UW Biostatistics Working Paper Series

10-26-2006

Large Cluster Asymptotics for GEE: Working Correlation Models

Hyoju Chung
University of Washington, hyojuch@u.washington.edu

Thomas Lumley
University of Washington, tlumley@u.washington.edu

1 Introduction

In this paper, we consider a specific type of correlated data that commonly arise: multiple measurements on each of a large number of independent units. Longitudinal data, multivariate response data and clustered data are of this type. Analysis of correlated data has been a challenge in statistical modeling because the assumption of independence of observations is violated and because multivariate parametric models for correlated data are limited. However, when a marginal mean regression parameter is of primary interest and dependence structure is a nuisance, the generalized estimating equations (GEE) model of Liang and Zeger [4] has been widely used. The GEE model is a semiparametric moment-based estimating equations method. In the GEE model, a mean-zero estimating equation is constructed for each measurement under the first- and second- moment assumptions, and correlated estimating equations are combined through a "working" correlation matrix. The working correlation matrix is simply a weight matrix which takes the inter-correlated data feature into account when combining mean-zero estimating equations, and it may contain a nuisance parameter.

Let $Y_i^t = (Y_{i1}, \cdots, Y_{im_i})$ be a vector of m_i response measurements on the i-th cluster, for $i = 1, \cdots, n$. Suppose that each response measurement Y_{ij} has a corresponding $(p \times 1)$ covariate vector X_{ij} and that $X_i = (X_{i1}, \cdots, X_{im_i})$ represent a $(p \times m_i)$ matrix of covariates for the i-th cluster. The conditional mean of Y_{ij} given X_i, denoted by μ_{ij}, is assumed to have the form:

$$\mu_{ij} = \mathrm{E}(Y_{ij}|X_i) = g(X_{ij}^t \boldsymbol{\beta}),$$

where $g(\cdot)$ is a known link function and $\boldsymbol{\beta}$ is a marginal regression parameter of p dimension. Let $\mu_i = (\mu_{i1}, \cdots, \mu_{im_i})$. The parameter of interest $\boldsymbol{\beta}_0$ is defined by

$$\mathrm{E}((Y_i - \mu_i(\boldsymbol{\beta}_0))|X_i) = 0, \forall i.$$

The mean-variance relationship is assumed to be known up to a constant:

$$\mathrm{Var}(Y_{ij}) = \phi v(\mu_{ij}),$$

where $v(\cdot)$ is a known function and ϕ is an unknown dispersion parameter. Let $A_i(\boldsymbol{\beta}) = \mathrm{diag}(v(\mu_{i1}), \cdots, v(\mu_{im_i}))$ be the diagonal matrix of marginal variances for i-th cluster. The true correlation structure and the probability distribution are not specified in the GEE model. Liang ang Zeger [4] defined the GEE estimator $\hat{\boldsymbol{\beta}}_n$ as a solution of the following estimating equations:

$$U(\boldsymbol{\beta}) = U_n(\boldsymbol{\beta}) = U_n(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} D_i^t(\boldsymbol{\beta}) V_i^{-1}(\boldsymbol{\beta}, \boldsymbol{\alpha})(Y_i - \mu_i(\boldsymbol{\beta})), \qquad (1)$$

where $D_i(\boldsymbol{\beta}) = \partial \mu_i(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$ and $V_i(\boldsymbol{\beta}, \boldsymbol{\alpha}) = A_i^{1/2}(\boldsymbol{\beta}) R_i(\boldsymbol{\alpha}) A_i^{1/2}(\boldsymbol{\beta})$ with a working correlation matrix $R(\boldsymbol{\alpha})$. Here $\boldsymbol{\alpha}$ is a working correlation parameter of r dimension. Let $R(\boldsymbol{\alpha}_n)$ be the working correlation matrix $R(\boldsymbol{\alpha})$ in $U_n(\boldsymbol{\beta}, \boldsymbol{\alpha})$. The GEE estimator $\hat{\boldsymbol{\beta}}_n$ depends on the working correlation matrix $R(\boldsymbol{\alpha}_n)$.

Due to lack of likelihood function (or more generally, objective function to be maximized), statistical inference about the GEE estimator relies on large sample theory. When establishing asymptotic properties of $\hat{\boldsymbol{\beta}}_n$, it is a key assumption that there exists a limit of $R(\boldsymbol{\alpha}_n)$. We denote this limit by $R(\boldsymbol{\alpha}^*)$. Under the asymptotic setting in which the maximum cluster size and the dimension of $\boldsymbol{\alpha}^*$ are finite, Liang and Zeger [4] derived asymptotic properties of $\hat{\boldsymbol{\beta}}_n$ via the Taylor expansion of $n^{-1/2} U_n(\boldsymbol{\beta}, \boldsymbol{\alpha})$ at $(\boldsymbol{\beta}, \boldsymbol{\alpha}) = (\boldsymbol{\beta}_0, \boldsymbol{\alpha}^*)$. The GEE estimator $\hat{\boldsymbol{\beta}}_n$ is asymptotically consistent and asymptotically normally distributed. The asymptotic variance of $\hat{\boldsymbol{\beta}}_n$, which indicates the efficiency of $\hat{\boldsymbol{\beta}}_n$, is minimized when $R_i(\boldsymbol{\alpha}_n)$ is the same as the true correlation matrix of Y_i. Such flexibility of working correlation matrix with regard to the asymptotic consistency of $\hat{\boldsymbol{\beta}}_n$ is a strength of the GEE model. However, in order to achieve a reasonable degree of efficiency for the GEE estimator, it is still important to choose a plausible working correlation matrix. Furthermore, if the working correlation matrix is poorly chosen, $R(\boldsymbol{\alpha}^*)$ could fail to exist even with fixed cluster size (see [1]). We note that in some cases, the estimating equations (1) are closely related to the score equations for a multivariate exponential family. Asymptotic results for maximum likelihood estimator with misspecified correlation structure were also established by White [8] and Gouriéroux, Monfort and Trognon [2].

When applying large sample theory to finite-sample data, it is an interesting but open question how large sample size is large enough. In the standard large sample theory based on Taylor expansion, this question relates to the relative size of random fluctuations in the remainder terms (e.g. [3]). In the GEE setting of Liang and Zeger [4], this question relates to whether the cluster size and the complexity of working correlation model are small enough compared to the number of independent clusters. In particular, the relative size of r to n is important because the size of a vector $(\boldsymbol{\alpha}_n - \boldsymbol{\alpha}^*)$ is closely related to the remainder terms of the Taylor expansion of $n^{-1/2}U_n(\boldsymbol{\beta}, \boldsymbol{\alpha})$.

In practice we occasionally encounter the large cluster data such as dental data (32 teeth per person) and panel time series data (e.g [6]). For example, see the Childhood Asthma Management Program (CAMP) air pollution ancillary study [10], the aim of which is to examine short-term air pollution effects on asthma symptoms in children. This data set contains daily self-reports of asthma symptoms of 133 children in Seattle area for an average of 58 days (range: 28 - 112 days) at the screening phase of the CAMP study. Note here that the average cluster size of 58 and the maximum cluster size of 112 are not negligibly small compared to the number of children, 133. It is also interesting to note that there are $\binom{112}{2} \approx 6200$ possibly distinct pairwise correlation components even if a common correlation structure across children is assumed.

When the cluster size is relatively large, it seems relevant to consider an asymptotic setting where the maximum cluster size increases with the number of clusters. Furthermore, since the true correlation structure is likely to be complicated for large cluster data, it seems more appropriate to consider an asymptotic setting where both the maximum cluster size and the complexity of working correlation model increase with the number of clusters. Note that this asymptotic setting is different from that in [4]. Recently, Xie and Yang [9] considered the large cluster GEE model and presented asymptotic results. However, in the approach of Xie and Yang [9], the working correlation matrix is assumed to be fixed. Although an independent working correlation matrix satisfies this assumption, generally speaking, Xie and Yang [9] did not address an issue of modeling and estimating a possibly high-dimensional working correlation model in the large cluster GEE model. The asymptotic results for large cluster GEE model with independence working correlation matrix were independently presented by Lumley and Mayer-Hamblett [5].

This paper is concerned with the large cluster GEE with a high-dimensional working correlation model. To be specific, under the asymptotic setting where both the maximum cluster size and the complexity of working correlation model increase with the number of clusters, we derive asymptotic properties of the GEE estimator, using the results of empirical process theory and the work of Xie and Yang [9]. The outline of this paper is as follows. In Section 2 we introduce basic notation and regularity conditions. The asymptotic existence and the weak consistency of the GEE estimator are established in Section 3. We show the asymptotic normality of the GEE estimator and the weak consistency of sandwich variance estimator in Sections 4 and 5, respectively. In Section 6 we discuss regularity conditions on the complexity of working correlation model and give sufficient conditions on the number of working correlation parameters. The final section provides concluding remarks.

2 Notation and Assumptions

Let m denote the maximum cluster size over n independent clusters:

$$m = m(n) = \max_{1 \le i \le n} m_i.$$

We note that r can be also viewed as a function of n:

$$r = r(n, m) = r(n, m(n)).$$

In order to put an emphasis on the relation of r and n, we shall hereafter add subscript n to r. In our asymptotic setting, both m and r_n can increase with n.

To deal with a sequence of different-dimensional working parameter spaces, we consider a single embedding space of larger dimension. Here we define a working correlation parameter space on the ℓ_2 space which is an infinite-dimensional analogue of Euclidean space. Let a point $(\boldsymbol{\beta}_0, \boldsymbol{\alpha}^*) \in \mathbb{R}^p \times \ell_2$ be fixed. We define the regression parameter space $T(\boldsymbol{\beta}_0)$ as

$$T(\boldsymbol{\beta}_0) = \{\boldsymbol{\beta} \in \mathbb{R}^p; ||\boldsymbol{\beta} - \boldsymbol{\beta}_0|| \le \delta_\beta\}$$

with a fixed $\delta_\beta > 0$, and define the working correlation parameter space $T(\boldsymbol{\alpha}^*; q)$ as

$$T(\boldsymbol{\alpha}^*; q) = \{\boldsymbol{\alpha} \in \ell_2; |\boldsymbol{\alpha}^* - \boldsymbol{\alpha}|_i \le a i^{-q}\}$$

with fixed $a > 0$ and $q > 0$. Without loss of generality, the constant a is usually set to one. The parameter space T is then defined as a product space of the regression parameter space $T(\boldsymbol{\beta}_0)$ and the working correlation parameter space $T(\boldsymbol{\alpha}^*; q)$:

$$T = T(\boldsymbol{\beta}_0, \boldsymbol{\alpha}^*; q) = T(\boldsymbol{\beta}_0) \times T(\boldsymbol{\alpha}^*; q) \subset \mathbb{R}^p \times \ell_2.$$

A distance measure d_T on the parameter space T is naturally defined as:

$$d_T((\boldsymbol{\beta}, \boldsymbol{\alpha}), (\boldsymbol{\beta}', \boldsymbol{\alpha}')) = \sqrt{d_1^2(\boldsymbol{\beta}, \boldsymbol{\beta}') + d_2^2(\boldsymbol{\alpha}, \boldsymbol{\alpha}')},$$

where d_1 is the Euclidean distance in $\mathbb{R}^p$ and d_2 is the ℓ_2-distance. A few remarks on the working correlation parameter space $T(\boldsymbol{\alpha}^*; q)$ are in order. First, the infinite-dimensional working correlation parameter space $T(\boldsymbol{\alpha}^*; q)$ can be approximated to a subspace of a finite-dimensional Euclidean space with a small margin of errors. Second, the size of the working parameter space $T(\boldsymbol{\alpha}^*; q)$ is controlled by q as $T(\boldsymbol{\alpha}^*; q_1) \subset T(\boldsymbol{\alpha}^*; q_2)$ for $q_1 > q_2 > 0$.

Also note that each summand in (1) needs to be modified accordingly:

$$g_{n\boldsymbol{\alpha},i}(\boldsymbol{\beta}) = g_{n\boldsymbol{\alpha}_{(n)},i}(\beta) = D_i^t(\boldsymbol{\beta}) V_i^{-1}(\boldsymbol{\beta}, \tilde{\boldsymbol{\alpha}}_{(n)})(Y_i - \mu_i(\boldsymbol{\beta})),$$

where $\boldsymbol{\alpha}_{(n)} = (\alpha_1, \cdots, \alpha_{r_n}, 0, 0, \cdots) \in \ell_2$ and $\tilde{\boldsymbol{\alpha}}_{(n)} = (\alpha_1, \cdots, \alpha_{r_n}) \in \mathbb{R}^{r_n}$ with $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots) \in \ell_2$. This redefinition is only for notational as it is governed only by $\boldsymbol{\beta} \in \mathbb{R}^p$ and the first r_n elements of $\boldsymbol{\alpha} \in \ell_2$. The sum of n estimating equations is now defined as

$$g_{n\boldsymbol{\alpha}}(\beta) = \sum_{i=1}^n g_{n\boldsymbol{\alpha},i}(\boldsymbol{\beta}) = \sum_{i=1}^n D_i^t(\boldsymbol{\beta}) V_i^{-1}(\boldsymbol{\beta}, \tilde{\boldsymbol{\alpha}}_{(n)})(Y_i - \mu_i(\boldsymbol{\beta})).$$

Similarly as in Xie and Yang [9], we denote

$$\begin{aligned}
M_{n\boldsymbol{\alpha}}(\boldsymbol{\beta}) &= \mathrm{Cov}(g_{n\boldsymbol{\alpha}}(\boldsymbol{\beta})), \\
H_{n\boldsymbol{\alpha}}(\boldsymbol{\beta}) &= \sum_{i=1}^n D_i^t(\boldsymbol{\beta}) V_i^{-1}(\boldsymbol{\beta}, \tilde{\boldsymbol{\alpha}}_{(n)}) D_i^t(\boldsymbol{\beta}), \\
\mathcal{D}_{n\boldsymbol{\alpha}}(\boldsymbol{\beta}) &= -\frac{\partial g_{n\boldsymbol{\alpha}}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^t}.
\end{aligned}$$

To alleviate the notation, we suppress the argument of $\boldsymbol{\beta}_0$ when $\boldsymbol{\beta} = \boldsymbol{\beta}_0$. For example, $g_{n\boldsymbol{\alpha}} = g_{n\boldsymbol{\alpha}}(\boldsymbol{\beta}_0)$. Also, for a fixed $\boldsymbol{\alpha}^* \in \ell_2$, we denote

$$\begin{aligned}
\gamma_n^{(D)} &= \max_{i=1,\cdots,n} \lambda_{\max}(H_{n\boldsymbol{\alpha}^*}^{-1/2} D_i^t V_i^{-1}(\boldsymbol{\alpha}^*) D_i H_{n\boldsymbol{\alpha}^*}^{-1/2}), \\
c_n &= \lambda_{\max}(M_{n\boldsymbol{\alpha}^*}^{-1} H_{n\boldsymbol{\alpha}^*}), \\
\tilde{\lambda}_n &= \max_{i=1,\cdots,n} \lambda_{\max}(R_i^{-1}(\boldsymbol{\alpha}^*)).
\end{aligned}$$

Note that c_n is related to the discrepancy between the working correlation matrix at $\boldsymbol{\alpha} = \boldsymbol{\alpha}^*$ and the true correlation matrix, and that $\tilde{\lambda}_n$ is related to the eigenvalues of the working correlation matrix at $\boldsymbol{\alpha} = \boldsymbol{\alpha}^*$. It is interesting to notice that $c_n = 1$ when $H_{n\boldsymbol{\alpha}^*} = M_{n\boldsymbol{\alpha}^*}$ and that $\tilde{\lambda}_n = 1$ with independent working correlation matrix. We also follow the matrix notations used in

[9]. Therefore, $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ represent the minimum and maximum eigenvalues of the matrix A, respectively. The ordering between two square matrices A_1 and A_2 is defined as: $A_1 \geq A_2$ if and only if $\lambda^T A_1 \lambda \geq \lambda^T A_2 \lambda$ for all vectors λ with $||\lambda|| = 1$.

In this paper, the following regularity conditions will be always assumed:

A.1 $(\boldsymbol{\beta}, \boldsymbol{\alpha}) \in T = T(\boldsymbol{\beta}_0, \boldsymbol{\alpha}^*; q) \subset \mathbb{R}^p \times \ell_2$ for some $q > 1$.

A.2 $H_{n\boldsymbol{\alpha}}$ and $M_{n\boldsymbol{\alpha}}$ are positive definite.

A.3 $\frac{\partial^2}{\partial \boldsymbol{\beta}^2} g_{n\boldsymbol{\alpha}}(\boldsymbol{\beta})$ exists at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$.

The condition A.3 is about the smoothness property of the estimating equations with respect to $\boldsymbol{\beta}$. Additional smoothness conditions of the estimating equations with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ will be presented later.

3 Asymptotic existence and weak consistency of the GEE estimator

For existence and consistency of $\hat{\boldsymbol{\beta}}_n$, we present the following conditions:

B.1

$$\sup_{\boldsymbol{\alpha} \in T(\boldsymbol{\alpha}^*; q)} \frac{\tau_{n\boldsymbol{\alpha}}}{\lambda_{\min}(H_{n\boldsymbol{\alpha}})} \rightarrow 0,$$

where $\tau_{n\boldsymbol{\alpha}} = \max_{i=1,\cdots,n} \lambda_{\max}(R_i^{-1}(\boldsymbol{\alpha})\bar{R}_i)$ with the true correlation matrix $\bar{R}_i$.

B.2 For any given $r > 0$ and $\eta > 0$,

$$\Pr(\sup_{\boldsymbol{\alpha} \in T(\boldsymbol{\alpha}^*; q)} \sup_{\boldsymbol{\beta} \in B_{n\boldsymbol{\alpha}}(r)} ||H_{n\boldsymbol{\alpha}}^{-1/2} \mathcal{D}_{n\boldsymbol{\alpha}}(\boldsymbol{\beta}) H_{n\boldsymbol{\alpha}}^{-1/2} - I|| < \eta) \rightarrow 1,$$

where $B_{n\boldsymbol{\alpha}}(r) = \{\boldsymbol{\beta} \in \mathbb{R}^p : ||H_{n\boldsymbol{\alpha}}^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)|| \leq r\sqrt{\tau_{n\boldsymbol{\alpha}}}\}$.

REMARK 1. When a single fixed $\boldsymbol{\alpha}$ is considered, the conditions B.1 and B.2 are equivalent to the conditions of Xie and Yang [9].

When none of the n independent summands of $g_{n\boldsymbol{\alpha}}$ dominates the rest, $H_{n\boldsymbol{\alpha}}$ increases at least at the rate of n. Hence the condition B.1 is typically met as $\tau_{n\boldsymbol{\alpha}}$ is determined by a single summand in $g_{n\boldsymbol{\alpha}}$. Also note that

$$B_{n\boldsymbol{\alpha}}(r) \subset \{\boldsymbol{\beta} : ||\boldsymbol{\beta} - \boldsymbol{\beta}_0|| \leq r\sqrt{\frac{\tau_{n\boldsymbol{\alpha}}}{\lambda_{\min}(H_{n\boldsymbol{\alpha}})}}\}.$$

Combined with the condition B.1, the radius of the sphere $B_{n\boldsymbol{\alpha}}(r)$ becomes small for large n. Based on a simple observation that

$$\mathrm{E}\mathcal{D}_{n\boldsymbol{\alpha}} = H_{n\boldsymbol{\alpha}} = H_{n\boldsymbol{\alpha}}(\boldsymbol{\beta}_0),$$

the condition B.2 is about the law of large numbers for the random matrices $\mathcal{D}_{n\boldsymbol{\alpha}}(\boldsymbol{\beta})$ and the uniform continuity of $H_{n\boldsymbol{\alpha}}(\boldsymbol{\beta})$ at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$.

The following theorem asserts that there exists a consistent sequence of roots of the estimating equations.

Theorem 1. *Suppose the conditions (B.1 - B.2) hold. Then, for every sequence* $\{\boldsymbol{\alpha}_n \in T(\boldsymbol{\alpha}^*; q)\}$*, there exists a sequence of random variables* $\hat{\boldsymbol{\beta}}_n$*, such that*

$$\Pr(g_{n\boldsymbol{\alpha}_n}(\hat{\boldsymbol{\beta}}_n) = 0) \to 1$$

and

$$\hat{\boldsymbol{\beta}}_n \to \boldsymbol{\beta}_0 \qquad \textit{in probability.}$$

Proof. See Appendix A. □

4 Asymptotic distribution of the GEE estimator

In this section, we derive the asymptotic distributional properties of $\hat{\boldsymbol{\beta}}_n$ of Theorem 1 from the asymptotic properties of the corresponding class of estimating equations indexed by $(\boldsymbol{\beta}, \boldsymbol{\alpha})$. Let us define $\mathcal{G}_n$ as:

$$\mathcal{G}_n = \{M_{n\boldsymbol{\alpha}}^{-1/2} g_{n\boldsymbol{\alpha}}(\boldsymbol{\beta}) : (\boldsymbol{\beta}, \boldsymbol{\alpha}) \in T\}.$$

Under suitable conditions on the size of the class $\mathcal{G}_n$, we first show that $\mathcal{G}_n$ is a Donsker class. To measure the size of the class $\mathcal{G}_n$, we follow the notation and terminology of van der Vaart and Wellner [7]. For example, for a class $\mathcal{F}$ equipped with a metric $||\cdot||$, we define the covering number $N(\epsilon, \mathcal{F}, ||\cdot||)$ and the bracketing number $N_{[\,]}(\epsilon, \mathcal{F}, ||\cdot||)$ as in [7].

The following conditions will be assumed later in this section:

C.1 Let $Y_i^* = (Y_{i1}^*, \cdots, Y_{im_i}^*)^t = A_i^{-1/2}(Y_i - \mu_i(\boldsymbol{\beta}_0))$. Then, there exists $\delta_0 > 0$ such that $\mathrm{E}|Y_{ij}^*|^{2+2/\delta_0}$ is uniformly bounded above and that $(c_n \tilde{\lambda}_n m)^{1+\delta_0} \gamma_n^{(D)} \to 0$.

C.2 For every n and i, there exists an $m_i \times m_i$ matrix W_{ni} such that

$$|M_{n\boldsymbol{\alpha}}^{-1/2} g_{n\boldsymbol{\alpha},i}(\boldsymbol{\beta}) - M_{n\boldsymbol{\alpha}'}^{-1/2} g_{n\boldsymbol{\alpha}',i}(\boldsymbol{\beta}')| \le d_T((\boldsymbol{\beta}, \boldsymbol{\alpha}), (\boldsymbol{\beta}', \boldsymbol{\alpha}'))\, M_{n\boldsymbol{\alpha}^*}^{-1/2} D_i^T(\boldsymbol{\beta}_0) W_{ni}(Y_i - \mu_i(\boldsymbol{\beta}_0)),$$

for all $(\boldsymbol{\beta}, \boldsymbol{\alpha}), (\boldsymbol{\beta}', \boldsymbol{\alpha}') \in T$. Further, the following moments conditions hold:

- $\mathrm{Cov}\left(M_{n\boldsymbol{\alpha}^*}^{-1/2} \sum_{i=1}^n D_i^T(\boldsymbol{\beta}_0) W_{ni}(Y_i - \mu_i(\boldsymbol{\beta}_0)) \right) = O(1)$,
- $\sum_{i=1}^n \mathrm{E}|M_{n\boldsymbol{\alpha}^*}^{-1/2} D_i^T(\boldsymbol{\beta}_0) W_{ni}(Y_i - \mu_i(\boldsymbol{\beta}_0))|^{2+\delta} = o(1)$ for some $\delta > 0$.

REMARK 2. As $n \to \infty$, $\gamma_n^{(D)}$ will typically decrease at the order of n^{-1} (i.e., $\gamma_n^{(D)} = O(n^{-1})$), whereas c_n and $\tilde{\lambda}_n$ are either bounded above or divergent to infinity (as a function of m). Therefore, the condition C.1 essentially places a restriction on the increasing rate of m versus n:

$$m = o(n^{1-\delta'}),$$

where δ' depends on the moment conditions of response measurements Y.

REMARK 3. The Lipschitz property in the condition C.2 implies the existence of envelope function G_n for $\mathcal{G}_n$, where $G_n = \sum_{i=1}^n G_{ni}$ with

$$G_{ni} = M_{n\boldsymbol{\alpha}^*}^{-1/2} |g_{n\boldsymbol{\alpha}^*,i}| + diam(T) \times \left(M_{n\boldsymbol{\alpha}^*}^{-1/2} D_i^T(\boldsymbol{\beta}_0) W_{ni}(Y_i - \mu_i(\boldsymbol{\beta}_0)) \right).$$

The following three lemmas will be useful in showing the Donsker property of $\mathcal{G}_n$ in Theorem 2.

LEMMA 1. *Suppose the condition C.2 holds. Then,*

$$N_{[\,]}(4\epsilon ||G_n||_{P,2}, \mathcal{G}_n, L_2(P)) \le N(2\epsilon, T, d_T).$$

Proof. See Theorem 2.7.11 in [7]. □

LEMMA 2. *For any $q > 0.5$,*

$$\log N(2\epsilon, T, d_T) \le \left(p - \frac{q}{2q-1}\right) \log\left(\frac{1}{\epsilon}\right) + q \left(\frac{1}{\epsilon}\right)^{\frac{1}{q-0.5}} + C,$$

where C is a constant depending on p and q.

The proof of Lemma 2 is given in Appendix B. The key idea is to cover the ball of radius of ϵ in the parameter space by the Cartesian product of the ball of radius $\epsilon/\sqrt{2}$ in the regression parameter space (in $\mathbb{R}^p$) and the ball of radius $\epsilon/\sqrt{2}$ in the working correlation parameter space (in ℓ_2). Then the working correlation parameter space is approximated into a subspace in a finite-dimensional Euclidean space $\mathbb{R}^{k(\epsilon)}$.

LEMMA 3. *Suppose the conditions (C.1 - C.2) hold. G_n satisfy the Lindeberg condition.*

Proof. Let λ be a fixed $p\times 1$ vector with $||\lambda|| = 1$. Let $Z_{ni,A} = \lambda^t M_{n\boldsymbol{\alpha}^*}^{-1/2}|g_{n\boldsymbol{\alpha}^*,i}|$ and let $Z_{ni,B} = \lambda^t M_{n\boldsymbol{\alpha}^*}^{-1/2} D_i^T(\boldsymbol{\beta}_0)W_{ni}(Y_i - \mu_i(\boldsymbol{\beta}_0))$. In the proof of Lemma 2 in Xie and Yang [9], it is shown that $\sum_{i=1}^n \mathrm{E}Z_{ni,A}^2 \mathrm{I}(|Z_{ni,A}| > \epsilon) \to 0$ for any $\epsilon > 0$. The Linderberg condition for the double arrays $Z_{ni,B}$ follows from the Lyapounov condition in C.2. Therefore, for $Z_{ni} = Z_{ni,A} + diam(T)Z_{ni,B}$, we have $\sum_{i=1}^n \mathrm{E}Z_{ni}^2 \mathrm{I}(|Z_{ni}| > \epsilon) \to 0$ for any $\epsilon > 0$. □

Lemmas 1 and 2 give a bracketing entropy condition:

$$\int_0^\tau \sqrt{\log N_{[\,]}(\epsilon, \mathcal{G}_n, L_2(P))}d\epsilon < \infty, \tag{2}$$

for any $\tau > 0$ and $q > 1$. This condition (2) is crucial in asserting that the class $\mathcal{G}_n$ is Donsker. Lemma 3 gives the Lindeberg condition for the envelope functions G_n, which is also essential in asserting that the limit process is the Gaussian process.

Theorem 2. *Suppose the conditions (C.1 - C.2) hold. Then, $\mathcal{G}_n$ is a Donsker class.*

Proof. By Theorem 2.11.9 in [7], the proof of this theorem follows from (2) and Lemma 3. □

The following is an immediate consequence of Theorem 2.

Theorem 3. *Let all the assumptions of Theorem 2 be fulfilled. Then, for any consistent estimator $(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\alpha}}_n)$ of $(\boldsymbol{\beta}_0, \boldsymbol{\alpha}^*)$ with $\Pr((\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\alpha}}_n) \in T) \to 1$,*

$$M_{n\hat{\boldsymbol{\alpha}}_n}^{-1/2}\left(g_{n\hat{\boldsymbol{\alpha}}_n}(\hat{\boldsymbol{\beta}}_n) - \mathrm{E}g_{n\hat{\boldsymbol{\alpha}}_n}(\hat{\boldsymbol{\beta}}_n)\right) - M_{n\boldsymbol{\alpha}^*}^{-1/2}\left(g_{n\boldsymbol{\alpha}^*} - \mathrm{E}g_{n\boldsymbol{\alpha}^*}\right) = o_p(1).$$

Note that $M_{n\boldsymbol{\alpha}^*}^{-1/2}(g_{n\boldsymbol{\alpha}^*} - \mathrm{E}g_{n\boldsymbol{\alpha}^*})$ asymptotically follows a standard p-dimensional normal distribution. Combined with differentiability of a map $\boldsymbol{\beta} \mapsto M_{n\boldsymbol{\alpha}}^{-1/2}\mathrm{E}g_{n\boldsymbol{\alpha}}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ uniformly in $\boldsymbol{\alpha}$, the asymptotic normality of the GEE estimator $\hat{\boldsymbol{\beta}}_n$ can be established. For notational simplicity, we define $\phi_n(\boldsymbol{\beta}, \boldsymbol{\alpha})$ and $\dot{\phi}_n(\boldsymbol{\beta}, \boldsymbol{\alpha})$ as $M_{n\boldsymbol{\alpha}}^{-1/2}\mathrm{E}g_{n\boldsymbol{\alpha}}(\boldsymbol{\beta})$ and $\frac{\partial}{\partial \boldsymbol{\beta}} M_{n\boldsymbol{\alpha}}^{-1/2}\mathrm{E}g_{n\boldsymbol{\alpha}}(\boldsymbol{\beta})$, respectively.

Theorem 4. *Suppose the conditions (B.1 - B.2) and (C.1 - C.2) hold. Further, assume that*

$$\sup_n \sup_{\boldsymbol{\beta}:||\boldsymbol{\beta}-\boldsymbol{\beta}_0||<\eta} |\frac{\partial^2}{\partial \boldsymbol{\beta}^2}\phi_n(\boldsymbol{\beta}, \boldsymbol{\alpha}^*)| < \infty, \tag{3}$$

for small $\eta > 0$. Then, given any sequence $\{\boldsymbol{\alpha}_n \in T(\boldsymbol{\alpha}^*; q)\}$ with $||\boldsymbol{\alpha}_n - \boldsymbol{\alpha}^*|| = o(1)$, the corresponding consistent GEE estimator $\hat{\boldsymbol{\beta}}_n = \hat{\boldsymbol{\beta}}_n(\boldsymbol{\alpha}_n)$ has the following asymptotic distribution:

$$\dot{\phi}_n(\boldsymbol{\beta}_0, \boldsymbol{\alpha}^*)(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{d} \mathrm{N}_p(0, \mathrm{I}_p).$$

Proof. Note that

$$\begin{aligned}
& M_{n\boldsymbol{\alpha}_n}^{-1/2} g_{n\boldsymbol{\alpha}_n}(\hat{\boldsymbol{\beta}}_n) - M_{n\boldsymbol{\alpha}_n}^{-1/2}\mathrm{E}g_{n\boldsymbol{\alpha}_n}(\hat{\boldsymbol{\beta}}_n) \\
&= -\phi_n(\hat{\boldsymbol{\beta}}_n, \boldsymbol{\alpha}_n) + o_p(1) \\
&= -\Big(\phi_n(\hat{\boldsymbol{\beta}}_n, \boldsymbol{\alpha}_n) - \phi_n(\boldsymbol{\beta}_0, \boldsymbol{\alpha}^*)\Big) + o_p(1) \\
&= -\Big(\phi_n(\hat{\boldsymbol{\beta}}_n, \boldsymbol{\alpha}_n) - \phi_n(\hat{\boldsymbol{\beta}}_n, \boldsymbol{\alpha}^*)\Big) - \Big(\phi_n(\hat{\boldsymbol{\beta}}_n, \boldsymbol{\alpha}^*) - \phi_n(\boldsymbol{\beta}_0, \boldsymbol{\alpha}^*)\Big) + o_p(1) \\
&= O(||\boldsymbol{\alpha}_n - \boldsymbol{\alpha}^*||) - \Big(\dot{\phi}_n(\boldsymbol{\beta}_0, \boldsymbol{\alpha}^*)(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) + O(||\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0||^2)\Big) + o_p(1) \\
&= -\Big(\dot{\phi}_n(\boldsymbol{\beta}_0, \boldsymbol{\alpha}^*)(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)\Big) + o_p(1).
\end{aligned}$$

The first equality above follows from the definition of the GEE estimator $\hat{\boldsymbol{\beta}}_n$ and the second equality above follows from the definition of the true parameter $\boldsymbol{\beta}_0$. The last second equality follows from the smoothness of $\phi_n(\boldsymbol{\beta}, \boldsymbol{\alpha})$ in the parameter $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ as in the condition C.2 and the second-order Taylor expansion of p-dimensional function $\phi_n(\boldsymbol{\beta}, \boldsymbol{\alpha}^*)$ at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$. The proof is completed by applying Theorem 3. □

5 Weak consistency of the sandwich variance estimator

Let us define $A_n(\boldsymbol{\beta},\boldsymbol{\alpha})$, $B_n(\boldsymbol{\beta},\boldsymbol{\alpha})$ and $\Xi_n(\boldsymbol{\beta},\boldsymbol{\alpha})$ as:

$$
\begin{aligned}
A_n(\boldsymbol{\beta},\boldsymbol{\alpha}) &= M_{n\boldsymbol{\alpha}}^{-1/2}\left(\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\beta}}g_{n\boldsymbol{\alpha},i}(\boldsymbol{\beta})\right),\\
B_n(\boldsymbol{\beta},\boldsymbol{\alpha}) &= M_{n\boldsymbol{\alpha}}^{-1/2}\left(\sum_{i=1}^{n}g_{n\boldsymbol{\alpha},i}(\boldsymbol{\beta})g_{n\boldsymbol{\alpha},i}^{t}(\boldsymbol{\beta})\right)M_{n\boldsymbol{\alpha}}^{-1/2},\\
\Xi_n(\boldsymbol{\beta},\boldsymbol{\alpha}) &= \left(\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\beta}}g_{n\boldsymbol{\alpha},i}(\boldsymbol{\beta})\right)^{-1}\left(\sum_{i=1}^{n}g_{n\boldsymbol{\alpha},i}(\boldsymbol{\beta})g_{n\boldsymbol{\alpha},i}^{t}(\boldsymbol{\beta})\right)\left(\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\beta}}g_{n\boldsymbol{\alpha},i}(\boldsymbol{\beta})\right)^{-1}\\
&= A_n^{-1}(\boldsymbol{\beta},\boldsymbol{\alpha})B_n(\boldsymbol{\beta},\boldsymbol{\alpha})A_n^{-1}(\boldsymbol{\beta},\boldsymbol{\alpha}).
\end{aligned}
$$

The sandwich variance estimator for $\hat{\boldsymbol{\beta}}_n$ in Theorem 4 is then given by $\Xi_n(\hat{\boldsymbol{\beta}}_n,\boldsymbol{\alpha}_n)$. In the following theorem, we show the weak consistency of $\Xi_n(\hat{\boldsymbol{\beta}}_n,\boldsymbol{\alpha}_n)$ to the asymptotic variance of $\hat{\boldsymbol{\beta}}_n$ by using the consistency of $(\hat{\boldsymbol{\beta}}_n,\boldsymbol{\alpha}_n)$ to $(\boldsymbol{\beta}_0,\boldsymbol{\alpha}^*)$ and the Glivenko-Cantelli property. The Glivenko-Cantelli property of a class of functions refers to the law of large numbers uniformly over the class. For a class to be Glivenko-Cantelli, finite bracketing number of the class and law of large numbers for each function are sufficient.

Theorem 5. *Suppose all the assumptions in Theorem 4 hold. We also assume that, for all* $(\boldsymbol{\beta},\boldsymbol{\alpha}),(\boldsymbol{\beta}',\boldsymbol{\alpha}')\in T$,

$$
\begin{aligned}
|A_n(\boldsymbol{\beta},\boldsymbol{\alpha})-A_n(\boldsymbol{\beta}',\boldsymbol{\alpha}')| &\le d_T((\boldsymbol{\beta},\boldsymbol{\alpha}),(\boldsymbol{\beta}',\boldsymbol{\alpha}'))\tilde{A}_n \qquad (4a)\\
|B_n(\boldsymbol{\beta},\boldsymbol{\alpha})-B_n(\boldsymbol{\beta}',\boldsymbol{\alpha}')| &\le d_T((\boldsymbol{\beta},\boldsymbol{\alpha}),(\boldsymbol{\beta}',\boldsymbol{\alpha}'))\tilde{B}_n, \qquad (4b)
\end{aligned}
$$

for some $\tilde{A}_n$ *and* $\tilde{B}_n$ *with* $\sup_n \mathrm{E}\tilde{A}_n<\infty$ *and* $\sup_n \mathrm{E}\tilde{B}_n<\infty$. *Then we have*

$$\sup_{(\boldsymbol{\beta},\boldsymbol{\alpha})\in T}|A_n(\boldsymbol{\beta},\boldsymbol{\alpha})-\mathrm{E}A_n(\boldsymbol{\beta},\boldsymbol{\alpha})|\xrightarrow{P}0,$$

and

$$\sup_{(\boldsymbol{\beta},\boldsymbol{\alpha})\in T}|B_n(\boldsymbol{\beta},\boldsymbol{\alpha})-\mathrm{E}B_n(\boldsymbol{\beta},\boldsymbol{\alpha})|\xrightarrow{P}0.$$

Consequently, for $(\hat{\boldsymbol{\beta}}_n,\boldsymbol{\alpha}_n)$ *in Theorem 4,* $\Xi_n(\hat{\boldsymbol{\beta}}_n,\boldsymbol{\alpha}_n)$ *is consistent for the asymptotic variance of* $\hat{\boldsymbol{\beta}}_n$.

Proof. See Appendix C. □

REMARK 4. The assumption (4) is closely related to the condition C.2 and condition (3), but slightly stronger than those.

6 Complexity of working correlation model

In the preceding sections, a restriction of the increasing rate of r_n is implicitly imposed through

$$\Pr(\hat{\boldsymbol{\alpha}}_n \in T(\boldsymbol{\alpha}^*; q)) \to 1,$$

where $\hat{\boldsymbol{\alpha}}_n$ has r_n components to be estimated from the data. Note that

$$\begin{aligned} &\Pr(\hat{\boldsymbol{\alpha}}_n \notin T(\boldsymbol{\alpha}^*; q)) \\ &\quad = \Pr(\max_{1 \le j \le r_n} j^q |\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}^*|_j \ge 1) \\ &\quad \le \Pr(\max_{1 \le j \le r_n} |\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}^*|_j \ge \frac{1}{r_n^q}). \end{aligned} \tag{5}$$

In this section, by using a maximal inequality (Lemma 2.2.2 of [7]), we present more explicit sufficient conditions on r_n with suitable moments assumption on $\hat{\boldsymbol{\alpha}}_n$. Throughout this section we assume the following assumption:

D.1 Each element of $\hat{\boldsymbol{\alpha}}_n$ is estimable individually at $\sqrt{n}$ rate:

$$\sqrt{n}(\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}^*)_j = O_p(1),$$

for $j = 1, 2, \cdots$.

Here we restate Lemma 2.2.2 of [7].

LEMMA 4. *Let ψ be a convex, nondecreasing, nonzero function with $\psi(0) = 0$ and*
$\limsup_{z_1, z_2 \to \infty} \psi(z_1)\psi(z_2)/\psi(cz_1z_2) < \infty$ *for some constant c. Then, for any random variables $Z_1, \cdots, Z_m$,*

$$|| \max_{1 \le i \le m} Z_i ||_\psi \le C\psi^{-1}(m) \max_i ||Z_i||_\psi,$$

for a constant C depending on ψ.

Examples of ψ in Lemmma 4 include $\psi_k(z) = z^k$ for any $k \ge 1$ and $\psi_\infty(z) = e^z - 1$. In the following two theorems, we give sufficient conditions on r_n by applying Lemma 4 with either $\psi_k(\cdot)$ or $\psi_\infty(\cdot)$.

Theorem 6. *Let $k \geq 1$ be fixed. Suppose that*

$$\limsup_n \max_{1\leq j\leq r_n} \mathrm{E}(|\hat{\boldsymbol{\alpha}}_n|_j)^k < \infty.$$

Then, for $r_n = o(n^{\frac{1}{2(q+1/k)}})$, we have

$$\Pr(\hat{\boldsymbol{\alpha}}_n \in T(\boldsymbol{\alpha}^*; q)) \to 1.$$

Proof. By (5), it suffices to show that $\Pr(\max_{1\leq j\leq r_n} |\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}^*|_j \geq r_n^{-q}) \to 0$. By Markov's inequality,

$$\begin{aligned}\Pr(\max_{1\leq j\leq r_n} |\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}^*|_j \geq r_n^{-q}) &\leq \frac{1}{\psi_k(r_n^{-q}/||\max_{1\leq j\leq r_n} |\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}^*|_j||_{\psi_k})} \\ &= (r_n^q \times || \max_{1\leq j\leq r_n} |\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}^*|_j||_{\psi_k})^k \qquad (6)\end{aligned}$$

By Lemma 4 with $\psi_k(\cdot)$,

$$|| \max_{1\leq j\leq r_n} |\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}^*|_j||_{\psi_k} \leq C r_n^{1/k} \max_{1\leq j\leq r_n} ||(\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}^*)_j||_{\psi_k} = r_n^{1/k} O_p(1/\sqrt{n}). \quad (7)$$

Therefore, Theorem 6 follows from combination of (6) and (7). □

Theorem 7. *Suppose that*

$$\limsup_n \max_{1\leq j\leq r_n} \mathrm{E}e^{(\hat{\boldsymbol{\alpha}}_n)_j} < \infty.$$

Then, for $r_n = o(n^{\frac{1}{2q}-\eta})$ with any small $\eta > 0$, we have

$$\Pr(\hat{\boldsymbol{\alpha}}_n \in T(\boldsymbol{\alpha}^*; q)) \to 1.$$

Proof. The proof of Theorem 7 is almost same as the proof of Theorem 6, except for the use of $\psi_\infty(\cdot)$ instead of $\psi_k(\cdot)$. By Markov's inequality and Lemma 4,

$$\Pr(\max_{1\leq j\leq r_n} |\hat{\boldsymbol{\alpha}}_n - \boldsymbol{\alpha}^*|_j \geq r_n^{-q}) \leq [\exp(\frac{1}{r_n^q \log(1+r_n) O_p(1/\sqrt{n})}) - 1]^{-1} = o_p(1).$$

This concludes the proof. □

It is noted that the assumption D.1 of elementwise $\sqrt{n}$-consistency of $\hat{\boldsymbol{\alpha}}$ is critical in Theorems 6 and 7. Since any shrinkage parameter q of slightly greater than 1 can be chosen, the conditions on r_n may be viewed as $r_n = o(n^{\frac{1}{2}-\Delta})$, where Δ indicates how many finite moments $(\hat{\boldsymbol{\alpha}}_n)_j$ has.

7 Concluding Remarks

In this paper we consider the asymptotic setting for GEE in which both m and r increase with n and present large cluster asymptotics for GEE. The standard inference approaches for GEE models (such as the normal approximation and the use of sandwich variance estimate) are shown to be valid unless the working correlation model is extremely overparameterized. Also we obtain the following sufficient conditions on the increasing rates of m and r:

$$\begin{aligned} m &= o(n^{1-\delta^*}), \\ r &= o(n^{\frac{1}{2}-\Delta}), \end{aligned}$$

where δ^* depends on the moment conditions of response measurements Y and Δ indicates how many finite moments $(\hat{\boldsymbol{\alpha}}_n)_j$ has.

APPENDIX

A Proof of Theorem 1

The proof is based on that of Theorem 2 in [9]. Let

$$E_n = \{\omega : ||H_{n\boldsymbol{\alpha}_n}^{-1/2} g_{n\boldsymbol{\alpha}_n}|| \leq \inf_{\boldsymbol{\beta} \in \partial B_{n\boldsymbol{\alpha}_n}(r)} ||H_{n\boldsymbol{\alpha}_n}^{-1/2}(g_{n\boldsymbol{\alpha}_n}(\boldsymbol{\beta}) - g_{n\boldsymbol{\alpha}_n})||\},$$

where $\partial B_{n\boldsymbol{\alpha}_n}(r)$ is the boundary of the sphere $B_{n\boldsymbol{\alpha}_n}(r)$. When considering a sequence of injective functions $H_{n\boldsymbol{\alpha}_n}^{-1/2} g_{n\boldsymbol{\alpha}_n}(\boldsymbol{\beta})$ of $\boldsymbol{\beta}$, we notice that there exists a $\hat{\boldsymbol{\beta}}_n \in B_{n\boldsymbol{\alpha}_n}(r)$ on the set E_n such that $g_{n\boldsymbol{\alpha}_n}(\hat{\boldsymbol{\beta}}_n) = 0$. We first show that, for any small $\epsilon > 0$, $\Pr(E_n) > 1 - \epsilon$ for large n. Then the proof is completed by showing that, for large n, on E_n we have $\Pr(||\hat{\boldsymbol{\beta}}_n(\omega) - \boldsymbol{\beta}_0|| < \epsilon) > 1 - \epsilon$.

By Taylor's expansion around $\boldsymbol{\beta}_0$,

$$H_{n\boldsymbol{\alpha}_n}^{-1/2}(g_{n\boldsymbol{\alpha}_n}(\boldsymbol{\beta}) - g_{n\boldsymbol{\alpha}_n}) = H_{n\boldsymbol{\alpha}_n}^{-1/2}\mathcal{D}_{n\boldsymbol{\alpha}_n}(\bar{\boldsymbol{\beta}})(\boldsymbol{\beta} - \boldsymbol{\beta}_0) = H_{n\boldsymbol{\alpha}_n}^{-1/2}\mathcal{D}_{n\boldsymbol{\alpha}_n}(\bar{\boldsymbol{\beta}})H_{n\boldsymbol{\alpha}_n}^{-1/2}H_{n\boldsymbol{\alpha}_n}^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0),$$

where $\bar{\boldsymbol{\beta}}$ lies between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$. In particular, when $\boldsymbol{\beta} \in \partial B_{n\boldsymbol{\alpha}_n}(r)$,

$$||H_{n\boldsymbol{\alpha}_n}^{-1/2}(g_{n\boldsymbol{\alpha}_n}(\boldsymbol{\beta}) - g_{n\boldsymbol{\alpha}_n})|| \geq \sqrt{z_{n,\lambda}}\ ||H_{n\boldsymbol{\alpha}_n}^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)|| = r\sqrt{z_{n,\lambda}}\sqrt{\tau_{n\boldsymbol{\alpha}_n}},$$

with $z_{n,\lambda} = z_{n,\lambda}(\omega) = \lambda_{\min}(H_{n\boldsymbol{\alpha}_n}^{-1/2}\mathcal{D}_{n\boldsymbol{\alpha}_n}^t(\bar{\boldsymbol{\beta}})H_{n\boldsymbol{\alpha}_n}^{-1}\mathcal{D}_{n\boldsymbol{\alpha}_n}(\bar{\boldsymbol{\beta}})H_{n\boldsymbol{\alpha}_n}^{-1/2})$. By the condition B.2, we can choose $c_0 > 0$ such that $\Pr(z_{n,\lambda} > c_0) > 1 - \epsilon/2$ for large n. Therefore,

$$\begin{aligned}
\Pr(E_n) &\geq \Pr(||H_{n\boldsymbol{\alpha}_n}^{-1/2}g_{n\boldsymbol{\alpha}_n}|| \leq r\sqrt{z_{n,\lambda}}\sqrt{\tau_{n\boldsymbol{\alpha}_n}}) \\
&\geq \Pr(\{z_{n,\lambda} > c_0\} \cap \{||H_{n\boldsymbol{\alpha}_n}^{-1/2}g_{n\boldsymbol{\alpha}_n}|| \leq r\sqrt{c_0}\sqrt{\tau_{n\boldsymbol{\alpha}_n}}\}) \\
&\geq \Pr(z_{n,\lambda} > c_0) + \Pr(||H_{n\boldsymbol{\alpha}_n}^{-1/2}g_{n\boldsymbol{\alpha}_n}|| \leq r\sqrt{c_0}\sqrt{\tau_{n\boldsymbol{\alpha}_n}}) - 1 \\
&\geq \Pr(z_{n,\lambda} > c_0) + 1 - \frac{\mathrm{E}||H_{n\boldsymbol{\alpha}_n}^{-1/2}g_{n\boldsymbol{\alpha}_n}||^2}{\tau_{n\boldsymbol{\alpha}_n}c_0r^2} - 1 \\
&= \Pr(z_{n,\lambda} > c_0) - \frac{\mathrm{tr}(H_{n\boldsymbol{\alpha}_n}^{-1}M_{n\boldsymbol{\alpha}_n})}{\tau_{n\boldsymbol{\alpha}_n}c_0r^2} \\
&\geq \Pr(z_{n,\lambda} > c_0) - \frac{p}{c_0r^2}
\end{aligned}$$

Note that the last inequality above follows because $M_{n\boldsymbol{\alpha}} \leq \tau_{n\boldsymbol{\alpha}}H_{n\boldsymbol{\alpha}}$.

Thus, by taking $r = r(\epsilon) = \sqrt{\frac{2p}{c_0\epsilon}}$, we have $\Pr(E_n) > 1 - \epsilon$ for sufficiently large n. The weak consistency of the GEE estimator $\hat{\boldsymbol{\beta}}_n \in B_{n\boldsymbol{\alpha}_n}(r)$ on the set E_n follows immediately, since the condition B.1 implies

$$B_{n\boldsymbol{\alpha}_n}(r) \subset \{\boldsymbol{\beta} : ||\boldsymbol{\beta} - \boldsymbol{\beta}_0|| \leq r\sqrt{\frac{\tau_{n\boldsymbol{\alpha}_n}}{\lambda_{\min}(H_{n\boldsymbol{\alpha}_n})}}\} \subset \{\boldsymbol{\beta} : ||\boldsymbol{\beta} - \boldsymbol{\beta}_0|| \leq \epsilon\},$$

for large n.

B Proof of Lemma 2

Since T is defined as the Cartesian product $T(\boldsymbol{\beta}_0) \times T(\boldsymbol{\alpha}^*; q)$ with the metric

$$d_T((\boldsymbol{\beta}_1, \boldsymbol{\alpha}_1), (\boldsymbol{\beta}_2, \boldsymbol{\alpha}_2)) = \sqrt{d_1^2(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) + d_2^2(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)},$$

we have

$$\log N(2\epsilon, T, d_T) \leq \log N(\sqrt{2}\epsilon, T(\boldsymbol{\beta}_0), d_1) + \log N(\sqrt{2}\epsilon, T(\boldsymbol{\alpha}^*; q), d_2). \quad (8)$$

A bound for the first term in the right hand side of (8) is given as:

$$N(\sqrt{2}\epsilon, T(\boldsymbol{\beta}_0), d_1) \leq C_1 \times (\frac{1}{\epsilon})^p, \quad (9)$$

with a constant C_1 depending only on p and the diameter of $T(\boldsymbol{\beta}_0)$.

A bound for the second term in the right hand side of (8) is obtained by approximating $T(\boldsymbol{\alpha}^*; q)$ into a finite-dimensional Euclidean space. Note that

$$\log N(\sqrt{2}\epsilon, T(\boldsymbol{\alpha}^*; q), d_2) \leq \log N(\epsilon, T_{k_0(\epsilon)}(\boldsymbol{\alpha}^*; q), d_2),$$

where $k_0(\epsilon) = \inf\{k \in \mathbb{Z}^+; \sum_{i>k} i^{-2q} < \epsilon^2\}$, and $T_k(\boldsymbol{\alpha}^*; q)$ is the projection of $T(\boldsymbol{\alpha}^*; q)$ onto $R^k \times \{0\} \times \{0\} \times \cdots$. Also note that $k_0(\epsilon) \approx \epsilon^{-\frac{1}{q-0.5}}$. Since a maximum length of hypercube in $\mathbb{R}^k$ contained in a ball of radius ϵ is $\frac{2\epsilon}{\sqrt{k}}$, we have

$$\begin{aligned}
N(\epsilon, T_{k_0(\epsilon)}(\boldsymbol{\alpha}^*; q), d_2) &\leq \prod_{i=1}^{k_0(\epsilon)} \frac{i^{-q}}{\epsilon/\sqrt{k_0(\epsilon)}} \\
&= \frac{1}{\{k_0(\epsilon)!\}^q} \times \{\frac{\sqrt{k_0(\epsilon)}}{\epsilon}\}^{k_0(\epsilon)} \\
&\approx \{\frac{1}{\sqrt{2\pi k_0(\epsilon)}}\}^q \times \{\frac{e}{k_0(\epsilon)}\}^{q \cdot k_0(\epsilon)} \times \{\frac{\sqrt{k_0(\epsilon)}}{\epsilon}\}^{k_0(\epsilon)} \\
&= \{2\pi k_0(\epsilon)\}^{-q/2} \times \{\frac{e^q}{\epsilon \cdot k_0(\epsilon)^{q-0.5}}\}^{k_0(\epsilon)} \\
&\approx \{2\pi k_0(\epsilon)\}^{-q/2} \times (e^q)^{k_0(\epsilon)} \\
&\approx \{2\pi\}^{-q/2} \times (\frac{1}{\epsilon})^{-\frac{q}{2q-1}} \times \exp\{q(\frac{1}{\epsilon})^{\frac{1}{q-0.5}}\}. \qquad (10)
\end{aligned}$$

Stirling's approximation to the factorial is used in (10). The proof is completed by combining (8), (9) and (10).

C Proof of Theorem 5

We work with $A_n(\boldsymbol{\beta}, \boldsymbol{\alpha})$ and $B_n(\boldsymbol{\beta}, \boldsymbol{\alpha})$ separately.

We first consider $A_n(\boldsymbol{\beta}, \boldsymbol{\alpha})$. Let $\mathcal{A}_n$ be the class of functions $\{A_n(\boldsymbol{\beta}, \boldsymbol{\alpha}) : (\boldsymbol{\beta}, \boldsymbol{\alpha}) \in T\}$, and let $\bar{A}_n$ be the corresponding envelope function in the form of sum of n independent, but not necessarily identically distributed, functions (i.e., $\bar{A}_n = M_{n\boldsymbol{\alpha}^*}^{-1/2} \sum_{i=1}^n A_{ni}$). From the condition C.2, we can assume that $\mathrm{E}\bar{A}_n^{2+\delta} = o(1)$ for $\delta > 0$ in the condition C.2, which immediately yields the weak convergence of $A_n(\boldsymbol{\beta}, \boldsymbol{\alpha})$ for each $(\boldsymbol{\beta}, \boldsymbol{\alpha}) \in T$. Note that, by the same arguments as in Lemma 1, the condition (4a) implies that $\sup_n N_{[\,]}(\epsilon, \mathcal{A}_n, L_1(P)) < \infty$. It then follows from the standard proof of the

Glivenko-Cantelli theorem (Theorem 2.4.1 in [7]) that

$$\sup_{(\boldsymbol{\beta},\boldsymbol{\alpha})\in T} |A_n(\boldsymbol{\beta},\boldsymbol{\alpha}) - \mathrm{E}A_n(\boldsymbol{\beta},\boldsymbol{\alpha})| \xrightarrow{P} 0.$$

Since the condition (4a) also implies the uniform equicontinuity of $\mathrm{E}A_n(\boldsymbol{\beta},\boldsymbol{\alpha})$, we have

$$|A_n(\hat{\boldsymbol{\beta}}_n,\boldsymbol{\alpha}_n) - \mathrm{E}A_n(\boldsymbol{\beta}_0,\boldsymbol{\alpha}^*)| \xrightarrow{P} 0. \tag{11}$$

Note that

$$\mathrm{E}A_n(\boldsymbol{\beta}_0,\boldsymbol{\alpha}) = \mathrm{E}M_{n\boldsymbol{\alpha}}^{-1/2}\left(\frac{\partial}{\partial\boldsymbol{\beta}}g_{n\boldsymbol{\alpha}}(\boldsymbol{\beta})\right)_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} = \left(\frac{\partial}{\partial\boldsymbol{\beta}}M_{n\boldsymbol{\alpha}}^{-1/2}\mathrm{E}g_{n\boldsymbol{\alpha}}(\boldsymbol{\beta})\right)_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} = \dot{\phi}_n(\boldsymbol{\beta}_0,\boldsymbol{\alpha}),$$

where the interchange of differentiation and integration is justified by the integrability of the envelope function $\bar{A}_n$.
We now consider $B_n(\boldsymbol{\beta},\boldsymbol{\alpha})$. In a manner analogous to the above, we define the corresponding class $\mathcal{B}_n = \{B_n(\boldsymbol{\beta},\boldsymbol{\alpha}) : (\boldsymbol{\beta},\boldsymbol{\alpha}) \in T\}$ and the envelope function $\bar{B}_n$ for $\mathcal{B}_n$. Note that $\mathrm{E}\bar{B}_n^{1+\delta/2} = o(1)$ for $\delta > 0$ in the condition C.2, and hence that $|B_n(\boldsymbol{\beta},\boldsymbol{\alpha}) - \mathrm{E}B_n(\boldsymbol{\beta},\boldsymbol{\alpha})| \xrightarrow{P} 0$ for each $(\boldsymbol{\beta},\boldsymbol{\alpha}) \in T$. Under the condition (4b), $\mathcal{B}_n$ is Glivenko-Cantelli and $\mathrm{E}B_n(\boldsymbol{\beta},\boldsymbol{\alpha})$ is uniformly equicontinuous. Therefore, we have

$$|B_n(\hat{\boldsymbol{\beta}}_n,\boldsymbol{\alpha}_n) - \mathrm{E}B_n(\boldsymbol{\beta}_0,\boldsymbol{\alpha}^*)| \xrightarrow{P} 0. \tag{12}$$

Note that $\mathrm{E}B_n(\boldsymbol{\beta}_0,\boldsymbol{\alpha}) = \mathrm{I}_p$.
The proof is completed by combining (11) and (12).

References

[1] Crowder, M. (1995). On the use of a working correlation matrix in using generalised linear models for repeated measures. *Biometrika* **82**, 2, 407–410.

[2] Gouriéroux, C., Monfort, A., and Trognon, A. (1984). Pseudo-maximum likelihood methods: theory. *Econometrica* **52**, 3, 681–700.

[3] Huber, P. J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. *Ann. Statist.* **1**, 5, 799–821.

[4] Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 1, 13–22.

[5] Lumley, T. and Mayer-Hamblett, N. (2003). Asymptotics for marginal generalized linear models with sparse correlations. *UW Biostatistics Working Paper Series*, Working Paper 207. http://www.bepress.com/uwbiostat/paper207/.

[6] Schildcrout, J. and Heagerty, P. (2005). Regression analysis of longitudinal binary data with time-dependent environmental covariates: bias and efficiency. *Biostatistics* **6**, 4, 633–652.

[7] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes.* Springer-Verlag, New York.

[8] White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1, 1–25.

[9] Xie, M. and Yang, Y. (2003). Asymptotics for generalized estimating equations with large cluster sizes. *Ann. Statist.* **31**, 1, 310–347.

[10] Yu, O., Sheppard, L., Lumley, T., Koenig, J., and Shapiro, G. (2000). Effects of ambient air pollution on symptoms of asthma in seattle-area children enrolled in the camp study. *Environmental Health Perspectives 108*, 1209–1214.