HIGH DIMENSION GENERALIZED ESTIMATING EQUATION ESTIMATION CONSISTENCY AND MODEL SELECTION CONSISTENCY

SHICHENG WU

A DISSERTATION SUBMITTED TO
THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN MATHEMATICS AND STATISTICS
YORK UNIVERSITY
TORONTO, ONTARIO

December 2018

Abstract

We consider the problem of asymptotic theory and model selection for high dimensional Generalized Estimating Equation (GEE) on marginal regression analysis for clustered or longitudinal data. In this "large n, diverging p" framework, we firstly establish the existence and consistency of the GEE estimator. Next we discuss the model selection and its consistency. As the GEE method only makes assumptions about the first two moments, the full likelihood is not specified. The likelihood based model selection criteria cannot be directly applied. This paper proposes two different information criteria. The first one applies simplified quasi-likelihood and the second one introduces a generalized model selection criterion based on a quadratic form of the residuals. Using the large deviation result of quadratic forms, we choose the appropriate penalty terms on the model complexity. The model selection consistency of the proposed criteria for divergent number of covariates is established for both simplified quasi-likelihood and the quadratic form of the residuals.

Keywords: GEE, Model Selection, Large Deviation, Quasi-likelihood.

Acknowledgements

First and foremost, I would like to express my greatest appreciation to my supervisor, Xin Gao. Without her valuable guidance, constant encouragement, extensive knowledge and generous weekend meeting time, this dissertation would not be completed. I truly respect her brilliant insights and enthusiasm on research.

I would like to extend my sincere appreciation to Professor Neal Madras and Professor Edward Furman, as members of my supervisory committee. My appreciation also goes to all faculty members, staff, and fellow graduate students in the Department of Mathematics and Statistics at York University. I also would like to thank Professor Raymond Carroll for his suggestions on the GIC part of this dissertation.

Last but not least, I deeply thank my family and my employers for their continuous support. Without their help and support, I cannot finish the study smoothly. A special thank goes to my mentor Dr. Henry Li from Manulife Financial Corporation for his understanding and support that I have taken half year leave from employment to finish the dissertation. A special thank goes to Marathon Cafe where my super-

visor and I meet there almost every weekend to update the progress and to discuss research topics. For me, the PhD study is a kind of Marathon.

Table of Contents

List of Tables

List of Figures

1 Introduction

Liang and Zeger (1986) extended the generalized linear models (McCullough and Nelder, 1989) to correlated data and proposed the Generalized Estimating Equation (GEE). The GEE estimator is consistent even when the working correlation matrix is mis-specified. Li (1997) investigated the consistency of GEE via a minimax approach. Xie and Yang (2003) established a more comprehensive large-sample theory for GEE including consistency and asymptotic normality. Balan and Schiopu-Kratina (2005) provided a rigorous study on GEE under a pseduo-likelihood framework. All of these papers assume the number of covariates p is fixed and the number of clusters n goes to infinity. Recently, a great amount of work has been devoted to the high-dimensional data analysis. Readers are referred to Donoho (2000), Fan and Li (2001); Fan and Lv (2008), and Lv and Fan (2009) for a more comprehensive review of the development. For correlated data, Wang (2011) established the consistency of GEE under the "large n, diverging p" scenario and the consistency requires that p grows to infinity at polynomial rate $o(n^{1/2})$.

With the advent of large collection of information, it is an essential problem to perform model selection to determine a subset of useful covariates. We consider the problem of model selection on generalized estimating equations for clustered or longitudinal data. The lack of likelihood formulation imposes the challenge on the use of traditional likelihood based model selection criterion. Based on the GEE approach, several model selection methods for marginal models have been developed. Pan (2001) extended Akaike (1974)'s work on Akaike Information Criterion (AIC) and proposed the Quasi-likelihood Information Criterion (QIC). The QIC combines the quasi-likelihoods of each observations using independent assumption, whereas each observation's quasi-likelihood is evaluated at the GEE estimates under any working correlation. Cantoni et al. (2005) proposed a generalized version of Mallows' C_p from Mallows (1973) by minimizing the prediction error. Wang and Qu (2009) developed a Bayesian Information Criterion (BIC) type of criterion Bayesian Information Quadratic Inference Function (BIQIF) using Qu et al. (2000)'s Quadratic Inference Function. The model selection consistency of BIQIF was established for finite number of covariates. It remains an open question of developing a GEE model selection criterion which is consistent for an unbounded number of predictors.

For model selection using full likelihood, Chen and Chen (2008) developed the Extended Bayesian Information Criterion (EBIC) for high dimensional linear regression.

Gao and Song (2010) developed the Composite Likelihood Bayesian Information Criterion (CLBIC) for high dimensional correlated data. Both EBIC and CLBIC are proved to have selection consistency when the total number of predictors tends to infinity and the number of true predictors is bounded by a constant. To deal with the situation where the true number of predictors is unbounded, Zhang and Shen (2010) proposed a corrected risk inflation criterion. Kim et al. (2012) proposed a Generalized Information Criterion (GIC) with modified penalty terms. The consistency of both criteria are established for linear regression model with unbounded true model size. In a more general setup including linear regression, generalized linear models and data integration of several correlated models, Gao and Carroll (2017) proposed a likelihood based information criterion with appropriately chosen penalty term and demonstrated its model selection consistency for unbounded true model size .

In Chapter 2 of this paper, we discuss the high dimensional GEE estimation consistency. Applying large deviation tools from Spokoiny and Zhilova (2013), we follow the technique from Gao and Carroll (2017) to estimate the distance between true and estimated parameters. To prove the existence and consistency, the same approach from Portnoy (1988) and Wang (2011) are used. We prove that the high dimensional GEE estimation consistency is still valid no matter the choice of working correlation matrix as long as it is positive definite.

In Chapter 3 of this paper, we aim to develop an information criterion for GEE with divergent number of predictors and unbounded true model size. As there is no likelihood available to evaluate the model fitting, we consider Pan (2001)'s style quasi-likelihood, which assumes the independent working correlation matrix. In Spokoiny and Zhilova (2013), exact large deviation results were established for quadratic forms based on random vector satisfying the exponential moment conditions. Gao and Carroll (2017) extends the large deviation results to asymptotic setting for quadratic forms based on sample mean type of random vectors. Studying the large deviation result of Pan (2001)'s style quasi-likelihood enables us to choose the appropriate penalty size on the model complexity to ensure the model selection consistency under independent working correlation matrix. But it comes with a limitation that the working correlation matrix has to be independent. In addition this model selection robustness is an extension to the estimation consistency of the GEE estimator under the mis-specification of the underlying working correlation.

In Chapter 4 of this paper, we introduce General Information Criterion (GIC) and prove the model selection consistency that does not require independent working correlation matrix, which overcomes the limitation of QBIC at Chapter 3. Here we consider a goodness-of-fit measure instead of Pan (2001)'s style quasi-likelihood. Since the working covariance matrix is used to model the within cluster covariance

structure, we use the working covariance matrix and the fitted residuals together to construct a quadratic form which serve as the goodness-of-fit measure of the candidate model. Following the same technique, exact large deviation results from quadratic forms based on random vector satisfying the exponential moment conditions, we are able to choose the appropriate penalty size on the model complexity to ensure the model selection consistency for any working correlation matrix. Rather surprisingly, we are able to show the proposed information criterion has selection consistency for the marginal mean model even if the working correlation is mis-specified.

In Chapter 5, we summarize the contribution of this thesis that we firstly introduce the model selection consistency information criteria on GEE under the "large n, diverging p" framework. And we also discuss potential future work.

2 High Dimensional GEE

In this Chapter, we discuss the high dimensional GEE estimation consistency. In Section 2.1, we briefly summarize the previous work on GEE and its estimation consistency. In Section 2.2, we prove the high dimensional GEE consistency when we can choose any structure of working correlation matrix. Lastly we discuss the numerical simulation results in Section 2.3.

2.1 GEE and its Consistency

Suppose n clusters are randomly selected for the study. These could be people with repeated measurement. The size of the ith cluster is m_i. For cluster $i \in \{1, 2 \dots n\}$, let $Y_i = (Y_{i1} \dots \ Y_{im_i})^T$ be an $m_i \times 1$ response vector with mean $\mathrm{E}(Y_i) = \mu_i$, where $\mu_i = (\mu_{i1}, \mu_{i2} \dots \mu_{im_i})^T$. Let $X_i = (X_{i1}, X_{i2} \dots X_{im_i})^T$ denote the $m_i \times p_n$ design matrix of covariates for the ith cluster. We consider a marginal regression model: $g(\mu_{ij}) = x_{ij}^T \beta$, where $g(.)$ is a known link function, and $\beta = (\beta_1, \beta_2 \dots \beta_{p_n})^T$ de-

notes the p_n dimensional regression coefficients. Let A_i be a diagonal matrix with elements $\text{Var}(Y_{ij}) = \nu(\mu_{ij})/\phi$, where ϕ is the dispersion parameter and ν is the variance function. Let A_{ij} be the value of j-th row and j-th column of matrix A_i. We usually assume the size of different clusters is the same, that is m. Let R be the working correlation matrix and $V_i = A_i^{1/2} R A_i^{1/2}$ be the working covariance matrix. The true correlation matrix is denoted as R^*, which is usually unknown or not exist. The working correlation matrix R is user defined and could be either unstructured or structured such as independent, Autoregressive-1, or exchangeable (compound symmetry). The working correlation matrix $R(\varrho)$ involves unknown correlation parameter ϱ, which can be estimated through the method of moments or another set of estimating equations. Liang and Zeger (1986) proposed to use the following generalized estimating equation to solve for the unknown regression parameter:

$$U(\beta)|_{\beta=\widehat{\beta}} = \sum_{i=1}^{n} D_i(\beta)^T V_i(\beta)^{-1} \{Y_i - \mu_i(\beta)\}|_{\beta=\widehat{\beta}} = 0, \tag{2.1}$$

where $D_i(\beta) = \partial \mu_i(\beta)/\partial \beta^T$. When p_n is finite, the GEE solution $\widehat{\beta}$ satisfies that $||\widehat{\beta} - \beta^*|| = n^{-1/2}$ even with the mis-specified working correlation matrix R. Here β^* is the true parameter.

Wang (2011) further proved that under certain regularity conditions, if the number of regression parameters p_n is diverging and $p_n^2/n \to 0$, the GEE estimator $\widehat{\beta}$ is

$(p_n/n)^{1/2}$ consistent with following requirement and assumptions. The user defined working correlation matrix has to be of the particular format which is suggested by Balan and Schiopu-Kratina (2005),

$$\widehat{R} = \frac{1}{n}\sum_{i=1}^{n} A_i^{-1/2}(\widetilde{\beta})\{Y_i - \mu_i(\widetilde{\beta})\}\{Y_i - \mu_i(\widetilde{\beta})\}^T A_i^{-1/2}(\widetilde{\beta}), \tag{2.2}$$

where $\widetilde{\beta}$ is a preliminary consistent estimator when the working correlation matrix is identity matrix such that $\sum_{i=1}^{n} D_i(\widetilde{\beta})^T A_i(\widetilde{\beta})^{-1}\{Y_i - \mu_i(\widetilde{\beta})\} = 0$. Wang (2011) proved that under "large n diverging p_n" situation, the estimated working correlation matrix is $(p_n/n)^{1/2}$ consistent to the true correlation matrix. The consistency also requires below assumptions:

Assumption (A1) $\sup_{i,j} ||X_{ij}|| = O(p_n^{1/2})$. In this paper, $||.||$ denotes Euclidean norm.

Assumption (A2) the unknown parameter β belongs to a compact subset $\mathcal{B} \in R^{p_n}$, the true parameter value β^* lies in the interior of $\mathcal{B}$ and there exist two positive constants, b_{21} and b_{22}, such that $b_{21} \leq A_{ij}(\beta^*) \leq b_{22}$ for all i and j.

Assumption (A3) there exist two positive constants, b_{23} and b_{24}, such that $b_{23} \leq \lambda_{\min}\{n^{-1}\sum_{i=1}^{n} X_i^T X_i\} \leq \lambda_{\max}\{n^{-1}\sum_{i=1}^{n} X_i^T X_i\} \leq b_{24}$. In this paper $\lambda_{\max}(.)$ and $\lambda_{\min}(.)$ denote the largest and smallest eigenvalue of the matrix.

Assumption (A4) the common true correlation matrix R^* has eigenvalues bounded away from zero and infinity; the estimated working correlation matrix $\widehat{R}$ satisfies

$||\bar{R} - \widehat{R}|| = O_p\{(p_n/n)^{1/2}\}$, where $\bar{R}$ is a constant positive definite matrix with eigenvalues bounded away from zero and infinity; we do not require $\bar{R}$ to be the true correlation matrix R^*.

Assumption (A5) there exist finite positive constants $b_{25} > 0$ and $b_{26} > 0$ such that $\mathrm{E}[A_i^{-1/2}(\beta)\{Y_i - \mu_i(\beta)\}^{2+b_{26}} \leq b_{25}]$ for all i.

Assumption (A6) if $\mathcal{B} = \{\beta : ||\beta - \beta^*|| \leq (p_n/n)^{1/2}\}$, then $\partial\mu(X_{ij}^T\beta)/\partial(X_{ij}^T\beta)$ are uniformly bounded away from 0 and infinity on $\mathcal{B}$ for all i, j ; and the second and the third derivatives $\partial^2\mu(X_{ij}^T\beta)/\partial(X_{ij}^T\beta)^2$ and $\partial^3\mu(X_{ij}^T\beta)/\partial(X_{ij}^T\beta)^3$ are uniformly bounded by a finite positive constant b_{27} on $\mathcal{B}$ for all i, j.

Here we introduce some notation: $||.||_{\max}$ denotes the largest absolute value in the matrix or vector; $[.]_{[i,j]}$ denotes a element that is the ith row and jth column of matrix; $[.]_{[i,]}$ denotes a row vector that is the ith row matrix; $[.]_{[,j]}$ denotes a column vector that is the jth column matrix. The subscript indexes have the following convention: $i = 1, 2 \ldots n$ indexes the cluster number; $j, h = 1, 2 \ldots m$ indexes the number of observations in the given cluster; and $k, l, r = 1, 2 \ldots p_n$ indexes the different covariates.

2.2 High Dimensional GEE

This section discuss the GEE estimation consistency with any format of working correlation matrix. We introduce some assumptions.

Assumption 2.1 *We assume that* $p_n \to \infty$, $n \to \infty$, *and* $p_n^{4+\alpha'}/n \to 0$ *for every small* $\alpha' > 0$.

Assumption 2.2 *We define the Matrix* $\Omega(\beta^*) = -E\{U(\beta^*)^{(1)}\}/n$ *and assume all eigenvalues are bounded away from zero and infinity. Here* β^* *is the true parameters.*

Note that $\mathrm{E}\{Y - \mu(\beta^*)\} = 0$, we get $\Omega(\beta^*) = n^{-1}\sum_{i=1}^n D_i(\beta^*)^T V_i(\beta^*)^{-1} D_i(\beta^*) = n^{-1}\sum_{i=1}^n D_i(\beta^*)^T A_i(\beta^*)^{-1/2} R^{-1} A_i(\beta^*)^{-1/2} D_i(\beta^*)$. Therefore the Assumption 2.2 also implies that the user defined working correlation matrix R is positive definite and $A_{ij}(\beta^*)$ are uniformly bounded away from zero and infinity for all i and j.

In this thesis, large deviation results are used as an important tool to establish the estimation consistency and model selection consistency in large p settings. We define cumulant boundedness condition in Definition 1. Let ψ denote a random vector, O denote a positive definite matrix, and $\psi^T O \psi$ denote a quadratic form. Large deviation results for quadratic form $\psi^T O \psi$ were established by Spokoiny and Zhilova (2013) under exponential moment condition:

$$\log[\mathrm{E}\{\exp(t^T\psi)\}] \le ||t||^2/2, ||t|| \le b_{28},$$

where b_{28} is a positive constant. We first prove that such an exponential moment condition can be satisfied asymptotically by sample mean types of statistics, if the original random variables satisfy a cumulant boundedness condition.

Definition 1 *For a random vector Z of dimension m, let $C(t) = \log[E\{\exp(t^T Z)\}]$ denote its cumulant generating function, with t being a m-dimensional real vector. The cumulant boundedness condition requires that the first two derivatives of the cumulant generating function satisfy $|\partial C(0)/\partial t_k| \leq b_{29}$ and $|\partial^2 C(0)/\partial t_k \partial t_l| \leq b_{29}$. Furthermore there exists a constant b_t such that with $||t|| \leq b_t$, the absolute value of all the third derivatives of its cumulant generating function satisfy $|\partial^3 C(t)/\partial t_k \partial t_l \partial t_r| \leq b_{29}$.*

We define some terminology below. For $i \in \{1, 2 \ldots n\}$, let $U_i(\beta) = D_i(\beta)^T V_i(\beta)^{-1} \{Y_i - \mu_i(\beta)\}$. For $r \in \{1, 2 \ldots p_n\}$, let $U_i(\beta)_{[r]}$ denote the rth element of vector $U_i(\beta)$. For $k \in \{1, 2 \ldots p_n\}$ and $l \in \{1, 2 \ldots p_n\}$, $U_i(\beta)^{(1)}_{[rk]}$ denote $\partial U_i(\beta)_{[r]}/\partial \beta_k$, and $U_i(\beta)^{(2)}_{[rkl]}$ denote $\partial U_i(\beta)^{(1)}_{[rk]}/\partial \beta_l$. Here $U(\beta) = \sum_{i=1}^n U_i(\beta)$, $U(\beta)_{[r]} = \sum_{i=1}^n U_i(\beta)_{[r]}$, $U(\beta)^{(1)}_{[rk]} = \sum_{i=1}^n U_i(\beta)^{(1)}_{[rk]}$, and $U(\beta)^{(2)}_{[rkl]} = \sum_{i=1}^n U_i(\beta)^{(2)}_{[rkl]}$.

Assumption 2.3 *Assume for all $i \in \{1, 2 \ldots n\}$, and $r, k, l \in \{1, 2 \ldots p_n\}$ that each element of score function $U_i(\beta^*)_{[r]}$ and each element of its first derivatives $U_i(\beta^*)^{(1)}_{[rk]}$ satisfy the cumulant boundedness condition in Definition 1 uniformly. Also assume*

that there exists an neighborhood $||\beta - \beta^*|| \leq b_{nbh}$*, such that all the second derivatives of the score function* $U_i(\beta)^{(2)}_{[rkl]}$ *in that neighborhood satisfy the cumulant boundedness condition in Definition 1 uniformly*

Based on the cumulant boundedness condition in Assumption 2.3, using large deviation result in Spokoiny and Zhilova (2013) and Gao and Carroll (2017), we obtain the asymptotic orders of the following terms.

Lemma 2.1 *Under Assumption 2.3, for all* $k, l, r \in \{1, 2 \dots p_n\}$*, any* $\alpha' > 0$*,* $p_n \to \infty$*, and* β *in the neighborhoods* $||\beta - \beta^*|| \leq (p_n^{1+\alpha'}/n)^{1/2}$*, the zero-centered terms* $|U(\beta^*)_{[k]} - E\{U(\beta^*)_{[k]}\}|$*,* $|U(\beta^*)^{(1)}_{[kl]} - E\{U(\beta^*)^{(1)}_{[kl]}\}|$ *and* $|U(\beta)^{(2)}_{[klr]} - E\{U(\beta)^{(2)}_{[klr]}\}|$ *are of order* $O_p\{(np_n^{\alpha'})^{1/2}\}$ *uniformly.*

Proof of Lemma 2.1. We first prove the following claim Equation (2.3). Given independent variables $z_1, z_2 \dots z_n$ with mean zero and unit variance satisfying Cumulant Bounded Condition, for any real number b_c we have

$$\Pr\{\sum_{i=1}^{n} z_i > (nb_c^2)^{1/2}\} \leq \exp[-(1/2)b_c^2\{1 + o(1)\}]. \tag{2.3}$$

By Taylor expansion, the cumulant generating function for z_i is

$$C_i(t) = t^2/2 + C_i^{(3)}(t^*)t^3/6,$$

for some $0 < |t^*| < |t|$. Let $\bar{C}^{(3)}(t) = n^{-1}\sum_{i=1}^{n} C_i^{(3)}(t)$. Because each $C_i^{(3)}(t)$ is uniformly bounded for $|t| < b_t$, the average $\bar{C}^{(3)}(t)$ is also bounded for $|t| < b_t$. For

any $|t|/n^{1/2} < b_t$, the moment generating function of $n^{-1/2}\sum_{i=1}^n z_i$ is

$$\widetilde{C}_n(t) = \exp\{t^2/2 + \bar{C}_n^{(3)}(t^*/n^{1/2})t^3/(6n^{1/2})\}.$$

For any real number b_c and any $t > 0$, if $n^{-1/2}\sum_{i=1}^n z_i < b_c$, we know $I(n^{-1/2}\sum_{i=1}^n z_i > b_c) = 0 < \exp\{t(n^{-1/2}\sum_{i=1}^n z_i - b_c)\}$. If $n^{-1/2}\sum_{i=1}^n z_i \geq b_c$, we know $I(n^{-1/2}\sum_{i=1}^n z_i > b_c) = 1 \leq \exp\{t(n^{-1/2}\sum_{i=1}^n z_i - b_c)\}$. Combing above two situation, it follows that

$$I(n^{-1/2}\sum_{i=1}^n z_i > b_c) \leq \exp\{t(n^{-1/2}\sum_{i=1}^n z_i - b_c)\}.$$

Taking the expectation for both side, we have following inequality:

$$\begin{aligned}
&\Pr(n^{-1/2}\sum_{i=1}^n z_i > b_c) \leq \mathrm{E}[\exp\{t(n^{-1/2}\sum_{i=1}^n z_i - b_c)\}] \\
&= \exp\{t^2/2 + \bar{C}^{(3)}(t^*/n^{1/2})t^3/(6n^{1/2}) - b_c t\} \\
&= \exp\{[(t^2/2)\{1 + o(1)\} - b_c t]\}.
\end{aligned}$$

Let $b_c = t$, we prove Equation (2.3):

$$\Pr\{\sum_{i=1}^n z_i > (nb_c^2)^{1/2}\} \leq \exp[-(1/2)b_c^2\{1 + o(1)\}].$$

Let $b_c = p_n^{\alpha'}$, we get $\Pr\{\sum_{i=1}^n z_i > (np_n^{\alpha'})^{1/2}\} \leq \exp[-(1/2)p_n^{\alpha'}\{1 + o(1)\}] = o(1)$, when $p_n \to \infty$ and $\alpha' > 0$.

Because $U(\beta)$ satisfies the cumulant boundedness condition, its first and second moments are uniformly bounded. By Equation 2.3 we get $\Pr(\sum_{i=1}^n [U_i(\beta) - E\{U_i(\beta)\}]/\mathrm{Var}\{U_i(\beta)\} > (np_n^{\alpha'})^{1/2}) \to 0$. And b_{var} is the upper bound for $\mathrm{Var}\{U_i(\beta)\}$.

Similar arguments apply to the result for each element of its first and second derivatives. ■

Theorem 2.1 *Under Assumption 2.1 - 2.3, as* $p_n \to \infty$, $n \to \infty$, *there exists a solution* $\widehat{\beta}$ *to the score equation* $U(\widehat{\beta}) = 0$ *such that it falls within an* $(p_n^{1+\alpha'}/n)^{1/2}$ *neighborhood of* β^* *for all* $s \in S$ *with probability tending to 1. In other words, we have following estimation:*

$$||\widehat{\beta} - \beta^*|| = O_p\{(p_n^{1+\alpha'}/n)^{1/2}\}.$$

Proof of Theorem 2.1. To establish the existence of a consistent estimator $\widehat{\beta}$ within the specified neighborhood, we follow the approach from Portnoy (1988) and Wang (2011). It suffices to verify the following condition: $\forall \epsilon > 0$, there exists a constant $\Delta > 0$ such that for all n sufficiently large

$$\Pr\{\sup_{||\beta-\beta^*||=\Delta(p_n^{1+\alpha'}/n)^{1/2}} (\beta - \beta^*)^T U(\beta) < 0\} \geq 1 - \epsilon.$$

Let $\beta - \beta^* = \Delta(p_n^{1+\alpha'}/n)^{1/2}v$ and v is a unit vector such that $||v|| = 1$. By Taylor expansion, there exist a $\widetilde{\beta}$ between β and β^* such that $U(\beta) = U(\beta^*)+U(\widetilde{\beta})^{(1)}(\beta-\beta^*)$. We can rewrite $U(\widetilde{\beta})^{(1)}$ in terms of below format:

$$U(\widetilde{\beta})^{(1)} = n(\frac{1}{n}E\{U(\beta^*)^{(1)}\} + \frac{1}{n}[U(\beta^*)^{(1)} - \mathrm{E}\{U(\beta^*)^{(1)}\}] + \frac{1}{n}\{U(\widetilde{\beta})^{(1)} - U(\beta^*)^{(1)}\}).$$

By Assumption 2.2, $-\mathrm{E}[U(\beta^*)^{(1)}]/n = \Omega(\beta^*)$ is positive definite with bounded eigenvalues. From Lemma 2.1 we also know that for rth row and kth column of matrix

$n^{-1}[U(\beta^*)^{(1)} - E\{U(\beta^*)^{(1)}\}]_{[rk]} = O_p\{(p_n^{\alpha'}/n)^{1/2}\}$. We also know that there is a $\check{\beta}$ between $\widetilde{\beta}$ and β^* such that

$$|\frac{1}{n}\{U(\widetilde{\beta})^{(1)}_{[rk]} - U(\beta^*)^{(1)}_{[rk]}\}| = |\frac{1}{n}U(\check{\beta})^{(2)}_{[rk]}(\widetilde{\beta} - \beta^*)| \leq \frac{1}{n}||U(\check{\beta})^{(2)}_{[rk]}|| \times ||\widetilde{\beta} - \beta^*||,$$

where $U(\check{\beta})^{(2)}_{[rk]} = (U(\check{\beta})^{(2)}_{[rk1]}, U(\check{\beta})^{(2)}_{[rk2]} \ldots U(\check{\beta})^{(2)}_{[rkp_n]})^T$ is a $p_n \times 1$ vector. Since $\check{\beta}$ is between $\widetilde{\beta}$ and β^* and $\widetilde{\beta}$ is between β and β^*, therefore $||\widetilde{\beta} - \beta^*|| = O_p\{(p_n^{1+\alpha'}/n)^{1/2}\}$. We can reformulate:

$$\frac{1}{n}U(\check{\beta})^{(2)}_{[rk]} = \frac{1}{n}\{U(\check{\beta})^{(2)}_{[rk]} - \mathrm{E}[U(\check{\beta})^{(2)}_{[rk]}]\} + \frac{1}{n}\mathrm{E}[U(\check{\beta})^{(2)}_{[rk]}].$$

From Lemma 2.1, $n^{-1}\{U(\check{\beta})^{(2)}_{[rkl]} - \mathrm{E}[U(\check{\beta})^{(2)}_{[rkl]}]\} = O\{(p_n^{\alpha'}/n)^{1/2}\}$. This entails $||U(\check{\beta})^{(2)}_{[rkl]} - \mathrm{E}[U(\check{\beta})^{(2)}_{[rkl]}]||/n = O_p\{(p_n^{1+\alpha'}/n)^{1/2}\}$. From Assumption 2.3, the cumulant bounded condition indicates that $\mathrm{E}[U_i^{(2)}(\check{\beta})_{[rkl]}] = C_{U_i^{(2)}(\check{\beta})_{[rkl]}}(0)$ is uniformly bounded. Then $n^{-1}||\mathrm{E}[U(\check{\beta})^{(2)}_{[rk]}]|| = O_p(p_n^{1/2})$. It indicates that $n^{-1}\{U(\widetilde{\beta})^{(1)} - U(\beta^*)^{(1)}\}_{[rk]} = O_p\{(p_n^{2+\alpha'}/n)^{1/2}\}$. From above equations, we get $U(\widetilde{\beta})^{(1)} = -n\{\Omega(\beta^*) + Res'\}$, where each element in Res' is $O_p\{(p_n^{2+\alpha'}/n)^{1/2}\}$. We also know that $\mathrm{E}[U(\beta^*)] = 0$ since $\mathrm{E}[Y_i] = \mu_i(\beta^*)$. From Lemma 2.1, we know $||U(\beta^*)|| = ||U(\beta^*) - \mathrm{E}[U(\beta^*)]|| = O_p\{(np_n^{1+\alpha'})^{1/2}\}$. There exists a constant number b_u such that $||U(\beta^*)|| \leq b_u(np_n^{1+\alpha'})^{1/2}$. In addition, we have $|v^T Res'v| = |\sum_{kr} v_k v_r Res'_{kr}| \leq \max_{kr}|Res'_{kr}| \times p_n \times ||v||^2 = O_p\{(p_n^{4+\alpha'}/n)^{1/2}\} =$

$o_p(1)$. Combining the result above, we have

$$
\begin{aligned}
&(\beta-\beta^*)^T U(\beta) \\
&= (\beta-\beta^*)^T U(\beta^*) + (\beta-\beta^*)^T U(\widetilde{\beta})^{(1)}(\beta-\beta^*) \\
&= \Delta(p_n^{1+\alpha'}/n)^{1/2} v^T U(\beta^*) - \Delta^2(p_n^{1+\alpha'}/n) v^T n\{\Omega(\beta^*) + Res'\}v \\
&\leq \Delta(p_n^{1+\alpha'}/n)^{1/2}||v|| \times ||U(\beta^*)|| - \Delta^2 p_n^{1+\alpha'}[\lambda_{\min}\{\Omega(\beta^*)\} + o_p(1)]||v||^2 \\
&\leq \Delta(p_n^{1+\alpha'}/n)^{1/2}(np_n^{1+\alpha'})^{1/2} b_u - \Delta^2 p_n^{1+\alpha'}[\lambda_{\min}\{\Omega(\beta^*)\} + o_p(1)] \\
&= p_n^{1+\alpha'}(b_u\Delta - [\lambda_{\min}\{\Omega(\beta^*)\} + o_p(1)]\Delta^2).
\end{aligned}
$$

Therefore by choosing Δ large enough, $(\beta-\beta^*)^T U(\beta)$ on $\{\beta : ||\beta-\beta^*|| = \Delta(p_n^{1+\alpha'}/n)^{1/2}\}$ is negative. ■

Comparing to previous work from Wang (2011) which requires choice of working correlation matrix has to be the form from Equation (2.2), the Theorem 2.1 has demonstrated that GEE estimation is still consistency no matter the choice of working correlation matrix as long as it is positive definite.

2.3 Numerical Study

In this section, we conduct simulations on both clustered binary variables for discrete case and clustered Gaussian variable for continuous case. Regarding to working correlation matrix, Section 2.2 illustrates that the estimation is consistency with the working correlation matrix R being any arbitrary positive definite matrices.

In our simulation, we compare different choices of R including Independent (Ind), Exchangeable (Exc), Autoregressive-1 (AR1), and unstructured (Uns) working correlation matrix. The independent correlation matrix is a m by m identity matrix. The exchangeable working correlation matrix has diagonal elements equal to 1 and off-diagonal elements equal to a constant number b_ρ. The Autoregressive-1 working correlation matrix has diagonal elements equal to 1 and off-diagonal elements equal to $b_\rho^{|j-j'|}$, where j and j' are column and row numbers. Here b_ρ can be estimated through the method of moments or another set of estimating equations. The unstructured working correlation matrix comes from Equation (2.2).

We simulate both binary and Gaussian response variables Y_{ij}. We consider different settings with sample size $n = 500$, 1000, or 2000, the number of covariates $p_n = 20$, 60, or 200, and the cluster size $m = 10$. For $j = 1, 2 \dots p_n/2$, β_j is drawn from the uniform distribution $U(0.05, 0.5)$. For $j = p_n/2 + 1, p_n/2 + 2 \dots p_n$, we define $\beta_j = -\beta_1$. For the jth observation in the ith cluster, we simulate the associated covariates $X_{ij} = (x_{ij1}, x_{ij2} \dots x_{ijp_n})^T$, and the mean parameter is denoted as $\mu_{ij} = \text{logit}^{-1}(X_{ij}^T\beta)$ for binary response. Similarly for Gaussian response, the mean parameter is denoted as $\mu_{ij} = X_{ij}^T\beta$ and the variance is set as 1. The covariates X_{ijk} are partitioned into independent blocks of 10 covariates, and within each block the 10 covariates are simulated from the multivariate Gaussian distribution with variances

equal to 1 and off-diagonal covariances all equal to $0.5^{|k-k'|}$, where k and k' index for the covariates. For each cluster i, Y_i is simulated from a multivariate binary or Gaussian distribution with mean μ_i and an unstructured correlation matrix. The r package "SimCorMultRes" is used to simulate the correlated multivariate binary distribution.

We measure the accuracy of estimation by the simulated average mean square error, which is obtained by averaging $||\widehat{\beta} - \beta^*||^2/p_n$ over 100 simulated samples. Table 2.1 and 2.2 illustrate the simulated mean and standard deviation of binary and Gaussian response. We observe that the smaller p_n/n ratio the better convergence of the estimation. The choice of unstructured correlation matrix from Equation (2.2) usually has better convergence comparing to other structured correlation matrices. The rationale is that the choice of working correlation from Equation (2.2) converges to the true correlation (Wang, 2011). Although the type of working correlation matrix will not impact the convergence, choosing the approximately true correlation matrix usually has smaller estimation error. Theoretically Chaganty and Joe (2004) demonstrated that GEE for binary distribution with an appropriately chosen working correlation matrix does have good efficiency.

Table 2.1: GEE Simulation Result for Binary Response

The average mean squared error for 100 simulation

		p 20		p 60		p 200	
		mean	std	mean	std	mean	std
n 500	Ind	0.0025	0.0010	0.0043	0.0010	0.0156	0.0036
	AR1	0.0025	0.0010	0.0043	0.0010	0.0156	0.0036
	Exc	0.0024	0.0010	0.0043	0.0010	0.0156	0.0036
	Uns	0.0023	0.0009	0.0042	0.0010	0.0155	0.0036
n 1000	Ind	0.0013	0.0004	0.0022	0.0005	0.0055	0.0011
	AR1	0.0013	0.0004	0.0022	0.0005	0.0055	0.0011
	Exc	0.0013	0.0004	0.0022	0.0005	0.0055	0.0011
	Uns	0.0011	0.0004	0.0021	0.0004	0.0054	0.0011
n 2000	Ind	0.0006	0.0002	0.0011	0.0002	0.0023	0.0004
	AR1	0.0006	0.0002	0.0011	0.0002	0.0023	0.0004
	Exc	0.0006	0.0002	0.0011	0.0002	0.0023	0.0004
	Uns	0.0005	0.0002	0.0011	0.0002	0.0023	0.0004

Table 2.2: GEE Simulation Result for Normal Response

The average mean squared error for 100 simulation

Note all mean and std have been amplified 1000 times

		p 20		p 60		p 200	
		mean	std	mean	std	mean	std
n=500	Ind	0.3218	0.1107	0.3274	0.0710	0.3479	0.0416
	AR1	0.3213	0.1108	0.3268	0.0708	0.3478	0.0416
	Exc	0.3088	0.1129	0.3201	0.0698	0.3355	0.0403
	Uns	0.0733	0.0241	0.0814	0.0167	0.1065	0.0154
n=1000	Ind	0.1554	0.0531	0.1577	0.0372	0.1697	0.0200
	AR1	0.1553	0.0530	0.1576	0.0371	0.1697	0.0200
	Exc	0.1513	0.0508	0.1539	0.0367	0.1650	0.0189
	Uns	0.0359	0.0131	0.0379	0.0085	0.0433	0.0050
n=2000	Ind	0.0790	0.0277	0.0788	0.0155	0.0815	0.0100
	AR1	0.0791	0.0277	0.0788	0.0154	0.0815	0.0100
	Exc	0.0781	0.0279	0.0777	0.0151	0.0797	0.0099
	Uns	0.0194	0.0077	0.0192	0.0039	0.0201	0.0024

3 Quasi-likelihood Bayesian Information Criterion (QBIC)

In this Chapter, we introduce Quasi-likelihood Bayesian Information Criterion, named QBIC, a new model selection criterion on GEE. Under the "large n and diverging p_n" framework, it is the first time that there is a model selection consistency information criterion on GEE. But the new proposed QBIC has a limitation that the choice of working correlation matrix has to be an identity matrix. The chapter is structured as follows. The Section 3.1 introduces the previous work on GEE model selection and some notations. The Section 3.2 introduces the new proposed QBIC. When we discuss the model selection, it is unavoidable to mention the underfitting and overfitting models. The Section 3.3 defines the underfitting or overfitting models. The Section 3.4 proves the model selection consistency of QBIC. The Section 3.5 shows the numerical simulation results. Lastly Section 3.6 lists few extra lemmas and provide the proof details of all Lemmas which are too long to put in the main

text.

3.1 Previous Model Selection on GEE

One of the challenging part of GEE model selection is that the distribution and likelihood of GEE is unknown given that GEE only requires on the first two moments of outcome variables. Given the observations Y_{ij} with expectations μ_{ij} and variances $\nu(\mu_{ij})$, where $\nu(.)$ is a known variance function, Wedderburn (1974) introduced the quasi-likelihood $QL'(Y_{ij}, \mu_{ij})$ in the partial derivative format:

$$\frac{\partial QL'(Y_{ij}, \mu_{ij})}{\partial \mu_{ij}} = \frac{Y_{ij} - \mu_{ij}}{\nu(\mu_{ij})}. \tag{3.1}$$

Even though there is partial derivative style quasi-likelihood, the integration of quasi-likelihood is a line integral and path dependent. In other words, it is impossible to write down the full quasi-likelihood explicitly in an integral format. Pan (2001) has simplified the quasi-likelihood definition by assuming the correlation matrix in each cluster is an identity matrix and extending the AIC to Quasi-likelihood Information Criterion (QIC):

$$QIC(s) = -2QL_{\phi}\{\widehat{\beta}_s(R)\} + 2\,\mathrm{Tr}[\Omega\{\beta_s(I)\}^{-1}W\{\beta_s(R)\}]. \tag{3.2}$$

McCullagh and Nelder (1989) and Pan (2001) have indicated that the integration of quasi-likelihood is no longer path dependent when the identity working correlation

matrix. The simplified quasi-likelihood QL_ϕ has below form:

$$QL_\phi\{\widehat{\beta}_s(R)\} = \sum_{i=1}^{n}\sum_{j=1}^{m}\int_{Y_{ij}}^{g^{-1}\{X_{ij}\widehat{\beta}_s(R)\}} \frac{Y_{ij}-t}{\phi A_{ij}(t)}dt. \tag{3.3}$$

Here GEE estimator $\widehat{\beta}_s(R)$ satisfies that $\sum_{i=1}^{n} D_i^T\{\widehat{\beta}_s(R)\}V_i^{-1}\{\widehat{\beta}_s(R)\}[Y_i-\mu_i\{\widehat{\beta}_s(R)\}] = 0$ with working correlation matrix R. And GEE estimator $\widehat{\beta}_s(I)$ satisfies that $\sum_{i=1}^{n} D_i^T\{\widehat{\beta}_s(I)\}A_i^{-1}\{\widehat{\beta}_s(I)\}[Y_i-\mu_i\{\widehat{\beta}_s(I)\}] = 0$ with identity working correlation matrix I. ϕ is the dispersion parameter, which is useful in modeling under-dispersion or over-dispersion scenario. $\mathrm{Tr}[\Omega\{\beta_s(I)\}W\{\beta_s(R)\}]$ is the effective degrees of freedom of the model s (Varin and Vidoni, 2005; Gao and Song, 2010) with the sensitivity matrix $\Omega\{\beta_s(I)\} = n^{-1}\mathrm{E}[-\partial^2 QL\{\beta_s(I)\}/\partial\beta\partial\beta^T]$ and the variability matrix $W\{\beta_s(R)\} = n^{-1}\mathrm{Cov}[U\{\beta_s(R)\}]$.

Pan (2001)'s QIC is AIC type of information criterion which minimizes the K-L divergence and is not model selection consistent. Wang and Qu (2009) proposed BIC style of information criterion based on Qu et al. (2000)'s Quadratic Inference Function (QIF) and called Bayesian Information Quadratic Inference Function (BIQIF):

$$BIQIF(s) = Q_{QIF}(\widehat{\beta}_s) + d_s \log n. \tag{3.4}$$

Here d_s is the number of covariates in model s. The QIF type of goodness of fitting is defined as $Q_{QIF}(\beta) = nG(\beta)^T\{n^{-1}\sum_{i=1}^{n} G_i(\beta)G_i(\beta)^T\}^{-1}G(\beta)$. And the function

G has following format:

$$G(\beta) = n^{-1} \sum_{i=1}^{n} G_i(\beta) = n^{-1} \sum_{i=1}^{n} \begin{bmatrix} D_i(\beta) A_i^{-1/2}(\beta) T_1 A_i^{-1/2}(\beta) \{Y_i - \mu_i(\beta)\} \\ \dots \\ D_i(\beta) A_i^{-1/2}(\beta) T_k A_i^{-1/2}(\beta) \{Y_i - \mu_i(\beta)\} \end{bmatrix}.$$

The model selection consistency has been approved for bounded size of covariates $p = O(1)$. The QIF approach is motivated by the observation that the inverse of the commonly used working correlation structures can be exactly represented or approximated by a linear combination of basis matrices $R^{-1} = \sum_{i=1}^{k} a_i T_i$, where $\{a_i\}_{i=1}^{k}$ are constant coefficients. T_1 is identity matrix and $T_2, T_3 \dots T_k$ are basis matrices such that for any pairs of (i, j) there exists l satisfying $T_i T_j = T_l$ for $i, j, l \in \{1, 2 \dots k\}$. The most challenging part of BIQIF is that it is not easy and common to decompose the inverse of correlation matrix into a list of basis matrices. And therefore the BIQIF is unpopular and rarely used.

3.2 Introduction to QBIC

We propose the below Quasi-likelihood Bayesian Information Criterion (QBIC) for model selection on GEE models:

$$QBIC(s) = -2QL(\widehat{\beta}_s) + d_s^* \gamma_n. \tag{3.5}$$

The first term of QBIC is the simplified quasi-likelihood

$$QL(\widehat{\beta}_s) = \sum_{i=1}^{n}\sum_{j=1}^{n_i}\int_{Y_{ij}}^{g^{-1}(X_{ij}\widehat{\beta}_s)} \frac{Y_{ij}-t}{A_{ij}(t)}dt, \tag{3.6}$$

which is similar to Equation (3.3) in Pan (2001)'s QIC. The difference between Equation (3.3) and Equation (3.6) is that we set the dispersion parameter $\phi = 1$ in Equation (3.6). Quasi-likelihood reflects the goodness-of-fit for a given model s. The second term is the penalty for model complexity, which enforces sparsity on the selected model. The γ_n is a sequence of penalties on the complexity of the model, and d_s^* is the effective degrees of freedom of the model s (Pan, 2001; Varin and Vidoni, 2005; Gao and Song, 2010). We define $d_s^* = \mathrm{Tr}\{W_s(\beta_s^*)\Omega_s^{-1}(\beta_s^*)\}$, where the variability matrix $W(\beta_s^*) = n^{-1}\mathrm{Cov}\{U(\beta_s^*)\}$ and the sensitivity matrix $\Omega(\beta_s^*) = -n^{-1}\mathrm{E}\{\partial U(\beta_s^*)/\partial\beta_s^T\}$. If the true correlation is identity matrix, and the marginal regression model is the true model, the variability matrix and sensitivity matrix are the same. If the model s is the overfitting model, as $\mathrm{E}\{Y_i-\mu_i(\beta_s^*)\} = 0$, the variability matrix and sensitivity matrix can be expressed as $W(\beta_s^*) = n^{-1}\sum_{i=1}^{n} D_i(\beta_s^*)^T A_i(\beta_s^*)^{-1}\mathrm{Cov}(Y_i)A_i(\beta_s^*)^{-1}D_i(\beta_s^*)$ and $\Omega(\beta_s^*) = n^{-1}\sum_{i=1}^{n} D_i(\beta_s^*)^T A_i(\beta_s^*)^{-1}D_i(\beta_s^*)$.

3.3 Overfitting and Underfitting Model

Model selection is the task of selecting a statistical model from a set of candidate models. It is unavoidable to discuss the overfitting models and underfitting models

among all candidate models. This section defines the overfitting models and underfitting models and their corresponding pseudo true parameters. Let s be a subset of $(1, 2 \dots p_n)$. The kth element of vector β denotes as $\beta_{[k]}$. The model with $\beta_{[k]} = 0$ for all $k \notin s$ is called model s.

Let T denote the true model and d_T be the size of the true model T. Let β_T^* denote the true values of the parameters under the model T. Consider all the competing models s in the model space S. Let d_s denote the number of covariates in the model s, with s_n being the upper bound of d_s in S, and $d_T \leq s_n \leq p_n$. If s is overfitting, $T \subseteq s$; whereas if s is underfitting, $T \nsubseteq s$. The sets of underfitting models and overfitting models are denoted as S_- and S_+ respectively. Note that $T \in S_+$.

The true parameter values under an overfitting model s are denoted as β_s^*, where the common d_T elements are the same as β_T^* and the rest of $d_s - d_T$ elements are zero. For any underfitting model $s \in S_-$, we assume there exists a unique pseudo true parameters β_s^* such that $\sum_{i=1}^n D_i(\beta_s^*)^T V_i(\beta_s^*)^{-1}\{\mu_i(\beta_T^*) - \mu_i(\beta_s^*)\} = 0$. This definition of pseudo true parameter values is similar to the definition used in the maximum likelihood estimation under mis-specified models (White, 1981, 1982).

In this section, we can prove that the GEE estimator will still converge to the pseudo true parameter for any given model s. Comparing to Chapter 2 which discusses the consistency under one true model, this chapter discusses the model se-

lection under larger model space $s \in S$. There could be up to $p_n^{s_n}$ models. The Assumption 2.2, Lemma 2.1, and Theorem 2.1 have new version under model space $s \in S$. Similarly to Assumption 2.2, which only applies to the true model, we extend the previous assumption to all models $s \in S$.

Assumption 3.1 *For all models $s \in S$, the variability matrix and sensitivity matrix $\Omega(\beta_s^*)$ and $W(\beta_s^*)$ are positive definite matrices and their eigenvalues are uniformly bounded away from zero and infinity.*

Assumption 3.2 *There exists a neighborhood $||\beta_s - \beta_s^*|| \leq (s_n^2 \log p_n/n)^{1/2}$, such that $QL_i(\beta_s^*)$,$U_i(\beta_s^*)_{[k]}$,$U_i(\beta_s^*)_{[kl]}^{(1)}$, and $U_i(\beta_s)_{[klr]}^{(2)}$ satisfy the cumulant boundedness condition in Definition 1 uniformly for all models $s \in S$, $i \in \{1, 2 \dots n\}$ and $k, l, r \in \{1, 2 \dots p_n\}$.*

Based on the cumulant boundedness condition in Assumption 3.2, using large deviation result in Spokoiny and Zhilova (2013) and Gao and Carroll (2017), we obtain the asymptotic orders of the following terms.

Lemma 3.1 *Under Assumption 3.2, for all $k, l, r \in \{1, 2 \dots p_n\}$, all models $s \in S$, and β_s in the neighborhoods $||\beta_s - \beta_s^*|| \leq (s_n^2 \log p_n/n)^{1/2}$, the zero-centered terms $|QL(\beta_s^*) - E\{QL(\beta_s^*)\}|$, $|U(\beta_s^*)_{[k]} - E\{U(\beta_s^*)_{[k]}\}|$, $|U(\beta_s^*)_{[kl]}^{(1)} - E\{U(\beta_s^*)_{[kl]}^{(1)}\}|$, and $|U(\beta_s)_{[klr]}^{(2)} - E\{U(\beta_s)_{[klr]}^{(2)}\}|$ are of order $O_p\{(ns_n \log p_n)^{1/2}\}$ uniformly.*

Lemma 3.1 has extra order terms comparing to Lemma 2.1 due to we have up to $p_n^{s_n}$ model selection choice instead of 1 true model. Theorem 2.1 illustrates that the GEE of true model is consistent estimator; we can also extend it to the overfitting and underfitting models.

Theorem 3.1 *Under Assumptions 3.1 and 3.2, as* $p_n \to \infty$, $s_n^5 \log p_n = o(n)$, *there exists a solution* $\widehat{\beta}_s$ *to the score equation* $U(\widehat{\beta}_s) = 0$ *such that it falls within* $(s_n^2 \log p_n/n)^{1/2}$ *neighborhood of* β_s^* *for all* $s \in S$ *with probability tending to 1. Mathematically we have:*

$$||\widehat{\beta}_s - \beta_s^*|| = O_p\{(s_n^2 \log p_n/n)^{1/2}\}.$$

Proof of Theorem 3.1. To establish the existence of a consistent estimator $\widehat{\beta}_s$ within the specified neighborhood, we follow the approach from Portnoy (1988) and Wang (2011). It suffices to verify the following condition: $\forall\, \epsilon > 0$, there exists a constant $\Delta > 0$ such that for all n sufficiently large,

$$\Pr[\cap_{s\in S}\{\sup_{||\beta_s-\beta_s^*||=\Delta(s_n^2 \log p_n/n)^{1/2}} (\beta_s - \beta_s^*)^T U(\beta_s) < 0\}] \geq 1 - \epsilon.$$

Let $\beta_s - \beta_s^* = \Delta(s_n^2 \log p_n/n)^{1/2} v$, where v is a unit vector with $||v|| = 1$. By Taylor expansion, there exists a $\widetilde{\beta}_s$ between β_s and β_s^* such that $U(\beta_s) = U(\beta_s^*) + U(\widetilde{\beta}_s)^{(1)}(\beta_s - \beta_s^*)$. We reformulate $U(\widetilde{\beta}_s)^{(1)}$ as

$$n(\frac{1}{n}\mathrm{E}\{U(\beta_s^*)^{(1)}\} + \frac{1}{n}[U(\beta_s^*)^{(1)} - \mathrm{E}\{U(\beta_s^*)^{(1)}\}] + \frac{1}{n}\{U(\widetilde{\beta}_s)^{(1)} - U(\beta_s^*)^{(1)}\}).$$

By Assumption 3.1, $-n^{-1}\mathrm{E}[U(\beta_s^*)^{(1)}] = \Omega(\beta_s^*)$, which is a positive definite matrix with bounded eigenvalues. From Lemma 3.1, the $(r,k)th$ entry of the matrix $n^{-1}[U(\beta_s^*)^{(1)} - \mathrm{E}\{U(\beta_s^*)^{(1)}\}]_{[rk]} = O_p\{(s_n \log p_n/n)^{1/2}\}$. There exists a $\check{\beta}_s$ between $\widetilde{\beta}_s$ and β_s^* such that

$$|\frac{1}{n}\{U(\widetilde{\beta}_s)^{(1)}_{[rk]} - U(\beta_s^*)^{(1)}_{[rk]}\}| = |\frac{1}{n}U(\check{\beta}_s)^{(2)}_{[rk]}(\widetilde{\beta}_s - \beta_s^*)| \leq \frac{1}{n}||U(\check{\beta}_s)^{(2)}_{[rk]}|| \times ||\widetilde{\beta}_s - \beta_s^*||,$$

where $U(\check{\beta}_s)^{(2)}_{[rk]} = (U(\check{\beta}_s)^{(2)}_{[rk1]}, U(\check{\beta}_s)^{(2)}_{[rk2]} \ldots U(\check{\beta}_s)^{(2)}_{[rkp_n]})^T$ is a $d_s \times 1$ vector. Since $\check{\beta}_s$ is between $\widetilde{\beta}_s$ and β_s^*,$||\widetilde{\beta}_s - \beta_s^*|| = O_p\{(s_n^2 \log p_n/n)^{1/2}\}$. We reformulate:

$$\frac{1}{n}U(\check{\beta}_s)^{(2)}_{[rk]} = \frac{1}{n}[U(\check{\beta}_s)^{(2)}_{[rk]} - \mathrm{E}\{U(\check{\beta}_s)^{(2)}_{[rk]}\}] + \frac{1}{n}\mathrm{E}\{U(\check{\beta}_s)^{(2)}_{[rk]}\}.$$

From Lemma 3.1,$n^{-1}[U(\check{\beta}_s)^{(2)}_{[rkl]} - \mathrm{E}\{U(\check{\beta}_s)^{(2)}_{[rkl]}\}] = O_p\{(s_n \log p_n/n)^{1/2}\}$. This entails $n^{-1}||U(\check{\beta}_s)^{(2)}_{[rk]} - \mathrm{E}\{U(\check{\beta}_s)^{(2)}_{[rk]}\}|| = O_p\{(s_n^2 \log p_n/n)^{1/2}\}$. From Assumption 3.2, $\mathrm{E}\{U_i^{(2)}(\check{\beta}_s)_{[rkl]}\}$ is bounded. Then $n^{-1}||\mathrm{E}\{U(\check{\beta}_s)^{(2)}_{[rk]}\}|| = O_p\{(s_n)^{1/2}\}$. This implies $n^{-1}\{U(\widetilde{\beta}_s)^{(1)} - U(\beta_s^*)^{(1)}\}_{[rk]} = O_p\{(s_n^2 \log p_n/n)^{1/2}\}$. Thus, $U(\widetilde{\beta}_s)^{(1)} = -n\{\Omega(\beta_s^*) + \mathrm{Res}\}$, where I is an identity matrix and each element in the residual matrix Res is $O_p\{(s_n^3 \log p_n/n)^{1/2}\}$. For overfitting models, $\mathrm{E}\{U(\beta_s^*)\} = 0$. For underfitting models, based on the definition of β_s^*, we know $\mathrm{E}\{U(\beta_s^*)\} = \mathrm{E}[\sum_{i=1}^n D_i(\beta_s^*)^T V_i(\beta_s^*)\{Y_i - \mu_i(\beta_s^*)\}] = \mathrm{E}[\sum_{i=1}^n D_i(\beta_s^*)^T V_i(\beta_s^*)\{Y_i - \mu_i(\beta_T^*)\}] + \sum_{i=1}^n D_i(\beta_s^*)^T V_i(\beta_s^*)\{\mu_i(\beta_T^*) - \mu_i(\beta_s^*)\} = 0$ as well. From Lemma 3.1, we have $||U(\beta_s^*)|| = ||U(\beta_s^*) - \mathrm{E}\{U(\beta_s^*)\}|| = O_p\{(ns_n^2 \log p_n)^{1/2}\}$. Thus there exists a constant number b_u such that $||U(\beta_s^*)|| \leq$

$b_u(ns_n^2 \log p_n)^{1/2}$ for n sufficiently large. In addition, we have

$$\begin{aligned} v^T Res v &= \sum_{kr} v_k v_r Res_{kr} \\ &\geq -\max_{kr} |Res_{kr}| \sum_{kr} |v_k| \times |v_r| \\ &\geq -\max_{kr} |Res_{kr}| \times s_n \times ||v||^2 \\ &= O_p\{(s_n^5 \log p_n/n)^{1/2}\} \end{aligned}$$

Combining the results above, we have

$$\begin{aligned} &(\beta_s - \beta_s^*)^T U(\beta_s) \\ =&(\beta_s - \beta_s^*)^T U(\beta_s^*) + (\beta_s - \beta_s^*)^T U(\widetilde{\beta}_s)^{(1)}(\beta_s - \beta_s^*) \\ =&\Delta(s_n^2 \log p_n/n)^{1/2} v^T U(\beta_s^*) - \Delta^2 (s_n^2 \log p_n/n) v^T n\{\Omega(\beta_s^*) + Res\} v \\ \leq&\Delta(s_n^2 \log p_n/n)^{1/2} ||v|| \times ||U(\beta_s^*)|| - \Delta^2 s_n^2 \log p_n ||v||^2 [\lambda_{\min}\{\Omega(\beta_s^*)\} + o_p(1)] \\ \leq&\Delta(s_n^2 \log p_n/n)^{1/2} b_u (ns_n^2 \log p_n)^{1/2} - \Delta^2 s_n^2 \log p_n [\lambda_{\min}\{\Omega(\beta_s^*)\} + o_p(1)] \\ =&s_n^2 \log p_n (b_u \Delta - [\lambda_{\min}\{\Omega(\beta_s^*)\} + o_p(1)]\Delta^2). \end{aligned}$$

Therefore by choosing Δ large enough, $(\beta_s - \beta_s^*)^T U(\beta_s)$ is negative for all $\{\beta_s : ||\beta_s - \beta_s^*|| = \Delta(s_n^2 \log p_n/n)^{1/2}\}$ and all $s \in S$. ■

Theorem 3.1 implies that the GEE estimator has a convergence rate of $(s_n^2 \log p_n/n)^{1/2}$ uniformly for all $s \in S$. Compared to the convergence rate of $(p_n/n)^{1/2}$ established in Wang (2011) and convergence rate of $(p_n^{1+\tau}/n)^{1/2}$ in Theorem 2.1 for the true model, this uniform convergence rate has an extra factor of $(s_n \log p_n)^{1/2}$ due to the

multitude of competing models that causes Lemma 3.1 having extra order terms comparing to Lemma 2.1.

3.4 QBIC Model Selection Consistency

In this section, we establish the model selection consistency of the proposed QBIC under "large n and divergent p_n scenario". Our approach consists of two steps. First, we show that the difference in the simplified quasi-likelihood measures between two competing models s and T can be approximated by quadratic forms and the approximation errors are uniformly bounded across the model spaces. Second, based on the quadratic forms, we apply the large deviation result to quantify the size of the penalty γ_n.

Assumption 3.3 *We assume the working independence model. In other words, the working correlation matrix is identity matrix* $R = I$.

The rationale of the identity correlation matrix is that the simplified quasi-likelihood assumes the independent inner cluster structure to make the integral path independence. To prove the consistency also requires identity working correlation matrix. Mathematically we have $R = I$ and $V_i(\beta) = A_i(\beta)$ at the remaining part of this chapter. And the GEE estimator $\widehat{\beta}$ satisfies that $\sum_{i=1}^{n} D_i(\widehat{\beta})^T A_i(\widehat{\beta})^{-1}\{Y_i - \mu_i(\widehat{\beta})\} = 0$.

Assumption 3.4 *The cluster size m is finite, the number of covariates goes to infinity $p_n \to \infty$, and $s_n^5 \log p_n/n \to 0$. Assume that* $\liminf_n \min_{s \in S_-} n^{-1}[\sum_{i=1}^n \{\mu_i(\beta_T^*) - \mu_i(\beta_s^*)\}^T \{\mu_i(\beta_T^*) - \mu_i(\beta_s^*)\}]/(s_n^3 \log p_n/n)^{1/2} = \infty$.

Assumption 3.4 implies that the minimum distance between the true model T and any competing underfitting model s is allowed to converge to zero but at a rate slower than $(s_n^3 \log p_n/n)^{1/2}$.

Assumption 3.5 *For all $s \in S$ and any β in the neighborhood of $||\beta - \beta_s^*|| \leq (s_n^2 \log p_n/n)^{1/2}$, there exist positive values b_{33} and b_{34} such that $A_{ij}(\beta_s^*) > b_{33}$, and $|\mu_{ij}(\beta)|$, $|\partial \mu_{ij}(\beta)/\partial \beta_{[k]}|$, $|A_{ij}(\beta)|$, $|Cov(Y_{ij})|$ exist and bounded from above by b_{34}, for all $i \in \{1, 2 \ldots n\}$, $j \in \{1, 2 \ldots m\}$, $k \in \{1, 2 \ldots p_n\}$.*

Lemma 3.2 *Under Assumptions 4.1 - 4.4, there exists a matrix Res_d that all elements in the matrix are at the order of $O_p\{(s_n^3 \log p_n/n)^{1/2}\}$ such that $\widehat{\beta}_s - \beta_s^* = n^{-1}\{\Omega(\beta_s^*) + Res_d\}^{-1} U(\beta_s^*)$, where the $O_p\{(s_n^3 \log p_n/n)^{1/2}\}$ term uniformly holds for all models $s \in S_+$.*

Lemma 3.2 approximates the distance of $\widehat{\beta}_s$ to β_s^* as the product of a small perturbation of information matrix and the score vector.

Lemma 3.3 *Under Assumption 3.1 - 3.5, for all models $s \in S_+$ the difference of quasi-likelihood between true and estimated parameter can be approximated as*

quadratic forms:

$$2[QL(\widehat{\beta}_s) - QL(\beta_s^*)] = \frac{1}{n} U(\beta_s^*)^T \Omega(\beta_s^*)^{-1} U(\beta_s^*)\{1 + o_p(1)\}. \tag{3.7}$$

Lemma 3.4 *Under Assumption 3.1 - 3.5, for an overfitting model $s \in S_+$, the difference of quasi-likelihood between true model estimator and overfitting model estimator can be expressed as a quadratic form:*

$$2[QL(\widehat{\beta}_s) - QL(\widehat{\beta}_T)] = \eta_I^T B_I \eta_I \{1 + o_p(1)\}, \tag{3.8}$$

here $\eta_I = n^{-1/2} W(\beta_s^*)^{-1/2} U(\beta_s^*)$; $B_I = W(\beta_s^*)^{1/2}\{\Omega^{-1}(\beta_s^*) - D_s^T \Omega^{-1}(\beta_T^*) D_s\} W(\beta_s^*)^{1/2}$.

Lemma 3.5 *For $s \in S_+$, let $\eta = n^{-1/2} W(\beta_s^*)^{-1/2} \sum_{i=1}^n U_i(\beta_s^*)$. The random vectors $U_1(\beta_s^*), U_2(\beta_s^*) \ldots U_n(\beta_s^*)$ are independently distributed random vectors of dimension d_s with zero mean and satisfy the cumulant boundedness condition. Under Assumptions 3.1 - 3.5, $\log E[e^{t^T \eta}] \leq a^2 t^T t/2$ for $||t||^2 \leq s_n^2 \log p_n$ and some constant $a^2 > 1$.*

In this article, large deviation results are used as an important tool to establish the estimation consistency and model selection consistency in large p_n settings. Let ψ denote a random vector and O denote a positive definite matrix. Large deviation results for quadratic form $\psi^T O \psi$ were established by Spokoiny and Zhilova (2013) under an exponential moment condition:

$$\log[\mathrm{E}\{\exp(t^T \psi)\}] \leq ||t||^2/2, ||t|| \leq \rho, \tag{3.9}$$

where ρ is a positive constant. Define $P_G = \text{Tr}[O]$ and $V_G^2 = \text{Tr}[O^2]$. Based on Corollary 4.2 in Spokoiny and Zhilova (2013), for $\rho^2/4 > K > V_G/3$,

$$\Pr(\psi^T O\psi > P_G + K) \leq 10.4 \exp(-K/6). \tag{3.10}$$

This key result establishes the exponential decay of the tail probability for a quadratic form. Such exponential decay rate is crucial for the control of the overall model selection error. We will show that by choosing an appropriate penalty term, the model selection error rate for each competing model can be derived using equation (3.10), which is exponentially small . The total number of competing models is of the order of $s_n^{p_n}$. By Bonferroni inequality, the overall model selection error rate will be less than the sum of each individual error and the sum can be controlled to have the limiting value of zero. Gao and Carroll (2017) show that the exponential moment condition in equation (3.9) can be satisfied asymptotically by sample mean types of statistics if the original random vector satisfies the above cumulant boundedness condition.

Lemma 3.5 indicates that η satisfies cumulant boundedness condition and implies that we will be able to apply large deviation results to the quasi-likelihood difference type of statistics arising in our analysis from equation (3.9).

Lemma 3.6 *Under Assumption 3.3 and 3.5, the expectation of differences of quasi-likelihood between true model's parameter and any candidate model's pseudo-true*

parameter can be estimated by:

$$E\{QL(\beta_T^*) - QL(\beta_s^*)\} = O[\sum_{i=1}^{n}\{\mu_i(\beta_T^*) - \mu_i(\beta_s^*)\}^T\{\mu_i(\beta_T^*) - \mu_i(\beta_s^*)\}].$$

For underfitting model $s \in S_-$, Assumption 3.4 assumes that $\sum_{i=1}^{n}\{\mu_i(\beta_T^*) - \mu_i(\beta_s^*)\}^T\{\mu_i(\beta_T^*) - \mu_i(\beta_s^*)\}/(ns_n^3 \log p_n)^{1/2} \to \infty$. So we know that $n^{-1}\mathrm{E}\{QL(\beta_T^*) - QL(\beta_s^*)\}$ has higher order of $(s_n^3 \log p_n/n)^{1/2}$. For overfitting model $s \in S_+$, β_s^* and β_T^* are the same for the non-zero components and therefore we have $\mu_i(\beta_s^*) = \mu_i(\beta_T^*)$ and $QL(\beta_s^*) = QL(\beta_T^*)$.

Lemma 3.7 *Under Assumption 3.1 - 3.5, and for all models $s \in S$, we can estimate the order of quasi-likelihood between true and estimated parameter:*

$$|QL(\widehat{\beta}_s) - QL(\beta_s^*)| = O_p\{(ns_n^3 \log p_n)^{1/2}\}.$$

Let $\omega = \max_{s\in S}(d_s^* - d_T^*)/(d_s - d_T)$, the ratio of effective degrees of freedom over the true degrees of freedom. For true likelihood setting, we have $\omega = 1$.

Lemma 3.8 *Assume ω is bounded away from zero and infinity. Let $\gamma_n = 6\omega(1 + \log\log p_n)\log p_n$ or $\gamma_n = 6\omega(1+\gamma)\log p_n$ for some $\gamma > 0$. Under Assumption 3.1 - 3.5, we have*

$$\Pr\{\max_{s\in S_+\setminus T} \Delta_{s/T}/(d_s^* - d_T^*) \geq \gamma_n\} = o(1),$$

where $\Delta_{s/T} = \eta_I^T B_I \eta_I$.

Theorem 3.2 *Assume ω is bounded away from zero and infinity. Let $\gamma_n = 6\omega(1 + \log\log p_n)\log p_n$ or $\gamma_n = 6\omega(1+\gamma)\log p_n$ for some $\gamma > 0$. Under Assumption 3.1 - 3.5, we have*

$$\Pr\{\min_{s\in S} QBIC(s) \geq QBIC(T)\} \to 1.$$

Proof. Firstly we consider the overfitting situation $s \in S_+ \setminus T$. We have

$$\begin{aligned}
&\min_{s\in S_+\setminus T} QBIC(s) - QBIC(T) \\
&= \{\min_{s\in S_+\setminus T} -2QL(\widehat{\beta}_s) + 2QL(\widehat{\beta}_T) + (d_s^* - d_T^*)\gamma_n\} \\
&\geq -\max_{s\in S_+\setminus T}\{\Delta_{s/T} + (d_s^* - d_T^*)\gamma_n + o_p(1)\}.
\end{aligned}$$

According to Lemma 3.8, we have $\Pr\{\max_{s\in S_+\setminus T}\Delta_{s/T}/(d_s^* - d_T^*) \geq \gamma_n\} = o(1)$. Therefore we get $\Pr\{\min_{s\in S} QBIC(s) \geq QBIC(T)\} \to 1$. Next we consider the underfitting situation $s \in S_-$. We have

$$\begin{aligned}
&\min_{s\in S_-} QBIC(s) - QBIC(T) \\
&= \min_{s\in S_-}[-2\{QL(\widehat{\beta}_s) - QL(\widehat{\beta}_T)\} + (d_s^* - d_T^*)\gamma_n] \\
&= -\max_{s\in S_-}[2\{QL(\widehat{\beta}_s) - QL(\widehat{\beta}_T)\} - (d_s^* - d_T^*)\gamma_n] \\
&\geq -\max_{s\in S_-} 2\{QL(\widehat{\beta}_s) - QL(\beta_s^*)\} + 2\{QL(\widehat{\beta}_T) - QL(\beta_T^*)\} - 2\max_{s\in S_-}|QL(\beta_s^*) - \mathrm{E}\{QL(\beta_s^*)\}| \\
&+ [QL(\beta_T^*) - \mathrm{E}\{QL(\beta_T^*)\}] + 2\min_{s\in S_-}\mathrm{E}\{QL(\beta_T^*) - QL(\beta_s^*)\} + (d_s^* - d_T^*)\gamma_n.
\end{aligned}$$

Lemma 3.6 and Assumption 3.4 show that $\mathrm{E}\{QL(\beta_T^*)-QL(\beta_s^*)\}/(ns_n^3\log p_n)^{1/2}=\infty$. Lemma 3.1 shows that $QL(\beta_s^*)-\mathrm{E}\{QL(\beta_s^*)\}=O_p\{(ns_n\log p_n)^{1/2}\}$ and $QL(\beta_T^*)-\mathrm{E}\{QL(\beta_T^*)\}=O_p\{(ns_n\log p_n)^{1/2}\}$. Lemma 3.7 shows that for all models $s\in S$, we have $QL(\widehat{\beta}_s)-QL(\beta_s^*)=O_p\{(ns_n^3\log p_n)^{1/2}\}$, $QL(\widehat{\beta}_T)-QL(\beta_T^*)=O_p\{(ns_n^3\log p_n)^{1/2}\}$. In addition, $|d_s^*-d_T^*|\leq\omega|d_s-d_T|=O_p(s_n)$. Therefore $\mathrm{E}\{QL(\beta_T^*)-QL(\beta_s^*)\}$ is the leading term. And we can get:

$$\Pr\{\min_{s\in S_-}QBIC(s)\geq QBIC(T)\}\to 1.$$

■

3.5 Numerical Study

Section 3.4 illustrates that the QBIC is model selection consistency with the working correlation matrix R being an identity matrix I. In the simulation part, we set the $R=I$. We simulate both clustered binary and clustered Gaussian response variables Y_{ij}. We consider different settings with sample size $n=1000$ or 500, the number of covariates $p_n=1000$ or 500, and the cluster size $m=10$ or 20. The number of true covariates d_T is set be 50. For $j=1,...,d_T$, β_j is drawn from the uniform distribution $U(0.05,0.5)$, whereas for $j=d_T+1,d_T+2\ldots p_n$, β_j are set to zero. For the jth observation in the ith cluster, we simulate the associated covariates $X_{ij}=(x_{ij1},x_{ij2}\ldots x_{ijp_n})^T$, and the mean parameter is denoted as

$\mu_{ij} = \text{logit}^{-1}(X_{ij}^T\beta)$ for binary response and $\mu_{ij} = X_{ij}^T\beta$ for Gaussian response. The covariates X_{ijk} are partitioned into independent blocks of 50 covariates, and within each block the 50 covariates are simulated from the multivariate normal distribution with variances equal to 1 and off-diagonal covariances equal to $0.5^{|k-k'|}$, where k and k' index for the covariates. In each cluster i, for binary response Y_i is simulated from a multivariate binary distribution with mean μ_i and an unstructured correlation matrix; for Gaussian response Y_i is simulated from a multivariate Gaussian distribution with mean μ_i, variance 1, and an unstructured correlation matrix. The r package "SimCorMultRes" is used to simulate the correlated multivariate binary distribution. We use the LASSO to generate a sequence of subset models and use the proposed QBIC to select the best subset model. With regard to the penalty term, Theorem 3.1 provides a theoretical value of $6\omega \times d_s^* \log p_n$. We set the penalty term to be $c \times d_s^* \log p_n$, where c is a constant multiplicative factor and we vary c from 1 to 4. This penalty term has the same asymptotic order as the theoretical penalty term. We run 100 simulations and evaluate the mean and standard deviation of the Positive Selection Rates (PSR) and False Discovery Rates (FDR) of Pan (2001)'s QIC and our proposed QBIC.

Table 3.1 and Table 3.2 compare the PSR and FDR of QIC with the proposed QBIC when $c = 1, 2, 3, 4$. It is shown that QIC has largely inflated FDR, whereas

the proposed QBIC has a good error rate control. For example, when $n = 1000$ and $p_n = 1000$, the FDR of the QIC can be as high as 70%, while the FDR of the QBIC is about 5%. This demonstrates that with large p_n, QIC tends to select overfitting models. This is due to the fact that the QIC uses the AIC type of penalty, which is too small to control the error rate. We vary the multiplicative factor of c from 1 to 4 and examine how the sensitivity and selectivity of our method changes. When c changes from 1 to 4, we found that the QBIC's PSR and FDR both decrease slightly. More simulation details refer to Table A.1 and A.2 at Appendix part.

Table 3.1: QBIC Simulation Result for Binary Response

The free multiplicative constant c for the penalty varies from 1 to 4. QBIC assumes the identity working correlation matrix. The cluster size m is 10 and true model size d_T is 50.

		n 1000	p 1000			n 500	p 500		
		mean	std	mean	std	mean	std	mean	std
		psr	psr	fdr	fdr	psr	psr	fdr	fdr
QIC	Ind	1.0000	0.0000	0.7093	0.0241	0.9974	0.0084	0.5677	0.0735
	Exc	1.0000	0.0000	0.7099	0.0241	0.9974	0.0084	0.5677	0.0735
	AR1	1.0000	0.0000	0.7093	0.0241	0.9974	0.0084	0.5677	0.0735
	Uns	1.0000	0.0000	0.7234	0.0179	0.9976	0.0082	0.6163	0.0677
QBIC	c=1	0.9994	0.0060	0.1192	0.0585	0.9576	0.0562	0.0653	0.0778
	c=2	0.9774	0.0399	0.0029	0.0092	0.7920	0.0781	0.0000	0.0000
	c=3	0.8984	0.0779	0.0000	0.0000	0.6698	0.0876	0.0000	0.0000
	c=4	0.8042	0.0808	0.0000	0.0000	0.5992	0.0844	0.0000	0.0000

Table 3.2: QBIC Simulation Result for Gaussian Response

The free multiplicative constant c for the penalty varies from 1 to 4. QBIC assumes the identity working correlation matrix. The cluster size m is 10 and true model size d_T is 50.

		n 1000	p 1000			n 500	p 500		
		mean	std	mean	std	mean	std	mean	std
		psr	psr	fdr	fdr	psr	psr	fdr	fdr
QIC	Ind	1.0000	0.0000	0.6937	0.0326	1.0000	0.0000	0.5401	0.0607
	Exc	1.0000	0.0000	0.6937	0.0326	1.0000	0.0000	0.5395	0.0599
	AR1	1.0000	0.0000	0.6937	0.0326	1.0000	0.0000	0.5401	0.0607
	Uns	1.0000	0.0000	0.7281	0.0136	1.0000	0.0000	0.7109	0.0324
QBIC	c=1	1.0000	0.0000	0.1077	0.0460	1.0000	0.0000	0.0871	0.0449
	c=2	1.0000	0.0000	0.0012	0.0047	1.0000	0.0000	0.0008	0.0038
	c=3	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
	c=4	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000

3.6 Proofs of Related Lemmas

Since the proofs of some Lemmas are too long to put them into the main text, we list the proofs in this separate section to make the main text succinct.

Proof of Lemma 3.1. The proof is similar to Lemma 2.1 given the similar assumptions. According to Equation (2.3), for independent variables $z_1, z_2 \dots z_n$ with mean zero and unit variance satisfy Cumulant Bounded Condition and for any real number $b_{cbc} > 0$, we have $\Pr\{\sum_{i=1}^{n} z_i > (nb_{cbc}^2)^{1/2}\} \leq \exp[-(1/2)b_{cbc}^2\{1 + o(1)\}]$. Because $QL_i(\beta_s)$ satisfies the cumulant boundedness condition for all $i \in \{1, 2 \dots n\}$. And $QL_i(\beta_s)$'s first and second moments are uniformly bounded. Given a model s, let $b_{cbc}^2 = p_n^{s_n}$, we have $\Pr[(\sum_{i=1}^{n}[QL_i(\beta_s) - \mathrm{E}\{QL_i(\beta_s)\}]/\mathrm{Var}\{QL_i(\beta_s)\}) > (2ns_n \log p_n)^{1/2}] = o(p_n^{-s_n})$. According to Bonferroni inequality, $\Pr(\max_{s \in S} \sum_{i=1}^{n} [QL_i(\beta_s) - \mathrm{E}\{QL_i(\beta_s)\}] > b_{var}(2ns_n \log p_n)^{1/2}) \leq o(p_n^{-s_n})p_n^{s_n} = 0$, as there are $p_n^{s_n}$ models in the model space, and b_{var} is the upper bound for $\mathrm{Var}\{QL_i(\beta_s)\}$. Similar arguments apply to the result for each element of score function, and its first and second derivatives. ■

Proof of Lemma 3.2. From Taylor expansion around β_s^*, there exists $\widetilde{\beta}_s$ between β_s^* and $\widehat{\beta}_s$ such that $(1/n)U(\widehat{\beta}_s)_{[r]} = 0$. Therefore we know

$$\frac{1}{n}U(\beta_s^*)_{[r]} + \sum_k \frac{1}{n}U(\beta_s^*)_{[rk]}^{(1)}(\widehat{\beta}_s - \beta_s^*)_{[k]} + \sum_{k,l} \frac{1}{n}U(\widetilde{\beta}_s)_{[rkl]}^{(2)}(\widehat{\beta}_s - \beta_s^*)_{[k]}(\widehat{\beta}_s - \beta_s^*)_{[l]} = 0.$$

According to Lemma 3.1 $\max_{s\in S}|n^{-1}U(\beta_s^*)^{(1)}_{[rk]}+\Omega(\beta_s^*)_{[rk]}| = \max_{s\in S}|n^{-1}U(\beta_s^*)^{(1)}_{[rk]}-\mathrm{E}[U(\beta_s^*)^{(1)}_{[rk]}]| = O_p\{(s_n\log p_n/n)^{1/2}\}$, then we have

$$\begin{aligned}
&\sum_k \frac{1}{n}U(\beta_s^*)^{(1)}_{[rk]}(\widehat{\beta}_s-\beta_s^*)_{[k]}\\
&=\sum_k[-\Omega(\beta_s^*)_{[rk]}+\{\Omega(\beta_s^*)_{[rk]}+\frac{1}{n}U(\beta_s^*)^{(1)}_{[rk]}\}](\widehat{\beta}_s-\beta_s^*)_{[k]}\\
&=\sum_k[-\Omega(\beta_s^*)_{[rk]}+O_p\{(s_n\log p_n/n)^{1/2}\}](\widehat{\beta}_s-\beta_s^*)_{[k]}.
\end{aligned}$$

Similarly from Lemma 3.1, $n^{-1}U(\widetilde{\beta}_s)^{(2)}_{[rkl]} = n^{-1}\mathrm{E}[U(\widetilde{\beta}_s)^{(2)}_{[rkl]}]\{1+o_p(1)\}$. From Assumption 3.2, $n^{-1}\mathrm{E}[U(\widetilde{\beta}_s)^{(2)}_{[rkl]}]$ is bounded. So $n^{-1}U(\widetilde{\beta}_s)^{(2)}_{[rkl]} = O_p(1)$. According to Theorem 3.1 $||\widehat{\beta}_s-\beta_s^*|| = O_p\{(s_n^2\log p_n/n)^{1/2}\}$. Then

$$\begin{aligned}
&|\sum_l n^{-1}U(\beta_s^*)^{(2)}_{[rkl]}(\widehat{\beta}_s-\beta_s^*)_{[l]}|\\
&\le \max_l|n^{-1}U(\beta_s^*)^{(2)}_{[rkl]}|\sum_l|(\widehat{\beta}_s-\beta_s^*)_{[l]}|\\
&= O_p(1)\times\sum_l|(\widehat{\beta}_s-\beta_s^*)_{[l]}|\\
&\le O_p(1)\times d_s^{1/2}\times||\widehat{\beta}_s-\beta_s^*||\\
&= O_p\{(s_n^3\log p_n/n)^{1/2}\}.
\end{aligned}$$

Combining the second and the third order terms of Taylor expansion, we have

$$
\begin{aligned}
0 &= \frac{1}{n} U(\widehat{\beta}_s)_{[r]} \\
&= \frac{1}{n} U(\beta_s^*)_{[r]} - \sum_k [\Omega(\beta_s^*)_{[rk]} + O_p\{(s_n \log p_n/n)^{1/2}\}](\widehat{\beta}_s - \beta_s^*)_{[k]} \\
&+ \sum_k O_p\{(s_n^3 \log p_n/n)^{1/2}\}(\widehat{\beta}_s - \beta_s^*)_{[k]}.
\end{aligned}
$$

We can reformat it as

$$
\frac{1}{n} U(\beta_s^*) - \{\Omega(\beta_s^*) + Res_d\}(\widehat{\beta}_s - \beta_s^*) = 0,
$$

where Res_d is a matrix that all elements are at order of $O_p\{(s_n^3 \log p_n/n)^{1/2}\}$ uniformly. Let v_{min} be the corresponding eigenvector of smallest eigenvalue $\lambda_{\min}\{\Omega(\beta_s^*)\}$. According to matrix perturbation theory (Stewart, 1990), we have

$$
\begin{aligned}
&\lambda_{\min}\{\Omega(\beta_s^*) + Res_d\} \\
&= \lambda_{\min}\{\Omega(\beta_s^*)\} + v_{min}^T Res_d v_{min} + o(||Res_d||^2) \\
&\geq \lambda_{\min}\{\Omega(\beta_s^*)\} + d_s \times ||Res_d||_{\max} + o(1) \\
&= \lambda_{\min}\{\Omega(\beta_s^*)\} + O_p\{(s_n^5 \log p_n/n)^{1/2}\} + o(1)
\end{aligned}
$$

Since $\lambda_{\min}\{\Omega(\beta_s^*)\} > 0$ and $s_n^5 \log p_n/n \to 0$, we have $\lambda_{\min}\{\Omega(\beta_s^*) + Res_d\} > 0$ and therefore $\Omega(\beta_s^*) + Res_d$ is invertible. This entails

$$
\widehat{\beta}_s - \beta_s^* = \frac{1}{n}\{\Omega(\beta_s^*) + Res_d\}^{-1} U(\beta_s^*).
$$

■

Proof of Lemma 3.3. According to Taylor expansion, there exist a $\widetilde{\beta}_s$ between $\widehat{\beta}_s$ and β_s^* such that

$$QL(\widehat{\beta}_s) - QL(\beta_s^*) = U(\beta_s^*)(\widehat{\beta}_s - \beta_s^*) - \frac{n}{2}(\widehat{\beta}_s - \beta_s^*)^T \Omega(\beta_s^*)(\widehat{\beta}_s - \beta_s^*) + Residual,$$

where the residual term has below format:

$$\begin{aligned} Residual = & \sum_{r,k} \frac{1}{2}\{QL_{rt}^{(2)}(\beta_s^*) + n\Omega_{rt}(\beta_s^*)\}(\widehat{\beta}_s - \beta_s^*)_{[r]}(\widehat{\beta}_s - \beta_s^*)_{[k]} \\ & + \sum_{r,k,l} \frac{1}{n} QL_{rkl}^{(3)}(\widetilde{\beta}_s)(\widehat{\beta}_s - \beta_s^*)_{[r]}(\widehat{\beta}_s - \beta_s^*)_{[k]}(\widehat{\beta}_s - \beta_s^*)_{[l]}. \end{aligned}$$

Similarly from Lemma 3.1, we know that $QL_{rt}^{(2)}(\beta_s^*) + n\Omega_{rk}(\beta_s^*) = O_p\{(s_n \log p_n/n)^{1/2}\}$. From Assumption 3.1, Lemma 3.1 and Theorem 3.1, we also know that $|\sum_l Q_{rkl}^{(3)}(\widetilde{\beta}_s^*)(\widehat{\beta}_s - \beta_s^*)_{[l]}| \leq \{\sum_l Q_{rkl}^{(3)}(\widetilde{\beta}_s^*)^2\}^{1/2} \times ||\widehat{\beta}_s - \beta_s^*|| = O_p\{(s_n^3 \log p_n/n)^{1/2}\}$. Therefore we have below equation:

$$QL(\widehat{\beta}_s) - QL(\beta_s^*) = U(\beta_s^*)^T(\widehat{\beta}_s - \beta_s^*) - (n/2)(\widehat{\beta}_s - \beta_s^*)^T\{\Omega(\beta_s^*) + Res_Q\}(\widehat{\beta}_s - \beta_s^*),$$

where Res_Q is a matrix that each elements are at order of $O_p\{(s_n^3 \log p_n/n)^{1/2}\}$. For a unit vector $||v||^2 = 1$, we have $|v^T Res_Q v| = |\sum_{kr} v_k v_r [Res_Q]_{kr}| \leq \max_{kr} |[Res_Q]_{kr}| \times p_n \times ||v||^2 = O_p\{(p_n^5 \log p_n/n)^{1/2}\} = o_p(1)$. Then we know

$$(\widehat{\beta}_s - \beta_s^*)^T\{\Omega(\beta_s^*) + Res_Q\}(\widehat{\beta}_s - \beta_s^*) = (\widehat{\beta}_s - \beta_s^*)^T\Omega(\beta_s^*)(\widehat{\beta}_s - \beta_s^*)\{1 + o_p(1)\}.$$

In addition, it can be shown that

$$\begin{aligned}
&\sup_{||v||^2=1} v^T[\Omega(\beta_s^*)^{-1} - \{\Omega(\beta_s^*) + Res_d\}^{-1}]v \\
&= \sup_{||v||^2=1} v^T\Omega(\beta_s^*)^{-1/2}[I - \{I + \Omega(\beta_s^*)^{-1/2}Res_d\Omega(\beta_s^*)^{-1/2}\}^{-1}]\Omega(\beta_s^*)^{-1/2}v \\
&\leq \lambda_{\max}\{\Omega(\beta_s^*)^{-1}\}\lambda_{\max}(I - [\{I + \Omega(\beta_s^*)^{-1/2}Res_d\Omega(\beta_s^*)^{-1/2}\}^{-1}])||v||^2 \\
&= \lambda_{\max}\{\Omega(\beta_s^*)^{-1}\}(1 - \lambda_{\min}[\{I + \Omega(\beta_s^*)^{-1/2}Res_d\Omega(\beta_s^*)^{-1/2}\}^{-1}]) \\
&= \lambda_{\max}\{\Omega(\beta_s^*)^{-1}\}[1 - \frac{1}{1 + \lambda_{\max}\{\Omega(\beta_s^*)^{-1/2}Res_d\Omega(\beta_s^*)^{-1/2}\}}] \\
&= \lambda_{\max}\{\Omega(\beta_s^*)^{-1}\}[\frac{\lambda_{\max}\{\Omega(\beta_s^*)^{-1/2}Res_d\Omega(\beta_s^*)^{-1/2}\}}{1 + \lambda_{\max}\{\Omega(\beta_s^*)^{-1/2}Res_d\Omega(\beta_s^*)^{-1/2}\}}].
\end{aligned}$$

Furthermore,

$$\begin{aligned}
&\lambda_{\max}\{\Omega(\beta_s^*)^{-1/2}Res_d\Omega(\beta_s^*)^{-1/2}\} \\
&= \sup_{||v||=1} v^T\{\Omega(\beta_s^*)^{-1/2}Res_d\Omega(\beta_s^*)^{-1/2}\}v \\
&\leq \lambda_{\max}\{\Omega(\beta_s^*)^{-1}\}\lambda_{\max}\{Res_d\}||v||^2 \\
&= o_p(1).
\end{aligned}$$

Thus $\sup_{||v||^2=1} v^T[\Omega(\beta_s^*)^{-1} - \{\Omega(\beta_s^*) + Res_d\}^{-1}]v = o_p(1)$. And it equally means that

$$U(\beta_s^*)^T\{\Omega(\beta_s^*) + Res_d\}^{-1}U(\beta_s^*) = U(\beta_s^*)^T\Omega(\beta_s^*)^{-1}U(\beta_s^*)\{1 + o_p(1)\}.$$

Then from Lemma 3.2 and above equations, we can rewrite above equation:

$$
\begin{aligned}
& QL(\widehat{\beta}_s) - QL(\beta_s^*) \\
&= \frac{1}{n} U(\beta_s^*)^T \{\Omega(\beta_s^*) + Res_d\}^{-1} U(\beta_s^*) \\
&- \frac{1}{2n} U(\beta_s^*)^T \{\Omega(\beta_s^*) + Res_d\}^{-1} \{\Omega(\beta_s^*) + Res_Q\} \{\Omega(\beta_s^*) + Res_d\}^{-1} U(\beta_s^*) \\
&= U(\beta_s^*)^T \Omega(\beta_s^*)^{-1} U(\beta_s^*) \{1 + o_p(1)\} \\
&- \frac{1}{2n} U(\beta_s^*)^T \{\Omega(\beta_s^*) + Res_d\}^{-1} \Omega(\beta_s^*) \{\Omega(\beta_s^*) + Res_d\}^{-1} U(\beta_s^*) \{1 + o_p(1)\} \\
&= \frac{1}{2n} U(\beta_s^*)^T \Omega^{-1}(\beta_s^*) U(\beta_s^*) \{1 + o_p(1)\}.
\end{aligned}
$$

■

Proof of Lemma 3.4. The differences can be re-structured into three components: $QL(\widehat{\beta}_s) - QL(\widehat{\beta}_T) = \{QL(\widehat{\beta}_s) - QL(\beta_s^*)\} - \{QL(\widehat{\beta}_T) - QL(\beta_T^*)\} + \{QL(\beta_s^*) - QL(\beta_T^*)\}$.

Firstly we prove that the third term is zero. Since β_s^* should be the same as β_T^* for first T elements and the rest of parameters should be zero. Then the overfitting model and true model share the same mean equation $\mu_{ij}(\beta_s^*) = \mu_{ij}(\beta_T^*)$. We get the

following:

$$\begin{aligned} QL(\beta_s^*) - QL(\beta_T^*) &= \sum_{i=1}^{n}\sum_{j=1}^{m}\int_{Y_{ij}}^{\mu_{ij}(\beta_T^*)} \frac{Y_{ij}-t}{A_{ij}(t)}dt - \sum_{i=1}^{n}\sum_{j=1}^{m}\int_{Y_{ij}}^{\mu_{ij}(\beta_S^*)} \frac{Y_{ij}-t}{A_{ij}(t)}dt \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n_i}\int_{\mu_{ij}(\beta_S^*)}^{\mu_{ij}(\beta_T^*)} \frac{Y_{ij}-t}{A_{ij}(t)}dt \\ &= 0. \end{aligned}$$

From Lemma 3.3, we know $2\{QL(\widehat{\beta}_T) - QL(\beta_T^*)\} = n^{-1}U(\beta_T^*)^T\Omega^{-1}(\beta_T^*)U(\beta_T^*)\{1+o_p(1)\}$. Define matrix $D_s = (I_T, 0_{d_T,d_s-d_T})$ with I_T be a d_T by d_T identity matrix and $0_{d_T,d_s-d_T}$ be zeros matrix. Then we have $2\{QL(\widehat{\beta}_T) - QL(\beta_T^*)\} = n^{-1}U(\beta_s^*)^T D_s^T\Omega^{-1}(\beta_T^*) D_s U(\beta_T^*)\{1+o_p(1)\}$. Therefore we can get following equation:

$$2\{QL(\widehat{\beta}_s) - QL(\widehat{\beta}_T)\} = U(\beta_s^*)^T M_{s/T} U(\beta_s^*)\{1+o_p(1)\},$$

where $M_{s/T} = n^{-1}\Omega^{-1}(\beta_s^*) - n^{-1}D_s^T\Omega^{-1}(\beta_T^*)D_s$. Define $\eta_I = n^{-1/2}W(\beta_s^*)^{-1/2}U(\beta_s^*)$ and $B_I = W(\beta_s^*)^{1/2}\{\Omega^{-1}(\beta_s^*) - D_s^T\Omega^{-1}(\beta_T^*)D_s\}W(\beta_s^*)^{1/2}$. We can rewrite the differences for two quasilikelihood equation into below formate:

$$2\{QL(\widehat{\beta}_s) - QL(\widehat{\beta}_T)\} = \eta_I^T B_I \eta_I\{1+o_p(1)\}.$$

Here the trace of matrix B_I is $\mathrm{Tr}(B_I) = \mathrm{Tr}\{W(\beta_s^*)\Omega^{-1}(\beta_s^*) - W(\beta_T^*)\Omega^{-1}(\beta_T^*)\} = d_s^* - d_T^*$. ■

Proof of Lemma 3.5. For $s \in S_+$, $\mathrm{E}\{U_i(\beta_s^*)\} = \mathrm{E}[D_i(\beta_s^*)^T V_i^{-1}\{Y_i - \mu_i(\beta_s^*)\}] = 0$. Let $W_i = \mathrm{Cov}\{U_i(\beta_s^*)\}$ be the covariance matrix of $U_i(\beta_s^*)$ and $W = \sum_{i=1}^n W_i/n$.

The cumulant generating function of U is

$$\begin{aligned} C_{U_i(\beta_s^*)}(t) &= \log \mathrm{E}\{e^{t^T U_i(\beta_s^*)}\} \\ &= C(0) + t^T C^{(1)}(0) + \frac{1}{2} t^T \mathrm{Cov}\{U_i(\beta_s^*)\} t + \frac{1}{6} \sum_{lrk} t_l t_r t_k C^{(3)}_{lrk}(t^*), \end{aligned}$$

with a t^* such that $||t^*|| \leq ||t||$. Let $\eta_1 = n^{-1/2} \sum_{i=1}^n U_i(\beta_s^*)$, then the cumulant generating function of η_1 is

$$\begin{aligned} C_{\eta_1}(t) &= \sum_{i=1}^n C_{U_i(\beta_s^*)}(\frac{t}{n^{1/2}}) \\ &= \sum_{i=1}^n \{\frac{1}{2n} t^T \mathrm{Cov}\{U_i(\beta_s^*)\} t + \sum_{lrk} \frac{1}{6n^{3/2}} t_l t_r t_k C^{(3)}_{lrk}(\frac{t^*}{n^{1/2}})\} \\ &= C_1 + C_2. \end{aligned}$$

First, C_1 can be simplified as $C_1 = t^T\{n^{-1} \sum_{i=1}^n \mathrm{Cov}(U_i)\} t/2 = t^T (n^{-1} \sum_{i=1}^n W_i) t/2$ $= t^T W t/2$. Next, C_2 has the bound as follows:

$$C_2 \leq \frac{b_c}{6n^{1/2}} s_n^{3/2} ||t||^3 = \frac{b_c}{6} (\frac{s_n^3 ||t||^2}{n})^{1/2} ||t||^2 = O_p[\{s_n^5 \log p_n / n\}^{1/2}] ||t||^2 = o_p(1) ||t||^2.$$

This entails

$$C_{\eta_1}(t) \leq \frac{1}{2} a^2 t^T W t + ||t||^2 O_p[\{s_n^5 \log p_n / n\}^{1/2}] \leq \frac{1}{2} a^2 t^T W t,$$

for some a with $a^2 > 1$ and $||t|| \leq \{(s_n^2 \log p_n)^{1/2}\}$. Here a is able to be chosen as close to 1 as n is large. Let $\eta = W^{-1/2} \eta_1$, then the cumulant generating function of η is $\log \mathrm{E}[e^{t^T \eta}] \leq a^2 t^T t/2$. ■

Proof of Lemma 3.6. We know that $\mathrm{E}\{QL(\beta_T^*) - QL(\beta_s^*)\} = \sum_{i=1}^n \sum_{j=1}^m \int_{\mu_{ij}(\beta_s^*)}^{\mu_{ij}(\beta_T^*)} \{\mathrm{E}[Y_{ij}] - t\}/\{A_{ij}(t)\}dt = \sum_{i=1}^n \sum_{j=1}^m \int_{\mu_{ij}(\beta_s^*)}^{\mu_{ij}(\beta_T^*)} \{\mu_{ij}(\beta_T^*) - t\}/\{A_{ij}(t)\}dt$. According to Mean Value Theorem, there is a β_{ij} between β_T^* and β_s^* such that $\int_{\mu_{ij}(\beta_s^*)}^{\mu_{ij}(\beta_T^*)} \{\mu_{ij}(\beta_T^*) - t\}/\{A_{ij}(t)\}dt = A_{ij}(\beta_{ij})^{-1} \int_{\mu_{ij}(\beta_s^*)}^{\mu_{ij}(\beta_T^*)} \{\mu_{ij}(\beta_T^*) - t\}dt$. Then we can have following:

$$\begin{aligned}
\mathrm{E}[QL(\beta_T^*) - QL(\beta_s^*)] &= \sum_{i=1}^n \sum_{j=1}^m \int_{\mu_{ij}(\beta_s^*)}^{\mu_{ij}(\beta_T^*)} \frac{\mu_{ij}(\beta_T^*) - t}{A_{ij}(t)} dt \\
&= \sum_{i=1}^n \sum_{j=1}^m A_{ij}(\beta_{ij})^{-1} \int_{\mu_{ij}(\beta_s^*)}^{\mu_{ij}(\beta_T^*)} \{\mu_{ij}(\beta_T^*) - t\} dt \\
&= \sum_{i=1}^n \sum_{j=1}^m 0.5 A_{ij}(\beta_{ij})^{-1} \{\mu_{ij}(\beta_T^*) - \mu_{ij}(\beta_s^*)\}^2 \\
&= O[\sum_{i=1}^n \{\mu_i(\beta_T^*) - \mu_i(\beta_s^*)\}^T \{\mu_i(\beta_T^*) - \mu_i(\beta_s^*)\}]
\end{aligned}$$

Here according to Assumption 3.5, $A_{ij}(\beta_{ij})$ are uniformly bounded away from zero and infinity. ■

Proof of Lemma 3.7. According to Mean Value Theorem, there is a β_{tij} between β_s^* and $\widehat{\beta}_s$ such that $\int_{\mu_{ij}(\widehat{\beta}_s)}^{\mu_{ij}(\beta_s^*)} (Y_{ij} - t)/\{A_{ij}(t)\}dt = A_{ij}(\beta_{tij})^{-1} \int_{\mu_{ij}(\beta_T^*)}^{\mu_{ij}(\beta_s^*)} \{\mu_{ij}(\beta_T^*) - t\}dt$. Then we have following equation:

$$\begin{aligned}
|QL(\widehat{\beta}_s) - QL(\beta_s^*)| &= |\sum_{i=1}^n \sum_{j=1}^m \int_{\mu_{ij}(\beta_s^*)}^{\mu_{ij}(\widehat{\beta}_s)} \frac{Y_{ij} - t}{A_{ij}(t)} dt| \\
&= |\sum_{i=1}^n \sum_{j=1}^m A_{ij}(\beta_{tij})^{-1} \int_{\mu_{ij}(\beta_s^*)}^{\mu_{ij}(\widehat{\beta}_s)} (Y_{ij} - t) dt| \\
&= |\sum_{i=1}^n \sum_{j=1}^m 0.5 A_{ij}(\beta_{tij})^{-1} \{\mu_{ij}(\widehat{\beta}_s) - \mu_{ij}(\beta_s^*)\} \{2Y_{ij} - \mu_{ij}(\widehat{\beta}_s) - \mu_{ij}(\beta_s^*)\}|.
\end{aligned}$$

According to Taylor Expansion, we know that there is a β_{sij} between $\widehat{\beta}_s$ and β_s^* such that $\mu_{ij}(\widehat{\beta}_s) - \mu_{ij}(\beta_s^*) = (\widehat{\beta}_s - \beta_s^*)^T\{\partial\mu_{ij}(\beta_{sij})/\partial\beta\}$. And from Cauchy-Schwarz Inequality, we know that: $|\mu_{ij}(\widehat{\beta}_s) - \mu_{ij}(\beta_s^*)| = |(\widehat{\beta}_s - \beta_s^*)^T\partial\mu_{ij}(\beta_{sij})/\partial\beta| \leq ||\widehat{\beta}_s - \beta_s^*|| \times ||\partial\mu_{ij}(\beta_{sij})/\partial\beta||$. Theorem 3.1 indicates that $||\widehat{\beta}_s - \beta_s^*|| = O_p\{(s_n \log p_n/n)^{1/2}\}$. And from Assumption 3.5, we know that each element of $\partial\mu_{ij}(\beta_{sij})/\partial\beta$ is bounded and there are s_n none zero elements. Then we get $||\partial\mu_{ij}(\beta_{sij})/\partial\beta|| = O_p(s_n^{1/2})$. Therefore we get below order $|\mu_{ij}(\widehat{\beta}_s) - \mu_{ij}(\beta_s^*)| = O_p\{(s_n^3 \log p_n/n)^{1/2}\}$ uniformly for all i, j and all models $s \in S_+$. From Lemma 3.9 , we know that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} |2Y_{ij} - \mu_{ij}(\widehat{\beta}_s) - \mu_{ij}(\beta_s^*)| \\
&\leq \frac{1}{n}\sum_{i=1}^{n} |2Y_{ij} - 2\mu_{ij}(\beta_s^*)| + |\mu_{ij}(\beta_s^*) - \mu_{ij}(\widehat{\beta}_s)| \\
&= O_p(1) + O_p\{(s_n^3 \log p_n/n)^{1/2}\}.
\end{aligned}$$

Assumption 3.5 indicates that $A_{ij}(\beta_{tij})^{-1}$ is uniformly bounded. Then we have

$$\begin{aligned}
&|QL(\widehat{\beta}_s) - QL(\beta_s)| \\
&= |\sum_{i=1}^{n}\sum_{j=1}^{m} 0.5A_{ij}(\beta_{tij})^{-1}\{\mu_{ij}(\widehat{\beta}_s) - \mu_{ij}(\beta_s^*)\}\{2Y_{ij} - \mu_{ij}(\widehat{\beta}_s) - \mu_{ij}(\beta_s^*)\}| \\
&\leq \frac{m}{2}\max_{i,j} A_{ij}(\beta_{tij})^{-1} \times \max_{i,j}|\mu_{ij}(\widehat{\beta}_s) - \mu_{ij}(\beta_s^*)| \times \max_{j}\sum_{i=1}^{n}|2Y_{ij} - \mu_{ij}(\widehat{\beta}_s) - \mu_{ij}(\beta_s^*)| \\
&= O_p\{(s_n^3 \log p_n/n)^{1/2}\}.
\end{aligned}$$

■

Proof of Lemma 3.8. Let $\eta_s = n^{-1/2}W^{-1/2}(\beta_s^*)U(\beta_s^*)$. According to Lemma

4.14, it satisfies the exponential moment condition,

$$\log[\mathrm{E}\{\exp(t^T\eta_s)\}] \leq a^2||t||^2/2,$$

with $t \in R^{d_s}$, $||t||^2 \leq s_n^2 \log p_n$ and some constant $a^2 > 1$. Denote $\rho = s_n(\log p_n)^{1/2}$. We scale the vector η as $\eta^* = \eta/a$, so that $\log[\mathrm{E}\{\exp(t^T\eta^*)\}] \leq ||t||^2/2$ with $||t|| \leq \{a^2 s_n^2 \log p_n\}^{1/2} = a \times \rho$. Given matrix $B_{s/T} = W^{1/2}(\beta_s^*)M_{s/T}(\beta_s^*)W^{1/2}(\beta_s^*)$ and $\mathrm{Tr}(B_{s/T}) = d_s^* - d_T^*$, we define $B_{s/T}^* = B_{s/T}/\tau$ where $\tau = \lambda_{max}(B_{s/T})$. Therefore $\lambda_{max}(B_{s/T}^*) = 1$. We scale the quadratic form $\Delta_s^* = \Delta_s/a^2\tau = (\eta^*)^T B_s^* \eta^*$. Let $\Delta_{s/T} = \eta^T B\eta$ and $\Delta_{s/T}^* = \Delta/a^2\tau = (\eta^*)B^*\eta^*$. Define $P_G = \mathrm{Tr}[B^*] = (d_s^* - d_T^*)/\tau$ and $V_G^2 = \mathrm{Tr}[(B^*)^2]$. From Lemma 3.10, we know B^* is a positive definite matrix, we have $B^* = B_e^{-1}HB_e$ where $H = diag\{H_i, i = 1, \ldots, d_s\}$ and H_i are positive eigenvalues. We have $V_G^2 = \mathrm{Tr}[(B^*)^2] = \mathrm{Tr}[B_e^{-1}H^2B_e] = \mathrm{Tr}[H^2] = \sum_{k=1}^{d_s} H_i^2 \leq (\sum_{k=1}^{d_s} H_i)^2 = [\mathrm{Tr}(B^*)]^2$. Therefore we get $V_G^{1/2} \leq P_G = O(s_n)$. Choosing $K = \{(d_s^* - d_T^*)/\tau\}\{\gamma_n/a^2 - 1\}$, we know $K = O(s_n \log p_n)$. Given $\rho^2 = s_n^2 \log p_n$, we have $3/2\rho^2 > K > V_G/3$.

We apply the large deviation result from Corollary 4.2 of Spokoiny and Zhilova (2013) for $3/2\rho^2 > K > V_G/3$

$$\Pr(\Delta_{s/T}^* > P_G + K) \leq 10.4\exp(-K/6).$$

Let $\check{\tau} = (d_s^* - d_T^*)/(d_s - d_T)$. We have

$$\begin{aligned}
&\Pr\{\max_{s\in S_+} \Delta_{s/T} > (d_s^* - d_T^*)\gamma_n\} \\
&\leq \sum_{s\in S_+} \Pr\{\Delta_{s/T}^* > [(d_s^* - d_T^*)\gamma_n/(a^2\tau)]\} \\
&= \sum_{s\in S_+} \Pr\{\Delta_{s/T}^* > P_G + P_G(\frac{\gamma_n}{a^2} - 1)\} \\
&= \sum_{s\in S_+} \Pr\{\Delta_{s/T}^* > P_G + K\} \\
&\leq \sum_{s\in S_+} 10.4\exp\{-\frac{(d_s - d_T)\check{\tau}}{6\tau}(\frac{\gamma_n}{a^2} - 1)\} \\
&\leq \sum_{d_s=d_T+1}^{p_n} C_{p_n-d_T}^{d_s-d_T} 10.4\exp\{-\frac{(d_s - d_T)\check{\tau}}{6\tau}(\frac{\gamma_n}{a^2} - 1)\} \\
&\leq \sum_{m'=1}^{p_n-d_T} C_{p_n-d_T}^{m'} 10.4\exp\{-\frac{m'}{6\omega}(\frac{\gamma_n}{a^2} - 1)\} \\
&\leq \{1 + 10.4\exp(-\frac{\gamma_n/a^2 - 1}{6\omega})\}^{p_n-d_T} - 1.
\end{aligned}$$

As a^2 can be chosen as close to 1 as possible with increasing sample size n, the choices of $\gamma_n = 6\omega(1+\gamma)\log p_n$ for some $\gamma > 0$ or $\gamma_n = 6\omega(1 + \log\log p_n)\log p_n$ lead to $\lim_{n\to\infty}(1 + 10.4\exp\{-(\gamma_n/a^2 - 1)/(6\omega)\})^{p_n-d_T} = 1$. This entails

$$P\{-\max_{s\in S_+\setminus T}[\Delta_{s/T} - (d_s^* - d_T^*)\log p_n] > 0\} \to 1.$$

■

Lemma 3.9 *Under Assumptions 3.5, for all models $s \in S$,*

$$\max_{j=1}^{m} \frac{1}{n}\sum_{i=1}^{n} |Y_{ij} - \mu_{ij}(\beta_s^*)| = O_p(1).$$

Proof of Lemma 3.9 . First we consider the true and overfitting models $s \in \{T, S_+\}$. For any given j and $i \in \{1, 2 \dots n\}$, Y_{ij} are independent and their variances are uniformly bounded. This entails

$$\begin{aligned}
&\mathrm{Var}\{|Y_{ij} - \mu_{ij}(\beta_s^*)|\} = \mathrm{E}\{|Y_{ij} - \mu_{ij}(\beta_s^*)|^2\} - [\mathrm{E}\{|Y_{ij} - \mu_{ij}(\beta_s^*)|\}]^2 \\
&\leq \mathrm{E}[\{Y_{ij} - \mu_{ij}(\beta_s^*)\}^2] = \mathrm{Var}(Y_{ij}) \\
&= O(1).
\end{aligned}$$

By the Law of Large Numbers

$$\frac{1}{n}\sum_{i=1}^{n} |Y_{ij} - \mu_{ij}(\beta_s^*)| \xrightarrow{p} \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\{|Y_{ij} - \mu_{ij}(\beta_s^*)|\}.$$

Furthermore,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\{|Y_{ij} - \mu_{ij}(\beta_s^*)|\} \leq \frac{1}{2n}\sum_{i=1}^{n} \mathrm{E}[\{Y_{ij} - \mu_{ij}(\beta_s^*)\}^2 + 1] \\
&\leq \frac{1}{2n}\{\sum_{i=1}^{n} \mathrm{Var}(Y_{ij}) + 1\} \\
&= O(1).
\end{aligned}$$

For all $j \in \{1, 2 \dots m\}$, we have $n^{-1}\sum_{i=1}^{n} |Y_{ij} - \mu_{ij}(\beta_s^*)| = O_p(1)$.

For underfitting models $s \in S_-$, Assumption 3.5 implies that both $\mu_{ij}(\beta_T^*)$ and $\mu_{ij}(\beta_s^*)$ are bounded. Thus $n^{-1}\sum_{i=1}^{n} |\mu_{ij}(\beta_T^*) - \mu_{ij}(\beta_s^*)| = O(1)$. For $j = 1, 2 \dots m$, we have $n^{-1}\sum_{i=1}^{n} |Y_{ij} - \mu_{ij}(\beta_s^*)| \leq n^{-1}\sum_{i=1}^{n} |Y_{ij} - \mu_{ij}(\beta_T^*)| + n^{-1}\sum_{i=1}^{n} |\mu_{ij}(\beta_T^*) - \mu_{ij}(\beta_s^*)| = O_p(1)$. ■

Lemma 3.10 *Under Assumption 3.1, for overfitting model* $s \in S_+$, $M_{s/T} = \Omega^{-1}(\beta_s^*) - D_s^T \Omega^{-1}(\beta_T^*) D_s$ *is non-negative definite.*

Proof. From Assumption 3.1 both $\Omega(\beta_T^*)^{-1}$ and $\Omega(\beta_s^*)^{-1}$ are positive definite. The $\Omega(\beta_T^*)$ is sub-block of $\Omega(\beta_s^*)$, we define $\Omega = \Omega(\beta_T^*)$ and define block matrix $\Omega(\beta_s^*) = \begin{bmatrix} \Omega & \check{\Omega} \\ \check{\Omega}^T & \widetilde{\Omega} \end{bmatrix}$, where $\widetilde{\Omega}$ is a positive definite $d_s \times d_s$ matrix and $\check{\Omega}$ a $d_s \times (d_s - d_T)$ matrix. From the inverse of block matrix, we have

$$\begin{bmatrix} \Omega & \check{\Omega} \\ \check{\Omega}^T & \widetilde{\Omega} \end{bmatrix}^{-1} = \begin{bmatrix} \Omega^{-1} + \Omega^{-1}\check{\Omega}(\widetilde{\Omega} - \check{\Omega}^T\Omega^{-1}\check{\Omega})^{-1}\check{\Omega}^T\Omega^{-1} & -\Omega^{-1}\check{\Omega}(\widetilde{\Omega} - \check{\Omega}^T\Omega^{-1}\check{\Omega})^{-1} \\ -(\widetilde{\Omega} - \check{\Omega}^T\Omega^{-1}\check{\Omega})^{-1}\check{\Omega}^T\Omega^{-1} & (\widetilde{\Omega} - \check{\Omega}^T\Omega^{-1}\check{\Omega})^{-1} \end{bmatrix},$$

where the sub matrix $\widetilde{\Omega} - \check{\Omega}^T\Omega^{-1}\check{\Omega}$ is also invertible. In addition by the definition of D_s, we have $D_s^T\Omega^{-1}(\beta_T^*)D_s = \begin{bmatrix} \Omega^{-1} & 0 \\ 0 & 0 \end{bmatrix}$. For any $d_s \times 1$ vector ι, let ι_1 be a $d_T \times 1$ vector denoting the first d_T elements and ι_2 be a $d_s - d_T \times 1$ vector denoting the next $d_s - d_T$ elements.

we can show that $\iota^T M_{s/T} \iota$ is non-negative and therefore matrix $M_{s/T}$ is non-

negative

$$
\begin{aligned}
&\iota^T M_{s/T}\iota \\
&= \begin{bmatrix} \iota_1 \\ \iota_2 \end{bmatrix}^T \left(\begin{bmatrix} \Omega & \check{\Omega} \\ \check{\Omega}^T & \widetilde{\Omega} \end{bmatrix}^{-1} - \begin{bmatrix} \Omega^{-1} & 0 \\ 0 & 0 \end{bmatrix} \right) \begin{bmatrix} \iota_1 \\ \iota_2 \end{bmatrix} \\
&= \begin{bmatrix} \iota_1 \\ \iota_2 \end{bmatrix}^T \begin{bmatrix} \Omega^{-1}\check{\Omega}(\widetilde{\Omega} - \check{\Omega}^T\Omega^{-1}\check{\Omega})^{-1}\check{\Omega}^T\Omega^{-1} & -\Omega^{-1}\check{\Omega}(\widetilde{\Omega} - \check{\Omega}^T\Omega^{-1}\check{\Omega})^{-1} \\ -(\widetilde{\Omega} - \check{\Omega}^T\Omega^{-1}\check{\Omega})^{-1}\check{\Omega}^T\Omega^{-1} & (\widetilde{\Omega} - \check{\Omega}^T\Omega^{-1}\check{\Omega})^{-1} \end{bmatrix} \begin{bmatrix} \iota_1 \\ \iota_2 \end{bmatrix} \\
&= (\check{\Omega}^T\Omega^{-1}\iota_1 - \iota_2)^T(\widetilde{\Omega} - \check{\Omega}^T\Omega^{-1}\check{\Omega})^{-1}(\check{\Omega}^T\Omega^{-1}\iota_1 - \iota_2) \\
&\geq 0.
\end{aligned}
$$

■

4 Generalized Information Criterion (GIC)

The Chapter 3 has introduced the QBIC and proved its model selection consistency. However QBIC has a limitation that the working correlation matrix has to be identity matrix. To overcome this limitation, we introduce a new information criterion Generalized Information Criterion (GIC) in this chapter. We show that GIC is model selection consistent and with any arbitrary working correlation matrix. The following Chapter is structured as following. The Section 4.1 introduces the new information criterion. The Section 4.2 proves the model selection consistency of GIC. And the Section 4.3 covers the numerical analysis. Lastly Section 4.4 lists few extra lemmas and provides the proof details of all Lemmas which are too long to put in the main text.

4.1 Introduction to GIC

Since the model of GEE only requires assumptions on the first and second moments, the true likelihood is not specified. Alternatively, one can integrate the multivariate quasi score vectors to obtain the quasi-likelihood. However, such multivariate integration is path-dependent and does not lead to a unique quasi-likelihood. In Pan (2001)'s QIC and Chapter 3's QBIC, the quasi-likelihood of each observation from a cluster is added together under a working independence model assumption.

Consider a divergent number p_n of covariates where $p_n \to \infty$, and $p_n \leq n$. Let s be a subset of $\{1, 2 \dots p_n\}$. The model with $\beta_j = 0$ for all $j \notin s$ is denoted as a model s. Let $\widehat{\beta}_s$ denote the GEE estimator under the model s. We propose to use the working covariance matrix and the fitted residual vectors to form a quadratic form and use it as a goodness-of-fit measure for the model s:

$$Q(\widehat{\beta}_s) = \frac{1}{2}\sum_{i=1}^{n}\{Y_i - \mu(\widehat{\beta}_s)\}^T A_i(\widehat{\beta}_F)^{-1/2} R^{-1} A_i(\widehat{\beta}_F)^{-1/2}\{Y_i - \mu(\widehat{\beta}_s)\}, \tag{4.1}$$

where $\widehat{\beta}_F$ denotes the GEE estimates under the full model. Using $\widehat{\beta}_F$ in the variance function is to ensure that the variances are consistently estimated. In the quadratic form, the working correlation matrix R can be any positive definite matrix with diagonal entries equal to one. Note that both $A_i(\widehat{\beta}_F)$ and R remain the same for different competing models in equation (4.1). The estimated variances $A_i(\widehat{\beta}_F)$ are evaluated under the full model. This is in spirit similar to the Mallows

C_p statistics using standard error obtained from the model using all predictors. Let $\widehat{V}_i = A_i(\widehat{\beta}_F)^{-1/2} R^{-1} A_i(\widehat{\beta}_F)^{-1/2}$, equation (4.1) can be reformulated as:

$$Q(\widehat{\beta}_s) = \frac{1}{2} \sum_{i=1}^{n} \{Y_i - \mu(\widehat{\beta}_s)\}^T \widehat{V}_i^{-1} \{Y_i - \mu(\widehat{\beta}_s)\}. \tag{4.2}$$

Equation (4.2) is similar to Carey and Wang (2011)'s Gaussian pseudo-likelihood which takes the form of $-2^{-1}\{\sum_{i=1}^{n}\{Y_i - \mu(\widehat{\beta}_s)\}^T V_i(\widehat{\beta}_s)^{-1}\{Y_i - \mu(\widehat{\beta}_s)\} + \log(|V_i(\widehat{\beta}_s)|)\}$, whose V_i depends on $\widehat{\beta}_s$. And $|V_i(\widehat{\beta}_s)|$ denotes the determinant of matrix $V_i(\widehat{\beta}_s)$. Similarly, Kim et al. (2012) used weighted sum of squares of residuals as a goodness-of-fit measure to construct information criteria in linear regression. The quadratic form can be considered as the extension of weighted sum of squares of residuals to incorporate the within cluster correlation among the observations.

Let T denote the true model and d_T be the size of the true model T. Let β_T^* denote the true values of the parameters under the model T. Consider all the competing models s in the model space S. Let d_s denote the number of covariates in the model s, with s_n being the upper bound of d_s in S, and $d_T \leq s_n \leq p_n$. s_n can go up to p_n. If s is overfitting, $T \subseteq s$; whereas if s is underfitting, $T \nsubseteq s$. The sets of underfitting models and overfitting models are denoted as S_- and S_+ respectively. Since GIC uses the full model to estimate β_F and $\widehat{V}_i$, we will use p_n as the upper bound of candidate models. In the later part of this chapter we have $s_n = p_n$.

The true parameter values under an overfitting model s are denoted as β_s^*, where

the common d_T elements are the same as β_T^* and the rest of $d_s - d_T$ elements are zero. For any underfitting model $s \in S_-$, we assume there exists a unique pseudo true parameters β_s^* such that $\sum_{i=1}^n D_i(\beta_s^*)^T V_i(\beta_s^*)^{-1}\{\mu_i(\beta_T^*) - \mu_i(\beta_s^*)\} = 0$. This definition of pseudo true parameter values is similar to the definition used in the maximum likelihood estimation under mis-specified models (White, 1981, 1982).

We propose the following Generalized Information Criterion for model selection on GEE models:

$$GIC(s) = 2Q(\widehat{\beta}_s) + d_s^* \gamma_n. \tag{4.3}$$

The first term of GIC is the quadratic form, which reflects the goodness-of-fit for a given model s, while the second term is the penalty for model complexity, which enforces sparsity on the selected model. The γ_n is a sequence of penalties on the complexity of the model, and d_s^* is the effective degrees of freedom of the model s (Pan, 2001; Varin and Vidoni, 2005; Gao and Song, 2010). We define $d_s^* = \mathrm{tr}\{W_s(\beta_s^*)\Omega_s^{-1}(\beta_s^*)\}$, where the variability matrix $W(\beta_s^*) = n^{-1}\mathrm{Cov}\{U(\beta_s^*)\}$ and the sensitivity matrix $\Omega(\beta_s^*) = -n^{-1}\mathrm{E}\{\partial U(\beta_s)/\partial \beta_s^T|_{\beta_s^*}\}$. If the working correlation is correctly specified, and the marginal regression model is the true model T, the variability matrix and sensitivity matrix are the same and $d_s^* = d_T$. If the model s is the true or overfitting model, as $\mathrm{E}\{Y_i - \mu_i(\beta_s^*)\} = 0$, the variability matrix and sensitivity matrix can be expressed as $W(\beta_s^*) = n^{-1}\sum_{i=1}^n D_i(\beta_s^*)^T V_i(\beta_s^*)^{-1}\mathrm{Cov}(Y_i)V_i(\beta_s^*)^{-1}D_i(\beta_s^*)$ and

$\Omega(\beta_s^*) = n^{-1} \sum_{i=1}^n D_i(\beta_s^*)^T V_i(\beta_s^*)^{-1} D_i(\beta_s^*)$.

4.2 GIC Model Selection Consistency

Assumption 4.1 *The cluster size m is finite, the number of covariates goes to infinity $p_n \to \infty$ and $p_n^5 \log p_n / n \to 0$. Assume that* $\liminf_n \min_{s \in S_-} n^{-1} [\sum_{i=1}^n \{\mu_i(\beta_T^*) - \mu_i(\beta_s^*)\}^T \{\mu_i(\beta_T^*) - \mu_i(\beta_s^*)\}] / (p_n^3 \log p_n / n)^{1/2} = \infty$.

Assumption 4.1 implies that the minimum distance between the true model T and any competing underfitting model s is allowed to converge to zero but at a rate slower than $(p_n^3 \log p_n / n)^{1/2}$.

Assumption 4.2 *For any model $s \in S$ and any β_s in the small neighborhood $||\beta_s - \beta_s^*|| \leq (p_n^2 \log p_n / n)^{1/2}$, there exist two positive value b_1 and b_2 that all the eigenvalues of $\Omega(\beta_s)$, $W(\beta_s)$, $n^{-1} \sum_{i=1}^n X_i^T X_i$ and $Cov(Y_i)$, $i \in \{1, 2 \ldots n\}$, are bounded from below by b_1 and bounded from above by b_2.*

We define the linear predictor function $\zeta_{ij}(\beta) = X_{ij}^T \beta$, the mean function $\mu_{ij}(\beta) = g^{-1}\{\zeta_{ij}(\beta)\}$ and the variance function $A_{ij}(\beta) = \nu\{\mu_{ij}(\beta)\} = \nu[g^{-1}\{\zeta_{ij}(\beta)\}]$. Let $\Lambda_{ij}(\beta) = \partial \mu_{ij} / \partial \zeta_{ij}$ and $\Lambda_i(\beta) = diag\{\Lambda_{ij}(\beta), j = 1, 2 \ldots m\}$, a diagonal matrix of dimension m.

Assumption 4.3 *For all* $s \in S$ *and any* β *in the neighborhood of* $||\beta - \beta_s^*|| \leq (p_n^2 \log p_n/n)^{1/2}$, *there exist positive values* b_3 *and* b_4 *such that* $b_3 < A_{ij}(\beta_s^*) < b_4$, $|\Lambda_{ij}(\beta_s^*)| < b_4$, $|\mu_{ij}(\beta_s^*)| < b_4$, *and derivatives* $|\partial \Lambda_{ij}(\beta)/\partial \beta_{[k]}|$, $|\partial^2 A_{ij}(\beta)/\partial \beta_{[k]} \partial \beta_{[l]}|$, $|\partial^3 \mu_{ij}(\beta)/\ \partial \beta_{[k]} \partial \beta_{[l]} \partial \beta_{[r]}|$ *exist and bounded from above by* b_4, *for all* $i \in \{1, 2 \dots n\}$, $j \in \{1, 2 \dots m\}$, $k, l, r \in \{1, 2 \dots p_n\}$.

In this article, large deviation results are used as an important tool to establish the estimation consistency and model selection consistency in large p_n settings. Let ψ denote a random vector and O denote a positive definite matrix. Large deviation results for quadratic form $\psi^T O \psi$ were established by Spokoiny and Zhilova (2013) under an exponential moment condition:

$$\log[\mathrm{E}\{\exp(t^T \psi)\}] \leq ||t||^2/2, ||t|| \leq \rho, \tag{4.4}$$

where ρ is a positive constant. Define $P_G = \mathrm{Tr}[O]$ and $V_G^2 = \mathrm{Tr}[O^2]$. Based on Corollary 4.2 in Spokoiny and Zhilova (2013), for $\rho^2/4 > K > V_G/3$,

$$\Pr(\psi^T O \psi > P_G + K) \leq 10.4 \exp(-K/6). \tag{4.5}$$

This key result establishes the exponential decay of the tail probability for a quadratic form. Such exponential decay rate is crucial for the control of the overall models selection error. We will show that by choosing an appropriate penalty term, the model selection error rate for each competing model can be derived using equation

(4.5), which is exponentially small . The total number of competing model is of the order of $p_n^{p_n}$. By Bonferroni inequality, the overall models selection error rate will be less than the sum of each individual error and the sum can be controlled to have the limiting value of zero. Gao and Carroll (2017) show that the exponential moment condition in equation (4.4) can be satisfied asymptotically by sample mean types of statistics if the original random vector satisfies the following cumulant boundedness condition.

Let $Q_i(\beta) = \{Y_i - \mu_i(\beta)\}^T \widehat{V}_i^{-1} \{Y_i - \mu_i(\beta)\}$ and $U_i(\beta) = D_i(\beta)^T V_i(\beta)^{-1} \{Y_i - \mu_i(\beta_s)\}$, with $\widehat{V}_i = A_i(\widehat{\beta}_F)^{-1/2} R^{-1} A_i(\widehat{\beta}_F)^{-1/2}$. Let $U_i(\beta)_{[k]}$ denote the kth element of vector $U_i(\beta)$, $U_i(\beta)^{(1)}_{[kl]}$ denote $\partial U_i(\beta)_{[k]}/\partial \beta_{[l]}$, and $U_i(\beta)^{(2)}_{[klr]}$ denote $\partial U_i(\beta)^{(1)}_{[kl]}/\partial \beta_{[r]}$.

Assumption 4.4 *There exists an neighborhood* $||\beta_s - \beta_s^*|| \leq b_7$, *such that* $Q_i(\beta_s^*)$, $U_i(\beta_s^*)_{[k]}$, $U_i(\beta_s^*)^{(1)}_{[kl]}$, *and* $U_i(\beta_s)^{(2)}_{[klr]}$ *satisfy the cumulant boundedness condition in Definition 1 uniformly for all models* $s \in S$.

Based on the cumulant boundedness condition in Assumption 4.4, using large deviation result in Spokoiny and Zhilova (2013) and Gao and Carroll (2017), we obtain the asymptotic orders of the following terms.

Lemma 4.1 *Under Assumption 4.4, for all* $k, l, r \in \{1, 2 \dots p_n\}$, *all models* $s \in S$, *and* β_s *in the neighborhoods* $||\beta_s - \beta_s^*|| \leq (p_n^2 \log p_n / n)^{1/2}$, *the zero-centered*

terms $|Q(\beta_s^*) - E\{Q(\beta_s^*)\}|$, $|U(\beta_s^*)_{[k]} - E\{U(\beta_s^*)_{[k]}\}|$, $|U(\beta_s^*)^{(1)}_{[kl]} - E\{U(\beta_s^*)^{(1)}_{[kl]}\}|$ *and* $|U(\beta_s)^{(2)}_{[klr]} - E\{U(\beta_s)^{(2)}_{[klr]}\}|$ *are of order* $O_p\{(np_n \log p_n)^{1/2}\}$ *uniformly.*

Next we investigate the consistency of the GEE estimator under different competing models.

Theorem 4.1 *Under Assumptions 4.1 - 4.4, as* $n \to \infty$, *there exists a solution* $\widehat{\beta}_s$ *to the score equation* $U(\widehat{\beta}_s) = 0$ *such that it falls within an* $(p_n^2 \log p_n/n)^{1/2}$ *neighborhood of* β_s^* *for all* $s \in S$ *with probability tending to 1.*

The proof of Theorem 4.1 is the same as Theorem 3.1 given the same assumptions. The only change is to replace all of s_n to p_n. s_n is the model size up bound and can go as large as p_n. The replacement from s_n to p_n does not change the proofs. So we do not repeat the proof here.

Lemma 4.2 *Under Assumptions 4.1 - 4.4, for all models* $s \in S_+$, *and* $i = 1, 2 \ldots n$, $\max[|\lambda_{\max}\{V_i(\widehat{\beta}_F)^{-1} - V_i(\widehat{\beta}_s)^{-1}\}|, |\lambda_{\min}\{V_i(\widehat{\beta}_F)^{-1} - V_i(\widehat{\beta}_s)^{-1}\}|] = O_p\{(p_n^3 \log p_n/n)^{1/2}\}$ *and* $\max[|\lambda_{\max}\{V_i(\widehat{\beta}_F)^{-1} - V_i(\beta_s^*)^{-1}\}|, |\lambda_{\min}\{V_i(\widehat{\beta}_F)^{-1} - V_i(\beta_s^*)^{-1}\}|] = O_p\{(p_n^3 \log p_n/)^{1/2}\}$.

For true and overfitting models, Lemma 4.2 measures the distance between the two matrices $V_i(\widehat{\beta}_s)$ and $V_i(\beta_s^*)$.

Next we will establish the model selection consistency of the proposed GIC under "large n and divergent p_n scenario". Our approach consists of two steps. First,

we show that the difference in the goodness-of-fit measures between two competing models s and T can be approximated by quadratic forms and the approximation errors are uniformly bounded across the model spaces. Second, based on the quadratic forms, we apply the large deviation result to quantify the size of the penalty γ_n.

Lemma 4.3 *Under Assumptions 4.1 - 4.4, there exists a matrix Res_d that all elements in the matrix are at the order of $O_p\{(p_n^3 \log p_n/n)^{1/2}\}$ such that $\widehat{\beta}_s - \beta_s^* = n^{-1}\{\Omega(\beta_s^*) + Res_d\}^{-1}U(\beta_s^*)$, where the $O_p\{(p_n^3 \log p_n/n)^{1/2}\}$ term uniformly holds for all models $s \in S_+$.*

Lemma 4.3 approximates the distance of $\widehat{\beta}_s$ to β_s^* as the product of a small perturbation of information matrix and the score vector.

Lemma 4.4 *Under Assumptions 4.1 - 4.4, the differences between the goodness-of-fit measures can be approximated as quadratic forms:*

$$\begin{aligned} 2\{Q(\widehat{\beta}_s) - Q(\beta_s^*)\} &= -n(\beta_s^* - \widehat{\beta}_s)^T\Omega(\beta_s^*)(\beta_s^* - \widehat{\beta}_s)\{1 + o_p(1)\} \\ &= -n^{-1}U^T(\beta_s^*)\Omega^{-1}(\beta_s^*)U(\beta_s^*)\{1 + o_p(1)\}, \end{aligned}$$

where the $o_p(1)$ term holds for all models $s \in S_+$.

Lemma 4.4 show that the differences in the goodness-of-fit measures can be approximated by the score type and the Wald type quadratic forms. Next Lemma 4.5 establishes the asymptotic order of these quadratic forms.

Lemma 4.5 *Under Assumptions 4.1 - 4.4,* $\sup_{s\in S_+} |Q(\widehat{\beta}_s) - Q(\beta_s^*)| = O_p(p_n^2 \log p_n)$, *and* $\sup_{s\in S_-} |Q(\widehat{\beta}_s) - Q(\beta_s^*)| = O_p\{(np_n^3 \log p_n)^{1/2}\}$.

We now establish the consistency result for the proposed generalized information criterion. For any overfitting model s, define a matrix $D_s = (I_{d_T}, 0_{d_T, d_s - d_T})$, with I_{d_T} being an identity matrix with dimension $d_T \times d_T$, and $0_{d_T,d_s-d_T}$ denoting a matrix of zeros with dimension of $d_T \times (d_s - d_T)$. For every overfitting model s, let Δ_s denote the quadratic form $n^{-1}U^T(\beta_s^*)\ \Omega^{-1}(\beta_s^*)U(\beta_s^*)$. According to Lemma 4.4, we have $2Q(\widehat{\beta}_s) - 2Q(\widehat{\beta}_T) = -\Delta_{s/T}\ \{1 + o_p(1)\}$,with $\Delta_{s/T} = n^{-1}U^T(\beta_s^*)M_{s/T}U(\beta_s^*)$, where $M_{s/T}$ denotes the difference matrix $\Omega^{-1}(\beta_s^*) - D_s^T\Omega^{-1}(\beta_T^*)D_s$.

Lemma 4.6 *Under Assumptions 4.1 - 4.4, for overfitting model* $s \in S_+$, $M_{s/T} = \Omega^{-1}(\beta_s^*) - D_s^T\Omega^{-1}(\beta_T^*)D_s$ *is non-negative definite.*

Define $C_s = W^{1/2}(\beta_s^*)M_{s/T}W^{1/2}(\beta_s^*)$. It can be shown that $\mathrm{Tr}(C_s) = d_s^* - d_T^*$. Let $\omega = \max_{s\in S}(d_s^* - d_T^*)/(d_s - d_T)$, the ratio of effective degrees of freedom over the true degrees of freedom. For true likelihood setting, $\omega = 1$.

Lemma 4.7 *Assume* ω *is bounded away from zero and infinity. Let* $\gamma_n = 6\omega(1 + \gamma)\log p_n$ *for some* $\gamma > 0$ *or* $\gamma_n = 6\omega(1 + \log\log p_n)\log p_n$. *Under Assumptions 4.1 - 4.4,*

$$\Pr\{\max_{s\in S_+\setminus T} \Delta_{s/T}/(d_s^* - d_T^*) \geq \gamma_n\} = o(1).$$

Theorem 4.2 *Assume ω is bounded bounded away from zero and infinity. Let $\gamma_n = 6\omega(1+\gamma)\log p_n$ for some $\gamma > 0$ or $\gamma_n = 6\omega(1+\log\log p_n)\log p_n$. Under Assumptions 4.1 - 4.4, as $n \to \infty$,*

$$\Pr\{\min_{s\in S} GIC(s) \geq GIC(T)\} \to 1.$$

Proof of Theorem 4.2. First for overfitting models $s \in S_+ \setminus T$, we have

$$\begin{aligned}
&\min_{s\in S_+\setminus T} \mathrm{GIC}(s) - \mathrm{GIC}(T) \\
=&\min_{s\in S_+\setminus T} [2\{Q(\widehat{\beta}_s) - Q(\widehat{\beta}_T)\} + (d_s^* - d_T^*)\gamma_n] \\
\geq& -\max_{s\in S_+\setminus T} [\Delta_{s/T} - (d_s^* - d_T^*)\gamma_n + o_p(1)].
\end{aligned}$$

According to Lemma 4.7, $\Pr\{\max_{s\in S_+\setminus T} \Delta_{s/T}/(d_s^* - d_T^*) \geq \gamma_n\} = o(1)$. Therefore $\Pr\{\min_{s\in S+} \mathrm{GIC}(s) \geq \mathrm{GIC}(T)\} \to 1$. Next for the underfitting models, we have $\min_{s\in S_-} \mathrm{GIC}(s) \geq \mathrm{GIC}(T) = \min_{s\in S_-}[2\{Q(\widehat{\beta}_s) - Q(\widehat{\beta}_T)\} + (d_s^* - d_T^*)\gamma_n]$. We further decompose the difference in the quadratic forms:

$$\begin{aligned}
&Q(\widehat{\beta}_s) - Q(\widehat{\beta}_T) \\
=&Q(\widehat{\beta}_s) - Q(\beta_s^*) + Q(\beta_s^*) - Q(\beta_T^*) + Q(\beta_T^*) - Q(\widehat{\beta}_T) \\
=&\{Q(\widehat{\beta}_s) - Q(\beta_s^*)\} + \{Q(\beta_T^*) - Q(\widehat{\beta}_T)\} + [Q(\beta_s^*) - Q(\beta_T^*) - \mathrm{E}\{Q(\beta_s^*) - Q(\beta_T^*)\}] \\
&+ [\mathrm{E}\{Q(\beta_s^*) - Q(\beta_T^*)\}].
\end{aligned}$$

Based on Lemma 4.1, $Q(\beta_s^*) - Q(\beta_T^*) - \mathrm{E}\{Q(\beta_s^*) - Q(\beta_T^*)\} = O_p\{(np_n \log p_n)^{1/2}\}$. Lemma 4.5 implies that $Q(\widehat{\beta}_T) - Q(\beta_T^*) = O_p(p_n^2 \log p_n)$ and $Q(\beta_s^*) - Q(\widehat{\beta}_s) =$

$O_p\{(np_n^3 \log p_n)^{1/2}\}$. Next we determine the order of $\mathrm{E}\{Q(\beta_s^*) - Q(\beta_T^*)\}$. First we estimate the order of following term.

$$\begin{aligned}
&\sum_{i=1}^{n} 2\mathrm{E}[\{Y_i - \mu_i(\beta_T^*)\}^T \widehat{V}_i^{-1}\{\mu_i(\beta_T^*) - \mu_i(\beta_s^*)\}] \\
&= \sum_{i=1}^{n} 2\mathrm{E}[\{Y_i - \mu_i(\beta_T^*)\}^T (\widehat{V}_i^{-1} - V_i^{*^{-1}})\{\mu_i(\beta_T^*) - \mu_i(\beta_s^*)\}] \\
&+ \sum_{i=1}^{n} 2\mathrm{E}[\{Y_i - \mu_i(\beta_T^*)\}^T V_i^{*^{-1}}\{\mu_i(\beta_T^*) - \mu_i(\beta_s^*)\}] \\
&= \sum_{i=1}^{n} 2\mathrm{E}[\{Y_i - \mu_i(\beta_T^*)\}^T (\widehat{V}_i^{-1} - V_i^{*^{-1}})\{\mu_i(\beta_T^*) - \mu_i(\beta_s^*)\}].
\end{aligned}$$

According to Lemma 4.12, $\mathrm{E}\{n^{-1}\sum_{i=1}^{n} |Y_{ij} - \mu_{ij}(\beta_T^*)|\}$ is bounded. Based on Lemma 4.8 and Lemma 4.2, $||\mu_i(\beta_T^*) - \mu_i(\beta_s^*)||_{\max}$ is bounded for all i and $||\widehat{V}_i^{-1} - V_i^{*^{-1}}||_{\max} = O_p\{(p_n^3 \log p_n/n)^{1/2}\}$. This means $\sum_{i=1}^{n} 2\mathrm{E}[\{Y_i - \mu_i(\beta_T^*)\}^T \widehat{V}_i^{-1}\{\mu_i(\beta_T^*) - \mu_i(\beta_s^*)\}] = O_p\{(np_n^3 \log p_n)^{1/2}\}$. Next we estimate the order of $\mathrm{E}\{Q(\beta_s^*) - Q(\beta_T^*)\}$ and show that it is the leading term.

$$
\begin{aligned}
&2\mathrm{E}\{Q(\beta_s^*) - Q(\beta_T^*)\} \\
&= \mathrm{E}[\sum_{i=1}^{n}\{Y_i - \mu_i(\beta_T^*) + \mu_i(\beta_T^*) - \mu_i(\beta_s^*)\}^T \widehat{V}_i^{-1}\{Y_i - \mu_i(\beta_T^*) + \mu_i(\beta_T^*) - \mu_i(\beta_s^*)\} \\
&\quad - \{Y_i - \mu_i(\beta_T^*)\}^T \widehat{V}_i^{-1}\{Y_i - \mu_i(\beta_T^*)\}] \\
&= \mathrm{E}[\sum_{i=1}^{n}\{\mu_i(\beta_s^*) - \mu_i(\beta_T^*)\}^T \widehat{V}_i^{-1}\{\mu_i(\beta_s^*) - \mu_i(\beta_T^*)\}] \\
&\quad + \sum_{i=1}^{n} 2\mathrm{E}[\{Y_i - \mu_i(\beta_T^*)\}^T \widehat{V}_i^{-1}\{\mu_i(\beta_T^*) - \mu_i(\beta_s^*)\}] \\
&\geq \mathrm{E}\{\lambda_{\min_i}(\widehat{V}_i^{-1})\} \sum_{i=1}^{n}\{\mu_i(\beta_s^*) - \mu_i(\beta_T^*)\}^T\{\mu_i(\beta_s^*) - \mu_i(\beta_T^*)\} + O_p\{(np_n^3 \log p_n)^{1/2}\}.
\end{aligned}
$$

Assumption 4.2 implies that $\lambda_{\min_i}(\widehat{V}_i^{-1})$ is a positive value bounded away from zero. Furthermore based on Assumption 4.1, $\sum_{i=1}^{n}\{\mu_i(\beta_s^*) - \mu_i(\beta_T^*)\}^T \ \{\mu_i(\beta_s^*) - \mu_i(\beta_T^*)\}/\ (np_n^3 \log p_n)^{1/2} \to \infty$. This means $\mathrm{E}\{Q(\beta_s^*) - Q(\beta_T^*)\}/(np_n^3 \log p_n)^{1/2} \to \infty$. As ω is bounded, $|d_s^* - d_T^*| \leq \omega|d_s - d_T| = O(p_n)$. So $\mathrm{E}\{Q(\beta_s^*) - Q(\beta_T^*)\}$ is the leading term in the difference between the two information criteria. Thus we have

$$
\Pr\{\min_{s \in S_+} \mathrm{GIC}(s) \geq \mathrm{GIC}(T)\} \to 1.
$$

■

Through all the asymptotic discussions above, we rely on the full model of size p_n to obtain the consistent variance estimate $\widehat{V}_i$. Alternatively, we can constrain the competing models all bounded by size s_n and assume $s_n \ll p_n$. As long as we can

identify a model of size s_n which is an overfitting model, we can obtain a consistent variance estimate at this model. If so, the sample size requirement of $p_n^5 \log p_n/n \to 0$ can be relaxed to $s_n^5 \log p_n/n \to 0$, where p_n can be allowed to be greater than n.

4.3 Numerical Study

4.3.1 Estimation of working correlation matrix

Section 4.2 illustrates that the generalized information criterion is selection consistent with the working correlation matrix R being any arbitrary positive definite matrix. Hence the selection consistency is robust against mis-specification of the working correlation. This matrix R needs to be fixed when we compare the generalized information criterion across different competing models. In practice, the choice of working correlation matrix R used in the criterion could impact its model selection efficiency. In our simulation, we compare different choices of R including independent, exchangeable, AR-1, and unstructured working correlation. Balan and Schiopu-Kratina (2005) suggested using the formula below to estimate the unstructured working correlation matrix

$$\widehat{R}_B = \frac{1}{n}\sum_{i=1}^{n} A_i^{-1/2}(\widetilde{\beta}_F)\{Y_i - \mu_i(\widetilde{\beta}_F)\}\{Y_i - \mu_i(\widetilde{\beta}_F)\}^T A_i^{-1/2}(\widetilde{\beta}_F),$$

where $\widetilde{\beta}_F$ is a preliminary consistent estimator under the full model using the independent working correlation matrix. Wang (2011) proved that under "large n diverging p_n" situation, the estimated working correlation matrix is $(p_n/n)^{1/2}$ consistent to the true correlation matrix.

4.3.2 Simulations

We simulate both clustered binary response for discrete case and clustered Gaussian response for continuous case. We consider different settings with sample size $n = 1000$ or 500, the number of covariates $p_n = 1000$ or 500, and the cluster size $m = 10$ or 5. The number of true covariates d_T is set be 50. For $j = 1, ..., d_T$, β_j is drawn from the uniform distribution $U(0.05, 0.5)$, whereas for $j = d_T+1, d_T+2 \ldots p_n$, β_j is set to zero. For the jth observation in the ith cluster, we simulate the associated covariates $X_{ij} = (x_{ij1} \ldots x_{ijp_n})^T$, and the mean parameter is denoted as $\mu_{ij} = \text{logit}^{-1}(X_{ij}^T\beta)$ for binary response or $\mu_{ij} = X_{ij}^T\beta$ for Gaussian response. The covariates X_{ijk} are partitioned into independent blocks of 50 covariates, and within each block the 50 covariates are simulated from the multivariate normal distribution with variances equal to 1 and off-diagonal covariances all equal to $0.5^{|k-k'|}$, where k and k' index for the covariates. In each cluster i, for binary response Y_i is simulated from a multivariate binary distribution with mean μ_i and an unstructured corre-

lation matrix; for Gaussian response Y_i is simulated from a multivariate Gaussian distribution with mean μ_i, variance 1, and an unstructured correlation matrix. The r package "SimCorMultRes" is used to simulate the correlated multivariate binary distribution. We use the LASSO to generate a sequence of subset models and use the proposed generalized information criterion to select the best subset model. With regard to the penalty term, Theorem 4.2 provides a theoretical value of $6\omega \times d_s^* \log p_n$. We set the penalty term to be $c \times d_s^* \log p_n$, where c is a constant multiplicative factor and we vary c from 1 to 4. This penalty term has the same asymptotic order as the theoretical penalty term. We run 100 simulations and evaluate the mean and standard deviation of the Positive Selection Rates (PSR) and False Discovery Rates (FDR) of Pan (2001)'s QIC and our proposed QBIC and GIC.

Table 4.1 and Table 4.2 compares the PSR and FDR of the proposed QBIC and GIC with QIC when $c = 1$. It is shown that QIC has largely inflated FDR, whereas the proposed QBIC and GIC has a good error rate control. For example, when $n = 1000$ and $p_n = 1000$, the FDR of the QIC can be as high as 70%, while the FDR of the QBIC and GIC are about 5%. This demonstrates that with large p_n, QIC tends to select overfitting models. This is due to the fact that the QIC uses the AIC type of penalty, which is too small to control the error rate. Although both QBIC and GIC largely outperform QIC, GIC's simulation result is slightly better comparing

to QBIC. The reason could spring from the choice of working correlation matrix. QBIC's model selection consistency requires independent working correlation matrix, which may not accurately measure the goodness of fitting. We also observe that as the underlying true correlation matrix is an unstructured correlation matrix, the choice of R using the formula from Balan and Schiopu-Kratina (2005) outperforms the independent (Ind), exchangeable (Exc), and Autoregressive-1 (AR1) correlation matrices. We vary the multiplicative factor of c from 1 to 4 and examine how the sensitivity and selectivity of our method changes. When c changes from 1 to 4, we found that the QBIC's PSR and FDR and GIC's PSR and FDR both decrease slightly. More simulation details refer to Table A.1 and Table A.2 at Appendix part.

4.3.3 Real Data Analysis

We apply our proposed model selection method to the University of Michigan Health and Retirement Study (HRS) data. The data is generated from a longitudinal study which surveys approximately 20,000 senior people in America. Information about their financial situations, family structures and different health factors were collected every two years in the last two decades. We use the proposed model selection method to choose important predictors on the depression status of seniors. In total, there are 2,652 individuals who provided 10 repeated depression status measurements from 1996 to 2014. There are 316 valid covariates with less than 4% of missing data. We perform five-fold cross validation to evaluate the prediction power of the selected subset model. The Receiver Operating Characteristic (ROC) curve from Figure 4.1 demonstrates that the prediction power of the selected subset models by GIC, QBIC, and QIC are satisfactory. GIC and QBIC are generating the same model selection result. QIC has slightly higher sensitivity rate and higher specificity rate than that of GIC and QBIC. In terms of variable selection, GIC and QBIC chooses a smaller subset model with 57 covariates in comparison to QIC's selected model with 84 covariates. Using different choices of working correlation matrices in the proposed GIC leads to similar performance.

Table 4.1: GIC, QBIC, and QIC Simulation Result for Binary Response

The true parameters size d_T is 50 and the cluster size m is 10. The free multiplicative constant c for the penalty is 1.

		n 1000	p 1000			n 500	p 500		
		mean psr	std psr	mean fdr	std fdr	mean psr	std psr	mean fdr	std fdr
QIC	Ind	1.0000	0.0000	0.7093	0.0241	0.9974	0.0084	0.5677	0.0735
	Exc	1.0000	0.0000	0.7099	0.0241	0.9974	0.0084	0.5677	0.0735
	AR1	1.0000	0.0000	0.7093	0.0241	0.9974	0.0084	0.5677	0.0735
	Uns	1.0000	0.0000	0.7234	0.0179	0.9976	0.0082	0.6163	0.0677
GIC	Ind	0.9982	0.0081	0.0596	0.0476	0.9182	0.0710	0.0250	0.0548
	Exc	0.9982	0.0081	0.0574	0.0454	0.9194	0.0715	0.0265	0.0561
	AR1	0.9980	0.0083	0.0584	0.0471	0.9180	0.0708	0.0246	0.0550
	Uns	0.9990	0.0066	0.0445	0.0399	0.9498	0.0538	0.0242	0.0381
QBIC	Ind	0.9994	0.0060	0.1192	0.0585	0.9576	0.0562	0.0653	0.0778

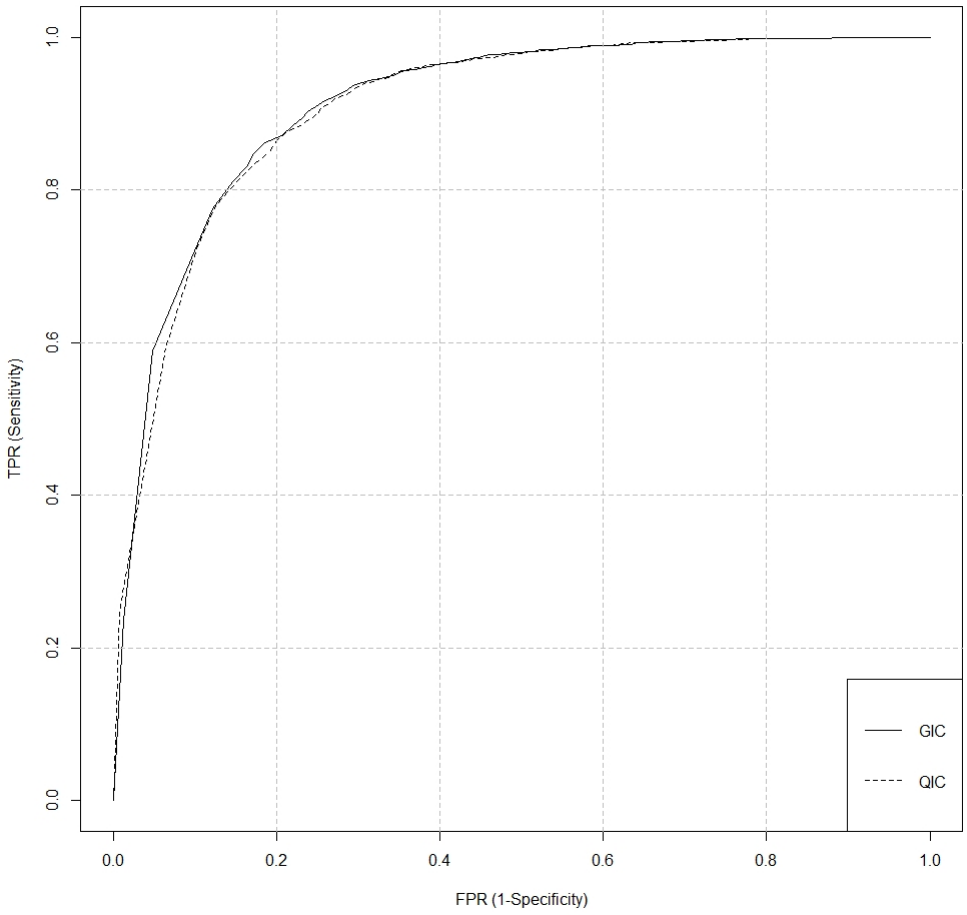
Table 4.2: GIC, QBIC, and QIC Simulation Result for Normal Response

The true parameters size d_T is 50 and the cluster size m is 10. The free multiplicative constant c for the penalty is 1.

		n 1000	p 1000			n 500	p 500		
		mean psr	std psr	mean fdr	std fdr	mean psr	std psr	mean fdr	std fdr
QIC	Ind	1.0000	0.0000	0.6937	0.0326	1.0000	0.0000	0.5401	0.0607
	Exc	1.0000	0.0000	0.6937	0.0326	1.0000	0.0000	0.5395	0.0599
	AR1	1.0000	0.0000	0.6937	0.0326	1.0000	0.0000	0.5401	0.0607
	Uns	1.0000	0.0000	0.7281	0.0136	1.0000	0.0000	0.7109	0.0324
GIC	Ind	1.0000	0.0000	0.1077	0.0460	1.0000	0.0000	0.0871	0.0449
	Exc	1.0000	0.0000	0.0961	0.0511	1.0000	0.0000	0.0705	0.0450
	AR1	1.0000	0.0000	0.1073	0.0471	1.0000	0.0000	0.0860	0.0461
	Uns	1.0000	0.0000	0.0226	0.0295	1.0000	0.0000	0.0272	0.0355
QBIC	Ind	1.0000	0.0000	0.1077	0.0460	1.0000	0.0000	0.0871	0.0449

Figure 4.1: ROC Curve of the selected subset models for the HRS Study

The unstructured correlation structure is used. True Positive Rate (TPR) is defined as sensitivity. False Positive Rate (FPR) is defined as 1-specificity. GIC and QBIC curve are the same and overlaped. That is why we do not label QBIC curve.



4.4 Proofs of Related Lemmas

There are few extra Lemmas that we did not list them in the Section 4.2, but they are useful. We firstly introduce these useful Lemmas and prove them at the beginning part of this section. Later we illustrate the proof of Lemmas that are mentioned in Section 4.2 which have not been proved yet.

Lemma 4.8 *Under Assumptions 4.1 - 4.4, for all β in the neighborhood of $||\beta - \beta_s^*|| < (p_n^2 \log p_n/n)^{1/2}$, $A_{ij}(\beta)$ are uniformly bounded away from zero and infinity as $n \to \infty$. Furthermore, $|\partial A_{ij}(\beta)/\partial\beta_{[k]}|$, $|A_{ij}^{-1/2}(\beta)|$, $|\partial A_{ij}^{-1/2}(\beta)/\ \partial\beta_{[k]}|$, $|\partial^2 A_{ij}^{-1/2}(\beta)/\ \partial\beta_{[k]}\partial\beta_{[l]}|$, $|\mu_{ij}(\beta)|$, $|\partial\mu_{ij}(\beta)/\partial\beta_{[k]}|$, $|\partial^2\mu_{ij}(\beta)/\ \partial\beta_{[k]}\partial\beta_{[l]}|$, $|\Lambda_{ij}(\beta)|$ are all bounded by b_a for all $i \in \{1, 2 \dots n\}$, $j \in \{1, 2 \dots m\}$, $k, l \in \{1, 2 \dots p_n\}$.*

Proof of Lemma 4.8. For all β in the neighborhood of $||\beta - \beta_s^*|| < (p_n^2 \log p_n/n)^{1/2}$, the boundedness of second derivative $|\partial^2 A_{ij}(\beta)/\ \partial\beta_{[k]}\partial\beta_{[l]}|$ in a compact set and boundedness of $A_{ij}(\beta_s^*)$ implies the boundedness of the first derivative $|\partial A_{ij}(\beta)/\partial\beta_{[k]}|$ and $|A_{ij}(\beta)|$. Based on Assumption 4.3, $A_{ij}(\beta_s^*) > b_3$, a positive constant, and the smoothness and boundedness of the $|\partial A_{ij}(\beta)/\partial\beta_{[k]}|$, $A_{ij}(\beta)$, and $A_{ij}^{-1/2}(\beta)$ are all bounded away from zero and infinity for all β in the neighborhood of β_s^*. By similar argument, $|\partial^2\mu_{ij}(\beta)/\partial\beta_{[k]}\partial\beta_{[l]}|$, $|\partial\mu_{ij}(\beta)/\partial\beta_{[k]}|$, $|\mu_{ij}(\beta)|$, $|\partial A_{ij}^{-1/2}(\beta)/\ \partial\beta_{[k]}|$, $|\partial^2 A_{ij}^{-1/2}(\beta)/\ \partial\beta_{[k]}\partial\beta_{[l]}|$, and $|\Lambda_{ij}(\beta)|$ are also uniformly bounded. ■

We introduce some extra notations. Let B and $\widetilde{B}$ denote $d_s \times m$ matrices. Let $D_i^{(1)}(\beta, \check{\beta}, B)$ be an $m \times d_s$ matrix and its jth row and kth column entry is $D_i^{(1)}(\beta, \check{\beta}, B)_{[jk]} = (\beta - \check{\beta})^T\{\partial^2 \mu_{ij}(B_{[,j]})/\partial\beta_{[k]}\partial\beta\}$. Let $D_i^{(2)}(\beta, \widetilde{\beta}, B, \widetilde{B})$ be an $m \times d_s$ matrix with the jth row and kth column entry as $D_i^{(2)}(\beta, \widetilde{\beta}, B, \widetilde{B})_{[jk]} = (\beta - \widetilde{\beta})^T\{\partial^3 \mu_{ij}(\widetilde{B}_{[,j]})/\partial\beta_{[k]}\partial\beta\partial\beta^T\}(B_{[,j]} - \beta)$.

Lemma 4.9 *Let $B_s^* = (\beta_s^*, \beta_s^* \dots \beta_s^*)$. Under Assumptions 4.1 - 4.4, for all model $s \in S$, $i \in \{1, 2 \dots n\}$, there exist $d_s \times m$ matrices $B_s^{\mu_i}$, $B_s^{\widetilde{\mu}_i}$, $B_s^{\check{\mu}_i}$, and $B_s^{D_i}$ such that each column of these four matrices is between $\widehat{\beta}_s$ and β_s^* and they satisfy:*

$$\mu_i(\widehat{\beta}_s) - \mu_i(\beta_s^*) = D_i(\beta_s^*)(\widehat{\beta}_s - \beta_s^*) + \frac{1}{2} D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{\mu_i})(\widehat{\beta}_s - \beta_s^*); \tag{4.6}$$

$$\mu_i(\beta_s^*) - \mu_i(\widehat{\beta}_s) = D_i(\widehat{\beta}_s)(\beta_s^* - \widehat{\beta}_s) + \frac{1}{2} D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^{\widetilde{\mu}_i})(\beta_s^* - \widehat{\beta}_s); \tag{4.7}$$

$$D_i(\widehat{\beta}_s) = D_i(\beta_s^*) + D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{D_i}); \tag{4.8}$$

$$D_i^{(2)}(\beta_s^*, \widehat{\beta}_s, B_s^{\widetilde{\mu}_i}, B_s^{\check{\mu}_i}) = D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^{\widetilde{\mu}_i}) - D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^*). \tag{4.9}$$

The max norms of the matrices have the following uniform bounds for all model

$s \in S$, $i \in \{1, 2 \dots n\}$:

$$||D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{\mu_i})||_{\max} = O_p\{(p_n^3 \log p_n/n)^{1/2}\},$$

$$|D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^{\widetilde{\mu}_i})||_{\max} = O_p\{(p_n^3 \log p_n/n)^{1/2}\},$$

$$||D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{D_i})||_{\max} = O_p\{(p_n^3 \log p_n/n)^{1/2}\},$$

$$||D_i^{(2)}(\beta_s^*, \widehat{\beta}_s, B_s^{\widetilde{\mu}_i}, B_s^{\breve{\mu}_i})||_{\max} = O_p(p_n^3 \log p_n/n).$$

Proof of Lemma 4.9. From Taylor expansion, there exists a $\beta_s^{\mu_{ij}}$ between β_s^* and $\widehat{\beta}_s$ such that

$$\mu_{ij}(\widehat{\beta}_s) - \mu_{ij}(\beta_s^*) = \frac{\partial \mu_{ij}(\beta_s^*)}{\partial \beta^T}(\widehat{\beta}_s - \beta_s^*) + \frac{1}{2}(\widehat{\beta}_s - \beta_s^*)^T \frac{\partial^2 \mu_{ij}(\beta_s^{\mu_{ij}})}{\partial \beta \partial \beta^T}(\widehat{\beta}_s - \beta_s^*). \quad (4.10)$$

Let $B_s^{\mu_i} = (\beta_s^{\mu_{i1}}, \beta_s^{\mu_{i2}} \dots \beta_s^{\mu_{im}})$ and each column of $B_s^{\mu_i}$ is between β_s^* and $\widehat{\beta}_s$. Define

$$D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, \beta_s^{\mu_i}) = \begin{bmatrix} (\widehat{\beta}_s - \beta_s^*)^T\{\partial^2 \mu_{i1}(\beta_s^{\mu_{i1}})/\partial \beta \partial \beta^T\} \\ (\widehat{\beta}_s - \beta_s^*)^T\{\partial^2 \mu_{i2}(\beta_s^{\mu_{i2}})/\partial \beta \partial \beta^T\} \\ \dots \\ (\widehat{\beta}_s - \beta_s^*)^T\{\partial^2 \mu_{im}(\beta_s^{\mu_{im}})/\partial \beta \partial \beta^T\} \end{bmatrix}.$$

Then Equation (4.10) can be reformulated as

$$\mu_i(\widehat{\beta}_s) - \mu_i(\beta_s^*) = D_i(\beta_s^*)(\widehat{\beta}_s - \beta_s^*) + \frac{1}{2} D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{\mu_i})(\widehat{\beta}_s - \beta_s^*).$$

Similarly if we perform Taylor Expansion at $\mu_i(\widehat{\beta}_s)$, we obtain

$$\mu_i(\beta_s^*) - \mu_i(\widehat{\beta}_s) = D_i(\widehat{\beta}_s)(\beta_s^* - \widehat{\beta}_s) + \frac{1}{2} D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^{\widetilde{\mu}_i})(\beta_s^* - \widehat{\beta}_s).$$

By similar argument, there exists a $\beta_s^{D_{ij}}$ between β_s^* and $\widehat{\beta}_s$ such that $\partial\mu_{ij}(\widehat{\beta}_s)/\partial\beta_{[k]} = \partial\mu_{ij}(\beta_s^*)/\partial\beta_{[k]} + (\widehat{\beta}_s - \beta_s^*)^T\{\partial^2\mu_{ij}(\beta_s^{D_{ij}})/\partial\beta_{[k]}\partial\beta\}$. Define $B_s^{D_i} = (\beta_s^{D_{i1}}, \beta_s^{D_{i2}} \dots \beta_s^{D_{im}})$ and we can reformulate the equation above as

$$D_i(\widehat{\beta}_s) = D_i(\beta_s^*) + D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, \beta_s^{D_i}).$$

According to Taylor Expansion, there exists a $\beta_s^{\check{\mu}_{ij}}$ between β_s^* and $\beta_s^{\mu_{ij}}$ such that

$$\begin{aligned}
&D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^{\tilde{\mu}_i})_{[jk]} - D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^*)_{[jk]} \\
&= (\beta_s^* - \widehat{\beta}_s)^T \frac{\partial^2\mu_{ij}(\beta_s^{\tilde{\mu}_{ij}})}{\partial\beta_{[k]}\partial\beta} - (\beta_s^* - \widehat{\beta}_s)^T \frac{\partial^2\mu_{ij}(\beta_s^*)}{\partial\beta_{[k]}\partial\beta} \\
&= (\beta_s^* - \widehat{\beta}_s)^T \frac{\partial^3\mu_{ij}(\beta_s^{\check{\mu}_{ij}})}{\partial\beta_{[k]}\partial\beta\partial\beta^T}(\beta_s^{\tilde{\mu}_{ij}} - \beta_s^*).
\end{aligned}$$

Define $B_s^{\check{\mu}_i} = (\beta_s^{\check{\mu}_{i1}}, \beta_s^{\check{\mu}_{i2}} \dots \beta_s^{\check{\mu}_{im}})$. Then the equation above can be simplified as

$$D_i^{(2)}(\beta_s^*, \widehat{\beta}_s, B_s^{\tilde{\mu}_i}, B_s^{\check{\mu}_i}) = D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^{\tilde{\mu}_i}) - D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^*).$$

Next we estimate the orders of $D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{\mu_i})$ and $D_i^{(2)}(\beta_s^*, \widehat{\beta}_s B_s^{\tilde{\mu}_i}, B_s^{\check{\mu}_i})$. According to Cauchy-Schwarz inequality, we have

$$|D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{\mu_i})_{[jk]}| = |(\widehat{\beta}_s - \beta_s^*)^T \frac{\partial^2\mu_{ij}(\beta_s^{\mu_{ij}})}{\partial\beta_{[k]}\partial\beta}| \leq ||\widehat{\beta}_s - \beta_s^*|| \times ||\frac{\partial^2\mu_{ij}(\beta_s^{\mu_{ij}})}{\partial\beta_{[k]}\partial\beta}||.$$

Here $||\partial^2\mu_{ij}(\beta_s^{\mu_{ij}})/\partial\beta_{[k]}\partial\beta|| = [\sum_{l=1}^{p_n}\{\partial^2\mu_{ij}(\beta_s^{\mu_{ij}})/\partial\beta_{[k]}\partial\beta_{[l]}{}^2\}]^{1/2} = O_p(p_n^{1/2})$. Thus

$D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, \beta_s^{\mu_i})_{[jk]} = O_p(p_n^{1/2}||\beta_s^* - \widehat{\beta}_s||)$, for all i, j, and s. Similarly we have

$$|D_i^{(2)}(\beta_s^*, \widehat{\beta}_s, B_s^{\widetilde{\mu}_i}, B_s^{\check{\mu}_i})_{[jk]}| = |(\beta_s^* - \widehat{\beta}_s)^T \frac{\partial^3 \mu_{ij}(\beta_s^{\check{\mu}_{ij}})}{\partial \beta_{[k]} \partial \beta \partial \beta^T}(B_s^{\widetilde{\mu}_{ij}} - \beta_s^*)|$$

$$\leq p_n||\beta_s^* - \widehat{\beta}_s|| \times ||B_s^{\widetilde{\mu}_{ij}} - \beta_s^*|| \times ||\frac{\partial^3 \mu_{ij}(\beta_s^{\check{\mu}_{ij}})}{\partial \beta_{[k]} \partial \beta \partial \beta^T}||_{\max}$$

$$= O_p(p_n^3 \log p_n / n).$$

■

Lemma 4.10 *Let β_s, $\dot{\beta}_s$, $\widetilde{\beta}_s$, $\check{\beta}_s$ and every column of B_i be a $d_s \times 1$ vector that falls within a $(p_n^2 \log p_n/n)^{1/2}$ neighborhood of β_T^*, $i = 1, 2 \ldots n$. Under Assumptions 4.1 - 4.4, for any unit vector $||v||^2 = 1$ and any model $s \in S_+$ we have the following bounds:*

$$\max_{||v||^2=1} v^T \{\frac{1}{n} \sum_{i=1}^n D_i(\beta_s)^T D_i(\beta_s)\} v = O(1),$$

$$\max_{||v||^2=1} v^T \{\frac{1}{n} \sum_{i=1}^n D_i^{(1)}(\beta_s, \check{\beta}_s, B_i)^T D_i^{(1)}(\beta_s, \check{\beta}_s, B_i)\} v = O_p(p_n^3 \log p_n / n),$$

$$\max_{||v||^2=1} v^T \{\frac{1}{n} \sum_{i=1}^n D_i(\dot{\beta}_s)^T V_i(\widetilde{\beta}_s)^{-1} D_i^{(1)}(\beta_s, \check{\beta}_s, B_i)\} v = O_p\{(p_n^3 \log p_n / n)^{1/2}\}.$$

Proof of Lemma 4.10. First we have the following bound:

$$\max_{||v||^2=1} |v^T \frac{1}{n} \sum_{i=1}^n D_i(\beta_s)^T D_i(\beta_s) v| = \max_{||v||^2=1} |v^T \frac{1}{n} \sum_{i=1}^n X_i^T \Lambda_i(\beta_s)^2 X_i v|$$

$$\leq \max_i \lambda_{\max}\{\Lambda_i(\beta_s)^2\} \lambda_{\max}\{\frac{1}{n} \sum_{i=1}^n X_i^T X_i\} ||v||^2$$

$$\leq \max_{i,j} \{\Lambda_{ij}(\beta_s)^2\} \lambda_{\max}\{\frac{1}{n} \sum_{i=1}^n X_i^T X_i\} ||v||^2$$

$$= O(1).$$

As $\mu_{ij}(\beta_s)$ is differentiable to the third order, we rewrite $\partial^2\mu_{ij}(\beta_s)/\partial\beta_s\partial\beta_s^T = \{\partial\Lambda_{ij}(\beta_s)/\partial\beta_s\}X_{ij}^T$. Here $\{\partial\Lambda_{ij}(\beta_s)/\partial\beta_s\}$ is a $d_s \times 1$ column vector and X_{ij}^T is a $1 \times d_s$ row vector. We have $D_i^{(1)}(\beta_s, \check{\beta}_s, B)_{[j,]} = (\beta_s - \check{\beta}_s)^T\{\partial^2\mu_{ij}(B_{[,j]})/\partial\beta_s\partial\beta_s^T\} = (\beta_s - \check{\beta}_s)^T\{\partial\Lambda_{ij}(B_{[,j]})/\partial\beta_s\}X_{ij}^T$. Therefore we have

$$D_i^{(1)}(\beta_s, \check{\beta}_s, B) = \begin{bmatrix} (\beta_s - \check{\beta}_s)^T\{\partial\Lambda_{i1}(B_{[,1]})/\partial\beta_s\}X_{i1}^T \\ (\beta_s - \check{\beta}_s)^T\{\partial\Lambda_{i2}(B_{[,2]})/\partial\beta_s\}X_{i2}^T \\ \cdots \\ (\beta_s - \check{\beta}_s)^T\{\partial\Lambda_{im}(B_{[,m]})/\partial\beta_s\}X_{im}^T \end{bmatrix}.$$

Let $diag_{j=1}^m\{(\beta_s - \check{\beta}_s)^T[\partial\{\Lambda_{ij}(\beta_s)\}/\partial\beta_s]\}$ represent a diagonal matrix with the jth diagonal entry equal to $(\beta_s - \check{\beta}_s)^T[\partial\{\Lambda_{ij}(\beta_s)\}/\partial\beta_s]$. Then we can reformat $D_i^{(1)}(\beta_s, \check{\beta}_s, B)$ $= diag_{j=1}^m\{(\beta_s - \check{\beta}_s)^T[\partial\{\Lambda_{ij}(\beta_s)\}/\partial\beta_s]\}X_i$. From Assumption 4.2, we have the boundedness of $\lambda_{\max}\{n^{-1}\sum_{i=1}^n X_i^T X_i\}$. This entails

$$\max_{||v||^2=1} |v^T \frac{1}{n} \sum_{i=1}^{n} D_i^{(1)}(\beta_s, \check{\beta}_s, B_i)^T D_i^{(1)}(\beta_s, \check{\beta}_s, B_i) v|$$

$$= \max_{||v||^2=1} |v^T \frac{1}{n} \sum_{i=1}^{n} X_i^T diag_{j=1}^{m} \{(\beta_s - \check{\beta}_s)^T \frac{\partial \Lambda_{ij}(B_{i[,j]})}{\partial \beta_s}\}^2 X_i v|$$

$$\leq \max_i \lambda_{\max} \{diag_{j=1}^{m} \{(\beta_s - \check{\beta}_s)^T \frac{\partial \Lambda_{ij}(B_{i[,j]})}{\partial \beta_s}\}^2\} \lambda_{\max}(\frac{1}{n} \sum_{i=1}^{n} X_i^T X_i) ||v||^2$$

$$\leq \max_{i,j} \{(\beta_s - \check{\beta}_s)^T \frac{\partial \Lambda_{ij}(B_{i[,j]})}{\partial \beta_s}\}^2 \lambda_{\max}(\frac{1}{n} \sum_{i=1}^{n} X_i^T X_i)$$

$$\leq ||\beta_s - \check{\beta}_s||^2 \times \max_{i,j} \{||\frac{\partial \Lambda_{ij}(B_{i[,j]})}{\partial \beta_s}||^2\} \times \lambda_{\max}(\frac{1}{n} \sum_{i=1}^{n} X_i^T X_i)$$

$$\leq ||\beta_s - \check{\beta}_s||^2 \times p_n \times \max_{i,j,k} [\{\frac{\partial \Lambda_{ij}(B_{i[,j]})}{\partial \beta_{s_{[k]}}}\}^2] \times \lambda_{\max}(\frac{1}{n} \sum_{i=1}^{n} X_i^T X_i)$$

$$= O_p(p_n^3 \log p_n / n);$$

$$\max_{||v||^2=1} |v^T \frac{1}{n} \sum_{i=1}^{n} D_i(\dot{\beta}_s)^T V_i(\widetilde{\beta}_s)^{-1} D_i^{(1)}(\beta_s, \check{\beta}_s, B_i) v|$$

$$= \max_{||v||^2=1} |v^T \frac{1}{n} \sum_{i=1}^{n} X_i^T \Lambda_i(\dot{\beta}_s) V_i(\widetilde{\beta}_s)^{-1} diag_{j=1}^{m} \{(\beta_s - \check{\beta}_s)^T \frac{\partial \Lambda_{ij}(B_{i[,j]})}{\partial \beta_s}\} X_i v|$$

$$\leq \max_{i,j} \{\Lambda_{ij}(\dot{\beta}_s)\} \times \max_i \lambda_{\max} \{V_i(\widetilde{\beta}_s)^{-1}\} \times \max_{i,j} \{(\beta_s - \check{\beta}_s)^T \frac{\partial \Lambda_{ij}(B_{i[,j]})}{\partial \beta_s}\}$$

$$\times \lambda_{\max}(\frac{1}{n} \sum_{i=1}^{n} X_i^T X_i) \times ||v||^2$$

$$\leq ||\beta_s - \check{\beta}_s|| \times \max_{i,j} ||\frac{\partial \Lambda_{ij}(B_{i[,j]})}{\partial \beta_s}|| \times O(1)$$

$$\leq ||\beta_s - \check{\beta}_s|| \times (p_n^{1/2}) \times \max_{i,j,k} ||\frac{\partial \Lambda_{ij}(B_{i[,j]})}{\partial \beta_{s_{[k]}}}|| \times O(1)$$

$$= O_p\{(p_n^3 \log p_n / n)^{1/2}\}.$$

■

Lemma 4.11 *Under Assumption 4.1 - 4.4, the estimated inverse working covariance matrices can be decomposed into the sum of serveral matrices of the same dimension* $V_i(\widehat{\beta}_s)^{-1} = V_i^{-1}(\beta_s^*)+V_i^{(1)}(\widehat{\beta}_s, \beta_s^*)+V_i^{(2)}(\widehat{\beta}_s, \beta_s^*, B_s^{A_i})$, *where* $B_s^{A_i} = (\beta_s^{A_{i1}}, \beta_s^{A_{i2}} \dots \beta_s^{A_{im}})$, *and each* $\beta_s^{A_{ij}}$, $j = 1, 2 \dots m$, *is a vector between* $\widehat{\beta}_s$ *and* β_s^*. *Let* $\check{V}_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{A_i}) = V_i^{(1)}(\widehat{\beta}_s, \beta_s^*)+V_i^{(2)}(\widehat{\beta}_s, \beta_s^*, B_s^{A_i})$. *The bounds* $||V_i^{(1)}(\widehat{\beta}_s, \beta_s^*)||_{\max} = O_p\{(p_n^3 \log p_n/n)^{1/2}\}$, $||\check{V}_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{A_i})||_{\max} = O_p\{(p_n^3 \log p_n/n)^{1/2}\}$, $||V_i^{(2)}\ (\widehat{\beta}_s, \beta_s^*, B_s^{A_i})||_{\max} = O_p\{p_n^3 \log p_n/n\}$ *are uniformly held for all model* $s \in S$, *and* $i = 1, 2 \dots n$.

Proof of Lemma 4.11. According to Taylor expansion, there exists a $\beta_s^{A_{ij}}$ between $\widehat{\beta}_s$ and β_s^* such that

$$A_{ij}^{-1/2}(\widehat{\beta}_s) = A_{ij}^{-1/2}(\beta_s^*) + (\widehat{\beta}_s - \beta_s^*)^T \frac{\partial A_{ij}^{-1/2}(\beta_s^*)}{\partial \beta} + \frac{1}{2}(\widehat{\beta}_s - \beta_s^*)^T \frac{\partial^2 A_{ij}^{-1/2}(\beta_s^{A_{ij}})}{\partial \beta \partial \beta^T}(\widehat{\beta}_s - \beta_s^*).$$

For the jth row and hth column of matrix $V_i^{-1}(\widehat{\beta}_s) - V_i^{-1}(\beta_s^*)$, we apply the formula above and obtain

$$\begin{aligned}
&[V_i^{-1}(\widehat{\beta}_s) - V_i^{-1}(\beta_s^*)]_{[jh]} = A_{ij}^{-1/2}(\widehat{\beta}_s)[R^{-1}]_{[jh]}A_{ih}^{-1/2}(\widehat{\beta}_s) - A_{ij}^{-1/2}(\beta_s^*)[R^{-1}]_{[jh]}A_{ih}^{-1/2}(\beta_s^*) \\
&= (\widehat{\beta}_s - \beta_s^*)^T[R^{-1}]_{[jh]}\{A_{ih}^{-1/2}(\beta_s^*)\frac{\partial A_{ij}^{-1/2}(\beta_s^*)}{\partial \beta} + A_{ij}^{-1/2}(\beta_s^*)\frac{\partial A_{ih}^{-1/2}(\beta_s^*)}{\partial \beta}\} \\
&+ (\widehat{\beta}_s - \beta_s^*)^T[R^{-1}]_{[jh]}\{\frac{1}{2}A_{ih}^{-1/2}(\beta_s^*)\frac{\partial^2 A_{ij}^{-1/2}(\beta_s^{A_{ij}})}{\partial \beta \partial \beta^T} + \frac{1}{2}A_{ij}^{-1/2}(\beta_s^*)\frac{\partial^2 A_{ih}^{-1/2}(\beta_s^{A_{ih}})}{\partial \beta \partial \beta^T} \\
&+ \frac{\partial A_{ij}^{-1/2}(\beta_s^*)}{\partial \beta}\frac{\partial A_{ih}^{-1/2}(\beta_s^*)}{\partial \beta^T}\}(\widehat{\beta}_s - \beta_s^*) + O_p(||\widehat{\beta}_s - \beta_s^*||^3).
\end{aligned}$$

Denote the first term in the expansion as $V_i^{(1)}(\widehat{\beta}_s, \beta_s^*)_{[jh]}$ and the remaining three terams as $V_i^{(2)}(\widehat{\beta}_s, \beta_s^*, B_s^{A_i})_{[jh]}$. Based on the Cauchy-Schwarz inequality, the bounds determined in Lemma 4.8 and Assumption 4.2, we have

$$
\begin{aligned}
&|V_i^{(1)}(\widehat{\beta}_s, \beta_s^*)_{[jh]}| \\
&\leq ||\widehat{\beta}_s - \beta_s^*|| \times ||[R^{-1}]_{[jh]}\{A_{ih}^{-1/2}(\beta_s^*)\frac{\partial A_{ij}^{-1/2}(\beta_s^*)}{\partial \beta} + A_{ij}^{-1/2}(\beta_s^*)\frac{\partial A_{ih}^{-1/2}(\beta_s^*)}{\partial \beta}\}|| \\
&\leq p_n^{1/2}||\widehat{\beta}_s - \beta_s^*|| \times \max_l |[R^{-1}]_{[jh]}\{A_{ih}^{-1/2}(\beta_s^*)\frac{\partial A_{ij}^{-1/2}(\beta_s^*)}{\partial \beta_{[l]}} + A_{ij}^{-1/2}(\beta_s^*)\frac{\partial A_{ih}^{-1/2}(\beta_s^*)}{\partial \beta_{[l]}}\}| \\
&= O_p(p_n^{1/2}||\widehat{\beta}_s - \beta_s^*||) \\
&= O_p\{(p_n^3 \log p_n/n)^{1/2}\};
\end{aligned}
$$

and

$$
\begin{aligned}
&|V_i^{(2)}(\widehat{\beta}_s, \beta_s^*, B_s^{A_i})_{[jh]}| \leq ||\widehat{\beta}_s - \beta_s^*|| \times p_n^{1/2} \max_l ||[R^{-1}]_{[jh]}\{\frac{1}{2}A_{ij}^{-1/2}(\beta_s^*)\frac{\partial^2 A_{ih}^{-1/2}(\beta_s^{A_{ih}})}{\partial \beta_{[l]}\partial \beta^T} \\
&+ \frac{1}{2}A_{ih}^{-1/2}(\beta_s^*)\frac{\partial^2 A_{ij}^{-1/2}(\beta_s^{A_{ij}})}{\partial \beta_{[l]}\partial \beta^T} + \frac{\partial A_{ij}^{-1/2}(\beta_s^*)}{\partial \beta_{[l]}}\frac{\partial A_{ih}^{-1/2}(\beta_s^*)}{\partial \beta^T}\}(\widehat{\beta}_s - \beta_s^*)|| + O_p(||\widehat{\beta}_s - \beta_s^*||^3) \\
&\leq ||\widehat{\beta}_s - \beta_s^*||^2 \times p_n \max_l \max_r |[R^{-1}]_{[jh]}\{\frac{1}{2}A_{ij}^{-1/2}(\beta_s^*)\frac{\partial^2 A_{ih}^{-1/2}(\beta_s^{A_{ih}})}{\partial \beta_{[l]}\partial \beta_{[r]}} \\
&+ \frac{1}{2}A_{ih}^{-1/2}(\beta_s^*)\frac{\partial^2 A_{ij}^{-1/2}(\beta_s^{A_{ij}})}{\partial \beta_{[l]}\partial \beta_{[r]}} + \frac{\partial A_{ij}^{-1/2}(\beta_s^*)}{\partial \beta_{[l]}}\frac{\partial A_{ih}^{-1/2}(\beta_s^*)}{\partial \beta_{[r]}}\}| + O_p(||\widehat{\beta}_s - \beta_s^*||^3) \\
&= O_p(||\widehat{\beta}_s - \beta_s^*||^2 \times p_n) = O_p(p_n^3 \log p_n/n).
\end{aligned}
$$

■

Lemma 4.12 *Under Assumptions 4.1 - 4.4, for all models $s \in S$,*

$$\max_{j=1}^{m} \frac{1}{n}\sum_{i=1}^{n} |Y_{ij} - \mu_{ij}(\beta_s^*)| = O_p(1).$$

Proof of Lemma 4.12. The proof is the same as Lemma 3.9 given the same assumptions. So we do not repeat here. ■

Lemma 4.13 *Under Assumptions 4.1 - 4.4, for the true and overfitting models,*

$$\sum_{i=1}^{n}\{\mu_i(\beta_s^*) - \mu_i(\widehat{\beta}_s)\}^T \widehat{V}_i^{-1}\{Y_i - \mu_i(\widehat{\beta}_s)\} = n||\beta_s^* - \widehat{\beta}_s||^2 o_p(1),$$

where the $o_p(1)$ *term holds for all models* $s \in S_+$.

Proof of Lemma 4.13.

From Equation (4.7) of Lemma 4.9, we have

$$
\begin{aligned}
&\sum_{i=1}^{n}\{\mu_i(\beta_s^*)-\mu_i(\widehat{\beta}_s)\}^T\widehat{V}_i^{-1}\{Y_i-\mu_i(\widehat{\beta}_s)\}\\
&=\sum_{i=1}^{n}\{D_i(\widehat{\beta}_s)(\beta_s^*-\widehat{\beta}_s)+\frac{1}{2}D_i^{(1)}(\beta_s^*,\widehat{\beta}_s,B_s^{\widetilde{\mu}_i})(\beta_s^*-\widehat{\beta}_s)\}^T\widehat{V}_i^{-1}\{Y_i-\mu_i(\widehat{\beta}_s)\}\\
&=(\beta_s^*-\widehat{\beta}_s)^T\sum_{i=1}^{n}D_i(\widehat{\beta}_s)^T\{\widehat{V}_i^{-1}-V_i(\widehat{\beta}_s)^{-1}+V_i(\widehat{\beta}_s)^{-1}\}\{Y_i-\mu_i(\widehat{\beta}_s)\}\\
&+\frac{1}{2}(\beta_s^*-\widehat{\beta}_s)^T\sum_{i=1}^{n}D_i^{(1)}(\beta_s^*,\widehat{\beta}_s,B_s^{\widetilde{\mu}_i})^T\widehat{V}_i^{-1}\{Y_i-\mu_i(\widehat{\beta}_s)\}\\
&=(\beta_s^*-\widehat{\beta}_s)^T\sum_{i=1}^{n}D_i(\widehat{\beta}_s)^T\{\widehat{V}_i^{-1}-V_i(\widehat{\beta}_s)^{-1}\}\{Y_i-\mu_i(\widehat{\beta}_s)\}\\
&+(\beta_s^*-\widehat{\beta}_s)^TU(\widehat{\beta}_s)\\
&+\frac{1}{2}(\beta_s^*-\widehat{\beta}_s)^T\sum_{i=1}^{n}D_i^{(1)}(\beta_s^*,\widehat{\beta}_s,B_s^{\widetilde{\mu}_i})^T\widehat{V}_i^{-1}\{Y_i-\mu_i(\widehat{\beta}_s)\}\\
&=(\beta_s^*-\widehat{\beta}_s)^T\sum_{i=1}^{n}D_i(\widehat{\beta}_s)^T\{\widehat{V}_i^{-1}-V_i(\widehat{\beta}_s)^{-1}\}\{Y_i-\mu_i(\widehat{\beta}_s)\}\\
&+\frac{1}{2}(\beta_s^*-\widehat{\beta}_s)^T\sum_{i=1}^{n}D_i^{(1)}(\beta_s^*,\widehat{\beta}_s,B_s^{\widetilde{\mu}_i})^T\widehat{V}_i^{-1}\{Y_i-\mu_i(\widehat{\beta}_s)\}\\
&=Res_1+Res_2.
\end{aligned}
$$

We expand the residual terms as follows.

$$
\begin{aligned}
Res_1 &= (\beta_s^* - \widehat{\beta}_s)^T \sum_{i=1}^{n} D_i(\widehat{\beta}_s)^T \{\widehat{V}_i^{-1} - V_i(\widehat{\beta}_s)^{-1}\}\{Y_i - \mu_i(\widehat{\beta}_s)\} \\
&= (\beta_s^* - \widehat{\beta}_s)^T \sum_{i=1}^{n} D_i(\widehat{\beta}_s)^T \{\widehat{V}_i^{-1} - V_i(\widehat{\beta}_s)^{-1}\}\{Y_i - \mu_i(\beta_s^*)\} \\
&+ (\beta_s^* - \widehat{\beta}_s)^T \sum_{i=1}^{n} D_i(\widehat{\beta}_s)^T \{\widehat{V}_i^{-1} - V_i^{-1}(\widehat{\beta}_s)\}\{\mu_i(\beta_s^*) - \mu_i(\widehat{\beta}_s)\} \\
&= (\beta_s^* - \widehat{\beta}_s)^T \sum_{i=1}^{n} D_i(\widehat{\beta}_s)^T \{\widehat{V}_i^{-1} - (V_i^*)^{-1} + (V_i^*)^{-1} - V_i(\widehat{\beta}_s)^{-1}\}\{Y_i - \mu_i(\beta_s^*)\} + Res_{11} \\
&= Res_{11} + (\beta_s^* - \widehat{\beta}_s)^T \sum_{i=1}^{n} D_i(\widehat{\beta}_s)^T \{\widehat{V}_i^{-1} - (V_i^*)^{-1}\}\{Y_i - \mu_i(\beta_s^*)\} \\
&- (\beta_s^* - \widehat{\beta}_s)^T \sum_{i=1}^{n} D_i(\widehat{\beta}_s)^T \{V_i(\widehat{\beta}_s)^{-1} - (V_i^*)^{-1}\}\{Y_i - \mu_i(\beta_s^*)\} \\
&= Res_{11} + Res_{12} - Res_{13}.
\end{aligned}
$$

The first term can be further decomposed:

$$
\begin{aligned}
Res_{11} &= (\beta_s^* - \widehat{\beta}_s)^T \sum_{i=1}^{n} D_i(\widehat{\beta}_s)^T \{\widehat{V}_i^{-1} - V_i(\widehat{\beta}_s)^{-1}\}\{\mu_i(\beta_s^*) - \mu_i(\widehat{\beta}_s)\} \\
&= (\beta_s^* - \widehat{\beta}_s)^T \sum_{i=1}^{n} D_i(\widehat{\beta}_s)^T \{\widehat{V}_i^{-1} - V_i(\widehat{\beta}_s)^{-1}\}\{D_i(\widehat{\beta}_s) + \frac{1}{2} D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^{\tilde{\mu}_i})\}(\beta_s^* - \widehat{\beta}_s) \\
&= (\beta_s^* - \widehat{\beta}_s)^T \sum_{i=1}^{n} D_i(\widehat{\beta}_s)^T \{\widehat{V}_i^{-1} - V_i(\widehat{\beta}_s)^{-1}\} D_i(\widehat{\beta}_s)(\beta_s^* - \widehat{\beta}_s) \\
&+ \frac{1}{2}(\beta_s^* - \widehat{\beta}_s)^T \sum_{i=1}^{n} D_i(\widehat{\beta}_s)^T \widehat{V}_i^{-1} D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^{\tilde{\mu}_i})(\beta_s^* - \widehat{\beta}_s) \\
&- \frac{1}{2}(\beta_s^* - \widehat{\beta}_s)^T \sum_{i=1}^{n} D_i(\widehat{\beta}_s)^T V_i(\widehat{\beta}_s)^{-1} D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^{\tilde{\mu}_i})(\beta_s^* - \widehat{\beta}_s) \\
&= Res_{111} + Res_{112} + Res_{113}.
\end{aligned}
$$

We obtain the bounds for each of the residual terms:

$$|Res_{111}| = |(\beta_s^* - \widehat{\beta}_s)^T \sum_{i=1}^{n} D_i(\widehat{\beta}_s)^T \{\widehat{V}_i^{-1} - V_i(\widehat{\beta}_s)^{-1}\} D_i(\widehat{\beta}_s)(\beta_s^* - \widehat{\beta}_s)|$$

$$\leq n \max\{|\lambda_{\max}\{\widehat{V}_i^{-1} - V_i(\widehat{\beta}_s)^{-1}\}|, |\lambda_{\min}\{\widehat{V}_i^{-1} - V_i(\widehat{\beta}_s)^{-1}\}|\}$$

$$\times (\beta_s^* - \widehat{\beta}_s)^T \frac{1}{n} \sum_{i=1}^{n} D_i(\widehat{\beta}_s)^T D_i(\widehat{\beta}_s)(\beta_s^* - \widehat{\beta}_s)$$

$$\leq n||\beta_s^* - \widehat{\beta}_s||^2 \times \max\{|\lambda_{\max}\{\widehat{V}_i^{-1} - V_i(\widehat{\beta}_s)^{-1}\}|, |\lambda_{\min}\{\widehat{V}_i^{-1} - V_i(\widehat{\beta}_s)^{-1}\}|\}$$

$$\times \lambda_{\max}\{\frac{1}{n} \sum_{i=1}^{n} D_i(\widehat{\beta}_s)^T D_i(\widehat{\beta}_s)\}$$

$$= n||\beta_s^* - \widehat{\beta}_s||^2 O_p\{(p_n^3 \log p_n/n)^{1/2}\};$$

$$|2Res_{112}| = |(\beta_s^* - \widehat{\beta}_s)^T \sum_{i=1}^{n} D_i(\widehat{\beta}_s)^T \widehat{V}_i^{-1} D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^{\tilde{\mu}_i})(\beta_s^* - \widehat{\beta}_s)|$$

$$= |n(\beta_s^* - \widehat{\beta}_s)^T \frac{1}{n} \sum_{i=1}^{n} D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^{\tilde{\mu}_i})^T \widehat{V}_i^{-1} D_i(\widehat{\beta}_s)(\beta_s^* - \widehat{\beta}_s)|$$

$$\leq n||\beta_s^* - \widehat{\beta}_s||^2 \max_{||v||^2=1} \{v^T \frac{1}{n} \sum_{i=1}^{n} D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^{\tilde{\mu}_i})^T \widehat{V}_i^{-1} D_i(\widehat{\beta}_s) v\}$$

$$= n||\beta_s^* - \widehat{\beta}_s||^2 O_p\{(p_n^3 \log p_n/n)^{1/2}\}.$$

Following the same technique on Res_{112}, we obtain $|Res_{113}| = n||\beta_s^* - \widehat{\beta}_s||^2$

$O_p\{(p_n^3 \log p_n/n)^{1/2}\}$. Applying Lemma 4.9 and 4.11 to Res_{12}, we have

$$
\begin{aligned}
Res_{12} &= (\beta_s^* - \widehat{\beta}_s)^T \sum_{i=1}^{n} D_i(\widehat{\beta}_s)^T (\widehat{V}_i^{-1} - (V_i^*)^{-1})\{Y_i - \mu_i(\beta_s^*)\} \\
&= (\beta_s^* - \widehat{\beta}_s)^T \sum_{i=1}^{n} \{D_i(\beta_s^*)^T + D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{D_i})\}\{V_i^{(1)}(\widehat{\beta}_F, \beta_F^*) \\
&+ V_i^{(2)}(\widehat{\beta}_F, \beta_F^*, B_F^{A_i})\}\{Y_i - \mu_i(\beta_s^*)\} \\
&= (\beta_s^* - \widehat{\beta}_s)^T \sum_{i=1}^{n} \{D_i(\beta_s^*)^T V_i^{(1)}(\widehat{\beta}_F, \beta_F^*) + D_i(\beta_s^*)^T V_i^{(2)}(\widehat{\beta}_F, \beta_F^*, B_F^{A_i}) \\
&+ D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{D_i})^T V_i^{(1)}(\widehat{\beta}_F, \beta_F^*) + D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{D_i})^T V_i^{(2)}(\widehat{\beta}_F, \beta_F^*, B_F^{A_i})\}\{Y_i - \mu_i(\beta_s^*)\} \\
&= Res_{121} + Res_{122} + Res_{123} + Res_{124}.
\end{aligned}
$$

For Res_{121}, define the $d_s \times 1$ vector $\Gamma = \sum_{i=1}^{n} D_i(\beta_s^*)^T V_i^{(1)}(\widehat{\beta}_F, \beta_F^*)\{Y_i - \mu_i(\beta_s^*)\}$. Res_{121} can be reformulated as $(\widehat{\beta}_s - \beta_s^*)^T\Gamma$. The kth element of Γ is denoted as Γ_k. The kth row of $D_i(\beta_s^*)^T$ is denoted as $[D_i(\beta_s^*)^T]_{[k,]}$.

$$
\begin{aligned}
\Gamma_k &= \sum_{i=1}^{n} [D_i(\beta_s^*)^T]_{[k,]} V_i^{(1)}(\widehat{\beta}_F, \beta_F^*)\{Y_i - \mu_i(\beta_s^*)\} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{\bar{j}=1}^{m} [D_i(\beta_s^*)^T]_{[kj]} V_i^{(1)}(\widehat{\beta}_F, \beta_F^*)_{[j\bar{j}]}\{Y_{i\bar{j}} - \mu_{i\bar{j}}(\beta_s^*)\} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{\bar{j}=1}^{m} D_i(\beta_s^*)_{[jk]}(\widehat{\beta}_F - \beta_F^*)^T [R^{-1}]_{[j\bar{j}]}\{A_{i\bar{j}}^{-1/2}(\beta_F^*)\frac{\partial A_{ij}^{-1/2}(\beta_F^*)}{\partial \beta} \\
&+ A_{ij}^{-1/2}(\beta_F^*)\frac{\partial A_{i\bar{j}}^{-1/2}(\beta_F^*)}{\partial \beta}\}\{Y_{i\bar{j}} - \mu_{i\bar{j}}(\beta_s^*)\} \\
&= (\widehat{\beta}_F - \beta_F^*)^T \Pi_k(\beta_s^*).
\end{aligned}
$$

Note that for overfitting model, $\mu_i(\beta_F^*) = \mu_i(\beta_s^*)$. Here $\Pi_k(\beta_s^*) = \sum_{i=1}^n \sum_{j=1}^m \sum_{\bar{j}=1}^m$ $D_i(\beta_s^*)_{[jk]}[R^{-1}]_{[j\bar{j}]}[A_{i\bar{j}}^{-1/2}(\beta_s^*)\{\partial A_{ij}^{-1/2}(\beta_s^*)/\partial\beta\} + A_{ij}^{-1/2}(\beta_s^*)\{\partial A_{i\bar{j}}^{-1/2}(\beta_s^*)/\partial\beta\}]\{Y_{i\bar{j}} - \mu_{i\bar{j}}(\beta_s^*)\}$ represents a $d_s \times 1$ vector. The rth element of $\Pi_k(\beta_s^*)$ is denoted as $\Pi_{kr}(\beta_s^*)$. Then we have

$$\Pi_{kr}(\beta_s^*) = \sum_{i=1}^n \sum_{j=1}^m \sum_{\bar{j}=1}^m D_i(\beta_s^*)_{[jk]}[R^{-1}]_{[j\bar{j}]}\{A_{i\bar{j}}^{-1/2}(\beta_s^*)\frac{\partial A_{ij}^{-1/2}(\beta_s^*)}{\partial\beta_{[r]}}$$
$$+ A_{ij}^{-1/2}(\beta_s^*)\frac{\partial A_{i\bar{j}}^{-1/2}(\beta_s^*)}{\partial\beta_{[r]}}\}\{Y_{i\bar{j}} - \mu_{i\bar{j}}(\beta_s^*)\}.$$

Given that $\mathrm{E}[Y_{i\bar{j}}] = \mu_{i\bar{j}}(\beta_s^*)$, then $\mathrm{E}[n^{-1}\Pi_{kr}(\beta_s^*)] = 0$. According to Lemma 4.8, $A_{ih}^{-1/2}(\beta_s^*)$ and its first derivative are uniformly bounded for all i, h and all model s. Therefore there exists a b_Π for all model s such that:

$$Var\{\frac{1}{n}\Pi_{kr}(\beta_s^*)\} = \frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^m \sum_{\bar{j}=1}^m \sum_{h=1}^m \sum_{\bar{h}=1}^m D_i(\beta_s^*)_{[jk]}[R^{-1}]_{[j\bar{j}]}[D_i(\beta_s^*)]_{hk}[R^{-1}]_{h\bar{h}}$$
$$\{A_{i\bar{j}}^{-1/2}(\beta_s^*)\frac{\partial A_{ij}^{-1/2}(\beta_s^*)}{\partial\beta_r} + A_{ij}^{-1/2}(\beta_s^*)\frac{\partial A_{i\bar{j}}^{-1/2}(\beta_s^*)}{\partial\beta_r}\}$$
$$\{A_{i\bar{h}}^{-1/2}(\beta_s^*)\frac{\partial A_{ih}^{-1/2}(\beta_s^*)}{\partial\beta_r} + A_{ih}^{-1/2}(\beta_s^*)\frac{\partial A_{i\bar{h}}^{-1/2}(\beta_s^*)}{\partial\beta_r}\}\mathrm{Cov}(Y_{i\bar{j}}, Y_{i\bar{h}}) \le \frac{b_\Pi}{n}.$$

According to Chebyshev's inequality,

$$\Pr\{|\frac{1}{n}\Pi_{kr}(\beta_s^*)| \ge (\frac{b_\Pi}{n}p_n^2\log p_n)^{1/2}\} \le \frac{1}{p_n^2\log p_n}.$$

When $p_n \to \infty$, according to Bonferroni inequality,

$$\Pr\{\max_{k,r}|\frac{1}{n}\Pi_{kr}(\beta_s^*)| \ge (\frac{b_\Pi}{n}p_n^2\log p_n)^{1/2}\} \le \frac{p_n^2}{p_n^2\log p_n} = (\log p_n)^{-1} \to 0,$$

Or equivalently we have

$$\max_{kr} |\Pi_{kr}(\beta_s^*)| = O_p\{(p_n^2 \log p_n)^{1/2}\}.$$

According to Cauchy-Schwarz inequality

$$|Res_{121}| \leq ||\widehat{\beta}_s - \beta_s^*|| \times ||\Gamma|| \leq p_n^{1/2} \times ||\widehat{\beta}_s - \beta_s^*|| \times \max_k |\Gamma_k|,$$

$$|\Gamma_k| \leq ||\widehat{\beta}_s - \beta_s^*|| \times ||\Pi_k|| \leq p_n^{1/2} \times ||\widehat{\beta}_s - \beta_s^*|| \times \max_{kr} |\Pi_r(\beta_s^*)| = ||\widehat{\beta}_s - \beta_s^*|| O_p\{(p_n^3 \log p_n)^{1/2}\}.$$

Therefore we have

$$|Res_{121}| = n||\widehat{\beta}_s - \beta_s^*||^2 \times O_p\{(p_n^4 \log p_n/n)^{1/2}\}.$$

For the term Res_{122}, Lemma 4.11 implies that the largest elements of matrix $||V_i^{(2)}(\widehat{\beta}_F, \beta_F^*, B_F^{A_i})||_{\max}$ is $O_p(p_n^3 \log p_n/n)$. Lemma 4.8 implies that all elements from $D_i(\beta_s^*)$ are bounded. Lemma 4.12 demonstrates that $\sum_{i=1}^n |Y_{ij} - \mu_{ij}(\beta_s^*)|/n$ are

bounded for all $j \in \{1, 2 \dots m\}$. Therefore we have

$$
\begin{aligned}
|Res_{122}| &= |(\widehat{\beta}_s - \beta_s^*)^T \sum_{i=1}^{n} D_i(\beta_s^*)^T V_i^{(2)}(\widehat{\beta}_F, \beta_F^*, B_F^{A_i})\{Y_i - \mu_i(\beta_s^*)\}| \\
&\leq n||\widehat{\beta}_s - \beta_s^*|| \times ||\frac{1}{n}\sum_{i=1}^{n} D_i(\beta_s^*)^T V_i^{(2)}(\widehat{\beta}_F, \beta_F^*, B_F^{A_i})\{Y_i - \mu_i(\beta_s^*)\}|| \\
&\leq np_n^{1/2}||\widehat{\beta}_s - \beta_s^*|| \times \max_k |\frac{1}{n}\sum_{i=1}^{n} [D_i(\beta_s^*)^T]_{[k,]} V_i^{(2)}(\widehat{\beta}_F, \beta_F^*, B_F^{A_i})\{Y_i - \mu_i(\beta_s^*)\}| \\
&\leq np_n^{1/2}||\widehat{\beta}_s - \beta_s^*|| \times m||D_i(\beta_s^*)||_{\max}||V_i^{(2)}(\widehat{\beta}_F, \beta_F^*, B_F^{A_i})||_{\max}\{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}|Y_{ij} - \mu_{ij}(\beta_s^*)|\} \\
&\leq np_n^{1/2}||\widehat{\beta}_s - \beta_s^*|| \times m^2||D_i(\beta_s^*)||_{\max}||V_i^{(2)}(\widehat{\beta}_F, \beta_F^*, B_F^{A_i})||_{\max} \max_j\{\frac{1}{n}\sum_{i=1}^{n}|Y_{ij} - \mu_{ij}(\beta_s^*)|\} \\
&= O(np_n^{1/2})||\widehat{\beta}_s - \beta_s^*|| \times ||V_i^{(2)}(\widehat{\beta}_F, \beta_F^*, B_F^{A_i})||_{\max} \\
&= O(np_n^{1/2}p_n^3 \log p_n/n)||\widehat{\beta}_s - \beta_s^*|| \\
&= n||\widehat{\beta}_s - \beta_s^*||^2 O_p\{(p_n^5 \log p_n/n)^{1/2}\}.
\end{aligned}
$$

Similarly we can estimate the orders of Res_{123} and Res_{124}.

$$
\begin{aligned}
Res_{123} &= (\widehat{\beta}_s - \beta_s^*)^T \sum_{i=1}^{n} D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{D_i})^T V_i^{(1)}(\widehat{\beta}_F, \beta_F^*)\{Y_i - \mu_i(\beta_s^*)\} \\
&\leq ||\widehat{\beta}_s - \beta_s^*|| \times || \sum_{i=1}^{n} D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{D_i})^T V_i^{(1)}(\widehat{\beta}_F, \beta_F^*)\{Y_i - \mu_i(\beta_s^*)\}|| \\
&\leq ||\widehat{\beta}_s - \beta_s^*|| \times p_n^{1/2} \times \max_k | \sum_{i=1}^{n} D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{D_i})_{[k,]}^T V_i^{(1)}(\widehat{\beta}_F, \beta_F^*)\{Y_i - \mu_i(\beta_s^*)\}| \\
&\leq ||\widehat{\beta}_s - \beta_s^*|| \times p_n^{1/2} \times ||D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{D_i})||_{\max} \times ||V_i^{(1)}(\widehat{\beta}_F, \beta_F^*)||_{\max} \\
&\times m^2 \max_j \sum_{i=1}^{n} |Y_{ij} - \mu_{ij}(\beta_s^*)| \\
&= O(np_n^{1/2}) \times ||\widehat{\beta}_s - \beta_s^*|| \times ||D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{D_i})||_{\max} \times O_p\{(p_n^3 \log p_n/n)^{1/2}\} \\
&= n||\widehat{\beta}_s - \beta_s^*||^2 O_p\{(p_n^5 \log p_n/n)^{1/2}\};
\end{aligned}
$$

$$
\begin{aligned}
Res_{124} &= (\widehat{\beta}_s - \beta_s^*)^T \sum_{i=1}^{n} D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{D_i})^T V_i^{(2)}(\widehat{\beta}_F, \beta_F^*, B_F^{A_i})\{Y_i - \mu_i(\beta_s^*)\} \\
&\leq ||\widehat{\beta}_s - \beta_s^*|| p_n^{1/2} ||D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{D_i})||_{\max} \times ||V_i^{(2)}(\widehat{\beta}_F, \beta_F^*, B_F^{A_i})||_{\max} \\
&\times m^2 \max_j \sum_{i=1}^{n} |Y_{ij} - \mu_{ij}(\beta_s^*)| \\
&= O_p\{np_n^{3/2}\}||\widehat{\beta}_s - \beta_s^*||^3 \times ||\widehat{\beta}_F - \beta_F^*|| \\
&= n||\widehat{\beta}_s - \beta_s^*||^2 O_p\{p_n^{3.5} \log p_n/n\}.
\end{aligned}
$$

According to Lemma 4.11, both $||V_i^* - \widehat{V}_i||_{\max}$ and $||V_i^* - V_i(\widehat{\beta}_s)||_{\max}$ have the same order. Similar to $|Res_{12}|$, we have $|Res_{13}| = n||\beta_s^* - \widehat{\beta}_s||^2 o_p(1)$. Next we analyze the other residual terms.

$$2Res_2 = (\beta_s^* - \widehat{\beta}_s)^T \sum_{i=1}^{n} D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^{\widetilde{\mu}_i})^T \widehat{V}_i^{-1}\{Y_i - \mu_i(\widehat{\beta}_s)\}$$

$$= (\beta_s^* - \widehat{\beta}_s)^T \sum_{i=1}^{n} D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^{\widetilde{\mu}_i})^T \widehat{V}_i^{-1}\{Y_i - \mu_i(\beta_s^*) + \mu_i(\beta_s^*) - \mu_i(\widehat{\beta}_s)\}$$

$$= (\beta_s^* - \widehat{\beta}_s)^T \sum_{i=1}^{n} D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^{\widetilde{\mu}_i})^T \widehat{V}_i^{-1}\{Y_i - \mu_i(\beta_s^*)\}$$

$$+ (\beta_s^* - \widehat{\beta}_s)^T \sum_{i=1}^{n} D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^{\widetilde{\mu}_i})^T \widehat{V}_i^{-1}\{\mu_i(\beta_s^*) - \mu_i(\widehat{\beta}_s)\}$$

$$= Res_{21} + Res_{22}.$$

$$Res_{21} = (\beta_s^* - \widehat{\beta}_s)^T \sum_{i=1}^{n} D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^{\widetilde{\mu}_i})^T \widehat{V}_i^{-1}\{Y_i - \mu_i(\beta_s^*)\}$$

$$= (\beta_s^* - \widehat{\beta}_s)^T \sum_{i=1}^{n} \{D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^*) + D_i^{(2)}(\beta_s^*, \widehat{\beta}_s, B_s^{\widetilde{\mu}_i}, B_s^{\check{\mu}_i})\}^T$$

$$\{V_i^{-1}(\beta_F^*) + \check{V}_i^{(1)}(\widehat{\beta}_F, \beta_F^*, B_s^{A_i})\}\{Y_i - \mu_i(\beta_s^*)\}$$

$$= (\beta_s^* - \widehat{\beta}_s)^T \sum_{i=1}^{n} \{D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^*)V_i^{-1}(\beta_F^*) + D_i^{(2)}(\beta_s^*, \widehat{\beta}_s, B_s^{\widetilde{\mu}_i}, B_s^{\check{\mu}_i})V_i^{-1}(\beta_F^*) +$$

$$D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^*)\check{V}_i^{(1)}(\widehat{\beta}_F, \beta_F^*, B_s^{A_i}) + D_i^{(2)}(\beta_s^*, \widehat{\beta}_s, B_s^{\widetilde{\mu}_i}, B_s^{\check{\mu}_i})\check{V}_i^{(1)}(\widehat{\beta}_F, \beta_F^*, B_s^{A_i})\}\{Y_i - \mu_i(\beta_s^*)\}$$

$$= Res_{211} + Res_{212} + Res_{213} + Res_{214}.$$

By similar arguments as above, we are able to show Res_{211}, Res_{212}, Res_{213}, and Res_{214} are all of the order $n||\widehat{\beta}_s - \beta_s^*||^2 o_p(1)$. For Res_{22}, there exists a $\check{\beta}_s$ between β_s^*

and $\widehat{\beta}_s$ such that $\mu_i(\beta_s^*) - \mu_i(\widehat{\beta}_s) = D_i(\check{\beta}_s)(\beta_s^* - \widehat{\beta}_s)$. This entails

$$
\begin{aligned}
|Res_{22}| &= |(\widehat{\beta}_s - \beta_s^*)^T \sum_{i=1}^{n} D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^{\tilde{\mu}_i})^T \widehat{V}_i^{-1}\{\mu_i(\beta_s^*) - \mu_i(\widehat{\beta}_s)\}| \\
&= |(\beta_s^* - \widehat{\beta}_s)^T \sum_{i=1}^{n} D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^{\tilde{\mu}_i})^T \widehat{V}_i^{-1} D_i(\check{\beta}_s)(\beta_s^* - \widehat{\beta}_s)| \\
&= |n(\beta_s^* - \widehat{\beta}_s)^T \frac{1}{n} \sum_{i=1}^{n} D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^{\tilde{\mu}_i})^T \widehat{V}_i^{-1} D_i(\check{\beta}_s)(\beta_s^* - \widehat{\beta}_s)| \\
&\leq n||\beta_s^* - \widehat{\beta}_s||^2 \max_{||v||^2=1} \{v^T \frac{1}{n} \sum_{i=1}^{n} D_i^{(1)}(\beta_s^*, \widehat{\beta}_s, B_s^{\tilde{\mu}_i})^T \widehat{V}_i^{-1} D_i(\check{\beta}_s) v\} \\
&= n||\beta_s^* - \widehat{\beta}_s||^2 O_p\{(p_n^3 \log p_n/n)^{1/2}\}.
\end{aligned}
$$

Combining all the orders for each of the terms, results in the lemma follows. ■

Lemma 4.14 *For $s \in S_+$, let $\eta = n^{-1/2} W(\beta_s^*)^{-1/2} \sum_{i=1}^{n} U_i(\beta_s^*)$. The random vectors $U_1(\beta_s^*), U_2(\beta_s^*) \ldots U_n(\beta_s^*)$ are independently distributed random vectors of dimension d_s with zero mean and satisfy the cumulant boundedness condition. Under Assumptions 4.1 - 4.4, $\log E[e^{t^T \eta}] \leq a^2 t^T t/2$ for $||t||^2 \leq p_n^2 \log p_n$ and some constant $a^2 > 1$.*

This implies that if the cumulant boundedness condition in Definition 1 holds, we will be able to apply large deviation results to the modified residual sum of squares difference type of statistics arising in our analysis.

Proof of Lemma 4.14. The proof is the same as Lemma 3.5 given the same assumptions. The only change is to replace all of s_n to p_n. s_n is the model size up bound and can go as large as p_n. The replacement from s_n to p_n does not change the proofs. So we do not repeat the proof here. ■

Proof of Lemma 4.1. The proof is the same as Lemma 3.1 given the same assumptions. The only change is to replace all of s_n to p_n. s_n is the model size up bound and can go as large as p_n. The replacement from s_n to p_n does not change the proofs. So we do not repeat the proof here. ■

Proof of Lemma 4.2.

We know for true model $V_i(\beta_s^*) = V_i(\beta_F^*)$ for $s \in S_+$. Therefore we know that $||\widehat{V}_i^{-1} - V_i(\beta_s^*)^{-1}||_{\max} = ||V_i(\widehat{\beta}_F)^{-1} - V_i(\beta_s^*)^{-1}||_{\max} = ||V_i(\widehat{\beta}_F)^{-1} - V_i(\beta_F^*)^{-1}||_{\max} = O_p\{(p_n^3 \log p_n/n)^{1/2}\}$, according to Lemma 4.11. In addition $||\widehat{V}_i^{-1} - V_i(\widehat{\beta}_s)^{-1}||_{\max} \leq ||\widehat{V}_i^{-1} - V_i(\beta_s^*)^{-1}||_{\max} + ||V_i(\beta_s^*)^{-1} - V_i(\widehat{\beta}_s)^{-1}||_{\max} = O_p\{(p_n^3 \log p_n/n)^{1/2}\}$. This implies:

$$\max |\lambda_{\max}\{\widehat{V}_i^{-1} - V_i(\widehat{\beta}_s)^{-1}\}, \lambda_{\min}\{\widehat{V}_i^{-1} - V_i(\widehat{\beta}_s)^{-1}\}| = O_p\{(p_n^3 \log p_n/n)^{1/2}\},$$

$$\max |\lambda_{\max}\{\widehat{V}_i^{-1} - V_i(\beta_s^*)^{-1}\}, \lambda_{\min}(\widehat{V}_i^{-1} - V_i(\beta_s^*)^{-1})| = O_p\{(p_n^3 \log p_n/n)^{1/2}\}.$$

■

Proof of Lemma 4.3. The proof is the same as Lemma 3.2 given the same assumptions. The only change is to replace all of s_n to p_n. s_n is the model size up bound and can go as large as p_n. The replacement from s_n to p_n does not change the proofs. So we do not repeat the proof here. ■

Proof of Lemma 4.4. Considering a competing model s.

$$
\begin{aligned}
2Q(\beta_s^*) &= \sum_{i=1}^{n}\{Y_i - \mu_i(\beta_s^*)\}^T\widehat{V}_i^{-1}\{Y_i - \mu_i(\beta_s^*)\} \\
&= \sum_{i=1}^{n}\{Y_i - \mu_i(\widehat{\beta}_s) + \mu_i(\widehat{\beta}_s) - \mu_i(\beta_s^*)\}^T\widehat{V}_i^{-1}\{Y_i - \mu_i(\widehat{\beta}_s) + \mu_i(\widehat{\beta}_s) - \mu_i(\beta_s^*)\} \\
&= 2Q(\widehat{\beta}_s) + \sum_{i=1}^{n}\{\mu_i(\beta_s^*) - \mu_i(\widehat{\beta}_s)\}^T\widehat{V}_i^{-1}\{\mu_i(\beta_s^*) - \mu_i(\widehat{\beta}_s)\} \\
&+ 2\sum_{i=1}^{n}\{\mu_i(\widehat{\beta}_s) - \mu_i(\beta_s^*)\}^T\widehat{V}_i^{-1}\{Y_i - \mu_i(\widehat{\beta}_s)\}.
\end{aligned}
$$

Lemma 4.13 shows that the last term $\sum_{i=1}^{n}\{\mu_i(\widehat{\beta}_s) - \mu_i(\beta_s^*)\}^T\widehat{V}_i^{-1}\{Y_i - \mu_i(\widehat{\beta}_s)\} = n||\beta_s^* - \widehat{\beta}_s||^2 o_p(1)$. We consider the second term. Applying Equation (4.6) from Lemma 4.9 to the second term, we have

$$
\begin{aligned}
&\sum_{i=1}^{n}\{\mu_i(\beta_s^*) - \mu_i(\widehat{\beta}_s)\}^T\widehat{V}_i^{-1}\{\mu_i(\beta_s^*) - \mu_i(\widehat{\beta}_s)\} \\
&= \sum_{i=1}^{n}(\beta_s^* - \widehat{\beta}_s)^T\{D_i(\beta_s^*) + \frac{1}{2}D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{\mu_i})\}^T\widehat{V}_i^{-1}\{D_i(\beta_s^*) + \frac{1}{2}D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{\mu_i})\}(\beta_s^* - \widehat{\beta}_s) \\
&= (\beta_s^* - \widehat{\beta}_s)^T\{\sum_{i=1}^{n} D_i(\beta_s^*)^T\widehat{V}_i^{-1}D_i(\beta_s^*) + D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{\mu_i})^T\widehat{V}_i^{-1}D_i(\beta_s^*) \\
&+ \frac{1}{4}D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{\mu_i})^T\widehat{V}_i^{-1}D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{\mu_i})\}(\beta_s^* - \widehat{\beta}_s) \\
&= n(\beta_s^* - \widehat{\beta}_s)^T[\Omega(\beta_s^*) + \frac{1}{n}\sum_{i=1}^{n} D_i(\beta_s^*)^T\{\widehat{V}_i^{-1} - V_i(\beta_s^*)\}D_i(\beta_s^*) \\
&+ D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{\mu_i})^T\widehat{V}_i^{-1}D_i(\beta_s^*) + \frac{1}{4}D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{\mu_i})^T\widehat{V}_i^{-1}D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{\mu_i})](\beta_s^* - \widehat{\beta}_s) \\
&= n(\beta_s^* - \widehat{\beta}_s)^T\{\Omega(\beta_s^*) + Res_3\}(\beta_s^* - \widehat{\beta}_s).
\end{aligned}
$$

Let $Res_3 = Res_{31} + Res_{32} + Res_{33}$, with $Res_{31} = \sum_{i=1}^{n} D_i(\beta_s^*)^T\{\widehat{V}_i^{-1} - V_i(\beta_s^*)\}D_i(\beta_s^*)/n$, $Res_{32} = \sum_{i=1}^{n} D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{\mu_i})^T\widehat{V}_i^{-1}\ D_i(\beta_s^*)/n$, and $Res_{33} = \sum_{i=1}^{n} D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{\mu_i})^T$

$\widehat{V}_i^{-1} D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{\mu_i})/(4n)$. Let v be a d_s dimensional unit vector with $||v||^2 = 1$. Then

$$|v^T Res_{31} v| = |v^T \frac{1}{n} \sum_{i=1}^{n} D_i(\beta_s^*)^T \{\widehat{V}_i^{-1} - V_i(\beta_s^*)\} D_i(\beta_s^*) v|$$

$$\leq \max[|\lambda_{\max}\{\widehat{V}_i^{-1} - V_i(\beta_s^*)\}|, |\lambda_{\min}\{\widehat{V}_i^{-1} - V_i(\beta_s^*)\}|] \max_{||v||^2=1} \{v^T \frac{1}{n} \sum_{i=1}^{n} D_i(\beta_s^*)^T D_i(\beta_s^*) v\}$$

$$\leq O_p\{(p_n^3 \log p_n/n)^{1/2}\}.$$

From Lemma 4.10, we have $|v^T Res_{32} v| = |v^T (1/n) \sum_{i=1}^{n} D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{\mu_i})^T V_i^{-1} D_i(\beta_s^*) v| = O_p\{(p_n^3 \log p_n/n)^{1/2}\}$. For Res_{33},

$$|v^T Res_{33} v| = |v^T \frac{1}{n} \sum_{i=1}^{n} D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{\mu_i})^T V_i^{-1} D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{\mu_i})^T v|$$

$$\leq \lambda_{\max_i}(V_i^{-1}) \max_{||v||^2=1} \{v^T \frac{1}{n} \sum_{i=1}^{n} D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{\mu_i})^T D_i^{(1)}(\widehat{\beta}_s, \beta_s^*, B_s^{\mu_i}) v\}$$

$$= O_p(p_n^3 \log p_n/n).$$

From Assumption 4.2, the eigenvalues of $\Omega(\beta_s^*)$ are bounded from zero to infinity. We have

$$\sup_{||v||=1} |v^T\{\Omega(\beta_s^*) + Res_3\} v| = \sup_{||v||=1} |v^T \Omega(\beta_s^*) v|(1 + O_p\{(p_n^3 \log p_n/n)^{1/2}\}).$$

Combining the above equation and Lemma 4.13, we have $2[Q(\widehat{\beta}_s) - Q(\beta_s^*)] = -n(\beta_s^* - \widehat{\beta}_s)^T\{\Omega(\beta_s^*) + Res_3\}(\beta_s^* - \widehat{\beta}_s) + n||\beta_s^* - \widehat{\beta}_s||^2 O_p\{(p_n^3 \log p_n/n)^{1/2}\} = -n(\beta_s^* - \widehat{\beta}_s)^T \Omega(\beta_s^*)$ $(\beta_s^* - \widehat{\beta}_s)\{1 + o_p(1)\}$. According to Lemma 4.3, $\widehat{\beta}_s - \beta_s^* = \{\Omega(\beta_s^*) + Res_d\}^{-1} U(\beta_s^*)$.

We rewrite the equation as

$$
\begin{aligned}
&2\{Q(\widehat{\beta}_s) - Q(\beta_s^*)\} = -n(\beta_s^* - \widehat{\beta}_s)^T \Omega(\beta_s^*)(\beta_s^* - \widehat{\beta}_s)\{1 + o_p(1)\} \\
&= -\frac{1}{n} U(\beta_s^*)^T \{\Omega(\beta_s^*) + Res_d^T\}^{-1} \Omega(\beta_s^*) \{\Omega(\beta_s^*) + Res_d\}^{-1} U(\beta_s^*)\{1 + o_p(1)\}. \\
&= -\frac{1}{n} U(\beta_s^*)^T [\{\Omega(\beta_s^*) + Res_d^T\}\Omega(\beta_s^*)^{-1}\{\Omega(\beta_s^*) + Res_d\}]^{-1} U(\beta_s^*)\{1 + o_p(1)\}. \\
&= -\frac{1}{n} U(\beta_s^*)^T \{\Omega(\beta_s^*) + Res_d + Res_d^T + Res_d^T \Omega(\beta_s^*)^{-1} Res_d\}^{-1} U(\beta_s^*)\{1 + o_p(1)\}.
\end{aligned}
$$

Let $Res_s = Res_d + Res_d^T + Res_d^T \Omega(\beta_s^*)^{-1} Res_d$ and we have

$$
2\{Q(\widehat{\beta}_s) - Q(\beta_s^*)\} = -1/n U(\beta_s^*)^T \{\Omega(\beta_s^*) + Res_s\}^{-1} U(\beta_s^*)\{1 + o_p(1)\}.
$$

We estimate the order of the matrix Res_s as follows:

$$
\begin{aligned}
&\sup_{||v||=1} v^T (Res_d + Res_d^T) v \leq 2 \sup_{||v||=1} v^T Res_d v \\
&= \sup_{||v||=1} \sum_{kr} v_k v_r [Res_d]_{[kr]} \\
&\leq \max_{k,r} |[Res_d]_{[kr]}| \sum_{kr} |v_k| \times |v_r| \\
&\leq \max_{k,r} |[Res_d]_{[kr]}| \times d_s \times ||v||^2 \\
&= O_p\{(p_n^5 \log p_n)^{1/2}\};
\end{aligned}
$$

$$\inf_{||v||=1} v^T(Res_d + Res_d^T)v \geq 2 \inf_{||v||=1} v^T Res_d v$$

$$= \inf_{||v||=1} \sum_{kr} v_k v_r [Res_d]_{[kr]}$$

$$\geq -\max_{k,r} |[Res_d]_{[kr]}| \sum_{kr} |v_k| \times |v_r|$$

$$\geq -\max_{k,r} |[Res_d]_{[kr]}| \times d_s \times ||v||^2$$

$$= -O_p\{(p_n^5 \log p_n)^{1/2}\};$$

$$\sup_{||v||=1} v^T(Res_d^T \Omega(\beta_s^*)^{-1} Res_d)v$$

$$\leq \lambda_{\max}\{\Omega(\beta_s^*)^{-1}\} \sup_{||v||=1} v^T(Res_d^T Res_d)v$$

$$= \lambda_{\max}\{\Omega(\beta_s^*)^{-1}\} \sum_{k,r} v_k v_r (Res_d^T Res_d)_{[kr]}$$

$$= \lambda_{\max}\{\Omega(\beta_s^*)^{-1}\} \sum_{k,r} v_k v_r \sum_{l} [Res_d^T]_{[kl]} [Res_d]_{[lr]}$$

$$\leq \lambda_{\max}\{\Omega(\beta_s^*)^{-1}\} d_s \times ||Res_d||_{\max}^2 \times \sum_{k,r} |v_k| \times |v_r|$$

$$\leq \lambda_{\max}\{\Omega(\beta_s^*)^{-1}\} d_s \times ||Res_d||_{\max}^2 \times d_s \times ||v||^2$$

$$= O_p\{p_n^5 \log p_n / n\};$$

$$\inf_{||v||=1} v^T(Res_d^T\Omega(\beta_s^*)^{-1}Res_d)v$$

$$\geq \lambda_{\min}\{\Omega(\beta_s^*)^{-1}\}\inf_{||v||=1} v^T(Res_d^T Res_d)v$$

$$= \lambda_{\min}\{\Omega(\beta_s^*)^{-1}\}\sum_{k,r} v_k v_r (Res_d^T Res_d)_{[kr]}$$

$$= \lambda_{\min}\{\Omega(\beta_s^*)^{-1}\}\sum_{k,r} v_k v_r \sum_l [Res_d^T]_{[kl]}[Res_d]_{[kr]}$$

$$\geq -\lambda_{\min}\{\Omega(\beta_s^*)^{-1}\}d_s \times ||Res_d||^2_{\max} \times \sum_{k,r} |v_k| \times |v_r|$$

$$\geq -\lambda_{\min}\{\Omega(\beta_s^*)^{-1}\}d_s \times ||Res_d||^2_{\max} \times d_s \times ||v||^2$$

$$= -O_p\{p_n^5 \log p_n/n\}.$$

Thus we have $\sup_{||v||=1} |v^T Res_s v| = O_p\{(p_n^5 \log p_n/n)^{1/2}\} = o_p(1)$. This implies that the eigenvalues of Res_s are of the order of $o_p(1)$. It can be shown that

$$\sup_{||v||^2=1} v^T[\Omega(\beta_s^*)^{-1} - \{\Omega(\beta_s^*) + Res_s\}^{-1}]v$$

$$= \sup_{||v||^2=1} v^T\Omega(\beta_s^*)^{-1/2}[I - \{I + \Omega(\beta_s^*)^{-1/2}Res_s\Omega(\beta_s^*)^{-1/2}\}^{-1}]\Omega(\beta_s^*)^{-1/2}v$$

$$\leq \lambda_{\max}\{\Omega(\beta_s^*)^{-1}\}\lambda_{\max}(I - [\{I + \Omega(\beta_s^*)^{-1/2}Res_s\Omega(\beta_s^*)^{-1/2}\}^{-1}])||v||^2$$

$$= \lambda_{\max}\{\Omega(\beta_s^*)^{-1}\}(1 - \lambda_{\min}[\{I + \Omega(\beta_s^*)^{-1/2}Res_s\Omega(\beta_s^*)^{-1/2}\}^{-1}])$$

$$= \lambda_{\max}\{\Omega(\beta_s^*)^{-1}\}[1 - \frac{1}{1 + \lambda_{\max}\{\Omega(\beta_s^*)^{-1/2}Res_s\Omega(\beta_s^*)^{-1/2}\}}]$$

$$= \lambda_{\max}\{\Omega(\beta_s^*)^{-1}\}[\frac{\lambda_{\max}\{\Omega(\beta_s^*)^{-1/2}Res_s\Omega(\beta_s^*)^{-1/2}\}}{1 + \lambda_{\max}\{\Omega(\beta_s^*)^{-1/2}Res_s\Omega(\beta_s^*)^{-1/2}\}}].$$

Furthermore,

$$\begin{aligned}
&\lambda_{\max}\{\Omega(\beta_s^*)^{-1/2} Res_s \Omega(\beta_s^*)^{-1/2}\} \\
&= \sup_{||v||=1} v^T \{\Omega(\beta_s^*)^{-1/2} Res_s \Omega(\beta_s^*)^{-1/2}\} v \\
&\leq \lambda_{\max}\{\Omega(\beta_s^*)^{-1}\} \lambda_{\max}\{Res_s\} ||v||^2 \\
&= o_p(1).
\end{aligned}$$

Thus $\sup_{||v||^2=1} v^T[\Omega(\beta_s^*)^{-1} - \{\Omega(\beta_s^*) + Res_s\}^{-1}]v = o_p(1)$. Therefore,

$$\begin{aligned}
2\{Q(\widehat{\beta}_s) - Q(\beta_s^*)\} &= -1/n U(\beta_s^*)^T \{\Omega(\beta_s^*) + Res_s\}^{-1} U(\beta_s^*)\{1 + o_p(1)\} \\
&= -\frac{1}{n} U(\beta_s^*)^T \Omega(\beta_s^*)^{-1} U(\beta_s^*)\{1 + o_p(1)\}.
\end{aligned}$$

■

Proof of Lemma 4.5. We first consider the true and overfitting situation. By Lemma 4.4 shows that $|Q(\widehat{\beta}_s) - Q(\beta_s^*)| = (n/2)(\beta_s^* - \widehat{\beta}_s)^T \Omega(\beta_s^*)(\beta_s^* - \widehat{\beta}_s)\{1 + o_p(1)\}$. Theorem 4.1 shows that $||\beta_s^* - \widehat{\beta}_s|| = O_p\{(p_n^2 \log p_n / n)^{1/2}\}$. And Assumption 4.2 indicates that all eigenvalue of $\Omega(\beta_s^*)$ is bounded.

$$\begin{aligned}
&|Q(\widehat{\beta}_s) - Q(\beta_s^*)| = \frac{n}{2}(\beta_s^* - \widehat{\beta}_s)^T \Omega(\beta_s^*)(\beta_s^* - \widehat{\beta}_s)\{1 + o_p(1)\} \\
&\leq \frac{n}{2} \lambda_{\max}\{\Omega(\beta_s^*)\} ||\beta_s^* - \widehat{\beta}_s||^2 \{1 + o_p(1)\} \\
&\leq O_p(p_n^2 \log p_n).
\end{aligned}$$

Then we consider the underfitting situation.

$$|2Q(\widehat{\beta}_s) - 2Q(\beta_s^*)| = |\sum_{i=1}^{n}\{Y_i - \mu_i(\widehat{\beta}_s)\}^T\widehat{V}_i^{-1}\{Y_i - \mu_i(\widehat{\beta}_s)\} - 2Q(\beta_s^*)|$$

$$= |\sum_{i=1}^{n}\{Y_i - \mu_i(\beta_s^*) + \mu_i(\beta_s^*) - \mu_i(\widehat{\beta}_s)\}^T\widehat{V}_i^{-1}\{Y_i - \mu_i(\beta_s^*) + \mu_i(\beta_s^*) - \mu_i(\widehat{\beta}_s)\} - 2Q(\beta_s^*)|$$

$$= |\sum_{i=1}^{n}\{\mu_i(\beta_s^*) - \mu_i(\widehat{\beta}_s)\}^T\widehat{V}_i^{-1}\{\mu_i(\beta_s^*) - \mu_i(\widehat{\beta}_s)\} + 2\sum_{i=1}^{n}\{\mu_i(\beta_s^*) - \mu_i(\widehat{\beta}_s)\}^T\widehat{V}_i^{-1}\{Y_i - \mu_i(\beta_s^*)\}|$$

$$\leq |\sum_{i=1}^{n}\{\mu_i(\beta_s^*) - \mu_i(\widehat{\beta}_s)\}^T\widehat{V}_i^{-1}\{\mu_i(\beta_s^*) - \mu_i(\widehat{\beta}_s)\}| + 2|\sum_{i=1}^{n}\{\mu_i(\beta_s^*) - \mu_i(\widehat{\beta}_s)\}^T\widehat{V}_i^{-1}\{Y_i - \mu_i(\beta_s^*)\}|.$$

We consider the first term from the above formula. According to Taylor expansion, there exists a $\check{\beta}_s$ between β_s^* and $\widehat{\beta}_s$ such that $\mu_i(\beta_s^*) - \mu_i(\widehat{\beta}_s) = D_i(\check{\beta})(\beta_s^* - \widehat{\beta}_s)$. Then we have

$$\begin{aligned}
&|\sum_{i=1}^{n}\{\mu_i(\beta_s^*) - \mu_i(\widehat{\beta}_s)\}^T\widehat{V}_i^{-1}\{\mu_i(\beta_s^*) - \mu_i(\widehat{\beta}_s)\}| \\
&= \sum_{i=1}^{n}(\beta_s^* - \widehat{\beta}_s)^T D_i(\check{\beta})^T\widehat{V}_i^{-1}D_i(\check{\beta})(\beta_s^* - \widehat{\beta}_s) \\
&\leq n\lambda_{\max}\{\widehat{V}_i^{-1}\} \times ||\beta_s^* - \widehat{\beta}_s||^2 \times \max_{||v||^2=1}\{v^T\frac{1}{n}\sum_{i=1}^{n}D_i(\check{\beta})^T D_i(\check{\beta})v\} \\
&= O_p(p_n^2 \log p_n).
\end{aligned}$$

Next we consider the second term. Lemma 4.12 implies that $\max_j\{(1/n)\sum_{i=1}^{n}|Y_{ij} - \mu_{ij}(\beta_s^*)|\} = O_p(1)$. Assumption 4.3 implies that $\max_i\{||D_i(\check{\beta})||_{\max}||\widehat{V}_i^{-1}||_{\max}\} =$

$O_p(1)$. Combining these results, we have

$$
\begin{aligned}
&|\sum_{i=1}^{n}\{\mu_i(\beta_s^*)-\mu_i(\widehat{\beta}_s)\}^T\widehat{V}_i^{-1}\{Y_i-\mu_i(\beta_s^*)\}| \\
&=|\sum_{i=1}^{n}(\beta_s^*-\widehat{\beta}_s)^T D_i(\check{\beta})^T\widehat{V}_i^{-1}\{Y_i-\mu_i(\beta_s^*)\}| \\
&\leq ||\beta_s^*-\widehat{\beta}_s||\times||\sum_{i=1}^{n}D_i(\check{\beta})^T\widehat{V}_i^{-1}\{Y_i-\mu_i(\beta_s^*)\}|| \\
&\leq ||\beta_s^*-\widehat{\beta}_s||\times p_n^{1/2}\max_k|\sum_{i=1}^{n}[D_i(\check{\beta})^T]_{[k,]}\widehat{V}_i^{-1}\{Y_i-\mu_i(\beta_s^*)\}| \\
&\leq np_n^{1/2}||\beta_s^*-\widehat{\beta}_s||\times\frac{1}{n}\sum_{i=1}^{n}m^2||D_i(\check{\beta})||_{\max}||\widehat{V}_i^{-1}||_{\max}\times\max_j|Y_{ij}-\mu_{ij}(\beta_s^*)| \\
&\leq np_n^{1/2}m^2||\beta_s^*-\widehat{\beta}_s||\times\max_i\{||D_i(\check{\beta})||_{\max}||\widehat{V}_i^{-1}||_{\max}\}\times\max_j\{\frac{1}{n}\sum_{i=1}^{n}|Y_{ij}-\mu_{ij}(\beta_s^*)|\} \\
&=O_p\{(p_n^3\log p_n/n)^{1/2}\}.
\end{aligned}
$$

■

Proof of Lemma 4.6. The proof is exactly the same as Lemma 3.10 given the same assumptions. So we do not repeat here. ■

Proof of Lemma 4.7. The proof is the same as Lemma 3.8 given the same assumptions. The only change is to replace all of s_n to p_n. s_n is the model size up bound and can go as large as p_n. The replacement from s_n to p_n does not change the proofs. So we do not repeat the proof here. ■

5 Conclusions and Future Work

In this chapter, we summarize the conclusion and contribution in this dissertation and discuss potential future works.

5.1 Conclusions

It is the first time that high dimensional GEE estimation consistency has been proved under any arbitrary positive definite working correlation matrix. Liang and Zeger (1986) has proved the GEE estimation consistency for any arbitrary positive definite working correlation matrix when the size of covariates is finite. Although Wang (2011) has demonstrated that GEE estimation is still consistent with diverging number of covariates p_n, it requires a particular unstructured working correlation matrix from Equation (2.2). The Table 5.1 compares the differences among Liang and Zeger (1986), Wang (2011) and this dissertation.

Furthermore the dissertation launches two new GEE information criteria QBIC

and GIC. It is the first time that GEE model selection is consistency for ultra-high dimensional data when the size of p_n could go to infinity. Table 5.2 compares the different type of information criteria under GEE framework.

Table 5.1: Comparison of Different Estimation Consistency

α' is a small positve value

	Liang and Zeger (1986)	Wang (2011)	This Dissertation
Size of p_n	$p_n = O(1)$	$p_n^2/n \to 0$	$p_n^{4+\alpha'}/n \to 0$
Consistency	$(1/n)^{1/2}$	$(p_n/n)^{1/2}$	$(p_n^{1+\alpha'}/n)^{1/2}$
Corr Matrix	Any	Equation 2.2	Any

In traditional statistical study, the variable size of p is usually bounded. And the size of observation n will go to infinity to make sure the asymptotic feature. If $p_n = O(n^\alpha)$ for some positive number α, we say it is the high dimensional setting. If $\log p_n = O(n^\alpha)$ for some positive number α, we say it is ultra high dimensional setting (Fan and Lv, 2011). Here QBIC could extend to ultra high dimensional setting as long as $s_n^5 \log p_n = o(n)$. If s_n is small, $\log p_n = o(n)$. But the GIC is not able to extend to ultra high dimensional setting given we need full model to find out β_F and requires $p < n$.

Table 5.2: Comparison of Different Information Criteria

	QIC	BIQIF	QBIC	GIC
Consistency	No	Yes	Yes	Yes
Corr Matrix	Any	Basis	Identity	Any
Size of p_n	$p_n = O(1)$	$p_n = O(1)$	$p_n \to \infty$	$p_n \to \infty$
Size of n	$n \to \infty$	$n \to \infty$	$s_n^5 \log p_n / n \to 0$	$p_n^5 \log p_n / n \to 0$

5.2 Future Work

There are two potential future works. Firstly we consider the future work of QBIC. QBIC requires the identity working correlation matrix. The reason springs from the simplified quasi-likelihood assuming the identity correlation matrix. Equation (3.7) in Lemma 3.3 indicates that for an overfitting model $s \in S_+$, $QL\{\widehat{\beta}_s(I)\} - QL(\beta_s^*)$ can be approximated by a quadratic form. However we cannot rewrite $QL\{\widehat{\beta}_s(R)\} - QL(\beta_s^*)$ in a similar quadratic form easily. That is why the approach from Chapter 3 is not able to prove the model selection consistency for any arbitrary working correlation matrix. According to the numerical simulation result, the choice of working correlation matrix has merely impacts to the simplified quasi-likelihood and QBIC value. In fact according to limited numerical simulations, I observe that

$QL\{\widehat{\beta}_s(R)\} - QL\{\widehat{\beta}_T(R)\}$ approximately equals to $Q\{\widehat{\beta}_s(R)\} - Q\{\widehat{\beta}_T(R)\}$. If we can find a solution to prove the observation which indicates that QBIC and GIC are equivalent, we can also prove the model selection consistency for QBIC under any arbitrary working correlation matrix. In my opinion, it is highly possible that the QBIC style model selection consistency is still valid for any arbitrary positive definite working correlation matrix. To prove the QBIC model selection consistency for flexible choice of working correlation matrix is a potential future work.

In addition, GIC has flexibility to use any arbitrary working correlation matrix but with a new limitation that $p_n \leq n$. The reason is that GIC need to determine the volatility $A_i(\widehat{\beta}_F)$ which is at $(p_n^3 \log p_n/n)^{1/2}$ neighborhood of $A_i(\beta_T^*)$. If the size of variables is larger than the sample size $p_n > n$, it is full rank situation that GEE estimator is not stable or exist and therefore it is not able to find out $\widehat{\beta}_F$. For Gaussian distribution whose volatility is not a function of β, we can safely remove the requirement $p_n \leq n$ and still able to show the model selection consistency. In fact Kim et al. (2012) has proved the model selection consistency for Gaussian distribution under linear regression, a special case of this GIC which does not require $p_n \leq n$. In modern high dimensional data situation, the model selection in $p_n > n$ scenario is a popular topic. Although using LASSO or SCAD type of penalty could reduce the dimensional of variables size p_n, it is not guaranteed that the true model is included in

the selected sub-models from LASSO or SCAD type of technique. Another potential future work is to extend the GIC into the $p_n > n$ scenario.

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723.

Balan, R. M. and Schiopu-Kratina, I. (2005). Asymptotic results with generalized estimating equations for longitudinal data. *Annals of Statistics*, 33(2):522–541.

Cantoni, E., Flemming, J. M., and Ronchetti, E. (2005). Variable selection for marginal longitudinal generalized linear models. *Biometrics*, 61(2):507–514.

Carey, V. J. and Wang, Y.-G. (2011). Working covariance model selection for generalized estimating equations. *Statistics in medicine*, 30(26):3117–3124.

Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771.

Donoho, D. L. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. *AMS Math Challenges Lecture*, 1:32.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.

Gao, X. and Carroll, R. J. (2017). Data integration with high dimensionality. *Biometrika*, 104(2):251–272.

Gao, X. and Song, P. X.-K. (2010). Composite likelihood bayesian information criteria for model selection in high-dimensional data. *Journal of the American Statistical Association*, 105(492):1531–1540.

Kim, Y., Kwon, S., and Choi, H. (2012). Consistent model selection criteria on high dimensions. *Journal of Machine Learning Research*, 13(Apr):1037–1057.

Li, B. (1997). On the consistency of generalized estimating equations. *Lecture Notes-Monograph Series*, pages 115–136.

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

Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Annals of Statistics*, pages 3498–3528.

Mallows, C. L. (1973). Some comments on c p. *Technometrics*, 15(4):661–675.

McCullagh, P. and Nelder, J. A. (1989). Generalized linear models, no. 37 in monograph on statistics and applied probability.

McCullough, P. and Nelder, J. (1989). Generalized linear models chapman and hall. *New York*.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics*, 57(1):120–125.

Portnoy, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Annals of Statistics*, 16(1):356–366.

Qu, A., Lindsay, B. G., and Li, B. (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika*, 87(4):823–836.

Spokoiny, V. and Zhilova, M. (2013). Sharp deviation bounds for quadratic forms. *Mathematical Methods of Statistics*, 22(2):100–113.

Stewart, G. W. (1990). Matrix perturbation theory.

Varin, C. and Vidoni, P. (2005). A note on composite likelihood inference and model selection. *Biometrika*, 92(3):519–528.

Wang, L. (2011). Gee analysis of clustered binary data with diverging number of covariates. *Annals of Statistics*, 39(1):389–417.

Wang, L. and Qu, A. (2009). Consistent model selection and data-driven smooth tests for longitudinal data in the estimating equations approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(1):177–190.

Wedderburn, R. W. (1974). Quasi-likelihood functions, generalized linear models, and the gauss—newton method. *Biometrika*, 61(3):439–447.

White, H. (1981). Consequences and detection of misspecified nonlinear regression models. *Journal of the American Statistical Association*, 76(374):419–433.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, pages 1–25.

Xie, M. and Yang, Y. (2003). Asymptotics for generalized estimating equations with large cluster sizes. *Annals of Statistics*, 31(1):310–347.

Zhang, Y. and Shen, X. (2010). Model selection procedure for high-dimensional data.

Statistical Analysis and Data Mining: The ASA Data Science Journal, 3(5):350–358.

A Appendix: More Tables

The Table A.1 and A.2 illustrate the simulation results of QIC, QBIC and GIC for both Binary response and Gaussian response under 100 repeated measurements. We simulate 4 different group of the scenarios: $(n = 1000, p = 1000, d_T = 50, m = 10)$; $(n = 1000, p = 500, d_T = 50, m = 10)$; and $(n = 500, p = 500, d_T = 50, m = 10)$. The Table A.3 compares the simulation results between Binary response and Gaussian response for parameters setting $(n = 500, p = 500, d_T = 50, m = 20)$. The different correlation matrix Independent (I), Exchangeable (E), AR1 (A), and Unstructured (U) from Equation (2.2) are considered. We also list the results when the free penalty parameter c varies from 1 to 4.

Table A.1: Different Information Criteria Simulation Result for Binary Response

	n 1000	p 1000	d_T 50	m 10	n 1000	p 500	d_T 50	m 10	n 500	p 500	d_T 50	m 10
	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
	psr	psr	fdr	fdr	psr	psr	fdr	fdr	psr	psr	fdr	fdr
QIC.I	1.0000	0.0000	0.7093	0.0241	1.0000	0.0000	0.4984	0.0585	0.9974	0.0084	0.5677	0.0735
QIC.E	1.0000	0.0000	0.7099	0.0241	1.0000	0.0000	0.4988	0.0582	0.9974	0.0084	0.5677	0.0735
QIC.A	1.0000	0.0000	0.7093	0.0241	1.0000	0.0000	0.4984	0.0585	0.9974	0.0084	0.5677	0.0735
QIC.B	1.0000	0.0000	0.7234	0.0179	1.0000	0.0000	0.5151	0.0707	0.9976	0.0082	0.6163	0.0677
QBIC.1	0.9994	0.0060	0.1192	0.0585	1.0000	0.0000	0.0831	0.0472	0.9576	0.0562	0.0653	0.0778
QBIC.2	0.9774	0.0399	0.0029	0.0092	0.9897	0.0201	0.0033	0.0099	0.7920	0.0781	0.0000	0.0000
QBIC.3	0.8984	0.0779	0.0000	0.0000	0.9321	0.0656	0.0000	0.0000	0.6698	0.0876	0.0000	0.0000
QBIC.4	0.8042	0.0808	0.0000	0.0000	0.8339	0.0821	0.0000	0.0000	0.5992	0.0844	0.0000	0.0000
GIC.I.1	0.9982	0.0081	0.0596	0.0476	0.9988	0.0069	0.1194	0.0932	0.9182	0.0710	0.0250	0.0548
GIC.I.2	0.9554	0.0552	0.0012	0.0065	0.9822	0.0339	0.0147	0.0233	0.7378	0.0881	0.0000	0.0000
GIC.I.3	0.8398	0.0854	0.0000	0.0000	0.9317	0.0898	0.0018	0.0068	0.6288	0.0794	0.0000	0.0000
GIC.I.4	0.7568	0.0802	0.0000	0.0000	0.8566	0.1222	0.0008	0.0047	0.5524	0.0906	0.0000	0.0000
GIC.E.1	0.9982	0.0081	0.0574	0.0454	0.9986	0.0071	0.1217	0.0939	0.9194	0.0715	0.0265	0.0561
GIC.E.2	0.9554	0.0552	0.0012	0.0065	0.9822	0.0339	0.0147	0.0233	0.7410	0.0880	0.0000	0.0000
GIC.E.3	0.8414	0.0858	0.0000	0.0000	0.9323	0.0900	0.0018	0.0068	0.6316	0.0833	0.0000	0.0000
GIC.E.4	0.7560	0.0810	0.0000	0.0000	0.8592	0.1211	0.0006	0.0043	0.5546	0.0906	0.0000	0.0000
GIC.A.1	0.9980	0.0083	0.0584	0.0471	0.9988	0.0069	0.1168	0.0915	0.9180	0.0708	0.0246	0.0550
GIC.A.2	0.9554	0.0552	0.0014	0.0068	0.9822	0.0339	0.0146	0.0230	0.7378	0.0881	0.0000	0.0000
GIC.A.3	0.8414	0.0837	0.0000	0.0000	0.9317	0.0898	0.0023	0.0088	0.6278	0.0784	0.0000	0.0000
GIC.A.4	0.7568	0.0802	0.0000	0.0000	0.8558	0.1215	0.0008	0.0047	0.5536	0.0910	0.0000	0.0000
GIC.U.1	0.9990	0.0066	0.0445	0.0399	0.9998	0.0020	0.0990	0.0885	0.9498	0.0538	0.0242	0.0381
GIC.U.2	0.9820	0.0296	0.0017	0.0085	0.9911	0.0200	0.0098	0.0193	0.7978	0.0848	0.0000	0.0000
GIC.U.3	0.9060	0.0773	0.0000	0.0000	0.9588	0.0651	0.0025	0.0085	0.6782	0.0909	0.0000	0.0000
GIC.U.4	0.8146	0.0820	0.0000	0.0000	0.9143	0.0969	0.0006	0.0034	0.6094	0.0841	0.0000	0.0000

Table A.2: Different Information Criteria Simulation Result for Gaussian Response

	n 1000	p 1000	d_T 50	m 10	n 1000	p 500	d_T 50	m 10	n 500	p 500	d_T 50	m 10
	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
	psr	psr	fdr	fdr	psr	psr	fdr	fdr	psr	psr	fdr	fdr
QIC.I	1.0000	0.0000	0.6937	0.0326	1.0000	0.0000	0.5308	0.0652	1.0000	0.0000	0.5401	0.0607
QIC.E	1.0000	0.0000	0.6937	0.0326	1.0000	0.0000	0.5381	0.0648	1.0000	0.0000	0.5395	0.0599
QIC.A	1.0000	0.0000	0.6937	0.0326	1.0000	0.0000	0.5308	0.0652	1.0000	0.0000	0.5401	0.0607
QIC.B	1.0000	0.0000	0.7281	0.0136	1.0000	0.0000	0.7077	0.0354	1.0000	0.0000	0.7109	0.0324
QBIC.1	1.0000	0.0000	0.1077	0.0460	1.0000	0.0000	0.0784	0.0415	1.0000	0.0000	0.0871	0.0449
QBIC.2	1.0000	0.0000	0.0012	0.0047	1.0000	0.0000	0.0024	0.0065	1.0000	0.0000	0.0008	0.0038
QBIC.3	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0004	0.0028	1.0000	0.0000	0.0000	0.0000
QBIC.4	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
GIC.I.1	1.0000	0.0000	0.1077	0.0460	1.0000	0.0000	0.0784	0.0415	1.0000	0.0000	0.0871	0.0449
GIC.I.2	1.0000	0.0000	0.0012	0.0047	1.0000	0.0000	0.0024	0.0065	1.0000	0.0000	0.0008	0.0038
GIC.I.3	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0004	0.0028	1.0000	0.0000	0.0000	0.0000
GIC.I.4	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
GIC.E.1	1.0000	0.0000	0.0961	0.0511	1.0000	0.0000	0.0709	0.0404	1.0000	0.0000	0.0705	0.0450
GIC.E.2	1.0000	0.0000	0.0028	0.0095	1.0000	0.0000	0.0024	0.0065	1.0000	0.0000	0.0015	0.0065
GIC.E.3	1.0000	0.0000	0.0004	0.0028	1.0000	0.0000	0.0004	0.0028	1.0000	0.0000	0.0000	0.0000
GIC.E.4	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
GIC.A.1	1.0000	0.0000	0.1073	0.0471	1.0000	0.0000	0.0784	0.0415	1.0000	0.0000	0.0860	0.0461
GIC.A.2	1.0000	0.0000	0.0012	0.0047	1.0000	0.0000	0.0024	0.0065	1.0000	0.0000	0.0008	0.0038
GIC.A.3	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0004	0.0028	1.0000	0.0000	0.0000	0.0000
GIC.A.4	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
GIC.B.1	1.0000	0.0000	0.0226	0.0295	1.0000	0.0000	0.0116	0.0220	1.0000	0.0000	0.0272	0.0355
GIC.B.2	1.0000	0.0000	0.0012	0.0047	1.0000	0.0000	0.0012	0.0047	1.0000	0.0000	0.0004	0.0027
GIC.B.3	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0004	0.0028	1.0000	0.0000	0.0000	0.0000
GIC.B.4	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000

Table A.3: Comparison Between Binary Response VS Gaussian Response

	Binary				Gaussian			
	n 500	p 500	d_T 50	m 20	n 500	p 500	d_T 50	m 20
	mean	std	mean	std	mean	std	mean	std
	psr	psr	fdr	fdr	psr	psr	fdr	fdr
QIC.I	1.0000	0.0000	0.5482	0.0629	1.0000	0.0000	0.5262	0.0628
QIC.E	1.0000	0.0000	0.5490	0.0634	1.0000	0.0000	0.5262	0.0628
QIC.A	1.0000	0.0000	0.5490	0.0634	1.0000	0.0000	0.5262	0.0628
QIC.B	1.0000	0.0000	0.5978	0.0553	1.0000	0.0000	0.7294	0.0207
QBIC.1	1.0000	0.0000	0.0876	0.0482	1.0000	0.0000	0.0779	0.0406
QBIC.2	0.9854	0.0252	0.0039	0.0099	1.0000	0.0000	0.0036	0.0086
QBIC.3	0.9308	0.0593	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
QBIC.4	0.8646	0.0837	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
GIC.I.1	0.9986	0.0051	0.0452	0.0375	1.0000	0.0000	0.0779	0.0406
GIC.I.2	0.9708	0.0380	0.0019	0.0070	1.0000	0.0000	0.0036	0.0086
GIC.I.3	0.8952	0.0755	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
GIC.I.4	0.7966	0.1006	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
GIC.E.1	0.9986	0.0051	0.0456	0.0375	1.0000	0.0000	0.0775	0.0409
GIC.E.2	0.9710	0.0379	0.0019	0.0070	1.0000	0.0000	0.0036	0.0086
GIC.E.3	0.8952	0.0755	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
GIC.E.4	0.7966	0.1006	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
GIC.A.1	0.9986	0.0051	0.0448	0.0375	1.0000	0.0000	0.0789	0.0411
GIC.A.2	0.9706	0.0379	0.0019	0.0070	1.0000	0.0000	0.0032	0.0083
GIC.A.3	0.8952	0.0755	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
GIC.A.4	0.7966	0.1006	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
GIC.U.1	0.9992	0.0039	0.0346	0.0370	1.0000	0.0000	0.0147	0.0292
GIC.U.2	0.9916	0.0187	0.0020	0.0065	1.0000	0.0000	0.0016	0.0067
GIC.U.3	0.9562	0.0525	0.0004	0.0028	1.0000	0.0000	0.0000	0.0000
GIC.U.4	0.9012	0.0733	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000