\beta\|^2 \\ &\quad + \sum_j\sum_k E\|W_i\|\sup_{\Theta}\left\|\frac{\partial\mathrm{vec}(D)}{\partial\theta}\right\|^2\|z_{ij}\|^2\|z_{ik}\|^2 + nE\|W_i\| \\ &< \infty, \end{aligned}$$

and

$$\begin{aligned} &E\sup_{\Gamma}\left\|(\rho_i'(\Upsilon)W_i\otimes I)\frac{\partial\mathrm{vec}(\partial\rho_i'(\Upsilon)/\partial\Upsilon)}{\partial\Upsilon'}\right\| \\ &\le \sqrt{2m(p+r+1)}E\|W_i\|\sup_{\Gamma}\|\rho_i(\Upsilon)\|\left\|\frac{\partial\mathrm{vec}(\partial\rho_i'(\Upsilon)/\partial\Upsilon)}{\partial\Upsilon'}\right\| \\ &\le \sqrt{2m(p+r+1)}\left(E\|W_i\|\sup_{\Gamma}\|\rho_i(\Upsilon)\|^2\right)^{1/2}\left(E\|W_i\|\sup_{\Gamma}\left\|\frac{\partial\mathrm{vec}(\partial\rho_i'(\Upsilon)/\partial\Upsilon)}{\partial\Upsilon'}\right\|^2\right)^{1/2} \\ &\le \sqrt{2m(p+r+1)}\left(E\|W_i\|\sup_{\Gamma}\|\rho_i(\Upsilon)\|^2\right)^{1/2}\left(2\sum_j\sum_k E\|W_i\|\|x_{ij}\|^2\|x_{ik}\|^2\right)^{1/2} \\ &< \infty. \end{aligned}$$

It follows from the ULLN, that $(1/m)\partial^2 Q_m(\Upsilon)/\partial\Upsilon\partial\Upsilon' \xrightarrow{a.s.} \partial^2 Q(\Upsilon)/\partial\Upsilon\partial\Upsilon'$ uniformly for all Υ in Γ, where $\partial^2 Q(\Upsilon)/\partial\Upsilon\partial\Upsilon' = 2E[\frac{\partial\rho_i'(\Upsilon)}{\partial\Upsilon}W_i\frac{\partial\rho_i(\Upsilon)}{\partial\Upsilon'} + (\rho_i'(\Upsilon)W_i\otimes I)\frac{\partial\mathrm{vec}(\partial\rho_i'(\Upsilon)/\partial\Upsilon)}{\partial\Upsilon'}]$. Thus, it follows Lemma 4 of Amemiya (1973)

$$\frac{1}{m}\frac{\partial^2 Q_m(\tilde{\Upsilon}_m)}{\partial\Upsilon\partial\Upsilon'} \xrightarrow{a.s.} \frac{\partial^2 Q(\Upsilon_0)}{\partial\Upsilon\partial\Upsilon} = 2B,$$

which is due to the fact that

$$\begin{aligned} &E\left[(\rho_i'(\Upsilon_0)W_i\otimes I)\frac{\partial\mathrm{vec}(\partial\rho_i'(\Upsilon_0)/\partial\Upsilon)}{\partial\Upsilon'}\right] \\ &= E\left[(E(\rho_i'(\Upsilon_0)|X_i, Z_i)W_i\otimes I)\frac{\partial\mathrm{vec}(\partial\rho_i'(\Upsilon_0)/\partial\Upsilon)}{\partial\Upsilon'}\right] = 0. \end{aligned}$$

Since B is non singular, for sufficiently large m, we have

$$\sqrt{m}(\widehat{\Upsilon}_m - \Upsilon_0) = -\left(\frac{1}{m}\frac{\partial^2 Q_m(\tilde{\Upsilon}_m)}{\partial\Upsilon\partial\Upsilon'}\right)^{-1}\frac{1}{\sqrt{m}}\frac{\partial Q_m(\Upsilon_0)}{\partial\Upsilon}$$

Therefore, by, Assumption 4 and Slutsky's theorem, we have $\sqrt{m}(\hat{\Upsilon}_m - \Upsilon_0) \xrightarrow{L} N(0, B^{-1}CB^{-1})$.

Proof of Equation (2.14). The IF (2.12) is bounded if and only if $G(v; \widehat{\Upsilon}_m, F)$ is bounded. Write

$$\widehat{U} = \frac{1}{m}\sum_{i=1}^{m}\rho_i\rho_i' = \frac{1}{m}(V_l + \rho_l\rho_l'),$$

where $V_l = \sum_{i\neq l}\rho_i\rho_i'$. Then by Sherman-Morrison-Woodbury formula, we have

$$\widehat{U}^{-1} = m(V_l + \rho_l\rho_l')^{-1} = m\left(V_l^{-1} - \frac{V_l^{-1}\rho_l\rho_l'V_l^{-1}}{1+\rho_l'V_l^{-1}\rho_l}\right)$$

if V_l is non singular, V_l^{-1} and U^{-1} exist. Therefore,

$$U^{-1}\rho_l = m\left(V_l^{-1}\rho_l - \frac{V_l^{-1}\rho_l\rho_l'V_l^{-1}\rho_l}{1+\rho_l'V_l^{-1}\rho_l}\right) = m\left(\frac{V_l^{-1}\rho_l}{1+\rho_l'V_l^{-1}\rho_l}\right),$$

and accordingly,

$$\left\|\frac{\partial\rho_l'(\Upsilon)}{\partial\Upsilon}U_i^{-1}\rho_l\right\|^2 = m^2\left(\frac{\rho_l'V_l^{-1}\frac{\partial\rho_i(\Upsilon)}{\partial\Upsilon}\frac{\partial\rho_i'(\Upsilon)}{\partial\Upsilon}V_l^{-1}\rho_l}{1+\rho_l'V_l^{-1}\rho_l}\frac{1}{1+\rho_l'V_l^{-1}\rho_l}\right) \to 0$$

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

Acknowledgments

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

Abarin, T., Wang, L. (2006). Comparison of GMM with second-order least squares estimator in nonlinear models. *Far East J. Theoret. Statist.* 20:179–196.

Amemiya, T. (1973). Regression analysis when the dependent variable is truncated normal. *Econometrica* 41:997–1016.

Demidenko, E. (2004). *Mixed Models: Theory and Applications*. New York: Wiley.

Nelder, J. A., Mead, R. (1965). A simplex method for function minimization. *Comput. J.* 7:308–313.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. New York: Wiley.

Huber, P. J. (2004). *Robust Statistics*. New York: Wiley.

Jacqmin-Gadda, H., Sibillot, S., Proust, C., Molina, J., Thiebau, R. (2006). Robustness of the linear mixed model to missspecified error distribution. *Computat. Statist. Data Anal.* 51:5142–5154.

Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *Ann. Mathemat. Statist.* 40:633–643.

Laird, N. M., Wair, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* 38:963–974.

Lin, T. I., Lee, J. C. (2008). Estimation and prediction in linear mixed models with Skew-normal random effects for longitudinal data. *Statist. Med.* 27:1490–1507.

Pinheiro, J. C., Liu, C., Wu, Y. N. (2001). Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate *t* distribution. *J. Computat. Graph. Statist.* 10:249–276.

Verbeke, G., Lesaffre, E. (1997). The effect of misspecifying the random effects distribution in linear mixed models for longitudinal data. *Computat. Statist. Data Anal.* 23:541–556.

Wang, L. (2007). A unified approach to estimation of nonlinear mixed effects and berkson measurement error models. *Can. J. Statist.* 35:233–248.

Wang, L., Leblanc, A. (2008). Second-order nonlinear least squares estimation. *Ann. Instit. Statist. Math.* 60:883–900.

Zhang, D., Davidian, M. (2001). Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics* 57:795–802.