The Annals of Statistics
1999, Vol. 27, No. 6, 1974–2007

CONDITIONAL INFERENCE ABOUT GENERALIZED LINEAR MIXED MODELS[1]

BY JIMING JIANG

Case Western Reserve University

We propose a method of inference for generalized linear mixed models (GLMM) that in many ways resembles the method of least squares. We also show that adequate inference about GLMM can be made based on the conditional likelihood on a subset of the random effects. One of the important features of our methods is that they rely on weak distributional assumptions about the random effects. The methods proposed are also computationally feasible. Asymptotic behavior of the estimates is investigated. In particular, consistency is proved under reasonable conditions.

1. Introduction. Inference about generalized linear mixed models (GLMM) has received much attention. These models take into account the fact that in many practical problems responses are both discrete and correlated, and therefore are useful in statistical application [e.g., McCullagh and Nelder (1989), Section 14.5, Breslow and Clayton (1993), Lee and Nelder (1996) and Malec, Sedransk, Moriarity and LeClere (1997)]. Several methods of inference about GLMM have been proposed, which will be summarized below.

In this paper, we shall consider these models more generally and propose a method of inference about these models which in many ways resembles the method of least squares (LS) in linear models. An important feature of our method is that it relies on weak distributional assumptions about the random effects. In particular, to apply the method one does not have to assume that the random effects are normally distributed. In practice, one is almost never sure about normality. In fact, in many problems little is known about the distribution of the random effects. Therefore, it is of practical interest to develop methods that do not require strong distributional assumptions.

It is interesting to note a difference between linear and nonlinear models. In the linear case, assuming normality brings technical convenience, because one can then write out the likelihood function in a closed form. This advantage disappears in GLMM. To see this, consider the following example.

EXAMPLE 1.1. Suppose that given the random effects a_i, $1 \le i \le m_1$ and b_j, $1 \le j \le m_2$, binary responses y_{ij}'s are independent with

$$\text{logit}\big(P\big(y_{ij} = 1|a, b\big)\big) = \mu + a_i + b_j. \tag{1.1}$$

Received June 1997; revised August 1999.

[1] Supported in part by NSA Grant MDA904-98-1-0038.

AMS 1991 *subject classification.* Primary 62J12; secondary 62F12.

Key words and phrases. Asymptotic properties of estimates, maximum conditional likelihood, penalized generalized WLS, semiparametric inference.

Assume that the a_i's and b_j's are independent with $a_i \sim N(0, \sigma^2)$, $b_j \sim N(0, \tau^2)$. The log-likelihood for estimating the parameters μ, σ^2 and τ^2 has the form

$$\begin{aligned}
&\text{constant} - \frac{m_1}{2}\log\sigma^2 - \frac{m_2}{2}\log\tau^2 + \mu y_{..}\\
&\quad + \log\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left\{\prod_{i=1}^{m_1}\prod_{j=1}^{m_2}\bigl(1+\exp(\mu+a_i+b_j)\bigr)^{-1}\right\}\\
&\qquad\times\exp\left\{\sum_{i=1}^{m_1}a_i y_{i\cdot} - \frac{1}{2\sigma^2}\sum_{i=1}^{m_1}a_i^2\right.\\
&\qquad\qquad\left. + \sum_{j=1}^{m_2}b_j y_{\cdot j} - \frac{1}{2\tau^2}\sum_{j=1}^{m_2}b_j^2\right\}\prod_{i=1}^{m_1}da_i\prod_{j=1}^{m_2}db_j,
\end{aligned}\tag{1.2}$$

where $y = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}y_{ij}$, $y_{i\cdot} = \sum_{j=1}^{m_2}y_{ij}$ and $y_{\cdot j} = \sum_{i=1}^{m_1}y_{ij}$. If $m_1 = m_2 = 40$, as in the salamander mating problem discussed in McCullagh and Nelder [(1989), Section 14.5], the integral in (1.2) will be 80-dimensional. Obviously, such an expression is much more difficult to evaluate than the log-likelihood under a linear mixed model, so one advantage of assuming normality is much reduced.

To overcome the computational difficulty, several authors have proposed alternatives. These include approximate inference methods [e.g., Schall (1991), Breslow and Clayton (1993), McGilchrist (1994), Kuk (1995), Lin and Breslow (1996) and Lee and Nelder (1996)]; Bayesian inference based on Gibbs sampling [e.g., Zeger and Karim (1991), Karim and Zeger (1992), Malec, Sedransk, Moriarity and LeClere (1997)]; Monte Carlo EM [McCulloch (1994, 1997)] and the method of simulated moments [Jiang (1998)]. However, these approaches have two characteristics. First, strong distributional assumptions about the random effects (e.g., normality or conjugate distributions) are often made. It should be pointed out that Schall [(1991), page 720] has indicated that it is not necessary to assume the random effects to be normal when computing the estimates. However, it is not clear, from a theoretical point of view, how the estimates behave when strong distributional assumptions do not hold. Second, the estimates of the (fixed and random) effects are tied up with those of the variance components. In other words, one has to simultaneously estimate the effects and variance components, or estimate the variance components first, then compute estimates of the effects. In the following we shall propose a method which is different from all the above in exactly these two aspects.

Linear models (LM) have been known as a special case of generalized linear models [GLM, e.g., McCullagh and Nelder (1989), Section 2.2]. However, this is the case only when normality is assumed. On the other hand, the definition of LM does not have to be associated with normality. In other words, GLM by their classic definition do not necessarily include LM as a special case. A similar paradox exists between linear mixed models [LMM, e.g., Searle, Casells and McCulloch (1992)] and GLMM. Therefore, we need to extend the definition of GLMM so that it includes LMM as a special case regardless of the normality assumption.

Suppose that, given a vector $\alpha = (\alpha_k)_{1 \le k \le m}$ of unobservable random variables (the random effects) satisfying

$$E(\alpha) = 0, \tag{1.3}$$

responses $y_1, \dots, y_N$ are independent with conditional expectation

$$E(y_i | \alpha) = b_i'(\eta_i), \tag{1.4}$$

where $b_i(\cdot)$ is a differentiable function. Furthermore, suppose

$$\eta_i = x_i^t \beta + z_i^t \alpha, \tag{1.5}$$

where $\beta = (\beta_j)_{1 \le j \le p}$ is a vector of unknown constants (the fixed effects), and $x_i = (x_{ij})_{1 \le j \le p}$, $z_i = (z_{ik})_{1 \le k \le m}$ are known vectors, $1 \le i \le N$. This generalizes the classic definition of GLMM, in which it is assumed that the conditional density

$$f(y_i | \alpha) = \exp\left\{ \frac{y_i \eta_i - b_i(\eta_i)}{a_i(\phi)} + c_i(y_i, \phi) \right\}, \tag{1.6}$$

$i = 1, \dots, N$, where $b_i(\cdot)$'s and $c_i(\cdot, \cdot)$'s are specific functions corresponding to the type(s) of the exponential family, ϕ is a dispersion parameter, and $a_i(\cdot)$'s are functions of weights. Let $\eta = (\eta_i)_{1 \le i \le N}$, $X = (x_{ij})_{1 \le i \le N,\, 1 \le j \le p}$ and $Z = (z_{ik})_{1 \le i \le N,\, 1 \le k \le m}$. We assume wlog that rank$(X) = p$ and no column of Z is 0.

In LM, which correspond to (1.4) and (1.5) with $b_i(\eta_i) = \eta_i^2/2$ and $m = 0$ (i.e., there are no random effects), a well-known method is weighted least squares (WLS) which defines the estimate of β as the minimizer of

$$\sum_{i=1}^{N} w_i (y_i - \eta_i)^2, \tag{1.7}$$

where w_i, $1 \le i \le N$ are weights, or equivalently, the maximizer of

$$\sum_{i=1}^{N} w_i \left(y_i \eta_i - \frac{\eta_i^2}{2} \right). \tag{1.8}$$

A straight generalization of this method to the case of GLMM would suggest the maximizer of the following function as the estimates of β and α:

$$\sum_{i=1}^{N} w_i (y_i \eta_i - b_i(\eta_i)). \tag{1.9}$$

However, conditionally, the individual fixed and random effects may not be identifiable. For example, the rhs of (1.1) $= (\mu + c + d) + (a_i - c) + (b_j - d)$, $1 \le i \le m_1$, $1 \le j \le m_2$ for any c and d. In LM there are two remedies when the identifiability problem arises, namely, reparametrization and constraints. We shall, for now, focus on the latter. A set of linear constraints on α may be expressed as $P\alpha = 0$ for some matrix P. By Lagrange's method of multipliers, maximizing (1.9) subject to $P\alpha = 0$ is equivalent to maximizing

$$\sum_{i=1}^{N} w_i(y_i\eta_i - b_i(\eta_i)) - \lambda|P\alpha|^2 \tag{1.10}$$

without constraint, where λ is an additional variable. On the other hand, for fixed λ the last term in (1.10) may be regarded as a penalizer. The only thing that needs to be specified is the matrix P. For any matrix M and vector space V, let $\mathscr{B}(V) = \{B\text{: } B$ is a matrix whose columns constitute a base for $V\}$; $\mathscr{N}(M) =$ the null-space of $M = \{v\text{: } Mv = 0\}$; $P_M = M(M^tM)^-M^t$, and $P_{M^\perp} = I - P_M$. Let $A \in \mathscr{B}(\mathscr{N}(P_{X^\perp}Z))$. We define the penalized generalized WLS (PGWLS) estimate of $\gamma = (\beta, \alpha)$ as the maximizer of

$$l_P(\gamma) = \sum_{i=1}^{N} w_i(y_i\eta_i - b_i(\eta_i)) - \frac{\lambda}{2}|P_A\alpha|^2, \tag{1.11}$$

where λ is a positive constant. The notation l_P is used because (1.11) may also be viewed as a penalized conditional quasi-log-likelihood. In Section 2 we shall further explain why the penalizer is chosen this way.

Several questions arise immediately. First, the method seems to ignore the information about the distribution of the random effects. Our view is different. Such information is useful only when it is available. For example, in some rare case one might know for sure that the random effects are normal, which is a lot of information. On the other hand, if one has little knowledge about the random effects, there will not be much information loss by using PGWLS. Plus, PGWLS does not completely ignore the information about α. Note that so far the only assumption about α is (1.3), which implies that $E(P_A\alpha_0) = 0$, where α_0 is the vector of true realizations of the random effects. This means that the constraints $P_A\alpha = 0$ are, on average, satisfied by α_0. In fact, it will be seen that quite often one has $|P_A\alpha_0| \to_P 0$ as sample size increases. Therefore, these constraints are also satisfied asymptotically. PGWLS also uses the fact that since the first moments of the random effects are finite, the α's should be relatively concentrated around their means, and therefore cannot be too large in absolute values, or they will be penalized.

A related observation is that PGWLS seems to treat the random effects as fixed. If so, does one always have sufficient information, in large samples, about all the random effects? The answer is not necessarily. But if there is sufficient information about all the random effects, the latter may, in some sense, be treated as fixed. For example, in order to consistently estimate μ, σ^2 and τ^2 in Example 1.1, it is necessary that $m_1, m_2 \to \infty$. In such a case there is sufficient information about all the random effects, because each one of them appears a large number of times. However, there is a different case.

EXAMPLE 1.2. Suppose that given the random effects a_i, b_{ij}, $1 \le i \le m_1$, $1 \le j \le n$, binary responses y_{ijk}'s are independent with

$$\operatorname{logit}\big(P\big(y_{ijk} = 1|a, b\big)\big) = \mu + a_i + b_{ij}, \tag{1.12}$$

$k = 1, \dots, r$. Suppose the a_i's and b_{ij}'s are independent with $\sigma_a^2 = \mathrm{var}(a_i)$ and $\sigma_b^2 = \mathrm{var}(b_{ij})$, and that m_1, $n \to \infty$ but r remains fixed. Then there is sufficient information about the a_i's but not the b_{ij}'s. Nevertheless, in such a case one should be able to estimate μ, σ_a^2 and σ_b^2 consistently.

The question now is what to do in situations like Example 1.2? One idea is to "give up" the individual random effects that cannot be estimated with adequacy, no matter what. Traditionally, this is done by integrating out all the random effects. However, this requires knowledge about the distributions of all the random effects. Furthermore, it is often possible to estimate, with adequacy, a subset of the random effects, such as the a_i's in Example 1.2. In fact, in this example, one only has to specify the distribution of the b_{ij}'s, because they are the ones to be integrated out. Thus, a natural idea is to divide the random effects into two groups: those that can be estimated with adequacy and those that cannot. The integration will be carried out, but only with respect to the second group, leaving the first group to be estimated individually. This will only require distributional knowledge about a subset of the random effects. To illustrate this method, let us consider a special case, and the more general setting will be similar. Suppose

$$\eta = X\beta + Z\alpha + U\zeta, \tag{1.13}$$

where $\alpha = (\alpha_k)_{1 \le k \le l}$ is independent of $\zeta = (\zeta_k)_{1 \le j \le M}$. [Note that this corresponds to (1.5) with α replaced by (α, ζ).] Furthermore, suppose that U is standard in the sense that it consists of 0's and 1's and there is exactly one 1 in each row and at least one 1 in each column and that $\zeta_1, \dots, \zeta_M$ are independent $\sim \psi(\cdot/\tau)/\tau$, where $\psi(\cdot)$ is a known density function and $\tau > 0$ is an unknown scale parameter, and

$$f(y_i|\alpha, \zeta) = f(y_i|\eta_i), \qquad 1 \le i \le N, \tag{1.14}$$

where $f(\xi_2|\xi_1)$ denotes the conditional density of ξ_2 given ξ_1. Let u_i^t be the ith row of U and $e_{M,j}$ the M-dimensional vector whose jth component is 1 and other components are 0. Let $S_j = \{1 \le i \le N\colon u_i = e_{M,j}\}$, and $y^{(j)} = (y_i)_{i \in S_j}$, $1 \le j \le M$. Then, it is easy to show that

$$f(y|\alpha) = \prod_{j=1}^{M} f(y^{(j)}|\alpha), \tag{1.15}$$

where

$$f(y^{(j)}|\alpha) = E\left(\prod_{i \in S_j} f\big(y_i|x_i^t\beta + z_i^t\alpha + \tau\xi\big)\right), \tag{1.16}$$

and the expectation in (1.16) is taken with respect to $\xi \sim \psi(\cdot)$. Intuitively, ζ is a subset of the random effects about which it is impossible to make adequate inference. It often corresponds to the random effect factor of highest level of interaction or hierarchy (nesting), for example, the b_{ij}'s in Example 1.2. In Section 2 we shall consider inferences about $\tilde{\beta}$ and $\tilde{\alpha}$, which are reparametrization of β and α such that $X\beta + Z\alpha = \tilde{X}\tilde{\beta} + \tilde{Z}\tilde{\alpha}$ for some known matrices $\tilde{X}$ and $\tilde{Z}$. Since (1.15) is the likelihood function conditional on a subset of the random effects, these estimates will be referred as maximum conditional likelihood (MCL) estimates.

Because of the computational difficulty associated with GLMM, mentioned earlier, it is natural to ask whether PGWLS and MCL are computationally feasible. Although in practice the number of fixed effects in a GLMM is often fairly small, the number of random effects can be quite large, which means that one may have to solve a large system of nonlinear equations to obtain the PGWLS (MCL) estimates. A Gauss–Seidel type algorithm is proposed by Jiang (1999) for computing the maximum posterior (MP) estimates of the fixed and random effects. The MP is similar to PGWLS but with $w_i = 1/a_i(\phi)$, $\lambda = 1$ and $|P_A \alpha|^2$ replaced by $\alpha^t D^{-1}\alpha$, where D is the covariance matrix of α, which is assumed normal here. It is shown by Jiang (1999) that the algorithm converges in all typical situations of GLMM (1.6). It seems promising that such types of algorithms may provide effective ways of computing the estimates introduced in this paper.

In GLMM, the variance components associated with the random effects are often of interest. Since our methods are based on conditional inference, estimates of the variance components are not directly obtained [except for the scale parameter τ in (1.16)]. However, since the variance components are closely related to the random effects, adequate inference about the random effects often easily results in that about the variance components. We shall discuss this in Section 4. Note that our method is different from the previous approaches, for inference about GLMM in that our estimates of the (fixed and random) effects do not depend on the estimates of the variance components. This is, again, similar to the LS method. In LM, the LS estimates of the regression coefficients do not depend on the estimate of the variance of the errors, say, σ^2. On the other hand, the estimate of σ^2 is based on the residuals which are analogous to the estimates of the random effects.

Finally, there is, of course, a question about the behavior of the PGWLS and MCL estimates. The main goal of this paper is to study the behavior of these estimates from an asymptotic point of view. In particular, we shall prove the consistency of these estimates under reasonable conditions. The main theorems are stated in Section 2 and further illustrated by examples in Section 3. In Section 4, we make a number of observations regarding, in addition to estimation of the variance components, choice of the penalizer and the connection between PGWLS and the penalized-likelihood method based on Laplace approximation. The proofs of the theorems are given in Section 5.

Notation. In addition to those that have been introduced, we have the following.

Let $B = (b_{ij})_{1 \le i \le k,\, 1 \le j \le l}$ be a matrix, $v = (v_i)_{1 \le i \le k}$ a vector and V a vector space. Define $|v| = (v^t v)^{1/2}$, $\|v\| = \max_{1 \le i \le k}|v_i|$; $\lambda_{\min}(B)$ $(\lambda_{\max}(B))$ = the smallest (largest) eigenvalue of B, $\|B\| = \lambda_{\max}^{1/2}(B^t B)$, $\|B\|_R = (\mathrm{tr}(B^t B))^{1/2}$, $\|B\|_\infty = \max_{1 \le i \le k} \sum_{j=1}^{l} |b_{ij}|$; $BV = \{Bv\colon\ v \in V\}$, $\lambda_{\min}(B)|_V = \inf_{v \in V \setminus \{0\}} (v^t Bv / v^t v)$.

Let $v_{(1)}, \dots, v_{(n)}$ be vectors and $B_1, \dots, B_n$ be matrices. We use the symbol $(v_{(1)}, \dots, v_{(n)})$ for the vector $(v_{(1)}^t \cdots v_{(n)}^t)^t$. To avoid confusion, a row vector will be written as $(v_1 \cdots v_n)$, that is, without commas in between. Let $\mathrm{diag}(B_1, \dots, B_n)$ be the block-diagonal matrix with B_i being its ith diagonal block.

Let X_u be the uth column of X, $1 \le u \le p$, Z_k the kth column of Z, $1 \le k \le m$ and $H = (X\ Z)^t(X\ Z)$. Let β_0 and τ_0 be the true β and τ, respectively, and $\eta_0 = X\beta_0 + Z\alpha_0$.

2. Asymptotic properties of the estimates. In two ways, the asymptotic theory regarding random effects is different from that about fixed parameters. First, the individual random effects are typically not identifiable [see the discussion in Section 1]. Therefore, any asymptotic theory must take care, in particular, of the identifiability problem. Second, the total number of random effects m often increases with the sample size N. Asymptotic properties of estimates of fixed parameters when the number of parameters increases with the sample size have been studied by Portnoy in a series of papers. There are several major differences between our results and those of Portnoy. Besides the fact that the effects we are interested in may be random and that we are typically dealing with correlated responses, the design matrix Z often has the ANOVA structure, which is more general than that considered by Portnoy [e.g., (1984), Section 5].

Also, we note that for the asymptotic results in this paper to hold as $N \to \infty$, m may be, wlog, considered as a function of N. This is because such results hold iff they hold for each sequence with N increasing strictly monotonically, in which case m may readily be regarded as a function of N. Similarly, the number p, the matrices X, Z, A, etc., may be regarded as dependent on N.

To explore the asymptotic behavior of the estimates of the fixed and random effects, one has to distinguish two different cases: the case where there is enough information about the random effects and the case where there is not. The first case is characterized by $m/N \to 0$, while the second by m/N not $\to 0$.

2.1. *The case $m/N \to 0$.* In this case we consider the asymptotic behavior of the PGWLS estimates. A basic technique here is penalization. As discussed in Section 1, the purpose of the penalization is to make the individual effects conditionally identifiable, which may be different from that of many traditional uses of the penalizers. More specifically, we first explain why P_A, defined above (1.11), is chosen this way. The main result states that, under suitable conditions, the PGWLS estimates of the fixed and random effects are consistent, where the convergence of the estimates of the random effects is in an overall sense.

Consider the expression (1.11). The reason that one needs a penalizer here is because the first term, $l_C(\gamma) = \sum_{i=1}^N w_i(y_i\eta_i - b_i(\eta_i))$, depends on $\gamma = (\beta, \alpha)$ only through η. However, γ cannot be identified by η, so there may be many vectors γ for which $\eta = X\beta + Z\alpha$ is the same. The idea is therefore to consider a restricted space $S = \{\gamma\colon P_A \alpha = 0\}$, such that within this subspace, γ is uniquely determined by η.

Define the map $T\colon \gamma = (\beta, \alpha) \to \tilde{\gamma} = (\tilde{\beta}, \tilde{\alpha})$ as follows: $\tilde{\alpha} = P_{A^\perp}\alpha$, $\tilde{\beta} = \beta + (X^tX)^{-1}X^tZP_A\alpha$. Obviously, T does not depend on the choice of A. Since $X\tilde{\beta} + Z\tilde{\alpha} = X\beta + Z\alpha - P_{X^\perp}ZP_A\alpha = X\beta + Z\alpha$, we have $l_C(\gamma) = l_C(\tilde{\gamma})$. Let $G_A = \left(\begin{smallmatrix} X & Z \\ 0 & A^t \end{smallmatrix}\right)$. The proofs of the following results will be given in Section 5.

LEMMA 2.1. $\operatorname{rank}(G_A) = p + m$.

COROLLARY 2.1. *Suppose that $b_i''(\cdot) > 0$, $1 \le i \le N$. Then there can be only one maximizer of l_P.*

LEMMA 2.2. *For any positive numbers b_u, $1 \le u \le p$ and a_k, $1 \le k \le m$,*

$$\lambda_{\min}(W^{-1}HW^{-1})|_{WS} \ge \frac{\lambda_{\min}\left(G_A^tG_A\right)}{\left(\max_{1 \le u \le p} b_u^2\right) \vee \left(\max_{1 \le k \le m} a_k^2\right)} > 0,$$

where $W = diag(b_1, \dots, b_p, a_1, \dots, a_m)$.

THEOREM 2.1. *Let $b_i''(\cdot)$ be continuous,* $\max_{1 \le i \le N}\{w_i^2 E \operatorname{var}(y_i|\alpha_0)\}$ *be bounded and*

$$(2.1) \qquad \frac{1}{N}\left[\left(\max_{1 \le u \le p} |X_u|^2\right)\|(X^tX)^{-1}X^tZ\|^2 + \left(\max_{1 \le k \le m} |Z_k|^2\right)\right]|P_A\alpha_0|^2 \to_P 0.$$

Let $c_N, d_N > 0$ be any sequences such that $\limsup\|\beta_0\|/c_N < 1$ *and* $P(\|\alpha_0\|/d_N < 1) \to 1$, $M_i \ge c_N\sum_{u=1}^p|x_{iu}| + d_N\sum_{k=1}^m|z_{ik}|$, $1 \le i \le N$ *and* $\hat{\gamma} = (\hat{\beta}, \hat{\alpha})$ *be the maximizer of l_P over* $\Gamma(M) = \{\gamma\colon |\eta_i| \le M_i, 1 \le i \le N\}$. *Then*

$$(2.2) \qquad \frac{1}{N}\left(\sum_{u=1}^p |X_u|^2\left(\hat{\beta}_u - \beta_{0u}\right)^2 + \sum_{k=1}^m |Z_k|^2\left(\hat{\alpha}_k - \alpha_{0k}\right)^2\right) \to_P 0,$$

provided that

$$(2.3) \qquad \frac{p + m}{N} = o(\omega^2),$$

where $\omega = \lambda_{\min}(W^{-1}HW^{-1})|_{WS} \min_{1 \le i \le N}\{w_i \inf_{|h| \le M_i} b_i''(h)\}$ *with* $W = \operatorname{diag}(|X_1|, \dots, |X_p|, |Z_1|, \dots, |Z_m|)$.

The following shows that, under further assumptions about the design matrices, $\hat{\beta}$ is a consistent estimate, and the convergence of $\hat{\alpha}$ can be expressed more intuitively.

COROLLARY 2.2. *Let the conditions of Theorem* 2.1 [*including* (2.3)] *hold.*

(i) *Suppose p is fixed, and*

$$(2.4) \qquad \liminf \lambda_{\min}(X^tX)/N > 0,$$

then $\hat{\beta} \to_P \beta_0$.

(ii) *Suppose* $Z = (Z_{(1)} \cdots Z_{(q)})$ *and correspondingly*, $\alpha = (\alpha_1, \dots, \alpha_q)$, *where* $\alpha_u = (\alpha_{uv})_{1 \le v \le m_u}$, *and each* $Z_{(u)}$ *is a standard design matrix in the same sense as for U described below* (1.13), $1 \le u \le q$. *Let* Z_{uv} *be the vth column of* $Z_{(u)}$ *and* $n_{uv} = |Z_{uv}|^2$ = *the number of appearances of the vth component of* α_u. *Then*

$$\left(\sum_{v=1}^{m_u} n_{uv} \right)^{-1} \sum_{v=1}^{m_u} n_{uv} (\hat{\alpha}_{uv} - \alpha_{0uv})^2 \to_P 0, \qquad 1 \le u \le q, \tag{2.5}$$

where $\hat{\alpha}_{uv}$ *and* α_{0uv}, $1 \le v \le m_u$, $1 \le u \le q$ *are the corresponding components of* $\hat{\alpha}$ *and* α_0, *respectively.*

One special case of GLMM is the LMM. Theorem 2.1 and Corollary 2.1 imply the following.

COROLLARY 2.3. *Suppose* $b_i''(\cdot) = 1$; $\max_{1 \le i \le N}\{w_i^2 E \operatorname{var}(y_i|\alpha_0)\}$ *is bounded, and* (2.1) *holds. Then* (2.2) *holds provided* (2.3), *where* $\hat{\gamma}$ *is the unique maximizer of* l_P *and* $\omega = \lambda_{\min}(W^{-1}HW^{-1})|_{WS} \min_{1 \le i \le N} w_i$.

Note that one difference between Theorem 2.1 and Corollary 2.3 is that in the former case $\hat{\gamma}$ is the maximizer of l_P over $\Gamma(M)$, while in the latter case $\hat{\gamma}$ is the global maximizer of l_P. In general, we have the following.

LEMMA 2.3. *Suppose that* $b_i''(\cdot) > 0$, $1 \le i \le N$. *Let* $\hat{\gamma}$ *be as in Theorem* 2.1. *If* $\hat{\gamma} \in R_N^o = \{\gamma: \|\beta\| < c_N, \|\alpha\| < d_N\}$, *then* $\hat{\gamma}$ *is identical to the unique global maximizer of* l_P.

This is because if $\hat{\gamma} \in R_N^o \subset \Gamma(M)$, then $\hat{\gamma}$ is a local maximizer of l_P and hence a root to

$$\frac{\partial l_P}{\partial \gamma} = 0. \tag{2.6}$$

On the other hand, by the proof of Corollary 2.1, it is easy to show that the root to (2.6) is identical to the unique maximizer of l_P.

In the following, we consider a special class of GLMM in which the responses are clustered into groups with each group associated with a single random effect (possibly vector valued). Suppose that given unobservable random vectors $\alpha_1, \dots, \alpha_m$ satisfying $E(\alpha_i) = 0$, the responses y_{ij}, $1 \le i \le m$, $1 \le j \le n_i$ $(n_i \ge 1)$ are independent with $E(y_{ij}|\alpha) = b_{ij}'(\eta_{ij})$, where $b_{ij}(\cdot)$ is differentiable. Furthermore,

$$\eta_{ij} = a + x_{ij}^t \beta + z_i^t \alpha_i, \tag{2.7}$$

where a is an unknown intercept, $\beta = (\beta_k)_{1 \le k \le s}$ (s is fixed) is an unknown vector of regression coefficients, and $x_{ij} = (x_{ijk})_{1 \le k \le s}$ and z_i are known vectors. Such models are useful, for example, in the context of small-area estimation [e.g., Ghosh and Rao (1994)] in which α_i represents a random effect associated with the ith selected area. Here we are interested in the estimation of the fixed effects a, β_k, $1 \le k \le s$, and the "area-specific" random effects $v_i = z_i^t \alpha_i$, $1 \le i \le m$. Therefore, wlog, we may assume that in the above model η_{ij} has the following expression:

$$\eta_{ij} = a + x_{ij}^t \beta + v_i, \tag{2.8}$$

where $v_1, \dots, v_m$ are random variables with $E(v_i) = 0$. Note that in (2.8), $a_i = a + v_i$ may be regarded as a random intercept. It is clear that this is a special case of the GLMM (1.3)–(1.5) with β replaced by (a, β), $\alpha = v = (v_i)_{1 \le i \le m}$, and the design matrices $X = (1_N \ X_1 \ \cdots \ X_s)$, where $N = \sum_{i=1}^m n_i$, $X_k = (X_{ik})_{1 \le i \le m}$, and $X_{ik} = (x_{ijk})_{1 \le j \le n_i}$, and $Z = \operatorname{diag}(1_{n_i}, 1 \le i \le m)$. Furthermore, it is easy to show that $A = 1_m \in \mathscr{B}(\mathscr{N}(P_{X^\perp} Z))$, $S = \{\gamma\colon v.= 0\}$, where $v.= \sum_{i=1}^m v_i$. Thus, (1.11) has a more explicit expression,

$$l_P(\gamma) = \sum_{i=1}^m \sum_{j=1}^{n_i} w_{ij}\big(y_{ij}\eta_{ij} - b_{ij}(\eta_{ij})\big) - \frac{\lambda}{2} m \bar{v}^2, \tag{2.9}$$

where $\bar{v} = v./m$. For such models, we have the following more explicit result. Let $\delta_N = \min_{i,j} w_{ij} \inf_{|h| \le M_{ij}} b_{ij}''(h)$, $\lambda_N = \lambda_{\min}(\sum_{i=1}^m \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)^t)$ with $\bar{x}_i = n_i^{-1} \sum_{j=1}^{n_i} x_{ij}$.

THEOREM 2.2. *Let $b_{ij}''(\cdot)$ be continuous*; $w_{ij}^2 E \operatorname{var}(y_{ij}|v_0)$, $|x_{ij}|$ *be bounded*, $\liminf(\lambda_N/N) > 0$ *and* $\bar{v}_0 \to_P 0$. *Let* $c_n, d_n > 0$ *be such that* $\limsup(|a_0| \vee |\beta_0|)/c_N < 1$ *and* $P(\|v_0\|/d_N < 1) \to 1$, $M_{ij} \ge c_N(1 + |x_{ij}|) + d_N$ *and* $\hat{\gamma} = (\hat{a}, \hat{\beta}, \hat{v})$ *be the maximizer of* l_P *over* $\Gamma(M) = \{\gamma\colon |\eta_{ij}| \le M_{ij},\ all\ i, j\}$. *Then*, $\hat{\beta} \to_P \beta_0$, *and*

$$\frac{1}{N} \sum_{i=1}^m n_i (\hat{a}_i - a_{0i})^2 \to_P 0, \tag{2.10}$$

where $\hat{a}_i = \hat{a} + \hat{v}_i$ *and* $a_{0i} = a_0 + v_{0i}$, *provided that* $m/N = o(\delta_N^2)$. *If the latter is strengthened to* $(\min_{1 \le i \le m} n_i)^{-1} = o(\delta_N^2)$, *then, in addition*, $\hat{a} \to_P a_0$, *and*

$$\frac{1}{N} \sum_{i=1}^m n_i (\hat{v}_i - v_{0i})^2 \to_P 0, \qquad \frac{1}{m} \sum_{i=1}^m (\hat{v}_i - v_{0i})^2 \to_P 0. \tag{2.11}$$

Note. It can be shown, by a simple example, that $\hat{a} \to_P a_0$ and (2.11) may not hold without $\min_{1 \le i \le m} n_i \to \infty$, even if $m/N \to 0$.

2.2. *The case m/N not → 0.* In this case we consider the asymptotic behavior of the MCL estimates. As noted in Section 1, a basic technique here is reparametrization, because, conditionally, the individual effects may not be identifiable. We first introduce the reparametrization, which is a map from (β, α) to $(\tilde{\beta}, \tilde{\alpha})$. The main result states that, under suitable conditions, the MCL estimates of $\tilde{\beta}$, $\tilde{\alpha}$ and τ are consistent with a certain convergence rate.

We consider the model defined by (1.13) with all the assumptions. Furthermore, we assume that there are no random effects nested within ζ. In notation, this means that $z_i = z_{*j}$, $i \in S_j$, $1 \le j \le M$, where $z_{*j} = (z_{*jk})_{1 \le k \le l}$. The following lemma defines the reparametrization.

LEMMA 2.4. *There is a map $\beta \mapsto \tilde{\beta}$, $\gamma \mapsto \tilde{\alpha}$ such that*:

(i) *$X\beta + Z\alpha = \tilde{X}\tilde{\beta} + \tilde{Z}\tilde{\alpha}$, where $(\tilde{X}\ \tilde{Z})$ is a known matrix of full column rank.*

(ii) *$\tilde{z}_i = \tilde{z}_{*j}$, $i \in S_j$ for some known vector $\tilde{z}_{*j}$, where $\tilde{z}_j^t$ is the ith row of $\tilde{Z}$ and S_j is defined above* (1.15).

By Lemma 2.4, we have

$$\eta = W\gamma + U\zeta, \tag{2.12}$$

where $W = (\tilde{X}\ \tilde{Z})$, $\gamma = (\tilde{\beta}, \tilde{\alpha})$. Let $\varphi = (\tilde{\beta}, \tau)$, $\psi = (\tilde{\alpha}, \varphi)$. By (1.14) we have

$$f(y|\psi) = \prod_{j=1}^{M} f(y^{(j)}|\psi), \tag{2.13}$$

and it is easy to show that $f(y^{(j)}|\psi) = g_j(\tilde{z}^t{}_{*j}\tilde{\alpha}, \tilde{\beta}, \tau)$, where

$$g_j(s) = E\left(\prod_{i \in S_j} f\left(y_i | s_1 + \tilde{x}_i^t s_{(2)} + s_{r+2}\,\xi\right)\right) \tag{2.14}$$

with $s_{(2)} = (s_2 \cdots s_{r+1})$ and $r = \dim(\tilde{\beta})$. Note that $r \le p$. Let $n = \dim(\tilde{\alpha})$ (Note that n is the same as t in the proof of Lemma 2.4.) Let $h_j(s) = \log g_j(s)$, $l_C(\psi) = \log f(y|\psi)$ and $l_{C,j}(\psi) = \log f(y^{(j)}|\psi) = h_j(\tilde{z}^t{}_{*j}\tilde{\alpha}, \tilde{\beta}, \tau)$. Then

$$l_C(\psi) = \sum_{j=1}^{M} l_{C,j}(\psi). \tag{2.15}$$

Let $\tilde{Z}_*$ be the matrix whose jth row is $\tilde{z}^t{}_{*j}$, $1 \le j \le M$. Let φ_0 and ψ_0 be the vectors corresponding to the true parameters and realizations of random effects. Define $s_{M,k}^{(l)} = \sum_{j=1}^{M} |\tilde{z}_{*jk}|^l$, $l = 1, 2, \dots, t_{M,k} = \sum_{j=1}^{M} \sum_{l \ne k} |\tilde{z}_{*jk}\tilde{z}_{*jl}|$,

$$H_j(\psi) = \left.\frac{\partial^2 h_j}{\partial s^2}\right|_{s_1 = \tilde{z}^t{}_{*j}\tilde{\alpha},\, s_{(2)} = \tilde{\beta},\, s_{r+2} = \tau}, \tag{2.16}$$

$$A_2 = \sum_{j=1}^{M} \begin{pmatrix} \tilde{z}_{*j} & 0 \\ 0 & I_{r+1} \end{pmatrix} \left(H_j(\psi_0) - E\left(H_j(\psi_0)|\psi_0\right)\right) \begin{pmatrix} \tilde{z}^t{}_{*j} & 0 \\ 0 & I_{r+1} \end{pmatrix}, \tag{2.17}$$

where I_l represents the l-dimensional identity matrix,

$$G = \begin{pmatrix} \tilde{Z}^t{}_*\tilde{Z}_* & 0 \\ 0 & MI_{r+1} \end{pmatrix} = \sum_{j=1}^{M} \begin{pmatrix} \tilde{z}_{*j}\tilde{z}^t{}_{*j} & 0 \\ 0 & I_{r+1} \end{pmatrix}, \tag{2.18}$$

$$\lambda_M(\psi) = \min_{1 \le j \le M} \lambda_{\min}\left(\operatorname{Var}\left(\frac{\partial h_j}{\partial s}\Big|_{(\tilde{z}^t{}_{*j}\tilde{\alpha},\, \tilde{\beta},\, \tau)} \Big| \psi\right)\right), \tag{2.19}$$

and $\lambda_M = \lambda_M(\psi_0)$. Let $\xi_j^{(l)}(\psi) = (\partial^l h_j/\partial s_1^l)(\tilde{z}^t{}_{*j}\tilde{\alpha}, \tilde{\beta}, \tau)$, $l = 1, 2$,

$$V_k^{(1)}(\varepsilon) = \max_{1\le j\le M} E\Big(|\xi_j^{(1)}(\psi_0)|1_{(|\tilde{z}_{*jk}\xi_j^{(1)}(\psi_0)|>1/2\varepsilon)}|\psi_0\Big),$$

$$V_k^{(2)}(\varepsilon) = \max_{1\le j\le M} E\Big(|\xi_j^{(2)}(\psi_0) - E\big(\xi_j^{(2)}(\psi_0)|\psi_0\big)|1_{(\tilde{z}^2{}_{*jk}|\cdots - \cdots|>1/2\varepsilon)}|\psi_0\Big).$$

THEOREM 2.3. *Suppose*:

(i) *the conditional densities* $f(y^{(j)}|\psi)$, $1 \le j \le M$ *are with respect to a common measure* μ *and have common support, and the first and second partial derivatives of* $\int f(y^{(j)}|\psi)\, d\mu$ *with respect to components of* ψ *exist and can be taken under the integral sign.*

(ii) $h_j(s)$, $1 \le j \le M$ *are three times differentiable and there exist* δ, $B > 0$ *such that*

$$\max_{1\le j\le M}\left\{\left(\max_{1\le u\le r+2}\left|\frac{\partial^2 h_j}{\partial s_1\,\partial s_u}\right|\right) \vee \left(\max_{1\le u,v,w\le r+2}\left|\frac{\partial^3 h_j}{\partial s_u\,\partial s_v\,\partial s_w}\right|\right)\right\} \le B$$

for all ψ *such that* $\|\varphi - \varphi_0\| \le \delta$.

(iii) $\tilde{Z}_{*k} \ne 0$, $1 \le k \le n$, *where* $\tilde{Z}_{*k}$ *is the kth column of* $\tilde{Z}_*$, *and the following are bounded*:

$$\|\tilde{Z}_*\|_\infty,\ \max_{1\le k\le n}\left(\frac{s_{M,k}^{(1)}}{s_{M,k}^{(2)}}\right),\ \max_{1\le k\le n}\left(\frac{s_{M,k}^{(2)}}{s_{M,k}^{(4)}}\right),\ \max_{1\le k\le n}\left(\frac{s_{M,k}^{(4)}}{s_{M,k}^{(2)}}\right) \text{ and}$$

$$\max_{1\le j\le M}\left(|\tilde{z}_{*j}|^2 E\left(\frac{\partial h_j}{\partial s_1}\bigg|_{s_0}\right)^2\right) \vee \left(\max_{2\le u\le r+2} E\left(\frac{\partial h_j}{\partial s_u}\bigg|_{s_0}\right)^2\right), \tag{2.20}$$

where $s_0 = (\tilde{z}^t{}_{*j}\tilde{\alpha}_0, \tilde{\beta}_0, \tau_0)$ *and*

(iv) $\lambda_M > 0$, *and there is a sequence* ρ_M *such that* $0 < \rho_M \le \lambda_M \wedge 1$, *and the following* $\to 0$ *in probability*:

$$\lambda_{\max}\big(G^{-1/2}A_2G^{-1/2}\big)/\rho_M,\ \max_{1\le k\le n}\left(\frac{t_{M,k}}{s_{M,k}^{(2)}}\right)\Big/\rho_M,\ \left(\frac{n}{M}\right)\Big/\rho_M^4 \text{ and}$$

$$\max_{l=1,2}\left(\log n/\rho_M^{2l}\min_{1\le k\le n} s_{M,k}^{(6-2l)}\right) \vee \left(n\max_{1\le k\le n} V_k^{(3-l)}(\rho_M^l)/\rho_M^l\right).$$

Then, with probability approaching 1, *there is a sequence* $\hat{\psi}$ *satisfying* $(\partial l_C/\partial\psi)(\hat{\psi}) = 0$ *and* $\|\hat{\psi} - \psi_0\| = o_P(\rho_M)$.

Note. In fact, it is seen from the proof of the theorem that $\|\hat{\varphi} - \varphi_0\| = o_P(\rho_M^2)$.

Consider a special case in which there is only one random effect factor. In such a case, one may integrate out all the random effects, if necessary. The resulting MCL estimates are the maximum likelihood estimates for the fixed parameters. We have the following.

COROLLARY 2.4. *Suppose that in* (1.13) $\alpha = 0$ (*i.e., there are no random effects besides* ζ), *and that*:

(i) *Part* (i) *of Theorem* 2.3 *holds with* ψ *replaced by* φ.

(ii) $h_j(\varphi)$, $1 \le j \le M$ *are three times differentiable and there is* δ, $B > 0$ *such that*

$$\max_{1 \le j \le M} \sup_{\|\varphi - \varphi_0\| \le \delta} |(\textit{any third derivative of } h_j)(\varphi)| \le B.$$

(iii) $\lambda_M = \min_{1 \le j \le M} \lambda_{\min}(\mathrm{Var}((\partial h_j/\partial\varphi)(\varphi_0)|\varphi_0)) > 0$ *and*

$$\frac{1}{M^2(\lambda_M \wedge 1)^2} \sum_{j=1}^{M} E\big(\|H_j(\varphi_0) - EH_j(\varphi_0)\|_R^2\big) \to 0,$$

where $H_j(\varphi) = \partial^2 h_j/\partial\varphi^2$. *Then, with probability approaching* 1, *there is a sequence* $\hat\varphi$ *such that* $(\partial l_C/\partial\varphi)(\hat\varphi) = 0$ *and* $\|\hat\varphi - \varphi_0\| = o_P((\lambda_M \wedge 1)^2)$.

This follows easily from Theorem 2.3 and the note following the theorem. Note that since there is no α, most of the assumptions in Theorem 2.3 [e.g., (iii) and most of (iv)] are not needed.

3. Examples.

First, we use an example to illustrate Theorem 2.1.

EXAMPLE 3.1. Consider the logit random effects model (1.1). We have $X = 1_{m_1} \otimes 1_{m_2}$, $Z_{(1)} = I_{m_1} \otimes 1_{m_2}$, $Z_{(2)} = 1_{m_1} \otimes I_{m_2}$, where $\otimes$ means Kronecker product. Then, $A = \mathrm{diag}(1_{m_1}, 1_{m_2}) \in \mathscr{B}(\mathscr{N}(P_{X^\perp}(Z_{(1)}\ Z_{(2)})))$. Also $W = \sqrt{m_1 m_2}\,\mathrm{diag}(1, (1/\sqrt{m_1})I_{m_1}, (1/\sqrt{m_2})I_{m_2})$. For any $(\mu, a, b) \in S = \{a_\cdot = b_\cdot = 0\}$, let $(h, u, v) = W(\mu, a, b)$. Then,

$$\begin{aligned}(h, u, v)^t W^{-1} H W^{-1} (h, u, v) &\\ = (\mu, a, b)^t H(\mu, a, b) &= \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} (\mu + a_i + b_j)^2 \\ = m_1 m_2 \mu^2 + m_2 \sum_{i=1}^{m_1} a_i^2 &+ m_1 \sum_{j=1}^{m_2} b_j^2 = (h, u, v)^t (h, u, v).\end{aligned}$$

Therefore, $\lambda_{\min}(W^{-1}HW^{-1})|_{WS} = 1$. It is easy to show that (2.1) is satisfied if $m_1 \wedge m_2 \to \infty$.

Suppose $m_1, m_2 \to \infty$ such that $\log m_1/(\log m_2)^2 \to 0$, $\log m_2/(\log m_1)^2 \to 0$. Let $\{c_N\}$, $\{e_N\}$ be any sequences such that $c_N, e_N \to \infty$, $c_N/\log(m_1 \wedge m_2) \to 0$ and $e_N\sqrt{\log(m_1 \vee m_2)}\,/\log(m_1 \wedge m_2) \to 0$. Let $d_N = e_N\sqrt{\log(m_1 \vee m_2)}$; $M_{N,(i,j)} = b_N = c_N + 2d_N$. Then, by Lemma 3.1 in the following, $(\|a_0\| \vee \|b_0\|)/a_N \to 0$. Also, $\omega = \min_{i,j}\{\inf_{|h| \le M_{N,(i,j)}} b''_{i,j}(h)\} = \inf_{|h| \le b_N}\{e^h/(1 + e^h)^2\} \ge (1/4)\exp(-2b_N)$. Thus it is easy to show that (2.3) is satisfied.

It follows from Theorem 2.1 that $\hat\mu \to_P \mu_0$, $(1/m_1)\Sigma_{i=1}^{m_1}(\hat a_i - a_{0i})^2 \to_P 0$, and $(1/m_2)\Sigma_{j=1}^{m_2}(\hat b_j - b_{0j})^2 \to_P 0$.

LEMMA 3.1. *Suppose* $\alpha_k \sim N(0, \sigma_k^2)$, $1 \le k \le m$, *and* $\max_{1 \le k \le m} \sigma_k^2$ *is bounded. Then* $P(\|\alpha\| \le d_N) \to 1$ *provided* $\log m / d_N^2 \to 0$.

We now use an example to illustrate Theorem 2.2.

EXAMPLE 3.2. Suppose y_{ij}, $1 \le i \le m$, $1 \le j \le n_i$ are binary with $\text{logit}(P(y_{ij} = 1|\alpha)) = \eta_{ij} = \beta_0 + \beta_1 x_{ij} + \alpha_i$, where x_{ij}'s are covariates and $\alpha_1, \dots, \alpha_m \sim N(0, \sigma^2)$.

Suppose the x_{ij}'s are bounded, and $\liminf (1/N)\sum_{i=1}^m \sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2 > 0$, that is, asymptotically, there is variation within the groups.

Suppose that $\log m/(\log(N/m))^2 \to 0$. Let $M_N = (2 + c)d_N$, where $d_N = (\log m)^{1/4}(\log(N/m))^{1/2}$, and $c \ge \max_{i,j}|x_{ij}|$. Then, by Lemma 3.1, $P(\|\alpha_0\| < d_N) \to 1$. Let $c_N = d_N$ and $M_{ij} = M_N$. Then, $M_{ij} \ge c_N(1 + |x_{ij}|) + d_N$. Also, as in Example 3.1, $\delta_N \ge (1/4)\exp(-2M_N)$. Thus, it is easy to show that $(m/N)\delta_N^{-2} \to 0$. It follows, by Theorem 2.2, that $\hat{\beta}_1 \to_P \beta_{01}$ and $N^{-1}\sum_{i=1}^m n_i(\hat{a}_i - a_{0i})^2 \to_P 0$, where $a_i = \beta_0 + \alpha_i$.

If, furthermore, $\log m/(\log n_*)^2 \to 0$, where $n_* = \min_i n_i$, we can choose, instead, $M_N = (2 + c)d_N$, where $d_N = c_N = (\log m)^{1/4}[(\log(N/m))^{1/2} \wedge (\log n_*)^{1/2}]$. Then, by the same argument, we have $(n_*)^{-1}\delta_N^{-2} \to 0$. Thus, by Theorem 2.2, $\hat{\beta}_0 \to_P \beta_{00}$ and $N^{-1}\sum_{i=1}^m n_i(\hat{\alpha}_i - \alpha_{0i})^2 \to_P 0$, $m^{-1}\sum_{i=1}^m(\hat{\alpha}_i - \alpha_{0i})^2 \to_P 0$.

Finally, we use one example to illustrate Theorem 2.3.

EXAMPLE 3.3. Consider the model defined by (1.12). Suppose the a_i's and b_{ij}'s are normally distributed with $\sigma_{0b}^2 > 0$ and $r \ge 2$. Take $\alpha = a = (a_i)_{1 \le i \le m_1}$, $\zeta = b = (b_{ij})_{1 \le i \le m_1, 1 \le j \le n}$. We have $X = 1_{m_1 nr}$ and $Z = I_{m_1} \otimes 1_{nr}$. The transformation of Lemma 2.4 results in $\tilde{a}_i = \mu + a_i$, $1 \le i \le m_1$, and there is no $\tilde{\beta}$. Furthermore, we have $\tilde{Z} = Z$, $\tilde{Z}_* = I_{m_1} \otimes 1_n$, and $S_{i,j} = \{(i, j, k)\colon 1 \le k \le r\}$. Suppose m_1, $n \to \infty$, and there is $1/2 < \rho < 1$ such that

$$(\log m_1)^{\rho}/\log n \to 0. \tag{3.1}$$

We shall verify the conditions of Theorem 2.3.

(i) is obvious, and (ii) follows from Lemma 3.3 in the following.

(iii) $\tilde{Z}_{*i}$ is a $m_1 n$-dimensional vector whose components are divided into m_1-blocks of equal length n with the ith block being 1_n and other blocks 0; $\|\tilde{Z}_*\| = 1$; $s_{M,i}^{(l)} = n$, $1 \le i \le m_1$, $l = 1, 2, \dots$; $|\tilde{z}_{*ij}| = 1$, which, combined with Lemma 3.3, implies the boundedness of (2.20).

(iv) By Lemma 3.2 in the following, there is $c_0 > 0$ such that

$$\lambda_M \ge d_0 \exp(-2(2r + 5)\|a_0\|), \tag{3.2}$$

where $d_0 = c_0 \exp(-2(2r + 5)|\mu_0|)$, μ_0 is the true μ and a_0 is the vector of true realizations of a. By Lemma 3.1,

$$P\big(\|a_0\| \le (\log m_1)^{\rho}\big) \to 1. \tag{3.3}$$

Let $\rho_M = d_0 \exp(-2(2r+5)(\log m_1)^{\rho}) \wedge \lambda_M \wedge 1$. It is easy to see that $t_{M,i} = 0$, $1 \le i \le m_1$ and we have, with probability approaching 1,

$$\left(\frac{n}{M}\right)\Big/\rho_M^4 \le \left(\frac{1}{m_1}\right) \vee d_0^{-4} \exp\left(-\log m_1 + 8(2r+5)(\log m_1)^{\rho}\right) \to_P 0;$$

$$\begin{aligned}\log n/\rho_M^{2l} &\min_{1 \le i \le m_1} s_{M,i}^{(6-2l)}\\ &\le \left(\frac{\log n}{n}\right) \vee d_0^{-2l} \exp\big(-\log n + \log\log n\\ &\qquad\qquad + 4l(2r+5)(\log m_1)^{\rho}\big) \to_P 0;\end{aligned}$$

and

$$n \max_{1 \le i \le m_1} V_i^{(3-l)}(\rho_M^l)/\rho_M^l = 0 \quad \text{for large } m_1,\ l = 1, 2.$$

Finally, we have

$$\begin{aligned}&\lambda_{\max}\left(G^{-1/2} A_2 G^{-1/2}\right)\\ &\qquad \le \max_{1 \le i \le m_1} \left\| \frac{1}{n} \sum_{j=1}^{n} \left(H_{ij}(\psi_0) - E\left(H_{ij}(\psi_0)|\psi_0\right)\right)\right\|_R .\end{aligned} \tag{3.4}$$

Let $\xi_{i,j,c,d} = (\partial^2 h_{ij}/\partial s_c\, \partial s_d)(\psi_0) - E(\cdots|\psi_0)$, where $h_{ij}(s) = \log E \exp((s_1 + s_2 \xi) y_{ij}.- r \log(1 + \exp(s_1 + s_2 \xi)))$ with $\xi \sim N(0,1)$. By Lemma 5.3 in Section 5, there is a constant $B > 0$ such that $\forall\ \delta > 0$, whenever $0 < \delta\rho_M \le 4/B$,

$$\begin{aligned}P&\left(\max_{1 \le i \le m_1} \left\| \frac{1}{n} \sum_{j=1}^{n} \left(H_{ij}(\psi_0) - E\left(H_{ij}(\psi_0)|\psi_0\right)\right)\right\|_R > \delta\rho_M B^2 \middle| \psi_0\right)\\ &\le \sum_{c,d=1}^{2} \sum_{i=1}^{m_1} P\left(\left|\sum_{j=1}^{n} \xi_{i,j,c,d}\right| > \frac{\delta\rho_M}{4} n B^2 \middle| \psi_0\right)\\ &\le \sum_{c,d=1}^{2} \sum_{i=1}^{m_1} 2\exp\left(-\frac{\delta^2\rho_M^2}{64} n B^2\right) = 8\exp\left[n\rho_M^2\left(-\frac{\delta^2 B^2}{64} + o_P(1)\right)\right] \to_P 0.\end{aligned}$$

Therefore, by the dominated convergence theorem, $\lambda_{\max}(G^{-1/2} A_2 G^{-1/2})/\rho_M \to_P 0$.

LEMMA 3.2. *Let* $\theta = (\mu, \tau)$, $h(\theta, k) = \log E \exp((\mu + \tau\xi)k - r\log(1 + \exp(\mu + \tau\xi)))$, $k = 0, 1, \dots, r$, *where* $\tau > 0$, $r \ge 2$ *and* $\xi \sim N(0,1)$. *Let* Y *be a random variable taking values in* $\{0, 1, \dots, r\}$ *such that* $P(Y = k) = \binom{r}{k}\exp(h(\theta, k))$. *Then, there is a constant* $c > 0$ *which may depend on* τ *such that for all* μ,

$$\lambda(\theta) \equiv \lambda_{\min}\left(\operatorname{Var}\left(\frac{\partial h}{\partial \theta}(\theta, Y)\right)\right) \ge c\exp(-2(2r+5)|\mu|). \tag{3.5}$$

LEMMA 3.3. *For any $b > 0$, the first, second and third derivatives of $h(\theta, k)$ (see Lemma 3.2) are uniformly bounded for $\mu \in R$, $0 \le \tau \le b$ and $0 \le k \le r$.*

4. Remarks.

1. In many cases the variance components of the random effects are of interest. The methods developed in Sections 1 and 2 provide an easy way to consistently estimate these parameters. To see this, let us first consider the case $m/N \to 0$. Suppose that α and Z have the structures described by (ii) of Corollary 2.2. Suppose that $\alpha_{u1}, \dots, \alpha_{um_u}$ are i.i.d. with $\sigma_u^2 = \mathrm{var}(\alpha_{uv})$, $E(\alpha_{uv}^4) < \infty$ and that $(\sum_{v=1}^{m_u} n_{uv})^{-2} \sum_{v=1}^{m_u} n_{uv}^2 \to 0$, $1 \le u \le q$. Then (2.5) implies that $(\sum_{v=1}^{m_u} n_{uv})^{-1} \sum_{v=1}^{m_u} n_{uv} \hat{\alpha}_{uv}^2 \to_P \sigma_u^2$, $1 \le u \le q$. In the case of m/N not $\to 0$, the MCL estimate $\hat{\tau}$ of τ, which often corresponds to a dispersion parameter, is consistent under the conditions of Theorem 2.3. If, furthermore, α and Z in (1.13) are as described above, the same consistent results hold for the variances of the α_{uv}'s.

2. As pointed out in Section 1, our procedures are different from the traditional approaches, in which one needs to get the variance components right before going to the effects, or one has to estimate the effects and variance components simultaneously. It may seem surprising that one can estimate the fixed and random effects without first getting the variance components right. Here, we need to clarify a few points. First, there have been other occasions in which one estimates the effects without "getting the variance components right." A well-known example is WLS [e.g., Diggle, Liang and Zeger (1996), Section 4.3], in which an estimate of β, the vector of regression coefficients, is obtained by minimizing the quadratic form $(y - X\beta)^t W(y - X\beta)$, where W is a weighting matrix. In cases where the responses are correlated, it can be shown that the optimal weighting matrix, in the sense of mean squared errors, is given by $W = V^{-1}$, where V is the covariance matrix of the errors. Note that V involves not only the variance components but also the correlation structures. Therefore, to identify this optimal weighting matrix, one needs not only to get the variance components right, but also to know the complete correlation structure of the data. The latter is often more difficult to do and requires more assumptions. On the other hand, the WLS estimate with an arbitrary W is unbiased, consistent, and asymptotically normal, even if it is not efficient. Our procedures, in a sense, are similar to WLS. Note that we do not assume the covariance matrix of the random effects α is known up to a number of variance components [(1.3) is the basic assumption]. In fact, with only (1.3), it may not be clear what are the "variance components." Also, our results only show the consistency of these estimates, not the asymptotic optimality. Second, the consistency of the PGWLS (or MCL) estimates holds only in large sample cases where there is enough information in the data about the random effects (or a subset of the random effects). For example, in Example 1.1, if both m_1 and m_2 are large, one has enough information about the a_i's and b_j's. One does

not need to know, in addition, the variances of the a_i's and b_j's, because these can be deduced from the previous information (see Remark 1). Finally, one should always be aware that large sample results may not apply to small sample cases.

3. In PGWLS we have chosen the penalizer as in the form $\lambda|P\alpha|^2$ [see (1.11)]. One reason for choosing such a penalizer is computational convenience, because it is a quadratic function of α, which corresponds to a linear function of α on the left side of the estimating equations (2.6). A penalizer of the form $\lambda|P\alpha|^2$ also has some ideal theoretical properties. For example, with $P = P_A$, where A is defined above (1.11), the maximizer of l_P, if it exists, is unique (Corollary 2.1). On the other hand, it is seen from the proof given in the next section that the result of Theorem 2.1 would hold for a variety of penalizers not necessarily quadratic, for example, those of the form $|P\alpha|^k$, where $k > 0$. It would be interesting to know what is the "best" choice of penalizer. Note that although many penalties would lead to consistent estimates, there may be a difference in terms of efficiency and small sample properties.

The PGWLS procedure also involves a constant λ which is assumed known. For the consistency of the estimates, λ does not make a difference. On the other hand, the choice of λ might affect the (asymptotic) efficiency as well as the small sample behavior of the estimates. It would be interesting to know to what extent this is the case.

4. There is some connection between PGWLS and the penalized-likelihood method based on Laplace approximation [e.g., Breslow and Clayton (1993), Shun and McCullagh (1995), Lee and Nelder (1996), Vonesh (1996)]. The Laplace-based penalized-likelihood method leads to a penalizer which typically involves unknown parameters such as variance components. PGWLS, on the other hand, is simpler in the sense that the penalizer is completely specified. It can be shown that [Jiang, Jia and Chen (1999)], in terms of consistency, the unknown variance components involved in the Laplace-based penalizer do not make a difference, provided that $m/N \to 0$. Therefore, assuming normality of the random effects and with any given values of the variance components, the Laplace-based method leads to a penalized log-likelihood in the form of (1.11). Note that the Laplace approximations considered here are nonstandard in the sense that the dimension of integrals increases with the sample size. Earlier results have indicated that, under more restricted limiting process than $m/N \to 0$, the Laplace approximation is asymptotically exact, and the estimates of the fixed parameters are consistent. For example, Shun and McCullagh (1995) requires that $m/N^{1/3} \to 0$; Vonesh (1996) considers a special case of a nonlinear mixed model with single random factor and requires that $\min_i p_i \to \infty$, where p_i is the number of observations at the ith level of the random factor. Note that the latter assumption is, in fact, sufficient for the estimates of individual random effects to be consistent, while under $m/N \to 0$, one can only expect consistency of the estimates of the random effects in an "overall" sense (see Theorem 2.1).

Nevertheless, the PGWLS estimates of the fixed effects are consistent under $m/N \to 0$.

5. To apply the MCL method, the conditional density $f(y_i|\eta_i)$ in (1.14) must be known. Sometimes, such as in (1.6), $f(y_i|\eta_i)$ may contain an additional dispersion parameter ϕ, although, in some cases such as the binomial and Poisson models, ϕ is known. When ϕ is unknown, it, too, has to be estimated. An obvious way to do this is to include ϕ as part of φ defined below (2.12) so that it can be estimated jointly with other parameters.

5. Proofs.

PROOF OF LEMMA 2.1. Let $B = -(X^tX)^{-1}X^tZA$. Then $XB + ZA = P_{X^\perp}$ ZA = 0. *On the other hand*, $\textit{rank}(\binom{B}{A}) = \text{rank}(A) = m - \textit{rank}(P_{X^\perp}Z) = m -$ $(\text{rank}((X\ Z)) - \text{rank}(X)) = p + m - \text{rank}((X\ Z))$. Thus $\binom{B}{A} \in \mathscr{B}(\mathscr{N}((X\ Z)))$. Suppose, (i) $X\beta + Z\alpha = 0$, (ii) $A^t\alpha = 0$. (i) $\Rightarrow \binom{\beta}{\alpha} = \binom{B}{A}l$ for some vector l. Thus by (ii), $A^tAl = 0 \Rightarrow l = 0$. □

PROOF OF COROLLARY 2.1. It is enough to show that $(\partial^2 l_P/\partial\gamma^2) < 0$, $\forall\ \gamma$. Simple calculation shows that

$$\begin{aligned} \frac{\partial^2 l_C}{\partial\gamma^2} &= -\sum_{i=1}^N w_i b_i''(\eta_i)\begin{pmatrix} x_i \\ z_i \end{pmatrix}\begin{pmatrix} x_i \\ z_i \end{pmatrix}^t \\ &\le -\delta\sum_{i=1}^N \begin{pmatrix} x_i \\ z_i \end{pmatrix}\begin{pmatrix} x_i \\ z_i \end{pmatrix}^t = -\delta(X\ Z)^t(X\ Z), \end{aligned} \tag{5.1}$$

where $\delta = \min_{1\le i\le N}\{w_i b_i''(\eta_i)\} > 0$. Thus for any $v \in R^p$, $u \in R^m$,

$$\begin{aligned} \begin{pmatrix} v \\ u \end{pmatrix}^t \frac{\partial^2 l_P}{\partial\gamma^2}\begin{pmatrix} v \\ u \end{pmatrix} &= \begin{pmatrix} v \\ u \end{pmatrix}^t \frac{\partial^2 l_C}{\partial\gamma^2}\begin{pmatrix} v \\ u \end{pmatrix} - \lambda\begin{pmatrix} v \\ u \end{pmatrix}^t\begin{pmatrix} 0 & 0 \\ 0 & P_A \end{pmatrix}\begin{pmatrix} v \\ u \end{pmatrix} \\ &\le -\delta|Xv + Zu|^2 - \lambda|P_A u|^2 \le 0. \end{aligned}$$

If " = " in the above holds, then (i) $Xv + Zu = 0$; (ii) $P_A u = 0$. (ii) $\Rightarrow A^t u = 0$ $\Rightarrow \begin{pmatrix} X & Z \\ 0 & A^t \end{pmatrix}\binom{v}{u} = 0 \Rightarrow \binom{v}{u} = 0$ by Lemma 2.1. □

PROOF OF LEMMA 2.2.

$\forall\ \gamma \in S$, $\gamma^t H\gamma = |X\beta + Z\alpha|^2 = \gamma^t G_A^t G_A\gamma \ge \lambda_{\min}(G_A^t G_A)|\gamma|^2$. $\forall\ \gamma_* \in WS$, we have

$$\begin{aligned} \gamma = W^{-1}\gamma_* \in S \Rightarrow \gamma_*{}^t W^{-1}HW^{-1}\gamma_* &= \gamma^t H\gamma \ge \lambda_{\min}(G_A^t G_A)|\gamma|^2 \\ &= \lambda_{\min}(G_A^t G_A)\gamma_*{}^t W^{-2}\gamma_* \\ &\ge \frac{\lambda_{\min}(G_A^t G_A)}{\left(\max_{1\le u\le p} b_u^2\right) \vee \left(\max_{1\le k\le m} a_k^2\right)}|\gamma_*|^2. \end{aligned}$$

The result then follows from Lemma 2.1. □

PROOF OF THEOREM 2.1. First we show that $\hat{\gamma} \in S$. This is because $\tilde{\hat{\gamma}} \in \Gamma(M)$ since $\tilde{\hat{\eta}} = X\tilde{\hat{\beta}} + Z\tilde{\hat{\alpha}} = X\hat{\beta} + Z\hat{\alpha} = \hat{\eta} \Rightarrow l_P(\hat{\gamma}) = l_C(\hat{\gamma}) - (\lambda/2)|P_A\hat{\alpha}|^2 \geq l_P(\tilde{\hat{\gamma}}) = l_C(\tilde{\hat{\gamma}}) - (\lambda/2)|P_A\tilde{\hat{\alpha}}|^2 = l_C(\hat{\gamma}) - (\lambda/2)|P_A P_{A^{\perp}}\hat{\alpha}|^2 = l_C(\hat{\gamma}) \Rightarrow (\lambda/2)|P_A\hat{\alpha}|^2 \leq 0 \Rightarrow P_A\hat{\alpha} = 0$, that is, $\hat{\gamma} \in S$.

By Taylor series expansion and the fact that $\tilde{\eta}_0 = \eta_0$, we have

$$\begin{aligned} l_C(\gamma) - l_C(\tilde{\gamma}_0) &= \sum_{u=1}^{p} \frac{\partial l_C}{\partial \beta_u}\bigg|_{\gamma_0} \left(\beta_u - \tilde{\beta}_{0u}\right) + \sum_{k=1}^{m} \frac{\partial l_C}{\partial \alpha_k}\bigg|_{\gamma_0} \left(\alpha_k - \tilde{\alpha}_{0k}\right) \\ &\quad + \frac{1}{2}(\gamma - \tilde{\gamma}_0)^t \frac{\partial^2 l_C}{\partial \gamma^2}\bigg|_{\gamma^*} (\gamma - \tilde{\gamma}_0) \qquad (5.2) \\ &= I_1 - \frac{1}{2} I_2, \end{aligned}$$

where $\gamma^* = (1-t)\tilde{\gamma}_0 + t\gamma$ for some $0 \leq t \leq 1$. We have

$$\begin{aligned} E\left(\sum_{u=1}^{p} |X_u|^{-2} \left(\frac{\partial l_C}{\partial \beta_u}\bigg|_{\gamma_0}\right)^2\right) &= \sum_{u=1}^{p} |X_u|^{-2} E\left(E\left[\left(\sum_{i=1}^{N} w_i x_{iu}\left(y_i - b_i'(\eta_{0i})\right)\right)^2 \bigg| \alpha_0\right]\right) \\ &= \sum_{u=1}^{p} |X_u|^{-2} \sum_{i=1}^{N} w_i^2 x_{iu}^2 E \operatorname{var}(y_i|\alpha_0) \\ &\leq \left(\max_{1 \leq i \leq N} w_i^2 E \operatorname{var}(y_i|\alpha_0)\right) p. \end{aligned}$$

Similarly,

$$\begin{aligned} E\left(\sum_{k=1}^{m} |Z_k|^{-2} \left(\frac{\partial l_C}{\partial \alpha_k}\bigg|_{\gamma_0}\right)^2\right) &= \sum_{k=1}^{m} |Z_k|^{-2} \sum_{i=1}^{N} w_i^2 z_{ik}^2 E \operatorname{var}(y_i|\alpha_0) \\ &\leq \left(\max_{1 \leq i \leq N} w_i^2 E \operatorname{var}(y_i|\alpha_0)\right) m. \end{aligned}$$

Therefore, by (2.3),

$$\begin{aligned} I_1 &\leq \left(\sum_{u=1}^{p} |X_u|^{-2} \left(\frac{\partial l_c}{\partial \beta_u}\bigg|_{\gamma_0}\right)^2\right)^{1/2} \left(\sum_{u=1}^{p} |X_u|^2 \left(\beta_u - \tilde{\beta}_{0u}\right)^2\right)^{1/2} \\ &\quad + \left(\sum_{k=1}^{m} |Z_k|^{-2} \left(\frac{\partial l_c}{\partial \alpha_k}\bigg|_{\gamma_0}\right)^2\right)^{1/2} \left(\sum_{k=1}^{m} |Z_k|^2 \left(\alpha_k - \tilde{\alpha}_{0k}\right)^2\right)^{1/2} \qquad (5.3) \\ &\leq \omega\sqrt{N}\, o_P(1) |W(\gamma - \tilde{\gamma}_0)|. \end{aligned}$$

On the other hand, if $\gamma_0, \gamma \in \Gamma(M)$, then, since $\eta^* = (X\ Z)\gamma^* = (1-t)(X\ Z)\tilde{\gamma}_0 + t(X\ Z)\gamma = (1-t)\tilde{\eta}_0 + t\eta = (1-t)\eta_0 + t\eta$, $|\eta_i^*| \leq M_i$, $1 \leq i \leq N$, hence $\gamma^* \in \Gamma(M)$. Therefore, we have, by (5.1) and the fact that $\tilde{\gamma} \in S$, $\forall\ \gamma$, that when $\gamma_0 \in \Gamma(M)$ and $\gamma \in \Gamma(M) \cap S$,

$$\begin{aligned} I_2 &= (\gamma - \tilde{\gamma}_0)^t \left(- \frac{\partial^2 l_C}{\partial \gamma^2} \bigg|_{\gamma^*} \right) (\gamma - \tilde{\gamma}_0) \\ &= \big(W(\gamma - \tilde{\gamma}_0)\big)^t \left(-W^{-1} \frac{\partial^2 l_C}{\partial \gamma^2} \bigg|_{\gamma^*} W^{-1} \right) \big(W(\gamma - \tilde{\gamma}_0)\big) \\ &\geq \lambda_{\min} \left(-W^{-1} \frac{\partial^2 l_C}{\partial \gamma^2} \bigg|_{\gamma^*} W^{-1} \right) \Big|_{WS} |W(\gamma - \tilde{\gamma}_0)|^2 \\ &\geq \omega |W(\gamma - \tilde{\gamma}_0)|^2. \end{aligned} \tag{5.4}$$

Since $l_P(\gamma) = l_C(\gamma)$, $\gamma \in S$, we have, by combining (5.2)–(5.4), that when $\gamma_0 \in \Gamma(M)$ and $\gamma \in \Gamma(M) \cap S$,

$$\begin{aligned} l_P(\gamma) - l_P(\tilde{\gamma}_0) &= l_C(\gamma) - l_C(\tilde{\gamma}_0) \\ &\leq \omega \sqrt{N} o_P(1) |W(\gamma - \tilde{\gamma}_0)| - \frac{\omega}{2} |W(\gamma - \tilde{\gamma}_0)|^2 \\ &= \omega \sqrt{N} |W(\gamma - \tilde{\gamma}_0)| \left(o_P(1) - \frac{1}{2} \left(\frac{|W(\gamma - \tilde{\gamma}_0)|}{\sqrt{N}} \right) \right). \end{aligned} \tag{5.5}$$

Note that the $o_P(1)$ in (5.5) does not depend on γ. Let $R_\varepsilon = \{\gamma\colon |W(\gamma - \tilde{\gamma}_0)| < \varepsilon\sqrt{N}\}$ $(\varepsilon > 0)$. Since $\gamma_0 \in \Gamma(M) \Rightarrow \tilde{\gamma}_0 \in \Gamma(M) \cap S$, we have, by (5.5) and the proved fact that $\hat{\gamma} \in S$, that

$$\begin{aligned} &\left\{ \gamma_0 \in \Gamma(M),\ o_P(1) < \frac{\varepsilon}{2} \right\} \\ &\qquad \subset \{ \tilde{\gamma}_0 \in \Gamma(M) \cap S,\ l_P(\gamma) < l_P(\tilde{\gamma}_0),\ \forall\, \gamma \in \Gamma(M) \cap S \cap R_\varepsilon^c \} \subset \{ \hat{\gamma} \in R_\varepsilon \}. \end{aligned}$$

Since $\{\gamma_0 \in \Gamma(M)\} \supset \{\|\beta_0\| \leq c_N,\ \|\alpha_0\| \leq d_N\}$, we have $P(\hat{\gamma} \in R_\varepsilon) \to 1$. By the arbitrariness of ε, we have

$$\frac{1}{N} |W(\hat{\gamma} - \tilde{\gamma}_0)|^2 \to_P 0. \tag{5.6}$$

Finally, let $W_X = \operatorname{diag}(|X_u|,\ 1 \leq u \leq p)$, $W_Z = \operatorname{diag}(|Z_k|,\ 1 \leq k \leq m)$. Then,

$$\begin{aligned} |W(\tilde{\gamma}_0 - \gamma_0)|^2 &= |W_X(\tilde{\beta}_0 - \beta_0)|^2 + |W_Z(\tilde{\alpha}_0 - \alpha_0)|^2 \\ &= |W_X (X^t X)^{-1} X^t Z P_A \alpha_0|^2 + |W_Z P_A \alpha_0|^2 \\ &\leq \|W_X\|^2 \|(X^t X)^{-1} X^t Z\|^2 |P_A \alpha_0|^2 + \|W_Z\|^2 |P_A \alpha_0|^2 \\ &= \left[\left(\max_{1 \leq u \leq p} |X_u|^2 \right) \|(X^t X)^{-1} X^t Z\|^2 + \left(\max_{1 \leq k \leq m} |Z_k|^2 \right) \right] |P_A \alpha_0|^2. \end{aligned}$$

(2.2) thus follows from (5.6) and (2.1). □

Proof of Corollary 2.3. Take $c_N = (1 + \|\beta_0\|)N$, $d_N = (1 + E\|\alpha_0\|)N$ and $M_i = \infty$, $1 \leq i \leq N$. □

Proof of Theorem 2.2. In the following, we use the abbreviation "PT2.1" for "the proof of Theorem 2.1." As in PT2.1, $\hat{\gamma} \in S$. By (5.2), it is easy to show that

$$\begin{aligned} l_C(\gamma) - l_C(\tilde{\gamma}_0) &= \sum_{i=1}^{m} \sum_{j=1}^{n_i} w_{ij}\big(y_{ij} - E(y_{ij}|v_0)\big)\big(\eta_{ij} - \tilde{\eta}_{0ij}\big) \\ &\quad - \tfrac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n_i} w_{ij} b''_{ij}(\eta_{*ij})\big(\eta_{ij} - \tilde{\eta}_{0ij}\big)^2 \\ &= I_1 - \tfrac{1}{2} I_2. \end{aligned} \tag{5.7}$$

Note that

$$\eta_{ij} - \tilde{\eta}_{0ij} = a_i - \tilde{a}_{0i} + \bar{x}_i\big(\beta - \tilde{\beta}_0\big) + \big(x_{ij} - \bar{x}_i\big)^t\big(\beta - \tilde{\beta}_0\big). \tag{5.8}$$

Thus we have, similar to PT2.1, that

$$\begin{aligned} |I_1| &\le \left(\sum_{i=1}^{m} n_i^{-1} \left(\sum_{j=1}^{n_i} w_{ij}\big(y_{ij} - E(y_{ij}|v_0)\big) \right)^2 \right)^{1/2} \\ &\quad \times \left(\sum_{i=1}^{m} n_i\big(a_i - \tilde{a}_{0i} + \bar{x}_i^t\big(\beta - \tilde{\beta}_0\big)\big)^2 \right)^{1/2} \\ &\quad + \left| \sum_{i=1}^{m} \sum_{j=1}^{n_i} w_{ij}\big(y_{ij} - E(y_{ij}|v_0)\big)\big(x_{ij} - \bar{x}_i\big) \right| |\beta - \tilde{\beta}_0| \end{aligned} \tag{5.9}$$

$$= O_P(1)\sqrt{m} \left(\sum_{i=1}^{m} n_i\big(a_i - \tilde{a}_{0i} + \bar{x}_i^t\big(\beta - \tilde{\beta}_0\big)\big)^2 \right)^{1/2} + O_P(1)\sqrt{N}|\beta - \tilde{\beta}_0|;$$

$$\begin{aligned} I_2 &\ge \delta_N \sum_{i=1}^{m} \sum_{j=1}^{n_i} \Big(a_i - \tilde{a}_{0i} + \bar{x}_i^t\big(\beta - \tilde{\beta}_0\big) + \big(x_{ij} - \bar{x}_i\big)^t\big(\beta - \tilde{\beta}_0\big)\Big)^2 \\ &\ge \delta_N \left(\sum_{i=1}^{m} n_i\big(a_i - \tilde{a}_{0i} + \bar{x}_i^t\big(\beta - \tilde{\beta}_0\big)\big)^2 + \lambda_N |\beta - \tilde{\beta}_0|^2 \right). \end{aligned} \tag{5.10}$$

Let $r_N^2 = \sum_{i=1}^m n_i(a_i - \tilde{a}_{0i} + \bar{x}_i^t(\beta - \tilde{\beta}_0))^2 + N|\beta - \tilde{\beta}_0|^2$, we have, by (5.7), (5.9) and (5.10) that, when $\gamma_0 \in \Gamma(M)$, $\gamma \in \Gamma(M) \cap S$,

$$\begin{aligned} l_P(\gamma) - l_P(\tilde{\gamma}_0) &\le O_P(1)\sqrt{m}\, r_N + O_P(1) r_N - \left(1 \wedge \frac{\lambda_N}{N}\right) \frac{\delta_N}{2} r_N^2 \\ &= \delta_N r_N \sqrt{N} \left(O_P(1)\, \delta_N^{-1} \sqrt{\frac{m}{N}} - \frac{1}{2}\left(1 \wedge \frac{\lambda_N}{N}\right) \frac{r_N}{\sqrt{N}} \right). \end{aligned} \tag{5.11}$$

For any $\varepsilon > 0$, let $R_\varepsilon = \{r_N^2 < \varepsilon^2 N\}$. By (5.11) and the same argument as in PT2.1, we have $P(\hat{\gamma} \in R_\varepsilon) \to 1$. Thus, $\hat{r}_N^2/N \to_P 0$, where $\hat{r}_N$ is r_N with γ replaced by $\hat{\gamma}$. It follows that

$$|\hat{\beta} - \tilde{\beta}_0|^2 \to_P 0 \quad \text{and} \quad \frac{1}{N} \sum_{i=1}^{m} n_i\big(\hat{a}_i - \tilde{a}_{0i}\big)^2 \to_P 0. \tag{5.12}$$

We now show that

$$|\tilde{\beta}_0 - \beta_0|^2 \to_P 0 \quad \text{and} \quad \frac{1}{N}\sum_{i=1}^{m} n_i(\tilde{a}_{0i} - a_{0i})^2 \to_P 0. \tag{5.13}$$

Therefore, the conclusions without $\min_{1 \le i \le m} n_i \to \infty$ follow. To show (5.13), we let $\beta^* = (a, \beta)$, and note that, by the definition, $\tilde{\beta}_0^* - \beta_0^* = (X^tX)^{-1}X^tZP_Av_0$, $\tilde{v}_0 - v_0 = -P_Av_0$ and $|\tilde{\beta}_0 - \beta_0| \le |\tilde{\beta}_0^* - \beta_0^*|$, $|\tilde{a}_{0i} - a_{0i}| \le |\tilde{\beta}_0^* - \beta_0^*| + |\tilde{v}_{0i} - v_{0i}|$. It is easy to show that the first row of X^tZ consists of n_i, $1 \le i \le m$, and the $(u+1)$'s row $(1 \le u \le s)$ of X^tZ consists of $x_{i\cdot u}$, $1 \le i \le m$, where $x_{i\cdot u} = \sum_{j=1}^{n_i} x_{iju}$. It follows that $|\tilde{\beta}_0^* - \beta_0^*| \le \|(N^{-1}X^tX)^{-1}\|\,|B|$, where $B = N^{-1}X^tZP_Av_0$. Now, $\|(N^{-1}X^tX)^{-1}\| = (N^{-1}\lambda_{\min}(X^tX))^{-1}$, and |the first row of $B| \le \|P_Av_0\|$, |the $(u+1)$'s row of $B| \le \|P_Av_0\|\max_{i,j,k}|x_{ijk}|$, $1 \le u \le s$ and $\|P_Av_0\| = |\bar{v}_0|$. Thus, $|\tilde{\beta}_0^* - \beta_0^*| \to_P 0$. (5.13) thus follows.

Finally, let $n_* = \min_{1 \le i \le m} n_i$, and suppose $n_*^{-1} = o(\delta_N^2)$. Let $s_N^2 = \sum_{i=1}^m n_i(a_i - \tilde{a}_{0i} + \bar{x}_i^t(\beta - \tilde{\beta}_0))^2 + (\lambda_N/2)|\beta - \tilde{\beta}_0|^2$, and $S_\varepsilon = \{s_N^2 < \varepsilon^2 mn_*\}$ $(\varepsilon > 0)$. By (5.7), (5.9) and (5.10), we have that, when $\gamma_0 \in \Gamma(M)$ and $\gamma \in \Gamma(M) \cap S$,

$$\begin{aligned} &l_P(\gamma) - l_P(\tilde{\gamma}_0) \\ &\quad \le O_P(1)\sqrt{m}\,s_N - \frac{\delta_N}{2}s_N^2 + O_P(1)\sqrt{N}|\beta - \tilde{\beta}_0| \\ &\qquad - \frac{\delta_N\lambda_N}{4N}\left(\sqrt{N}|\beta - \tilde{\beta}_0|\right)^2 \\ &\quad \le \delta_N s_N\sqrt{mn_*}\left(O_P(1)\,\delta_N^{-1}n_*^{-1/2}\right. \\ &\qquad\qquad \left. + O_P(1)\,\delta_N^{-2}n_*^{-1}\frac{\sqrt{n_*}}{\sqrt{m}\,s_N} - \frac{s_N}{2\sqrt{mn_*}}\right). \end{aligned} \tag{5.14}$$

Here we use the fact that the function $\lambda x - \mu x^2$ $(\mu > 0)$ is bounded by $\lambda^2/4\mu$. By (5.14) and a similar argument as in PT2.1, we have $P(\hat{\gamma} \in S_\varepsilon) \to 1$. Thus, $\hat{s}_N^2/mn_* \to_P 0$, where $\hat{s}_N$ is s_N with γ replaced by $\hat{\gamma}$. It follows that [using the first half of (5.12)] $m^{-1}\sum_{i=1}^m(\hat{a}_i - \tilde{a}_{0i})^2 \to_P 0$. Also, by the same argument as before, we have $m^{-1}\sum_{i=1}^m(\tilde{a}_{0i} - a_{0i})^2 \to_P 0$. Therefore, by the fact that $\hat{\gamma} \in S$, we have

$$\begin{aligned} (\hat{a} - a_0 - \bar{v}_0)^2 &= \left(\frac{1}{m}\sum_{i=1}^m(\hat{a}_i - a_{0i})\right)^2 \\ &\le \frac{1}{m}\sum_{i=1}^m(\hat{a}_i - a_{0i})^2 \to_P 0. \end{aligned}$$

The rest of the conclusions follow easily. □

PROOF OF LEMMA 2.4. Let $\alpha^* = \alpha + (Z^tZ)^- Z^tX\beta$. Then

$$\begin{aligned} X\beta + Z\alpha &= P_{Z^\perp}X\beta + P_Z X\beta + Z\alpha \\ &= P_{Z^\perp}X\beta + Z(Z^tZ)^- Z^tX\beta + Z\alpha \\ &= P_{Z^\perp}X\beta + Z\alpha^*, \end{aligned} \tag{5.15}$$

Let $S = \{j_1, \dots, j_s\}$ be a set of indexes such that $\tilde{X} = (P_{Z^\perp}X_j)_{j \in S} \in \mathscr{B}(\mathscr{L}(P_{Z^\perp}X))$. Then for $k \notin S$ there are numbers λ_{jk}, $j \in S$ such that $P_{Z^\perp}X_k = \sum_{j \in S} \lambda_{jk} P_{Z^\perp}X_j$. Thus

$$\begin{aligned} P_{Z^\perp}X\beta &= \sum_{k=1}^{p} P_{Z^\perp}X_k \beta_k \\ &= \sum_{k \in S} P_{Z^\perp}X_k \beta_k + \sum_{k \notin S} \left(\sum_{j \in S} \lambda_{jk} P_{Z^\perp}X_j \right) \beta_k \\ &= \sum_{j \in S} P_{Z^\perp}X_j \left(\beta_j + \sum_{k \notin S} \lambda_{jk} \beta_k \right) = \tilde{X}\tilde{\beta}, \end{aligned} \tag{5.16}$$

where $\tilde{\beta} = (\tilde{\beta}_j)_{j \in S}$ with $\tilde{\beta}_j = \beta_j + \sum_{k \notin S} \lambda_{jk} \beta_k$.

Let $T = \{k_1, \dots, k_t\}$ be a set of indexes such that $\tilde{Z} = (Z_k)_{k \in T} \in \mathscr{B}(\mathscr{L}(Z))$. Then for $j \notin T$ there are numbers μ_{kj}, $k \in T$ such that $Z_j = \sum_{k \in T} \mu_{kj} Z_k$. Thus

$$\begin{aligned} Z\alpha^* &= \sum_{j=1}^{m} Z_j \alpha_j^* \\ &= \sum_{j \in T} Z_j \alpha_j^* + \sum_{j \notin T} \left(\sum_{k \in T} \mu_{kj} Z_k \right) \alpha_j^* \\ &= \sum_{k \in T} Z_k \left(\alpha_k^* + \sum_{j \notin T} \mu_{kj} \alpha_j^* \right) = \tilde{Z}\tilde{\alpha}, \end{aligned} \tag{5.17}$$

where $\tilde{\alpha} = (\tilde{\alpha}_k)_{k \in T}$ with $\tilde{\alpha}_k = \alpha_k^* + \sum_{j \notin T} \mu_{kj} \alpha_j^*$.

Suppose that $\tilde{X}a + \tilde{Z}b = 0$. Then since $\tilde{X}^t\tilde{Z} = (X_j)_{j \in S}^t P_{Z^\perp} (Z_k)_{k \in T} = 0$, we have $|\tilde{X}a|^2 = a^t\tilde{X}^t\tilde{X}a = 0 \Rightarrow a = 0 \Rightarrow \tilde{Z}b = 0 \Rightarrow b = 0$. (i) thus follows from (5.15)–(5.17).

Since $z_i = (z_{ik})_{1 \le k \le m} = z_{*j} = (z_{*jk})_{1 \le k \le m}$, $i \in S_j$, we have $\tilde{z}_i = (z_{ik})_{k \in T} = (z_{*jk})_{k \in T} \equiv \tilde{z}_{*j}$, $i \in S_j$. □

The proof of Theorem 2.3 is fairly long; therefore we divide it by lemmas.

LEMMA 5.1. *Let $f(x) = f(x_1, \dots, x_s)$ be a differentiable function, where $x_i = (x_{ij})_{1 \le j \le n_i}$, $1 \le i \le s$; $R = \{x \in R^n\colon |x_i - x_{0i}| \le \delta_i,\ 1 \le i \le s\}$, where $n = \sum_{i=1}^{s} n_i$, $\delta_i > 0$, $1 \le i \le s$. Let $x^* \in R$ such that $f(x^*) = \max_{x \in R} f(x)$. Then, $x^* \notin \partial R_i$ provided that*

$$(x_i - x_{0i})^t \frac{\partial f}{\partial x_i} < 0, \qquad x \in \partial R_i, \tag{5.18}$$

where $\partial f / \partial x_i = (\partial f / \partial x_{ij})_{1 \le j \le n_i}$, and $\partial R_i = \{x \in R^n\colon |x_i - x_{0i}| = \delta_i, |x_{i'} - x_{0i'}| \le \delta_{i'},\ i' \ne i\}$.

PROOF. For any $x \in \partial R_i$, define the function $x(u)$: $[0,1] \to R^s$ as follows: $x(u)_i = x_{0i} + u(x_i - x_{0i})$, $x(u)_{i'} = x_{i'}$, $i' \neq i$. Let $\varphi(u) = f(x(u))$; then $\varphi'(u) = \sum_{j=1}^{n_i}(\partial f/\partial x_{ij})(x(u))(x_{ij} - x_{0ij})$, hence $\varphi'(1) = (x_i - x_{0i})^t(\partial f/\partial x_i)(x) < 0$. Therefore, there is $0 < u < 1$ such that $f(x(u)) = \varphi(u) > \varphi(1) = f(x)$, and $|x(u)_i - x_{0i}| = u|x_i - x_{0i}| < \delta_i$, $|x(u)_{i'} - x_{0i'}| = |x_{i'} - x_{0i'}| \leq \delta_{i'}$, $i' \neq i$. Therefore, $x(u) \in R \setminus \partial R_i$, hence $x^* \notin \partial R_i$. □

Given $0 < \varepsilon < 1$, let $R_\varepsilon = \{\psi\colon \|\tilde{\alpha} - \tilde{\alpha}_0\| \leq \varepsilon,\ |\varphi - \varphi_0| \leq \varepsilon^2\}$, $\partial R_{\varepsilon,0} = \{\psi\colon \|\tilde{\alpha} - \tilde{\alpha}_0\| \leq \varepsilon,\ |\varphi - \varphi_0| = \varepsilon^2\}$, $\partial R_{\varepsilon,k} = \{\psi\colon |\tilde{\alpha}_k - \tilde{\alpha}_{0k}| = \varepsilon,\ |\tilde{\alpha}_l - \tilde{\alpha}_{0l}| \leq \varepsilon,\ l \neq k,\ |\varphi - \varphi_0| \leq \varepsilon^2\}$, $1 \leq k \leq n$.

LEMMA 5.2. *Under the conditions of Theorem* 2.3 *we have that whenever* $\lambda_M > \lambda_{\max}(G^{-1/2}A_2G^{-1/2}) + d$,

$$\begin{aligned}&\sup_{\psi\in\partial R_{\varepsilon,0}}\{l_C(\psi) - l_C(\psi_0)\}\\ &\qquad\leq \varepsilon\sqrt{Mn}\,O_P(1) + \varepsilon^2\sqrt{M}\,O_P(1)\\ &\qquad\quad -(\varepsilon^2/2)M\big(\lambda_M - \lambda_{\max}(G^{-1/2}A_2G^{-1/2}) - d\big),\end{aligned}\tag{5.19}$$

where $d = \max_{1\leq j\leq M}\sup_{\|\psi-\psi_0\|\leq\varepsilon}\|H_j(\psi) - H_j(\psi_0)\|_R$, *and the* $O_P(1)$*'s do not depend on* ε.

PROOF. For any $\psi \in \partial R_{\varepsilon,0}$, we have

$$\begin{aligned}l_C(\psi) - l_C(\psi_0) &= \left(\frac{\partial l_C}{\partial\tilde{\alpha}}\bigg|_{\psi_0}\right)^t(\tilde{\alpha} - \tilde{\alpha}_0)\\ &\quad+\left(\frac{\partial l_C}{\partial\varphi}\bigg|_{\psi_0}\right)^t(\varphi - \varphi_0)\\ &\quad+\frac{1}{2}(\psi - \psi_0)^t\frac{\partial^2 l_C}{\partial\psi^2}\bigg|_{\psi_*}(\psi - \psi_0) = I_{11} + I_{12} + \frac{1}{2}I_2,\end{aligned}\tag{5.20}$$

where $\psi_* = \psi_0 + t(\psi - \psi_0)$ for some $0 \leq t \leq 1$.

Given ψ, $\partial l_{C,j}/\partial\psi = (\partial/\partial\psi)\log f(y^{(j)}|\psi)$, $1 \leq j \leq M$ are independent with, by (i), $E(\partial l_{C,j}/\partial\psi|\psi) = 0$. Therefore,

$$\begin{aligned}E\left(\left|\frac{\partial l_C}{\partial\tilde{\alpha}}\bigg|_{\psi_0}\right|^2\right) &= \sum_{k=1}^n E\left(E\left[\left(\frac{\partial l_C}{\partial\tilde{\alpha}_k}\bigg|_{\psi_0}\right)^2\bigg|\psi_0\right]\right)\\ &= \sum_{k=1}^n E\left(\sum_{j=1}^M E\left[\left(\frac{\partial l_{C,j}}{\partial\tilde{\alpha}_k}\bigg|_{\psi_0}\right)^2\bigg|\psi_0\right]\right)\\ &= \sum_{j=1}^M|\tilde{z}_{*j}|^2E\left(\frac{\partial h_j}{\partial s_1}\bigg|_{s_0}\right)^2 \leq KM,\end{aligned}$$

where K is given by (2.20). Thus,

$$|I_{11}| \le \left| \frac{\partial l_C}{\partial \tilde{\alpha}} \bigg|_{\psi_0} \right| |\tilde{\alpha} - \tilde{\alpha}_0| \le \varepsilon \sqrt{Mn}\, O_P(1). \tag{5.21}$$

Similarly, one can show

$$E\left(\left| \frac{\partial l_C}{\partial \varphi} \bigg|_{\psi_0} \right|^2 \right) \le (r+1) KM.$$

Thus,

$$|I_{12}| \le \left| \frac{\partial l_C}{\partial \varphi} \bigg|_{\psi_0} \right| |\varphi - \varphi_0| = \varepsilon^2 \sqrt{M}\, O_P(1). \tag{5.22}$$

On the other hand, we have $\partial^2 l_{C,j} / \partial \psi^2 = C_j H_j(\psi) C_j^t$, where $C_j = \left(\begin{smallmatrix} \tilde{z}_{*j} & 0 \\ 0 & I_{r+1} \end{smallmatrix}\right)$ and $H_j(\psi)$ is defined by (2.16). Thus,

$$\begin{aligned} \frac{\partial^2 l_{C,j}}{\partial \psi^2} \bigg|_{\psi_*} &= \sum_{j=1}^M C_j E\big(H_j(\psi_0) | \psi_0\big) C_j^t + \sum_{j=1}^M C_j\big(H_j(\psi_0) - E\big(H_j(\psi_0)|\psi_0\big)\big) C_j^t \\ &\quad + \sum_{j=1}^M C_j\big(H_j(\psi_*) - H_j(\psi_0)\big) C_j^t \\ &= A_1 + A_2 + A_3. \end{aligned} \tag{5.23}$$

It is easy to show, by (i), that

$$C_j E\left(\frac{\partial^2 h_j}{\partial s^2} \bigg|_{s_0} \bigg| \psi_0 \right) C_j^t = -C_j \operatorname{Var}\left(\frac{\partial h_j}{\partial s} \bigg|_{s_0} \bigg| \psi_0 \right) C_j^t. \tag{5.24}$$

It follows that

$$A_1 \le -\lambda_M(\psi_0) \sum_{j=1}^M \begin{pmatrix} \tilde{z}_{*j} \tilde{z}^t{}_{*j} & 0 \\ 0 & I_{r+1} \end{pmatrix} = -\lambda_M G. \tag{5.25}$$

In addition, we have

$$A_2 \le \lambda_{\max}\big(G^{-1/2} A_2 G^{-1/2}\big) G, \tag{5.26}$$

$$A_3 \le \sum_{j=1}^M \lambda_{\max}\big(H_j(\psi_*) - H_j(\psi_0)\big) C_j C_j^t \le dG. \tag{5.27}$$

Note that $\lambda_{\max}(A) \le \|A\|_R$ for any symmetric matrix A. (5.19) thus follows easily from (5.20)–(5.23) and (5.25)–(5.27). □

LEMMA 5.3. *Suppose $X_1, \dots, X_n$ are independent with $EX_i = 0$ and $|X_i| \le B$, $1 \le i \le n$, where $B > 0$. Then for any $0 < \varepsilon \le 1/B$ and $a_i \ge EX_i^2$, $1 \le i \le n$,*

$$P\left(\left| \sum_{i=1}^n X_i \right| > \varepsilon A \right) \le 2 \exp\left(-\left(\frac{\varepsilon}{2} \right)^2 A \right),$$

where $A = \sum_{i=1}^n a_i$.

The proof is straightforward [e.g., Stout (1974), Lemma 5.4.1].

COROLLARY 5.1. *Suppose $X_1, \dots, X_n$ are independent with $EX_i = 0$ and $EX_i^2 < \infty$, $1 \le i \le n$. Then for any $\varepsilon > 0$ and $a_i \ge EX_i^2$, $1 \le i \le n$,*

$$P\left(\left|\sum_{i=1}^{n} X_i\right| > 2\varepsilon A\right) \le 2\left[\exp\left(-\left(\frac{\varepsilon}{2}\right)^2 A\right) + \frac{1}{\varepsilon A}\sum_{i=1}^{n} E|X_i| 1_{(|X_i| > 1/2\varepsilon)}\right],$$

where $A = \sum_{i=1}^{n} a_i$.

PROOF. We have $X_i = U_i + V_i$, where $U_i = X_i 1_{(|X_i| \le 1/2\varepsilon)} - EX_i 1_{(\dots)}$, $V_i = X_i - U_i$. By Lemma 5.3, $P(|\sum_{i=1}^{n} U_i| > \varepsilon A) \le 2\exp(-(\varepsilon/2)^2 A)$. By Chebyshev inequality, $P(|\sum_{i=1}^{n} V_i| > \varepsilon A) \le (2/\varepsilon A)\sum_{i=1}^{n} E|X_i| 1_{(|X_i| > 1/2\varepsilon)}$. The result thus follows. □

LEMMA 5.4. *Under the conditions of Theorem 2.3 there is a set S_ε with*

$$\begin{aligned} P(S_\varepsilon^c|\psi_0) \le 2n \sum_{l=1}^{2} \Bigg\{ &\exp\left(-\frac{B_{3-l}}{4}\varepsilon^{2l} \min_{1\le k\le n} s_{M,k}^{(6-2l)}\right) \\ &+ \left[\max_{1\le k\le n}\left(\frac{s_{M,k}^{(3-l)}}{s_{M,k}^{(6-2l)}}\right)\right]\left(\frac{\max_{1\le k\le n} V_k^{(3-l)}(\varepsilon^l)}{\varepsilon^l B_{3-l}}\right)\Bigg\}, \end{aligned} \tag{5.28}$$

where $B_1 = 1 \vee \max_{1\le j\le M}|E(\xi_j^{(2)}(\psi_0)|\psi_0)|$ and $B_2 = 1 \vee \max_{1\le j\le M}$ $\mathrm{var}(\xi_j^{(2)}(\psi_0)|\psi_0)$ such that on S_ε,

$$\begin{aligned} &\max_{1\le k\le n}\left\{|\tilde{Z}_{*k}|^{-2} \sup_{\psi\in\partial R_{\varepsilon,k}}\left[(\tilde{\alpha}_k - \tilde{\alpha}_{0k})\frac{\partial l_C}{\partial\tilde{\alpha}_k}\right]\right\} \\ &\quad \le \varepsilon^2\Bigg\{-\lambda_M + d_1 + b_1 \max_{1\le k\le n}\left(\frac{t_{M,k}}{s_{M,k}^{(2)}}\right) \\ &\qquad + \varepsilon\Bigg[2B_1 + 2B_2 \max_{1\le k\le n}\left(\frac{s_{M,k}^{(4)}}{s_{M,k}^{(2)}}\right) \\ &\qquad\qquad + (r+1)b_2 \max_{1\le k\le n}\left(\frac{s_{M,k}^{(1)}}{s_{M,k}^{(2)}}\right)\Bigg]\Bigg\}, \end{aligned} \tag{5.29}$$

where

$$b_1 = \max_{1\le j\le M} \sup_{\|\psi-\psi_0\|\le\varepsilon} |\xi_j^{(2)}(\psi)|,$$

$$b_2 = \max_{1\le j\le M} \max_{2\le u\le r+2} \sup_{\|\psi-\psi_0\|\le\varepsilon} \left|\frac{\partial^2 h_j}{\partial s_1\,\partial s_u}\Bigg|_{(\tilde{z}^t_{*j}\tilde{\alpha},\,\tilde{\beta},\,\tau)}\right|,$$

and

$$d_1 = \max_{1\le j\le M} \sup_{\|\psi-\psi_0\|\le\varepsilon} |\xi_j^{(2)}(\psi) - \xi_j^{(2)}(\psi_0)|.$$

PROOF. For any $1 \le k \le n$ and $\psi \in R_{\varepsilon, k}$ we have

$$
\begin{aligned}
(\tilde{\alpha}_k - \tilde{\alpha}_{0k}) \frac{\partial l_C}{\partial \tilde{\alpha}_k} &= (\tilde{\alpha}_k - \tilde{\alpha}_{0k}) \frac{\partial l_C}{\partial \tilde{\alpha}_k}\bigg|_{\psi_0} + (\tilde{\alpha}_k - \tilde{\alpha}_{0k})^2 \frac{\partial^2 l_C}{\partial \tilde{\alpha}_k^2}\bigg|_{\psi(k)} \\
&\quad + \sum_{l \ne k} (\tilde{\alpha}_k - \tilde{\alpha}_{0k})(\tilde{\alpha}_l - \tilde{\alpha}_{0l}) \frac{\partial^2 l_C}{\partial \tilde{\alpha}_k \, \partial \tilde{\alpha}_l}\bigg|_{\psi(k)} \\
&\quad + (\tilde{\alpha}_k - \tilde{\alpha}_{0k}) \frac{\partial^2 l_C}{\partial \tilde{\alpha}_k \, \partial \varphi}\bigg|_{\psi_{(k)}} (\varphi - \varphi_0) \\
&= I_{k1} + I_{k2} + I_{k3} + I_{k4},
\end{aligned}
\tag{5.30}
$$

where $\psi_{(k)} = \psi_0 + t_k(\psi - \psi_0)$ for some $0 \le t_k \le 1$.

It follows from (5.24) that $\operatorname{var}(\xi_j^{(1)}(\psi_0)|\psi_0)\tilde{z}^2{}_{*jk} = -E(\xi_j^{(2)}(\psi_0)|\psi_0)\tilde{z}^2{}_{*jk}$, $1 \le k \le n$. Thus,

$$
E\left[\left(\frac{\partial l_{C,j}}{\partial \tilde{\alpha}_k}\bigg|_{\psi_0}\right)^2 \Bigg| \psi_0\right] = \operatorname{var}\bigl(\xi_j^{(1)}(\psi_0)|\psi_0\bigr)\tilde{z}^2{}_{*jk} \le B_1 \tilde{z}^2{}_{*jk}.
$$

Let $S_{\varepsilon,1} = \bigcap_{k=1}^n \{|(\partial l_C/\partial \tilde{\alpha}_k)(\psi_0)| \le 2\varepsilon^2 B_1 s_{M,k}^{(2)}\}$. Then, on $S_{\varepsilon,1}$,

$$
|I_{k1}| \le 2\varepsilon^3 B_1 s_{M,k}^{(2)}
\tag{5.31}
$$

and by Corollary 5.1,

$$
\begin{aligned}
P(S_{\varepsilon,1}^c|\psi_0) &\le \sum_{k=1}^n P\left(\left|\frac{\partial l_C}{\partial \tilde{\alpha}_k}\bigg|_{\psi_0}\right| > 2\varepsilon^2 B_1 s_{M,k}^{(2)} \Bigg| \psi_0\right) \\
&\le 2 \sum_{k=1}^n \Biggl[\exp\left(-\frac{\varepsilon^4}{4} B_1 s_{M,k}^{(2)}\right) \\
&\qquad + \frac{1}{\varepsilon^2 B_1 s_{M,k}^{(2)}} \sum_{j=1}^M |\tilde{z}_{*jk}| E\Bigl(|\xi_j^{(1)}(\psi_0)| 1_{(|\tilde{z}_{*jk}\xi_j^{(1)}(\psi_0)| > 1/2\varepsilon^2)} |\psi_0\Bigr)\Biggr] \\
&\le 2 \sum_{k=1}^n \left[\exp\left(-\frac{\varepsilon^4}{4} B_1 s_{M,k}^{(2)}\right) + \left(\frac{V_k^{(1)}(\varepsilon^2)}{\varepsilon^2 B_1}\right)\left(\frac{s_{M,k}^{(1)}}{S_{M,k}^{(2)}}\right)\right].
\end{aligned}
\tag{5.32}
$$

Next, we have

$$
E\left(\frac{\partial^2 l_C}{\partial \tilde{\alpha}_k^2}\bigg|_{\psi_0} \Bigg| \psi_0\right) = -\sum_{j=1}^M \tilde{z}^2{}_{*jk} \operatorname{var}\bigl(\xi_j^{(1)}(\psi_0)|\psi_0\bigr) \le -\lambda_M s_{M,k}^{(2)}
$$

and

$$
\operatorname{var}\left(\frac{\partial^2 l_{C,j}}{\partial \tilde{\alpha}_k^2}\bigg|_{\psi_0} \Bigg| \psi_0\right) \le B_2 \tilde{z}^4{}_{*jk}.
$$

Let $S_{\varepsilon,2} = \bigcap_{k=1}^{n}\{|(\partial^2 l_C/\partial\tilde{\alpha}_k^2)(\psi_0) - E((\partial^2 l_C/\partial\tilde{\alpha}_k^2)(\psi_0)|\psi_0)| \le 2\varepsilon B_2 s_{M,k}^{(4)}\}$. Since

$$\left|\frac{\partial^2 l_C}{\partial\tilde{\alpha}_k^2}\bigg|_{\psi_{(k)}} - \frac{\partial^2 l_C}{\partial\tilde{\alpha}_k^2}\bigg|_{\psi_0}\right| \le \sum_{j=1}^{M} \tilde{z}^2{}_{*jk}|\xi_j^{(2)}(\psi_{(k)}) - \xi_j^{(2)}(\psi_0)| \le d_1 s_{M,k}^{(2)},$$

we have, on $S_{\varepsilon,2}$, that

$$I_{k2} \le \varepsilon^2\big(-\lambda_M s_{M,k}^{(2)} + 2\varepsilon B_2 s_{M,k}^{(4)} + d_1 s_{M,k}^{(2)}\big) \tag{5.33}$$

and by Corollary 5.1,

$$\begin{aligned}
P(S_{\varepsilon,2}^c|\psi_0) &\le \sum_{k=1}^{n} P\left(\left|\frac{\partial^2 l_C}{\partial\tilde{\alpha}_k^2}\bigg|_{\psi_0} - E\left(\frac{\partial^2 l_C}{\partial\tilde{\alpha}_k^2}\bigg|_{\psi_0}\bigg|\psi_0\right)\right| > 2\varepsilon B_2 s_{M,k}^{(4)}\bigg|\psi_0\right)\\
&\le 2\sum_{k=1}^{n}\Bigg[\exp\left(-\frac{\varepsilon^2}{4}B_2 s_{M,k}^{(4)}\right) + \frac{1}{\varepsilon B_2 s_{M,k}^{(4)}}\sum_{j=1}^{M}\tilde{z}^2{}_{*jk}E\Big(|\xi_j^{(2)}(\psi_0)\\
&\qquad\qquad -E\big(\xi_j^{(2)}(\psi_0)|\psi_0\big)|1_{(\tilde{z}^2{}_{*jk}|\cdots - \cdots|>1/2\varepsilon)}|\psi_0\Big)\Bigg]\\
&\le 2\sum_{k=1}^{n}\left[\exp\left(-\frac{\varepsilon^2}{4}B_2 s_{M,k}^{(4)}\right) + \left(\frac{V_k^{(2)}(\varepsilon)}{\varepsilon B_2}\right)\left(\frac{s_{M,k}^{(2)}}{s_{M,k}^{(4)}}\right)\right].
\end{aligned} \tag{5.34}$$

Also, we have

$$|I_{k3}| \le \sum_{l\ne k}\big|(\tilde{\alpha}_k - \tilde{\alpha}_{0k})(\tilde{\alpha}_l - \tilde{\alpha}_{0l})\big|\sum_{j=1}^{M}\left|\frac{\partial^2 l_{C,j}}{\partial\tilde{\alpha}_k\,\partial\tilde{\alpha}_l}\bigg|_{\psi_{(k)}}\right| \le \varepsilon^2 b_1 t_{M,k}, \tag{5.35}$$

$$\begin{aligned}
|I_{k4}| &\le \varepsilon^2\left(\sum_{u=1}^{r}\sum_{j=1}^{M}\left|\frac{\partial^2 l_{C,j}}{\partial\tilde{\alpha}_k\,\partial\tilde{\beta}_u}\bigg|_{\psi_{(k)}}\right| + \sum_{j=1}^{M}\left|\frac{\partial^2 l_{C,j}}{\partial\tilde{\alpha}_k\,\partial\tau}\bigg|_{\psi_{(k)}}\right|\right)\\
&\le \varepsilon^3(r+1)b_2 s_{M,k}^{(1)}.
\end{aligned} \tag{5.36}$$

Inequalities (5.28) and (5.29) then follow by combining (5.30)–(5.36) and letting $S_\varepsilon = S_{\varepsilon,1}\cap S_{\varepsilon,2}$. □

PROOF OF THEOREM 2.3. (iv) implies that there is a sequence δ_M such that $0 < \delta_M \le 1$, $\delta_M \to 0$ and

$$\begin{aligned}
&\left(\rho_M^{-2}\sqrt{\frac{n}{M}}\right) \vee \max_{l=1,2}\left\{\left(\frac{\log n}{\rho_M^{2l}\,\min_{1\le k\le n} s_{M,k}^{(6-2l)}}\right)^{1/2l}\right.\\
&\qquad\qquad \left.\vee\left(\frac{n\max_{1\le k\le n} V_k^{(3-l)}(\rho_M^l)}{\rho_M^l}\right)^{1/l}\right\}\bigg/\delta_M \to_P 0.
\end{aligned}$$

Let $\varepsilon_M = \delta_M \rho_M$. Then we have

$$(\varepsilon_M \delta_M)^{-1}(n/M)^{1/2} = \left(\rho_M^{-2}(n/M)^{1/2}\right)/\delta_M \to_P 0,$$

$$\log n/\varepsilon_M^{2l} \min_{1\le k\le n} s_{M,k}^{(6-2l)} = \left(\log n/\rho_M^{2l} \min_{1\le k\le n} s_{M,k}^{(6-2l)}\right)\Big/\delta_M^{2l} \to_P 0,$$

$$n \max_{1\le k\le n} V_k^{(3-l)}(\varepsilon_M^l)/\varepsilon_M^l \le \left(n \max_{1\le k\le n} V_k^{(3-l)}(\rho_M^l)/\rho_M^l\right)\Big/\delta_M^l \to_P 0.$$

Replace ε in Lemmas 5.2 and 5.4 by ε_M. (ii) implies that there is a constant C such that $B_1 \vee B_2 \vee b_1 \vee b_2 \le C$ when $\varepsilon_M \le \delta$. (ii) and (iii) imply that there is $D > 0$ such that $d \le D\varepsilon_M$ when $\varepsilon_M \le \delta$. Thus $d_1/\rho_M \le d/\rho_M \le D\delta_M \to 0$. (Note that b_1, b_2, d and d_1 all depend on ε_M.)

By Lemma 5.2, on $S_M^{(0)} = \{\lambda_M > \lambda_{\max}(G^{-1/2}A_2G^{-1/2}) + d\}$, we have

$$\begin{aligned}
&\sup_{\psi\in \partial R_{\varepsilon_M,0}} \{l_C(\psi) - l_C(\psi_0)\}\\
&\quad\le \varepsilon_M^2 \rho_M M\Bigg\{-\frac{1}{2}\left(1 - \frac{\lambda_{\max}(G^{-1/2}A_2G^{-1/2})}{\rho_M} - \frac{d}{\rho_M}\right)\\
&\qquad\qquad + 2\left(\sqrt{\frac{n}{M}}\Big/\varepsilon_M\rho_M\right)O_P(1)\Bigg\}\\
&\quad\le \varepsilon_M^2 \rho_M M\left(-\frac{1}{2} + o_P(1)\right).
\end{aligned}\tag{5.37}$$

By Lemma 5.4, on $S_M^{(1)} = S_{\varepsilon_M}$, we have

$$\begin{aligned}
&\max_{1\le k\le n}\left\{|\tilde{Z}_{*k}|^{-2} \sup_{\psi\in \partial R_{\varepsilon_M,k}}\left\{(\tilde{\alpha}_k - \tilde{\alpha}_{0k})\frac{\partial l_C}{\partial\tilde{\alpha}_k}\right\}\right\}\\
&\quad\le \varepsilon_M^2\rho_M\Bigg\{-1 + \frac{d_1}{\rho_M} + b_1\left[\max_{1\le k\le n}\left(\frac{t_{M,k}}{s_{M,k}^{(2)}}\right)\Big/\rho_M\right]\\
&\qquad + \delta_M\Bigg[2B_1 + 2B_2\max_{1\le k\le n}\left(\frac{s_{M,k}^{(4)}}{s_{M,k}^{(2)}}\right)\\
&\qquad\qquad + (r+1)b_2\max_{1\le k\le n}\left(\frac{s_{M,k}^{(1)}}{s_{M,k}^{(2)}}\right)\Bigg]\Bigg\}\\
&\quad= \varepsilon_M^2\rho_M(-1 + o_P(1)).
\end{aligned}\tag{5.38}$$

Finally, we have

$$1_{(S_M^{(0)})^c} \le 1_{((\lambda_{\max}(G^{-1/2}A_2G^{-1/2})/\rho_M)+(d/\rho_M)>1)} \to_P 0$$

and

$$P\Big(\big(S_M^{(1)}\big)^c|\psi_0\Big) \le 4\bigg\{\exp\bigg[\frac{\log n}{o_P(1)}\Big(-\frac{1}{4}+o_P(1)\Big)\bigg] + O_P(1)\bigg(\frac{n\max_{1\le k\le n}V_k^{(3-l)}\big(\varepsilon_M^l\big)}{\varepsilon_M^l}\bigg)\bigg\} \to_P\ 0.$$

Therefore, by the dominated convergence theorem, $P(S_M^c)\to 0$, where $S_M = S_M^{(0)}\cap S_M^{(1)}$.

Let $\hat{\psi}$ be the maximizer of l_C over R_{ε_M}. Then by Lemma 5.1, $\{\hat{\psi}\notin\partial R_{\varepsilon_M}\}\supset S_M\cap\{o_P(1)<1/2\}\cap\{o_P(1)<1\}\cap\{\delta_M\le\delta\}$, where the two $o_P(1)$'s are as in (5.37) and (5.38), respectively. Thus $P(\hat{\psi}\in R^o_{\varepsilon_M})\to 1$. Note that $\hat{\psi}$ satisfies $(\partial l_C/\partial\psi)(\hat{\psi})=0$. □

PROOF OF LEMMA 3.1. Suppose $\xi\sim N(0,\sigma^2)$; then it is easy to show that

$$P(|\xi|>\lambda)\le 2\sqrt{\frac{2}{\pi}}\Big(\frac{\sigma}{\lambda}\Big)\exp\bigg\{-\frac{1}{2}\Big(\frac{\lambda}{\sigma}\Big)^2\bigg\},\qquad \lambda>0. \tag{5.39}$$

Let $\sigma^2=\max_{1\le k\le m}\sigma_k^2$. By (5.39),

$$\begin{aligned}P(\|\alpha\|>a_N) &\le \sum_{k=1}^m P(|\alpha_k|>a_N)\\ &\le 2\sqrt{\frac{2}{\pi}}\,m\Big(\frac{\sigma}{a_N}\Big)\exp\bigg\{-\frac{1}{2}\frac{a_N^2}{\sigma^2}\bigg\}\\ &= 2\sqrt{\frac{2}{\pi}}\exp\bigg\{-\frac{1}{2}\frac{a_N^2}{\sigma^2}\bigg[1+\frac{\sigma^2}{a_N^2}\log\Big(\frac{a_N^2}{\sigma^2}\Big)-2\sigma^2\Big(\frac{\log m}{a_N^2}\Big)\bigg]\bigg\}\to 0.\end{aligned}$$

□

PROOF OF LEMMA 3.2. It is easy to see that $E((\partial h/\partial\theta)(\theta,Y))=0$. We shall first show that $\lambda(\theta)>0$, $\forall\ \theta$. Suppose that this is not true. Then there is λ such that $E\exp(u_k(\xi))v_k(\theta,\xi)\xi=\lambda E\exp(u_k(\xi))v_k(\theta,\xi)$, $0\le k\le r$, where $u_k(\xi)=(\mu+\tau\xi)k-r\log(1+\exp(\mu+\tau\xi))$, $v_k(\theta,\xi)=k-r\exp(\mu+\tau\xi)/(1+\exp(\mu+\tau\xi))$. By integration by parts we have

$$\begin{aligned}E\exp(u_k(\xi))v_k(\theta,\xi)&=\tau^{-1}E\varphi_k(\theta,\xi)\xi,\\ E\exp(u_k(\xi))v_k(\theta,\xi)\xi&=\tau^{-1}E\varphi_k(\theta,\xi)(\xi^2-1),\end{aligned} \tag{5.40}$$

where $\varphi_k(\theta,\xi)=(\exp(\mu+\tau\xi))^k/(1+\exp(\mu+\tau\xi))^r$. Therefore we have

$$E\varphi_k(\theta,\xi)(\xi^2-\lambda\xi-1)=0,\qquad k=0,1,\dots,r. \tag{5.41}$$

Let $\lambda_1 = (1/2)(\lambda - \sqrt{\lambda^2 + 4})$, $\lambda_2 = (1/2)(\lambda + \sqrt{\lambda^2 + 4})$. By (5.41) and the fact that $r \ge 2$ we have

$$\begin{aligned}
0 &= E\frac{\xi^2 - \lambda\xi - 1}{(1 + \exp(\mu + \tau\xi))^r}(\exp(\mu + \tau\xi))^2 \\
&\quad - \left(\sum_{j=1}^{2} \exp(\mu + \tau\lambda_j)\right) E\frac{\xi^2 - \lambda\xi - 1}{(1 + \exp(\mu + \tau\xi))^r}\exp(\mu + \tau\xi) \\
&\quad + \left(\prod_{j=1}^{2} \exp(\mu + \tau\lambda_j)\right) E\frac{\xi^2 - \lambda\xi - 1}{(1 + \exp(\mu + \tau\xi))^r} \\
&= E(1 + \exp(\mu + \tau\xi))^{-r} \prod_{j=1}^{2} (\xi - \lambda_j)\big(\exp(\mu + \tau\xi) - \exp(\mu + \tau\lambda_j)\big) > 0,
\end{aligned}$$

which is a contradiction.

Let $g(\theta, k) = E\varphi_k(\theta, \xi)$. Let $u, v \in R$ such that $u^2 + v^2 = 1$. By (5.40) it is easy to show that

$$\begin{aligned}
&\binom{u}{v}^t \operatorname{Var}\left(\frac{\partial h}{\partial\theta}(\theta, Y)\right)\binom{u}{v} \\
&\qquad \ge \tau^{-2} \min_{0 \le k \le r}\left\{\binom{r}{k}\Big/ g(\theta, k)\right\} \sum_{k=0}^{r} \big(E\varphi_k(\theta, \xi)(v\xi^2 + u\xi - v)\big)^2.
\end{aligned} \tag{5.42}$$

On the other hand, it is easy to show that there is a constant C such that

$$\begin{aligned}
&\sum_{k=0}^{r} \big(E\varphi_k(\theta, \xi)(v\xi^2 + u\xi - v)\big)^2 \\
&\qquad \ge \sum_{k=0}^{r} \big(E\varphi_k(\theta, \xi)(v\xi^2 + u\xi - v)1_{(|\xi| \le |\mu|/\tau)}\big)^2 \\
&\qquad\quad -(r + 1)C\exp\left(-\frac{\mu^2}{4\tau^2}\right).
\end{aligned} \tag{5.43}$$

Suppose that $|\mu| > \tau$ and $v \ne 0$. Let $\xi_1 = \lambda_1 \vee (-|\mu|/\tau)$, $\xi_2 = \lambda_2 \wedge (|\mu|/\tau)$. By the fact that for $|\xi| \le |\mu|/\tau$, $\prod_{j=1}^{2}(\xi - \lambda_j)(\exp(\tau\xi) - \exp(\tau\xi_j)) \ge 0$ and continuity, it is easy to show that

$$\begin{aligned}
&\bigg| E(1 + \exp(\mu + \tau\xi))^{-r}(v\xi^2 + u\xi - v) \\
&\qquad \times \prod_{j=1}^{2} \big(\exp(\mu + \tau\xi) - \exp(\mu + \tau\xi_j)\big)1_{(|\xi| \le |\mu|/\tau)}\bigg| \\
&\qquad = |v|e^{2\mu}E\bigg\{(1 + \exp(\mu + \tau\xi))^{-r} \\
&\qquad\qquad \times \prod_{j=1}^{2} (\xi - \lambda_j)\big(\exp(\tau\xi) - \exp(\tau\xi_j)\big)1_{(|\xi| \le |\mu|/\tau)}\bigg\} \\
&\qquad \ge 2^{-r}\eta\exp(-2(r + 1)|\mu|),
\end{aligned} \tag{5.44}$$

for some constant $\eta > 0$. On the other hand, let

$$\delta = \max_{0 \le k \le 2} |E\varphi_k(\theta, \xi)(v\xi^2 + u\xi - v)1_{(|\xi| \le |\mu|/\tau)}|.$$

Then, the lhs of (5.44) equals

$$\begin{aligned} &\Bigg| E\varphi_2(\theta, \xi)(v\xi^2 + u\xi - v)1_{(|\xi| \le |\mu|/\tau)} \\ &\quad - \left(\sum_{j=1}^{2} \exp(\mu + \tau\xi_j)\right) E\varphi_1(\theta, \xi) \cdots \\ &\quad + \left(\prod_{j=1}^{2} \exp(\mu + \tau\xi_j)\right) E\varphi_0(\theta, \xi) \cdots \Bigg| \\ &\quad \le 4\delta e^{3|\mu|}. \end{aligned} \tag{5.45}$$

Combining (5.42)–(5.45) and the fact that $g(\theta, k) \le 1$, we have that, when $|\mu| > \tau$,

$$\begin{aligned} &\binom{u}{v}^t \operatorname{Var}\left(\frac{\partial h}{\partial \theta}(\theta, Y)\right)\binom{u}{v} \\ &\quad \ge \frac{1}{\tau^2}\Bigg[2^{-2(r+2)}\eta^2 \exp(-2(2r+5)|\mu|) \\ &\qquad - (r+1)C \exp\left(-\frac{\mu^2}{4\tau^2}\right)\Bigg] \end{aligned} \tag{5.46}$$

for all u, v such that $u^2 + v^2 = 1$ and $v \ne 0$. By continuity, (5.46) holds even if $v = 0$ (and $u^2 = 1$). The rest of the proof is easy to complete. □

PROOF OF LEMMA 3.3. It is easy to show that for any integers m, n, $l \ge 0$ and $b > 0$,

$$\begin{aligned} \sup_{0 \le \tau \le b} \Bigg| &\exp((n - m)\mu) E\left\{\frac{(\exp(\mu + \tau\xi))^m}{(1 + \exp(\mu + \tau\xi))^n}|\xi|^l\right\} \\ &- E\{\exp((m - n)\tau\xi)|\xi|^l\}\Bigg| \to 0, \end{aligned}$$

as $\mu \to \infty$, and

$$\sup_{0 \le \tau \le b} \left|\exp(-m\mu) E\left\{\frac{(\exp(\mu + \tau\xi))^m}{(1 + \exp(\mu + \tau\xi))^n}|\xi|^l\right\} - E\{\exp(m\tau\xi)|\xi|^l\}\right| \to 0,$$

as $\mu \to -\infty$. It is then easy to demonstrate that for any integers k, l, $r \geq 0$, $0 \leq s \leq t$ and $b > 0$, the expression

$$\left(E\left\{ \frac{(\exp(\mu + \tau\xi))^{k}}{(1 + \exp(\mu + \tau\xi))^{r}} |\xi|^{l} \right\} \right)^{-1} E\left\{ \frac{(\exp(\mu + \tau\xi))^{k+s}}{(1 + \exp(\mu + \tau\xi))^{r+t}} |\xi|^{l} \right\} \tag{5.47}$$

is bounded for $\mu \in R$, $0 \leq \tau \leq b$. The conclusion then follows, because any derivative of $h(\theta, k)$, up to third, is a linear combination of terms of the form (5.47) or products of terms of such a form. □

Acknowledgments. The author is grateful to a Co-editor, an Associate Editor and a referee for carefully reviewing the manuscript and for their constructive and insightful comments, which helped greatly to improve the manuscript.

REFERENCES

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.* **88** 9–25.

Diggle, P. J., Liang, K. Y. and Zeger, S. L. (1996). *Analysis of Longitudinal Data*. Oxford Univ. Press.

Ghosh, M. and Rao, J. N. K. (1994). Small area estimation: an appraisal. *Statist. Sci.* **6** 15–51.

Jiang, J. (1998). Consistent estimators in generalized linear mixed models. *J. Amer. Statist. Assoc.* **93** 720–729.

Jiang, J. (1999). A nonlinear Gauss–Seidel algorithm for inference about GLMM. *Comp. Statist.* To appear.

Jiang, J., Jia, H. and Chen, H. (1999). Maximum posterior estimates of random effects in generalized linear mixed models. *Statist. Sinica.* To appear.

Karim, M. R. and Zeger, S. L. (1992). Generalized linear models with random effects; salamander mating revisited. *Biometrics* **48** 631–644.

Kuk, A. Y. C. (1995). Asymptotically unbiased estimation in generalized linear models with random effects. *J. Roy. Statist. Soc. Ser. B* **57** 395–407.

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models. *J. Roy. Statist. Soc. Ser. B* **58** 619–678.

Lin, X. and Breslow, N. E. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *J. Amer. Statist. Assoc.* **91** 1007–1016.

Malec, D., Sedransk, J., Moriarity, C. L. and LeClere, F. B. (1997). Small area inference for binary variables in the National Health Interview Survey. *J. Amer. Statist. Assoc.* **92** 815–826.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman and Hall, New York.

McCulloch, C. E. (1994). Maximum likelihood variance components estimation for binary data. *J. Amer. Statist. Assoc.* **89** 330–335.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *J. Amer. Statist. Assoc.* **92** 162–170.

McGilchrist, C. A. (1994). Estimation in generalized mixed models. *J. Roy. Statist. Soc. Ser. B* **56** 61–69.

Portnoy, S. (1984). Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. I. *Ann. Statist.* **12** 1298–1309.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika* **78** 719–727.

SEARLE, S. R., CASELLA, G. and MCCULLOCH, C. E. (1992). *Variance Components*. Wiley, New York.

SHUN, Z. and MCCULLAGH, P. (1995). Laplace approximation of high-dimensional integrals. *J. Roy. Statist. Soc. Ser. B* **57** 749–760.

STOUT, W. F. (1974). *Almost Sure Convergence*. Academic Press, New York.

VONESH, E. F. (1996). A note on the use of Laplace's approximation for nonlinear mixed-effects models. *Biometrika* **83** 447–452.

ZEGER, S. L. and KARIM, M. R. (1991). Generalized linear models with random effects: a Gibbs sampling approach. *J. Amer. Statist. Assoc.* **86** 79–86.

DEPARTMENT OF STATISTICS
CASE WESTERN RESERVE UNIVERSITY
10900 EUCLID AVENUE
CLEVELAND, OHIO 44106-7054
E-MAIL: jiang@eureka.cwru.edu