The Annals of Statistics
2003, Vol. 31, No. 1, 310–347

ASYMPTOTICS FOR GENERALIZED ESTIMATING EQUATIONS WITH LARGE CLUSTER SIZES

BY MINGE XIE[1] AND YANING YANG

Rutgers University

Generalized estimating equations are used in regression analysis of longitudinal data, where observations on each subject are correlated. Statistical analysis using such methods is based on the asymptotic properties of regression parameter estimators. This paper presents asymptotic results when either the number of independent subjects or the cluster sizes (the number of observations on each subject) or both go to infinity. A set of (information matrix based) general conditions is developed, which leads to the weak and strong consistency as well as the asymptotic normality of the estimators. Most of the results are parallel to the elegant work of Fahrmeir and Kaufmann on maximum likelihood estimators related to the generalized linear models. The conditions for weak consistency and asymptotic normality are verified for several examples of general interest.

1. Introduction. The class of generalized linear models [Nelder and Wedderburn (1972)] plays a central role in regression problems with discrete or nonnegative responses. This class of regression models was extended by Liang and Zeger (1986) to analyze longitudinal or batch correlated data. In biostatistics, the Liang and Zeger approach is known as the Generalized Estimating Equations (GEE) method [see, e.g., Diggle, Liang and Zeger (1996)]. In the past, attention has been paid to methodological development and modeling issues. Most of the work relies on the asymptotic results presented by Liang and Zeger (1986), in which exact conditions are not specified. There are some rigorous discussions of estimating equation approaches, which may be applicable to the GEE setting. For example, Crowder (1986) studied the (weak) consistency and inconsistency of the solutions of general estimating equations. Li (1996) used a minimax approach introduced by Cramér (1946) to identify a weakly consistent root of estimating equations. More recently, Yuan and Jennrich (1998) developed weak consistency and asymptotic normality conditions for estimating equations along the lines of Crowder (1986). However, the estimating equations considered in Crowder (1986), Li (1996) and Yuan and Jennrich (1998) are not particularly tailored for longitudinal data, and their asymptotics do not cover cases when the number of observations on each subject goes to infinity.

Received February 2000; revised October 2001.

[1]Supported in part by NSF Grant DMS-98-03273.

AMS 2000 subject classifications. 62F12, 62J12.

Key words and phrases. Generalized estimating equations (GEE), longitudinal data, cluster correlated observations, infinite cluster sizes.

Suppose $(y_{ij}, \mathbf{x}_{ij})$ are observations for the jth measurement on the ith subject, $j = 1, 2, \ldots, m_i$ and $i = 1, 2, \ldots, n$, where y_{ij} is a scalar response, $\mathbf{x}_{ij}$ is a $p \times 1$ covariate vector, and m_i is the cluster size. Assume that the observations on different subjects are independent and the observations on the same subject are correlated. For $i = 1, \ldots, n$, let $\mathbf{y}_i = (y_{i1}, \ldots, y_{im_i})^T$ and $\mathbf{X}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{im_i})^T$. Liang and Zeger (1986) used a generalized linear model to model the marginal density of y_{ij} (with respect to a σ-finite measure ξ),

$$f(y_{ij}|\mathbf{x}_{ij}, \beta, \phi) = \exp[\{y_{ij}\theta_{ij} - a(\theta_{ij}) + b(y_{ij})\}/\phi], \tag{1}$$

where $\theta_{ij} = u(\eta_{ij})$, u is a known injective function and $\eta_{ij} = \mathbf{x}_{ij}^T\beta$. The vector β contains the regression parameters of interest, and ϕ is a nuisance scale parameter. Under such a model specification, the first two moments of y_{ij} are given by

$$\begin{aligned} \mu_{ij}(\beta) &= \mathrm{E}(y_{ij}|\mathbf{x}_{ij}, \beta, \phi) = a'(\theta_{ij}), \\ \sigma_{ij}^2(\beta) &= \mathrm{Cov}(y_{ij}|\mathbf{x}_{ij}, \beta, \phi) = a''(\theta_{ij})\phi. \end{aligned} \tag{2}$$

Let $g(t) = (a' \circ u)^{-1}(t)$; then $g(\mu_{ij}(\beta)) = \mathbf{x}_{ij}^T\beta$. The function $g(t)$ is the link function and its inverse function $h(s) = (a' \circ u)(s)$ is called the inverse link function. Of importance are the canonical link functions, where $u(s) = s$, so $g(t) = (a')^{-1}(t)$ and $h(s) = a'(s)$.

Denote $\mu_i(\beta) = \mathrm{E}(\mathbf{y}_i|\mathbf{X}_i, \beta, \phi) = (\mu_{i1}(\beta), \ldots, \mu_{im_i}(\beta))^T$ and $\mathbf{\Sigma}_i(\beta) = \mathrm{Cov}(\mathbf{y}_i|\mathbf{X}_i, \beta, \phi)$. We write $\mathbf{A}_i(\beta) = \mathrm{diag}(\sigma_{i1}^2(\beta), \ldots, \sigma_{im_i}^2(\beta))$ and $\mathbf{\Delta}_i(\beta) = \mathrm{diag}(u'(\mathbf{x}_{i1}^T\beta), \ldots, u'(\mathbf{x}_{im_i}^T\beta))$, where, for any vector $\mathbf{v}$, $\mathrm{diag}(\mathbf{v})$ represents a diagonal matrix whose diagonal elements are the elements of $\mathbf{v}$. Let $\mathbf{D}_i(\beta) = \mathbf{A}_i(\beta)\mathbf{\Delta}_i(\beta)\mathbf{X}_i$ and $\mathbf{V}_i(\beta, \alpha) = \mathbf{A}_i^{1/2}(\beta)\mathbf{R}_i(\alpha)\mathbf{A}_i^{1/2}(\beta)$. Here $\mathbf{R}_i(\alpha)$ is the "working" correlation matrix that one can choose freely, which may possibly contain a nuisance parameter (or parameter vector) α. If $\mathbf{R}_i(\alpha)$ is equal to the true (often unspecified) correlation matrix $\bar{\mathbf{R}}_i$, then $\mathbf{V}_i(\beta_0, \alpha) = \mathbf{\Sigma}_i(\beta_0)$ at the true parameter β_0.

Liang and Zeger (1986) proposed solving the following equations:

$$\mathbf{g}_n(\beta) = \sum_{i=1}^n \mathbf{g}_{m_i,i} = \sum_{i=1}^n \mathbf{D}_i(\beta)^T \mathbf{V}_i^{-1}(\beta, \alpha)\big(\mathbf{y}_i - \mu_i(\beta)\big) = 0, \tag{3}$$

which are GEE. Note that (3) and its solution $\widehat{\beta}_n$ is "derived without specifying the joint distribution of a subject's observations" [Liang and Zeger (1986)], that is, without specifying the $\bar{\mathbf{R}}_i$'s. This is an appealing feature of the GEE approach according to Liang and Zeger (1986), since, as they pointed out, the $\bar{\mathbf{R}}_i$'s should be considered a "nuisance" in many applications and could be "difficult to specify." Because (3) depends only on the first and second moments of the marginal distributions of the individual observations, under a slightly more general GEE model setting, the density assumption (1) can be replaced by the first two moment assumptions, such as in (2); see Zeger and Liang (1986). In Liang and Zeger (1986), Zeger and Liang (1986) and subsequent literature, the asymptotic properties of the GEE estimator $\widehat{\beta}_n$ are studied under the assumption that the number of independent subjects n goes to infinity and the cluster sizes are finite with an upper bound.

In this paper, we study asymptotic properties of the GEE estimator $\widehat{\beta}_n$ under a broader setting when either the number of independent subjects n or the cluster sizes m_i go to infinity. The results, initially developed under the setting where n goes to infinity and the maximum cluster size $m = m(n) = \max_{1\le i\le n} m_i$, as a function of n, can either be bounded or go to infinity, were motivated by a consulting project in which the maximum cluster size is relatively large [see Xie and Simpson (1998) and Xie, Simpson and Carroll (2000)]. At the suggestion of the editors, we have also extended our discussion to cover the case when n is bounded but $m \to \infty$, where possible. So, in particular, we are interested in three large sample settings:

1. $n \to \infty$ and $m = m(n) = \max_{1\le i\le n} m_i$ is bounded above, for all n;
2. n is bounded but $m \to \infty$;
3. $m \to \infty$ as $n \to \infty$.

Under setting 3, in order that the GEE estimators have good large sample properties, restrictions on the speed at which the maximum cluster size m tends to infinity are usually required. For convenience, we will rewrite $\mathbf{g}_n(\beta)$ in (3) as $\mathbf{g}_{nm}(\beta)$ and $\widehat{\beta}_n$ as $\widehat{\beta}_{nm}$. Also, since m_i can possibly be a function of n (e.g., $m_i = n$), we shall treat $\mathbf{g}_{m_1,1}, \mathbf{g}_{m_2,2}, \dots, \mathbf{g}_{m_n,n}$, the summands of $\mathbf{g}_{nm}$, as a double array sequence, when $n \to \infty$.

We present in this paper a set of (unified) information matrix-based conditions which assures the weak consistency, the strong consistency and the asymptotic normality of the estimator $\widehat{\beta}_{nm}$. Most of the conditions parallel the elegant conditions presented by Fahrmeir and Kaufmann (1985) for maximum likelihood estimators in generalized linear models. Unlike many papers on estimating equations [Haberman (1977), Crowder (1986) and Yuan and Jennrich (1998)] and in the M-estimation literature [Huber (1981) and Yohai and Maronna (1979)], we do not use a fixed-point theorem to develop our consistency results. Instead, we use a lemma of Chen, Hu and Ying (1999) on injective functions, leading to simpler proofs. In addition, because $\mathbf{g}_{nm}(\beta)$ essentially contains double summations and both of them can tend to infinity, the GEE equations $\mathbf{g}_{nm}(\beta) = 0$ are not a set of M-estimation equations in the usual sense. The standard results for M-estimators are not applicable.

Let $\lambda_{\min}(\mathbf{T})$ ($\lambda_{\max}(\mathbf{T})$) denote the smallest (largest) eigenvalue of the matrix $\mathbf{T}$. Also, we denote

$$\text{(4a)}\quad \mathbf{M}_{nm}(\beta) = \mathrm{Cov}(\mathbf{g}_{nm}(\beta)) = \sum_{i=1}^{n} \mathbf{D}_i^T(\beta)\mathbf{V}_i^{-1}(\beta,\alpha)\boldsymbol{\Sigma}_i(\beta)\mathbf{V}_i^{-1}(\beta,\alpha)\mathbf{D}_i(\beta),$$

$$\text{(4b)}\quad \mathcal{D}_{nm}(\beta) = -\frac{\partial \mathbf{g}_{nm}(\beta)}{\partial \beta^T},$$

(4c) $\mathbf{H}_{nm}(\beta) = \sum_{i=1}^{n} \mathbf{D}_i^T(\beta)\mathbf{V}_i^{-1}(\beta, \alpha)\mathbf{D}_i(\beta).$

The matrix $\mathcal{D}_{nm}(\beta)$ is not symmetric in general. Let β_0 be the true regression parameter. In the sequel, when the terms of functions of β are evaluated at β_0, we will suppress β_0. For example, we let $\mathbf{g}_{nm} = \mathbf{g}_{nm}(\beta_0)$, $\mathbf{H}_{nm} = \mathbf{H}_{nm}(\beta_0)$ and $\mathbf{M}_{nm} = \mathbf{M}_{nm}(\beta_0)$, etc.

To prove the existence and weak consistency of $\widehat{\beta}_{nm}$, we present two sets of general conditions. The first set of conditions requires the sandwich information matrix $\mathbf{F}_{nm} = \mathbf{H}_{nm}\mathbf{M}_{nm}^{-1}\mathbf{H}_{nm}$ to be divergent and the second set of conditions requires the marginal information matrix $\mathbf{H}_{nm}$ to be divergent at a rate faster than $\tau_{nm} = \max_{1\le i\le n}\{\lambda_{\max}(\mathbf{R}_i^{-1}(\alpha)\bar{\mathbf{R}}_i)\}$, in additon to some mild local conditions; see Section 2 for details. The second set of conditions depends on the $\bar{\mathbf{R}}_i$'s only through τ_{nm}. Also, τ_{nm} can be replaced by $m\widetilde{\lambda}_{nm}$, a term that does not depend on the $\bar{\mathbf{R}}_i$'s, where $\widetilde{\lambda}_{nm} = \max_{1\le i\le n} \lambda_{\max}(\mathbf{R}_i^{-1}(\alpha))$.

We establish, as in Crowder (1986) but under a simpler set of conditions, that the asymptotic distributions of $\widehat{\beta}_{nm}$ and $\mathbf{g}_{nm}$ differ only by a scale matrix. So the asymptotic normality of $\widehat{\beta}_{nm}$ can be established by establishing asymptotic normality of $\mathbf{g}_{nm}$. When m goes to infinity, the rate of m versus n is critical. When n is bounded or m goes to infinity too fast, we usually do not have asymptotic normality of $\mathbf{g}_{nm}$ or $\widehat{\beta}_{nm}$ without specifying the dependence structure on each subject; see Example 3.1 in Section 3. When m is bounded above or tends to ∞ at a limited rate as $n \to \infty$, we present a set of sufficient conditions to ensure the asymptotic normality of $\mathbf{g}_{nm}$ and $\widehat{\beta}_{nm}$. This set of conditions relies mainly on the marginal moments of the individual distributions.

Our condition for almost sure existence and strong consistency is that, when $n \to \infty$ and β is in a neighborhood of β_0, $\lambda_{\min}(\mathcal{D}_{nm}(\beta)^T\mathbf{M}_{nm}^{-1}\mathcal{D}_{nm}(\beta))$ increases at the rate $(\log n)_{2(1+\delta)}$ or faster, almost surely, for some small $\delta > 0$, plus some additional minor conditions. A difficulty arises in proving the strong consistency because $\mathbf{g}_{m_1,1}, \mathbf{g}_{m_2,2}, \ldots, \mathbf{g}_{m_n,n}$, the summands of $\mathbf{g}_{nm}$, should be treated as a double array sequence and results like Lemma 2 of Wu (1981) cannot be applied.

The conditions for weak consistency and normality are verified for several examples of general interest, including (i) linear regression models, (ii) generalized linear models with regressors in a compact range, (iii) Poisson regression models, (iv) models with bounded responses, for example, logistic, probit and other categorical regression models, etc. Although our conditions may not be the best set of conditions for each specific model, the technical assumptions are very mild in general and are typically satisfied in practice.

Following Fahrmeir and Kaufmann (1985), we will ignore the nuisance parameter ϕ in (1) and (2). The estimator $\widehat{\beta}_{nm}$ remains the same with or without ϕ, and the information matrices defined in (4a)–(4c) and $\mathbf{F}_{nm}$ only change by a scale factor that can often be estimated consistently. For our asymptotic results, we do not need to estimate the nuisance parameter ϕ. Also, to simplify our discussion, we will not study the other nuisance parameters α that appear in the working correlation matrix $\mathbf{R}_i(\alpha)$. The standard assumption on the estimator of α is that it is consistent, given β and ϕ. The results presented here can be extended to this case by the standard arguments similar to those used in Liang and Zeger (1986).

Let Θ be the *natural parameter space* of the exponential family distributions presented in (1), that is, $\Theta = \{\theta | 0 < \int \exp\{y\theta + b(y)\}\, d\xi(y) < +\infty\}$. The interior of Θ is denoted as Θ^o. Throughout the paper, we assume the following *regularity assumptions*:

1. β is in an admissible set $\mathcal{B}$, where $\mathcal{B}$ is an open set in R^p.
2. $\mathbf{x}^T\beta \in g(\mathcal{M})$ for all $\beta \in \mathcal{B}$ and $\mathbf{x} \in \mathcal{X}$, where $\mathcal{M}$ is the image of $a'(\Theta^o)$ and $\mathcal{X}$ is the set of all possible covariate variables.
3. $a'(\theta)$ is three times continuously differentiable and $a''(\theta) > 0$ in Θ^o. Also, $u(\eta)$ is three times continuously differentiable and $u'(\eta) > 0$ in $g(\mathcal{M})^o$.
4. $\mathbf{H}_{nm}$ and $\mathbf{M}_{nm}$ are positive definite when n or m is large.

The rest of the paper is arranged as follows. Section 2 presents general theorems on the existence and weak consistency of the GEE estimator $\widehat{\beta}_{nm}$. Section 3 studies the asymptotic distributions of the GEE score function $\mathbf{g}_{nm}$ and the GEE estimator $\widehat{\beta}_{nm}$. Section 4 develops conditions to ensure the existence and strong consistency of the GEE estimator $\widehat{\beta}_{nm}$. Section 5 studies several examples of general interest in practice.

2. Asymptotic existence and weak consistency of the GEE estimator. In addition to the regularity assumptions listed at the end of Section 1, we need some further conditions to ensure the existence and weak consistency of the GEE estimator. These are:

(I_w) $\lambda_{\min}(\mathbf{F}_{\mathrm{nm}}) \to \infty$.

(L_w) There exists a constant $c_0 > 0$, for any $r > 0$, such that

$$P\big(\mathcal{D}_{nm}^T(\beta)\mathbf{M}_{nm}^{-1}\mathcal{D}_{nm}(\beta) \geq c_0\mathbf{F}_{nm} \text{ and } \mathcal{D}_{nm}(\beta) \text{ is nonsingular, for all } \beta \in B_{nm}(r)\big) \to 1,$$

where $B_{nm}(r) = \{\beta : \|\mathbf{M}_{nm}^{-1/2}\mathbf{H}_{nm}(\beta - \beta_0)\| \leq r\}$.

In (L_w) (also in the sequel), the square root of a positive definite matrix is the (unique) symmetric positive definite square root [see, e.g., Gourieroux and Monfort (1981)]. Also, we define ordering between two square matrices as: $\mathbf{C}_1 \geq \mathbf{C}_2$ if and only if $\lambda^T\mathbf{C}_1\lambda \geq \lambda^T\mathbf{C}_2\lambda$ for any vector λ, $\|\lambda\| = 1$. The two square matrices need not be symmetric. When both $\mathbf{C}_1$ and $\mathbf{C}_2$ are symmetric, $\mathbf{C}_1 \geq \mathbf{C}_2$ is equivalent to $\mathbf{C}_1 - \mathbf{C}_2$ being nonnegative definite.

REMARK 1. Condition (L_w) is a local condition. One can write

$$\mathcal{D}_{nm}(\beta) = \mathbf{H}_{nm}(\beta) + \mathbf{B}_{nm}(\beta) + \mathcal{E}_{nm}(\beta), \tag{5}$$

where $\mathbf{B}_{nm}(\beta) = \mathrm{E}\{\mathcal{D}_{nm}(\beta)\} - \mathbf{H}_{nm}(\beta)$ and $\mathcal{E}_{nm}(\beta) = \mathcal{D}_{nm}(\beta) - \mathrm{E}\{\mathcal{D}_{nm}(\beta)\}$. Exact formulas for $\mathbf{B}_{nm}(\beta)$ and $\mathcal{E}_{nm}(\beta)$ are provided in Appendix A. It is easy to see that

$$\mathbf{B}_{nm}(\beta_0) = 0 \quad \text{and} \quad \mathrm{E}\{\mathcal{E}_{nm}(\beta)\} = 0.$$

Intuitively, if $\mathbf{B}_{nm}(\beta)$ is continuous in β and one can apply a uniform law of large numbers to $\mathcal{E}_{nm}(\beta)$, $\mathcal{D}_{nm}(\beta)$ will be close to $\mathbf{H}_{nm}(\beta)$ for any β close to the true parameter β_0. As pointed out by a referee, $\mathcal{D}_{nm}^T(\beta)\mathbf{M}_{nm}^{-1}\mathcal{D}_{nm}(\beta) \geq c_0\mathbf{F}_{nm}$ in (L_w) may be replaced by $\mathbf{H}_{nm}(\beta)\mathbf{M}_{nm}^{-1}\mathbf{H}_{nm}(\beta) \geq c_0\mathbf{F}_{nm}$. The conditions required for $\mathbf{H}_{nm}(\beta)$ to be continuous around β_0 are almost the same as those required for $\mathbf{B}_{nm}(\beta)$ to be continuous around β_0. Typically, one would directly verify (L_w).

REMARK 2. Clearly, $\mathcal{D}_n(\beta)$ is a random matrix. Such a random matrix was also used by Fahrmeir and Kaufmann (1985) in their conditions (C*) and (S_δ^*) when they specify their consistency conditions for noncanonical link generalized linear models [their $\mathbf{H}_n(\beta)$ matrix].

THEOREM 1. *Under conditions* (I_w) *and* (L_w), *there exists a sequence of random variables* $\widehat{\beta}_{nm}$, *such that*

$$P\big(\mathbf{g}_{nm}(\widehat{\beta}_{nm}) = 0\big) \to 1$$

and

$$\widehat{\beta}_{nm} \to \beta_0 \qquad \textit{in probability}.$$

PROOF. For any $\varepsilon > 0$, take $r = \sqrt{\frac{2p}{c_0\varepsilon}}$, define $B_{nm}(r)$ as in (L_w) and let

$$E_{nm} = \left\{\omega : \|\mathbf{M}_{nm}^{-1/2}\mathbf{g}_{nm}\| \leq \inf_{\beta \in \partial B_{nm}(r)} \|\mathbf{M}_{nm}^{-1/2}(\mathbf{g}_{nm}(\beta) - \mathbf{g}_{nm})\|\right\},$$

where $\partial B_{nm}(r)$ is the boundary of sphere $B_{nm}(r)$. Under the regularity assumptions listed at the end of Section 1, it follows that the mapping $T : \beta \to \mathbf{M}_{nm}^{-1/2}\mathbf{g}_{nm}(\beta)$ is continuously differentiable. Since $\mathcal{D}_{nm}(\beta)$ is nonsingular for $\beta \in B_{nm}(r)$, it is clear that T is also an injection from $B_{nm}(r)$ to $T(B_{nm}(r))$. According to Lemma A of Chen, Hu and Ying (1999), on the set $E_{nm} \cap \{\mathcal{D}_{nm}(\beta)$ is nonsingular$\}$, there exists a $\widehat{\beta}_{nm} \in B_{nm}(r)$ such that $\mathbf{g}_{nm}(\widehat{\beta}_{nm}) = 0$. Therefore, we only need to prove $P(E_{nm}) > 1 - \varepsilon$ for n or m large enough. After that, we prove $P(\|\widehat{\beta}_{nm} - \beta_0\| \leq \delta) > 1 - \varepsilon$, for any $\delta > 0$.

By Taylor's expansion,

$$\begin{aligned}\mathbf{M}_{nm}^{-1/2}(\mathbf{g}_{nm}(\beta) - \mathbf{g}_{nm}) &= \mathbf{M}_{nm}^{-1/2}\mathcal{D}_{nm}(\bar{\beta})(\beta - \beta_0)\\ &= \mathbf{M}_{nm}^{-1/2}\mathcal{D}_{nm}(\bar{\beta})\mathbf{H}_{nm}^{-1}\mathbf{M}_{nm}^{1/2}\mathbf{M}_{nm}^{-1/2}\mathbf{H}_{nm}(\beta - \beta_0),\end{aligned}$$

where $\bar{\beta}$ lies between β and β_0. So, for $\beta \in \partial B_{nm}(r)$, we have $\bar{\beta} \in B_{nm}(r)$ and

$$\|\mathbf{M}_{nm}^{-1/2}(\mathbf{g}_{nm}(\beta) - \mathbf{g}_{nm})\| \geq r\mathbf{z}_{\lambda}^{1/2},$$

where $\mathbf{z}_{\lambda} = \lambda_{\min}(\mathbf{M}_n^{1/2}\mathbf{H}_{nm}^{-1}\mathcal{D}_{nm}^T(\bar{\beta})\mathbf{M}_{nm}^{-1}\mathcal{D}_{nm}(\bar{\beta})\mathbf{H}_{nm}^{-1}\mathbf{M}_{nm}^{1/2})$. By (L_w), $P(\mathbf{z}_{\lambda} \geq c_0) > 1 - \varepsilon/2$ when n or m is large enough. By the Chebyshev inequality, we have

$$P\big(\{\|\mathbf{M}_{nm}^{-1/2}\mathbf{g}_{nm}\| \leq c_0{}^{1/2}r\}\big) \geq 1 - \frac{\mathrm{E}\|\mathbf{M}_{nm}^{-1/2}\mathbf{g}_{nm}\|^2}{c_0 r^2} = 1 - \frac{p}{c_0 r^2} = 1 - \frac{\varepsilon}{2}.$$

Therefore,

$$\begin{aligned} P(E_{nm}) &\geq P\big(\|\mathbf{M}_{nm}^{-1/2}\mathbf{g}_{nm}\| \leq r\mathbf{z}_{\lambda}^{1/2}\big) \geq P\big(\{\|\mathbf{M}_{nm}^{-1/2}\mathbf{g}_{nm}\| \leq r\mathbf{z}_{\lambda}^{1/2}\} \cap \{\mathbf{z}_{\lambda} \geq c_0\}\big) \\ &\geq P\big(\{\|\mathbf{M}_{nm}^{-1/2}\mathbf{g}_{nm}\| \leq rc_0^{1/2}\}\big) + P(\mathbf{z}_{\lambda} \geq c_0) - 1 > 1 - \varepsilon. \end{aligned}$$

On E_{nm}, there exists a GEE estimator $\widehat{\beta}_{nm} \in B_{nm}(r)$. This implies $\|\mathbf{M}_{nm}^{-1/2} \times \mathbf{H}_{nm}(\widehat{\beta}_{nm} - \beta_0)\| < r$. Since $\lambda_{\min}(\mathbf{F}_{nm}) \to \infty$, we have $r/\lambda_{\min}(\mathbf{F}_{nm}) < \delta$ for any given δ, when n or m is large. Therefore, on E_{nm} we have $\|\widehat{\beta}_{nm}(\omega) - \beta_0\| \leq \delta$. This leads to

$$P\big(\omega : \|\widehat{\beta}_{nm}(\omega) - \beta_0\| \leq \delta\big) > 1 - \varepsilon,$$

when n or m is large enough. $\square$

The matrix $\mathbf{M}_{nm}$ in conditions (I_w) and (L_w) depends on the $\bar{\mathbf{R}}_i$'s, and often it is not completely specified in a GEE model. Next, we study an alternative set of conditions not depending on the matrix $\mathbf{M}_{nm}$. Although the conditions still depend on the $\bar{\mathbf{R}}_i$'s, the dependence is only through $\tau_{nm} = \max_{1\leq i\leq n}\{\lambda_{\max}(\mathbf{R}_i^{-1}(\alpha)\bar{\mathbf{R}}_i)\}$ and they can be further reduced to conditions free of the $\bar{\mathbf{R}}_i$'s (see Remark 5 for further details).

Note that $\mathbf{M}_{nm} \leq \tau_{nm}\mathbf{H}_{nm}$. Condition (I_w) is implied by the following assumption:

(I_w^*) $(\tau_{nm})^{-1}\lambda_{\min}(\mathbf{H}_{\mathrm{nm}}) \to \infty$.

When (I_w) is replaced by the stronger condition (I_w^*), we can replace (L_w) by the following (L_w^*), which does not depend on $\mathbf{M}_{nm}$.

(L_w^*) There exists a constant c_0, for any $\delta > 0$ and $r > 0$, such that

$$P\big(\mathcal{D}_{nm}(\beta) \geq c_0\mathbf{H}_{nm} \text{ and } \mathcal{D}_{nm}(\beta) \text{ is nonsingular, for } \beta \in B_{nm}^*(r)\big) \to 1,$$

where $B_{nm}^*(r) = \{\beta : \|\mathbf{H}_{nm}^{1/2}(\beta - \beta_0)\| \leq (\tau_{nm})^{1/2}r\}$.

REMARK 3. In response to a comment by a referee, we note that the eigenvalues of $\mathbf{R}_i^{-1}(\alpha)\bar{\mathbf{R}}_i$ are the same as the eigenvalues of $\mathbf{R}_i^{-1/2}(\alpha)\bar{\mathbf{R}}_i\mathbf{R}_i^{-1/2}(\alpha)$.

This can be proved using, for example, Theorem 3.2(d) of Schott (1997), page 88. By the same argument, the eigenvalues of $\mathbf{M}_{nm}^{-1}\mathbf{H}_{nm}$ and $\mathbf{M}_{nm}^{-1/2}\mathbf{H}_{nm}\mathbf{M}_{nm}^{-1/2}$ are the same, and the same applies to $\mathbf{H}_{nm}^{-1}\mathbf{M}_{nm}$ and $\mathbf{H}_{nm}^{-1/2}\mathbf{M}_{nm}\mathbf{H}_{nm}^{-1/2}$. In the sequel, we'll use these results without further comment.

REMARK 4. When $\mathcal{D}_{nm}(\beta)$ is positive definite, $\mathcal{D}_{nm}^T(\beta)\mathbf{M}_{nm}^{-1}\mathcal{D}_{nm}(\beta) \geq c_0\mathbf{F}_{nm}$ implies $\mathcal{D}_{nm}(\beta) \geq c_0^{1/2}\mathbf{H}_{nm}$. This is an immediate result from a basic matrix theorem: For $p \times p$ matrices A, B and G, if A and G are positive definite and $AGA \geq BGB$, then $A \geq B$ [Ni (1984), Theorem 2.5, page 107]. So, when $\mathcal{D}_{nm}(\beta)$ is positive definite, the statement in (L_w^*) is weaker than the statement in (L_w). However, in general, $\mathcal{D}_{nm}(\beta)$ is not symmetric and $B_n(r) \subset B_n^*(r)$, hence, neither (L_w) nor (L_w^*) can be implied by the other.

THEOREM 2. *The results of Theorem* 1 *hold if* (I_w) *and* (L_w) *are replaced by* (I_w^*) *and* (L_w^*), *respectively.*

We first need to prove a lemma.

LEMMA 1. *Suppose* $\mathbf{C}$ *is a* $p \times p$ *matrix. For any* $p \times 1$ *vector* λ, $\|\lambda\| = 1$, *we have* $\lambda^T\mathbf{C}^T\mathbf{C}\lambda \geq (\lambda^T\mathbf{C}\lambda)^2$.

PROOF. For any given $p \times 1$ vector λ with λ, $\|\lambda\| = 1$, construct an orthogonal matrix $\mathbf{\Gamma}$ such that its first column is λ. Denote $\mathbf{b} = \mathbf{\Gamma}^T\mathbf{C}\lambda$. The first element of $\mathbf{b}$ is then $b_1 = \lambda^T\mathbf{C}\lambda$ and $\mathbf{C}\lambda = \mathbf{\Gamma}\mathbf{b}$. So, $\lambda^T\mathbf{C}^T\mathbf{C}\lambda = \mathbf{b}^T\mathbf{b} \geq b_1^2 = (\lambda^T\mathbf{C}\lambda)^2$. $\square$

PROOF OF THEOREM 2. The proof is the same as the proof of Theorem 1, except for a few changes. We consider here the mapping $T^*: \beta \to \mathbf{H}_{nm}^{-1/2}\mathbf{g}_{nm}(\beta)$, instead of $T: \beta \to \mathbf{M}_{nm}^{-1/2}\mathbf{g}_{nm}(\beta)$. In addition, we replace the ball $B_{nm}(r)$ by $B_{nm}^*(r)$, replace the set E_{nm} by

$$E_{nm}^* = \left\{\omega : \|\mathbf{H}_{nm}^{-1/2}\mathbf{g}_{nm}\| \leq \inf_{\beta \in \partial B_{nm}^*(r)} \left\|\mathbf{H}_{nm}^{-1/2}\big(\mathbf{g}_{nm}(\beta) - \mathbf{g}_{nm}\big)\right\|\right\}$$

with $r = \sqrt{\frac{2p}{c_0^2\varepsilon}}$, and replace the $\mathbf{z}_\lambda$ by $\mathbf{z}_\lambda^* = \lambda_{\min}(\mathbf{H}_{nm}^{1/2}\mathcal{D}_{nm}^T(\bar{\beta})\mathbf{H}_{nm}^{-1}\mathcal{D}_{nm}(\bar{\beta})\mathbf{H}_{nm}^{-1/2})$.

By Taylor's expansion, we have

$$\left\|\mathbf{H}_{nm}^{-1/2}\big(\mathbf{g}_{nm}(\beta) - \mathbf{g}_{nm}\big)\right\| \geq r(\tau_{nm})^{1/2}(\mathbf{z}_\lambda^*)^{1/2},$$

for $\beta \in \partial B_{nm}^*(r)$ and $\bar{\beta}$ between β and β_0. By Lemma 1 and (L_w^*), we know that $\mathbf{z}_\lambda^* > c_0^2$. By the Chebyshev inequality, we have

$$\begin{aligned} &P\big(\|\mathbf{H}_{nm}^{-1/2}\mathbf{g}_{nm}\| \leq c_0 r(\tau_{nm})^{1/2}\big) \\ &\quad \geq 1 - \frac{\mathrm{E}\|\mathbf{H}_{nm}^{-1/2}\mathbf{g}_{nm}\|^2}{c_0^2 r^2\tau_{nm}} = 1 - \frac{\mathrm{tr}(\mathbf{H}_{nm}^{-1}\mathbf{M}_{nm})}{c_0^2 r^2\tau_{nm}} \geq 1 - \frac{\varepsilon}{2}. \end{aligned}$$

With these changes, the proof proceeds exactly the same as that of Theorem 1. □

REMARK 5. The results in Theorem 2 still hold if τ_{nm} in (I_w^*) and (L_w^*) is replaced by $m\widetilde{\lambda}_{nm}$ or $\lambda_{\max}(\mathbf{H}_{nm}^{-1}\mathbf{M}_{nm})$, where $\widetilde{\lambda}_{nm} = \max_{1\le i\le n} \lambda_{\max}(\mathbf{R}_i^{-1}(\alpha))$. The proof is similar to that of Theorem 2. Note, when τ_{nm} is replaced by $m\widetilde{\lambda}_{nm}$, the conditions ($\mathrm{I}_w^*$) and ($\mathrm{L}_w^*$) only involve terms that are related to the marginal distributions of the individual responses.

Before we end this section, let us give a brief discussion of the conditions (I_w) and (I_w^*) on the information matrices. In particular, we examine the divergence rate of the information matrices $\mathbf{F}_{nm}$ and $\mathbf{H}_{nm}$ in these conditions. Clearly, the rate depends on the correlation structure of clusters, the choice of the working correlation matrix, the covariates design matrix as well as the assumed marginal model, although the correlation structures may not be that critical for consistency when $n \to \infty$. Generally speaking, in the case when $n \to \infty$, with some mild restrictions on x_{ij}, there exists a set of working correlation matrices $\mathbf{R}_i$'s such that both $(\tau_{nm})^{-1}\lambda_{\min}(\mathbf{H}_{\mathrm{nm}})$ in (I_w^*) and $\lambda_{\min}(\mathbf{F}_{nm})$ in (I_w) diverge at the rate n or faster, regardless of the dependence structure for each cluster and whether $m = m(n) \to \infty$ or not; the rates can be nm, depending on the correlation structure of clusters. In the case when n is bounded and $m \to \infty$, our theorems can still apply. However, in order to obtain a consistent GEE estimator in this case, it is necessary to impose restrictions on the correlation structure. We use the following example of a simple linear regression model to illustrate some of the details. Note, conditions (L_w) and (L_w^*) are trivially true for this model.

EXAMPLE 2.1. Assume the random response y_{ij} follows a marginal regression model,

$$y_{ij} = x_{ij}\beta + \varepsilon_{ij} \qquad \text{for } j = 1, 2, \dots, m,\ i = 1, 2, \dots, n, \tag{6}$$

where $\mathrm{E}(\varepsilon_{ij}) = 0$ and, without loss of generality and to simplify our notation, we assume that $\mathrm{var}(\varepsilon_{ij}) = 1$ and the x_{ij} are scalars. The observations between clusters are independent and the observations within a cluster may or may not be correlated. The random effects model $y_{ij} = \mu + b_i + \varepsilon'_{ij}$, with random effect b_i and independent error ε'_{ij}, is a special case of model (6). We assume that the x_{ij} are from a compact set and $\sum_{i=1}^n \mathbf{x}_i^T\mathbf{x}_i = \sum_{i=1}^n\sum_{j=1}^m x_{ij}^2$ is of order nm. Denote $\bar{\rho}_{i;jj'} = \mathrm{corr}(\varepsilon_{ij}, \varepsilon_{ij'})$, so $\bar{\rho}_{i;jj'}$ is the (j, j') element of $\bar{\mathbf{R}}_i$, for $j \ne j'$. Under the current setting, $H_{nm} = \sum_{i=1}^n \mathbf{x}_i^T\mathbf{R_i}^{-1}(\alpha)\mathbf{x}_i$, $M_{nm} = \sum_{i=1}^n \mathbf{x}_i^T\mathbf{R_i}^{-1}(\alpha)\bar{\mathbf{R}}_\mathbf{i}\mathbf{R_i}^{-1}(\alpha)\mathbf{x}_i$ and $F_{nm} = (H_{nm})^2/M_{nm}$.

When $n \to \infty$, no matter what value $\bar{\rho}_{i;jj'}$ is and whether m tends to infinity or not, by taking the working correlation matrices $\mathbf{R_i}(\alpha)$'s to be an identity matrix, we always have $H_{nm}^{-1} = O((nm)^{-1})$. Therefore, the rate of $(\tau_{nm})^{-1}H_{nm}$ in (I_w^*) is at least n or faster. Since (I_w^*) implies (I_w), F_{nm} in (I_w) diverges at the rate of n or faster. When m tends to infinity and with some special correlation structures, we can get even faster divergence. For example, in the extreme case that $\bar{\rho}_{i;jj'} \equiv 0$, by taking the $\bar{\mathbf{R}}_\mathbf{i}$'s to be an identity matrix, both H_{nm} (now $\tau_{nm} = 1$) in (I_w^*) and F_{nm} in (I_w) diverge at the order of nm.

Suppose now the number of clusters n is finite. Without loss of generality, one can assume $n = 1$. In the case when $\bar{\rho}_{1;jj'} \equiv 1$, we essentially have only one random observation, $(\tau_{nm})^{-1} H_{nm}$ and F_{nm} are bounded, and we usually cannot get a consistent GEE estimator. Suppose, however, $\max_{j \neq j'} |\rho_{1;jj'}|$ is uniformly bounded away from 1. In particular, if $\rho_{1;jj'} \equiv \rho$ for $0 < \rho < 1$, we can easily conclude that there exists a consistent GEE estimator if and only if $x_{1j} \not\equiv x_{1j'}$, for some $j \neq j'$. Further, when $|\rho_{1;jj'}| \leq \rho_{|j-j'|}$ for $j \neq j'$ and $\lim_{k\to\infty} \rho_k = 0$, taking $\mathbf{R}_1(\alpha) = \mathbf{I}$, we have that H_m diverges at the order of m and $M_m \leq 2\max\{|x_{11}|^2, \ldots, |x_{1m}|^2\} \sum_{k=0}^m (m-k)\rho_k = o(m^2)$. So F_m diverges and we can always find a consistent GEE estimator in this case.

3. Asymptotic distribution of the GEE estimator. The asymptotic distribution of the GEE estimator $\widehat{\beta}_{nm}$ is closely related to the asymptotic distribution of the score function $\mathbf{g}_{nm}$. The following condition is used in establishing the relationship.

(CC) For any given $r > 0$ and $\delta > 0$,

$$P\Bigl(\sup_{\beta \in B_{nm}^*(r)} \|\mathbf{H}_{nm}^{-1/2} \mathcal{D}_{nm}(\beta) \mathbf{H}_{nm}^{-1/2} - I\| < \delta\Bigr) \to 1,$$

where $B_{nm}^*(r)$ is defined in (L_w^*) and the matrix norm is the Euclidean matrix norm.

Note (CC) implies (L_w^*). We have the following result.

THEOREM 3. *Suppose that conditions* (I_w), (L_w) *and* (CC) *hold*, *or the conditions* (I_w^*) *and* (CC) *hold. Then*, *there exists a sequence of solutions* $\widehat{\beta}_{nm}$ *to the GEE equation in* $B_{nm}^*(r)$ *such that* $\mathbf{M}_{nm}^{-1/2}\mathbf{H}_{nm}(\widehat{\beta}_{nm} - \beta_0)$ *and* $\mathbf{M}_{nm}^{-1/2}\mathbf{g}_{nm}$ *are asymptotically identically distributed.*

PROOF. By Theorem 1 or Theorem 2, there exists a GEE solution $\widehat{\beta}_{nm}$ such that $\widehat{\beta}_{nm} \in B_{nm}(r)$ or $\widehat{\beta}_{nm} \in B_{nm}^*(r)$, respectively. Since $B_{nm}(r) \subset B_{nm}^*(r)$, $\widehat{\beta}_{nm} \in B_{nm}^*(r)$ in either case. By Taylor's expansion, there exists $\bar{\beta}_{nm} \in B_{nm}^*(r)$, which lies between $\widehat{\beta}_{nm}$ and β_0, such that

$$\begin{aligned}\mathbf{H}_{nm}^{-1/2}\mathbf{g}_{nm} &= \mathbf{H}_{nm}^{-1/2}\mathcal{D}_{nm}(\bar{\beta}_{nm})(\widehat{\beta}_{nm} - \beta_0)\\ &= \{\mathbf{H}_{nm}^{-1/2}\mathcal{D}_{nm}(\bar{\beta}_{nm})\mathbf{H}_{nm}^{-1/2}\}\mathbf{H}_{nm}^{1/2}(\widehat{\beta}_{nm} - \beta_0).\end{aligned}$$

By (CC), the pair $\mathbf{H}_{nm}^{1/2}(\widehat{\beta}_{nm} - \beta_0)$ and $\mathbf{H}_{nm}^{-1/2}\mathbf{g}_{nm}$ are asymptotically identically distributed. Therefore, the pair $\mathbf{M}_{nm}^{-1/2}\mathbf{g}_{nm}$ and $\mathbf{M}_{nm}^{-1/2}\mathbf{H}_{nm}(\widehat{\beta}_{nm} - \beta_0)$ are asymptotically identically distributed. $\square$

In order to establish the asymptotic normality of $\widehat{\beta}_{nm}$ using Theorem 3, we need to establish that $\mathbf{g}_{nm}$ is asymptotically normally distributed. Since $\mathbf{g}_{nm}$ is a summation of n independent terms, when $n \to \infty$, one immediate sufficient condition is the Lindeberg condition [see, e.g., Billingsley (1986), page 369] on the double array series $Z_{nm;i} = \lambda^T \mathbf{M}_{nm}^{-1/2}\mathbf{D}_i^T \mathbf{V}_i^{-1}(\mathbf{y}_i - \mu_i) = \lambda^T \mathbf{M}_{nm}^{-1/2}\mathbf{g}_{m_i,i}$ where λ is any $p \times 1$ vector with $\|\lambda\| = 1$. However, direct verification of the Lindeberg condition for the $Z_{nm;i}$ requires knowledge of the $\bar{\mathbf{R}}_i$'s, which is usually unknown in a GEE model. It is natural to ask the question whether we can establish the asymptotic normality for $\mathbf{g}_{nm}$ without specifying the $\bar{\mathbf{R}}_i$'s. If m tends to infinity too fast compared to n (this includes the case when n is bounded and $m \to \infty$), the answer is no. In these cases, one or a few of the summands in $\mathbf{g}_{nm}$ can dominate the rest. Example 3.1 below demonstrates that, when m tends to infinity too fast, $\mathbf{g}_{nm}$ (however normalized) may not converge to a normal distribution, even if the marginal distributions of the responses are fully specified and have nice properties.

EXAMPLE 3.1. Suppose $\{Y_{ij},\ j = 1, \dots, m\}$, for $i = 1, \dots, n$, are n independent batches of random variables. The marginal distribution of Y_{ij} is a Poisson(μ_{ij}) distribution with $\mu_{ij} = \exp(\mathbf{x}_{ij}^T \beta)$, where $\mathbf{x}_{ij} \equiv 1$ (scalar). The m observations in the first cluster are identical, and the observations within the other $n-1$ clusters are independent. If we take $\mathbf{R}_i = \mathbf{I}$, then the GEE score function is

$$g_{nm}(\beta) = m(y_{11} - e^{\beta}) + \sum_{i=2}^{n}\sum_{j=1}^{m}(y_{ij} - e^{\beta}).$$

All terms in the above score function are independent of each other. When $m/n \to \infty$, the first term is dominant and $g_{nm}(\beta)$ has asymptotically a centered Poisson distribution (after proper normalization). When $m/n \to c \in (0, \infty)$, $g_{nm}(\beta)$ is asymptotically a linear combination of a centered Poisson distribution and a Gaussian distribution. In fact, $g_{nm}(\beta)$ is asymptotically normally distributed, if and only if $m/n \to 0$.

When $n \to \infty$, if m is bounded or $m \to \infty$ at a limited rate, we can establish asymptotic normality for $\mathbf{g}_{nm}$. Next, we present such a result.

For $t > 0$, let $\psi(t)$ be a positive nondecreasing function such that $\lim_{t\to\infty} \psi(t) = \infty$ and $t\psi(t)$ is a convex function. Examples are $\psi(t) = e^t$, or $t^{1/\delta}$ for a $\delta > 0$, etc. Also, we denote $c_{nm} = \lambda_{\max}(\mathbf{M}_{nm}^{-1}\mathbf{H}_{nm})$ and

$$\gamma_{nm}^{(D)} = \max_{1 \le i \le n} \lambda_{\max}(\mathbf{H}_{nm}^{-1/2}\mathbf{D}_i^T \mathbf{V}_i^{-1}\mathbf{D}_i \mathbf{H}_{nm}^{-1/2}).$$

The following lemma holds.

LEMMA 2. *Under the GEE setting, suppose there exist a constant K (independent of n) and an integer n_0 such that, for $j = 1, 2, \dots, m_i$ and $i = 1, 2, \dots n$, when $n > n_0$,*

$$\mathrm{E}\big[{y_{ij}^{*}}^{2}\psi({y_{ij}^{*}}^{2})\big] \le K,$$

where $\mathbf{y}_i^* = (y_{i1}^*, \dots, y_{im_i}^*)^T = \mathbf{A}_i^{-1/2}(\mathbf{y}_i - \mu_i)$. In addition, for any $\varepsilon > 0$,

$$c_{nm}\widetilde{\lambda}_{nm} m\left[\psi\left(\frac{\varepsilon^2}{c_{nm}\widetilde{\lambda}_{nm} m \gamma_{nm}^{(D)}}\right)\right]^{-1} \to 0. \tag{7}$$

Then, when $n \to \infty$, we have

$$\mathbf{M}_{nm}^{-1/2}\mathbf{g}_{nm} \to N(0, \mathbf{I}) \qquad \textit{in distribution.}$$

PROOF. We only need to verify the Lindeberg condition for the double arrays $Z_{nm,i} = \lambda^T \mathbf{M}_{nm}^{-1/2}\mathbf{D}_i^T\mathbf{V}_i^{-1}(\mathbf{y}_i - \mu_i)$, for any given $p \times 1$ vector λ, $\|\lambda\| = 1$. By the Cauchy–Schwarz inequality,

$$\begin{aligned} Z_{nm,i}^2 &\le \lambda_{\max}(\mathbf{M}_{nm}^{-1}\mathbf{H}_{nm})\lambda^T\mathbf{H}_{nm}^{-1/2}\mathbf{D}_i^T\mathbf{V}_i^{-1}\mathbf{D}_i\mathbf{H}_{nm}^{-1/2}\lambda\ (\mathbf{y}_i - \mu_i)^T\mathbf{V}_i^{-1}(\mathbf{y}_i - \mu_i) \\ &\le c_{nm}\widetilde{\lambda}_{nm}\gamma_{nm,i}^{(D)}\ {\mathbf{y}_i^*}^T\mathbf{y}_i^*. \end{aligned}$$

Here, $\gamma_{nm}^{(D)} = \lambda^T\mathbf{H}_{nm}^{-1/2}\mathbf{D}_i^T\mathbf{V}_i^{-1}\mathbf{D}_i\mathbf{H}_{nm}^{-1/2}\lambda$. Let $\mathbf{I}(c)$ be the indicator function of the set c. By the definition of $\psi(t)$ and the Jensen inequality, we have

$$\begin{aligned} &\sum_{i=1}^n \mathrm{E}\big[Z_{nm,i}^2\mathbf{I}(|Z_{nm,i}| > \varepsilon)\big] \\ &\quad\le \sum_{i=1}^n c_{nm}\widetilde{\lambda}_{nm}\gamma_{nm,i}^{(D)}\mathrm{E}\left[{\mathbf{y}_i^*}^T\mathbf{y}_i^*\mathbf{I}\left\{{\mathbf{y}_i^*}^T\mathbf{y}_i^* > \frac{\varepsilon^2}{c_{nm}\widetilde{\lambda}_{nm}\gamma_{nm,i}^{(D)}}\right\}\right] \\ &\quad\le c_{nm}\widetilde{\lambda}_{nm}\sum_{i=1}^n m_i\gamma_{nm,i}^{(D)}\mathrm{E}\left\{\frac{\sum_{j=1}^{m_i}{y_{ij}^*}^2}{m_i}\psi\left(\frac{\sum_{j=1}^{m_i}{y_{ij}^*}^2}{m_i}\right)\right\} \\ &\qquad\qquad\times\left\{\psi\left(\frac{\varepsilon^2}{c_{nm}\widetilde{\lambda}_{nm}m_i\gamma_{nm,i}^{(D)}}\right)\right\}^{-1} \\ &\quad\le c_{nm}\widetilde{\lambda}_{nm} m\sum_{i=1}^n \gamma_{nm,i}^{(D)}\left\{\frac{\sum_{j=1}^{m_i}\mathrm{E}\{{y_{ij}^*}^2\psi({y_{ij}^*}^2)\}}{m_i}\right\}\left\{\psi\left(\frac{\varepsilon^2}{c_{nm}\widetilde{\lambda}_{nm}m\gamma_{nm}^{(D)}}\right)\right\}^{-1} \\ &\quad\le Kc_{nm}\widetilde{\lambda}_n m\left\{\psi\left(\frac{\varepsilon^2}{c_{nm}\widetilde{\lambda}_{nm}m\gamma_{nm}^{(D)}}\right)\right\}^{-1} \to 0. \end{aligned}$$

Thus, the lemma holds, by the Lindeberg central limit theorem and the Cramér–Wold theorem. □

We introduce the following condition (N_δ). Aside from c_{nm}, condition (N_δ) only involves terms that are related to the marginal distributions of individual observations.

(N_δ) There exists a $\delta > 0$, such that $\mathrm{E}(y_{ij}^*)^{2+(2/\delta)}$ is uniformly bounded above, and

$$(c_{nm}\widetilde{\lambda}_{nm}m)^{1+\delta}\gamma_{nm}^{(D)} \to 0. \tag{8}$$

The next theorem proves asymptotic normality of the GEE estimator under condition (N_δ).

THEOREM 4. *Suppose the marginal distribution of each individual observation has a density of the form from* (1). *If condition* (N_δ) *is satisfied*, *then*, *when* $n \to \infty$, *we have*,

$$\mathbf{M}_{nm}^{-1/2}\mathbf{g}_{nm} \to N(0, \mathbf{I}) \qquad \textit{in distribution}.$$

Further, *under the conditions in Theorem* 3, *there exists a sequence of weakly consistent GEE estimators* $\widehat{\beta}_{nm}$ *and*

$$\mathbf{M}_{nm}^{-1/2}\mathbf{H}_{nm}(\widehat{\beta}_{nm} - \beta_0) \to N(0, \mathbf{I}) \qquad \textit{in distribution}. \tag{9}$$

PROOF. Taking $\psi(t) = t^{1/\delta}$ in Lemma 2, we can get the first result. The second result follows immediately from Theorem 3. □

REMARK 6. In many cases, $\gamma_{nm}^{(D)} \to 0$ at the rate $O(n^{-1})$. Condition (8) imposes a restriction on the rate of $m \to \infty$, as $n \to \infty$.

REMARK 7. In the nonlongitudinal case with $m_1 \equiv m_2 \equiv \cdots \equiv m_n \equiv 1$, (8) reduces to $\gamma_{nm}^{(D)} \to 0$. This condition is equivalent to the infinitesimal array condition in Feller's theorem [see, e.g., Billingsley (1986), pages 373 and 374]. Also, the moment restriction in (N_δ) is essentially the Lyapounov condition that $\max_{ij}\mathrm{E}(y_{ij}^*)^{2+\delta'}$ is bounded above, where $\delta' = 2/\delta$ [see also, e.g., Billingsley (1986), page 371].

REMARK 8. The normalization term in (9) involves $\mathbf{M}_{nm}$, which depends on the unknown covariance matrix $\mathbf{\Sigma}_i$. Following Liang and Zeger (1986), we suggest estimating $\mathbf{M}_{nm}$ by

$$\widehat{\mathbf{M}}_{nm} = \sum_{i=1}^{n} \mathbf{D}_i^T(\beta)\mathbf{V}_i^{-1}(\beta, \alpha)\widehat{\Sigma}_i\mathbf{V}_i^{-1}(\beta, \alpha)\mathbf{D}_i(\beta)\big|_{\beta=\hat{\beta}},$$

where $\widehat{\Sigma}_i = (\mathbf{y}_i - \mu_i(\hat{\beta}))(\mathbf{y}_i - \mu_i(\hat{\beta}))^T$. Using Theorem 1 and Corollary 2 of Section 10.1 in Chow and Teicher [(1988), pages 338 and 340], one can prove that under conditions such as those in Theorem 4, $\mathbf{M}_{nm}^{-1/2}\widehat{\mathbf{M}}_{nm}\mathbf{M}_{nm}^{-1/2} \to \mathbf{I}$ (elementwise) in probability as $n \to \infty$. Details of its proof are omitted.

4. Asymptotic existence and strong consistency. As mentioned in the introduction, when $n \to \infty$, we treat the summands $\mathbf{g}_{m_1,1}, \mathbf{g}_{m_2,2}, \dots, \mathbf{g}_{m_n,n}$ of the score function $\mathbf{g}_{nm}$ as a double array sequence. It is well known that the strong law of large numbers does not hold in general for an independent double array sequence [see Romano and Siegel (1986), pages 112–114]. The next lemma gives conditions under which a strong law of large numbers holds for double arrays. It can be viewed as an extension of Lemma 2 of Wu (1981) for single array sequences. We will use this result to give a strong consistency result for the GEE estimators when $n \to \infty$.

LEMMA 3. *Consider a double array sequence*,

$$\begin{array}{l} Z_{11} \\ Z_{21}, Z_{22} \\ \dots\dots \\ Z_{n1}, Z_{n2}, \dots, Z_{nn} \\ \dots\dots, \end{array}$$

where, *for each* n, $Z_{n1}, Z_{n2}, \dots, Z_{nn}$ *are independent and* $EZ_{n,j} = 0$, $j = 1, 2, \dots, n$. *Suppose there is a constant* $c_0 > 0$ (*independent of* n), *such that*

$$\limsup_{n\to\infty} \frac{\max_{1\le i\le n} |Z_{n,i}|}{s_n} \le c_0 \qquad a.s., \tag{10}$$

where $s_n^2 = \sum_{i=1}^n \sigma_{n,i}^2$, $\sigma_{n,i}^2 = \operatorname{var}(Z_{n,i})$. *Let* $\phi(n)$ *be a positive nondecreasing function satisfying*

$$\phi(n)/\log n \to \infty \qquad \text{when } n \to \infty. \tag{11}$$

Then, *for an* $A_n \to \infty$ *which satisfies*

$$\limsup_{n\to\infty} \frac{s_n\phi(n)}{A_n} = K < \infty, \tag{12}$$

we have

$$\lim_{n\to+\infty} \frac{1}{A_n} \sum_{i=1}^n Z_{n,i} = 0 \qquad a.s.$$

The proof of Lemma 3 is quite technical and can be found in Appendix B.

REMARK 9. In Lemma 2 of Wu (1981), (12) is replaced by $\limsup_{n\to\infty} s_n^{1+\delta}/A_n = K < \infty$. In a double array sequence, $\phi(n)$ cannot be replaced by s_n^δ and, also, $\phi(n)$ should tend to ∞ faster than $\log n$. This can be seen in a simple example: Suppose, for each given n, $Z_{n,i}$, $i = 1, \dots, n$, are n independent random variables, with $(Z_{n,i} + 1/n)$ following a Gamma$(n^{-1}, 1)$ distribution. In addition, the rows of the array $\{Z_{n,i},\ 1 \le i \le n\}$, for $n = 1, 2, \dots$, are independent. In this example, $s_n = 1$ and (10) can be directly verified. Because $\sum_{i=1}^n Z_{n,i} + 1$ follows an exponential distribution, $\sum_{n=1}^\infty P(|\sum_{i=1}^n Z_{n,i}| > \log n) \ge e^{-1}\sum_{n=1}^\infty (1/n) = \infty$. Thus, by the second Borel–Cantelli lemma, $\sum_{i=1}^n Z_{n,i}/\log n$ does not converge to 0 almost surely. This suggests that $\sum_{i=1}^n Z_{n,i}/A_n$ does not converge to 0 almost surely, for either $A_n = s_n^{1+\delta}$ or $A_n = s_n\phi(n)$ where $\phi(n) = O(\log n)$.

REMARK 10. Condition (10) cannot be omitted. To see this, we use a counterexample given in Romano and Siegel [(1986), page 114, Example 5.41(ii)]: Suppose, for each given n, $Z_{n,i}$, $i = 1, \dots, n$, are n independent identically distributed random variables taking on the values n, 0 and $-n$ with probabilities $1/(2n^2)$, $1 - 1/n^2$ and $1/(2n^2)$, respectively. In addition, the rows of the array $\{Z_{n,i},\ 1 \le i \le n\}$ for $n = 1, 2, \dots$ are independent. Note that $\mathrm{E}Z_{n,i} = 0$, $\mathrm{var}(Z_{n,i}) = 1$, $s_n^2 = n$ and (10) is violated. But (11) and (12) hold for $\phi(n) = (\log n)^{1+\delta}$, $\delta > 0$ and $A_n = n$. Romano and Siegel showed, however, that $\sum_{i=1}^n Z_{n,i}/n$ does not converge to 0 almost surely.

Condition (10) places a restriction on the tails of the random variables in the double array sequence. We call such kind of sequence a *strong infinitesimal double array sequence*. The definition provided next is for random vectors.

DEFINITION 4.1. An independent double array sequence of $p \times 1$ random vectors $\{\mathbf{z}_{n1}, \mathbf{z}_{n2}, \dots, \mathbf{z}_{nn}\}$, for $n = 1, 2, \dots$, is said to form a *strong infinitesimal double array sequence* if

$$\limsup_{n\to\infty} \frac{\max_{1\le i\le n} \|\mathbf{z}_{n,i} - \mathrm{E}\mathbf{z}_{n,i}\|}{\lambda_{\min}^{1/2}\{\sum_{i=1}^n \mathrm{var}(\mathbf{z}_{n,i})\}} \le c_0 \qquad \text{a.s.} \tag{13}$$

REMARK 11. When $p = 1$, $\mathbf{z}_{n,i}$ is a random variable, denote $Z_{n,i}$, and (13) becomes

$$\limsup_{n\to\infty} \frac{\max_{1\le i\le n} |Z_{n,i} - \mathrm{E}Z_{n,i}|}{s_n} \le c_0 \qquad \text{a.s.,} \tag{13$'$}$$

where $s_n^2 = \sum_{i=1}^n \sigma_{n,i}^2$ and $\sigma_{n,i}^2 = \mathrm{var}(Z_{n,i})$. Requirement (13)$'$ is the same as (10). By the Borel–Cantelli lemma, either

$$\frac{\mathrm{E}\{\max_{1\le i\le n} |Z_{n,i} - \mathrm{E}Z_{n,i}|^k\}}{{s_n}^k} = O\left(\frac{1}{n^{1+\varepsilon}}\right)$$

or

$$\frac{\max_{1\le i\le n}\{\mathrm{E}|Z_{n,i}-\mathrm{E}Z_{n,i}|^k\}}{{s_n}^k}=O\left(\frac{1}{n^{2+\varepsilon}}\right),$$

for some $k>2$ and $\varepsilon>0$, is sufficient to ensure $(13)'$. Note that, for an independent single array sequence $\{Z_1,Z_2,\dots,Z_n,\dots\}$, s_n^2 is a summation of (although not always) order n. In such a case, if $\sup_{1\le i\le\infty}\mathrm{E}|Z_i-\mathrm{E}Z_i|^{4+\varepsilon}$ has an upper bound for some $\varepsilon>0$, then $\{Z_1,Z_2,\dots,Z_n,\dots\}$ is a strong infinitesimal double array sequence.

Before presenting our main result, we first prove the following lemma.

LEMMA 4. *If* $\mathbf{g}_{m_i,i}$, $i=1,2,\dots,n$, *the summands of the GEE score function* $\mathbf{g}_{nm}$, *form a strong infinitesimal double array sequence*, *then*

$$\lim_{n\to\infty}\frac{\mathbf{M}_{nm}^{-1/2}\mathbf{g}_{nm}}{(\log n)^{1+\delta}}=0\qquad a.s.$$

where the convergence is elementwise.

PROOF. Without loss of generality, we look at the first element of the $p\times 1$ vector $\mathbf{M}_{nm}^{-1/2}\mathbf{g}_{nm}$. We define $\mathbf{g}_i^{(n)}=\mathbf{g}_{m_i,i}$ and $Z_{n,i}=\mathbf{e}_1^T\mathbf{M}_{nm}^{-1/2}\mathbf{g}_i^{(n)}$ where the $p\times 1$ vector $\mathbf{e}_1=(1,0,\dots,0)^T$. Then we have $\mathrm{E}Z_{n,i}=0$ and $s_n^2=\sum_{i=1}^n\mathrm{var}(Z_{n,i})=1$.

Since $|Z_{n,i}|^2\le\lambda_{\max}(\mathbf{M}_{nm}^{-1})\|\mathbf{g}_i^{(n)}\|^2$, $\max_{1\le i\le n}|Z_{n,i}|\le(\max_{1\le i\le n}\|\mathbf{g}_i^{(n)}\|)/\lambda_{\min}^{1/2}(\mathbf{M}_{nm})$. The double array $\{Z_{n,i}$, for $i=1,2,\dots,n$; $n=1,2,\dots\}$ satisfies (10). By Lemma 3, taking $\phi(n)=(\log n)^{1+\delta}$ and $A_n=s_n\phi(n)$, we have

$$\lim_{n\to\infty}\frac{\mathbf{e}_1^T\mathbf{M}_{nm}^{-1/2}\mathbf{g}_{nm}}{(\log n)^{1+\delta}}=0\qquad\text{a.s.}\qquad\square$$

To prove the existence and strong consistency of the GEE estimator when $n\to\infty$, we employ the following local condition.

(L_s) In a neighborhood of β_0, say N, there exists a constant $c_0>0$ (independent of n) and $\delta>0$ such that, when $n\to\infty$,

$$\lambda_{\min}\big(\mathcal{D}_{nm}(\beta)^T\mathbf{M}_{nm}^{-1}\mathcal{D}_{nm}(\beta)\big)\ge c_0(\log n)^{2(1+\delta)}$$

and

$$\mathcal{D}_{nm}(\beta)\text{ is nonsingular}\qquad\text{a.s. for }\beta\in N.$$

REMARK 12. When $\mathcal{D}_{nm}(\beta)$ is symmetric, (L_s) is implied by

$$\lambda_{\min}\{\mathcal{D}_{nm}(\beta)\}\ge c_0\lambda_{\max}^{1/2}(\mathbf{M}_{nm})(\log n)^{1+\delta}\qquad\text{a.s.}$$

This is comparable to the strong consistency conditions (S_δ) and (S_δ^*) of Fahrmeir and Kaufmann (1985) for multivariate generalized linear models [note, in their setting, $\mathcal{D}_n(\beta)$ is symmetric and $\mathbf{H}_n \equiv \mathbf{M}_n \equiv \mathbf{F}_n$]. The difference is that, in their conditions, $(\log n)^{1+\delta}$ is replaced by $\lambda_{\max}^{\delta}(\mathbf{M}_n)$. Under our setting, the term $(\log n)^{1+\delta}$ is unavoidable; see Remark 9.

Now, we state our main theorem for existence and strong consistency of the GEE estimator.

THEOREM 5. *Suppose $\mathbf{g}_{m_i,i}$, $i = 1, 2, \ldots, n$, the summands of the GEE score function $\mathbf{g}_{nm}$, form a strong infinitesimal double array sequence. Under conditions* (L_s), *there exist a sequence of random variables $\widehat{\beta}_{nm}$ and a random number n_0, such that*

$$P\bigl(\mathbf{g}_{nm}(\widehat{\beta}_{nm}) = 0, \textit{ for all } n \ge n_0\bigr) = 1,$$

and when $n \to \infty$,

$$\widehat{\beta}_{nm} \to \beta \qquad \textit{a.s.}$$

PROOF. For any $\varepsilon > 0$, define $B(\varepsilon) = \{\beta : \|\beta - \beta_0\| \le \varepsilon\}$. By Taylor's expansion, there exists a $\bar{\beta}$ between β and β_0 such that $\mathbf{M}_{nm}^{-1/2}(\mathbf{g}_{nm}(\beta) - \mathbf{g}_{nm}) = \mathbf{M}_{nm}^{-1/2}\mathcal{D}_{nm}(\bar{\beta})(\beta - \beta_0)$. Therefore,

$$\inf_{\beta \in \partial B(\varepsilon)} \bigl\|\mathbf{M}_{nm}^{-1/2}\bigl(\mathbf{g}_{nm}(\beta) - \mathbf{g}_{nm}\bigr)\bigr\| \ge \varepsilon \inf_{\beta \in B(\varepsilon)} \lambda_{\min}^{1/2}\bigl(\mathcal{D}_{nm}(\beta)^T \mathbf{M}_{nm}^{-1} \mathcal{D}_{nm}(\beta)\bigr). \tag{14}$$

By Lemma 4, we have

$$\frac{\|\mathbf{M}_{nm}^{-1/2}\mathbf{g}_{nm}\|}{(\log n)^{1+\delta}} \to 0 \qquad \text{when } n \to \infty \text{ a.s.}$$

So, by (L_s), when n is sufficiently large,

$$\|\mathbf{M}_{nm}^{-1/2}\mathbf{g}_{nm}\| \le \inf_{\beta \in \partial B_{nm}(r)} \bigl\|\mathbf{M}_{nm}^{-1/2}\bigl(\mathbf{g}_{nm}(\beta) - \mathbf{g}_{nm}\bigr)\bigr\| \qquad \text{a.s.} \tag{15}$$

By (L_s), $\mathcal{D}_{nm}(\beta)$ is nonsingular. Under the regularity assumptions listed at the end of Section 1, the mapping $T : \beta \to \mathbf{M}_{nm}^{-1/2}\mathbf{g}_{nm}(\beta)$ is a smooth injection from $B_{nm}(r)$ to $T(B_{nm}(r))$. According to Lemma A of Chen, Hu and Ying (1999), (15) implies that almost surely there exists a $\widehat{\beta}_{nm} \in B(\varepsilon)$ such that $\mathbf{g}_{nm}(\widehat{\beta}_{nm}) = 0$. Note, $\widehat{\beta}_{nm} \in B(\varepsilon)$ implies $\|\widehat{\beta}_{nm} - \beta_0\| < \varepsilon$ and, hence, the strong consistency conclusion also follows. $\square$

As in Section 2.1, condition (L_s) can be reduced to be stated only in terms of $\mathbf{H}_{nm}$ and $\mathcal{D}_{nm}(\beta)$.

(L_s^*) In a neighborhood of β_0, say N, there exist a constant $c_0 > 0$ and some $\delta > 0$, such that, when $n \to \infty$,

$$\lambda_{\min}\big(\mathcal{D}_{nm}(\beta)^T \mathbf{H}_{nm}^{-1} \mathcal{D}_{nm}(\beta)\big) \geq c_0 \tau_{nm} (\log n)^{2(1+\delta)}$$

and

$$\mathcal{D}_{nm}(\beta) \text{ is nonsingular} \qquad \text{a.s. for } \beta \in N.$$

Similar to Remark 11, when $\mathcal{D}_{nm}(\beta)$ is symmetric, (L_s^*) is implied by

$$\lambda_{\min}\{\mathcal{D}_{nm}(\beta)\} \geq c_0 \lambda_{\max}^{1/2}(\mathbf{H}_{nm})(\tau_{nm})^{1/2}(\log n)^{1+\delta} \qquad \text{a.s. for } \beta \in N.$$

THEOREM 6. *The results of Theorem* 5 *hold if condition* (L_s) *is replaced by* (L_s^*).

PROOF. The proof is similar to that of Theorem 5, except that we prove the inequality $\|\mathbf{H}_{nm}^{-1/2}\mathbf{g}_{nm}\| \leq \inf_{\beta \in \partial B(\varepsilon)} \|\mathbf{H}_{nm}^{-1/2}(\mathbf{g}_{nm}(\beta) - \mathbf{g}_{nm})\|$ here. Details are omitted. □

In a special case when $m = m(n)$ is bounded above as $n \to \infty$, to get a strong consistency result, Lemma 2 of Wu (1981) can be used, along with the following two alternative conditions:

(I_s)^{sp} $\lambda_{\min}(\mathbf{M}_n) \to \infty$ for some $\delta > 0$, when $n \to \infty$.

(L_s)^{sp} In a neighborhood of β_0, say N, there exists a constant $c_0 > 0$, independent of n, such that, when $n \to \infty$, for any $p \times 1$ vector λ, $\|\lambda\| = 1$,

$$\lambda^T \mathcal{D}_n(\beta)\lambda \geq c_0 \lambda_{\max}^{(1/2)+\delta}(\mathbf{M}_n) \qquad \text{a.s. for } \beta \in N.$$

THEOREM 7. *Suppose* $m = \max_{1 \leq i \leq n} m_i$ *is bounded above by a finite constant, independent of* n. *Under conditions* (I_s)^{sp} *and* (L_s)^{sp}, *there exist a sequence of random variables* $\widehat{\beta}_n$ *and a random number* n_0, *such that*,

$$P\big(\mathbf{g}_n(\widehat{\beta}_n) = 0, \textit{ for all } n \geq n_0\big) = 1$$

and when $n \to \infty$,

$$\widehat{\beta}_n \to \beta \qquad \textit{a.s.}$$

The proof is similar to the proofs of Theorem 5 and Theorem 6, and is therefore omitted.

Finally, before we end this section, let us briefly consider the case when n is bounded and $m \to \infty$. Without loss of generality, we can assume that $n = 1$, that is, we only have one single cluster. In this case, the score $\mathbf{g}_{1m}$ is a weighted sum of dependent random variables. We need to establish a strong law of large numbers similar to Lemma 2 of Wu (1981) for the weighted sum of dependent variables. To do so, obviously, we need to make assumptions on the dependence structure. There are many ways to do so. For example, one could impose martingale assumptions or mixing conditions on the dependent sequence. Or, one can use Serfling's (1970) stability results where only moment conditions are assumed. This is a topic that deserves further development. As the first step in this development, we provide below a result for a special set of dependent sequences that is fairly common. Here, the goal is to provide simple and "easy to verify" conditions, rather than the best set of conditions.

Suppose, for $j = 1, 2, \ldots, m$, y_j is the jth observation of a cluster of m dependent random variables, and μ_j and σ_j are its marginal mean and marginal standard deviation, respectively. Denote $e_j = (y_j - \mu_j)/\sigma_j$. We assume that, for any $j < j'$,

$$\mathrm{E}e_j^2 e_{j'}^2 \le \rho_{j'-j}(\mathrm{E}e_j^4 \mathrm{E}e_{j'}^4)^{1/2}, \tag{16}$$

where $\sum_{k=1}^{+\infty} \rho_k^{1/2} < +\infty$. We set $\rho_0 = 1$. We have the following lemma.

LEMMA 5. *Suppose $e_1, \ldots, e_m$ are a dependent sequence satisfying* (16) *and $\mathrm{E}e_j^4 \le M < \infty$, for $j = 1, \ldots, m$. If we have a set of double array coefficients $\{a_{m,j},\ j = 1, 2, \ldots, m\}$, a constant K and an integer l_0, such that*

$$\sum_{j=1}^{m} a_{m,j}^2 < K < +\infty \qquad \textit{for } m = l_0, l_0 + 1, \ldots,$$

then we have

$$\lim_{m \to +\infty} \frac{\sum_{j=1}^{m} a_{m,j} e_j}{m^{1/4}(\log m)^{(1+\delta)/4}} = 0.$$

The proof of Lemma 5 can be found in Appendix B.

In the case when $n = 1$ and $m \to \infty$, we employ the following condition to assure the almost sure existence and strong consistency of a GEE estimator.

$(\mathrm{L}_s)_{\mathrm{Bm}}$ In a neighborhood of β_0, say N, there exists a constant $c_0 > 0$, independent of m, such that, when $m \to \infty$,

$$\lambda_{\min}\big(\mathcal{D}_{1m}(\beta)^T \mathbf{M}_{1m}^{-1} \mathcal{D}_{1m}(\beta)\big) \ge c_0 m^{1/2} (\log m)^{(1+\delta)/2}$$

and

$$\mathcal{D}_{1m}(\beta) \text{ is nonsingular almost surely.}$$

THEOREM 8. *Suppose in a single cluster the dependent responses satisfy* (16) *and the fourth marginal moments of $(y_{1j} - \mu_{1j})/\sigma_{1j}$ are bounded above. If $\lambda_{\min}(\bar{\mathbf{R}}_1) \ge c_1$, for some constant $c_1 > 0$ independent of m, and condition* $(\mathrm{L}_s)_{\mathrm{Bm}}$

holds, then there exist a sequence of random variables $\hat{\beta}_m$ and a random number m_0, such that

$$P\big(\mathbf{g}_{1m}(\beta) = 0, \textit{ for all } m \geq m_0\big) = 1$$

and when $m \to +\infty$,

$$\hat{\beta}_m \to \beta_0 \qquad a.s.$$

PROOF. The proof is similar to that of Theorem 5, except that we use Lemma 5 instead of Lemmas 3 and 4. In Lemma 5, for any given $p \times 1$ vector λ, $\|\lambda\| = 1$, we take $a_{m,j}$ to be the jth element of the $1 \times m$ vector $\lambda^T \mathbf{M}_{1m}^{-1/2} D_1 V_1^{-1}(\alpha) A_1^{1/2}$. So $\sum_{i=1}^n a_{m,j}^2 = \lambda^T \mathbf{M}_{1m}^{-1/2} X_1 \Delta_1 A_1^{1/2} \mathbf{R}_1^{-1} \mathbf{R}_1^{-1} A_1^{1/2} \Delta_1 X_1 \mathbf{M}_{1m}^{-1/2} \lambda \leq \lambda_{\max}(\bar{\mathbf{R}}^{-1}) \leq {c_1}^{-1} < \infty$. From Lemma 5, we have

$$\limsup_{m \to \infty} \frac{\|\mathbf{M}_{1m}^{-1} \mathbf{g}_m(\beta)\|}{m^{1/4} (\log m)^{(1+\delta)/4}} = 0.$$

The rest of the proof proceeds as the proof of Theorem 5 from (14) on. $\square$

5. Examples. In this section, we first consider the linear regression model and show how consistency and asymptotic normality can be deduced in this special case. Subsequently, we provide some corollaries of general interest for cases of practical importance, such as (i) marginal generalized linear models with compact covariate set, (ii) marginal Poisson regression models and (iii) marginal generalized linear models with bounded responses (binomial or polytomous regression models). Poisson, binomial and polytomous regressions are the most commonly used models in categorical data analysis, and the assumption of covariate regressors in a compact set is satisfied in many applications.

EXAMPLE 5.1 (Linear regression model). Suppose the jth individual response in the ith cluster follows a linear regression model,

$$y_{ij} = \mathbf{x}_{ij}^T \beta + \varepsilon_{ij},$$

where ε_{ij} need not have a normal distribution. Without loss of generality, we assume $\mathrm{E}\varepsilon_{ij} = 0$ and $\mathrm{var}(\varepsilon_{ij}) = 1$. It is easy to see that the score function is

$$\mathbf{g}_{nm}(\beta) = \sum_{i=1}^n \mathbf{X}_i^T \mathbf{R_i}^{-1}(\alpha)(\mathbf{y}_i - \mathbf{X}_i \beta),$$

and the information matrices in (4a)–(4c) are

$$\mathcal{D}_{nm}(\beta) = \mathbf{H}_{nm}(\beta) = \mathbf{H}_{nm} = \sum_{i=1}^n \mathbf{X}_i^T \mathbf{R_i}^{-1}(\alpha) \mathbf{X}_i$$

and

$$\mathbf{M}_{nm} = \sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{R_i}^{-1}(\alpha)\bar{\mathbf{R}}_\mathbf{i}\mathbf{R_i}^{-1}(\alpha)\mathbf{X}_i.$$

Under this model, conditions (L_w), (L_w^*) and (CC) are trivially true. By Theorems 1, 2 and 3, if either (I_w) or (I_w^*) is true, then there exists a consistent GEE estimator $\widehat{\beta}_{nm}$, and $\mathbf{M}_{nm}^{-1/2}\mathbf{H}_{nm}(\widehat{\beta}_{nm} - \beta_0)$ and $\mathbf{M}_{nm}^{-1/2}\mathbf{g}_{nm}$ are asymptotically identically distributed. Although conditions (I_w) and (I_w^*) allow no general reduction, both $\mathbf{M}_n$ and $\mathbf{H}_n$ have a simpler form and they do not depend on the value of β_0.

If $\varepsilon_i = (\varepsilon_{i1}, \dots, \varepsilon_{im_i})^T$, for $i = 1, \dots, n$, are normal random vectors, $\mathbf{M}_{nm}^{-1/2} \times \mathbf{H}_{nm}(\widehat{\beta}_{nm} - \beta_0)$ will follow a standard multivariate normal distribution. Otherwise, we can use condition (N_δ) to ensure the asymptotic normality of the GEE estimator $\widehat{\beta}_{nm}$. In particular, (8) can be implied by

$$(c_{nm}m)^{1+\delta}\widetilde{\lambda}_{nm}^{2+\delta} \max_{1 \le i \le n} \lambda_{\max}(\mathbf{X}_i \mathbf{H}_{nm}^{-1}\mathbf{X}_i^T) \to 0 \qquad \text{for some } \delta > 0. \tag{17}$$

In the nonlongitudinal case, each subject only has one observation (i.e., $m_1 \equiv m_2 \equiv \cdots \equiv m_n \equiv 1$) and $\mathbf{X}_i$ is a $1 \times p$ vector $\mathbf{x}_i^T$. In this case, both (I_w) and (I_w^*) reduce to

$$\lambda_{\min}\left(\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^T\right) \to \infty$$

and (17) reduces to

$$\widetilde{\gamma}_n = \max_{1 \le i \le n} \mathbf{x}_i^T \left(\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^T\right)^{-1} \mathbf{x}_i \to 0.$$

These are exactly the conditions for consistency and normality of least squares estimators in the standard linear regression models [cf. Eicker (1967)].

Next, we provide several corollaries of general interest for some cases of practical importance. Unlike the case of the linear regression models, we need to verify conditions (L_w) or (L_w^*), as well as condition (CC). The verification of (L_w), (L_w^*) and (CC) involves lengthy calculations in most cases. To save computing effort, we provide in Appendix C (Theorems A1 and A2) sufficient conditions for (L_w), (L_w^*) and (CC) under a very general setting. In particular, we make the following (smoothness) assumption on the functions $a(\theta)$ and $h(\eta)$:

(AH) $k_{nm}^{[l]}$, for $l = 1, 2, 3, 4$, are bounded, where we denote

$$k_{nm}^{[1]} = \sup_{\beta \in B_n^*(r)} \max_{ij} \left\{ \left| \frac{a^{(3)}(\theta_{ij})}{a''(\theta_{ij})} \right| \right\},$$

$$k_{nm}^{[2]} = \sup_{\beta \in B_n^*(r)} \max_{ij} \left\{ \left| \frac{h''(\eta_{ij})}{\{h'(\eta_{ij})\}^2} a''(\theta_{ij}) \right| \right\},$$

$$k_{nm}^{[3]} = \sup_{\beta \in B_n^*(r)} \max_{ij} \left\{ \left| \frac{a^{(4)}(\theta_{ij})}{a''(\theta_{ij})} \right| \right\}$$

and

$$k_{nm}^{[4]} = \sup_{\beta \in B_n^*(r)} \max_{ij} \left\{ \left| \frac{h^{(3)}(\eta_{ij})}{\{h'(\eta_{ij})\}^3} \{a''(\theta_{ij})\}^2 \right| \right\}.$$

Assumption (AH) is usually satisfied in commonly used models. In models with canonical link functions, $k_{nm}^{[1]} \equiv k_{nm}^{[2]}$ and $k_{nm}^{[3]} \equiv k_{nm}^{[4]}$. Theorems A1 and A2 are developed under assumption (AH) and they are used to prove the corollaries in this section.

Define

$$\gamma_{nm}^{(0)} = \max_{1 \le i \le n} \max_{1 \le j \le m_i} \mathbf{x}_{ij}^T \mathbf{H}_{nm}^{-1} \mathbf{x}_{ij}.$$

We also adopt the notation

$$\kappa_{nm} = \max_{i,j} \left\{ u'(\mathbf{x}_{ij}^T \beta_0) \right\}^2 \quad \text{and} \quad \pi_{nm} = \frac{\max_{1 \le i \le n} \{\lambda_{\max}(\mathbf{R}_i^{-1}(\alpha))\}}{\min_{1 \le i \le n} \{\lambda_{\min}(\mathbf{R}_i^{-1}(\alpha))\}}.$$

In models with canonical link functions, $\kappa_{nm} \equiv 1$. When $\mathbf{R}_i(\alpha) \equiv \mathbf{I}$, we have $\pi_{nm} = 1$.

First, we consider the case of generalized linear models with compact regressors. We introduce the following assumptions:

(C1) The covariates $\{\mathbf{x}_{ij},\ j = 1, \dots, m_i,\ i = 1, \dots, n\}$, for $n = 1, 2, \dots$, are in a compact set $\mathcal{X}$ with $\mathbf{x}^T \beta \in \Theta^o$ for any $\mathbf{x} \in \mathcal{X}$ and $\beta \in \mathcal{B}$.

(C2) (i) $\tau_{nm} \lambda_{\max}(\mathbf{H}_{nm}^{-1}) \to 0$;

(ii) $\pi_{nm}^2 \tau_{nm} m \gamma_{nm}^{(0)} \to 0$.

(C3) $(c_{nm})^{1+\delta} (\widetilde{\lambda}_{nm} m)^{2+\delta} \gamma_{nm}^{(0)} \to 0$ for some $\delta > 0$.

We have the following corollary.

COROLLARY 1 (Generalized linear model with compact regressors). *Suppose assumptions* (C1) *and* (C2) *hold.*

(a) *There exists a sequence of random variables* $\widehat{\beta}_{nm}$ *such that* $\widehat{\beta}_{nm} \to \beta_0$ *in probability, and* $\mathbf{M}_{nm}^{-1/2} \mathbf{H}_{nm} (\widehat{\beta}_{nm} - \beta_0)$ *and* $\mathbf{M}_{nm}^{-1/2} \mathbf{g}_{nm}$ *have the same asymptotic distribution.*

(b) *If, further,* (C3) *is true, then, when* $n \to \infty$,

$$\mathbf{M}_{nm}^{-1/2} \mathbf{H}_{nm} (\widehat{\beta}_{nm} - \beta_0) \to N(0, \mathbf{I}) \qquad \text{in distribution.}$$

PROOF. It is clear that (C2)(i) is (I_w^*). To verify conditions (L_w^*) and (CC), we use Theorem A1(ii) and Theorem A2 provided in Appendix C. Note that $a''(u(t))$, $h'(t)$ and their first and second derivatives are all continuous functions. By conditions (C1) and (C2)(i), it is easy to see that, for $\beta \in B_{nm}^*(r)$, $a''(\theta_{ij})$, $|a^{(3)}(\theta_{ij})|$, $|a^{(4)}(\theta_{ij})|$, and $|h''(\eta_{ij})|$ are uniformly bounded above, and $a''(\theta_{ij})$, $|h'(\eta_{ij})|$ are uniformly bounded below away from zero. Therefore, assumption (AH) is true, κ_{nm} is bounded above, and $b_{nm} = \min_{i,j}\{\sigma_{ij}^2\}$ is bounded below away from zero. Thus, by Theorem A1(ii) and Theorem A2, conditions (L_w^*) and (CC) hold. By Theorem 2 and Theorem 3, part (a) follows.

To prove the asymptotic normality in part (b), we use Theorem 4. Since $\mathrm{E}(y_{ij}^{*\,2(1+1/\delta)})$ is a continuous function of $\mathbf{x}_{ij}$ for each i, j, it is easy to see from (C1) that it is uniformly bounded above. Also by (C1), there exists a finite constant K such that $\gamma_{nm}^{(D)} \le K\widetilde{\lambda}_{nm} m \gamma_{nm}^{(0)}$. Hence, (8) can be implied immediately from (C3). By Theorem 4, part (b) follows. $\square$

REMARK 13. One can prove that (C2) and (C3) are implied by the following conditions (in terms of the design matrix $\mathbf{X}_i$):

(C2)′ $(\pi_{nm})^3 m^2 \lambda_{\max}\{(\sum_{i=1}^n \mathbf{X}_i^T\mathbf{X}_i)^{-1}\} \to 0$.
(C3)′ $\pi_{nm}(c_{nm}\widetilde{\lambda}_{nm})^{1+\delta} m^{2+\delta}\lambda_{\max}\{(\sum_{i=1}^n \mathbf{X}_i^T\mathbf{X}_i)^{-1}\} \to 0$ for some $\delta > 0$.

REMARK 14. In the nonlongitudinal case, $m_1 \equiv m_2 \equiv \cdots \equiv m_n \equiv 1$. Condition (C1) is exactly condition (R_c^*)(i) of Fahrmeir and Kaufmann (1985). In addition, both conditions (C2)′ and (C3)′ reduce to

$$\lambda_{\min}\left(\sum_{i=1}^n \mathbf{X}_i^T\mathbf{X}_i\right) \to \infty.$$

This is the same as condition (R_c^*)(ii) of Fahrmeir and Kaufmann (1985).

In a *marginal Poisson model*, the marginal distribution of y_{ij} is

$$P(y_{ij} = y) = \exp(y\theta_{ij} - e^{\theta_{ij}})/y!, \tag{18}$$

where $\theta_{ij} = \mathbf{x}_{ij}^T\beta$. Under this model, $a(\theta_{ij}) = e^{\theta_{ij}}$, $\mu_{ij}(\beta) = e^{\eta_{ij}}$, $h(\eta_{ij}) = e^{\eta_{ij}}$ and $\eta_{ij} = \theta_{ij}$.

We consider the following assumptions:

(P1) (i) $\tau_{nm}\lambda_{\max}(\mathbf{H}_{nm}^{-1}) \to 0$;
(ii) $\pi_{nm}^2\tau_{nm} m\gamma_{nm}^{(0)} \to 0$;
(iii) $\{\mu_{ij}(\beta_0) = \exp(\mathbf{x}_{ij}^T\beta_0),\ j = 1, \ldots, m_n \text{ and } i = 1, \ldots, n\}$ are bounded above and below away from zero.

(P2) $(c_{nm})^{1+\delta}(\widetilde{\lambda}_{nm} m)^{2+\delta}\gamma_{nm}^{(0)} \to 0$ for some $\delta > 0$.

COROLLARY 2 (Poisson model). *Suppose model* (18) *and assumption* (P1) *hold.*

(a) *There exists a sequence of random variables* $\widehat{\beta}_{nm}$ *such that* $\widehat{\beta}_{nm} \to \beta_0$ *in probability, and* $\mathbf{M}_{nm}^{-1/2}\mathbf{H}_{nm}(\widehat{\beta}_n - \beta_0)$ *and* $\mathbf{M}_{nm}^{-1/2}\mathbf{g}_{nm}$ *have the same asymptotic distribution.*

(b) *If, further,* (P2) *is true, then, when* $n \to \infty$,

$$\mathbf{M}_{nm}^{-1/2}\mathbf{H}_{nm}(\widehat{\beta}_{nm} - \beta_0) \to N(0, \mathrm{I}) \qquad \textit{in distribution.}$$

PROOF. It is clear that (I_w^*) is (P1)(i). Under model (18), $\kappa_{nm} \equiv 1$ and $k_{nm}^{[l]} \equiv 1$, for $l = 1, 2, 3, 4$. Assumption (AH) is trivially true. By Theorem A1(ii) and Theorem A2, we have both (L_w^*) and (CC). By Theorem 2 and Theorem 3, we can conclude part (a).

To prove part (b), we only need to verify condition (N_δ). Note that, for δ in (P2), there exist some constants C_1, C_2 depending only on δ such that $\mathrm{E}(y_{ij}^*)^{2(1+1/\delta)} \le C_1 + C_2\{\mu_{ij}(\beta_0)\}^{2(1+1/\delta)+1}$, which is uniformly bounded above by (P1)(iii). Also, it is clear that (P2) implies (8). So, (N_δ) holds. By Theorem 4, we have the asymptotic normality result in part (b). □

REMARK 15. In the nonlongitudinal case with $m_1 \equiv m_2 \equiv \cdots \equiv m_n \equiv 1$, condition (P1)(i) becomes $\lambda_{\min}(\mathbf{H}_{nm}) \to \infty$, which is exactly the same as condition (D) of Fahrmeir and Kaufmann (1985). Both conditions (P1)(ii) and (P2) reduce to $\gamma_n^{(0)} = \max_i\{\mathbf{x}_i^T \mathbf{H}_n^{-1}\mathbf{x}_i\} \to 0$, which is equivalent to (3.14) of Fahrmeir and Kaufmann (1985). Although we require an extra condition (P1)(iii) here from Corollary 2, we find that condition (P1)(iii) can be dropped in the nonlongitudinal case, based on a step-by-step examination of the proofs of our related theorems. But, in general, in order to ensure the uniform convergence and asymptotic normality in the longitudinal case, we need to have condition (P1)(iii).

Finally, we consider regression models with bounded responses. These include binary and categorical regression models. First let us verify assumption (AH) for some special models:

Logistic regression models. In this case, $a(\theta_{ij}) = \log(1 + e^{\theta_{ij}})$, $h(\eta_{ij}) = e^{\eta_{ij}}/(1+e^{\eta_{ij}})$ and $\theta_{ij} = \eta_{ij}$. We have $k_{nm}^{[1]} = k_{nm}^{[2]} = \sup_{\beta\in B_n^*(r)} \max_{i,j} |1 - 2h(\eta_{ij})| \le 3$ and $k_{nm}^{[3]} = k_{nm}^{[4]} = \sup_{\beta\in B_n^*(r)} \max_{i,j} |1 - 6\{h(\eta_{ij})(1 - h(\eta_{ij}))\}| \le 2.5$. So, assumption (AH) holds.

Probit regression models. In this case, $a(\theta_{ij}) = \log(1 + e^{\theta_{ij}})$, $h(\eta_{ij}) = \Phi(\eta_{ij})$, and $h(\eta_{ij}) = a'(\theta_{ij})$. Note that $u'(\eta_{ij}) = \Psi(\eta_{ij})/\{\Phi(\eta_{ij})(1 - \Phi(\eta_{ij}))\}$. Here $\Phi(\eta_{ij})$ and $\Psi(\eta_{ij})$ are the cumulative distribution function and the density function of the standard normal distribution, respectively. From direct computation, $k_{nm}^{[1]} = \sup_{\beta\in B_n^*(r)} \max_{i,j} |1 - 2\Phi(\eta_{ij})|$, $k_{nm}^{[2]} = \sup_{\beta\in B_n^*(r)} \max_{i,j}\{|\eta_{ij}|\Phi(\eta_{ij}) \times (1 - \Phi(\eta_{ij}))/\Psi(\eta_{ij})\}$, $k_{nm}^{[3]} = \sup_{\beta\in B_n^*(r)} \max_{i,j} |1 - 6\Phi(\eta_{ij})(1 - \Phi(\eta_{ij}))|$ and $k_{nm}^{[4]} = \sup_{\beta\in B_n^*(r)} \max_{i,j}[|\eta_{ij}^2 - 1|\{\Phi(\eta_{ij})(1 - \Phi(\eta_{ij}))\}^2/\{\Psi(\eta_{ij})\}^2]$. Note that

$\lim_{\eta_{ij} \to \pm\infty} |\eta_{ij}|\Phi(\eta_{ij})(1-\Phi(\eta_{ij}))/\Psi(\eta_{ij}) = 1$. All the above terms are bounded, and hence assumption (AH) holds.

Models with canonical link functions. In this case, $\theta_{ij} = \eta_{ij}$ and $k_{nm}^{[1]} = k_{nm}^{[2]} = \sup_{\beta \in B_n^*(r)} \max_{i,j} |a^{(3)}(\theta_{ij})|/\{a''(\theta_{ij})\}$, $k_{nm}^{[3]} = k_{nm}^{[4]} = \sup_{\beta \in B_n^*(r)} \max_{i,j} |a^{(4)}(\theta_{ij})|/\{a''(\theta_{ij})\}$. When y_{ij} has a density function of form (1), we have $|a^{(3)}(\theta_{ij})| = |\mathrm{E}(y_{ij} - \mu_{ij})^3/\phi^2|$ and $|a^{(4)}(\theta_{ij})| = |\mathrm{E}(y_{ij} - \mu_{ij})^4/\phi^3 - 3\{a''(\theta_{ij})\}^2/\phi|$. If the responses y_{ij} are bounded, then $|a^{(3)}(\theta_{ij})|$ and $|a^{(4)}(\theta_{ij})|$ are bounded, for each fixed scale parameter ϕ. So, if one assumes that $a''(\theta_{ij})$ are uniformly bounded below away from zero, then assumption (AH) holds.

We set the following assumptions:

(B1) Responses y_{ij} are bounded.
(B2) (i) $\tau_{nm}\lambda_{\max}(\mathbf{H}_{nm}^{-1}) \to 0$;
(ii) $\pi_{nm}^2 \tau_{nm}\kappa_{nm} m \gamma_{nm}^{(0)} / \min_{i,j} \sigma_{ij}^2 \to 0$.
(B3) $c_{nm}\widetilde{\lambda}_{nm}^2 \kappa_{nm} m^2 \gamma_{nm}^{(0)} / \min_{i,j} \sigma_{ij}^2 \to 0$.

We now have the following corollary.

COROLLARY 3 (Generalized linear model with bounded responses). *Suppose assumptions* (AH), (B1) *and* (B2) *hold.*

(a) *There exists a sequence of random variables* $\widehat{\beta}_{nm}$ *such that* $\widehat{\beta}_{nm} \to \beta_0$ *in probability, and* $\mathbf{M}_{nm}^{-1/2}\mathbf{H}_{nm}(\widehat{\beta}_n - \beta_0)$ *and* $\mathbf{M}_{nm}^{-1/2}\mathbf{g}_{nm}$ *have the same asymptotic distribution.*

(b) *If further,* (B3) *is true, then, when* $n \to \infty$,

$$\mathbf{M}_{nm}^{-1/2}\mathbf{H}_{nm}(\widehat{\beta}_{nm} - \beta_0) \to N(0, \mathrm{I}) \qquad \textit{in distribution}.$$

PROOF. Under assumptions (AH) and (B2), the results in part (a) are immediate by Theorem A2, Theorem 2 and Theorem 3. In order to prove the result in part (b), we directly verify the Lindeberg condition on the double arrays sequence $Z_{nm,i} = \lambda^T \mathbf{M}_{nm}^{-1/2}\mathbf{D}_i^T \mathbf{V}_i^{-1}(\mathbf{y}_i - \mu_i)$. By (B1), $\mathbf{y}_i^{*T}\mathbf{y}_i^*$ is uniformly bounded above by $m/\min_{i,j}\sigma_{ij}^2$ times a constant. So, by (B3), for any given ε, when n is large enough, $\mathbf{y}_i^{*T}\mathbf{y}_i^* < \varepsilon^2\{c_{nm}\widetilde{\lambda}_{nm}^2\kappa_{nm} m \gamma_{nm}^{(0)}\}^{-1}$. Therefore, when $n \to \infty$,

$$\begin{aligned}
&\sum_{i=1}^n \mathrm{E}\big[Z_{nm,i}^2 \mathbf{I}(|Z_{nm,i}| > \varepsilon)\big] \\
&\qquad \le \sum_{i=1}^n c_{nm}\widetilde{\lambda}_{nm}\kappa_{nm} m \gamma_{nm}^{(0)} \mathrm{E}\left[\mathbf{y}_i^{*T}\mathbf{y}_i^* \mathbf{I}\left\{\mathbf{y}_i^{*T}\mathbf{y}_i^* > \frac{\varepsilon^2}{c_{nm}\widetilde{\lambda}_{nm}^2\kappa_{nm} m \gamma_{nm}^{(0)}}\right\}\right] \to 0.
\end{aligned}$$

By the Lindeberg central limit theorem and the Cramér–Wold theorem, part (b) follows. □

REMARK 16. In the nonlongitudinal case, with $m_1 \equiv m_2 \equiv \cdots \equiv m_n \equiv 1$, conditions (B2)(ii) and (B3) are implied by $\kappa_{nm}\gamma_n^{(0)}/\min_{i,j}\sigma_{ij}^2 \to 0$. In the special case of probit models, $\lim_{\eta_{ij}\to\pm\infty} u'(\eta_{ij})/|\eta_{ij}| = 1$. In such a case, if $\min_{i,j}\sigma_{ij}^2$ is bounded away from zero, then conditions (B2)(ii) and (B3) become $\max_{ij}\{|\eta_{ij}|^2\}\gamma_n^{(0)} \to 0$. This is equivalent to condition (A)(c) of Fahrmeir and Kaufmann (1986).

APPENDIX A

Formulas for $\mathbf{B}_{nm}(\boldsymbol{\beta})$ and $\mathcal{E}_{nm}(\boldsymbol{\beta})$. One can write $\mathbf{B}_{nm}(\beta) = \mathbf{B}_{nm}^{[1]}(\beta) + \mathbf{B}_{nm}^{[2]}(\beta)$ and $\mathcal{E}_{nm}(\beta) = \mathcal{E}_{nm}^{[1]}(\beta) + \mathcal{E}_{nm}^{[2]}(\beta)$, where

$$\mathbf{B}_{nm}^{[1]}(\beta) = \sum_{i=1}^{n} \mathbf{X}_i^T \operatorname{diag}\left[\mathbf{R}_i^{-1}(\alpha)\mathbf{A}_i^{-1/2}(\beta)\big(\mu_i - \mu_i(\beta)\big)\right]\mathbf{G}_i^{[1]}(\beta)\mathbf{X}_i,$$

$$\mathbf{B}_{nm}^{[2]}(\beta) = \sum_{i=1}^{n} \mathbf{X}_i^T \boldsymbol{\Delta}_i(\beta)\mathbf{A}_i^{1/2}(\beta)\mathbf{R}_i^{-1}(\alpha) \operatorname{diag}\left[\big(\mu_i - \mu_i(\beta)\big)\right]\mathbf{G}_i^{[2]}(\beta)\mathbf{X}_i,$$

$$\mathcal{E}_{nm}^{[1]}(\beta) = \sum_{i=1}^{n} \mathbf{X}_i^T \operatorname{diag}\left[\mathbf{R}_i^{-1}(\alpha)\mathbf{A}_i^{-1/2}(\beta)(\mathbf{y}_i - \mu_i)\right]\mathbf{G}_i^{[1]}(\beta)\mathbf{X}_i$$

and

$$\mathcal{E}_{nm}^{[2]} = \sum_{i=1}^{n} \mathbf{X}_i^T \boldsymbol{\Delta}_i(\beta)\mathbf{A}_i^{1/2}(\beta)\mathbf{R}_i^{-1}(\alpha) \operatorname{diag}[(\mathbf{y}_i - \mu_i)]\mathbf{G}_i^{[2]}(\beta)\mathbf{X}_i.$$

Here, $\mathbf{G}_i^{[k]}(\beta) = \operatorname{diag}(q_k'(\eta_{ij}), \dots, q_k'(\eta_{ij}))$, for $k = 1, 2$, where

$$q_1(\eta_{ij}) = [a''(\theta_{ij})]^{-1/2} h'(\eta_{ij}), \qquad q_2(\eta_{ij}) = [a''(\theta_{ij})]^{-1/2},$$

and their derivatives

$$q_1'(\eta_{ij}) = -\frac{1}{2}\frac{a^{(3)}(\theta_{ij})}{[a''(\theta_{ij})]^{5/2}}\{h'(\eta_{ij})\}^2 + \frac{h''(\eta_{ij})}{[a''(\theta_{ij})]^{1/2}},$$

$$q_2'(\eta_{ij}) = -\frac{1}{2}\frac{a^{(3)}(\theta_{ij})}{[a''(\theta_{ij})]^{5/2}} h'(\eta_{ij}).$$

APPENDIX B

Proof of Lemma 3. For $c = c_0/(K+1)$, denote $Z_{n,i}^c = Z_{n,i}\mathbf{I}(|Z_{n,i}| \le cA_n/\phi(n))$, where $\mathbf{I}(C)$ is the indicator function for set C. We have

$$\frac{1}{A_n}\sum_{i=1}^n Z_{n,i} = \frac{1}{A_n}\sum_{i=1}^n \mathrm{E}Z_{n,i}^c + \frac{1}{A_n}\sum_{i=1}^n (Z_{n,i}^c - \mathrm{E}Z_{n,i}^c) + \frac{1}{A_n}\sum_{i=1}^n (Z_{n,i} - Z_{n,i}^c)$$
$$= I + II + III.$$

Since $\mathrm{E}Z_{n,i}^c = -\mathrm{E}Z_{n,i}\mathbf{I}(|Z_{n,i}| > cA_n/\phi(n))$ and by (12), we have

$$|I| = \left|\frac{1}{A_n}\sum_{i=1}^n \mathrm{E}Z_{n,i}^c\right| \le \frac{1}{A_n}\sum_{i=1}^n \mathrm{E}|Z_{n,i}|\mathbf{I}\left(|Z_{n,i}| \ge c\frac{A_n}{\phi(n)}\right)$$
$$\le \frac{\phi(n)}{cA_n^2}\sum_{i=1}^n \mathrm{E}|Z_{n,i}|^2 \to 0.$$

Take $W_{n,i} = (Z_{n,i}^c - \mathrm{E}Z_{n,i}^c)\phi(n)/A_n$. We have $\mathrm{E}|W_{n,i}|^s = \mathrm{E}|W_{n,i}|^2|W_{n,i}|^{s-2} \le \mathrm{E}|W_{n,i}|^2(2c)^{s-2} \le \sigma_{n,i}^2\phi(n)^2(2c)^{s-2}A_n^{-2}$, for $s > 2$. By the Bernstein inequality [Chow and Teicher (1988), page 111],

$$\sum_{n=1}^\infty P\left(\left|\frac{1}{A_n}\sum_{i=1}^n (Z_{n,i}^c - \mathrm{E}Z_{n,i}^c)\right| > \varepsilon\right)$$
$$= \sum_{n=1}^\infty P\left(\left|\sum_{i=1}^n W_{n,i}\right| > \varepsilon\phi(n)\right)$$
$$\le \sum_{n=1}^\infty \exp\left\{\frac{-[\varepsilon\phi(n)]^2}{2[(\sum_{i=1}^n \mathrm{var}(Z_{n,i}^c)\phi(n)^2)/A_n^2 + \varepsilon\phi(n)]}\right\}$$
$$\le C\sum_{n=1}^\infty \exp\left\{-\frac{\varepsilon}{2}\phi(n)\right\} < \infty,$$

for some finite constant C. The last two inequalities follow from (12) and (11), respectively. Therefore, by the Borel–Cantelli Lemma, $|II| \to 0$, a.s.

Finally, $Z_{n,i} - Z_{n,i}^c = Z_{n,i}\mathbf{I}(|Z_{n,i}| > cA_n/\phi(n))$. By (12), when n is large enough,

$$\left|\frac{1}{A_n}\sum_{i=1}^n (Z_{n,i} - Z_{n,i}^c)\right| \le \frac{K+1}{s_n\phi(n)}\sum_{i=1}^n |Z_{n,i}|\mathbf{I}(|Z_{n,i}| > c_0 s_n)$$
$$\le (K+1)\frac{\sum_{i=1}^n |Z_{n,i}|}{s_n\phi(n)}\mathbf{I}\left(\max_{1\le i\le n}|Z_{n,i}| > c_0 s_n\right).$$

By (10), $\mathbf{I}(\max_{1\le i\le n}|Z_{n,i}| > c_0 s_n) = 0$, a.s., when n is large enough. Thus, $|III| \to 0$, a.s.

Proof of Lemma 5. Note that

$$\begin{aligned}
&\mathrm{E}\biggl(\sum_{j=1}^{m} a_{m,j}e_j\biggr)^4 \\
&\quad= \sum_{\ell<j<k<s} a_{m,\ell}a_{m,j}a_{m,k}a_{m,s}\mathrm{E}(e_\ell e_j e_k e_s) + 4\sum_{\ell\neq j\neq k} a_{m,\ell}^2 a_{m,j}a_{m,k}\mathrm{E}e_\ell^2 e_j e_k \\
&\qquad + 6\sum_{j<\ell} a_{m,j}^2 a_{m,\ell}^2 \mathrm{E}e_j^2 e_\ell^2 + \sum_{j=1}^{m} a_{m,j}^4 \mathrm{E}e_j^4 \\
&\quad= I + II + III + IV.
\end{aligned}$$

First, we have

$$\begin{aligned}
I &\le \sum_{\ell<j<k<s} |a_{m,\ell}||a_{m,j}||a_{m,k}||a_{m,s}|(\mathrm{E}e_\ell^2 e_j^2 \cdot \mathrm{E}e_k^2 e_s^2)^{1/2} \\
&\le K \sum_{\ell<j<k<s} |a_{m,\ell}||a_{m,j}||a_{m,k}||a_{m,s}|\rho_{j-\ell}^{1/2}\rho_{s-k}^{1/2} \\
&\le K\biggl(\sum_{l<j} |a_{m,\ell}||a_{m,j}|\rho_{j-\ell}^{1/2}\biggr)^2 \le K\biggl(\sum_{l<j} \frac{a_{m,\ell}^2 + a_{m,j}^2}{2}\rho_{j-\ell}^{1/2}\biggr)^2 \\
&\le K\Bigl(\sum a_{m,j}^2\Bigr)^2\Bigl(\sum \rho_j^{1/2}\Bigr)^2.
\end{aligned}$$

Similarly, $II \le 4K(\sum_{i=1}^m a_{m,i}^2)^2(\sum \rho_j^{1/2})$, $III \le 6K(\sum_{i=1}^m a_{m,i}^2)^2$ and $IV \le K(\sum_{i=1}^m a_{m,i}^2)^2$.

So, by the Markov inequality and for a constant C,

$$\begin{aligned}
&P\biggl(\biggl|\frac{\sum_{j=1}^m a_{m,j}e_j}{m^{1/4}(\log m)^{(1+\delta)/4}}\biggr| > \varepsilon\biggr) \\
&\quad\le \frac{1}{\varepsilon^4}\,\frac{1}{m(\log m)^{1+\delta}}\biggl(\sum_{\ell=1}^m a_{m,\ell}^2\biggr)^2 K\biggl[7 + 4\biggl(\sum_{u=1}^m \rho_u^{1/2}\biggr) + \biggl(\sum_{u=1}^m \rho_u^{1/2}\biggr)^2\biggr] \\
&\quad\le Cm^{-1}(\log m)^{-(1+\delta)}.
\end{aligned}$$

Therefore,

$$\sum_{m=1}^{\infty} P\biggl(\frac{\sum_{j=1}^m a_{m,j}e_j}{m^{1/4}(\log m)^{(1+\delta)/4}} > \varepsilon\biggr) < +\infty.$$

The result follows by the Borel–Cantelli lemma.

APPENDIX C

Sufficient conditions for ($\mathbf{L}_w$), ($\mathbf{L}_w^*$) and (CC). Denote

$$\gamma_{nm} = \max_{i,j}\{\mathbf{x}_{ij}^T \mathbf{F}_{nm}^{-1} \mathbf{x}_{ij}\} \quad \text{and} \quad \gamma_{nm}^* = \tau_{nm} \max_{i,j}\{\mathbf{x}_{ij}^T \mathbf{H}_{nm}^{-1} \mathbf{x}_{ij}\};$$

$$w_{nm} = \frac{\lambda_{\max}(\mathbf{H}_{nm}^{-1}\mathbf{M}_{nm})}{\lambda_{\min}(\mathbf{H}_{nm}^{-1}\mathbf{M}_{nm})} \quad \text{and} \quad w'_{nm} = \frac{\max_{1\le i\le n} \lambda_{\max}(\mathbf{R}_i^{-1}(\alpha)\bar{\mathbf{R}}_i)}{\lambda_{\min}(\mathbf{H}_{nm}^{-1}\mathbf{M}_{nm})}.$$

It is clear that $w_{nm} \le w'_{nm} \le \frac{\max_{1\le i\le n}\{\lambda_{\max}(\mathbf{R}_i^{-1}(\alpha)\bar{\mathbf{R}}_i)\}}{\min_{1\le i\le n}\{\lambda_{\min}(\mathbf{R}_i^{-1}(\alpha)\bar{\mathbf{R}}_i)\}}$.

First, we present three lemmas. Lemma A.1 provides conditions under which $\mathbf{H}_{nm}(\beta)$ is close to $\mathbf{H}_{nm}$. Lemma A.2 provides conditions under which $\mathbf{B}_{nm}(\beta)$ is dominated by $\mathbf{H}_{nm}$. Lemma A.3 provides conditions that assure the uniform convergence of $\mathcal{E}_{nm}(\beta)$.

LEMMA A.1. *Suppose assumption* (AH) *holds.*

(i) *If* $\pi_{nm}\kappa_{nm}\gamma_{nm} \to 0$, *then, for any* $p \times 1$ *vector* λ, $\|\lambda\| = 1$, *we have* $\sup_{\beta\in B_{nm}(r)} |\lambda^T \mathbf{H}_{nm}^{-1/2}\mathbf{H}_{nm}(\beta)\mathbf{H}_{nm}^{-1/2}\lambda - 1| = o(1)$.

(ii) *If* $\pi_{nm}\kappa_{nm}\gamma_{nm}^* \to 0$, *then, for any* $p \times 1$ *vector* λ, $\|\lambda\| = 1$, *we have* $\sup_{\beta\in B_{nm}^*(r)} |\lambda^T \mathbf{H}_{nm}^{-1/2}\mathbf{H}_{nm}(\beta)\mathbf{H}_{nm}^{-1/2}\lambda - 1| = o(1)$.

LEMMA A.2. *Suppose assumption* (AH) *holds.*

(i) *If* $w_{nm}(\pi_{nm})^2\kappa_{nm}\gamma_{nm} \to 0$, *then, for any* $p \times 1$ *vector* λ, $\|\lambda\| = 1$, $\sup_{\beta\in B_{nm(r)}}\{\lambda^T \mathbf{M}_{nm}^{-1/2}\mathbf{B}_{nm}(\beta)\mathbf{H}_{nm}^{-1}\mathbf{M}_{nm}^{1/2}\lambda\} = o(1)$.

(ii) *If* $(\pi_{nm})^2\kappa_{nm}\gamma_{nm}^* \to 0$, *then, for any* $p \times 1$ *vector* λ, $\|\lambda\| = 1$, $\sup_{\beta\in B_{nm}^*(r)}\{\lambda^T \mathbf{H}_{nm}^{-1/2}\mathbf{B}_{nm}(\beta)\mathbf{H}_{nm}^{-1/2}\lambda\} = o(1)$.

LEMMA A.3. *Suppose assumption* (AH) *holds. Denote* $b_{nm} = \min_{i,j}\sigma_{ij}^2$, $\nu_{nm} = \min\{(nm/w'_{nm})^{1/2}, (m\pi_{nm})/b_{nm}\}$ *and* $\nu_{nm}^* = \min\{(nm)^{1/2}, (m\pi_{nm})b_{nm}^{-1}\}$.

(i) *If* $\nu_{nm}w'_{nm}\pi_{nm}\kappa_{nm}\gamma_{nm} \to 0$, *then, for any* $p \times 1$ *vector* λ, $\|\lambda\| = 1$, $\sup_{\beta\in B_{nm}^*(r)}\{\lambda^T \mathbf{M}_{nm}^{-1/2}\mathcal{E}_{nm}(\beta)\mathbf{H}_{nm}^{-1}(\beta)\mathbf{M}_{nm}^{1/2}\lambda\} = o_p(1)$.

(ii) *If* $\nu_{nm}^*\pi_{nm}\kappa_{nm}\gamma_{nm}^* \to 0$, *then, for any* $p \times 1$ *vector* λ, $\|\lambda\| = 1$, $\sup_{\beta\in B_{n,m}^*(r)}\{\lambda^T \mathbf{H}_{nm}^{-1/2}\mathcal{E}_{nm}(\beta)\mathbf{H}_{nm}^{-1/2}\lambda\} = o_p(1)$.

REMARK A.1. There is an extra price one needs to pay for uniform convergence in Lemma A.3, compared to the assumptions in Lemmas A.1 and A.2. In Lemmas A.1 and A.2, we essentially require γ_{nm} or $\gamma_{nm}^* \to 0$; But in Lemma A.3, we require γ_{nm} or $\gamma_{nm}^* \to 0$ faster than either $(nm)^{-1/2}$ or m^{-1}.

We have the following two theorems. Theorem A.1 provides sufficient conditions for (L_w) and (L_w^*), and Theorem A.2 provides sufficient conditions for (CC).

THEOREM A.1. *Suppose assumption* (AH) *holds*.

(i) [Sufficient conditions for (L_w)]. *If the conditions of Lemmas* A.2(i) *and* A.3(i) *hold*, *then*

$$P\{\mathcal{D}_{nm}^T(\beta)\mathbf{M}_{nm}^{-1}\mathcal{D}_{nm}(\beta) \geq c_0\mathbf{F}_{nm},\ \textit{for all } \beta \in B_{nm}(r)\} \to 1.$$

(ii) [Sufficient conditions for (L_w^*)]. *If the conditions of Lemmas* A.2(ii) *and* A.3(ii) *hold*, *then*

$$P\{\mathcal{D}_{nm}(\beta) \geq c_0\mathbf{H}_{nm}, \textit{for all } \beta \in B_{nm}^*(r)\} \to 1.$$

THEOREM A.2 [Sufficient conditions for (CC)*]. *If the conditions of Theorem* A.1(ii) *hold*, *then condition* (CC) *holds*.

Proofs of Lemmas A.1–A.3 and Theorems A.1–A.2. In the sequel, we use the notation $\eta_{ij}^{(0)}$ and $\theta_{ij}^{(0)}$ to denote the values of η_{ij} and θ_{ij} evaluated at β_0, respectively. First, we state two additional lemmas, which are used in the proofs of Lemmas A.1–A.3.

LEMMA B.1. *Suppose assumption* (AH) *holds*.

(i) *When* $\kappa_{nm}\gamma_{nm} \to 0$, *we have*

$$\delta_{a,nm} = \sup_{\beta \in B_{nm}(r)} \left\{ \max_{1\leq i\leq n} \max_{1\leq j\leq m_i} \left| \frac{a''(\theta_{ij})}{a''(\theta_{ij}^{(0)})} - 1 \right| \right\} = O\big((\kappa_{nm}\gamma_{nm})^{1/2}\big)$$

and

$$\delta_{h,nm} = \sup_{\beta \in B_{nm}(r)} \left\{ \max_{1\leq i\leq n} \max_{1\leq j\leq m_i} \left| \frac{h'(\eta_{ij})}{h'(\eta_{ij}^{(0)})} - 1 \right| \right\} = O\big((\kappa_{nm}\gamma_{nm})^{1/2}\big).$$

(ii) *When* $\kappa_{nm}\gamma_{nm}^* \to 0$, *we have*

$$\delta_{a,nm}^* = \sup_{\beta \in B_{nm}*(r)} \left\{ \max_{1\leq i\leq n} \max_{1\leq j\leq m_i} \left| \frac{a''(\theta_{ij})}{a''(\theta_{ij}^{(0)})} - 1 \right| \right\} = O\big((\kappa_{nm}\gamma_{nm}^*)^{1/2}\big)$$

and

$$\delta_{h,nm}^* = \sup_{\beta \in B_{nm}*(r)} \left\{ \max_{1\leq i\leq n} \max_{1\leq j\leq m_i} \left| \frac{h'(\eta_{ij})}{h'(\eta_{ij}^{(0)})} - 1 \right| \right\} = O\big((\kappa_{nm}\gamma_{nm}^*)^{1/2}\big).$$

LEMMA B.2. *Suppose assumption* (AH) *holds*.

(i) *When* $\kappa_{nm}\gamma_{nm} \to 0$,

$$\sup_{\beta \in B_{nm}(r)} \left[\left| \frac{a^{(3)}(\theta_{ij})}{[a''(\theta_{ij})]^{5/2}} \{h'(\eta_{ij})\}^2 - \frac{a^{(3)}(\theta_{ij}^{(0)})}{[a''(\theta_{ij}^{(0)})]^{5/2}} \{h'(\eta_{ij}^{(0)})\}^2 \right| \left| \frac{\{a''(\theta_{ij}^{(0)})\}^{3/2}}{\{h'(\eta_{ij}^{(0)})\}^2} \right| \right]$$
$$= O\big((\kappa_{nm}\gamma_{nm})^{1/2}\big)$$

and

$$\sup_{\beta \in B_{nm}(r)} \left[\left| \frac{h''(\eta_{ij})}{[a''(\theta_{ij})]^{1/2}} - \frac{h''(\eta_{ij}^{(0)})}{[a''(\theta_{ij}^{(0)})]^{1/2}} \right| \left| \frac{\{a''(\theta_{ij}^{(0)})\}^{3/2}}{\{h'(\eta_{ij}^{(0)})\}^2} \right| \right] = O\big((\kappa_{nm}\gamma_{nm})^{1/2}\big).$$

(ii) *When* $\kappa_{nm}\gamma_{nm}^* \to 0$,

$$\sup_{\beta \in B_{nm}(r)} \left[\left| \frac{a^{(3)}(\theta_{ij})}{[a''(\theta_{ij})]^{5/2}} \{h'(\eta_{ij})\}^2 - \frac{a^{(3)}(\theta_{ij}^{(0)})}{[a''(\theta_{ij}^{(0)})]^{5/2}} \{h'(\eta_{ij}^{(0)})\}^2 \right| \left| \frac{\{a''(\theta_{ij}^{(0)})\}^{3/2}}{\{h'(\eta_{ij}^{(0)})\}^2} \right| \right]$$
$$= O\big((\kappa_{nm}\gamma_{nm}^*)^{1/2}\big)$$

and

$$\sup_{\beta \in B_{nm}(r)} \left[\left| \frac{h''(\eta_{ij})}{[a''(\theta_{ij})]^{1/2}} - \frac{h''(\eta_{ij}^{(0)})}{[a''(\theta_{ij}^{(0)})]^{1/2}} \right| \left| \frac{\{a''(\theta_{ij}^{(0)})\}^{3/2}}{\{h'(\eta_{ij}^{(0)})\}^2} \right| \right] = O\big((\kappa_{nm}\gamma_{nm}^*)^{1/2}\big).$$

PROOFS OF LEMMAS B.1 AND B.2. One can directly prove the results by Taylor's expansion around the true parameter β_0 and assumption (AH). Details are omitted.

PROOF OF LEMMA A.1. Write $s_\lambda(\beta) = \lambda^T \mathbf{H}_{nm}^{-1/2} \mathbf{H}_{nm}(\beta) \mathbf{H}_{nm}^{-1/2} \lambda - 1$. By the Cauchy–Schwarz inequality and direct computation, we have

$$\left| \sum_{i=1}^{n} \boldsymbol{\lambda}^T \mathbf{H}_{nm}^{-1/2} \mathbf{X}_i^T \big(\Delta_i(\beta)\mathbf{A}_i^{1/2}(\beta) - \Delta_i \mathbf{A}_i^{1/2}\big) \mathbf{R}_i^{-1}(\alpha) \mathbf{A}_i^{1/2}(\beta) \Delta_i(\beta) \mathbf{X}_i \mathbf{H}_{nm}^{-1/2} \lambda \right|$$
$$\le \left(\sum_{i=1}^{n} \lambda^T \mathbf{H}_{nm}^{-1/2} \mathbf{X}_i^T \Delta_i \mathbf{A}_i^{1/2} \mathbf{R}_i^{-1}(\alpha) \mathbf{A}_i^{1/2} \Delta_i \mathbf{X}_i \mathbf{H}_{nm}^{-1/2} \lambda \right)^{1/2}$$
$$\times \left[\sum_{i=1}^{n} \lambda^T \mathbf{H}_{nm}^{-1/2} \mathbf{X}_i \Delta_i(\beta) \mathbf{A}_i^{1/2}(\beta) \mathbf{R}_i^{-1}(\alpha) \right.$$

$$
\begin{aligned}
&\qquad \times \big(\Delta_i^{-1}\Delta_i(\beta)\mathbf{A}_i^{-1/2}\mathbf{A}_i^{1/2}(\beta) - I\big)\mathbf{R}_i(\alpha) \\
&\qquad \times \big(\Delta_i^{-1}\Delta_i(\beta)\mathbf{A}_i^{-1/2}\mathbf{A}_i^{1/2}(\beta) - I\big) \\
&\qquad\qquad \times \mathbf{R}_i^{-1}(\alpha)\Delta_i(\beta)\mathbf{A}_i^{1/2}(\beta)\mathbf{X}_i\mathbf{H}_{nm}^{-1/2}\lambda\Bigg]^{1/2} \\
&\le (\pi_{nm})^{1/2}\max_{1\le i\le n}\big[\lambda_{\max}\{|\Delta_i^{-1}\Delta_i(\beta)\mathbf{A}_i^{-1/2}\mathbf{A}_i^{1/2}(\beta) - I|\}\big] \\
&\quad \times \big[\lambda^T\mathbf{H}_{nm}^{-1/2}\mathbf{H}_{nm}(\beta)\mathbf{H}_{nm}^{-1/2}\lambda\big]^{1/2} \\
&\le (\pi_{nm})^{1/2}\max_{1\le i\le n}\max_{1\le j\le m_i}\left[\left|\frac{h'(\eta_{ij})}{h'(\eta_{ij}^{(0)})}\left\{\frac{a''(\theta_{ij}^{(0)})}{a''(\theta_{ij})}\right\}^{1/2} - 1\right|\right]\big(1+|s_\lambda(\beta)|\big)^{1/2}.
\end{aligned}
$$

Similarly, we have

$$
\begin{aligned}
&\left|\sum_{i=1}^n \lambda^T\mathbf{H}_{nm}^{-1/2}\mathbf{X}_i^T\Delta_i\mathbf{A}_i^{1/2}\mathbf{R}_i^{-1}(\alpha)(\mathbf{A}_i^{1/2}(\beta)\Delta_i(\beta) - \mathbf{A}_i^{1/2}\Delta_i)\mathbf{X}_i\mathbf{H}_{nm}^{-1/2}\lambda\right| \\
&\qquad \le \pi_{nm}^{1/2}\max_{1\le i\le n}\max_{1\le j\le m_i}\left|\frac{h'(\eta_{ij})}{h'(\eta_{ij}^{(0)})}\left\{\frac{a''(\theta_{ij}^{(0)})}{a''(\theta_{ij})}\right\}^{1/2} - 1\right|.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\sup_{\beta\in B_{nm}(r)}|s_\lambda(\beta)| \\
&\quad \le \left\{\left(1+\sup_{\beta\in B_{nm}(r)}|s_\lambda(\beta)|\right)^{1/2} + 1\right\} \\
&\qquad \times(\pi_{nm})^{1/2}\sup_{\beta\in B_{nm}(r)}\left\{\max_{1\le i\le n}\max_{1\le j\le m_i}\left[\left|\frac{h'(\eta_{ij})}{h'(\eta_{ij}^{(0)})}\left\{\frac{a''(\theta_{ij}^{(0)})}{a''(\theta_{ij})}\right\}^{1/2} - 1\right|\right]\right\}.
\end{aligned}
$$

From Lemma B.1(i), we have that

$$
(\pi_{nm})^{1/2}\sup_{\beta\in B_{nm}(r)}\left\{\max_{1\le i\le n}\max_{1\le j\le m_i}\left[\left|\frac{h'(\eta_{ij})}{h'(\eta_{ij}^{(0)})}\left\{\frac{a''(\theta_{ij}^{(0)})}{a''(\theta_{ij})}\right\}^{1/2} - 1\right|\right]\right\} = o(1).
$$

So, $\sup_{\beta\in B_{nm}(r)}|s_\lambda(\beta)| = o(1)$. This proves part (i). The proof of part (ii) is similar. □

PROOF OF LEMMA A.2. By Taylor's expansion, $(\mu_i(\beta) - \mu_i) = \mathbf{A}_i(\bar\beta) \times \Delta_i(\bar\beta)\mathbf{X}_i(\beta - \beta_0)$, where $\bar\beta$ is between β and β_0. So, by Lemma B.1 and

Lemma A.1, there exists a constant C_1 such that

$$\sum_{i=1}^{n}\big(\mu_i-\mu_i(\beta)\big)^T\mathbf{A}_i^{-1/2}(\beta)\mathbf{R}_i^{-1}\mathbf{A}_i^{-1/2}(\beta)\big(\mu_i-\mu_i(\beta)\big)$$

$$\le(\beta-\beta_0)^T\mathbf{H}_{nm}(\bar\beta)(\beta-\beta_0)\max_{1\le i\le n}\left\{\frac{\lambda_{\max}(\mathbf{R}_i(\alpha))}{\lambda_{\min}(\mathbf{R}_i(\alpha))}\lambda_{\max}\{\mathbf{A}_i(\bar\beta)\mathbf{A}_i^{-1}(\beta)\}\right\}$$

$$\le C_1\pi_{nm}(\beta-\beta_0)^T\mathbf{H}_{nm}(\bar\beta)(\beta-\beta_0)\le C_1r^2\pi_{nm}\lambda_{\max}(\mathbf{M}_{nm}\mathbf{H}_{nm}^{-1}).$$

By the Cauchy–Schwarz inequality, assumption (AH) and Lemma B.1, there exists a constant C_2 such that, for any $\beta\in B_{nm}(r)$,

$$\boldsymbol{\lambda}^T\mathbf{M}_{nm}^{-1/2}\mathbf{B}_{nm}^{[1]}(\beta)\mathbf{H}_{nm}^{-1}\mathbf{M}_{nm}^{1/2}\lambda$$

$$=\sum_{i=1}^{n}\lambda^T\mathbf{M}_{nm}^{-1/2}\mathbf{X}_i^T\mathbf{G}_i^{[1]}(\beta)\Delta_i^{-1}\operatorname{diag}\{\Delta_i\mathbf{X}_i\mathbf{H}_{nm}^{-1}\mathbf{M}_n^{1/2}\lambda\}$$

$$\times\mathbf{R}_i^{-1}(\alpha)\mathbf{A}_i^{-1/2}(\beta)\big(\mu_i-\mu_i(\beta)\big)$$

$$\le\pi_{nm}^{1/2}\big[\lambda^T\mathbf{M}_{nm}^{-1/2}\mathbf{H}_{nm}\mathbf{M}_{nm}^{-1/2}\lambda\big]^{1/2}$$

$$\times\left[\sum_{i=1}^{n}\big(\mu_i-\mu_i(\beta)\big)^T\mathbf{A}_i^{-1/2}(\beta)\mathbf{R}_i^{-1}(\alpha)\mathbf{A}_i^{-1/2}(\beta)\big(\mu_i-\mu_i(\beta)\big)\right]^{1/2}$$

$$\times\max_{1\le i\le n}\left\{\lambda_{\max}\{|\Delta_i^{-2}\mathbf{A}_i^{-1/2}\mathbf{G}_i^{[1]}(\beta)|\}\max_{1\le j\le m_i}\big[\{u'(\eta_{ij}^{(0)})\}^2\mathbf{x}_{ij}^T\mathbf{F}_{nm}^{-1}\mathbf{x}_{ij}\big]^{1/2}\right\}$$

$$\le C_2\pi_{nm}w_{nm}^{1/2}(\kappa_{nm}\gamma_{nm})^{1/2}.$$

In the above inequalities, we used the facts that $\sup_{\beta\in B_{nm}(r)}\lambda_{\max}\{|\Delta_i^{-2}(\beta)\times\{\mathbf{A}_i(\beta)\}^{-1/2}\mathbf{G}_i^{[1]}(\beta)|\}\le k_{nm}^{[1]}/2+k_{nm}^{[2]}$ and the absolute value of the jth diagonal element of $\operatorname{diag}\{\Delta_i\mathbf{X}_i\mathbf{H}_{nm}^{-1}\mathbf{M}_{nm}^{1/2}\lambda\}$ is less than $u'(\eta_{ij}^{(0)})(\mathbf{x}_{ij}^T\mathbf{F}_{nm}^{-1}\mathbf{x}_{ij})^{1/2}$.

Similarly, there exists a constant C_3 such that, for any $\beta\in B_{nm}(r)$,

$$\boldsymbol{\lambda}^T\mathbf{M}_{nm}^{-1/2}\mathbf{B}_{nm}^{[2]}(\beta)\mathbf{H}_{nm}^{-1}\mathbf{M}_{nm}^{1/2}\boldsymbol{\lambda}$$

$$=\sum_{i=1}^{n}\boldsymbol{\lambda}^T\mathbf{M}_{nm}^{-1/2}\mathbf{X}_i^T\Delta_i(\beta)\mathbf{A}_i^{1/2}(\beta)\mathbf{R}_i^{-1}(\alpha)\operatorname{diag}\{\Delta_i\mathbf{X}_i\mathbf{H}_{nm}^{-1}\mathbf{M}_{nm}^{1/2}\lambda\}$$

$$\times\Delta_i^{-1}\mathbf{G}_i^{[2]}(\beta)\big(\mu_i-\mu_i(\beta)\big)$$

$$\le C_3\pi_{nm}w_{nm}^{1/2}(\kappa_{nm}\gamma_{nm})^{1/2},$$

where we use the fact that $\sup_{\beta\in B_{nm}(r)}\lambda_{\max}\{|\{\Delta_i(\beta)\}^{-1}\{A_i(\beta)\}^{1/2}\mathbf{G}_i^{[2]}(\beta)|\}\le k_{nm}^{[1]}/2$. This proves part (i). Part (ii) can be proved in the same fashion. $\square$

PROOF OF LEMMA A.3.

$$
\begin{aligned}
&\boldsymbol{\lambda}^T \mathbf{M}_{nm}^{-1/2} \mathcal{E}_{nm}(\beta) \mathbf{H}_{nm}^{-1} \mathbf{M}_{nm}^{1/2} \lambda \\
&\quad = \sum_{i=1}^{n} \boldsymbol{\lambda}^T \mathbf{M}_{nm}^{-1/2} \mathbf{X}_i^T \mathbf{G}_i^{[1]} \Delta_i^{-1} \operatorname{diag}\{\Delta_i \mathbf{X}_i \mathbf{H}_{nm}^{-1} \mathbf{M}_{nm}^{1/2} \lambda\} \mathbf{R}_i^{-1}(\alpha) \mathbf{A}_i^{-1/2} (\mathbf{y}_i - \mu_i) \\
&\qquad + \sum_{i=1}^{n} \boldsymbol{\lambda}^T \mathbf{M}_{nm}^{-1/2} \mathbf{X}_i^T \Delta_i \mathbf{A}_i^{1/2} \mathbf{R}_i^{-1}(\alpha) \\
&\qquad\qquad \times \operatorname{diag}\{\Delta_i \mathbf{X}_i \mathbf{H}_{nm}^{-1} \mathbf{M}_{nm}^{1/2} \lambda\} \Delta_i^{-1} \mathbf{G}_i^{[2]} (\mathbf{y}_i - \mu_i) \\
&\qquad + \sum_{i=1}^{n} \boldsymbol{\lambda}^T \mathbf{M}_{nm}^{-1/2} \mathbf{X}_i^T \Delta_i^{-1} \mathbf{G}_i^{[1]}(\beta) \\
&\qquad\qquad \times \operatorname{diag}\{\Delta_i \mathbf{X}_i \mathbf{H}_{nm}^{-1} \mathbf{M}_{nm}^{1/2} \lambda\} \mathbf{R}_i^{-1}(\alpha) \big(\mathbf{A}_i^{-1/2}(\beta) - \mathbf{A}_i^{-1/2}\big) (\mathbf{y}_i - \mu_i) \\
&\qquad + \sum_{i=1}^{n} \boldsymbol{\lambda}^T \mathbf{M}_{nm}^{-1/2} \mathbf{X}_i^T \big(\Delta_i(\beta) \mathbf{A}_i^{1/2}(\beta) - \Delta_i \mathbf{A}_i^{1/2}\big) \mathbf{R}_i^{-1}(\alpha) \\
&\qquad\qquad \times \operatorname{diag}\{\Delta_i \mathbf{X}_i \mathbf{H}_{nm}^{-1} \mathbf{M}_{nm}^{1/2} \lambda\} \Delta_i^{-1} \mathbf{G}_i^{[2]}(\beta) (\mathbf{y}_i - \mu_i) \\
&\qquad + \sum_{i=1}^{n} \boldsymbol{\lambda}^T \mathbf{M}_{nm}^{-1/2} \mathbf{X}_i^T \Delta_i^{-1} \big(\mathbf{G}_i^{[1]}(\beta) - \mathbf{G}_i^{[1]}\big) \\
&\qquad\qquad \times \operatorname{diag}\{\Delta_i \mathbf{X}_i \mathbf{H}_{nm}^{-1} \mathbf{M}_{nm}^{1/2} \lambda\} \mathbf{R}_i^{-1}(\alpha) \mathbf{A}_i^{-1/2} (\mathbf{y}_i - \mu_i) \\
&\qquad + \sum_{i=1}^{n} \boldsymbol{\lambda}^T \mathbf{M}_{nm}^{-1/2} \mathbf{X}_i^T \Delta_i \mathbf{A}_i^{1/2} \mathbf{R}_i^{-1}(\alpha) \\
&\qquad\qquad \times \operatorname{diag}\{\Delta_i \mathbf{X}_i \mathbf{H}_{nm}^{-1} \mathbf{M}_{nm}^{1/2} \lambda\} \Delta_i^{-1} \big(\mathbf{G}_i^{[2]}(\beta) - \mathbf{G}_i^{[2]}\big) (\mathbf{y}_i - \mu_i) \\
&\quad = I + II + III + IV + V + VI.
\end{aligned}
$$

Note that I and II do not depend on β. By direct computation, we have

$$
\begin{aligned}
\mathrm{E} I^2 &= \sum_{i=1}^{n} E\{\lambda^T \mathbf{M}_{nm}^{-1/2} \mathbf{X}_i^T \mathbf{G}_i^{[1]} \Delta_i^{-1} \\
&\qquad \times \operatorname{diag}\{\Delta_i \mathbf{X}_i \mathbf{H}_{nm}^{-1} \mathbf{M}_{nm}^{1/2} \lambda\} \mathbf{R}_i^{-1}(\alpha) \mathbf{A}_i^{-1/2} (y_i - \mu_i)\}^2 \\
&= \sum_{i=1}^{n} \lambda^T \mathbf{M}_{nm}^{-1/2} \mathbf{X}_i^T \mathbf{G}_i^{[1]} \Delta_i^{-1} \operatorname{diag}\{\Delta_i \mathbf{X}_i \mathbf{H}_{nm}^{-1} \mathbf{M}_{nm}^{1/2} \lambda\} \mathbf{R}_i^{-1}(\alpha) \bar{\mathbf{R}}_i \mathbf{R}_i^{-1}(\alpha) \\
&\qquad \times \operatorname{diag}\{\Delta_i \mathbf{X}_i \mathbf{H}_{nm}^{-1}(\beta) \mathbf{M}_{nm}^{1/2} \lambda\} \Delta_i^{-1} \mathbf{G}_i^{[1]} \mathbf{X}_i \mathbf{M}_{nm}^{-1/2} \boldsymbol{\lambda} \\
&\le (k_{nm}^{[1]}/2 + k_{nm}^{[2]})^2 \lambda_{\max}(\mathbf{M}_{nm}^{-1} \mathbf{H}_{nm}) \max_{1 \le i \le n} \lambda_{\max}\big(\mathbf{R}_i^{-1}(\alpha) \bar{\mathbf{R}}_i\big) \pi_{nm} \kappa_{nm} \gamma_{nm} \\
&\le C_4 w'_{nm} \pi_{nm} \kappa_{nm} \gamma_{nm},
\end{aligned}
$$

where C_4 is a constant given by (AH). By the Chebyshev inequality, $I = o_p(1)$. Similarly, we can prove $II = o_p(1)$.

In *III*, *IV*, *V* and *VI*, the terms involve the parameter β. To prove uniform convergence, we need to separate the terms with $\mathbf{y}_i - \mu_i$ from the terms that involve β, so that expectations can be applied directly to the terms with $\mathbf{y}_i - \mu_i$. We consider here two different ways to separate the terms with $\mathbf{y}_i - \mu_i$ and the terms with β, which correspond to the two terms in the definition of ν_{nm}. The ν_{nm} has the same rate as the first term when $m \to \infty$ very fast (compared to n), and it has the same rate as the second term when m is bounded or $m \to \infty$ at a slow rate (compared to n). Next, we prove the uniform convergence for *III*. The proofs for *IV*, *V* and *VI* are similar.

By the Cauchy–Schwarz inequality, assumption (AH) and Lemma B.1, there exists a constant C_5 such that

$$\begin{aligned}
&\mathrm{E}\Big\{\sup_{\beta\in B_{nm}(r)} |III|^2\Big\}\\
&\quad = \mathrm{E}\Bigg\{\sup_{\beta\in B_{nm}(r)}\Bigg[\sum_{i=1}^{n} \boldsymbol{\lambda}^T \mathbf{M}_{nm}^{-1/2}\mathbf{X}_i^T\Delta_i^{-1}\mathbf{G}_i^{[1]}(\beta)\operatorname{diag}\{\Delta_i\mathbf{X}_i\mathbf{H}_{nm}^{-1}\mathbf{M}_{nm}^{1/2}\boldsymbol{\lambda}\}\\
&\qquad\qquad\qquad \times \mathbf{R}_i^{-1}(\alpha)\big(\mathbf{A}_i^{-1/2}(\beta) - \mathbf{A}_i^{-1/2}\big)(\mathbf{y}_i - \mu_i)\Bigg]^2\Bigg\}\\
&\quad \le (\pi_{nm})^2\{\boldsymbol{\lambda}^T\mathbf{M}_{nm}^{-1/2}\mathbf{H}_{nm}\mathbf{M}_{nm}^{-1/2}\boldsymbol{\lambda}\}\\
&\qquad \times \mathrm{E}\Bigg\{\sum_{i=1}^{n}(\mathbf{y}_i - \mu_i)^T\mathbf{A}_i^{-1/2}\mathbf{R}_i^{-1}(\alpha)\mathbf{A}_i^{-1/2}(\mathbf{y}_i - \mu_i)\Bigg\}\\
&\qquad \times \sup_{\beta\in B_{nm}(r)}\max_{1\le i\le n}\Big[\lambda_{\max}\big\{\big(\mathbf{A}_i^{-1/2}(\beta)\mathbf{A}_i^{1/2} - I\big)^2\big\}\\
&\qquad\qquad\qquad \times \lambda_{\max}\big\{\big(\mathbf{G}_i^{[1]}(\beta)\Delta_i^{-2}\mathbf{A}_i^{-1/2}\big)^2\big\}\\
&\qquad\qquad\qquad \times \lambda_{\max}\big\{\operatorname{diag}[\Delta_i\mathbf{X}_i\mathbf{H}_{nm}^{-1}\mathbf{M}_{nm}^{1/2}\lambda]^2\big\}\Big]\\
&\quad \le C_5(\pi_{nm})^2\lambda_{\max}(\mathbf{M}_{nm}^{-1}\mathbf{H}_{nm})\sum_{i=1}^{n}\operatorname{tr}\{\mathbf{R}_i^{-1}(\alpha)\bar{\mathbf{R}}_i\}\kappa_{nm}\gamma_{nm}^2\\
&\quad \le C_5\ (nm)w'_{nm}(\pi_{nm}\kappa_{nm}\gamma_{nm})^2.
\end{aligned}$$

Alternatively, we set $s_{ij} = \mathbf{e}_{ij}^T\mathbf{R}_i^{-1}(\alpha)(\mathbf{A}_i^{-1/2}(\beta)\mathbf{A}_i^{1/2} - I)^2\mathbf{R}_i^{-1}(\alpha)\mathbf{e}_{ij}$, where $\mathbf{e}_{ij}$ is the $m_i \times 1$ indicator vector with the jth element equal to 1. By the Cauchy–Schwarz inequality, assumption (AH) and Lemma B.1, there exists a constant C_6

such that

$$
\begin{aligned}
&\mathrm{E}\Big\{\sup_{\beta\in B_{nm}(r)}|III|^2\Big\}\\
&\quad=\mathrm{E}\Bigg\{\sup_{\beta\in B_{nm}(r)}\Bigg[\sum_{i=1}^{n}\boldsymbol{\lambda}^T\mathbf{M}_{nm}^{-1/2}\mathbf{X}_i^T\mathbf{G}_i^{[1]}(\beta)\\
&\qquad\qquad\times\operatorname{diag}\{\mathbf{R}_i^{-1}(\alpha)\big(\mathbf{A}_i^{-1/2}(\beta)-\mathbf{A}_i^{-1/2}\big)(\mathbf{y}_i-\mu_i)\}\\
&\qquad\qquad\times\mathbf{X}_i\mathbf{H}_{nm}^{-1}\mathbf{M}_{nm}^{1/2}\boldsymbol{\lambda}\Bigg]^2\Bigg\}\\
&\quad\le\max_{1\le i\le n}\big[\boldsymbol{\lambda}_{\max}\{\mathbf{R}_i(\alpha)\}\big]\{\boldsymbol{\lambda}^T\mathbf{M}_{nm}^{-1/2}\mathbf{H}_{nm}\mathbf{M}_{nm}^{-1/2}\boldsymbol{\lambda}\}\\
&\qquad\times\sup_{\beta\in B_{nm}(r)}\max_i\lambda_{\max}\{\mathbf{G}_i^{[1]}(\beta)\Delta_i^{-2}\mathbf{A}_i^{-1/2}\}^2\\
&\qquad\times\mathrm{E}\Bigg\{\sup_{\beta\in B_{nm}(r)}\sum_{i=1}^{n}\boldsymbol{\lambda}^T\mathbf{M}_{nm}^{1/2}\mathbf{H}_{nm}^{-1}\mathbf{X}_i^T\Delta_i\\
&\qquad\qquad\times\operatorname{diag}\{\mathbf{R}_i^{-1}(\alpha)\big(\mathbf{A}_i^{-1/2}(\beta)-\mathbf{A}_i^{-1/2}\big)(\mathbf{y}_i-\mu_i)\}^2\\
&\qquad\qquad\times\Delta_i\mathbf{X}_i\mathbf{H}_{nm}^{-1}\mathbf{M}_{nm}^{1/2}\boldsymbol{\lambda}\Bigg\}\\
&\quad\le C_6\max_{1\le i\le n}\big[\boldsymbol{\lambda}_{\max}\{\mathbf{R}_i(\alpha)\}\big]\boldsymbol{\lambda}_{\max}\{\mathbf{M}_{nm}^{-1}\mathbf{H}_{nm}\}\mathrm{E}\{(\mathbf{y}_i-\mu_i)^T\mathbf{A}_i^{-1}(\mathbf{y}_i-\mu_i)\}\\
&\qquad\times\sup_{\beta\in B_{nm}(r)}\Bigg\{\sum_{i=1}^{n}\boldsymbol{\lambda}^T\mathbf{M}_{nm}^{1/2}\mathbf{H}_{nm}^{-1}\mathbf{X}_i^T\Delta_i\operatorname{diag}\{s_{i1},\dots,s_{im_i}\}\Delta_i\mathbf{X}_i\mathbf{H}_{nm}^{-1}\mathbf{M}_{nm}^{1/2}\boldsymbol{\lambda}\Bigg\}\\
&\quad\le C_6\,m(\pi_{nm})^2w'_{nm}b_{nm}^{-1}\sup_{\beta\in B_{nm}(r)}\max_{1\le i\le n}\{\lambda_{\max}\{\mathbf{A}_i^{-1/2}(\beta)\mathbf{A}_i^{1/2}-I\}^2\}\\
&\quad\le C_6\,mb_{nm}^{-1}w'_{nm}(\pi_{nm})^2\kappa_{nm}\gamma_{nm}.
\end{aligned}
$$

Hence, $\mathrm{E}\{\sup_{\beta\in B_{nm}(r)}|III|^2\}\le\max\{C_5,C_6\}\nu_{nm}w'_{nm}\pi_{nm}\kappa_{nm}\gamma_{nm}$. By the Chebyshev inequality, $\sup_{\beta\in B_{nm}(r)}|III|=o_p(1)$. $\square$

PROOF OF THEOREM A.1. By (5), Lemma A.2(i) and Lemma A.3(i), we have, for any $p\times 1$ vector λ, $\|\lambda\|=1$,

$$\sup_{\beta\in B_{nm}(r)}\{|\lambda^T\mathbf{M}_{nm}^{-1/2}\mathcal{D}_{nm}(\beta)\mathbf{H}_{nm}^{-1}(\beta)\mathbf{M}_{nm}^{1/2}\lambda-1|\}=o_p(1).$$

By Lemma 1, it is easy to see that the result in part (i) is true. Part (ii) follows similarly, using the representation given in (5), Lemma A.2(ii) and Lemma A.3(ii). □

PROOF OF THEOREM A.2. Note p is fixed. We only need to prove that, for any k, l, $1 \le k, l \le p$,

$$\sup_{\beta \in B^*_{nm}(r)} \left| \mathbf{e}_k^T \{ \mathbf{H}_{nm}^{-1/2} \mathcal{D}_{nm}(\beta) \mathbf{H}_{nm}^{-1/2} - I \} \mathbf{e}_l \right| = o_p(1),$$

where $\mathbf{e}_k$ is a $p \times 1$ vector with the kth element equal to 1 and the other elements equal to 0. Using the representation given in (5), the proof can be broken into three pieces. The proofs of these three pieces follow as in the proofs of Lemma A1(ii), Lemma A2(ii) and Lemma A3(ii), respectively, except that λ must replaced by $\mathbf{e}_k$ or $\mathbf{e}_l$ in the appropriate places. Details are omitted. □

Acknowledgments. The authors thank Bing Li, Kesar Singh, Zhiliang Ying and Cun-Hui Zhang for their comments, suggestions, and references throughout the development of the paper. In particular, the authors thank Zhiliang Ying and John Marden, who kindly read the entire manuscript, and William Strawderman, who helped us to improve the presentation of the final manuscript; Their suggestions resulted in great improvements to the paper. The authors also thank the Editors, Associate Editors and the referees for their constructive suggestions and comments. Their suggestions have extended the scope of the paper and improved its structure and presentation.

REFERENCES

BILLINGSLEY, P. (1986). *Probability and Measure,* 2nd ed. Wiley, New York.

CHEN, K., HU, I. and YING, Z. (1999). Strong consistency of maximum quasi-likelihood estimators in generalized linear models with fixed and adaptive designs. *Ann. Statist.* **27** 1155–1163.

CHOW, Y. S. and TEICHER, H. (1988). *Probability Theory*: *Independence, Interchangeability, Martingales*, 2nd ed. Springer, New York.

CRAMÉR, H. (1946). *Mathematical Methods of Statistics*. Princeton Univ. Press.

CROWDER, M. (1986). On consistency and inconsistency of estimating equations. *Econometric Theory* **2** 305–330.

DIGGLE, P. J., LIANG, K.-Y. and ZEGER, S. L. (1996). *Analysis of Longitudinal Data*. Clarendon Press, Oxford.

EICKER, F. (1967). Limit theorems for regressions with unequal and dependent errors. *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* **1** 59–82. Univ. California Press, Berkeley.

FAHRMEIR, L. and KAUFMANN, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Ann. Statist.* **13** 342–368.

FAHRMEIR, L. and KAUFMANN, H. (1986). Asymptotic inference in discrete response models. *Statist. Hefte* **27** 179–205.

GOURIEROUX, C. and MONFORT, A. (1981). Asymptotic properties of the maximum likelihood estimator in dichotomous logit models. *J. Econometrics* **17** 83–97.

HABERMAN, S. J. (1977). Maximum likelihood estimates in exponential response models. *Ann. Statist.* **5** 815–841.

HUBER, P. J. (1981). *Robust Statistics*. Wiley, New York.

LI, B. (1996). A minimax approach to consistency and efficiency for estimating equations. *Ann. Statist.* **24** 1283–1297.

LIANG, K.-Y. and ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13–22.

NELDER, J. A. and WEDDERBURN, R. W. M. (1972). Generalized linear models. *J. Roy. Statist. Soc. Ser. A* **135** 370–384.

NI, G. X. (1984). *Basic Matrix Theories and Methods.* Shanghai Science and Technology Publisher. (In Chinese.)

ROMANO, J. P. and SIEGEL, A. F. (1986). *Counterexamples in Probability and Statistics*. Wadsworth, Monterey, CA.

SCHOTT, J. R. (1997). *Matrix Analysis for Statistics*. Wiley, New York.

SERFLING, R. J. (1970). Convergence properties of S_n under moment restrictions. *Ann. Math. Statist.* **41** 1235–1248.

WU, C. F. (1981). Asymptotic theory of nonlinear least squares estimation. *Ann. Statist.* **9** 501–513.

XIE, M. and SIMPSON, D. G. (1998). Categorical exposure-response regression analysis of toxicology experiments. *Case Studies in Environmental Statistics. Lecture Notes in Statist.* **132** 121–141. Springer, New York.

XIE, M., SIMPSON, D. G. and CARROLL, R. J. (2000). Random effects in interval-censored ordinal regression: Latent structure and Bayesian approach. *Biometrics* **56** 376–383.

YOHAI, V. J. and MARONNA, R. A. (1979). Asymptotic behavior of M-estimators for the linear model. *Ann. Statist.* **7** 258–268.

YUAN, K.-H. and JENNRICH, R. I. (1998). Asymptotics of estimating equations under natural conditions. *J. Multivariate Anal.* **65** 245–260.

ZEGER, S. L. and LIANG, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42** 121–130.

DEPARTMENT OF STATISTICS
RUTGERS UNIVERSITY
HILL CENTER, BUSCH CAMPUS
PISCATAWAY, NEW JERSEY 08854
E-MAIL: mxie@stat.rutgers.edu

LABORATORY OF STATISTICAL GENETICS
ROCKEFELLER UNIVERSITY
NEW YORK, NEW YORK 10021