The Annals of Statistics
2011, Vol. 39, No. 1, 389–417
DOI: 10.1214/10-AOS846

GEE ANALYSIS OF CLUSTERED BINARY DATA WITH DIVERGING NUMBER OF COVARIATES

BY LAN WANG[1]

University of Minnesota

Clustered binary data with a large number of covariates have become increasingly common in many scientific disciplines. This paper develops an asymptotic theory for generalized estimating equations (GEE) analysis of clustered binary data when the number of covariates grows to infinity with the number of clusters. In this "large n, diverging p" framework, we provide appropriate regularity conditions and establish the existence, consistency and asymptotic normality of the GEE estimator. Furthermore, we prove that the sandwich variance formula remains valid. Even when the working correlation matrix is misspecified, the use of the sandwich variance formula leads to an asymptotically valid confidence interval and Wald test for an estimable linear combination of the unknown parameters. The accuracy of the asymptotic approximation is examined via numerical simulations. We also discuss the "diverging p" asymptotic theory for general GEE. The results in this paper extend the recent elegant work of Xie and Yang [*Ann. Statist.* **31** (2003) 310–347] and Balan and Schiopu-Kratina [*Ann. Statist.* **32** (2005) 522–541] in the "fixed p" setting.

1. Introduction. A fundamental problem in statistical analysis is to characterize the effects of a set of covariates $X_1, \ldots, X_p$ on a response variable Y based on a sample of size n. Recently, there has been considerable interest in investigating this problem in the so-called "large n, diverging p" asymptotic framework, where the dimension of the covariates increases to infinity with the sample size. This setup allows statisticians to adopt a more complex statistical model as more abundant data become available, and thus to reduce the modeling bias.

The "large n, diverging p" framework can be traced back to the earlier pioneering work on M-estimators with a diverging number of parameter; see Huber (1973), Portnoy (1984, 1985, 1988), Mammen (1989), Welsh (1989), Bai and Wu (1994), He and Shao (2000) and the references therein. With the advent of high-dimensional data in many scientific areas, statistical theory developed in this new framework has become crucial for guiding practical data analysis with high-dimensional covariates, which relies heavily on asymptotic theory to justify its validity. By allowing the covariates' dimension to increase with the sample size, Fan

Received December 2009; revised July 2010.

[1]Supported by NSF Grant DMS-1007603.

AMS 2000 subject classifications. Primary 62F12; secondary 62J12.

Key words and phrases. Clustered binary data, generalized estimating equations (GEE), high-dimensional covariates, sandwich variance formula.

and Peng (2004) studied nonconcave penalized likelihood; Lam and Fan (2008) investigated profile-kernel likelihood inference with generalized varying coefficient partially linear models; Huang, Horowitz and Ma (2008) explored bridge estimators in linear regression; Hjort, McKeague and Van Keilegom (2009) and Chen, Peng and Qin (2009) studied the effects of data dimension on empirical likelihood; Zou and Zhang (2009) studied the adaptive elastic net, Zhu and Zhu (2009) investigated parameter estimation in a semiparametric regression model with highly correlated predictors. In the aforementioned literature, the number of covariates p grows to infinity at a polynomial rate $o(n^{\alpha})$ for some $0 < \alpha < 1$. In particular, most of these papers provide necessary conditions under which classical asymptotic theories remain valid for α in the range $[\frac{1}{5}, \frac{1}{2}]$.

A different line of research considers the case where p can be much larger than n and even grow at an exponential rate of n, in which case the sparsity assumption and other more stringent regularity conditions are generally required to investigate the large-sample properties. Furthermore, it is worth noting that much work has also been devoted to classification and multiple hypotheses testing problems with high-dimensional covariates, but these problems are different in nature from what is discussed in this paper. We refer to the review papers of Donoho (2000), Fan and Li (2006) and Fan and Lv (2010) for more comprehensive references on high-dimensional data analysis.

When the research focus is on modeling the relationship between Y and a high-dimensional vector of covariates, the existing literature in the "large n, diverging p" setting has been largely restricted to independent data. In many modern data sets, in addition to the large dimensionality of covariates, complexity also arises when the responses are correlated due to repeated measures or clustered design. One representative example is the Framingham Heart Study, where the researchers are interested in linking common risk factors to the occurrence of cardiovascular diseases. In this study, many variables, such as age, smoking status, cholesterol level and blood pressure, were recorded for the participants during their clinic visits over the years to describe their physical characteristics and lifestyles. Another example is the Chicago Longitudinal Study in social science, which investigated the educational and social development of about 1500 low income, minority youths in the Chicago area. The study collected a large amount of information on many variables that measure children's early antisocial behavior, individual-level attributes of the child, family attributes and social characteristics of both the child and the family, among others. In some other examples of clustered data, the number of variables measured for each individual or experimental unit may not be many, but when one considers various interaction effects, the actual number of predictors in the statistical model can still be large and better fits the "large p" setup.

The intrinsic complexity of clustered data raises challenging issues for statistical analysis, especially for correlated non-Gaussian data where it is difficult to specify the full likelihood. In this paper, we establish the asymptotic properties of

generalized estimating equations (GEE), a semiparametric procedure widely used in practice for clustered data analysis, while allowing the covariate dimension to grow to infinity with the sample size.

The GEE procedure was introduced in a seminal paper of Liang and Zeger (1986) as a useful extension of generalized linear models [McCullagh and Nelder (1989)] to correlated data. Instead of specifying the full likelihood, it only requires the knowledge of the first two marginal moments and a working correlation matrix. Thus, it is particularly effective for modeling clustered binary or count data. A key advantage of the GEE approach is that it yields a consistent estimator (in the classical "large n, fixed p" setup), even if the working correlation structure is misspecified. The GEE estimator is also asymptotically efficient if the correlation structure is indeed correctly specified. The original paper of Liang and Zeger focused mostly on the methodology development. Li (1997) adopted a minimax approach to study the consistency of GEE. A more complete and systematic large-sample theory for GEE, including consistency and asymptotic normality, was elegantly established by Xie and Yang (2003). Balan and Schiopu-Kratina (2005) also rigorously studied a closely related pseudo-likelihood framework for GEE. However, these papers all assume that p is fixed and that the number of clusters n goes to infinity. Xie and Yang (2003) also considered the case where the cluster size (number of observations within each cluster) is itself large, which corresponds to a large number of time points in the longitudinal setting.

This paper examines the effect of high-dimensional covariates on the GEE estimator in the "large n, diverging p" setup, where $p = p_n$ is a function of the sample size n. We focus on clustered binary data because binary response (e.g., disease status) is ubiquitous in many scientific applications and because of the relative transparency of technical derivation. We also discuss the related theory for general GEE in Section 5.1 The main technical challenges come from the high dimensionality of the covariates, the dependence among observations within each cluster and the nuisance parameters in the working correlation matrix. We provide a self-contained derivation and extend earlier theory in the literature on M-estimation with a large number of parameters, which is not tailored for clustered data and generally has not considered nuisance parameters.

We aim to answer the following essential questions. To what extent can the asymptotic results derived in the classical asymptotic framework for GEE still be deemed trustworthy when the number of covariates is large? How large can p_n be (relative to n)? The main findings in this paper reveal that under reasonable conditions, the GEE estimator $\widehat{\boldsymbol{\beta}}_n$ is $\sqrt{p_n/n}$-consistent when $p_n^2/n \to 0$ and that an arbitrary linear combination $\boldsymbol{\alpha}_n^T(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0})$ is asymptotically normal when $p_n^3/n \to 0$, where $\boldsymbol{\beta}_{n0}$ is the true parameter value. These findings resonate with those in the literature for independent data in the "large p" setting. Moreover, we also verify that the desirable robustness property against working correlation matrix misspecification still holds and that both the sandwich variance formula and the large-sample

Wald test still remain valid in this new context. Understanding these fundamental questions is essential to justifying asymptotic statistical inference based on GEE for analyzing real-world clustered data containing many covariates, such as the validity of the confidence intervals provided by the GEE package in R, SAS and other statistical software packages.

The rest of the paper is organized as follows. In Section 2, we provide a brief review of the GEE procedure for analyzing clustered binary data. Section 3 establishes the consistency and asymptotic normality of the GEE estimator, the consistency of the sandwich variance formula and the validity of the large-sample Wald test in the "large n, diverging p" framework. Section 4 examines the asymptotic results via numerical simulations. Section 5 discusses general GEE and related problems.

2. Generalized estimating equations. For the jth observation of the ith cluster, we observe a binary response variable Y_{ij} and a p_n-dimensional vector of covariates $\mathbf{X}_{ij}$, $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$. Observations from different clusters are independent, but those from the same clusters are correlated. Let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{im_i})^T$ denote the vector of responses for the ith cluster and let $\mathbf{X}_i = (\mathbf{X}_{i1}, \ldots, \mathbf{X}_{im_i})^T$ be the associated $m_i \times p_n$ matrix of covariates.

The marginal regression approach of GEE assumes that $\mathrm{E}(Y_{ij}|\mathbf{X}_{ij}) = \pi_{ij}$ and $\mathrm{Var}(Y_{ij}|\mathbf{X}_{ij}) = \pi_{ij}(1 - \pi_{ij})$, where a dispersion parameter may be added in the marginal variance function if overdispersion is suspected to be present. Furthermore, it relates the covariates to the marginal mean by specifying that

$$\mathrm{logit}(\pi_{ij}) = \mathbf{X}_{ij}^T \boldsymbol{\beta}_n, \tag{2.1}$$

where $\mathrm{logit}(\pi_{ij}) = \log(\frac{\pi_{ij}}{1-\pi_{ij}})$ is the link function and $\boldsymbol{\beta}_n$ is a p_n-dimensional vector of parameters. The true unknown parameter value is denoted by $\boldsymbol{\beta}_{n0}$.

Let $\boldsymbol{\pi}_i(\boldsymbol{\beta}_n) = (\pi_{i1}(\boldsymbol{\beta}_n), \ldots, \pi_{im_i}(\boldsymbol{\beta}_n))^T$, where $\pi_{ij}(\boldsymbol{\beta}_n) = \exp(\mathbf{X}_{ij}^T\boldsymbol{\beta}_n)/[1 + \exp(\mathbf{X}_{ij}^T\boldsymbol{\beta}_n)]$. Further, let $\mathbf{A}_i(\boldsymbol{\beta}_n)$ be the $m_i \times m_i$ diagonal matrix with the jth diagonal element $\mathbf{A}_{ij}(\boldsymbol{\beta}_n) = \pi_{ij}(\boldsymbol{\beta}_n)(1 - \pi_{ij}(\boldsymbol{\beta}_n))$, $j = 1, \ldots, m_i$. In what follows, we assume $m_i = m < \infty$, for simplicity. Liang and Zeger (1986) suggested to estimate $\boldsymbol{\beta}_{n0}$ by solving the following generalized estimating equation in $\boldsymbol{\beta}_n$:

$$\sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{A}_i(\boldsymbol{\beta}_n)\mathbf{V}_i^{-1}\big(\mathbf{Y}_i - \boldsymbol{\pi}_i(\boldsymbol{\beta}_n)\big) = 0, \tag{2.2}$$

where $\mathbf{V}_i$ is a working covariance matrix.

3. Asymptotic properties when $p_n \to \infty$.

3.1. *GEE estimator with estimated working correlation matrix.* In applications, the true correlation matrix of $\mathbf{Y}_i$, denoted by $\mathbf{R}_0$, is unknown. The working covariance matrix is often specified via a working correlation matrix $\mathbf{R}(\boldsymbol{\tau})$: $\mathbf{V}_i = \mathbf{A}_i^{1/2}(\boldsymbol{\beta}_n)\mathbf{R}(\boldsymbol{\tau})\mathbf{A}_i^{1/2}(\boldsymbol{\beta}_n)$, where $\boldsymbol{\tau}$ is a finite-dimensional parameter. Com-

monly used working correlation structures include AR-1, compound symmetry and unstructured working correlation, among others. Note that, in practice, the working correlation matrix is chosen to be independent of the covariates, for simplicity. However, for correlated non-normal data, the range of correlation generally depends on the univariate marginals. Thus, $\mathbf{R}(\boldsymbol{\tau})$ should be understood as a weight matrix [Chaganty and Joe (2004)]. Chaganty and Joe demonstrated that GEE with an appropriately chosen working correlation matrix does have good efficiency when compared with a proper likelihood model.

Given a working correlation structure, $\boldsymbol{\tau}$ is often estimated using a residual-based moment method, which requires an initial consistent estimator of $\boldsymbol{\beta}_{n0}$. We use $\widehat{\mathbf{R}}$ to denote the resulting estimated working correlation matrix, with the subscript "n" suppressed. Following (2.2), we formally define the *GEE estimator* $\widehat{\boldsymbol{\beta}}_n$ as the solution of

$$\mathbf{S}_n(\boldsymbol{\beta}_n) = \sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{A}_i^{1/2}(\boldsymbol{\beta}_n)\widehat{\mathbf{R}}^{-1}\mathbf{A}_i^{-1/2}(\boldsymbol{\beta}_n)\big(\mathbf{Y}_i - \boldsymbol{\pi}_i(\boldsymbol{\beta}_n)\big) = 0. \tag{3.1}$$

To solve for $\widehat{\boldsymbol{\beta}}_n$, we can iterate between a modified Fisher scoring algorithm for $\boldsymbol{\beta}_n$ and the moment estimation for $\boldsymbol{\tau}$. In the following, we provide examples of an initial consistent estimator and an estimated working correlation matrix.

EXAMPLE 1 (Initial estimator for $\boldsymbol{\beta}_{n0}$ when $p_n \to \infty$). A simple way to obtain an initial estimator for $\boldsymbol{\beta}_{n0}$ is to solve the generalized estimating equations under the working independence assumption

$$\widetilde{S}_n(\boldsymbol{\beta}_n) = \sum_{i=1}^{n} \mathbf{X}_i^T\big(\mathbf{Y}_i - \boldsymbol{\pi}_i(\boldsymbol{\beta}_n)\big) = 0. \tag{3.2}$$

Under conditions (A1)–(A3) in Section 3.2, we can show that if $p_n^2/n \to 0$ as $n \to \infty$, then the independence estimating equations in (3.2) have a root $\widetilde{\boldsymbol{\beta}}_n$ such that

$$\|\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}\| = O_p\big(\sqrt{p_n/n}\big), \tag{3.3}$$

where $\|\cdot\|$ denotes the Euclidean norm of a vector. A detailed derivation of (3.3) is given in the Appendix.

EXAMPLE 2 (Estimated working correlation matrix when $p_n \to \infty$). In Balan and Schiopu-Kratina (2005), it was suggested to use

$$\widehat{\mathbf{R}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{A}_i^{-1/2}(\widetilde{\boldsymbol{\beta}}_n)\big(\mathbf{Y}_i - \boldsymbol{\pi}_i(\widetilde{\boldsymbol{\beta}}_n)\big)\big(\mathbf{Y}_i - \boldsymbol{\pi}_i(\widetilde{\boldsymbol{\beta}}_n)\big)^T \mathbf{A}_i^{-1/2}(\widetilde{\boldsymbol{\beta}}_n),$$

where $\widetilde{\boldsymbol{\beta}}_n$ is a preliminary $\sqrt{n/p_n}$-consistent estimator of $\boldsymbol{\beta}_{n0}$, such as the one discussed in Example 1. This provides a moment estimator of the unstructured

working correlation matrix. Assuming conditions (A1)–(A3) of Section 3.2, we can prove that if $p_n^2/n \to 0$ as $n \to \infty$, then

$$\|\widehat{\mathbf{R}}^{-1} - \mathbf{R}_0^{-1}\| = O_p(\sqrt{p_n/n}), \tag{3.4}$$

where $\mathbf{R}_0$ denotes the true common correlation matrix. Here, and throughout the paper, for a matrix $\mathbf{B}$, $\|\mathbf{B}\| = [\mathrm{Tr}(\mathbf{B}\mathbf{B}^T)]^{1/2}$ denotes its Frobenius norm. A detailed derivation of (3.4) is given in the supplementary article [Wang (2010)].

3.2. *Existence and consistency.* In Fan and Peng (2004), Lam and Fan (2008) and Huang, Horowitz and Ma (2008), the estimator is defined as the minimizer of a certain objective function. We use alternative techniques here to establish the existence and consistency of the GEE estimator, which involve the roots of estimating equations. The approach we adopt here is also different from that of Xie and Yang (2003) and Balan and Schiopu-Kratina (2005), both of which rely on properties of injective functions.

We directly verify the following condition: $\forall \varepsilon > 0$, there exists a constant $\Delta > 0$ such that for all n sufficiently large,

$$P\Big(\sup_{\|\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}\| = \Delta\sqrt{p_n/n}} (\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0})^T \mathbf{S}_n(\boldsymbol{\beta}_n) < 0\Big) \geq 1 - \varepsilon. \tag{3.5}$$

Condition (3.5) is sufficient to ensure the existence of a sequence of roots $\widehat{\boldsymbol{\beta}}_n$ of the equation $\mathbf{S}_n(\boldsymbol{\beta}_n) = 0$ such that $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}\| = O_P(\sqrt{p_n/n})$. This approach follows from Theorem 6.3.4 of Ortega and Rheinboldt (1970). In Portnoy (1984), this technique was applied to establish the existence and consistency of an M-estimator for i.i.d. data; in a different setting, it was used by Wang et al. (2010) to study a partial linear single-index model. This leads to a more straightforward and elegant proof of weak consistency. On the other hand, the method relying on injective functions [Xie and Yang (2003); Balan and Schiopu-Kratina (2005)] can also be used to prove strong consistency.

To prove consistency and asymptotic normality, we need the following general regularity conditions:

(A1) $\sup_{i,j} \|\mathbf{X}_{ij}\| = O(\sqrt{p_n})$;

(A2) the unknown parameter $\boldsymbol{\beta}_n$ belongs to a compact subset $\mathcal{B} \subseteq R^{p_n}$, the true parameter value $\boldsymbol{\beta}_{n0}$ lies in the interior of $\mathcal{B}$ and there exist two positive constants, b_1 and b_2, such that $0 < b_1 \leq \pi_{ij}(\boldsymbol{\beta}_{n0}) \leq b_2 < 1$, $\forall i, j$;

(A3) there exist two positive constants, b_3 and b_4, such that

$$b_3 \leq \lambda_{\min}\left(n^{-1}\sum_{i=1}^n \mathbf{X}_i^T\mathbf{X}_i\right) \leq \lambda_{\max}\left(n^{-1}\sum_{i=1}^n \mathbf{X}_i^T\mathbf{X}_i\right) \leq b_4,$$

where $\lambda_{\min}$ (resp. $\lambda_{\max}$) denotes the minimum (resp. maximum) eigenvalue of a matrix;

(A4) the common true correlation matrix $\mathbf{R}_0$ has eigenvalues bounded away from zero and $+\infty$; the estimated working correlation matrix $\widehat{\mathbf{R}}$ satisfies $\|\widehat{\mathbf{R}}^{-1} - \overline{\mathbf{R}}^{-1}\| = O_p(\sqrt{p_n/n})$, where $\overline{\mathbf{R}}$ is a constant positive definite matrix with eigenvalues bounded away from zero and $+\infty$; we do not require $\overline{\mathbf{R}}$ to be the true correlation matrix $\mathbf{R}_0$.

REMARK 1. Condition (A1) is a common assumption in the literature on M-estimators with diverging dimension. For example, it is the same as assumption (3.9) of Portnoy (1985) and it is implied by conditions (C.9) and (C.10) of Welsh (1989). This condition holds almost surely under some weak moment conditions for X_{ij} from spherically symmetric distributions [see, e.g., the discussions in He and Shao (2000)]. When $m = 1$ (i.e., each cluster has only one observation), condition (A3) is also popularly adopted in the literature on high-dimensional regression for independent data. It can be shown that condition (A3) is implied by the following slightly stronger condition: there exist two positive constants, $c_1 \le c_2$, such that $\forall 1 \le j \le m$,

$$c_1 \le \lambda_{\min}\left(n^{-1}\sum_{i=1}^{n}\mathbf{X}_{ij}\mathbf{X}_{ij}^T\right) \le \lambda_{\max}\left(n^{-1}\sum_{i=1}^{n}\mathbf{X}_{ij}\mathbf{X}_{ij}^T\right) \le c_2.$$

Finally, condition (A4) is a direct extension of a similar assumption in the "fixed p" case. Liang and Zeger (1986) assumes that the estimator of the working correlation matrix parameter $\widehat{\boldsymbol{\tau}}$ satisfies $\sqrt{n}(\widehat{\boldsymbol{\tau}} - \boldsymbol{\tau}_0) = O_p(1)$ for some τ_0. Assumption (C2) of Chen and Jin (2006) is of similar nature, while Xie and Yang (2003) assumes the nuisance parameter $\boldsymbol{\tau}$ to be completely known. Note that Example 2 in Section 3.1 guarantees that (A4) is satisfied when a nonparametric moment estimator is used for the working correlation matrix, in which case $\overline{\mathbf{R}} = \mathbf{R}_0$.

We use notation similar to that in Xie and Yang (2003) and Balan and Schiopu-Kratina (2005). Consider the following estimating equation:

$$\overline{\mathbf{S}}_n(\boldsymbol{\beta}_n) = \sum_{i=1}^{n}\mathbf{X}_i^T\mathbf{A}_i^{1/2}(\boldsymbol{\beta}_n)\overline{\mathbf{R}}^{-1}\mathbf{A}_i^{-1/2}(\boldsymbol{\beta}_n)(\mathbf{Y}_i - \boldsymbol{\pi}_i(\boldsymbol{\beta}_n)).$$

If we let $\overline{\mathbf{M}}_n(\boldsymbol{\beta}_n)$ denote the covariance matrix of $\overline{\mathbf{S}}_n(\boldsymbol{\beta}_n)$, then

$$\overline{\mathbf{M}}_n(\boldsymbol{\beta}_n) = \sum_{i=1}^{n}\mathbf{X}_i^T\mathbf{A}_i^{1/2}(\boldsymbol{\beta}_n)\overline{\mathbf{R}}^{-1}\mathbf{R}_0\overline{\mathbf{R}}^{-1}\mathbf{A}_i^{1/2}(\boldsymbol{\beta}_n)\mathbf{X}_i.$$

To prove the consistency, the essential idea is to approximate $\mathbf{S}_n(\boldsymbol{\beta}_n)$ by $\overline{\mathbf{S}}_n(\boldsymbol{\beta}_n)$, whose moments are easier to evaluate. Lemma 3.1 below establishes the accuracy of this approximation, which also plays an important role in deriving the asymptotic normality in Section 3.3.

LEMMA 3.1. *Assume conditions* (A1)–(A4). *If* $n^{-1}p_n^2 = o(1)$, *then*

$$\|\mathbf{S}_n(\boldsymbol{\beta}_{n0}) - \overline{\mathbf{S}}_n(\boldsymbol{\beta}_{n0})\| = O_p(p_n).$$

To facilitate the Taylor expansion of the estimating function $\mathbf{S}_n(\boldsymbol{\beta}_n)$, we also use $\overline{\mathbf{D}}_n(\boldsymbol{\beta}_n) = -\frac{\partial}{\partial \boldsymbol{\beta}_n^T}\overline{\mathbf{S}}_n(\boldsymbol{\beta}_n)$ to approximate the negative gradient function $\mathbf{D}_n(\boldsymbol{\beta}_n) = -\frac{\partial}{\partial \boldsymbol{\beta}_n^T}\mathbf{S}_n(\boldsymbol{\beta}_n)$. Lemma 3.2 below provides a useful representation of $\overline{\mathbf{D}}_n(\boldsymbol{\beta}_n)$, based on which, Lemma 3.3 establishes the approximation of gradient functions.

LEMMA 3.2.

$$\overline{\mathbf{D}}_n(\boldsymbol{\beta}_n) = \overline{\mathbf{H}}_n(\boldsymbol{\beta}_n) + \overline{\mathbf{E}}_n(\boldsymbol{\beta}_n) + \overline{\mathbf{G}}_n(\boldsymbol{\beta}_n), \tag{3.6}$$

where

$$\overline{\mathbf{H}}_n(\boldsymbol{\beta}_n) = \sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{A}_i^{1/2}(\boldsymbol{\beta}_n)\overline{\mathbf{R}}^{-1}\mathbf{A}_i^{1/2}(\boldsymbol{\beta}_n)\mathbf{X}_i,$$

$$\overline{\mathbf{E}}_n(\boldsymbol{\beta}_n) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}(1 - 2\pi_{ij}(\boldsymbol{\beta}_n))\varepsilon_{ij}(\boldsymbol{\beta}_n)\mathbf{X}_i^T \mathbf{A}_i^{1/2}(\boldsymbol{\beta}_n)\overline{\mathbf{R}}^{-1}\mathbf{e}_j\mathbf{e}_j^T\mathbf{X}_i,$$

$$\overline{\mathbf{G}}_n(\boldsymbol{\beta}_n) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}(1 - 2\pi_{ij}(\boldsymbol{\beta}_n))\mathbf{A}_{ij}^{1/2}(\boldsymbol{\beta}_n)\mathbf{X}_{ij}\mathbf{X}_{ij}^T\mathbf{e}_j^T\overline{\mathbf{R}}^{-1}\boldsymbol{\varepsilon}_i(\boldsymbol{\beta}_n),$$

where $\varepsilon_{ij}(\boldsymbol{\beta}_n) = \mathbf{A}_{ij}^{-1/2}(\boldsymbol{\beta}_n)(Y_{ij} - \pi_{ij}(\boldsymbol{\beta}_n))$, $\boldsymbol{\varepsilon}_i(\boldsymbol{\beta}_n) = \mathbf{A}_i^{-1/2}(\boldsymbol{\beta}_n)(\mathbf{Y}_i - \boldsymbol{\pi}(\boldsymbol{\beta}_n))$ *and* $\mathbf{e}_j$ *denotes a unit vector of length* m *whose* j*th entry is* 1 *and all other entries of which are* 0.

LEMMA 3.3. *Assume conditions* (A1)–(A4). *If* $n^{-1}p_n^2 = o(1)$, *then* $\forall \Delta > 0$, *for* $\mathbf{b}_n \in R^{p_n}$, *we have*

$$\sup_{\|\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}\| \le \Delta\sqrt{p_n/n}} \sup_{\|\mathbf{b}_n\|=1} |\mathbf{b}_n^T[\mathbf{D}_n(\boldsymbol{\beta}_n) - \overline{\mathbf{D}}_n(\boldsymbol{\beta}_n)]\mathbf{b}_n| = O_p\big(\sqrt{np_n}\big).$$

REMARK 2. The matrix $\mathbf{D}_n(\boldsymbol{\beta}_n) - \overline{\mathbf{D}}_n(\boldsymbol{\beta}_n)$ is symmetric. The above lemma immediately implies that

$$\sup_{\|\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}\| \le \Delta\sqrt{p_n/n}} |\lambda_{\min}[\mathbf{D}_n(\boldsymbol{\beta}_n) - \overline{\mathbf{D}}_n(\boldsymbol{\beta}_n)]| = O_p\big(\sqrt{np_n}\big),$$

$$\sup_{\|\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}\| \le \Delta\sqrt{p_n/n}} |\lambda_{\max}[\mathbf{D}_n(\boldsymbol{\beta}_n) - \overline{\mathbf{D}}_n(\boldsymbol{\beta}_n)]| = O_p\big(\sqrt{np_n}\big).$$

Furthermore, we can use the leading term $\overline{\mathbf{H}}_n(\boldsymbol{\beta}_n)$ in (3.6) to approximate the negative gradient function $\overline{\mathbf{D}}_n(\boldsymbol{\beta}_n)$. This result is given by Lemma 3.4 below. Lemma 3.5 further establishes an equicontinuity result for $\overline{\mathbf{H}}_n(\boldsymbol{\beta}_n)$.

LEMMA 3.4. *Assume conditions* (A1)–(A4). *If* $n^{-1}p_n^2 = o(1)$, *then* $\forall \Delta > 0$, *for* $\mathbf{b}_n \in R^{p_n}$, *we have*

$$\sup_{\|\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}\| \le \Delta\sqrt{p_n/n}} \sup_{\|\mathbf{b}_n\|=1} |\mathbf{b}_n^T[\overline{\mathbf{D}}_n(\boldsymbol{\beta}_n) - \overline{\mathbf{H}}_n(\boldsymbol{\beta}_n)]\mathbf{b}_n| = O_p\big(\sqrt{n}\,p_n\big).$$

LEMMA 3.5. *Assume conditions* (A1)–(A4). *If* $n^{-1}p_n^2 = o(1)$, *then* $\forall \Delta > 0$, *for* $\mathbf{b}_n \in R^{p_n}$, *we have*

$$\sup_{\|\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}\| \le \Delta\sqrt{p_n/n}} \sup_{\|\mathbf{b}_n\|=1} |\mathbf{b}_n^T[\overline{\mathbf{H}}_n(\boldsymbol{\beta}_n) - \overline{\mathbf{H}}_n(\boldsymbol{\beta}_{n0})]\mathbf{b}_n| = O_p\big(\sqrt{n}\,p_n\big).$$

The proofs of Lemmas 3.1–3.4 are given in the Appendix; the proof of Lemma 3.5 is given in the supplementary article [Wang (2010)]. The following theorem ensures the existence and consistency of the GEE estimator when $p_n \to \infty$.

THEOREM 3.6 (Existence and consistency). *Assume conditions* (A1)–(A4) *and that* $n^{-1}p_n^2 = o(1)$. *Then*, $\mathbf{S}_n(\boldsymbol{\beta}_n) = 0$ *has a root* $\widehat{\boldsymbol{\beta}}_n$ *such that*

$$\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}\| = O_p\big(\sqrt{p_n/n}\big).$$

PROOF. We will prove that (3.5) holds. This requires us to evaluate the sign of $(\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0})^T\mathbf{S}_n(\boldsymbol{\beta}_n)$ on $\{\boldsymbol{\beta}_n : \|\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}\| = \Delta\sqrt{p_n/n}\}$. Note that

$$\begin{aligned}
&(\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0})^T\mathbf{S}_n(\boldsymbol{\beta}_n)\\
&\quad = (\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0})^T\mathbf{S}_n(\boldsymbol{\beta}_{n0}) - (\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0})^T\mathbf{D}_n(\boldsymbol{\beta}_n^*)(\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0})\\
&\quad \triangleq I_{n1} + I_{n2},
\end{aligned}$$

where $\boldsymbol{\beta}_n^*$ lies between $\boldsymbol{\beta}_n$ and $\boldsymbol{\beta}_{n0}$, that is, $\boldsymbol{\beta}_n^* = t\boldsymbol{\beta}_n + (1-t)\boldsymbol{\beta}_{n0}$ for some $0 < t < 1$. Next, we write

$$\begin{aligned}
I_{n1} &= (\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0})^T\overline{\mathbf{S}}_n(\boldsymbol{\beta}_{n0}) + (\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0})^T[\mathbf{S}_n(\boldsymbol{\beta}_{n0}) - \overline{\mathbf{S}}_n(\boldsymbol{\beta}_{n0})]\\
&\triangleq I_{n11} + I_{n12}.
\end{aligned}$$

We have $|I_{n11}| \le \|\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}\| \cdot \|\overline{\mathbf{S}}_n(\boldsymbol{\beta}_{n0})\| = \Delta\sqrt{p_n/n}\|\overline{\mathbf{S}}_n(\boldsymbol{\beta}_{n0})\|$ by the Cauchy–Schwarz inequality. Furthermore,

$$\begin{aligned}
&E[\|\overline{\mathbf{S}}_n(\boldsymbol{\beta}_{n0})\|^2]\\
&\quad = E\left\{\sum_{i=1}^n \boldsymbol{\varepsilon}_i^T(\boldsymbol{\beta}_{n0})\overline{\mathbf{R}}^{-1}\mathbf{A}_i^{1/2}(\boldsymbol{\beta}_{n0})\mathbf{X}_i\mathbf{X}_i^T\mathbf{A}_i^{1/2}(\boldsymbol{\beta}_{n0})\overline{\mathbf{R}}^{-1}\boldsymbol{\varepsilon}_i(\boldsymbol{\beta}_{n0})\right\}\\
&\quad \le \sum_{i=1}^n \lambda_{\max}(\mathbf{X}_i\mathbf{X}_i^T)\lambda_{\max}(\mathbf{A}_i(\boldsymbol{\beta}_{n0}))\lambda_{\max}(\overline{\mathbf{R}}^{-2})E[\boldsymbol{\varepsilon}_i^T(\boldsymbol{\beta}_{n0})\boldsymbol{\varepsilon}_i(\boldsymbol{\beta}_{n0})]\\
&\quad \le C\,\mathrm{Tr}\left(\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^T\right) = C\sum_{i=1}^n\sum_{j=1}^m \mathbf{X}_{ij}^T\mathbf{X}_{ij} = O(np_n).
\end{aligned}$$

Here, and throughout the paper, we use C to denote a generic positive constant which may vary from line to line. Thus, $\|\overline{\mathbf{S}}_n(\boldsymbol{\beta}_{n0})\| = O_p(\sqrt{np_n})$. This implies that $|I_{n11}| = \Delta O_p(p_n)$. For I_{n12}, we have

$$|I_{n12}| \le \|\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}\| \cdot \|\mathbf{S}_n(\boldsymbol{\beta}_{n0}) - \overline{\mathbf{S}}_n(\boldsymbol{\beta}_{n0})\| = \Delta\sqrt{p_n/n}\, O_p(p_n) = \Delta o_p(p_n),$$

by Lemma 3.1. Hence, $|I_{n1}| = \Delta O_p(p_n)$. In what follows, we evaluate I_{n2}:

$$\begin{aligned} I_{n2} &= -(\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0})^T \overline{\mathbf{D}}_n(\boldsymbol{\beta}_n^*)(\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}) \\ &\quad -(\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0})^T [\mathbf{D}_n(\boldsymbol{\beta}_n^*) - \overline{\mathbf{D}}_n(\boldsymbol{\beta}_n^*)](\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}) \\ &\triangleq I_{n21} + I_{n22}. \end{aligned}$$

First, note that

$$\begin{aligned} |I_{n22}| &\le \max\bigl(|\lambda_{\max}(\mathbf{D}_n(\boldsymbol{\beta}_n^*) - \overline{\mathbf{D}}_n(\boldsymbol{\beta}_n^*))|, |\lambda_{\min}(\mathbf{D}_n(\boldsymbol{\beta}_n^*) - \overline{\mathbf{D}}_n(\boldsymbol{\beta}_n^*))|\bigr) \\ &\quad \times \|\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}\|^2 \\ &= O_p(\sqrt{np_n})\Delta^2 \frac{p_n}{n} = \Delta^2 o_p(p_n), \end{aligned}$$

by Lemma 3.3. On the other hand,

$$\begin{aligned} I_{n21} &= -(\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0})^T \overline{\mathbf{H}}_n(\boldsymbol{\beta}_{n0})(\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}) \\ &\quad -(\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0})^T [\overline{\mathbf{H}}_n(\boldsymbol{\beta}_n^*) - \overline{\mathbf{H}}_n(\boldsymbol{\beta}_{n0})](\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}) \\ &\quad -(\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0})^T [\overline{\mathbf{D}}_n(\boldsymbol{\beta}_n^*) - \overline{\mathbf{H}}_n(\boldsymbol{\beta}_n^*)](\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}) \\ &\triangleq I_{n21}^a + I_{n21}^b + I_{n21}^c. \end{aligned}$$

From Lemma 3.5, we have $I_{n21}^b = \Delta^2 o_p(p_n)$; from Lemma 3.4, we have $I_{n21}^c = \Delta^2 o_p(p_n)$. Finally, we evaluate I_{n21}^a. We have

$$\begin{aligned} I_{n21}^a &= -(\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0})^T \left[\sum_{i=1}^n \mathbf{X}_i^T \mathbf{A}_i^{1/2}(\boldsymbol{\beta}_{n0}) \overline{\mathbf{R}}^{-1} \mathbf{A}_i^{1/2}(\boldsymbol{\beta}_{n0}) \mathbf{X}_i\right](\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}) \\ &\le -\lambda_{\min}(\overline{\mathbf{R}}^{-1}) \min_i \lambda_{\min}(\mathbf{A}_i(\boldsymbol{\beta}_{n0})) \lambda_{\min}\left(\sum_{i=1}^n \mathbf{X}_i^T \mathbf{X}_i\right) \|\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}\|^2 \\ &\le -C\Delta^2 p_n, \end{aligned}$$

by (A3). Thus, $(\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0})^T \mathbf{S}_n(\boldsymbol{\beta}_n)$ on $\{\boldsymbol{\beta}_n : \|\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}\| = \Delta\sqrt{p_n/n}\}$ is asymptotically dominated in probability by $I_{n11} + I_{n21}^a$, which is negative for Δ large enough. $\square$

3.3. *Asymptotic normality of the GEE estimator.* The asymptotic distribution of the GEE estimator $\widehat{\boldsymbol{\beta}}_n$ is closely related to that of the ideal estimating function $\overline{\mathbf{S}}_n(\boldsymbol{\beta}_{n0})$. When appropriately normalized, $\overline{\mathbf{S}}_n(\boldsymbol{\beta}_{n0})$ has an asymptotic normal distribution, as shown by the following lemma.

LEMMA 3.7. *Assume conditions* (A1)–(A4). *If* $n^{-1}p_n^3 = o(1)$, *then* $\forall \boldsymbol{\alpha}_n \in R^{p_n}$ *such that* $\|\boldsymbol{\alpha}_n\| = 1$, *we have*

$$\boldsymbol{\alpha}_n^T \overline{\mathbf{M}}_n^{-1/2}(\boldsymbol{\beta}_{n0})\overline{\mathbf{S}}_n(\boldsymbol{\beta}_{n0}) \to N(0,1) \quad \textit{in distribution}.$$

To prove Lemma 3.7, we write $\boldsymbol{\alpha}_n^T \overline{\mathbf{M}}_n^{-1/2}(\boldsymbol{\beta}_{n0})\overline{\mathbf{S}}_n(\boldsymbol{\beta}_{n0})$ as a sum of independent random variables and then check the Lindberg–Feller condition for the central limit theorem. The detailed proof is given in the Appendix. The following theorem ensures the asymptotic normality of the GEE estimator when $n^{-1}p_n^3 = o(1)$.

THEOREM 3.8 (Asymptotic normality). *Assume conditions* (A1)–(A4). *If* $n^{-1}p_n^3 = o(1)$, *then* $\forall \boldsymbol{\alpha}_n \in R^{p_n}$ *such that* $\|\boldsymbol{\alpha}_n\| = 1$, *we have*

$$\boldsymbol{\alpha}_n^T \overline{\mathbf{M}}_n^{-1/2}(\boldsymbol{\beta}_{n0})\overline{\mathbf{H}}_n(\boldsymbol{\beta}_{n0})(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}) \to N(0,1)$$

in distribution.

PROOF. We have

$$\begin{aligned}
&\boldsymbol{\alpha}_n^T \overline{\mathbf{M}}_n^{-1/2}(\boldsymbol{\beta}_{n0})\overline{\mathbf{S}}_n(\boldsymbol{\beta}_{n0}) \\
&\quad = \boldsymbol{\alpha}_n^T \overline{\mathbf{M}}_n^{-1/2}(\boldsymbol{\beta}_{n0})\mathbf{S}_n(\boldsymbol{\beta}_{n0}) + \boldsymbol{\alpha}_n^T \overline{\mathbf{M}}_n^{-1/2}(\boldsymbol{\beta}_{n0})[\overline{\mathbf{S}}_n(\boldsymbol{\beta}_{n0}) - \mathbf{S}_n(\boldsymbol{\beta}_{n0})] \\
&\quad = \boldsymbol{\alpha}_n^T \overline{\mathbf{M}}_n^{-1/2}(\boldsymbol{\beta}_{n0})\mathbf{D}_n(\boldsymbol{\beta}_n^*)(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}) \\
&\qquad + \boldsymbol{\alpha}_n^T \overline{\mathbf{M}}_n^{-1/2}(\boldsymbol{\beta}_{n0})[\overline{\mathbf{S}}_n(\boldsymbol{\beta}_{n0}) - \mathbf{S}_n(\boldsymbol{\beta}_{n0})] \\
&\quad = \boldsymbol{\alpha}_n^T \overline{\mathbf{M}}_n^{-1/2}(\boldsymbol{\beta}_{n0})\overline{\mathbf{H}}_n(\boldsymbol{\beta}_{n0})(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}) \\
&\qquad + \boldsymbol{\alpha}_n^T \overline{\mathbf{M}}_n^{-1/2}(\boldsymbol{\beta}_{n0})[\mathbf{D}_n(\boldsymbol{\beta}_n^*) - \overline{\mathbf{H}}_n(\boldsymbol{\beta}_{n0})](\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}) \\
&\qquad + \boldsymbol{\alpha}_n^T \overline{\mathbf{M}}_n^{-1/2}(\boldsymbol{\beta}_{n0})[\overline{\mathbf{S}}_n(\boldsymbol{\beta}_{n0}) - \mathbf{S}_n(\boldsymbol{\beta}_{n0})],
\end{aligned}$$

where, to obtain the second equality, we note that $\mathbf{S}_n(\widehat{\boldsymbol{\beta}}_n) = 0$ and thus, by a Taylor expansion, $\mathbf{S}_n(\boldsymbol{\beta}_{n0}) = \mathbf{D}_n(\boldsymbol{\beta}_n^*)(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0})$ for some $\boldsymbol{\beta}_n^*$ between $\widehat{\boldsymbol{\beta}}_n$ and $\boldsymbol{\beta}_{n0}$. By Lemma 3.7, $\boldsymbol{\alpha}_n^T \overline{\mathbf{M}}_n^{-1/2}(\boldsymbol{\beta}_{n0})\overline{\mathbf{S}}_n(\boldsymbol{\beta}_{n0}) \to N(0,1)$. Therefore, to prove the theorem, it is sufficient to verify that $\forall \Delta > 0$,

$$\begin{aligned}
&\sup_{\|\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}\| \le \Delta\sqrt{p_n/n}} |\boldsymbol{\alpha}_n^T \overline{\mathbf{M}}_n^{-1/2}(\boldsymbol{\beta}_{n0})[\mathbf{D}_n(\boldsymbol{\beta}_n) - \overline{\mathbf{H}}_n(\boldsymbol{\beta}_{n0})](\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0})| \\
&\quad = o_p(1)
\end{aligned} \tag{3.7}$$

and

$$|\boldsymbol{\alpha}_n^T \overline{\mathbf{M}}_n^{-1/2}(\boldsymbol{\beta}_{n0})[\overline{\mathbf{S}}_n(\boldsymbol{\beta}_{n0}) - \mathbf{S}_n(\boldsymbol{\beta}_{n0})]| = o_p(1). \tag{3.8}$$

We prove (3.8) first. Note that

$$\begin{aligned}
&\big[\boldsymbol{\alpha}_n^T \overline{\mathbf{M}}_n^{-1/2}(\boldsymbol{\beta}_{n0})[\overline{\mathbf{S}}_n(\boldsymbol{\beta}_{n0}) - \mathbf{S}_n(\boldsymbol{\beta}_{n0})]\big]^2 \\
&\quad = \boldsymbol{\alpha}_n^T \overline{\mathbf{M}}_n^{-1/2}(\boldsymbol{\beta}_{n0})[\overline{\mathbf{S}}_n(\boldsymbol{\beta}_{n0}) - \mathbf{S}_n(\boldsymbol{\beta}_{n0})][\overline{\mathbf{S}}_n(\boldsymbol{\beta}_{n0}) - \mathbf{S}_n(\boldsymbol{\beta}_{n0})]^T \overline{\mathbf{M}}_n^{-1/2}(\boldsymbol{\beta}_{n0})\boldsymbol{\alpha}_n \\
&\quad \le \lambda_{\max}(\overline{\mathbf{M}}_n^{-1}(\boldsymbol{\beta}_{n0}))\lambda_{\max}\big([\overline{\mathbf{S}}_n(\boldsymbol{\beta}_{n0}) - \mathbf{S}_n(\boldsymbol{\beta}_{n0})][\overline{\mathbf{S}}_n(\boldsymbol{\beta}_{n0}) - \mathbf{S}_n(\boldsymbol{\beta}_{n0})]^T\big) \\
&\quad \le \frac{\|\overline{\mathbf{S}}_n(\boldsymbol{\beta}_{n0}) - \mathbf{S}_n(\boldsymbol{\beta}_{n0})\|^2}{\lambda_{\min}(\overline{\mathbf{M}}_n(\boldsymbol{\beta}_{n0}))} \\
&\quad \le \frac{\|\overline{\mathbf{S}}_n(\boldsymbol{\beta}_{n0}) - \mathbf{S}_n(\boldsymbol{\beta}_{n0})\|^2}{C\lambda_{\min}(\sum_{i=1}^n \mathbf{X}_i^T \mathbf{X}_i)} \\
&\quad = O_p(p_n^2/n) = o_p(1),
\end{aligned}$$

by Lemma 3.1 and the fact that

$$\lambda_{\min}(\overline{\mathbf{M}}_n(\boldsymbol{\beta}_{n0})) \ge C\lambda_{\min}\left(\sum_{i=1}^n \mathbf{X}_i^T \mathbf{X}_i\right). \tag{3.9}$$

A justification of (3.9) is given in the proof of Lemma 3.7 in the Appendix. Thus, (3.8) holds. Next, we prove (3.7). We have

$$\begin{aligned}
&\sup_{\|\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}\| \le \Delta\sqrt{p_n/n}} |\boldsymbol{\alpha}_n^T \overline{\mathbf{M}}_n^{-1/2}(\boldsymbol{\beta}_{n0})[\mathbf{D}_n(\boldsymbol{\beta}_n) - \overline{\mathbf{H}}_n(\boldsymbol{\beta}_{n0})](\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0})| \\
&\quad \le \sup_{\|\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}\| \le \Delta\sqrt{p_n/n}} |\boldsymbol{\alpha}_n^T \overline{\mathbf{M}}_n^{-1/2}(\boldsymbol{\beta}_{n0})[\mathbf{D}_n(\boldsymbol{\beta}_n) - \overline{\mathbf{D}}_n(\boldsymbol{\beta}_n)](\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0})| \\
&\qquad + \sup_{\|\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}\| \le \Delta\sqrt{p_n/n}} |\boldsymbol{\alpha}_n^T \overline{\mathbf{M}}_n^{-1/2}(\boldsymbol{\beta}_{n0})[\overline{\mathbf{D}}_n(\boldsymbol{\beta}_n) - \overline{\mathbf{H}}_n(\boldsymbol{\beta}_n)](\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0})| \\
&\qquad + \sup_{\|\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}\| \le \Delta\sqrt{p_n/n}} |\boldsymbol{\alpha}_n^T \overline{\mathbf{M}}_n^{-1/2}(\boldsymbol{\beta}_{n0})[\overline{\mathbf{H}}_n(\boldsymbol{\beta}_n) - \overline{\mathbf{H}}_n(\boldsymbol{\beta}_{n0})](\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0})| \\
&\quad \triangleq I_{n1} + I_{n2} + I_{n3}.
\end{aligned}$$

By the Cauchy–Schwarz inequality and Remark 2, we have

$$\begin{aligned}
I_{n1} \le \sup_{\|\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}\| \le \Delta\sqrt{p_n/n}} &\big[\boldsymbol{\alpha}_n^T \overline{\mathbf{M}}_n^{-1/2}(\boldsymbol{\beta}_{n0})\big(\mathbf{D}_n(\boldsymbol{\beta}_n) - \overline{\mathbf{D}}_n(\boldsymbol{\beta}_n)\big)^2 \\
&\times \overline{\mathbf{M}}_n^{-1/2}(\boldsymbol{\beta}_{n0})\boldsymbol{\alpha}_n\big]^{1/2} \|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}\|
\end{aligned}$$

$$\le \sup_{\|\boldsymbol{\beta}_n-\boldsymbol{\beta}_{n0}\|\le\Delta\sqrt{p_n/n}} \max\bigl(\bigl|\lambda_{\min}(\mathbf{D}_n(\boldsymbol{\beta}_n)-\overline{\mathbf{D}}_n(\boldsymbol{\beta}_n))\bigr|,$$
$$\bigl|\lambda_{\max}(\mathbf{D}_n(\boldsymbol{\beta}_n)-\overline{\mathbf{D}}_n(\boldsymbol{\beta}_n))\bigr|\bigr)$$
$$\times \lambda_{\min}^{-1/2}(\overline{\mathbf{M}}_n(\boldsymbol{\beta}_{n0}))O_p(p_n^{1/2}n^{-1/2})$$
$$= O_p(\sqrt{n}p_n)O(n^{-1/2})O_p(n^{-1/2}p_n^{1/2}) = O_P(n^{-1/2}p_n^{3/2}) = o_p(1).$$

Hence, $I_{n1} = o_p(1)$. By the same argument and Lemmas 3.4 and 3.5, we also have $I_{n2} = o_p(1)$ and $I_{n3} = o_p(1)$. This proves (3.7). □

REMARK 3. Note that the condition $n^{-1}p_n^3 = o(1)$ is the same as that of Huber (1973) for an M-estimator with independent data and diverging number of parameters. It is weaker than the condition $n^{-1}p_n^5 = o(1)$ in Fan and Peng (2004) and Lam and Fan (2008) for asymptotic normality.

REMARK 4. Combining the result of Theorem 3.8 with the Cramér–Wold device, it is easy to see that for any $l \times p_n$ matrix $\mathbf{B}_n$ with l fixed and such that $\mathbf{B}_n\mathbf{B}_n^T \to \mathbf{F}$, a positive definite matrix, we have

$$\mathbf{B}_n\boldsymbol{\Sigma}_n^{-1/2}(\boldsymbol{\beta}_{n0})(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}) \to \mathbf{N}_l(\mathbf{0}, \mathbf{F}),$$

where

$$\boldsymbol{\Sigma}_n = \overline{\mathbf{H}}_n^{-1}(\boldsymbol{\beta}_{n0})\overline{\mathbf{M}}_n(\boldsymbol{\beta}_{n0})\overline{\mathbf{H}}_n^{-1}(\boldsymbol{\beta}_{n0}).$$

Now, take $\mathbf{B}_n = (\mathbf{L}_n\boldsymbol{\Sigma}_n\mathbf{L}_n^T)^{-1/2}\mathbf{L}_n\boldsymbol{\Sigma}_n^{1/2}$, where $\mathbf{L}_n$ is an $l \times p_n$ matrix such that $\mathbf{L}_n\boldsymbol{\Sigma}_n\mathbf{L}_n^T$ is invertible. Then, $\mathbf{B}_n\mathbf{B}_n^T = \mathbf{I}_l$ and we have the following corollary which gives the asymptotic distribution of $\mathbf{L}_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0})$.

COROLLARY 3.9. *Under the same conditions as in Theorem* 3.8, *if* $n^{-1}p_n^3 = o(1)$, *then*

$$(\mathbf{L}_n\boldsymbol{\Sigma}_n\mathbf{L}_n^T)^{-1/2}\mathbf{L}_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}) \to \mathbf{N}_l(\mathbf{0}, \mathbf{I}_l)$$

in distribution.

3.4. *Sandwich covariance formula and large-sample Wald test.* Theorem 3.8 and Corollary 3.9 suggest that the covariance matrix of $\widehat{\boldsymbol{\beta}}_n$ is approximately $\boldsymbol{\Sigma}_n$. To estimate $\boldsymbol{\Sigma}_n$, Liang and Zeger (1986) proposed, in the "fixed p" setup, the following well-known *sandwich covariance matrix estimator:*

$$\widehat{\boldsymbol{\Sigma}}_n = \mathbf{H}_n^{-1}(\widehat{\boldsymbol{\beta}}_n)\widehat{\mathbf{M}}_n(\widehat{\boldsymbol{\beta}}_n)\mathbf{H}_n^{-1}(\widehat{\boldsymbol{\beta}}_n),$$

where $\mathbf{H}_n(\boldsymbol{\beta}_n)$ is defined similarly as $\overline{\mathbf{H}}_n(\boldsymbol{\beta}_n)$, but with $\overline{\mathbf{R}}$ replaced by $\widehat{\mathbf{R}}$; $\widehat{\mathbf{M}}_n(\boldsymbol{\beta}_n)$ is defined similarly as $\overline{\mathbf{M}}_n(\boldsymbol{\beta}_n)$, except that $\overline{\mathbf{R}}$ is replaced by $\widehat{\mathbf{R}}$ and the unknown true correlation matrix $\mathbf{R}_0$ is replaced by $\boldsymbol{\varepsilon}_i(\boldsymbol{\beta}_n)\boldsymbol{\varepsilon}_i^T(\boldsymbol{\beta}_n)$, with $\boldsymbol{\varepsilon}_i(\boldsymbol{\beta}_n)$ defined in

Lemma 3.2. Based on Corollary 3.9 and the sandwich covariance matrix estimator, an asymptotic $(1-\alpha)\%$ confidence interval $(0<\alpha<1)$ for β_j is

$$\widehat{\boldsymbol{\beta}}_j \pm z_{\alpha/2}\mathbf{u}_j^T \widehat{\boldsymbol{\Sigma}}_n \mathbf{u}_j, \tag{3.10}$$

where $z_{\alpha/2}$ denotes the upper $\frac{\alpha}{2}$ quantile of the standard normal distribution and $\mathbf{u}_j$ is the unit vector of length p_n with the jth element equal to 1 and all the other elements equal to 0.

The sandwich covariance formula plays an important role in GEE methodology. In the "fixed p" setup, it is known that the sandwich covariance matrix estimator provides a consistent estimator for the variance of the GEE estimator, even when the working correlation matrix is misspecified. The following theorem shows that this appealing property is still valid when p_n converges to ∞ at an appropriate rate.

The proofs of Theorem 3.10 and Corollary 3.11 below are given in the Appendix.

THEOREM 3.10. *Assume conditions* (A1)–(A4) *and that* $n^{-1}p_n^3 = o(1)$. *Then,*

$$\mathbf{C}_n \widehat{\boldsymbol{\Sigma}}_n \mathbf{C}_n^T - \mathbf{C}_n \boldsymbol{\Sigma}_n \mathbf{C}_n^T = o_p(n^{-1}),$$

where $\mathbf{C}_n$ *is any* $l \times q_n$ *matrix such that* l *is fixed and* $\mathbf{C}_n\mathbf{C}_n^T = \mathbf{G}$ *with* $\mathbf{G}$ *being an* $l \times l$ *positive definite matrix.*

REMARK 5. It is worth pointing out a subtle issue that is sometimes overlooked in the existing literature on high-dimensional analysis of independent data. In order to justify the validity of the asymptotic confidence interval or large-sample test for estimable contrast, it is necessary to show that the convergence rate in Theorem 3.10 is $o_p(n^{-1})$. Note that the estimable contrast is asymptotically normal with convergence rate $O_p(n^{1/2})$; see, for example, Corollary 2.1 in He and Shao (2000) for the case of an M-estimator based on independent data. In the literature, sometimes only the $o_p(1)$ rate is provided, which is not adequate, but can be fixed.

Next, we consider the large-sample Wald test for testing the following linear hypothesis:

$$H_0 : \mathbf{L}_n \boldsymbol{\beta}_{n0} = \mathbf{0} \quad \text{vs.} \quad H_1 : \mathbf{L}_n \boldsymbol{\beta}_{n0} \neq \mathbf{0},$$

where $\mathbf{L}_n$ is an $l \times p_n$ matrix with l fixed and $\mathbf{L}_n\mathbf{L}_n^T = \mathbf{I}_l$. The Wald test statistic is defined as

$$W_n = (\mathbf{L}_n \widehat{\boldsymbol{\beta}}_n)^T (\mathbf{L}_n \widehat{\boldsymbol{\Sigma}}_n \mathbf{L}_n^T)^{-1} (\mathbf{L}_n \widehat{\boldsymbol{\beta}}_n).$$

The corollary below shows that the Wald test remains valid, even when the number of covariates diverges with the sample size.

COROLLARY 3.11. *Assume conditions* (A1)–(A4). *If* $n^{-1}p_n^3 = o(1)$, *then* $W_n \to \chi_l^2$ *in distribution under* H_0, *where* χ_l^2 *denotes the* χ^2 *distribution with* l *degrees of freedom.*

REMARK 6. For testing a high-dimensional hypothesis $H_0 : \boldsymbol{\beta}_n = \boldsymbol{\beta}_{n0}^*$ versus $H_1 : \boldsymbol{\beta}_n \neq \boldsymbol{\beta}_{n0}^*$, it can be shown that

$$\frac{(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}^*)^T \widehat{\boldsymbol{\Sigma}}_n^{-1} (\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}^*) - p_n}{\sqrt{2p_n}} \to N(0, 1) \tag{3.11}$$

in distribution under H_0, under some regularity conditions. A proof of this result is given in the supplementary article [Wang (2010)].

4. Numerical studies. We consider the following model for the marginal expectation of Y_{ij}, $i = 1, \ldots, n$, given $\mathbf{X}_{ij}$,

$$\operatorname{logit}(\pi_{ij}) = X_{ij}^T \boldsymbol{\beta}_{n0}, \quad j = 1, 2, 3, \tag{4.1}$$

where $\boldsymbol{\beta}_{n0}$ is a p_n-dimensional vector of parameters with $p_n = \lfloor 2.5n^{1/3} \rfloor$, with $\lfloor q \rfloor$ denoting the the largest integer not greater than q. In this example, $\boldsymbol{\beta}_{n0}^T = (0.4 \cdot \mathbf{1}_k^T, -0.3 \cdot \mathbf{1}_k^T, 0.2 \cdot \mathbf{1}_k^T, -0.1 \cdot \mathbf{1}_{p_n - 3k}^T)$, where $\mathbf{1}_k$ denotes a k-dimensional vector of 1's and $k = \lfloor p_n/4 \rfloor$. In addition, $X_{ij} = (x_{ij1}, \ldots, x_{ijp_n})^T$ has a multivariate normal distribution with mean zero, marginal variance 0.2 and an AR-1 correlation matrix with autocorrelation coefficient 0.5. The binary response vector for each cluster has the above marginal mean and an exchangeable (also called compound symmetry or CS) correlation structure with correlation coefficient 0.5. Such correlated binary data are generated using Bahadur's representation [see, e.g., Fitzmaurice (1995)].

Since, for different sample sizes, the parameter dimension is different, we measure the accuracy of estimation by the *simulated average mean square error*, which is obtained by averaging $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}\|^2/p_n$ over 500 simulated samples. Table 1 reports simulation results using four different working correlation structures: independence working correlation matrix (IN), unstructured working correlation matrix (UN), compound symmetry working correlation matrix (CS) and the first order autocorrelation working correlation matrix (AR-1), for sample sizes $n = 500$, 1000, 2000 and 3000. Table 1 demonstrates that when the covariate dimension grows at an appropriate rate with the sample size, the accuracy of GEE estimator is satisfactory. We also observe that when the true correlation matrix (CS in this case) is adopted, the estimator is more efficient.

We next examine the accuracy of the sandwich variance formula. The standard deviations of the estimated coefficients over 500 simulations are averaged and regarded as the true standard error (SD). Table 2 compares SD with the standard error obtained from the sandwich variance formula (SD2) when the unstructured working correlation matrix is used for estimating $\widehat{\beta}_k$, $\widehat{\beta}_{2k}$, $\widehat{\beta}_{3k}$ and $\widehat{\beta}_{p_n}$. We observe that the sandwich variance formula works remarkably well. Similar phenomena are

TABLE 1
The simulated average mean squared error ($\times 10$) *for estimating* $\boldsymbol{\beta}_{n0}$ *using four different working correlation structures*

		Working correlation structure			
n	p_n	IN	UN	CS	AR-1
500	19	0.265	0.156	0.154	0.179
1000	24	0.141	0.103	0.100	0.111
2000	31	0.090	0.074	0.071	0.075
3000	36	0.070	0.065	0.063	0.065

also observed for estimating other regression coefficients and with other working correlation structures, but, for reasons of brevity, these are not reported.

Finally, we investigate hypothesis testing based on the large-sample Wald test. We consider model (4.1) with $n = 1000$, $p_n = 24$ and $\boldsymbol{\beta}_{n0}^T = (0.4 \cdot \mathbf{1}_6^T, -0.3 \cdot \mathbf{1}_6^T, 0.2 \cdot \mathbf{1}_6^T, -0.1 \cdot \mathbf{1}_2^T, 0, 0, 0, 0)$. The left panel of Figure 1 depicts the density of the Wald test under the null hypothesis $H_0 : \beta_{21} = \beta_{22} = \beta_{23} = \beta_{24} = 0$ and compares it with the density curve of the χ_4^2 distribution. It demonstrates that the χ^2 approximation given in Corollary 3.11 is accurate. The right panel of Figure 1 gives the normal Q–Q plot for the Wald test statistic under the null hypothesis $\boldsymbol{\beta}_n = \boldsymbol{\beta}_{n0}$ and it shows that the null distribution can be approximated well by a normal distribution for testing a higher-dimensional alternative, as discussed in Remark 6.

5. Discussions.

5.1. *Extension to general GEE.* Although the focus of the paper is on clustered binary data, the approaches and techniques can be extended to general GEE. For general GEE, the decomposition of $\overline{\mathbf{D}}_n(\boldsymbol{\beta}_n)$ given in Lemma 3.2 has a more complex expression, and the potential unboundedness of Y_{ij} makes the derivation

TABLE 2
Standard deviation (*SD*) *and estimated standard deviation* (*SD*2) *using the sandwich variance formula*

		$\widehat{\beta}_k$		$\widehat{\beta}_{2k}$		$\widehat{\beta}_{3k}$		$\widehat{\beta}_{p_n}$	
n	p_n	SD	SD2	SD	SD2	SD	SD2	SD	SD2
500	19	0.126	0.111	0.114	0.110	0.117	0.111	0.089	0.098
1000	24	0.082	0.083	0.079	0.083	0.085	0.083	0.072	0.074
2000	31	0.073	0.060	0.063	0.060	0.065	0.060	0.051	0.053
3000	36	0.060	0.051	0.049	0.051	0.052	0.051	0.051	0.045

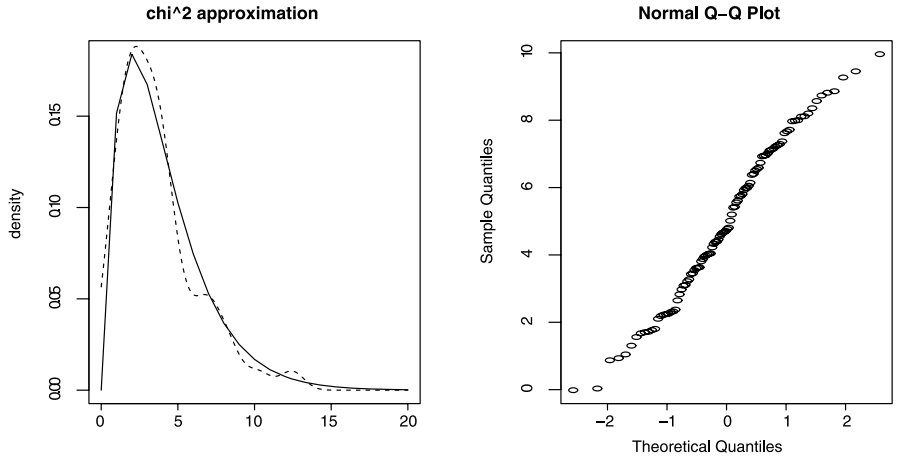

FIG. 1. *The left panel gives the estimated null density of the large-sample Wald test* (*dashed curve*) *and the density of the chi-square distribution with four degrees of freedom* (*solid curve*) *for testing* $H_0 : \beta_{21} = \beta_{22} = \beta_{23} = \beta_{24} = 0$. *The right panel gives the normal Q–Q plot of the Wald test statistic under the null hypothesis* $\boldsymbol{\beta}_n = \boldsymbol{\beta}_{n0}$.

of various probability bounds and asymptotic equivalence more delicate. Below, we give a brief discussion of the large-p asymptotics for general GEE.

Assume that the first two marginal moments of Y_{ij} are $\mu_{ij}(\boldsymbol{\beta}_n) := \mathrm{E}_{\boldsymbol{\beta}_n}(Y_{ij}) = \mu(\theta_{ij})$ and $\sigma_{ij}^2(\boldsymbol{\beta}_n) := \mathrm{Var}_{\boldsymbol{\beta}_n}(Y_{ij}) = \dot{\mu}(\theta_{ij})$, where $\theta_{ij} = \mathbf{X}_{ij}^T\boldsymbol{\beta}_n$. These moment assumptions would follow when the marginal response variable has a canonical exponential family distribution with scaling parameter 1. Let $\mathbf{A}_i(\boldsymbol{\beta}_n) = \mathrm{diag}(\sigma_{i1}^2(\boldsymbol{\beta}_n), \ldots, \sigma_{im}^2(\boldsymbol{\beta}_n))$ and $\boldsymbol{\mu}_i(\boldsymbol{\beta}_n) = (\mu_{i1}(\boldsymbol{\beta}_n), \ldots, \mu_{im}(\boldsymbol{\beta}_n))^T$. The *GEE estimator* $\widehat{\boldsymbol{\beta}}_n$ is the solution of

$$\sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{A}_i^{1/2}(\boldsymbol{\beta}_n)\widehat{\mathbf{R}}^{-1}\mathbf{A}_i^{-1/2}(\boldsymbol{\beta}_n)\big(\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_n)\big) = 0. \tag{5.1}$$

In addition to assumptions (A1)–(A4) in Section 3.2, we adopt two additional conditions:

(A5) there exists a finite constant $M_1 > 0$ such that $E(\|\mathbf{A}_i^{-1/2}(\boldsymbol{\beta}_n)(\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}_n))\|^{2+\delta}) \le M_1$ for all i and some $\delta > 0$;

(A6) if $B_n = \{\boldsymbol{\beta}_n : \|\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}\| \le \Delta\sqrt{p_n/n}\}$, then $\dot{\mu}(\mathbf{X}_{ij}^T\boldsymbol{\beta}_n)$, $1 \le i \le n$, $1 \le j \le m$, are uniformly bounded away from 0 and ∞ on B_n; $\ddot{\mu}(\mathbf{X}_{ij}^T\boldsymbol{\beta}_n)$ and $\mu^{(3)}(\mathbf{X}_{ij}^T\boldsymbol{\beta}_n)$, $1 \le i \le n$, $1 \le j \le m$, are uniformly bounded by a finite positive constant M_2 on B_n.

REMARK 7. Condition (A5) is similar to the condition in Lemma 2 of Xie and Yang (2003) and condition $(\tilde{N}_\delta)$ in Balan and Schiopu-Kratina (2005). Condition (A6) requires $\mu_{ij}^{(k)}(\mathbf{X}_{ij}^T\boldsymbol{\beta}_n)$, $k = 1, 2, 3$, to be uniformly bounded when $\boldsymbol{\beta}_n$ is in a local neighborhood around $\boldsymbol{\beta}_{n0}$. This condition is generally satisfied for GEE. For example, when the marginal model follows a Poisson distribution, $\mu(t) = \exp(t)$, thus $\mu_{ij}^{(k)}(\mathbf{X}_{ij}^T\boldsymbol{\beta}_n) = \exp(\mathbf{X}_{ij}^T\boldsymbol{\beta}_n)$, $k = 1, 2, 3$, are uniformly bounded on B_n.

THEOREM 5.1. *Assume conditions* (A1)–(A6) *and that* $n^{-1}p_n^2 = o(1)$. *The generalized estimating equation* (5.1) *then has a root* $\widehat{\boldsymbol{\beta}}_n$ *such that* $\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}\| = O_p(\sqrt{p_n/n})$. *Furthermore, if* $n^{-1}p_n^3 = o(1)$, *then* $\forall \boldsymbol{\alpha}_n \in R^{p_n}$ *such that* $\|\boldsymbol{\alpha}_n\| = 1$,

$$\boldsymbol{\alpha}_n^T\overline{\mathbf{M}}_n^{-1/2}(\boldsymbol{\beta}_{n0})\overline{\mathbf{H}}_n(\boldsymbol{\beta}_{n0})(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}) \to N(0, 1)$$

in distribution, where $\overline{\mathbf{M}}_n^{-1/2}(\boldsymbol{\beta}_{n0})$ *and* $\overline{\mathbf{H}}_n(\boldsymbol{\beta}_{n0})$ *have the same expressions as in Section* 3.2.

A sketch of the proof of Theorem 5.1 is given in the supplementary article [Wang (2010)].

5.2. *Related problems.* In some scenarios, a "large n, diverging m" asymptotic framework, where p is either fixed or also diverges at an appropriate rate, may be more appropriate. This corresponds to a real situation where the cluster size is itself large. For example, in a longitudinal study, doctors take measurements on the patients during each visit. Each patient forms a cluster. The cluster size is large if the number of visits is large. For a fixed p setting, this "large n, diverging m" asymptotic framework has been considered by Xie and Yang (2003). A future topic of interest is to consider large m together with large p.

Another interesting direction for future study is to consider a more flexible semiparametric specification for the generalized estimating equations in the large-p setting. In the classical "fixed p" setting, GEE with partially linear model specification has been investigated by Lin and Carroll (2001a, 2001b), Lin and Ying (2001), He, Zhu and Fung (2002), Fan and Li (2004), Chiou and Müller (2005), Wang, Carroll and Lin (2005), Chen and Jin (2006), He, Fung and Zhu (2006) and Huang, Zhang and Zhou (2007), among others.

APPENDIX

We use C to denote a generic positive constant that can vary from line to line.

PROOF OF (3.3). It suffices [Ortega and Rheinboldt (1970)] to show that $\forall \varepsilon > 0$, there exists a $\Delta > 0$ such that for all n sufficiently large,

$P(\sup_{\|\boldsymbol{\beta}_n-\boldsymbol{\beta}_{n0}\|=\Delta\sqrt{p_n/n}}(\boldsymbol{\beta}_n-\boldsymbol{\beta}_{n0})^T\widetilde{S}_n(\boldsymbol{\beta}_n)<0)\geq 1-\varepsilon$. We have

$$\begin{aligned}&(\boldsymbol{\beta}_n-\boldsymbol{\beta}_{n0})^T\widetilde{S}_n(\boldsymbol{\beta}_n)\\&\quad=(\boldsymbol{\beta}_n-\boldsymbol{\beta}_{n0})^T\widetilde{S}_n(\boldsymbol{\beta}_{n0})+(\boldsymbol{\beta}_n-\boldsymbol{\beta}_{n0})^T\frac{\partial}{\partial\boldsymbol{\beta}_n^T}\widetilde{S}_n(\boldsymbol{\beta}_n^*)(\boldsymbol{\beta}_n-\boldsymbol{\beta}_{n0})\\&\quad\triangleq I_{n1}+I_{n2},\end{aligned}$$

where $\boldsymbol{\beta}_n^*$ lies between $\boldsymbol{\beta}_{n0}$ and $\boldsymbol{\beta}_n$. We first consider I_{n1}. For any $\boldsymbol{\beta}_n$ such that $\|\boldsymbol{\beta}_n-\boldsymbol{\beta}_{n0}\|=\Delta\sqrt{\frac{p_n}{n}}$, we have $|I_{n1}|\leq\Delta\sqrt{\frac{p_n}{n}}\|\widetilde{S}_n(\boldsymbol{\beta}_{n0})\|$. Note that

$$\begin{aligned}E[\|\widetilde{S}_n(\boldsymbol{\beta}_{n0})\|^2]&=E\left[\sum_{i=1}^n\big(\mathbf{Y}_i-\boldsymbol{\pi}_i(\boldsymbol{\beta}_{n0})\big)^T\mathbf{X}_i\mathbf{X}_i^T\big(\mathbf{Y}_i-\boldsymbol{\pi}_i(\boldsymbol{\beta}_{n0})\big)\right]\\&\leq E\left[\sum_{i=1}^n\lambda_{\max}(\mathbf{X}_i\mathbf{X}_i^T)\|\mathbf{Y}_i-\boldsymbol{\pi}_i(\boldsymbol{\beta}_{n0})\|^2\right]\\&\leq C\operatorname{Tr}\left(\sum_{i=1}^n\mathbf{X}_i\mathbf{X}_i^T\right)=C\sum_{i=1}^n\sum_{j=1}^m\mathbf{X}_{ij}^T\mathbf{X}_{ij}=O(np_n),\end{aligned}$$

by assumption (A1). Thus, $|I_{n1}|\leq\Delta O_p(p_n)$. Next,

$$\begin{aligned}I_{n2}&=-(\boldsymbol{\beta}_n-\boldsymbol{\beta}_{n0})^T\left[\sum_{i=1}^n\mathbf{X}_i^TA_i(\boldsymbol{\beta}_{n0})\mathbf{X}_i\right](\boldsymbol{\beta}_n-\boldsymbol{\beta}_{n0})\\&\quad-(\boldsymbol{\beta}_n-\boldsymbol{\beta}_{n0})^T\left[\sum_{i=1}^n\mathbf{X}_i^T\big(A_i(\boldsymbol{\beta}^*)-A_i(\boldsymbol{\beta}_{n0})\big)\mathbf{X}_i\right](\boldsymbol{\beta}_n-\boldsymbol{\beta}_{n0})\\&\triangleq I_{n21}+I_{n22}.\end{aligned}$$

Note that $I_{n21}\leq-\lambda_{\min}(\mathbf{A}_i(\boldsymbol{\beta}_0))\lambda_{\min}(\sum_{i=1}^n\mathbf{X}_i^T\mathbf{X}_i)\|\boldsymbol{\beta}_n-\boldsymbol{\beta}_{n0}\|^2\leq-Cp_n\Delta^2$, by (A3). Since $\frac{\partial}{\partial\boldsymbol{\beta}_n}\mathbf{A}_{ij}(\boldsymbol{\beta}_n)=\pi_{ij}(\boldsymbol{\beta}_n)(1-\pi_{ij}(\boldsymbol{\beta}_n))(1-2\pi_{ij}(\boldsymbol{\beta}_n))\mathbf{X}_{ij}$, we have

$$\begin{aligned}|I_{n22}|&\leq(\boldsymbol{\beta}_n-\boldsymbol{\beta}_{n0})^T\left[\sum_{i=1}^n\sum_{j=1}^m|\mathbf{A}_{ij}(\boldsymbol{\beta}^*)-\mathbf{A}_{ij}(\boldsymbol{\beta}_{n0})|\mathbf{X}_{ij}\mathbf{X}_{ij}^T\right](\boldsymbol{\beta}_n-\boldsymbol{\beta}_{n0})\\&\leq\sup_{i,j}\|\mathbf{X}_{ij}\|\cdot\|\boldsymbol{\beta}^*-\boldsymbol{\beta}_0\|\cdot\|\boldsymbol{\beta}_n-\boldsymbol{\beta}_0\|^2\cdot\lambda_{\max}\left(\sum_{i=1}^n\mathbf{X}_i^T\mathbf{X}_i\right)\\&\leq O\big(\sqrt{p_n}\big)O_p\big(\sqrt{p_n/n}\big)(\Delta^2p_n/n)O(n)=\Delta^2o_p(p_n),\end{aligned}$$

by (A1)–(A3). Thus, for sufficiently large Δ, $(\boldsymbol{\beta}_n-\boldsymbol{\beta}_{n0})^T\widetilde{S}_n(\boldsymbol{\beta}_n)$ is dominated by I_{n21}, which is large and negative for all sufficiently large n. □

PROOF OF (3.4). The proof is given in the online supplement. □

PROOF OF LEMMA 3.1. Let $\mathbf{Q} = \{q_{j_1,j_2}\}_{1\le j_1,j_2\le m}$ denote the matrix $\widehat{\mathbf{R}}^{-1} - \overline{\mathbf{R}}^{-1}$. Then,

$$\begin{aligned}
&\mathbf{S}_n(\boldsymbol{\beta}_{n0}) - \overline{\mathbf{S}}_n(\boldsymbol{\beta}_{n0}) \\
&\quad= \sum_{i=1}^{n}\sum_{j_1=1}^{m}\sum_{j_2=1}^{m} q_{j_1,j_2}\mathbf{A}_{ij_1}^{1/2}(\boldsymbol{\beta}_{n0})\mathbf{A}_{ij_2}^{-1/2}(\boldsymbol{\beta}_{n0})\big(Y_{ij_2} - \pi_{ij_2}(\boldsymbol{\beta}_{n0})\big)\mathbf{X}_{ij_1} \\
&\quad= \sum_{j_1=1}^{m}\sum_{j_2=1}^{m} q_{j_1,j_2}\left[\sum_{i=1}^{n}\mathbf{A}_{ij_1}^{1/2}(\boldsymbol{\beta}_{n0})\varepsilon_{ij_2}(\boldsymbol{\beta}_{n0})\mathbf{X}_{ij_1}\right],
\end{aligned}$$

where $\varepsilon_{ij_2}(\boldsymbol{\beta}_{n0}) = \mathbf{A}_{ij_2}^{-1/2}(\boldsymbol{\beta}_{n0})(Y_{ij_2} - \pi_{ij_2}(\boldsymbol{\beta}_{n0}))$. Note that

$$\begin{aligned}
E\left[\left\|\sum_{i=1}^{n}\mathbf{A}_{ij_1}^{1/2}(\boldsymbol{\beta}_{n0})\varepsilon_{ij_2}(\boldsymbol{\beta}_{n0})\mathbf{X}_{ij_1}\right\|^2\right] &= \sum_{i=1}^{n}\mathbf{A}_{ij_1}(\boldsymbol{\beta}_{n0})E[\varepsilon_{ij_2}^2(\boldsymbol{\beta}_{n0})]\mathbf{X}_{ij_1}^T\mathbf{X}_{ij_1} \\
&\le \sum_{i=1}^{n}\mathbf{X}_{ij_1}^T\mathbf{X}_{ij_1} = O(np_n).
\end{aligned}$$

Thus, $\|\sum_{i=1}^{n}\mathbf{A}_{ij_1}^{1/2}(\boldsymbol{\beta}_{n0})\varepsilon_{ij_2}(\boldsymbol{\beta}_{n0})\mathbf{X}_{ij_1}\| = O_p(\sqrt{np_n})$ $\forall 1 \le j_1, j_2 \le m$. Since, by (A4), $q_{j_1,j_2} = O_p(\sqrt{p_n/n})$ $\forall 1 \le j_1, j_2 \le m$, the proof is complete. □

PROOF OF LEMMA 3.2. The derivation can be found in Pan (2002). □

PROOF OF LEMMA 3.3. Let $\mathbf{H}_n(\boldsymbol{\beta}_n)$, $\mathbf{E}_n(\boldsymbol{\beta}_n)$ and $\mathbf{G}_n(\boldsymbol{\beta}_n)$ be defined the same as $\overline{\mathbf{H}}_n(\boldsymbol{\beta}_n)$, $\overline{\mathbf{E}}_n(\boldsymbol{\beta}_n)$ and $\overline{\mathbf{G}}_n(\boldsymbol{\beta}_n)$, respectively, but with $\overline{\mathbf{R}}$ replaced by $\widehat{\mathbf{R}}$. By Lemma 3.2, it is sufficient to prove the following three results:

$$\begin{aligned}
&\sup_{\|\boldsymbol{\beta}_n-\boldsymbol{\beta}_{n0}\|\le\Delta\sqrt{p_n/n}}\ \sup_{\|\mathbf{b}_n\|=1}|\mathbf{b}_n^T[\mathbf{H}_n(\boldsymbol{\beta}_n) - \overline{\mathbf{H}}_n(\boldsymbol{\beta}_n)]\mathbf{b}_n| \\
&\qquad= O_p(\sqrt{np_n}),
\end{aligned} \tag{A.1}$$

$$\begin{aligned}
&\sup_{\|\boldsymbol{\beta}_n-\boldsymbol{\beta}_{n0}\|\le\Delta\sqrt{p_n/n}}\ \sup_{\|\mathbf{b}_n\|=1}|\mathbf{b}_n^T[\mathbf{E}_n(\boldsymbol{\beta}_n) - \overline{\mathbf{E}}_n(\boldsymbol{\beta}_n)]\mathbf{b}_n| \\
&\qquad= O_p(\sqrt{np_n}),
\end{aligned} \tag{A.2}$$

$$\begin{aligned}
&\sup_{\|\boldsymbol{\beta}_n-\boldsymbol{\beta}_{n0}\|\le\Delta\sqrt{p_n/n}}\ \sup_{\|\mathbf{b}_n\|=1}|\mathbf{b}_n^T[\mathbf{G}_n(\boldsymbol{\beta}_n) - \overline{\mathbf{G}}_n(\boldsymbol{\beta}_n)]\mathbf{b}_n| \\
&\qquad= O_p(\sqrt{np_n}).
\end{aligned} \tag{A.3}$$

We have

$$
\begin{aligned}
&|\mathbf{b}_n^T[\mathbf{H}_n(\boldsymbol{\beta}_n) - \overline{\mathbf{H}}_n(\boldsymbol{\beta}_n)]\mathbf{b}_n| \\
&\quad = \left|\sum_{i=1}^n \mathbf{b}_n^T \mathbf{X}_i^T \mathbf{A}_i^{1/2}(\boldsymbol{\beta}_n)[\widehat{\mathbf{R}}^{-1} - \overline{\mathbf{R}}^{-1}]\mathbf{A}_i^{1/2}(\boldsymbol{\beta}_n)\mathbf{X}_i\mathbf{b}_n\right| \\
&\quad \le \|\widehat{\mathbf{R}}^{-1} - \overline{\mathbf{R}}^{-1}\|\lambda_{\max}(\mathbf{A}_i(\boldsymbol{\beta}_n))\lambda_{\max}\left(\sum_{i=1}^n \mathbf{X}_i^T\mathbf{X}_i\right)\|\mathbf{b}_n\|^2.
\end{aligned}
$$

By assumptions (A2) and (A4), (A.1) is proved. Next, note that

$$
\begin{aligned}
&|\mathbf{b}_n^T[\mathbf{E}_n(\boldsymbol{\beta}_n) - \overline{\mathbf{E}}_n(\boldsymbol{\beta}_n)]\mathbf{b}_n| \\
&\quad = \frac{1}{2}\left|\sum_{i=1}^n\sum_{j=1}^m (1 - 2\pi_{ij}(\boldsymbol{\beta}_n))\varepsilon_{ij}(\boldsymbol{\beta}_n)\mathbf{b}_n^T\mathbf{X}_i^T\mathbf{A}_i^{1/2}(\boldsymbol{\beta}_n)\right. \\
&\qquad\qquad \left. \times [\widehat{\mathbf{R}}^{-1} - \overline{\mathbf{R}}^{-1}]\mathbf{e}_j\mathbf{e}_j^T\mathbf{X}_i\mathbf{b}_n\right| \\
&\quad \le \sum_{i=1}^n\sum_{j=1}^m \mathbf{A}_{ij}^{-1/2}(\boldsymbol{\beta}_n)|\mathbf{b}_n^T\mathbf{X}_i^T\mathbf{A}_i^{1/2}(\boldsymbol{\beta}_n)[\widehat{\mathbf{R}}^{-1} - \overline{\mathbf{R}}^{-1}]\mathbf{e}_j| \cdot |\mathbf{e}_j^T\mathbf{X}_i\mathbf{b}_n| \\
&\quad \le \sum_{i=1}^n\sum_{j=1}^m \mathbf{A}_{ij}^{-1/2}(\boldsymbol{\beta}_n)\|\widehat{\mathbf{R}}^{-1} - \overline{\mathbf{R}}^{-1}\| \cdot \|\mathbf{A}_i^{1/2}(\boldsymbol{\beta}_n)\| \cdot \|\mathbf{X}_i\mathbf{b}_n\|^2.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\sup_{\|\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}\| \le \Delta\sqrt{p_n/n}} \sup_{\|\mathbf{b}_n\|=1} |\mathbf{b}_n^T[\mathbf{E}_n(\boldsymbol{\beta}_n) - \overline{\mathbf{E}}_n(\boldsymbol{\beta}_n)]\mathbf{b}_n| \\
&\quad \le C\|\widehat{\mathbf{R}}^{-1} - \overline{\mathbf{R}}^{-1}\| \cdot \sum_{i=1}^n\sum_{j=1}^m \sup_{\|\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}\| \le \Delta\sqrt{p_n/n}} \mathbf{A}_{ij}^{-1/2}(\boldsymbol{\beta}_n) \sup_{\|\mathbf{b}_n\|=1} \|\mathbf{X}_i\mathbf{b}_n\|^2 \\
&\quad = O_p\big(\sqrt{p_n/n}\big)O(n) = O_p\big(\sqrt{np_n}\big),
\end{aligned}
$$

by assumption (A3). (A.3) is proved similarly. □

PROOF OF LEMMA 3.4. By (3.6), it is sufficient to verify that

$$
\sup_{\|\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}\| \le \Delta\sqrt{p_n/n}} \sup_{\|\mathbf{b}_n\|=1} |\mathbf{b}_n^T\overline{\mathbf{E}}_n(\boldsymbol{\beta}_n)\mathbf{b}_n| = O_p\big(\sqrt{n}p_n\big), \tag{A.4}
$$

$$
\sup_{\|\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n0}\| \le \Delta\sqrt{p_n/n}} \sup_{\|\mathbf{b}_n\|=1} |\mathbf{b}_n^T\overline{\mathbf{G}}_n(\boldsymbol{\beta}_n)\mathbf{b}_n| = O_p\big(\sqrt{n}p_n\big). \tag{A.5}
$$

First, note that we have the following decomposition of $\overline{\mathbf{E}}_n(\boldsymbol{\beta}_n)$:

$$\begin{aligned}
\overline{\mathbf{E}}_n(\boldsymbol{\beta}_n) &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}\bigl(1-2\pi_{ij}(\boldsymbol{\beta}_{n0})\bigr)\varepsilon_{ij}(\boldsymbol{\beta}_{n0})\mathbf{X}_i^T\mathbf{A}_i^{1/2}(\boldsymbol{\beta}_{n0})\overline{\mathbf{R}}^{-1}\mathbf{e}_j\mathbf{e}_j^T\mathbf{X}_i \\
&\quad+\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}\bigl(1-2\pi_{ij}(\boldsymbol{\beta}_{n0})\bigr)\varepsilon_{ij}(\boldsymbol{\beta}_{n0})\mathbf{X}_i^T[\mathbf{A}_i^{1/2}(\boldsymbol{\beta}_n)-\mathbf{A}_i^{1/2}(\boldsymbol{\beta}_{n0})] \\
&\qquad\qquad\times\overline{\mathbf{R}}^{-1}\mathbf{e}_j\mathbf{e}_j^T\mathbf{X}_i \\
&\quad+\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}\bigl[\bigl(1-2\pi_{ij}(\boldsymbol{\beta}_n)\bigr)\mathbf{A}_{ij}^{-1/2}(\boldsymbol{\beta}_n)-\bigl(1-2\pi_{ij}(\boldsymbol{\beta}_{n0})\bigr)\mathbf{A}_{ij}^{-1/2}(\boldsymbol{\beta}_{n0})\bigr] \\
&\qquad\qquad\times\bigl(Y_{ij}-\pi_{ij}(\boldsymbol{\beta}_{n0})\bigr)\mathbf{X}_i^T\mathbf{A}_i^{1/2}(\boldsymbol{\beta}_n)\overline{\mathbf{R}}^{-1}\mathbf{e}_j\mathbf{e}_j^T\mathbf{X}_i \\
&\quad+\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}\bigl(1-2\pi_{ij}(\boldsymbol{\beta}_n)\bigr)\mathbf{A}_{ij}^{-1/2}(\boldsymbol{\beta}_n)\bigl(\pi_{ij}(\boldsymbol{\beta}_{n0})-\pi_{ij}(\boldsymbol{\beta}_n)\bigr) \\
&\qquad\qquad\times\mathbf{X}_i^T\mathbf{A}_i^{1/2}(\boldsymbol{\beta}_n)\overline{\mathbf{R}}^{-1}\mathbf{e}_j\mathbf{e}_j^T\mathbf{X}_i \\
&\triangleq\overline{\mathbf{E}}_{1n}(\boldsymbol{\beta}_{n0})+\sum_{k=2}^{4}\overline{\mathbf{E}}_{kn}(\boldsymbol{\beta}_n).
\end{aligned}$$

Thus, to prove (A.4), it suffices to verify that $\sup_{\|\mathbf{b}_n\|=1}|\mathbf{b}_n^T\overline{\mathbf{E}}_{1n}(\boldsymbol{\beta}_{n0})\mathbf{b}_n| = O_P(\sqrt{n}p_n)$ and $\sup_{\|\boldsymbol{\beta}_n-\boldsymbol{\beta}_{n0}\|\le\Delta\sqrt{p_n/n}}\ \sup_{\|\mathbf{b}_n\|=1}|\mathbf{b}_n^T\overline{\mathbf{E}}_{kn}(\boldsymbol{\beta}_n)\mathbf{b}_n| = O_P(\sqrt{n}p_n)$. We first prove that $\sup_{\|\mathbf{b}_n\|=1}|\mathbf{b}_n^T\overline{\mathbf{E}}_{1n}(\boldsymbol{\beta}_{n0})\mathbf{b}_n| = O_P(\sqrt{n}p_n)$, by verifying that $\|\overline{\mathbf{E}}_{1n}(\boldsymbol{\beta}_{n0})\| = O_P(\sqrt{n}p_n)$, where $\|\overline{\mathbf{E}}_{1n}(\boldsymbol{\beta}_{n0})\| = \sqrt{\operatorname{trace}(\overline{\mathbf{E}}_{1n}(\boldsymbol{\beta}_{n0})\overline{\mathbf{E}}_{1n}^T(\boldsymbol{\beta}_{n0}))}$:

$$\begin{aligned}
&E[\|\overline{\mathbf{E}}_{1n}(\boldsymbol{\beta}_{n0})\|^2] \\
&\quad=\frac{1}{4}\sum_{i=1}^{n}\sum_{j_1=1}^{m}\sum_{j_2=1}^{m}\bigl(1-2\pi_{ij_1}(\boldsymbol{\beta}_{n0})\bigr)\bigl(1-2\pi_{ij_2}(\boldsymbol{\beta}_{n0})\bigr)E[\varepsilon_{ij_1}(\boldsymbol{\beta}_{n0})\varepsilon_{ij_2}(\boldsymbol{\beta}_{n0})] \\
&\qquad\qquad\times\operatorname{trace}[\mathbf{X}_i^T\mathbf{A}_i^{1/2}(\boldsymbol{\beta}_{n0})\overline{\mathbf{R}}^{-1}\mathbf{e}_{j_1}\mathbf{e}_{j_1}^T\mathbf{X}_i\mathbf{X}_i^T\mathbf{e}_{j_2}\mathbf{e}_{j_2}^T \\
&\qquad\qquad\qquad\qquad\times\overline{\mathbf{R}}^{-1}\mathbf{A}_i^{1/2}(\boldsymbol{\beta}_{n0})\mathbf{X}_i] \\
&\quad\le C\sum_{i=1}^{n}\sum_{j_1=1}^{m}\sum_{j_2=1}^{m}|\mathbf{e}_{j_1}^T\mathbf{X}_i\mathbf{X}_i^T\mathbf{e}_{j_2}\mathbf{e}_{j_2}^T\overline{\mathbf{R}}^{-1}\mathbf{A}_i^{1/2}(\boldsymbol{\beta}_{n0}) \\
&\qquad\qquad\times\mathbf{X}_i\mathbf{X}_i^T\mathbf{A}_i^{1/2}(\boldsymbol{\beta}_{n0})\overline{\mathbf{R}}^{-1}\mathbf{e}_{j_1}|
\end{aligned}$$

$$\le C\sum_{i=1}^{n}\sum_{j_1=1}^{m}\sum_{j_2=1}^{m}\|\mathbf{e}_{j_1}^T\mathbf{X}_i\|\cdot\|\mathbf{X}_i^T\mathbf{e}_{j_2}\|\cdot\|\mathbf{e}_{j_2}^T\overline{\mathbf{R}}^{-1}\mathbf{A}_i^{1/2}(\boldsymbol{\beta}_{n0})\mathbf{X}_i\|$$
$$\times\|\mathbf{X}_i^T\mathbf{A}_i^{1/2}(\boldsymbol{\beta}_{n0})\overline{\mathbf{R}}^{-1}\mathbf{e}_{j_1}\|.$$

Note that $\|\mathbf{e}_{j_1}^T\mathbf{X}_i\| = \|\mathbf{X}_{ij_1}\|$, $\|\mathbf{X}_i^T\mathbf{e}_{j_2}\| = \|\mathbf{X}_{ij_2}\|$, $\|\mathbf{e}_{j_2}^T\overline{\mathbf{R}}^{-1}\mathbf{A}_i^{1/2}(\boldsymbol{\beta}_{n0})\mathbf{X}_i\| \le C(\operatorname{trace}(\mathbf{X}_i\mathbf{X}_i^T))^{1/2}$ and $\|\mathbf{X}_i^T\mathbf{A}_i^{1/2}(\boldsymbol{\beta}_{n0})\overline{\mathbf{R}}^{-1}\mathbf{e}_{j_1}\| \le C(\operatorname{trace}(\mathbf{X}_i\mathbf{X}_i^T))^{1/2}$. Thus,

$$E[\|\overline{\mathbf{E}}_{1n}(\boldsymbol{\beta}_{n0})\|^2] \le C\sum_{i=1}^{n}\sum_{j_1=1}^{m}\sum_{j_2=1}^{m}\|\mathbf{X}_{ij_1}\|\cdot\|\mathbf{X}_{ij_2}\|\operatorname{trace}(\mathbf{X}_i\mathbf{X}_i^T)$$
$$\le C\cdot\max_{i,j}\|\mathbf{X}_{ij}\|^2\operatorname{trace}\left(\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^T\right) = O(np_n^2),$$

by assumptions (A1) and (A3). This implies that $\sup_{\|\mathbf{b}_n\|=1}|\mathbf{b}_n^T\overline{\mathbf{E}}_{1n}(\boldsymbol{\beta}_{n0})\mathbf{b}_n| = O_p(\sqrt{n}\,p_n)$. Next, we have

$$\begin{aligned}
&|\mathbf{b}_n^T\overline{\mathbf{E}}_{2n}(\boldsymbol{\beta}_n)\mathbf{b}_n|\\
&\quad=\left|\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}(1-2\pi_{ij}(\boldsymbol{\beta}_{n0}))\varepsilon_{ij}^{1/2}(\boldsymbol{\beta}_{n0})\mathbf{b}_n^T\mathbf{X}_i^T[\mathbf{A}_i^{1/2}(\boldsymbol{\beta}_n)-\mathbf{A}_i^{1/2}(\boldsymbol{\beta}_{n0})]\right.\\
&\qquad\left.\times\overline{\mathbf{R}}^{-1}\mathbf{e}_j\mathbf{e}_j^T\mathbf{X}_i\mathbf{b}_n\right|\\
&\quad\le C\sum_{i=1}^{n}\sum_{j=1}^{m}|\mathbf{b}_n^T\mathbf{X}_i^T[\mathbf{A}_i^{1/2}(\boldsymbol{\beta}_n)-\mathbf{A}_i^{1/2}(\boldsymbol{\beta}_{n0})]\overline{\mathbf{R}}^{-1}\mathbf{e}_j|\cdot|\mathbf{e}_j^T\mathbf{X}_i\mathbf{b}_n|\\
&\quad\le C\sum_{i=1}^{n}\sum_{j=1}^{m}\|\mathbf{X}_i\mathbf{b}_n\|^2\lambda_{\max}(\overline{\mathbf{R}}^{-1})\max_j|\mathbf{A}_{ij}^{1/2}(\boldsymbol{\beta}_n)-\mathbf{A}_{ij}^{1/2}(\boldsymbol{\beta}_{n0})|.
\end{aligned}$$

Note that there exists some $\boldsymbol{\beta}_n^*$ between $\boldsymbol{\beta}_n$ and $\boldsymbol{\beta}_{n0}$ such that

$$\begin{aligned}
\mathbf{A}_{ij}^{1/2}(\boldsymbol{\beta}_n)-\mathbf{A}_{ij}^{1/2}(\boldsymbol{\beta}_{n0}) &= \tfrac{1}{2}\mathbf{A}_{ij}^{1/2}(\boldsymbol{\beta}_n^*)(1-2\pi_{ij}(\boldsymbol{\beta}_n^*))\mathbf{X}_{ij}^T(\boldsymbol{\beta}_n-\boldsymbol{\beta}_{n0})\\
&\le C\|\mathbf{X}_{ij}\|\cdot\|\boldsymbol{\beta}_n-\boldsymbol{\beta}_{n0}\|.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\sup_{\|\boldsymbol{\beta}_n-\boldsymbol{\beta}_{n0}\|\le\Delta\sqrt{p_n/n}}\ \sup_{\|\mathbf{b}_n\|=1}|\mathbf{b}_n^T\overline{\mathbf{E}}_{2n}(\boldsymbol{\beta}_n)\mathbf{b}_n|\\
&\quad\le C\max_{i,j}\|\mathbf{X}_{ij}\|\sup_{\|\boldsymbol{\beta}_n-\boldsymbol{\beta}_{n0}\|\le\Delta\sqrt{p_n/n}}\|\boldsymbol{\beta}_n-\boldsymbol{\beta}_{n0}\|\cdot\lambda_{\max}\left(\sum_{i=1}^{n}\mathbf{X}_i^T\mathbf{X}_i\right)\\
&\quad= O\big(\sqrt{p_n}\big)O\big(\sqrt{p_n/n}\big)O(n) = O\big(\sqrt{n}\,p_n\big).
\end{aligned}$$

Similarly, we can show that $\sup_{\|\boldsymbol{\beta}_n-\boldsymbol{\beta}_{n0}\|\le\Delta\sqrt{p_n/n}}\sup_{\|\mathbf{b}_n\|=1}|\mathbf{b}_n^T\overline{\mathbf{E}}_{kn}(\boldsymbol{\beta}_n)\mathbf{b}_n| = O(\sqrt{n}p_n)$, $k=3,4$. This proves (A.4). Similarly, we can prove (A.5). $\square$

PROOF OF LEMMA 3.5. The proof is given in the online supplementary material. $\square$

PROOF OF LEMMA 3.7. We write $\boldsymbol{\alpha}_n^T\overline{\mathbf{M}}_n^{-1/2}(\boldsymbol{\beta}_{n0})\overline{\mathbf{S}}_n(\boldsymbol{\beta}_{n0}) = \sum_{i=1}^n Z_{ni}$, where $Z_{ni} = \boldsymbol{\alpha}_n^T\overline{\mathbf{M}}_n^{-1/2}(\boldsymbol{\beta}_{n0})\mathbf{X}_i^T\mathbf{A}_i^{1/2}(\boldsymbol{\beta}_{n0})\overline{\mathbf{R}}^{-1}\boldsymbol{\varepsilon}_i(\boldsymbol{\beta}_{n0})$. Since $\overline{\mathbf{M}}_n(\boldsymbol{\beta}_{n0}) = \mathrm{Cov}(\overline{\mathbf{S}}_n(\boldsymbol{\beta}_{n0}))$, we have $\mathrm{Var}(\boldsymbol{\alpha}_n^T\overline{\mathbf{M}}_n^{-1/2}(\boldsymbol{\beta}_{n0})\overline{\mathbf{S}}_n(\boldsymbol{\beta}_{n0})) = 1$. To establish the asymptotic normality, it suffices to check the Lindberg condition, that is, $\forall\varepsilon>0$, $\sum_{i=1}^n E[Z_{ni}^2 I(|Z_{ni}| > \varepsilon)] \to 0$. By the Cauchy–Schwarz inequality,

$$\begin{aligned}
Z_{ni}^2 &\le \|\boldsymbol{\alpha}_n^T\overline{\mathbf{M}}_n^{-1/2}(\boldsymbol{\beta}_{n0})\mathbf{X}_i^T\mathbf{A}_i^{1/2}(\boldsymbol{\beta}_{n0})\overline{\mathbf{R}}^{-1}\|^2\cdot\|\boldsymbol{\varepsilon}_i(\boldsymbol{\beta}_{n0})\|^2\\
&\le \lambda_{\max}(\overline{\mathbf{R}}^{-2})\lambda_{\max}(\mathbf{A}_i(\boldsymbol{\beta}_{n0}))(\boldsymbol{\alpha}_n^T\overline{\mathbf{M}}_n^{-1/2}(\boldsymbol{\beta}_{n0})\mathbf{X}_i^T\mathbf{X}_i\overline{\mathbf{M}}_n^{-1/2}(\boldsymbol{\beta}_{n0})\boldsymbol{\alpha}_n)\\
&\quad\times\|\boldsymbol{\varepsilon}_i(\boldsymbol{\beta}_{n0})\|^2\\
&\le C\gamma_{ni}\|\boldsymbol{\varepsilon}_i(\boldsymbol{\beta}_{n0})\|^2,
\end{aligned}$$

where $\gamma_{ni} \triangleq \boldsymbol{\alpha}_n^T\overline{\mathbf{M}}_n^{-1/2}(\boldsymbol{\beta}_{n0})\mathbf{X}_i^T\mathbf{X}_i\overline{\mathbf{M}}_n^{-1/2}(\boldsymbol{\beta}_{n0})\boldsymbol{\alpha}_n$. Next, we will show that $\max_{1\le i\le n}\gamma_{ni}\to 0$ as $n\to\infty$. Note that $\gamma_{ni}\le\lambda_{\max}(\mathbf{X}_i^T\mathbf{X}_i)\lambda_{\min}^{-1}(\overline{\mathbf{M}}_n(\boldsymbol{\beta}_{n0}))$. Since $\overline{\mathbf{M}}_n(\boldsymbol{\beta}_{n0})$ is symmetric, to evaluate $\lambda_{\min}(\overline{\mathbf{M}}_n(\boldsymbol{\beta}_{n0}))$, $\forall\mathbf{b}_n\in R^{p_n}$, we have

$$\begin{aligned}
\mathbf{b}_n^T\overline{\mathbf{M}}_n(\boldsymbol{\beta}_{n0})\mathbf{b}_n &\ge \lambda_{\min}(\mathbf{R}_0)\lambda_{\min}(\overline{\mathbf{R}}^{-2})\sum_{i=1}^n\lambda_{\min}(\mathbf{A}_i(\boldsymbol{\beta}_{n0}))\mathbf{b}_n^T\mathbf{X}_i^T\mathbf{X}_i\mathbf{b}_n\\
&\ge C\mathbf{b}_n^T\left(\sum_{i=1}^n\mathbf{X}_i^T\mathbf{X}_i\right)\mathbf{b}_n \ge C\lambda_{\min}\left(\sum_{i=1}^n\mathbf{X}_i^T\mathbf{X}_i\right)\|b_n\|^2.
\end{aligned}$$

Thus, $\inf_{\|\mathbf{b}_n\|=1}|\mathbf{b}_n^T\overline{\mathbf{M}}_n(\boldsymbol{\beta}_{n0})\mathbf{b}_n| \ge C\lambda_{\min}(\sum_{i=1}^n\mathbf{X}_i^T\mathbf{X}_i)$ and this implies that $\lambda_{\min}\times(\overline{\mathbf{M}}_n(\boldsymbol{\beta}_{n0})) \ge C\lambda_{\min}(\sum_{i=1}^n\mathbf{X}_i^T\mathbf{X}_i)$. Therefore, we have

$$\gamma_{ni}\le\frac{\lambda_{\max}(\mathbf{X}_i^T\mathbf{X}_i)}{C\lambda_{\min}(\sum_{i=1}^n\mathbf{X}_i^T\mathbf{X}_i)}\le\frac{\mathrm{Tr}(\mathbf{X}_i^T\mathbf{X}_i)}{C\lambda_{\min}(\sum_{i=1}^n\mathbf{X}_i^T\mathbf{X}_i)} = \frac{\sum_{j=1}^m\mathbf{X}_{ij}^T\mathbf{X}_{ij}}{C\lambda_{\min}(\sum_{i=1}^n\mathbf{X}_i^T\mathbf{X}_i)}.$$

It follows that $\max_{1\le i\le n}\gamma_{ni}\le O(n^{-1}p_n) = o(1)$. We have

$$\sum_{i=1}^n E[Z_{ni}^2 I(|Z_{ni}|>\varepsilon)] \le \sum_{i=1}^n C\gamma_{ni}E\left[\|\boldsymbol{\varepsilon}_i(\boldsymbol{\beta}_{n0})\|^2 I\left\{\|\boldsymbol{\varepsilon}_i(\boldsymbol{\beta}_{n0})\|^2 > \frac{\varepsilon^2}{C\gamma_{ni}}\right\}\right].$$

Note that $\|\boldsymbol{\varepsilon}_i(\boldsymbol{\beta}_{n0})\|^2$ is uniformly bounded, by assumption (A2). Thus, for all $\varepsilon>0$ and $\delta>0$, there exists a positive integer N such that (1) $I\{\|\boldsymbol{\varepsilon}_i(\boldsymbol{\beta}_{n0})\|^2 >$

$\frac{\varepsilon^2}{C\gamma_{ni}}\} = 0$ for all $n > N$; (2) $\sum_{i=1}^{N} C\gamma_{ni} \le \delta$ for all n sufficiently large. This ensures that

$$\sum_{i=1}^{n} C\gamma_{ni} E\left[\|\boldsymbol{\varepsilon}_i(\boldsymbol{\beta}_{n0})\|^2 I\left\{\|\boldsymbol{\varepsilon}_i(\boldsymbol{\beta}_{n0})\|^2 > \frac{\varepsilon^2}{C\gamma_{ni}}\right\}\right] \to 0.$$

Therefore, the Lindberg condition is verified. □

PROOF OF THEOREM 3.10. It is sufficient to show that for $\mathbf{b}_n \in R^{p_n}$,

$$\sup_{\|\mathbf{b}_n\|=1} |\mathbf{b}_n^T(\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}_n)\mathbf{b}_n| = o_p(n^{-1}). \tag{A.6}$$

We use the conclusion of Theorem 3.6 throughout the proof. Note that we can write $\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}_n = I_{n1} + I_{n2} + I_{n3}$, where

$$\begin{aligned}
I_{n1} &= \mathbf{H}_n^{-1}(\widehat{\boldsymbol{\beta}}_n)[\widehat{\mathbf{M}}_n(\widehat{\boldsymbol{\beta}}_n) - \overline{\mathbf{M}}_n(\boldsymbol{\beta}_{n0})]\mathbf{H}_n^{-1}(\widehat{\boldsymbol{\beta}}_n),\\
I_{n2} &= [\mathbf{H}_n^{-1}(\widehat{\boldsymbol{\beta}}_n) - \overline{\mathbf{H}}_n^{-1}(\boldsymbol{\beta}_{n0})]\overline{\mathbf{M}}_n(\boldsymbol{\beta}_{n0})\mathbf{H}_n^{-1}(\widehat{\boldsymbol{\beta}}_n),\\
I_{n3} &= \overline{\mathbf{H}}_n^{-1}(\boldsymbol{\beta}_{n0})\overline{\mathbf{M}}_n(\boldsymbol{\beta}_{n0})[\mathbf{H}_n^{-1}(\widehat{\boldsymbol{\beta}}_n) - \overline{\mathbf{H}}_n^{-1}(\boldsymbol{\beta}_{n0})].
\end{aligned}$$

Thus, (A.6) is implied by $\sup_{\|\mathbf{b}_n\|=1} |\mathbf{b}_n^T I_{ni}\mathbf{b}_n| = o_p(1)$. We have

$$\begin{aligned}
&\sup_{\|\mathbf{b}_n\|=1} |\mathbf{b}_n^T I_{n1}\mathbf{b}_n|\\
&\quad\le \frac{\max(|\lambda_{\max}(\widehat{\mathbf{M}}_n(\widehat{\boldsymbol{\beta}}_n) - \overline{\mathbf{M}}_n(\boldsymbol{\beta}_{n0}))|, |\lambda_{\min}(\widehat{\mathbf{M}}_n(\widehat{\boldsymbol{\beta}}_n) - \overline{\mathbf{M}}_n(\boldsymbol{\beta}_{n0}))|)}{\lambda_{\min}^2(\mathbf{H}_n(\widehat{\boldsymbol{\beta}}_n))}.
\end{aligned}$$

To evaluate the eigenvalues of $\widehat{\mathbf{M}}_n(\widehat{\boldsymbol{\beta}}_n) - \overline{\mathbf{M}}_n(\boldsymbol{\beta}_{n0})$, we have

$$\begin{aligned}
&|\mathbf{c}_n^T[\widehat{\mathbf{M}}_n(\widehat{\boldsymbol{\beta}}_n) - \overline{\mathbf{M}}_n(\boldsymbol{\beta}_{n0})]\mathbf{c}_n|\\
&\quad\le |\mathbf{c}_n^T[\widehat{\mathbf{M}}_n(\widehat{\boldsymbol{\beta}}_n) - \widehat{\mathbf{M}}_n(\boldsymbol{\beta}_{n0})]\mathbf{c}_n| + |\mathbf{c}_n^T[\widehat{\mathbf{M}}_n(\boldsymbol{\beta}_{n0}) - \overline{\mathbf{M}}_n(\boldsymbol{\beta}_{n0})]\mathbf{c}_n|
\end{aligned}$$

for $\mathbf{c}_n \in R^{p_n}$. Note that

$$\begin{aligned}
&\sup_{\|\mathbf{c}_n\|=1} |\mathbf{c}_n^T[\widehat{\mathbf{M}}_n(\widehat{\boldsymbol{\beta}}_n) - \widehat{\mathbf{M}}_n(\boldsymbol{\beta}_{n0})]\mathbf{c}_n|\\
&\quad\le \sup_{\|\mathbf{c}_n\|=1}\left|\sum_{i=1}^{n} \mathbf{c}_n^T\mathbf{X}_i^T[\mathbf{A}_i^{1/2}(\widehat{\boldsymbol{\beta}}_n) - \mathbf{A}_i^{1/2}(\boldsymbol{\beta}_{n0})]\widehat{\mathbf{R}}^{-1}\boldsymbol{\varepsilon}_i(\widehat{\boldsymbol{\beta}}_n)\boldsymbol{\varepsilon}_i^T(\widehat{\boldsymbol{\beta}}_n)\right.\\
&\qquad\qquad\left.\times \widehat{\mathbf{R}}^{-1}\mathbf{A}_i^{1/2}(\widehat{\boldsymbol{\beta}}_n)\mathbf{X}_i\mathbf{c}_n\right|\\
&\qquad + \sup_{\|\mathbf{c}_n\|=1}\left|\sum_{i=1}^{n} \mathbf{c}_n^T\mathbf{X}_i^T\mathbf{A}_i^{1/2}(\boldsymbol{\beta}_{n0})\widehat{\mathbf{R}}^{-1}\boldsymbol{\varepsilon}_i(\widehat{\boldsymbol{\beta}}_n)\boldsymbol{\varepsilon}_i^T(\widehat{\boldsymbol{\beta}}_n)\widehat{\mathbf{R}}^{-1}\right.
\end{aligned}$$

$$\times [\mathbf{A}_i^{1/2}(\widehat{\boldsymbol{\beta}}_n) - \mathbf{A}_i^{1/2}(\boldsymbol{\beta}_{n0})]\mathbf{X}_i\mathbf{c}_n\Bigg|$$

$$+ \sup_{\|\mathbf{c}_n\|=1}\Bigg|\sum_{i=1}^{n}\mathbf{c}_n^T\mathbf{X}_i^T\mathbf{A}_i^{1/2}(\boldsymbol{\beta}_{n0})\widehat{\mathbf{R}}^{-1}[\boldsymbol{\varepsilon}_i(\widehat{\boldsymbol{\beta}}_n)\boldsymbol{\varepsilon}_i^T(\widehat{\boldsymbol{\beta}}_n) - \boldsymbol{\varepsilon}_i(\boldsymbol{\beta}_{n0})\boldsymbol{\varepsilon}_i^T(\boldsymbol{\beta}_{n0})]$$

$$\times \widehat{\mathbf{R}}^{-1}\mathbf{A}_i^{1/2}(\boldsymbol{\beta}_{n0})\mathbf{X}_i\mathbf{c}_n\Bigg|$$

$$\triangleq \sup_{\|\mathbf{c}_n\|=1} J_{n1} + \sup_{\|\mathbf{c}_n\|=1} J_{n2} + \sup_{\|\mathbf{c}_n\|=1} J_{n3}.$$

Note that

$$J_{n1} \le \sum_{i=1}^{n}\|\mathbf{c}_n^T\mathbf{X}_i^T[\mathbf{A}_i^{1/2}(\widehat{\boldsymbol{\beta}}_n) - \mathbf{A}_i^{1/2}(\boldsymbol{\beta}_{n0})]\| \cdot \|\widehat{\mathbf{R}}^{-1}\boldsymbol{\varepsilon}_i(\widehat{\boldsymbol{\beta}}_n)\|^2 \cdot \|\mathbf{A}_i^{1/2}(\widehat{\boldsymbol{\beta}}_n)\mathbf{X}_i\mathbf{c}_n\|.$$

We have $\|\mathbf{A}_i^{1/2}(\widehat{\boldsymbol{\beta}}_n)\mathbf{X}_i\mathbf{c}_n\| \le \|\mathbf{X}_i\mathbf{c}_n\|$ and

$$\begin{aligned}\|\mathbf{c}_n^T\mathbf{X}_i^T[\mathbf{A}_i^{1/2}(\widehat{\boldsymbol{\beta}}_n) - \mathbf{A}_i^{1/2}(\boldsymbol{\beta}_{n0})]\| &\le \|\mathbf{X}_i\mathbf{c}_n\| \max_j |\mathbf{A}_{ij}^{1/2}(\widehat{\boldsymbol{\beta}}_n) - \mathbf{A}_{ij}^{1/2}(\boldsymbol{\beta}_{n0})| \\ &\le C\|\mathbf{X}_i\mathbf{c}_n\| \cdot \|\mathbf{X}_{ij}\| \cdot \|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}\|.\end{aligned}$$

Furthermore,

$$\begin{aligned}\|\widehat{\mathbf{R}}^{-1}\boldsymbol{\varepsilon}_i(\widehat{\boldsymbol{\beta}}_n)\|^2 &= \big(\mathbf{Y}_i - \boldsymbol{\pi}_i(\widehat{\boldsymbol{\beta}}_n)\big)^T\mathbf{A}_i^{-1/2}(\widehat{\boldsymbol{\beta}}_n)\widehat{\mathbf{R}}^{-2}\mathbf{A}_i^{-1/2}(\widehat{\boldsymbol{\beta}}_n)\big(\mathbf{Y}_i - \boldsymbol{\pi}_i(\widehat{\boldsymbol{\beta}}_n)\big) \\ &\le \lambda_{\max}(\widehat{\mathbf{R}}^{-2})\lambda_{\max}(\mathbf{A}_i^{-1}(\widehat{\boldsymbol{\beta}}_n))\|\mathbf{Y}_i - \boldsymbol{\pi}_i(\widehat{\boldsymbol{\beta}}_n)\|^2 \le C\,O_p(1).\end{aligned}$$

Thus,

$$\sup_{\|\mathbf{c}_n\|=1} J_{n1} \le O_p(1)\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}\| \max_{i,j}\|\mathbf{X}_{ij}\|\lambda_{\max}\Bigg(\sum_{i=1}^{n}\mathbf{X}_i^T\mathbf{X}_i\Bigg) = o_p(n).$$

Similarly, $\sup_{\|\mathbf{c}_n\|=1} J_{n2} = o_p(n)$ and $\sup_{\|\mathbf{c}_n\|=1} J_{n3} = o_p(n)$. Thus,

$$\sup_{\|\mathbf{c}_n\|=1}|\mathbf{c}_n^T[\widehat{\mathbf{M}}_n(\widehat{\boldsymbol{\beta}}_n) - \widehat{\mathbf{M}}_n(\boldsymbol{\beta}_{n0})]\mathbf{c}_n| = o_p(n).$$

Similarly, $\sup_{\|\mathbf{c}_n\|=1}|\mathbf{c}_n^T[\widehat{\mathbf{M}}_n(\boldsymbol{\beta}_{n0}) - \overline{\mathbf{M}}_n(\boldsymbol{\beta}_{n0})]\mathbf{c}_n| = o_p(n)$. Finally, note that

$$\begin{aligned}\lambda_{\min}(\mathbf{H}_n(\widehat{\boldsymbol{\beta}}_n)) &\ge \lambda_{\min}(\widehat{\mathbf{R}})\min_{i,j}\big(\pi_{ij}(\widehat{\boldsymbol{\beta}}_n)\big(1 - \pi_{ij}(\widehat{\boldsymbol{\beta}}_n)\big)\big)\lambda_{\min}\Bigg(\sum_{i=1}^{n}\mathbf{X}_i^T\mathbf{X}_i\Bigg) \\ &= O_p(n).\end{aligned}$$

Thus, $\sup_{\|\mathbf{b}_n\|=1}|\mathbf{b}_n^T I_{n1}\mathbf{b}_n| = o_p(n^{-1})$. We can also prove that $\sup_{\|\mathbf{b}_n\|=1}|\mathbf{b}_n^T I_{ni} \times \mathbf{b}_n| = o_p(n^{-1})$, $i = 2, 3$, by first noting that

$$\mathbf{H}_n^{-1}(\widehat{\boldsymbol{\beta}}_n) - \overline{\mathbf{H}}_n^{-1}(\boldsymbol{\beta}_{n0}) = [\mathbf{H}_n^{-1}(\widehat{\boldsymbol{\beta}}_n) - \overline{\mathbf{H}}_n^{-1}(\widehat{\boldsymbol{\beta}}_n)] + [\overline{\mathbf{H}}_n^{-1}(\widehat{\boldsymbol{\beta}}_n) - \overline{\mathbf{H}}_n^{-1}(\boldsymbol{\beta}_{n0})]$$

and then using the expressions

$$\mathbf{H}_n^{-1}(\widehat{\boldsymbol{\beta}}_n) - \overline{\mathbf{H}}_n^{-1}(\widehat{\boldsymbol{\beta}}_n) = \overline{\mathbf{H}}_n^{-1}(\widehat{\boldsymbol{\beta}}_n)[\overline{\mathbf{H}}_n(\widehat{\boldsymbol{\beta}}_n) - \mathbf{H}_n(\widehat{\boldsymbol{\beta}}_n)]\mathbf{H}_n^{-1}(\widehat{\boldsymbol{\beta}}_n),$$
$$\overline{\mathbf{H}}_n^{-1}(\widehat{\boldsymbol{\beta}}_n) - \overline{\mathbf{H}}_n^{-1}(\boldsymbol{\beta}_{n0}) = \overline{\mathbf{H}}_n^{-1}(\boldsymbol{\beta}_{n0})[\overline{\mathbf{H}}_n(\boldsymbol{\beta}_{n0}) - \overline{\mathbf{H}}_n(\widehat{\boldsymbol{\beta}}_n)]\overline{\mathbf{H}}_n^{-1}(\widehat{\boldsymbol{\beta}}_n). \qquad \square$$

PROOF OF COROLLARY 3.11. It is sufficient to show that

$$[(\mathbf{L}_n\widehat{\boldsymbol{\Sigma}}_n\mathbf{L}_n^T)^{-1/2} - (\mathbf{L}_n\boldsymbol{\Sigma}_n\mathbf{L}_n^T)^{-1/2}]\mathbf{L}_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0}) \to 0 \tag{A.7}$$

in probability. Note that the left-hand side can be written as

$$[(\mathbf{L}_n\widehat{\boldsymbol{\Sigma}}_n\mathbf{L}_n^T)^{-1/2}(\mathbf{L}_n\boldsymbol{\Sigma}_n\mathbf{L}_n^T)^{1/2} - \mathbf{I}_l](\mathbf{L}_n\boldsymbol{\Sigma}_n\mathbf{L}_n^T)^{-1/2}\mathbf{L}_n(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_{n0})$$

and thus (A.7) is implied by

$$(\mathbf{L}_n\widehat{\boldsymbol{\Sigma}}_n\mathbf{L}_n^T)^{-1}(\mathbf{L}_n\boldsymbol{\Sigma}_n\mathbf{L}_n^T) - \mathbf{I}_l = (\mathbf{L}_n\widehat{\boldsymbol{\Sigma}}_n\mathbf{L}_n^T)^{-1}\mathbf{L}_n(\boldsymbol{\Sigma}_n - \widehat{\boldsymbol{\Sigma}}_n)\mathbf{L}_n^T = o_p(1).$$

Let $\mathbf{u}_i$ denote the $l \times 1$ unit vector with the ith element being 1 and all of the other elements being 0. Then, for all $1 \le i, j \le l$, we have, by the Cauchy–Schwarz inequality,

$$\begin{aligned}
&|\mathbf{u}_i^T(\mathbf{L}_n\widehat{\boldsymbol{\Sigma}}_n\mathbf{L}_n^T)^{-1}\mathbf{L}_n(\boldsymbol{\Sigma}_n - \widehat{\boldsymbol{\Sigma}}_n)\mathbf{L}_n^T\mathbf{u}_j| \\
&\quad \le |\mathbf{u}_i^T(\mathbf{L}_n\widehat{\boldsymbol{\Sigma}}_n\mathbf{L}_n^T)^{-2}\mathbf{u}_i|^{1/2}|\mathbf{u}_j^T[\mathbf{L}_n(\boldsymbol{\Sigma}_n - \widehat{\boldsymbol{\Sigma}}_n)\mathbf{L}_n^T]^2\mathbf{u}_j|^{1/2} \\
&\quad \le \frac{\|\mathbf{L}_n(\boldsymbol{\Sigma}_n - \widehat{\boldsymbol{\Sigma}}_n)\mathbf{L}_n^T\|}{\lambda_{\min}(\mathbf{L}_n\widehat{\boldsymbol{\Sigma}}_n\mathbf{L}_n^T)}.
\end{aligned}$$

Now, for any l-dimensional vector such that $\|\mathbf{b}\| = 1$, we have

$$\begin{aligned}
|\mathbf{b}^T\mathbf{L}_n\widehat{\boldsymbol{\Sigma}}_n\mathbf{L}_n^T\mathbf{b}| &\ge |\mathbf{b}^T\mathbf{L}_n\boldsymbol{\Sigma}_n\mathbf{L}_n^T\mathbf{b}| - |\mathbf{b}^T\mathbf{L}_n(\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}_n)\mathbf{L}_n^T\mathbf{b}| \\
&\ge \lambda_{\min}(\boldsymbol{\Sigma}_n) + o_p(n^{-1}) \\
&\ge \frac{\lambda_{\min}(\overline{\mathbf{M}}_n(\boldsymbol{\beta}_{n0}))}{\lambda_{\max}^2(\overline{\mathbf{H}}_n(\boldsymbol{\beta}_{n0}))} + o_p(n^{-1}),
\end{aligned}$$

where the second inequality uses Theorem 3.10. By (3.9), $\lambda_{\min}(\overline{\mathbf{M}}_n(\boldsymbol{\beta}_{n0})) \ge c_1\lambda_{\min}(\sum_{i=1}^n \mathbf{X}_i^T\mathbf{X}_i)$ for some positive constant c_1. Similarly, we can show that $\lambda_{\max}(\overline{\mathbf{H}}_n(\boldsymbol{\beta}_{n0})) \le c_2\lambda_{\max}(\sum_{i=1}^n \mathbf{X}_i^T\mathbf{X}_i)$ for some positive constant c_2. Thus, $\lambda_{\min}(\mathbf{L}_n\widehat{\boldsymbol{\Sigma}}_n\mathbf{L}_n^T) \ge O_p(n^{-1})$. This proves that $\frac{\|\mathbf{L}_n(\boldsymbol{\Sigma}_n - \widehat{\boldsymbol{\Sigma}}_n)\mathbf{L}_n^T\|}{\lambda_{\min}(\mathbf{L}_n\widehat{\boldsymbol{\Sigma}}_n\mathbf{L}_n^T)} = o_p(1)$, by Theorem 3.10. $\square$

Acknowledgments. The author would like to thank the Associate Editor and two referees for their constructive and insightful comments that significantly improved this paper.

SUPPLEMENTARY MATERIAL

Supplement to "GEE analysis of clustered binary data with diverging number of covariates" (DOI: 10.1214/10-AOS846SUPP; .pdf). The proofs of (3.3), Lemma 3.5, (3.11) and Theorem 5.1 are provided in this supplementary article [Wang (2010)].

REFERENCES

BALAN, R. M. and SCHIOPU-KRATINA, I. (2005). Asymptotic results with generalized estimating equations for longitudinal data. *Ann. Statist.* **32** 522–541. MR2163150

BAI, Z. and WU, Y. (1994). Limiting behavior of M-estimators of regression coefficients in high dimensional linear models, I. Scale-dependent case. *J. Multivariate Anal.* **51** 211–239. MR1321295

CHAGANTY, N. R. and JOE, H. (2004). Efficiency of generalized estimating equations for binary responses. *J. Roy. Statist. Soc. Ser. B* **66** 851–860. MR2102468

CHEN, K. and JIN, Z. (2006). Partial linear regression models for clustered data. *J. Amer. Statist. Assoc.* **101** 195–204. MR2268038

CHEN, S. X., PENG, L. and QIN, Y-L. (2009). Effects of data dimension on empirical likelihood. *Biometrika* **96** 711–722. MR2538767

CHIOU, J.-M. and MÜLLER, H.-G. (2005). Estimated estimating equations: Semiparametric inference for clustered and longitudinal data. *J. Roy. Statist. Soc. Ser. B* **67** 531–553. MR2168203

DONOHO, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. In *Math Challenges of 21st Century* 1–32. Amer. Math. Soc., Providence, RI.

FAN, J. and LI, R. (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *J. Amer. Statist. Assoc.* **99** 710–723. MR2090905

FAN, J. and LI, R. (2006). Statistical challenges with high dimensionality: Feature selection in knowledge discovery. In *Proceedings of the International Congress of Mathematicians* (M. Sanz-Sole, J. Soria, J. L. Varona and J. Verdera eds.) **III** 595–622. Eur. Math. Soc., Zürich. MR2275698

FAN, J. and LV, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statist. Sinica* **20** 101–148. MR2640659

FAN, J. and PENG, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32** 928–961. MR2065194

FITZMAURICE, G. M. (1995). A caveat concerning independence estimating equations with multivariate binary data. *Biometrics* **51** 309–317.

HE, X., FUNG, W. K. and ZHU, Z. Y. (2006). Robust estimation in generalized partial linear models for clustered data. *J. Amer. Statist. Assoc.* **100** 1176–1184. MR2236433

HE, X. and SHAO, Q. M. (2000). On parameters of increasing dimensions. *J. Multivariate Anal.* **73** 120–135. MR1766124

HE, X., ZHU, Z. Y. and FUNG, W. K. (2002). Estimation in a semiparametric model for longitudinal data with unspecified dependence structure. *Biometrika* **89** 579–590. MR1929164

HJORT, H. L., MCKEAGUE, I. W. and VAN KEILEGOM, I. (2009). Extending the scope of empirical likelihood. *Ann. Statist.* **37** 1079–1111. MR2509068

HUANG, J., HOROWITZ, J. L. and MA, S. J. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist.* **36** 587–613. MR2396808

HUANG, J. Z., ZHANG, L. and ZHOU, L. (2007). Efficient estimation in marginal partially linear models for longitudinal/clustered data using splines. *Scand. J. Statist.* **34** 451–477. MR2368793

HUBER, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Ann. Statist.* **1** 799–821. MR0356373

LAM, C. and FAN, J. (2008). Profile-kernel likelihood inference with diverging number of parameters. *Ann. Statist.* **36** 2232–2260. MR2458186

LI, B. (1997). On the consistency of generalized estimating equations. In *Selected Proceedings of the Symposium on Estimating Functions* 115–136. *IMS Lecture Notes—Monograph Series* **32**. IMS, Hayward, CA. MR1837801

LIANG, K. Y. and ZEGER, S. L. (1986). Longitudinal data analysis using generalised linear models. *Biometrika* **73** 12–22. MR0836430

LIN, D. Y. and YING, Z. (2001). Semiparametric and nonparametric regression analysis of longitudinal data. *J. Amer. Statist. Assoc.* **96** 103–112. MR1952726

LIN, X. and CARROLL, R. J. (2001a). Semiparametric regression for clustered data using generalized estimating equations. *J. Amer. Statist. Assoc.* **96** 1045–1056. MR1947252

LIN, X. and CARROLL, R. J. (2001b). Semiparametric regression for clustered data. *Biometrika* **88** 1179–1185. MR1872228

MAMMEN, E. (1989). Asymptotics with increasing dimension for robust regression with applications to the bootstrap. *Ann. Statist.* **17** 382–400. MR0981457

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman and Hall, London. MR0727836

ORTEGA, J. M. and RHEINBOLDT, W. C. (1970). *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, San Diego. MR0273810

PAN, W. (2002). Goodness-of-fit tests for GEE with correlated binary data. *Scand. J. Statist.* **29** 101–110. MR1894384

PORTNOY, S. (1984). Asymptotic behavior of M estimators of p regression parameters when p^2/n is large. I. Consistency. *Ann. Statist.* **12** 1298–1309. MR0760690

PORTNOY, S. (1985). Asymptotic behavior of M estimators of p regression parameters when p^2/n is large. II. Normal approximation. *Ann. Statist.* **13** 1403–1417. MR0811499

PORTNOY, S. (1988). Asymptotic properties of likelihood methods for exponential families when the number of parameters tends to infinity. *Ann. Statist.* **16** 356–366. MR0924876

XIE, M. and YANG, Y. (2003). Asymptotics for generalized estimating equations with large cluster sizes. *Ann. Statist.* **31** 310–347. MR1962509

WANG, J. L., XUE, L. G., ZHU, L. X. and CHONG, Y. S. (2010). Estimation for a partial-linear single-index model. *Ann. Statist.* **38** 246–274. MR2589322

WANG, L. (2010). Supplement to "GEE analysis of clustered binary data with diverging number of covariates." DOI: 10.1214/10-AOS846SUPP.

WANG, N., CARROLL, R. J. and LIN, X. (2005). Efficient semiparametric marginal estimation for longitudinal/clustered data. *J. Amer. Statist. Assoc.* **100** 147–157. MR2156825

WELSH, A. H. (1989). On M-processes and M-estimation. *Ann. Statist.* **17** 337–361. MR0981455

ZHU, L. P. and ZHU, L. X. (2009). On distribution-weighted partial least squares with diverging number of highly correlated predictors. *J. Roy. Statist. Soc. Ser. B* **71** 525–548.

ZOU, H. and ZHANG, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Ann. Statist.* **37** 1733–1751. MR2533470

SCHOOL OF STATISTICS
UNIVERSITY OF MINNESOTA
224 CHURCH STREET, SE
MINNEAPOLIS, MINNESOTA 55455
USA
E-MAIL: lan@stat.umn.edu