Empirical Likelihood for Generalized Linear Models with Longitudinal Data*

YIN Changming · AI Mingyao · CHEN Xia · KONG Xiangshun

DOI: 10.1007/s11424-023-2022-2
Received: 24 September 2020 / Revised: 26 January 2022

Abstract Generalized linear models are usually adopted to model the discrete or nonnegative responses. In this paper, empirical likelihood inference for fixed design generalized linear models with longitudinal data is investigated. Under some mild conditions, the consistency and asymptotic normality of the maximum empirical likelihood estimator are established, and the asymptotic χ^2 distribution of the empirical log-likelihood ratio is also obtained. Compared with the existing results, the new conditions are more weak and easy to verify. Some simulations are presented to illustrate these asymptotic properties.

Keywords Empirical likelihood ratio, generalized linear model, longitudinal data, maximum empirical likelihood estimator.

1 Introduction

The analysis of longitudinal (clustered or panel) data is an important issue in biomedical, economical and social science studies. Under these backgrounds, repeated measurements within the same cluster are usually considered as correlated while the measurements from different

YIN Changming
School of Mathematics and Information Science, Guangxi University, Nanning 530004*, China.*
Email: 2814294510@qq.com.
AI Mingyao (Corresponding author)
LMAM, School of Mathematical Sciences and Center for Statistical Science, Peking University, Beijing 100871*, China.* Email: myai@math.pku.edu.cn.
CHEN Xia
School of Mathematics and Statistics, Shaanxi Normal University, Xi'an 710062*, China.*
Email: xchen80@snnu.edu.cn.
KONG Xiangshun
School of Mathematics and Statistics, Beijing Institute of Technology, Beijing 100081*, China.*
Email: xskong@bit.edu.cn.
*This work was supported by the Natural Science Foundation of China under Grant Nos. 12031016, 11061002, 11801033, 12071014 and 12131001, the National Social Science Fund of China under Grant No. 19ZDA121, the Natural Science Foundation of Guangxi under Grant Nos. 2015GXNSFAA139006 and LMEQF.
◇ *This paper was recommended for publication by Editor LI Qizhai.*

clusters are assumed to be independent. For example, in designing continuous investigations, the subjects are measured repeatedly over a given period of time, thus the measurements from the same subject are often correlated and form a cluster[1]. In a longitudinal study of the air pollution effect on the health of children, repeated categorical data was collected at ages 9, 10, 11, and 12 years. In a randomized comparative study of alternative treatments of patients with rheumatoid arthritis, repeated observations of self-assessment of arthritis were collected at one-, three- and five-month examination[2].

A popular approach to analyzing continuous or discrete clustered data is the generalized estimating equations (GEE) developed by Liang and Zeger[3] for generalized linear models (GLMs)[4]. The GEE method involves a working correlation matrix and gives consistent estimators even if the correlation matrix is misspecified. We refer to [1, 3, 5–8] for more details. However, the efficiency of parameter estimator in GEE is sensitive to the working correlation matrix.

Besides GEE method, another approach is the Empirical likelihood (EL) originally proposed by Owen[9]. It has become a state-of-art method with many advantages. For example, compared with the plug-in estimation based on the asymptotic variance in GEE, the shape and the orientation of the EL-based confidence regions are determined automatically and entirely by the data, the EL approach yields better coverage probability for small sample[10], and the EL approach is Bartlett correctable[11]. In the context of GLMs, the EL method has been studied in many researches including but not limited to [12–18]. Chen and Cui[12] introduced an extended EL method in analyzing GLMs by incorporating extra constraints. The extended EL method reduces the variance of the quasi-likelihood based parameter estimators. Xue, et al.[13] considered the EL inference for GLMs with missing data. Kolaczyk[14] studied the EL inference for GLMs with fixed designs, and Yan and Chen[15] further considered the EL inference for GLMs with fixed and adaptive designs, Qin and Lawless[16] studied the EL for general estimating equations with i.i.d. data, Bai, et al.[17], Li and Pan[18] obtained the EL inference for GLMs with longitudinal data. Their estimation procedure for longitudinal data uses a linear approximation to the inverse of the correlation matrix and empirical likelihood, and does not require estimating the covariance matrices.

For fixed design GLMs with longitudinal data or nonlongitudinal data, some of the above papers, such as [17], only studied the large sample properties of EL ratio, but not of maximum EL estimators (MELEs). Furthermore, the conditions that support the asymptotic results are either very strong or hard to verify in practice. In this paper, under some milder conditions than before, we establish the large sample properties of both EL ratio and, more importantly, the MELEs of the regression parameters in a GLM with longitudinal data. Compared with GEE, the confidence interval given by the new approach is data driven and so excludes any plug-in variance estimation. Being different from the EL approach in [18] with a linear approximation of the working correlation matrix, the new approach accommodates arbitrary working correlation matrix. Meanwhile, the theoretical results are derived under milder assumptions.

The rest of the paper is organized as follows. In Section 2, the definitions of the EL ratio and MELEs are reviewed for generalized estimating equations. The large sample theories for both the EL ratio and MELEs are presented in Section 3 as the main results. Some simulations are shown in Section 4 to verify the theoretical properties. The detailed proofs of the previous results are delegated to the Appendix.

2 EL Ratio and Maximum EL Estimator

Suppose (Y_{ij}, X_{ij}) are observations for the jth measurement on the ith subject, $i = 1, 2, \cdots, n$ and $j = 1, 2, \cdots, m_i$, where Y_{ij} is a scalar response, X_{ij} is a $p \times 1$ covariate vector, and m_i is the cluster size. Assume that the observations on different subjects are independent and the observations on the same subject are correlated. For $i = 1, \cdots, n$, let $Y_i = (Y_{i1}, \cdots, Y_{im_i})^\tau$ and $X_i = (X_{i1}, \cdots, X_{im_i})$, where τ denotes the transpose of a matrix or vector.

The first two marginal moments of Y_{ij} are given by

$$\begin{aligned} \mu_{ij}(\beta) &= E_\beta(Y_{ij}) = \mu(\theta_{ij}), \\ \sigma_{ij}(\beta) &= \mathrm{Var}_\beta(Y_{ij}) = \dot{\mu}(\theta_{ij}), \end{aligned}$$

where $\mu(\theta)$ is called link function and $\dot{\mu}(\theta) > 0$ is its derivative, $\theta_{ij} = X_{ij}^\tau \beta$, β is the regression parameter vector, that is, we only consider canonical link functions for simpleness. Here are some important canonical link functions. In the linear regression, we have $\mu(\theta) = \theta$. In the logistic regression for binary data, we have $\mu(\theta) = \mathrm{e}^\theta/(1 + \mathrm{e}^\theta)$. In the log regression for count data, we have $\mu(\theta) = \mathrm{e}^\theta$. We refer to [4] for more detailed discussion on GLM frameworks.

Let $\mu_i(\beta) = (\mu_{i1}(\beta), \cdots, \mu_{im_i}(\beta))^\tau$, and $A_i(\beta) = \mathrm{diag}(\sigma_{i1}(\beta), \cdots, \sigma_{im_i}(\beta))$, where $\mathrm{diag}(v)$ represents a diagonal matrix whose elements are the elements of v for any vector v. Liang and Zeger[3], and Xie and Yang[5] defined the GEE estimator $\widetilde{\beta}_n$ for GLMs with longitudinal data as the solution of $\sum_{i=1}^n s_i(\beta) = 0$, where

$$s_i(\beta) = X_i A_i^{1/2}(\beta) R_i^{-1} A_i^{-1/2}(\beta)(Y_i - \mu_i(\beta)), \quad i = 1, \cdots, n,$$

R_i is the working correlation matrix of Y_i.

The EM ratio function of β is defined by

$$l_n(\beta) := \max \left\{ \sum_{i=1}^n \log(np_i) | p_i \geq 0, \sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i s_i(\beta) = 0 \right\},$$

where the maximization is taken with respect to the probabilities $p_1, \cdots, p_n$[9]. By the Lagrange multiplier method,

$$l_n(\beta) = -\sum_{i=1}^n \log[1 + t^\tau(\beta) s_i(\beta)],$$

where $t = t(\beta)$ is determined by

$$Q_{1n}(\beta, t) := \sum_{i=1}^n s_i(\beta)/[1 + t^\tau(\beta) s_i(\beta)] = 0. \tag{1}$$

The MELE $\widehat{\beta}_n$ is defined as the maximizer of $l_n(\beta)$, i.e.,

$$\widehat{\beta}_n := \arg\max\Big\{-\sum_{i=1}^n \log[1+t^\tau(\beta)s_i(\beta)]\Big\}, \quad \text{subject to } \sum_{i=1}^n s_i(\beta)/[1+t^\tau(\beta)s_i(\beta)]=0.$$

3 Main Results

Let $C, C_1, C_2, \cdots$ be arbitrary positive constants independent of n, whose value may change from one expression to another. In order to obtain our results, we need the following conditions.

(A1) (i) $\underline{\lambda}_n \to \infty (n \to \infty)$,
(ii) $\|X_n\|^2/\underline{\lambda}_n \to 0 (n\to\infty)$,
(iii) $\sup_{i\geq 1} m_i < +\infty$,
where $\underline{\lambda}_n = \lambda_{\min}(\sum_{i=1}^n X_i X_i^\tau)$, $\|T\| = [\mathrm{trace}(TT^\tau)]^{1/2}$ denotes its Frobenius norm, and $\lambda_{\min}(T)$ (or $\lambda_{\max}(T)$) denotes the minimum (or maximum) eigenvalue of a matrix T.

(A2) (i) The unknown parameter vector β belongs to a compact subset $B \subset R^p$, the true value of β, denoted as β_0, lies in the interior of B.
(ii) $\inf_{i\geq 1, 1\leq j\leq m_i} \sigma_{ij}(\beta_0) > 0$.
(iii) There exists a $\varepsilon_0 > 0$ such that $\sup_{i\geq 1,1\leq j\leq m_i} E|e_{ij}|^{2+\varepsilon_0} < \infty$, where $e_{ij} = Y_{ij} - \mu_{ij}(\beta_0)$.

(A3) (i) The true correlation matrix R_{i0} and working correlation matrix R_i have eigenvalues bounded away from zero.
(ii) The second derivative $|\ddot{\mu}(\theta_{ij0})|$ and third derivative $|\mu^{(3)}(\theta_{ij})|$ of $\mu(\theta_{ij})$ are uniformly bounded by a finite positive constant C for all $1\leq i\leq n, 1\leq j\leq m_i$ and $\beta\in N_n(\Delta) = \{\beta : \|F_n^{1/2}(\beta-\beta_0)\| \leq \Delta\}$, where $\theta_{ij0} = X_{ij}^\tau\beta_0$, $F_n = \sum_{i=1}^n X_i X_i^\tau$ and Δ is any positive constant.

Note that Condition (A1) is easier to verify and milder than the regular conditions that (i) $n^{-1}\sum_{i=1}^n \|X_i\|^3 < C$, $n^{-1}M_n(\beta_0) \to K_0(W)$ and $n^{-1}H_n(\beta_0) \to \varOmega_0(W)$ in [17], and (ii) $n^{-1}\sum_{i=1}^n s_i(\beta_0)s_i^\tau(\beta_0) \to_{a.s} \varSigma_{11}$ and $n^{-1}\sum_{i=1}^n \partial s_i(\beta)/\partial\beta^\tau \to_p \varSigma_{12}$ in [18]. Here, $M_n(\beta_0)$ and $H_n(\beta_0)$ are defined in Theorem 3.2, $K_0(W), \varOmega_0(W)$ and $\varSigma_{11}$ are all constant positive matrices, and $\varSigma_{12}$ is a constant matrix. Moreover, Condition (A1) relaxes $C_1 n \leq \underline{\lambda}_n \leq \overline{\lambda}_n \leq C_2 n$ implied by the conditions in [14, 16–18], because the constraint on $\overline{\lambda}_n$ is not needed and $\underline{\lambda}_n(\overline{\lambda}_n)$ needs not to be the same order with the sample size n, where $\overline{\lambda}_n = \lambda_{\max}(\sum_{i=1}^n X_i X_i^\tau)$. We only need $\underline{\lambda}_n \to \infty$, which is certainly a minimum requirement on the Fisher information matrix for large sample properties of $\widehat{\beta}_n$. We also drop the condition that $\{X_i, i\geq 1\}$ is bounded in [15].

Conditions (A2) (i) and (A2) (ii) are common assumptions, see [7, 8, 14, 15]. (A2) (iii) is milder than $\sup_{i\geq 1, 1\leq j\leq m_i} E|e_{ij}|^3 < \infty$ in [18]. In the nonlongitudinal case with $m_i = 1, i\geq 1$, $E|e_{i1}|^3 < \infty$ is needed in [16] and $E|e_{i1}|^4 < \infty$ is needed in [14].

Recall that Condition (A3) is also a common assumption, see [6–8] for justification. If $\{X_i, i\geq 1\}$ is bounded, then the condition (A3) (ii) is satisfied. For $\mu(\theta)=\theta$ in the linear regression, and $\mu(\theta) = \mathrm{e}^\theta/(1+\mathrm{e}^\theta)$ in the logistic regression for binary data, the condition

(A3) (ii) is naturally satisfied. For $\mu(\theta) = e^\theta$ in the log regression for count data, $\mu(\theta) = \dot{\mu}(\theta)$, the condition (A3) (ii) can be replaced by $\{|\mu^{(3)}(\theta_{ij})| : 1 \le i \le n, 1 \le j \le m_i\}$ are uniformly bounded by a finite positive constant C on $N_n(\Delta)$.

Under the above conditions (A1)–(A3), we can obtain the following results.

Theorem 3.1 *Assume Conditions* (A1)–(A3) *hold. Then*

$$P(l_n(\beta) \text{ attains its maximum at some } \widehat{\beta}_n \in N_n(\delta)) \to 1, \tag{2}$$

and

$$\|\widehat{\beta}_n - \beta_0\| = O_p(\underline{\lambda}_n^{-1/2}). \tag{3}$$

Theorem 3.2 *Assume Conditions* (A1)–(A3) *hold. Then*

$$M_n^{-1/2}(\beta_0)H_n(\beta_0)(\widehat{\beta}_n - \beta_0) \to_d N(0, I),$$

where $M_n(\beta) = \sum_{i=1}^n X_i A_i^{1/2}(\beta) R_i^{-1} R_{i0} R_i^{-1} A_i^{1/2}(\beta)\, X_i^\tau$, $H_n(\beta) = \sum_{i=1}^n X_i A_i^{1/2}(\beta) R_i^{-1} A_i^{1/2}(\beta)$ X_i^τ.

Theorem 3.3 *Under Conditions* (A1)–(A3)*, the empirical likelihood ratio statistic for testing* $H_0 : \beta = \beta_0$,

$$2l_n(\widehat{\beta}_n) - 2l_n(\beta_0) \to_d \chi_p^2,$$

as $n \to \infty$*, when* H_0 *is true.*

From the equation (57) in the proof of Theorem 3.3, we obtain the following corollary, which coincides with the results in [14, 17].

Corollary 3.4 *Under Conditions* (A1)–(A3)*, as* $n \to \infty$*, we have*

$$-2l_n(\beta_0) \to_d \chi_p^2.$$

Remark 3.5 We do not require working correlation matrix R_i to be the true correlation matrix R_{i0} for $\widehat{\beta}_n$ to be asymptotic normal. If $R_i = R_{i0}$, then the $\widehat{\beta}_n$ has greater efficiency, that is,

$$[H_n(\beta_0)M_n^{-1}(\beta_0)H_n(\beta_0)]^{-1} \ge \left[\sum_{i=1}^n X_i \dot{\mu}_i(\beta_0) A_i^{-1/2}(\beta_0) R_{i0}^{-1} A_i^{-1/2}(\beta_0) \dot{\mu}_i^\tau(\beta_0) X_i^\tau\right]^{-1},$$

which can be proved by similar argument as in [20].

Remark 3.6 If $R_{i0} = R_{n0}, i = 1, \cdots, n$ as in [3], we can estimate the common true correlation matrix as follows:

$$\widehat{R}_n = \frac{1}{n}\sum_{i=1}^n A_i^{-1/2}(\widetilde{\beta}_n)\varepsilon_i(\widetilde{\beta}_n)\varepsilon_i^T(\widetilde{\beta}_n)A_i^{-1/2}(\widetilde{\beta}_n),$$

where $\varepsilon_i(\beta) = y_i - \mu(\beta), \widetilde{\beta}_n$ is the solution to the working independence equation

$$\sum_{i=1}^n X_i \varepsilon_i(\beta) = 0$$

as a preliminary estimator of β_0. Under Conditions (A1)–(A3), we have $\|\widehat{R}_n - R_{n0}\| = o_p(1)$, whose proof is similar as in [7] and so omitted here.

4 Numerical Studies

In this section, we carry out Monte-Carlo studies to assess the performance of the proposed approach in terms of the mean and standard error of the parameter estimator, and the coverage probability of confidence intervals. We then apply the proposed method to analyze Ohio Children wheeze status data introduced in [21]. A nested algorithm can be used to solve MELE, see Chapter 12 of [10] for more details.

Example 4.1 (Estimation of Parameters) We consider the correlated binary responses model for the marginal expectation of Y_{ij},

$$E(Y_{ij}|X_{ij}) = \frac{\mathrm{e}^{X_{ij}^{\tau}\beta_0}}{1+\mathrm{e}^{X_{ij}^{\tau}\beta_0}}, \quad i=1,\cdots,n, \quad j=1,\cdots,10.$$

Let $\beta_0 = (0.7, -0.7, 0.4)$, $X_{ij} = (X_{ij1}, X_{ij2}, X_{ij3})$, X_{ijk} has an independent uniform distribution $U(0,1)$. The binary response vector $Y_i = (Y_{i1}, \cdots, Y_{i9}, Y_{i10})^{\tau}$ has an exchangeable correlation structure with correlation coefficient 0.4, that is $R_{i0} = 0.4\overline{\mathbf{1}}_{\mathbf{10}} + 0.6I_{10}$, where I_m is the $m-$identity matrix and $\overline{\mathbf{1}}_{\mathbf{m}}$ is an $m \times m$ matrix with every element equals 1. We took sample sizes $n = 50, 100, 200$ and 400, respectively, under three different working correlation structures: exchangeable (exch) working correlation matrix, independence (indep) working correlation matrix and the first order autocorrelation (ar1) working correlation matrix, i.e., $R_i = (0.4^{|j-j'|})$.

First, we note that the response is binary thus contains much less information than a continuous response. Table 1 reports the mean and standard error of $\widehat{\beta}_n$ based on the EL and GEE approaches in 500 repetitions. Table 1 shows that the EL method and GEE method almost have the same mean and standard error in the estimator of the regression coefficient β. As sample size increases, the standard deviation for the estimator of β by those two approaches decreases. From Table 1, we also observe that when the true correlation matrix (exch in oure case) is adopted, the standard error of $\widehat{\beta}_n$ is a little bit smaller.

Table 2 shows that compared to the GEE approach, the empirical coverage probabilities for the confidence intervals given by the EL approach are a little more close to the nominal coverage probability 0.95. Table 2 demonstrates that when sample size increases, coverage probabilities tend to theoretical value 95%. We also observe that when the true correlation matrix (exch in our case) is adopted, the χ^2 approximation is bit more accurate.

Example 4.2 (Simulation Based on EL with Different m_i) Consider the following model which coincides with that used in Example 4.1.

$$E(Y_{ij}|X_{ij}) = \frac{\mathrm{e}^{X_{ij}^{\tau}\beta_0}}{1+\mathrm{e}^{X_{ij}^{\tau}\beta_0}}, \quad i=1,\cdots,n, \quad j=1,\cdots,m_i.$$

For the choice of different m_i, we assume $m_i = k$ $(k = 1, 2, \cdots, 10)$ for $i \in [n(k-1)/10 + 1, nk/10]$. It is shown in Table 5 that the empirical coverage probabilities for the confidence intervals given by the EL approach converge to the nominal coverage probability 0.95 as n increases.

Table 1 The parameter estimators and corresponding standard errors (in the parentheses)

n	Correlation	β_0	Mean (Standard error) of $\widehat{\beta}_n$ by EL	Mean (Standard error) of $\widehat{\beta}_n$ by GEE
50	exch	0.7	0.695103 (0.236169)	0.694902 (0.236249)
		−0.7	−0.705151 (0.232220)	−0.705350 (0.233210)
		0.4	0.394551 (0.243116)	0.394356 (0.243018)
	indep	0.7	0.695215 (0.298306)	0.695316 (0.298266)
		−0.7	−0.701320 (0.291291)	−0.701422 (0.291391)
		0.4	0.401101 (0.299684)	0.401211 (0.299724)
	ar1	0.7	0.697618 (0.255692)	0.697519 (0.255708)
		−0.7	−0.699395 (0.246601)	−0.699375 (0.246591)
		0.4	0.394696 (0.268394)	0.394494 (0.268424)
100	exch	0.7	0.704386 (0.165811)	0.704482 (0.165791)
		−0.7	−0.700224 (0.155099)	−0.700684 (0.155183)
		0.4	0.400215 (0.165784)	0.400249 (0.165800)
	indep	0.7	0.709301 (0.200361)	0.709646 (0.200381)
		−0.7	−0.700601 (0.188008)	−0.700514 (0.188013)
		0.4	0.398527 (0.206793)	0.398211 (0.206863)
	ar1	0.7	0.703902 (0.183399)	0.703919 (0.183409)
		−0.7	−0.694306 (0.168165)	−0.694204 (0.168295)
		0.4	0.399512 (0.179389)	0.399402 (0.179398)
200	exch	0.7	0.710112 (0.120771)	0.710402 (0.120861)
		−0.7	−0.701920 (0.121216)	−0.701919 (0.121304)
		0.4	0.398087 (0.116172)	0.398088 (0.116172)
	indep	0.7	0.715950 (0.151562)	0.715946 (0.151642)
		−0.7	−0.702503 (0.146539)	−0.702442 (0.146559)
		0.4	0.400481 (0.148198)	0.400475 (0.148238)
	ar1	0.7	0.710766 (0.132933)	0.710838 (0.132963)
		−0.7	−0.700301 (0.134907)	−0.700243 (0.134897)
		0.4	0.401151 (0.130486)	0.401148 (0.130562)
400	exch	0.7	0.699192 (0.0840211)	0.699252 (0.0841225)
		−0.7	−0.695539 (0.0831322)	−0.695551 (0.0831432)
		0.4	0.403931 (0.0816547)	0.403900 (0.0817557)
	indep	0.7	0.699298 (0.101493)	0.699300 (0.101583)
		−0.7	−0.692858 (0.100865)	−0.692873 (0.100788)
		0.4	0.402924 (0.097482)	0.402874 (0.097546)
	ar1	0.7	0.699305 (0.0901975)	0.699314 (0.0901985)
		−0.7	−0.695632 (0.0885271)	−0.695749 (0.0885269)
		0.4	0.404442 (0.0856206)	0.404473 (0.0857946)

Example 4.3 Consider a simple regressive model adapted from Subsection 3.2 of [**?**].

$$y_i = \mu(X_i\beta_0) + \sigma^{1/2}(X_i\beta_0)e_i, \tag{4}$$

where y_i and X_i are both m_i-dimensional vector, $\beta_0 = 1$, $\mu(\theta) = \sin\theta$, $\sigma(\theta)$ is $m_i \times m_i$ matrix. We assume $\sigma(X_i\beta_0) = A^{1/2}(X_i\beta_0)R_0A^{1/2}(X_i\beta_0)$ with $A(X_i\beta_0) = \mathrm{diag}(\sigma_{i1}^2, \cdots, \sigma_{im_i}^2)$ and $\sigma_{ij} = 2\cos(x_{ij\beta_0})$. R_0 is the true correlation matrix and $e_i \sim U(-\sqrt{3}, \sqrt{3})$. We compare three different methods (i) MELE in this manuscript, (ii) GEE introduced in [3] and (iii) the linear approximation based EL method introduced in [18], denoted by LAEL for convenience. The simulation uses 500 repetitions to compare the cover rate and the length of 95% confidence interval given by these three methods. The working correlation matrix (WCM) is chosen same as that in Example 4.1. For unified $m_i = 10$, the results are listed in Table 3. For different m_i of the same setup in Example 4.2, the results are listed in Table 4.

Table 2 Cover rate of the estimated 95% confidence interval, $m_i = 10$, $i = 1, \cdots, n$

Method	Correlation	$n = 50$	$n = 100$	$n = 200$	$n = 400$
EL	exch	0.918	0.960	0.954	0.950
	indep	0.898	0.940	0.958	0.956
	ar1	0.910	0.968	0.938	0.948
GEE	exch	0.905	0.940	0.955	0.950
	indep	0.893	0.942	0.956	0.942
	ar1	0.896	0.933	0.937	0.953

Table 3 The cover rate and the length of 95% confidence interval given by three methods with unified m_i. The parentheses show the (cover rate, length)

method	WCM	$n = 50$	$n = 100$	$n = 200$	$n = 400$
MELE	exch	(0.960, 0.37544)	(0.948, 0.25368)	(0.956, 0.17392)	(0.952, 0.12006)
	ar1	(0.964, 0.42512)	(0.956, 0.28658)	(0.954, 0.19776)	(0.940, 0.13702)
	indep	(0.954, 0.50636)	(0.964, 0.34702)	(0.956, 0.24036)	(0.954, 0.16716)
GEE	exch	(0.908, 0.34172)	(0.916, 0.23842)	(0.916, 0.16612)	(0.894, 0.11472)
	ar1	(0.892, 0.39336)	(0.920, 0.27648)	(0.920, 0.19158)	(0.900, 0.13200)
	indep	(0.928, 0.48706)	(0.964, 0.34148)	(0.956, 0.23860)	(0.944, 0.16590)
LAEL	exch	(0.842, 0.29656)	(0.876, 0.21110)	(0.856, 0.14102)	(0.862, 0.10018)
	ar1	(0.814, 0.32896)	(0.866, 0.23274)	(0.854, 0.15922)	(0.836, 0.11146)
	indep	(0.954, 0.50636)	(0.964, 0.34702)	(0.956, 0.24036)	(0.954, 0.16716)

Table 4 The cover rate and the length of 95% confidence interval given by three methods with different m_i. The parentheses show the (cover rate, length)

method	WCM	$n = 50$	$n = 100$	$n = 200$	$n = 400$
MELE	exch	(0.946, 0.47308)	(0.938, 0.32090)	(0.952, 0.22022)	(0.950, 0.15266)
	ar1	(0.952, 0.51060)	(0.946, 0.34988)	(0.958, 0.24090)	(0.954, 0.16652)
	indep	(0.952, 0.58358)	(0.964, 0.41020)	(0.968, 0.28430)	(0.946, 0.19724)
GEE	exch	(0.928, 0.43230)	(0.926, 0.30266)	(0.934, 0.21202)	(0.930, 0.14688)
	ar1	(0.882, 0.47148)	(0.898, 0.33466)	(0.936, 0.23598)	(0.922, 0.16088)
	indep	(0.944, 0.56866)	(0.930, 0.40200)	(0.952, 0.28094)	(0.952, 0.19622)
LAEL	exch	(0.806, 0.38618)	(0.852, 0.27626)	(0.860, 0.19848)	(0.862, 0.13706)
	ar1	(0.854, 0.40214)	(0.866, 0.28462)	(0.866, 0.20052)	(0.834, 0.13394)
	indep	(0.952, 0.58358)	(0.964, 0.41020)	(0.968, 0.28430)	(0.946, 0.19724)

Example 4.4 (Analysis of Ohio Children Wheeze Status Data) We apply the following logistic model to analyze Ohio Children wheeze status data[21].

$$E(Y_{ij}|X_{ij}) = \frac{\mathrm{e}^{\theta_{ij}}}{1+\mathrm{e}^{\theta_{ij}}}, \quad i = 1, \cdots, 537, \quad j = 1, 2, 3, 4.$$

where $\theta_{ij} = \beta_1 + \beta_2 X_{ij2} + \beta_3 X_{ij3} + \beta_4 X_{ij2} X_{ij3}$, Y_{ij} is a binary outcome with 0 and 1, indicating the absence or presence of respiratory illness. X_{ij2}, X_{ij3} and $X_{ij2}X_{ij3}$ are the age of the child, the maternal smoking habit indicator and their interaction, respectively. Table 6 shows that by EL method, the estimators of the regression parameters are very similar by EL method under the exch, indep and ar1 (see Example 4.1) working correlation structures. The p-value shows that age is a significant factor and should be included in the model. The negative sign for age means that older children are less likely to have respiratory disease. Similarly, the maternal smoking habit has a positive effect on children respiratory disease, although smoking habit seems not to be a significant factor. The p-value also suggests that the interaction between the age of the child and maternal smoking is not significant. Similar to Example 4.3, we also compare MELE with GEE and LAEL.

Table 5 95% CP, $m_i = k$ $(k = 1, 2, \cdots, 10)$ for $i \in [n(k-1)/10 + 1, nk/10]$

Correlation	$n = 50$	$n = 100$	$n = 200$	$n = 400$
exch	0.920	0.930	0.958	0.954
indep	0.892	0.930	0.960	0.958
ar1	0.892	0.908	0.964	0.958

Table 6 Analysis of Ohio Children wheeze status data

method	WCM	Parameter	Estimation	Standard errpr	p-value
MELE	exch	β_1	−1.8640178	0.27247829	0.0000000
		β_2	−0.1137681	0.07723437	0.1407445
		β_3	0.3336982	0.13748979	0.0152211
		β_4	0.1174116	0.16189600	0.4683118
	ar1	β_1	−1.89259681	0.2175773	0.0000000
		β_2	−0.09786787	0.1038217	0.3458583
		β_3	0.32173701	0.1893112	0.0892224
		β_4	0.09742312	0.1520196	0.5216143
	indep	β_1	−1.87153648	0.09704457	0.0000000
		β_2	−0.10344292	0.07777718	0.1835214
		β_3	0.33065656	0.05411684	0.0000000
		β_4	0.09748281	0.05765975	0.0909025
GEE	exch	β_1	−1.83203430	0.38386768	0.0000018
		β_2	−0.09216941	0.18015046	0.6089138
		β_3	0.33769548	0.07322155	0.0000040
		β_4	0.07756037	0.25705256	0.7628582
	ar1	β_1	−1.9419207	0.2318928	0.0000000
		β_2	−0.1363242	0.1410585	0.3338254
		β_3	0.3592722	0.1135207	0.0015518
		β_4	0.1648687	0.1809492	0.3622256
	indep	β_1	−1.8852236	0.13875206	0.0000000
		β_2	−0.1143024	0.08096646	0.1580309
		β_3	0.3466189	0.08062294	0.0000171
		β_4	0.1175401	0.13815067	0.3948744
LAEL	exch	β_1	−1.8924038	0.11560057	0.0000000
		β_2	−0.1260650	0.05862409	0.0315239
		β_3	0.3500669	0.22937335	0.1269630
		β_4	0.1399492	0.23956350	0.5590964
	ar1	β_1	−1.75731598	0.4680985	0.0001739
		β_2	0.08675113	0.3195274	0.7860078
		β_3	0.26812517	0.7517423	0.7213376
		β_4	−0.23085026	0.5978775	0.6994105
	indep	β_1	−1.87153648	0.09704457	0.0000000
		β_2	−0.10344292	0.07777718	0.1835214
		β_3	0.33065656	0.05411684	0.0000000
		β_4	0.09748281	0.05765975	0.0909025

5 Discussions

To keep notation simple, we only consider the empirical likelihood for generalized linear models with natural link functions, which are important link functions for GLMs. This method can be extended to generalized estimating equations with general link functions, and the derivation is more delicate. Further, we can study EM for high-dimensional GLMs with longitudinal data, that is, either $m_n = \sup_{i \le n} m_i$ or the dimension of covariate $p = p_n$ or both go to infinity, as simple size $n \to \infty$.

References

[1] Diggle P, Heagerty P, Liang K, et al., *Analysis of Longitudinal Data*, Second Edition, Oxford University Press, Oxford, 2002.

[2] Lipsitz S R, Kim K, and Zhao L, Analysis of repeated categorical data using generalized estimating equations, *Statist Med*, 1994, **13**: 1149–1163.

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

[4] Fahrmeir L and Tutz G, *Multivariate Statistical Modeling Based on Generalized Linear Models*, Springer-Verlag, New York, 1994.

[5] Xie M G and Yang Y N, Asymptotic for generalized estimating equations with large cluster sizes, *Ann. Statist*, 2003, **31**: 310–347.

[6] Balan R M and Schiopu-Kratina I, Asymptotic results with generalized estimating equations for longitudinal data, *Ann. Statist*, 2005, **33**: 522–541.

[7] Wang L, GEE analysis of clustered binary data with diverging number of covariates, *Ann. Statist*, 2011, **39**: 389–417.

[8] Wang L, Zhou J, and Qu A, Penalized generalized estimating equations for highdimensional longitudinal data analysis, *Biometrics*, 2012, **68**: 353–360.

[9] Owen A B, Empirical likelihood ratio confidence intervals for a single functional, *Biometrika*, 1988, **75**(1): 237–249.

[10] Owen A B, *Empirical Likelihood*, Chapman and Hall, New York, 2001.

[11] DiCiccio T, Hall P, and Romano J, Empirical likelihood is Bartlett-correctable, *Ann. Statist*, 1991, **19**: 1053–1061.

[12] Chen S X and Cui H J, An extended empirical likelihood for generalized linear models, *Statist Sinica*, 2003, **13**: 69–81.

[13] Xue D, Xue L G, and Cheng W H, Empirical likelihood for generalized linear models with missing responses. *J. Statist Plann Inference*, 2011, **141**: 2007–2020.

[14] Kolaczyk E D, Empirical likelihood for generalized linear models, *Statist Sinica*, 1994, **4**: 199–218.

[15] Yan L and Chen X, Empirical likelihood for generalized linear models with fixed and adaptive designs, *Statistics*, 2015, **49**(5): 978–988.

[16] Qin J and Lawless J, Empirical likelihood and general estimating equations, *Ann. Statist*, 1994, **22**(1): 300–325.

[17] Bai Y, Fung W K, and Zhu Z Y, Weighted empirical likelihood for generalized linear models with longitudinaldata, *J. Statist Plann Inference*, 2010, **140**: 3446–3456.

[18] Li D J and Pan J X, Empirical likelihood for generalized linear models with longitudinal data, *Journal of Multivariate Analysis*, 2013, **114**(1): 63–73.

[19] Shao J, *Mathematical Statistics*, Second edition, Springer, New York, 1988, 338–340.

[20] Yue L and Chen X R, Asymptotic normality of quasi maximum likelihood estimates in generalized linear models, *Chin. Ann. Math. Ser. B*, 2005, **26**: 467–474.

[21] Fitzmaurice G and Laird N, A likelihood-based method for analysing longitudinal binary responses, *Biometrika*, 1993, **80**: 141–151.

[22] Heuser H, *Lehrbuch der Analysis, Teil* 2, Teubner, Stuttgart, 1981, 278.

[23] Yin C M, Li Y M, and Wang P Y, Strong convergence rates of the maximum quasi-likelihood estimator in generalized linear models (in Chinese), *Acta Math. Sci.*, 2009, **2009A**(4): 1058–1064.

[24] Chow Y S and Teicher H, *Probability Theory: Independence, Inchangeability, Martingales*, Springer, New York, 1988, 338–340.

[25] Chen K N, Hu I C, and Ying Z L, Strong consistency of maximum quasi-likelihood estimators in generalized linear models with fixed and adaptive designs, *Ann. Statist*, 1999, **27**: 1155–1163.

Appendix A: Proofs of the Main Results

In order to prove the main results in Section 3, we need the following lemmas.

Lemma 1 (see [22]) *Let $X = (x_1, \cdots, x_p)^\tau$ and $F = (f_1, \cdots, f_q)^\tau$. If $f_i = f_i(x_1, \cdots, x_p)$, $i = 1, \cdots, q$, are continuously differentiable in a convex set $G \subset R^p$, then for any $\alpha, \beta \in G$,*

$$F(\beta) - F(\alpha) = \left(\int_0^1 \frac{\partial F}{\partial X^\tau}\Big|_{X=\alpha+\lambda(\beta-\alpha)} d\lambda \right) (\beta - \alpha),$$

where the integration is to be read element-wise.

Lemma 2 (see [23]) *Let matrix $A(x) = (a_{ij}(x))$, if every element $a_{ij}(x)$ is continuous on x in $[\alpha, \beta]$,then*

$$\left\| \int_\alpha^\beta A(x)dx \right\| \le \int_\alpha^\beta \|A(x)\|dx,$$

where the integration is to be read element-wise.

Lemma 3 (see [24]) *Let $\{Y_{nj} > 0, 1 \le j \le k_n \to \infty\}$ is independent random variable sequence for each $n \ge 1$. If $E\sum_{j=1}^{k_n} Y_{nj} = 1$, and $\sum_{j=1}^{k_n} EY_{nj} I(Y_{nj} > \varepsilon) = o(1)$ for any $\varepsilon > 0$, then*

$$\sum_{j=1}^{k_n} Y_{nj} = 1 + o_p(1).$$

All proofs of the following Lemmas 4–9 are delegated to the Appendix B.

Lemma 4 *Under the conditions (A1)–(A3), as $n \to \infty$*

$$M_n^{-1/2}(\beta_0) \sum_{i=1}^n s_i(\beta_0) \to_d N(0, I_p).$$

Lemma 5 *Under conditions* (A1)–(A3), *the elements of* $\dot{\mu}_i(\beta)$, $A_i^{1/2}(\beta)$ *and* $A_i^{-1/2}(\beta)$, *and together with their* 1*st and* 2*nd partial derivatives with respect to* $\theta_i = X_i^\tau \beta$ *are uniformly bounded by a finite positive constant* C *for all* $1 \le i \le n$ *on* $N_n(\Delta)$.

Lemma 6 *Write* $D_n(\beta) = -\sum_{i=1}^n \partial s_i(\beta)/\partial \beta^\tau$ *and* $H_n = H_n(\beta_0)$. *Under the conditions* (A1)–(A3),

$$\sup_{\beta \in N_n(\Delta)} \|H_n^{-1/2} D_n(\beta) H_n^{-1/2} - I_p\| = o_p(1), \tag{5}$$

and

$$\sup_{\beta \in N_n(\Delta)} \|H_n^{-1/2} D_n^*(\beta) H_n^{-1/2} - I_p\| = o_p(1), \tag{6}$$

where $D_n^*(\beta) = \int_0^1 D_n(\beta_0 + \lambda\beta) d\lambda$.

Lemma 7 *Write* $M_n = M_n(\beta_0)$. *Under the conditions* (A1)–(A3),

$$\sup_{\beta \in N_n(\Delta)} \left\| \sum_{i=1}^n M_n^{-1/2} s_i(\beta) \right\| = O_p(1), \tag{7}$$

$$\sup_{\beta \in N_n(\Delta)} \max_{1 \le i \le n} ||M_n^{-1/2} s_i(\beta)|| = o_p(1), \tag{8}$$

$$\sup_{\beta \in N_n(\Delta)} \sum_{i=1}^n ||M_n^{-1/2} s_i(\beta)||^2 = O_p(1), \tag{9}$$

$$\sup_{\beta \in N_n(\Delta)} \sum_{i=1}^n ||M_n^{-1/2} s_i(\beta)||^3 = o_p(1). \tag{10}$$

Lemma 8 *Under the conditions* (A1)–(A3),

$$\sup_{\beta \in N_n(\Delta)} \left\| M_n^{-1/2} \sum_{i=1}^n s_i(\beta) s_i^\tau(\beta) M_n^{-1/2} - I_p \right\| = o_p(1).$$

Lemma 9 *Under the conditions* (A1)–(A3). *If* (1) *is satisfied, then*

$$\sup_{\beta \in N_n(\Delta)} \|t^\tau(\beta) M_n^{1/2}\| = O_p(1) \quad \text{and} \quad \sup_{\beta \in N_n(\Delta)} \|t^\tau(\beta) F_n^{1/2}\| = O_p(1).$$

Proof of Theorem 3.1 By (1), we can get

$$t(\beta) = \left[\sum_{i=1}^n s_i(\beta) s_i^\tau(\beta)\right]^{-1} \sum_{i=1}^n s_i(\beta) + \left[\sum_{i=1}^n s_i(\beta) s_i^\tau(\beta)\right]^{-1} \sum_{i=1}^n \frac{s_i(\beta)[t^\tau(\beta) s_i(\beta)]^2}{1 + t^\tau(\beta) s_i(\beta)}. \tag{11}$$

By Lemma 7 and Lemma 9, we have, for all $\beta \in N_n(\Delta)$ and $0 \le \lambda \le 1$,

$$\max_{1 \le i \le n} |\lambda t^\tau(\beta) s_i(\beta)| \le \|t^\tau(\beta) M_n^{1/2}\| \max_{1 \le i \le n} \|M_n^{-1/2} s_i(\beta)\| = o_p(1), \tag{12}$$

and

$$\max_{1\le i\le n}\left|\frac{1}{1+\lambda t^\tau(\beta)s_i(\beta)}\right| \le \frac{1}{1-\max_{1\le i\le n}|\lambda t^\tau(\beta)s_i(\beta)|} = O_p(1). \tag{13}$$

From (13), Lemmas 7 and 9, it follows that for $\beta \in N_n(\Delta)$ and $0 \le \lambda \le 1$,

$$\begin{aligned}\left|\sum_{i=1}^n \frac{[t^\tau(\beta)s_i(\beta)]^3}{1+\lambda t^\tau(\beta)s_i(\beta)}\right| &\le \max_{1\le i\le n}\left|\frac{1}{1+\lambda t^\tau(\beta)s_i(\beta)}\right| \|t^\tau(\beta)M_n^{1/2}\|^3 \sum_{i=1}^n \|M_n^{-1/2}s_i(\beta)\|^3 \\ &= o_p(1). \end{aligned}\tag{14}$$

By (1) and (14), for all $\beta \in N_n(\Delta)$,

$$\begin{aligned} 0 &= \sum_{i=1}^n t^\tau(\beta)s_i(\beta) - \sum_{i=1}^n [t^\tau(\beta)s_i(\beta)]^2 + \sum_{i=1}^n \frac{[t^\tau(\beta)s_i(\beta)]^3}{1+t^\tau(\beta)s_i(\beta)} \\ &= \sum_{i=1}^n t^\tau(\beta)s_i(\beta) - \sum_{i=1}^n [t^\tau(\beta)s_i(\beta)]^2 + o_p(1). \end{aligned}\tag{15}$$

By Tailor's expansion, (11), (14) and (15),

$$\begin{aligned} -l_n(\beta) &= \sum_{i=1}^n t^\tau(\beta)s_i(\beta) - \frac{1}{2}\sum_{i=1}^n [t^\tau(\beta)s_i(\beta)]^2 + \frac{1}{3}\sum_{i=1}^n \frac{[t^\tau(\beta)s_i(\beta)]^3}{(1+\lambda t^\tau(\beta)s_i(\beta)_i)^3} \quad (0<\lambda<1) \\ &= \frac{1}{2}\sum_{i=1}^n [t^\tau(\beta)s_i(\beta)]^2 + o_p(1) = I_{n1} + I_{n2} + I_{n3} + o_p(1), \end{aligned}\tag{16}$$

where

$$\begin{aligned} I_{n1} &= \frac{1}{2}\sum_{i=1}^n s_i^\tau(\beta)\left[\sum_{i=1}^n s_i(\beta)s_i^\tau(\beta)\right]^{-1}\sum_{i=1}^n s_i(\beta), \\ I_{n2} &= \sum_{i=1}^n s_i^\tau(\beta)\left[\sum_{i=1}^n s_i(\beta)s_i^\tau(\beta)\right]^{-1}\sum_{i=1}^n \left[s_i(\beta)\frac{[t^\tau(\beta)s_i(\beta)]^2}{1+t^\tau(\beta)s_i(\beta)}\right], \\ I_{n3} &= \frac{1}{2}\sum_{i=1}^n \left[\frac{[t^\tau(\beta)s_i(\beta)]^2}{1+t^\tau(\beta)s_i(\beta)}s_i^\tau(\beta)\right]\left[\sum_{i=1}^n s_i(\beta)s_i^\tau(\beta)\right]^{-1}\sum_{i=1}^n \left[s_i(\beta)\frac{[t^\tau(\beta)s_i(\beta)]^2}{1+t^\tau(\beta)s_i(\beta)}\right]. \end{aligned}$$

By Lemmas 7 and 8, for all $\beta \in N_n(\Delta)$,

$$I_{n1} = O_p(1). \tag{17}$$

By (13), Lemmas 7 and 9,

$$\begin{aligned}\left|\sum_{i=1}^n M_n^{-1/2}s_i(\beta)\frac{[t^\tau(\beta)s_i(\beta)]^2}{1+t^\tau(\beta)s_i(\beta)}\right| &\le \max_{1\le i\le n}\left|\frac{1}{1+t^\tau(\beta)s_i(\beta)}\right| \|t^\tau(\beta)M_n^{1/2}\|^2 \sum_{i=1}^n \|M_n^{-1/2}s_i(\beta)\|^3 \\ &= o_p(1), \end{aligned}$$

which combined with Lemma 8 implies

$$I_{n3} = o_p(1). \tag{18}$$

By the Cauchy–Schwarz inequality, (17) and (18), for all $\beta \in N_n(\Delta)$,

$$I_{n2} = o_p(1). \tag{19}$$

From (16), (18) and (19), it follows that, for all $\beta \in N_n(\Delta)$,

$$-l_n(\beta) = \frac{1}{2}\sum_{i=1}^n s_i^\tau(\beta)\Big[\sum_{i=1}^n s_i(\beta)s_i^\tau(\beta)\Big]^{-1}\sum_{i=1}^n s_i(\beta) + o_p(1). \tag{20}$$

By (20) and Lemma 1, for $\beta \in \partial N_n(\Delta) = \{\beta : \|F_n(\beta-\beta_0)\| = \Delta\}$,

$$-l_n(\beta) = J_{n1} + J_{n2} + J_{n3} + o_p(1), \tag{21}$$

where

$$\begin{aligned} J_{n1} &= \frac{1}{2}\sum_{i=1}^n s_i^\tau(\beta_0)\Big[\sum_{i=1}^n s_i(\beta)s_i^\tau(\beta)\Big]^{-1}\sum_{i=1}^n s_i(\beta_0), \\ J_{n2} &= \sum_{i=1}^n s_i^\tau(\beta_0)\Big[\sum_{i=1}^n s_i(\beta)s_i^\tau(\beta)\Big]^{-1} D_n^*(\beta)(\beta-\beta_0), \\ J_{n3} &= \frac{1}{2}(\beta-\beta_0)^\tau (D_n^*(\beta))^\tau \Big[\sum_{i=1}^n s_i(\beta)s_i^\tau(\beta)\Big]^{-1} D_n^*(\beta)(\beta-\beta_0). \end{aligned}$$

By Lemmas 4 and 8, for all $\beta \in \partial N_n(\Delta)$, the boundary of $N_n(\Delta)$,

$$J_{n1} = O_p(1). \tag{22}$$

Write $\Sigma_i(\beta) = A_i^{1/2}(\beta)R_i^{-1}A_i^{1/2}(\beta)$. By the conditions (A2) and (A3) (i),

$$C_1 I \le \Sigma_i(\beta_0) \le C_2 I, \quad C_1 I \le A_i^{1/2}(\beta_0)R_i^{-1}R_{i0}R_i^{-1}A_i^{1/2}(\beta_0) \le C_2 I, \tag{23}$$

which implies

$$C_1 F_n \le H_n \le C_2 F_n, \quad C_3 H_n \le M_n \le C_4 H_n. \tag{24}$$

By Lemmas 6 and 8, and (24), we can prove

$$P\Big\{C_1\Delta^2 \le \inf_{\beta\in\partial N_n(\Delta)} J_{n3} \le \sup_{\beta\in\partial N_n(\Delta)} J_{n3} \le C_2\Delta^2\Big\} \to 1. \tag{25}$$

By the Cauchy–Schwarz inequality, (22) and (25),

$$P\Big\{\sup_{\beta\in\partial N_n(\Delta)} |J_{n2}| \le C\Delta\Big\} \to 1. \tag{26}$$

From (21), (22), (25) and (26), it follows that

$$P\Big\{\inf_{\beta\in\partial N_n(\Delta)} -l_n(\beta) \ge C\Delta^2, \text{ for large enough } \Delta\Big\} \to 1. \tag{27}$$

By (20), Lemmas 4 and 8,

$$-l_n(\beta_0) = O_p(1),$$

which combined with (27) implies

$$P\Big\{\inf_{\beta\in\partial N_n(\Delta)} -l_n(\beta) \geq -l_n(\beta_0),\ \text{for large enough } \Delta\Big\} \to 1.$$

So, there exists $\widehat{\beta}_n$ such that $l_n(\beta)$ attains its maximum value and

$$P\left\{\widehat{\beta}_n \in N_n(\Delta), \quad \frac{\partial l_n(\beta)}{\partial \beta^\tau}\bigg|_{\beta=\widehat{\beta}_n} = 0\right\} \to 1, \tag{28}$$

which means that (2) is proved.

By (28) and $\underline{\lambda}_n I_p \leq F_n$, we have

$$P\{\|\widehat{\beta}_n - \beta_0\| \leq \underline{\lambda}_n^{-1/2}\Delta\} \geq P\{\|F_n^{1/2}(\widehat{\beta}_n - \beta_0)\| \leq \Delta\} \to 1. \tag{29}$$

Thus (3) is proved. By (28) and (29), the proof of Theorem 3.1 is complete. ∎

Proof of Theorem 3.2 By (28), we have

$$Q_{1n}(\widehat{\beta}_n, \widehat{t}) = 0, \quad Q_{2n}(\widehat{\beta}_n, \widehat{t}) = 0, \tag{30}$$

where

$$\widehat{t} = t(\widehat{\beta}_n), \quad Q_{2n}(\beta, t) = \sum_{i=1}^n \frac{1}{1 + t^\tau s_i(\beta)} \frac{\partial s_i^\tau(\beta)}{\partial \beta} t, \quad t = t(\beta).$$

Noting $Q_{1n}(\beta_0, 0) = \sum_{i=1}^n s_i(\beta_0)$ and $Q_{2n}(\beta_0, 0) = 0$, by (30) and Lemma 1,

$$\begin{pmatrix} \sum_{i=1}^n s_i(\beta_0) \\ 0 \end{pmatrix} = \begin{pmatrix} -Q_{1n}(\beta_0, 0) \\ -Q_{2n}(\beta_0, 0) \end{pmatrix} = \begin{pmatrix} A & B \\ B^\tau & D \end{pmatrix} \begin{pmatrix} \widehat{t} - 0 \\ \widehat{\beta}_n - \beta_0 \end{pmatrix}, \tag{31}$$

where

$$A = \int_0^1 \overline{A}_n(\beta_0 + \lambda\widehat{\beta}_n, \lambda\widehat{t}) d\lambda, \quad \overline{A}_n(\beta, t) = \frac{\partial Q_{1n}(\beta, t)}{\partial t^\tau}, \tag{32}$$

$$B = \int_0^1 \overline{B}_n(\beta_0 + \lambda\widehat{\beta}_n, \lambda\widehat{t}) d\lambda, \quad \overline{B}_n(\beta, t) = \frac{\partial Q_{1n}(\beta, t)}{\partial \beta^\tau}, \tag{33}$$

$$D = \int_0^1 \overline{D}_n(\beta_0 + \lambda\widehat{\beta}_n, \lambda\widehat{t}) d\lambda, \quad \overline{D}_n(\beta, t) = \frac{\partial Q_{2n}(\beta, t)}{\partial \beta^\tau}. \tag{34}$$

By straightforward but tedious calculation of matrix differentiation, we can obtain that

$$\begin{aligned}
\overline{A}_n(\beta, t) &= \sum_{i=1}^n \frac{-1}{[1 + t^\tau s_i(\beta)]^2} s_i(\beta) s_i^\tau(\beta), \\
\overline{B}_n(\beta, t) &= \sum_{i=1}^n \frac{1}{[1 + t^\tau s_i(\beta)]^2} \frac{\partial s_i(\beta)}{\partial \beta^\tau}, \\
\overline{D}_n(\beta, t) &= \overline{D}_n^{(1)}(\beta, t) + \overline{D}_n^{(2)}(\beta, t),
\end{aligned} \tag{35}$$

with

$$\overline{D}_n^{(1)}(\beta,t)=\sum_{i=1}^{n}\frac{-1}{[1+t^\tau s_i(\beta)]^2}\frac{\partial s_i^\tau(\beta)}{\partial\beta}tt^\tau\frac{\partial s_i(\beta)}{\partial\beta^\tau},$$

$$\overline{D}_n^{(2)}(\beta,t)=\sum_{i=1}^{n}\frac{1}{1+t^\tau s_i(\beta)}\sum_{l=1}^{p}\frac{\partial^2 s_i^\tau(\beta)\alpha_l}{\partial\beta\partial\beta^\tau}\alpha_l^\tau t, \tag{36}$$

where α_l is the $p\times 1$ vector with the lth element equal 1 and the other elements equal to 0.

Let α_{ij} be the $m_i\times 1$ vector with the jth element equal 1 and the other elements equal to 0, and $\overline{R}_i(\beta)=A_i^{1/2}(\beta)R_i^{-1}A_i^{-1/2}(\beta)$, then

$$\partial s_i(\beta)/\partial\beta^\tau=-X_i\Sigma_i(\beta)X_i^\tau+\sum_{j=1}^{m_i}X_i\frac{\partial[\overline{R}_i(\beta)\alpha_{ij}]}{\partial\theta_i^\tau}X_i^\tau\left\{\alpha_{ij}^\tau[Y_i-\mu_i(\beta)]\right\}. \tag{37}$$

By (37) and calculation of matrix differentiation, we can get that

$$\begin{aligned}\frac{\partial s_i^\tau(\beta)\alpha_l}{\partial\beta\partial\beta^\tau}=&-\sum_{j=1}^{m_i}X_i\frac{\partial\left[\Sigma_i(\beta)\alpha_{ij}\right]}{\partial\theta_i^\tau}X_i^\tau\left[\alpha_{ij}^\tau X_i^\tau\alpha_l\right]-\sum_{j=1}^{m_i}X_i\frac{\partial\left[\alpha_{ij}^\tau\overline{R}_i(\beta)\right]}{\partial\theta_i}X_i^\tau\alpha_l\left\{\alpha_{ij}^\tau A_i(\beta)X_i^\tau\right\}\\&+\sum_{j=1}^{m_i}\sum_{k=1}^{m_i}X_i\frac{\partial\left[\alpha_{ij}^\tau\overline{R}_i(\beta)\alpha_{ik}\right]}{\partial\theta_i\partial\theta_i^\tau}X_i^\tau\times\left(\alpha_{ik}^\tau X_i^\tau\alpha_l\left\{\alpha_{ij}^\tau[Y_i-\mu_i(\beta)]\right\}\right).\end{aligned}$$

Which combined with (36) implies

$$\begin{aligned}\overline{D}_n^{(2)}(\beta,t)=&-\sum_{i=1}^{n}\frac{1}{1+t^\tau s_i(\beta)}X_i\sum_{l=1}^{p}\sum_{j=1}^{m_i}\frac{\partial\left[\Sigma_i(\beta)\alpha_{ij}\right]}{\partial\theta_i^\tau}\left[\alpha_{ij}^\tau X_i^\tau\alpha_l\times\alpha_l^\tau t\right]X_i^\tau\\&-\sum_{i=1}^{n}\frac{1}{1+t^\tau s_i(\beta)}X_i\sum_{l=1}^{p}\sum_{j=1}^{m_i}\frac{\partial\left[\alpha_{ij}^\tau\overline{R}_i(\beta)\right]}{\partial\theta_i}X_i^\tau\alpha_l\alpha_{ij}^\tau A_i(\beta)\left[\alpha_l^\tau t\right]X_i^\tau\\&+\sum_{i=1}^{n}\frac{1}{1+t^\tau s_i(\beta)}X_i\sum_{l=1}^{p}\sum_{j=1}^{m_i}\sum_{k=1}^{m_i}\frac{\partial\left[\alpha_{ij}^\tau\overline{R}_i(\beta)\alpha_{ik}\right]}{\partial\theta_i\partial\theta_i^\tau}\left[\alpha_{ik}^\tau X_i^\tau\alpha_l\alpha_{ij}^\tau(Y_i-\mu_i(\beta))\alpha_l^\tau t\right]X_i^\tau.\end{aligned}$$

By (12), for all $\beta\in N_n(\Delta)$, $0\le\lambda\le 1$,

$$\max_{1\le i\le n}\left|1-\frac{1}{[1+\lambda t^\tau(\beta)s_i(\beta)]^2}\right|\le\frac{\max_{1\le i\le n}[2\,|t^\tau(\beta)s_i(\beta)|+|t^\tau(\beta)s_i(\beta)|^2]}{[1-\max_{1\le i\le n}|t^\tau(\beta)s_i(\beta)|]^2}=o_p(1). \tag{38}$$

Noting that $\underline{\lambda}_n$ is a positive nondecreasing function, by (A1), we can prove

$$\max_{1\le i\le n}\|X_i\|^2/\underline{\lambda}_n\to 0, \tag{39}$$

which implies

$$\sup_{i\le n,j\le m_i,\beta\in N_n(\Delta)}\|X_{ij}^\tau(\beta-\beta_0)\|\le\sup_{i\le n,\beta\in N_n(\Delta)}\|X_i^\tau F_n^{-1/2}\|\|F_n^{1/2}(\beta-\beta_0)\|\to 0. \tag{40}$$

By the mean value theorem, Lemma 5 and (40), as $n \to \infty$,

$$\sup_{i \le n, j \le m_i, \beta \in N_n(\Delta)} \|\mu_{ij}(\beta) - \mu_{ij}(\beta_0)\| \to 0. \tag{41}$$

From Lemma 5 and (A3) (i), it follows that, for all $\beta \in N_n(\Delta), 1 \le i \le n$,

$$\|\overline{R}_i(\beta)\| \le C, \quad \left\| \frac{\partial \left[\Sigma_i(\beta)\alpha_{ij}\right]}{\partial \theta_i^\tau} \right\| \le C, \tag{42}$$

and

$$\left\| \frac{\partial \left[\overline{R}_i(\beta)\alpha_{ij}\right]}{\partial \theta_i^\tau} \right\| \le C, \quad \left\| \frac{\partial \left[\alpha_{ij}^\tau \overline{R}_i(\beta)\alpha_{ik}\right]}{\partial \theta_i \partial \theta_i^\tau} \right\| \le C. \tag{43}$$

By (23), we have

$$\sum_{i=1}^n \|M_n^{-1/2} X_i\|^2 = \operatorname{trace} \sum_{i=1}^n M_n^{-1/2} X_i X_i^\tau M_n^{-1/2} \le C, \tag{44}$$

and

$$\sum_{i=1}^n \|H_n^{-1/2} X_i\|^2 \le C. \tag{45}$$

By (38), (41), (42), (A2) (iii), (44) and Markov inequality, for all $\beta \in N_n(\Delta)$ and $0 \le \lambda \le 1$,

$$\begin{aligned}
&\left\| M_n^{-1/2} \left[\overline{A}_n(\beta, \lambda t) + \sum_{i=1}^n s_i(\beta) s_i^\tau(\beta) \right] M_n^{-1/2} \right\| \\
&\le \sum_{i=1}^n \left| 1 - \frac{1}{\left[1 + \lambda t^\tau(\beta) s_i(\beta)\right]^2} \right| \|M_n^{-1/2} X_i\|^2 \|\overline{R}_i(\beta)(Y_i - \mu_i(\beta))\|^2 \\
&\le \max_{1 \le i \le n} \left| 1 - \frac{1}{\left[1 + \lambda t^\tau(\beta) s_i(\beta)\right]^2} \right| \sum_{i=1}^n \|M_n^{-1/2} X_i\|^2 C(\|Y_i - \mu_i(\beta_0)\| + C)^2 \\
&= o_p(1) \times O_p(1) \\
&= o_p(1).
\end{aligned} \tag{46}$$

From (46), Lemma 2 and Lemma 8, it follows that

$$\sup_{\beta \in N_n(\Delta)} \left\| M_n^{-1/2} \int_0^1 \overline{A}_n(\beta_0 + \lambda \beta, \lambda t) d\lambda M_n^{-1/2} + I \right\| = o_p(1). \tag{47}$$

Similar arguments as in (46). By (38), (41), (43), (A2) (iii), (45) and Markov inequality, for all $\beta \in N_n(\Delta)$, $0 \le \lambda \le 1$,

$$\begin{aligned}
&\|H_n^{-1/2} \left(\overline{B}_n(\beta, \lambda t) + D_n(\beta)\right) H_n^{-1/2}\| \\
&\le \sum_{i=1}^n \left| 1 - \frac{1}{\left[1 + \lambda t^\tau(\beta) s_i(\beta)\right]^2} \right| \|H_n^{-1/2} X_i\|^2 \left\| \Sigma_i(\beta) - \sum_{j=1}^{m_i} \frac{\partial \left[\overline{R}_i(\beta)\alpha_{ij}\right]}{\partial \theta_i^\tau} \alpha_{ij}^\tau (Y_i - \mu_i(\beta)) \right\|, \\
&\le \max_{1 \le i \le n} \left| 1 - \frac{1}{\left[1 + t^\tau(\beta) s_i(\beta)\right]^2} \right| \sum_{i=1}^n \|H_n^{-1/2} X_i\|^2 \left[C_1 + \sum_{j=1}^{m_i} C_2(\|Y_i - \mu(\beta_0)\| + C_3) \right] \\
&= o_p(1).
\end{aligned} \tag{48}$$

By Lemmas 2 and 6, and (48),

$$\sup_{\beta\in N_n(\Delta)}\left\|H_n^{-1/2}\left(\int_0^1 \overline{B}_n(\beta_0+\lambda\beta,\lambda t)d\lambda\right)H_n^{-1/2}+I\right\|=o_p(1). \tag{49}$$

Similarly, by (13), (41), (43), Lemma 9, (A2) (iii), (45), Markov inequality and (39), for all $\beta\in N_n(\Delta)$, $0\le\lambda\le 1$,

$$\begin{aligned}
&\|H_n^{-1/2}\overline{D}_n^{(1)}(\beta,t\lambda)H_n^{-1/2}\|\\
&\le\sum_{i=1}^n\frac{1}{[1+\lambda t^\tau(\beta)s_i(\beta)]^2}\|H_n^{-1/2}X_i\|^2\left\|\Sigma_i(\beta)-\sum_{j=1}^{m_i}\frac{\partial\left[\overline{R}_i(\beta)\alpha_{ij}\right]}{\partial\theta_i^\tau}\alpha_{ij}^\tau(Y_i-\mu_i(\beta))\right\|^2\|X_i^\tau t\lambda\|^2,\\
&\le\max_{\le i\le n}\frac{1}{[1+\lambda t^\tau(\beta)s_i(\beta)]^2}\max_{1\le i\le n}\|X_i^\tau t\lambda\|^2\sum_{i=1}^n\|H_n^{-1/2}X_i\|^2\left[C_1+\sum_{j=1}^{m_i}C_2(\|Y_i-\mu_i(\beta_0)\|+C_3)\right]^2\\
&=O_p(1)\times\max_{1\le i\le n}\|X_i\|^2\underline{\lambda}_n^{-1}\times O_p(1)\\
&=o_p(1).
\end{aligned}\tag{50}$$

By (13), (41)–(43), Lemma 9, (A2) (iii), (45), Markov inequality and (39), for all $\beta\in N_n(\Delta)$, $0\le\lambda\le 1$,

$$\begin{aligned}
&\|H_n^{-1/2}\overline{D}_n^{(2)}(\beta,\lambda t)H_n^{-1/2}\|\\
&\le\sum_{i=1}^n\frac{1}{|1+\lambda t^\tau(\beta)s_i(\beta)|}\|H_n^{-1/2}X_i\|^2\\
&\quad\times\left[\sum_{l=1}^p\sum_{j=1}^{m_i}C\|X_i\|\|t\|+\sum_{l=1}^p\sum_{j=1}^{m_i}C\|X_i\|C\|t\|+\sum_{l=1}^p\sum_{j=1}^{m_i}\sum_{k=1}^{m_i}C\|X_i\|(\|Y_i-\mu_i(\beta_0)\|+C)\|t\|\right]\\
&\le\max_{1\le i\le n}\frac{1}{|1+\lambda t^\tau(\beta)s_i(\beta)|}\max_{1\le i\le n}\|X_i\|\|t\|\sum_{i=1}^n\|H_n^{-1/2}X_i\|^2\left[C+C+C\|Y_i-\mu_i(\beta_0)\|+C\right],\\
&=O_p(1)\times\max_{1\le i\le n}\|X_i\|\times\underline{\lambda}_n^{-1/2}\times O_p(1)\\
&=o_p(1).
\end{aligned}\tag{51}$$

By (35), (50), (51) and Lemma 2,

$$\sup_{\beta\in N_n(\Delta)}\left\|H_n^{-1/2}\int_0^1\overline{D}_n(\beta_0+\lambda\beta,\lambda t)d\lambda H_n^{-1/2}\right\|=o_p(1). \tag{52}$$

Note the fact that

$$\begin{aligned}
&\begin{pmatrix} A & B\\ B^\tau & D\end{pmatrix}^{-1}\\
&=\begin{pmatrix} A^{-1}+A^{-1}B(D-B^\tau A^{-1}B)^{-1}B^\tau A^{-1} & -A^{-1}B(D-B^\tau A^{-1}B)^{-1}\\ -(D-B^\tau A^{-1}B)^{-1}B^\tau A^{-1} & (D-B^\tau A^{-1}B)^{-1}\end{pmatrix}.
\end{aligned}\tag{53}$$

By (24), we can obtain

$$\|H_n^{\frac{1}{2}} M_n^{-\frac{1}{2}}\| \le C, \quad \|M_n^{\frac{1}{2}} H_n^{-\frac{1}{2}}\| \le C. \tag{54}$$

By (2), (32)–(34), (47), (49), (52), (54), and tedious computation,

$$M_n^{-\frac{1}{2}} H_n (D - B^\tau A^{-1} B)^{-1} B^\tau A^{-1} M_n^{1/2} = I + o_p(1). \tag{55}$$

From (31), (53), (55) and Lemma 4, it follows that

$$M_n^{-\frac{1}{2}} H_n(\widehat{\beta}_n - \beta_0) = M_n^{-\frac{1}{2}} H_n (D - B^\tau A^{-1} B)^{-1} B^\tau A^{-1} \sum_{i=1}^{n} s_i(\beta_0) \to_d N(0, I_p). \tag{56}$$

By (28), (29) and (56), the proof of Theorem 3.2 is complete. ∎

Proof of Theorem 3.3 By (20), Lemmas 4 and 8,

$$-2l_n(\beta_0) \to \chi_p^2. \tag{57}$$

Similar as (55), using (2), (32)–(34), (47), (49), (52) and (54), we can prove

$$M_n^{1/2}[A^{-1} + A^{-1}B(D - B^\tau A^{-1} B)^{-1} B^\tau A^{-1}] M_n^{1/2} = o_p(1). \tag{58}$$

By (31), (53), (58) and Lemma 4,

$$M_n^{1/2}\widehat{t} = -M_n^{1/2}[A^{-1} + A^{-1}B(D - B^\tau A^{-1} B)^{-1} B^\tau A^{-1}] \sum_{i=1}^{n} s_i(\beta_0) = o_p(1). \tag{59}$$

From (2) and (14), it follows that

$$\frac{1}{3}\sum_{i=1}^{n} \frac{[\widehat{t}^\tau s_i(\widehat{\beta}_n)]^3}{(1 + \lambda \widehat{t}^\tau s_i(\widehat{\beta}_n))^3} = o_p(1), \quad 0 \le \lambda \le 1. \tag{60}$$

By Taylor's expansion, (2), Lemmas 7 and 8, (59) and (60),

$$\begin{aligned}
-l_n(\widehat{\beta}_n) &= \sum_{i=1}^{n} \log\{1 + \widehat{t}^\tau s_i(\widehat{\beta}_n)\} \\
&= \sum_{i=1}^{n} \widehat{t}^\tau s_i(\widehat{\beta}_n) - \frac{1}{2}\sum_{i=1}^{n} [\widehat{t}^\tau s_i(\widehat{\beta}_n)]^2 + \frac{1}{3}\sum_{i=1}^{n} \frac{[\widehat{t}^\tau s_i(\widehat{\beta}_n)]^3}{(1 + \lambda \widehat{t}^\tau s_i(\widehat{\beta}_n))^3}, \quad 0 < \lambda < 1 \\
&= o_p(1).
\end{aligned} \tag{61}$$

By (57) and (61), the proof of Theorem 3.3 is complete. ∎

Appendix B: Proofs of Lemmas 4–9

Proof of Lemma 4 For any fixed $p \times 1$ unit vector α, let $e_i = Y_i - \mu_i(\beta_0)$, $a_{ni} = \alpha M_n^{-1/2}(\beta_0) X_i A_i^{1/2}(\beta_0) R_i^{-1} A_i^{-1/2}(\beta_0)$, $Z_{ni} = a_{ni} e_i$. Obviously, $\mathrm{E}(Z_{ni}) = 0$, $\mathrm{Var}(\sum_{i=1}^{n} Z_{ni}) = 1$.

By (A2) and (A3) (i), we have

$$C_1 I_{m_i} \geq A_i^{1/2}(\beta_0) R_i^{-1} R_{i0} R_i^{-1} A_i^{1/2}(\beta_0) \geq C_2 I_{m_i}, \tag{62}$$

$$A_i^{1/2}(\beta_0) R_i^{-1} A_i^{-1/2}(\beta_0) \leq C I_{m_i}, \quad A_i^{-1/2}(\beta_0) \leq C R_{i0}^{1/2}. \tag{63}$$

By (A1), we can prove

$$\max_{1\leq i\leq n} \|X_i\|^2/\underline{\lambda}_n \to 0. \tag{64}$$

From (62), (63) and (64), it follows that

$$\max_{1\leq i\leq n} \|a_{ni}\| \to 0, \quad \sum_{i=1}^n \|a_{ni}\|^2 = \mathrm{trace}\bigg(\sum_{i=1}^n a_{ni} a_{ni}^\tau\bigg) \leq C. \tag{65}$$

By (65) and (A2) (iii), for any fixed $\varepsilon > 0$, as $n \to \infty$,

$$\sum_{i=1}^n E[(Z_{ni})^2 I(|Z_{ni}| > \varepsilon)] \leq \sup_{i\geq 1} E\|e_i\|^{2+\varepsilon_0} \max_{1\leq i\leq n} \|a_{ni}\|^{\varepsilon_0} \sum_{i=1}^n \|a_{ni}\|^2/\varepsilon^{\varepsilon_0} \to 0, \tag{66}$$

the Lindberg condition is verified, so Lemma 4 holds. ∎

Proof of Lemma 5 By (64), we have

$$\sup_{i\leq n, j\leq m_i, \beta\in N_n(\Delta)} \|X_{ij}^\tau(\beta-\beta_0)\| \leq \sup_{i\leq n, \beta\in N_n(\Delta)} \|X_i^\tau F_n^{-1/2}\| \|F_n^{1/2}(\beta-\beta_0)\| \to 0. \tag{67}$$

By the mean value theorem, (67) and (A3) (ii),

$$\sup_{i\leq n, j\leq m_i, \beta\in N_n(\Delta)} |\ddot{\mu}(X_{ij}^\tau\beta)| \leq C. \tag{68}$$

By the mean value theorem, (67), (68) and (A2), for $i \leq n, j \leq m_i, \beta \in N_n(\Delta)$,

$$C_1 \leq \sigma_{ij}(\beta) = \dot{\mu}(X_{ij}^\tau\beta) \leq C_2. \tag{69}$$

By (68), (69) and (A3) (ii), Lemma 5 can be easily proved. ∎

Proof of Lemma 6

Write $\Sigma_i(\beta) = A_i^{1/2}(\beta) R_i^{-1} A_i^{1/2}(\beta)$, and $W_{ij}(\beta) = \partial[A_i^{1/2}(\beta) R_i^{-1} A_i^{-1/2}(\beta)\alpha_{ij}]/\partial\theta_i^\tau$, then

$$\partial s_i(\beta)/\partial\beta^\tau = -X_i \Sigma_i(\beta) X_i^\tau + \sum_{j=1}^{m_i} X_i W_{ij}(\beta) X_i^\tau \left\{\alpha_{ij}^\tau [Y_i - \mu_i(\beta)]\right\}. \tag{70}$$

By direct decomposition, we have

$$D_n(\beta) \equiv -\sum_{i=1}^n \partial s_i(\beta)/\partial\beta^\tau = H_n(\beta) - \sum_{j=1}^m [E_{nj} + G_{nj}(\beta) + J_{nj}(\beta)], \tag{71}$$

where

$$E_{nj} \equiv \sum_{i=1}^{n} X_i W_{ij}(\beta_0)\alpha_{ij}^{\tau}[Y_i - \mu_i(\beta_0)]I(j \le m_i)X_i^{\tau},$$

$$G_{nj}(\beta) \equiv \sum_{i=1}^{n} X_i\{W_{ij}(\beta) - W_{ij}(\beta_0)\}\alpha_{ij}^{\tau}[Y_i - \mu_i(\beta)]I(j \le m_i)X_i^{\tau},$$

$$J_{nj}(\beta) \equiv \sum_{i=1}^{n} X_i W_{ij}(\beta_0)\alpha_{ij}^{\tau}[\mu_i(\beta_0) - \mu_i(\beta)]I(j \le m_i)X_i^{\tau}.$$

Here, $I(j \le m_i)$ takes value 1 if $j \le m_i$ and 0 otherwise.

By (A2) and (A3) (i),

$$CI \le \Sigma_i(\beta_0) \le C_1 I, \tag{72}$$

which combined with (64) implies

$$\max_{1\le i\le n} \|H_n^{-1/2}X_i\| \le C^{-1/2} \max_{1\le i\le n} \|F_n^{-1/2}X_i\| \to 0, \tag{73}$$

$$\sum_{i=1}^{n} \|H_n^{-1/2}X_i\|^2 = \operatorname{trace}\sum_{i=1}^{n} H_n^{-1/2}X_iX_i^{\tau}H_n^{-1/2} \le C^{-1}p. \tag{74}$$

By the mean value theorem, Lemma 5, A(3) (i) and (64), for $\beta \in N_n(\Delta), 1 \le i \le n$

$$\|\Sigma_i(\beta) - \Sigma_i(\beta_0)\| \le C\|X_i^{\tau}F_n^{-1/2}\|\|F_n^{1/2}(\beta - \beta_0)\| \to 0, \tag{75}$$

which combined with (74) implies

$$\|H_n^{-1/2}H_n(\beta)H_n^{-1/2} - I\| \le \sum_{i=1}^{n} \|H_n^{-1/2}X_i\|\|\Sigma_i(\beta) - \Sigma_i(\beta_0)\|\|H_n^{-1/2}X_i\| \to 0. \tag{76}$$

From (A2) (iii), (A3) (i) and Lemma 5, it follows that

$$\|\mathrm{Cov}(Y_i)\| \le C, \quad \|W_{ij}(\beta_0)\|^2 \le C. \tag{77}$$

By (73), (74) and (77), for $1 \le l, q \le p$,

$$\mathrm{Var}(\alpha_l^{\tau}H_n^{-1/2}E_{nj}H_n^{-1/2}\alpha_q) \le \sum_{i=1}^{n} \|H_n^{-1/2}X_i\|^4\|W_{ij}(\beta_0)\|^2\|\mathrm{Cov}(Y_i)\| \to 0. \tag{78}$$

Similarly as (75), by the mean value theorem, Lemma 5, A(3) (i) and (64),

$$\sup_{\beta\in N_n(\Delta), i\le n} \|W_{ij}(\beta) - W_{ij}(\beta_0)\| \to 0, \qquad \sup_{\beta\in N_n(\Delta), i\le n} \|\mu_i(\beta_0) - \mu_i(\beta)\| \to 0. \tag{79}$$

By (74), (79) and (A2) (iii),

$$E \sup_{\beta\in N_n(\Delta), i\le n} \|H_n^{-1/2}G_{nj}(\beta)H_n^{-1/2}\| \to 0. \tag{80}$$

By (74), (77) and (79),

$$\sup_{\beta \in N_n(\Delta), i \le n} \|H_n^{-1/2} J_{nj}(\beta) H_n^{-1/2}\| \to 0. \tag{81}$$

By (71), (76), (78), (80), (81) and the Markov inequality, then (5) holds. By (5) and Lemma 2, we know that the (6) is true. ∎

Proof of Lemma 7 By (62) and (72), we have

$$C_1 F_n \le H_n \le C_2 F_n, \quad C_3 H_n \le M_n \le C_4 H_n. \tag{82}$$

By the mean value theorem, Lemma 4, Lemma 6 and (82),

$$\begin{aligned}
&\sup_{\beta \in N_n(\Delta)} \left\| \sum_{i=1}^n M_n^{-1/2} s_i(\beta) \right\| \\
\le &\left\| \sum_{i=1}^n M_n^{-1/2} s_i(\beta_0) \right\| \\
&+ \sup_{\beta \in N_n(\Delta)} \|M_n^{-1/2} H_n^{1/2} \times H_n^{-1/2} D_n^*(\overline{\beta}) H_n^{-1/2} \times H_n^{1/2} F_n^{-1/2} \times F_n^{1/2}(\beta - \beta_0)\| \\
= &O_p(1) + C \times O_p(1) \times C \times \Delta \\
= &O_p(1),
\end{aligned}$$

then the formula (7) is proved.

From (62) and (64), it follows that

$$\max_{1 \le i \le n} ||M_n^{-1/2} X_i\| \to 0, \quad \sum_{i=1}^n \|M_n^{-1/2} X_i\|^2 \le C. \tag{83}$$

By the Markov inequality, (A2) (iii) and (83), for any $\varepsilon > 0$,

$$\begin{aligned}
&P\Big(\max_{1 \le i \le n} \|M_n^{-1/2} X_i\| \|Y_i - \mu_i(\beta_0)\| > \varepsilon \Big) \\
= &P(\cup_{1 \le i \le n} \{\|M_n^{-1/2} X_i\| \|Y_i - \mu_i(\beta_0)\| > \varepsilon\}) \\
\le &\sum_{i=1}^n E[||M_n^{-1/2} X_i\| \|Y_i - \mu_i(\beta_0)||]^{2+\delta} / \varepsilon^{2+\delta} \\
\le &C \max_{1 \le i \le n} ||M_n^{-1/2} X_i\|^\delta \sum_{i=1}^n ||M_n^{-1/2} X_i\|^2 \sup_{i \ge 1} E\|Y_i - \mu_i(\beta_0)||^{2+\delta} / \varepsilon^{2+\delta} \\
= &o(1).
\end{aligned} \tag{84}$$

By Lemma 5 and A(3) (i), we have

$$\sup_{\beta \in N_n(\Delta), i \le n} \|\overline{R}_i(\beta)\| \le C. \tag{85}$$

By (79) and (83)–(85),

$$\begin{aligned}
&\sup_{\beta\in N_n(\Delta)}\max_{1\le i\le n}\|M_n^{-1/2}s_i(\beta)\|\\
\le&\sup_{\beta\in N_n(\Delta),i\le n}[\|M_n^{-1/2}X_i\|\|Y_i-\mu_i(\beta_0)\|\cdot\|\overline{R}_i(\beta)\|+\|M_n^{-1/2}X_i\|\|\overline{R}_i(\beta)\|\|\mu_i(\beta_0)-\mu_i(\beta)\|]\\
=&\,o_p(1),
\end{aligned}$$

now the formula (8) is proved.

From (79), (83), (A2) (iii) and (85), it follows that

$$\begin{aligned}
&E\sup_{\beta\in N_n(\Delta)}\sum_{i=1}^n\|M_n^{-1/2}s_i(\beta)\|^2\\
\le&\, E\sum_{i=1}^n\sup_{\beta\in N_n(\Delta)}[\|M_n^{-1/2}X_i\|\|\overline{R}_i(\beta)\|(\|Y_i-\mu_i(\beta_0)\|+\|\mu_i(\beta_0)-\mu_i(\beta)\|)]^2\\
\le&\, C\sum_{i=1}^n\|M_n^{-1/2}X_i\|^2E(\|Y_i-\mu_i(\beta_0)\|+C)^2\\
\le&\, C,
\end{aligned}$$

which combined with the Markov inequality implies that (9) holds.

By (8) and (9),

$$\begin{aligned}
\sup_{\beta\in N_n(\Delta)}\sum_{i=1}^n\|M_n^{-1/2}s_i(\beta)\|^3&=\sup_{\beta\in N_n(\Delta)}\max_{1\le i\le n}\|M_n^{-1/2}s_i(\beta)\|\sup_{\beta\in N_n(\Delta)}\sum_{i=1}^n\|M_n^{-1/2}s_i(\beta)\|^2\\
&=o_p(1),
\end{aligned}$$

Thus the formula (10) is proved. ∎

Proof of Lemma 8 By direct decomposition,we have

$$\sum_{i=1}^n s_i(\beta)s_i^\tau(\beta)=\widetilde{A}_n+\widetilde{B}_n(\beta)+\widetilde{C}_n(\beta)+\widetilde{D}_n(\beta)+\widetilde{E}_n(\beta),\tag{86}$$

where

$$\begin{aligned}
\widetilde{A}_n&=\sum_{i=1}^n X_i\overline{R}_i(\beta_0)e_ie_i^\tau\overline{R}_i(\beta_0)X_i^\tau,\\
\widetilde{B}_n(\beta)&=\sum_{i=1}^n X_i\overline{R}_i(\beta)[Y_i-\mu_i(\beta)][\mu_i(\beta_0)-\mu_i(\beta)]^\tau\overline{R}_i(\beta)X_i^\tau,\\
\widetilde{B}_n^*(\beta)&=\sum_{i=1}^n X_i\overline{R}_i(\beta)[\mu_i(\beta_0)-\mu_i(\beta)]e_i^\tau\overline{R}_i(\beta)X_i^\tau,\\
\widetilde{C}_n(\beta)&=\sum_{i=1}^n X_i\overline{R}_i(\beta)e_ie_i^\tau[\overline{R}_i(\beta)-\overline{R}_i(\beta_0)]X_i^\tau,\\
\widetilde{C}_n^*(\beta)&=\sum_{i=1}^n X_i[\overline{R}_i(\beta)-\overline{R}_i(\beta_0)]e_ie_i^\tau\overline{R}_i(\beta_0)X_i^\tau.
\end{aligned}$$

By (66) and Lemma 3 (obviously, $Z_{ni}^2 > 0$, $E\sum_{i=1}^n Z_{ni}^2 = 1$),

$$M_n^{-1/2}\widetilde{A}_n M_n^{-1/2} = I_p + o_p(1). \tag{87}$$

From (79), (83), (A2) (iii) and (85), it follows that

$$\begin{aligned}
&E\sup_{\beta\in N_n(\Delta)}\|M_n^{-1/2}\widetilde{B}_n(\beta)M_n^{-1/2}\| \\
&\le \sup_{\beta\in N_n(\Delta),1\le i\le n}\|\mu_i(\beta_0)-\mu_i(\beta)\|\sum_{i=1}^n\|M_n^{-1/2}X_i\|^2 C[\sup_i E\|Y_i-\mu_i(\beta_0)\|+C]C, \\
&= o_p(1)\times C\times C \\
&= o(1).
\end{aligned} \tag{88}$$

Similarly as (75), by the mean value theorem, Lemma 5, (A3) (i) and (64),

$$\sup_{\beta\in N_n(\Delta),i\le n}\|\overline{R}_i(\beta)-\overline{R}_i(\beta_0)\|\to 0. \tag{89}$$

Similarly as (88), noting (89),

$$E\sup_{\beta\in N_n(\Delta)}\|M_n^{-1/2}[\widetilde{B}_n^*(\beta)+\widetilde{C}_n(\beta)+\widetilde{C}_n^*(\beta)]M_n^{-1/2}\| = o(1). \tag{90}$$

By combining (86)–(88) and (90), Lemma 8 follows. ∎

Proof of Lemma 9 By (1), we have

$$\begin{aligned}
0 &= \sum_{i=1}^n \frac{t^\tau(\beta)s_i(\beta)}{1+t^\tau(\beta)s_i(\beta)} = \sum_{i=1}^n t^\tau(\beta)s_i(\beta) - \sum_{i=1}^n\frac{[t^\tau(\beta)s_i(\beta)]^2}{1+t^\tau(\beta)s_i(\beta)} \\
&\le \|t^\tau(\beta)M_n^{1/2}\|\left\|\sum_{i=1}^n M_n^{-1/2}s_i(\beta)\right\| - \sum_{i=1}^n\frac{[t^\tau(\beta)s_i(\beta)]^2}{1+\|t^\tau(\beta)M_n^{1/2}\|\max_{1\le i\le n}\|M_n^{-1/2}s_i(\beta)\|},
\end{aligned}$$

which implies

$$\begin{aligned}
&\|t^\tau(\beta)M_n^{1/2}\|\left[\alpha^\tau\sum_{i=1}^n M_n^{-1/2}s_i(\beta)s_i^\tau(\beta)M_n^{-1/2}\alpha - \left(\max_{1\le i\le n}\|M_n^{-1/2}s_i(\beta)\|\right)\left\|\sum_{i=1}^n M_n^{-1/2}s_i(\beta)\right\|\right] \\
&\le \left\|\sum_{i=1}^n M_n^{-1/2}s_i(\beta)\right\|,
\end{aligned} \tag{91}$$

where $\alpha = t^\tau(\beta)M_n^{1/2}/\|t^\tau(\beta)M_n^{1/2}\|$. From Lemma 7, Lemma 8 and (91), we can get

$$t^\tau(\beta)M_n^{1/2} = O_p(1),$$

which combined with (82) implies

$$t^\tau(\beta)F_n^{1/2} = t^\tau(\beta)M_n^{1/2}\times M_n^{-1/2}F_n^{1/2} = O_p(1).$$

Thus Lemma 9 follows. ∎