Statistica Sinica Preprint No: SS-2020-0197	
Title	Model Selection of Generalized Estimating Equation With Divergent Model Size
Manuscript ID	SS-2020-0197
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202020.0197
Complete List of Authors	Shicheng Wu, Xin Gao and Raymond James Carroll
Corresponding Author	Xin Gao
E-mail	xingao@mathstat.yorku.ca
Notice: Accepted version subject to English editing.	

Model Selection of Generalized Estimating Equation with Divergent Model Size

Shicheng Wu, Xin Gao and Raymond J. Carroll

York University and Texas A&M University

Abstract: We consider the problem of model selection for high dimensional generalized estimating equation (GEE) on marginal regression analysis for clustered or longitudinal data. As the GEE method only makes assumptions about the first two moments, the full likelihood is not specified. Therefore, the likelihood based model selection criteria cannot be directly applied. This paper 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 criterion for divergent number of covariates is established.

Key words and phrases: generalized estimation equation, generalized information criterion, large deviation, model selection consistency.

1. Introduction

With the advent of large collection of information, it is an essential problem to perform model selection to determine a subset of useful covari-

ates. We consider the problem of model selection on generalized estimating equations for clustered or longitudinal data. As the full likelihood of the multivariate clustered data is often difficult to specify, Liang and Zeger (1986) extended the generalized linear models (McCullough and Nelder, 1989) to correlated data and proposed the generalized estimating equation. The GEE estimate 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 pseudo-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 estimates under the "large n, diverging p" scenario under the true model. When the number of predictors, true and zero both diverge, there is a multitude of competing models. No study has been done to investigate

the properties of GEE estimates under various competing models including underfitting models. In this paper we develop the methodology for this problem along with rates of convergence and model selection strategies for high dimensional GEE.

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 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 type of criterion (BIQIF) using Qu et al. (2000)'s Quadratic Inference Function. The model selection consistency of BIQIF was established for finite number of covariates. Fang et al. (2020) proposed a new quadratic decorrelated inference function approach for high-dimensional generalized estimating equations. For divergent number of parameters, the limiting distribution of the estimator is established.

The proposed test can be used to perform variable selections while controlling the false discovery rate. 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 be selection consistent 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 this paper, we aim to develop an information criterion for GEE with divergent number of predictors and unbounded true model size. Different from the likelihood setting in Gao and Carroll (2017), there is no likelihood available to evaluate the model fitting under GEE. Instead of likelihood formulation, we consider a goodness-of-fit measure. Since the working covariance matrix is used to model the within cluster covariance structure, we will use the working covariance matrix and the fitted residuals together to construct a quadratic form which will serve as the goodness-of-fit measure of the candidate model. 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 the goodness-of-fit measure will enable us to choose the appropriate penalty size on the model complexity to ensure the model selection consistency. Rather surprisingly, we will be able to show the proposed information criterion is selection consistent for the marginal mean model even if the working correlation is mis-specified. This model selection robustness is an extension to the estimation consistency of the GEE estimator under the mis-specification of the underlying working correlation. To our knowl-

edge, this is the first result of model selection consistence for GEE under large n and diverging p_n situation.

The rest of the paper is structured as follows. In Section 2, we will investigate the convergence rate of the GEE estimates under various competing models. Then we introduce the Generalized Information Criterion (GIC) and establish the model selection consistence of GIC under the "large n, and divergent p" setting. In Section 3, the performance of the proposed model selection criterion will be evaluated in some numerical studies and a real data analysis.

2. Model selection for Generalized Estimating Equation

2.1 Generalized Estimating Equations

Suppose n clusters are randomly selected for the study. These could be subjects with repeated measurement. The size of the ith cluster is m_i. For cluster $i = 1, \ldots, n$, let $Y_i = (Y_{i1}, \ldots, Y_{im_i})^T$ be an $m_i \times 1$ response vector with mean $\mathrm{E}(Y_i) = \mu_i$, where $\mu_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{im_i})^T$. Let $X_i = (X_{i1}, X_{i2} \ldots X_{im_i})^T$ 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 $\beta = (\beta_1, \beta_2 \ldots \beta_{p_n})^T$ denotes the p_n dimensional regression coefficients. Let A_i be a diagonal matrix with ele-

ments $\text{Var}(Y_{ij}) = \nu(\mu_{ij})\phi$, where ϕ is the dispersion parameters and $\nu()$ is the variance function. Let R_i be the working correlation matrix and $V_i = A_i^{1/2} R_i A_i^{1/2} \phi$ be the working covariance matrix. For simplicity, it is assumed that $\phi = 1$ throughout the paper. Discussion on the situation with dispersion parameter $\phi \neq 1$ is provided in Section 3.

The true correlation matrix is denoted as R_i^*, which is usually unknown. The working correlation matrix R_i is user defined and could be either unstructured or structured such as independent, autocorrelation, or compound symmetry. The working correlation matrix $R_i(\varrho)$ 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 fixed, the GEE solution $\widehat{\beta}$ is $n^{-1/2}$ consistent even with the mis-specified working correlation matrix R_i. 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.

2.2 A Quadratic Form of Goodness-of-fit Measure

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 (2000)'s QIC, the quasi-likelihood of each observation from a cluster is added together under a working independence assumption. However, the consistency of QIC for model selection under either finite or diverging p_n is not established.

Consider a divergent number p_n of covariates where $p_n \to \infty$, and $p_n \leq n$. Let s be a subset of $\{1, 2 \ldots, p_n\}$. The model with $\beta_k = 0$ for all $k \notin s$ is denoted as a model s. Let $\widehat{\beta}_s$ denote the GEE estimate 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:

$$Q(\widehat{\beta}_s) = \frac{1}{2}\sum_{i=1}^{n}\{Y_i - \mu_i(\widehat{\beta}_s)\}^T A_i(\widehat{\beta}_F)^{-1/2} R_i^{-1} A_i(\widehat{\beta}_F)^{-1/2}\{Y_i - \mu_i(\widehat{\beta}_s)\}, \quad (2.2)$$

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_i can be

any positive definite matrix with diagonal entries equal to one. Note that both $A_i(\widehat{\beta}_F)$ and R_i will remain the same for different competing models in equation (2.2). 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 $\widehat{V}_i^{-1} = A_i(\widehat{\beta}_F)^{-1/2} R_i^{-1} A_i(\widehat{\beta}_F)^{-1/2}$, equation (2.2) can be reformulated as:

$$Q(\widehat{\beta}_s) = \frac{1}{2} \sum_{i=1}^{n} \{Y_i - \mu_i(\widehat{\beta}_s)\}^T \widehat{V}_i^{-1} \{Y_i - \mu_i(\widehat{\beta}_s)\}. \tag{2.3}$$

Throughout this article, $\widehat{V}_i$ stands for the estimated covariance matrix that is evaluated at the full model and $V_i(\widehat{\beta}_s)$ denotes the working covariance matrix evaluated at a competing model s. Equation (2.3) is similar to Carey and Wang (2011)'s Gaussian pseudo-likelihood which takes the form of $-2^{-1}\{\sum_{i=1}^{n}\{Y_i - \mu_i(\widehat{\beta}_s)\}^T V_i(\widehat{\beta}_s)^{-1}\{Y_i - \mu_i(\widehat{\beta}_s)\} + \log(|V_i(\widehat{\beta}_s)|)\}$. 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.

2.3 Generalized Information Criterion

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 $d_s \leq p_n$. 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. The true model T belongs to S_+. As n increases to infinity, the model space S, all sub models s and the true model T all depend on n and d_T is unbounded.

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 $s \in S_-$, it is assumed 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) and it depends on the sample size.

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

$$GIC(s) = 2Q(\widehat{\beta}_s) + d_s^* \gamma_n. \tag{2.4}$$

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 $d_s^* = \text{tr}\{W_s(\beta_s^*)\Omega_s^{-1}(\beta_s^*)\}$, where the variability matrix $W(\beta_s^*) = n^{-1}\text{Cov}\{U(\beta_s^*)\}$ and the sensitivity matrix $\Omega(\beta_s^*) = -n^{-1}\text{E}\{\partial U(\beta_s)/\partial\beta_s^T|_{\beta_s^*}\}$. 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 $\text{E}\{Y_i - \mu_i(\beta_s^*)\} = 0$, the variability matrix and sensitivity matrix can be expressed as $W(\beta_s^*) = n^{-1}\sum_{i=1}^n D_i(\beta_s^*)^T V_i(\beta_s^*)^{-1}\text{Cov}(Y_i)V_i(\beta_s^*)^{-1}D_i(\beta_s^*)$ and $\Omega(\beta_s^*) = n^{-1}\sum_{i=1}^n D_i(\beta_s^*)^T V_i(\beta_s^*)^{-1}D_i(\beta_s^*)$.

2.4 Estimation Consistency under Various Competing Models

In this section, we will investigate the estimation consistency of the GEE estimator under various competing models. We first introduce some notation. Let $||.||$ denote the Euclidean norm, $||.||_{\max}$ denote the largest absolute value in the matrix or vector, $\lambda_{\max}(.)$ and $\lambda_{\min}(.)$ denote the largest and smallest

eigenvalue of the matrix, and Tr(.) denote the trace of matrix. Let $[.]_{[i,j]}$, $[.]_{[i,]}$ and $[.]_{[,j]}$ denote the (i,j)th element, the ith row vector and the jth column vector of a matrix.

Assumption 1. *The maximum cluster size $m = \max_i m_i$ is assumed to be bounded. As $n \to \infty$, $p_n \to \infty$ and $p_n^5 \log p_n / n \to 0$, the distance between the true model T and any underfitting model s satisfies*

$$\liminf_n \min_{s \in S_-} n^{-1} [\sum_{i=1}^{n} \{\mu_i(\beta_T^*) - \mu_i(\beta_s^*)\}^T \{\mu_i(\beta_T^*) - \mu_i(\beta_s^*)\}] / (p_n^3 \log p_n / n)^{1/2} = \infty.$$

This assumption ensures the identifiability of the true model. Similar identifiability conditions were assumed (Chen and Chen, 2008; Fan and Lv, 2011; Gao and Carroll, 2017). The term $n^{-1} \sum_{i=1}^{n} \{\mu_i(\beta_T^*) - \mu_i(\beta_s^*)\}^T \{\mu_i(\beta_T^*) - \mu_i(\beta_s^*)\}$ measures the distance between the true model T and a competing model s. For example, consider a multivariate Gaussian distribution with an identity covariance matrix, the distance between the true model T and a competing model s takes the form $n^{-1} \sum_{i=1}^{n} \{\mu_i(\beta_T^*) - \mu_i(\beta_s^*)\}^T \{\mu_i(\beta_T^*) - \mu_i(\beta_s^*)\}$, which coincides with the Kullback Leibler distance $n^{-1} \mathrm{E}\{l(\beta_T^*) - l(\beta_s^*)\}$ based on the likelihood. By definition, the true model is the most parsimonious model which ensures $\mu_i(\beta_T^*) = \mathrm{E}(Y_i)$ for all i. In contrast, for an underfitting model $s \in S_-$, $\mu_i(\beta_s^*) \neq \mathrm{E}(Y_i) = \mu_i(\beta_T^*)$, for some i. If $n^{-1} [\sum_{i=1}^{n} \{\mu_i(\beta_T^*) - \mu_i(\beta_s^*)\}^T \{\mu_i(\beta_T^*) - \mu_i(\beta_s^*)\}]$ is as large as $O(1)$, then

the assumption is easily satisfied given $(p_n^3 \log p_n/n)^{1/2} \to 0$. For nontrivial cases, we allow the minimum distance between the true model T and any competing underfitting model s to approach zero provided that it converges to zero at a rate slower than $(p_n^3 \log p_n/n)^{1/2}$.

Assumption 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 exists two positive value b_1 and b_2 such that all the eigenvalues of $\Omega(\beta_s)$, $W(\beta_s)$, $n^{-1}\sum_{i=1}^n X_i^T X_i$ and $Cov(Y_i)$, $i = 1, \ldots, n$, are bounded from below by b_1 and bounded from above by b_2. The two constants b_1 and b_2 are universal for all $s \in S$.*

The condition of bounded eigenvalues is a common assumption in the literature on estimation with diverging dimension. Similar assumption can be found in Assumption (A3) of Wang (2011).

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 $\Lambda_{ij}(\beta) = \partial\mu_{ij}(\beta)/\partial\zeta_{ij}(\beta)$ and $\Lambda_i(\beta) = diag\{\Lambda_{ij}(\beta), j = 1, \ldots, m\}$, a diagonal matrix of dimension m_i.

Assumption 3. *For all $s \in S$ and all i, j, k, there exist positive values b_3 and b_4 such that the covariates and the linear predictors are uniformly bounded $|X_{ijk}| < b_4$, and $|\zeta_{ij}(\beta_s^*)| < b_4$. On the bounded region of $\zeta_{ij}(\beta_s^*)$, we assume the inverse of link function $g^{-1}(.)$ has continuous derivatives*

up to the third order which are all bounded by b_4. *We assume the variance functions are uniformly bounded away from zero with* $A_{ij}(\beta_s^*) > b_3$. *Furthermore, on the bounded region of* $\mu_{ij}(\beta_s^*)$, *the variance function* $\nu(.)$ *has continuous derivatives up to the second order which are all bounded by* b_4.

The commonly used link functions and variance functions all satisfy the continuity and smoothness conditions required in Assumption 3. For example, given that the linear predictors are bounded, the logistic link $g^{-1}(w) = \exp(w)/\{1 + \exp(w)\}$ and variance function $\nu(w) = w(1 - w)$ both have bounded second and third derivatives.

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) for sub-exponential random vector which satisfies an exponential moment condition:

$$\log[\mathrm{E}\{\exp(t^T\psi)\}] \leq ||t||^2/2, ||t|| \leq \rho, \tag{2.5}$$

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 $\rho^2/4 > K > V_G/3$,

$$\Pr(\psi^T O\psi > P_G + K) \leq 10.4\exp(-K/6). \tag{2.6}$$

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 (2.6), which is exponentially small . The total number of competing model is of the order of 2^{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. This sub-exponential condition is often used in high dimensional data analysis literature (Ning and Liu, 2017; Fang et al., 2020). Gao and Carroll (2017) show that the exponential moment condition in equation (2.5) can be satisfied asymptotically by sample mean types of statistics if the original random vector satisfies the following cumulant boundedness condition.

Definition 1. For a random vector Z of dimension m, let $C(t)$ 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 $|\partial C(0)/\partial t_k| \leq b_5$ and $|\partial^2 C(0)/\partial t_k \partial t_l| \leq b_5$. Furthermore there exists a constant b_6 such that with $||t|| \leq b_6$, the absolute value of all the third derivatives of its cumulant

generating function satisfy $|\partial^3 C(t)/\partial t_k \partial t_l \partial t_r| \leq b_5$.

Let $Q_i(\beta) = \{Y_i - \mu_i(\beta)\}^T \widehat{V}_i^{-1} \{Y_i - \mu_i(\beta)\}$ and $U_i(\beta) = D_i(\beta)^T V_i(\beta)^{-1} \{Y_i - \mu_i\}$. 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. *There exists a neighborhood* $||\beta_s - \beta_s^*|| \leq (p_n^2 \log p_n / n)^{1/2}$, *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 model* $s \in S$.

The cumulant boundedness condition holds for exponential family in generalized linear models (Gao and Carroll, 2017). Under GEE model, we use Lemma S2.1 to show that Assumption 4 is satisfied if the joint distribution of each cluster belongs to the multivariate exponential family and each observation is a sub-Gaussian random variable. Using large deviation result in Spokoiny and Zhilova (2013) and Gao and Carroll (2017), we obtain the asymptotic orders of the following terms.

Lemma 1. *Under Assumption 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 1. *Under Assumptions 1 - 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.*

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

Lemma 2. *Under Assumptions 1 - 4, for all model* $s \in S_+$ *and* $i = 1, 2 \ldots n$, $\max[|\lambda_{\max}\{V_i^{-1}(\widehat{\beta}_F) - V_i^{-1}(\widehat{\beta}_s)\}|, |\lambda_{\min}\{V_i^{-1}(\widehat{\beta}_F) - V_i^{-1}(\widehat{\beta}_s)\}|] = O_p\{(p_n^3 \log p_n/n)^{1/2}\}$, *and* $\max[|\lambda_{\max}\{V_i^{-1}(\widehat{\beta}_F) - V_i^{-1}(\beta_s^*)\}|, |\lambda_{\min}\{V_i^{-1}(\widehat{\beta}_F) - V_i^{-1}(\beta_s^*)\}|] = O_p\{(p_n^3 \log p_n/n)^{1/2}\}$.

For true and overfitting models, Lemma 2 measures the distance between the two matrices $V_i(\widehat{\beta}_s)$ and $V_i(\beta_s^*)$.

2.5 Model Selection Consistency

In this section, 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 space. Second, based on the quadratic forms, we apply the large deviation result to quantify the size of the penalty γ_n.

Lemma 3. *Under Assumptions 1 - 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 holds for all models $s \in S_+$.*

Lemma 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. *Under Assumptions 1 - 4, the differences between the goodness-of-fit measures can be approximated as quadratic forms:*

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

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

Lemma 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 5 establishes the asymptotic order of these quadratic forms.

Lemma 5. *Under Assumptions 1 - 4,*

$$\sup_{s\in S_+} |Q(\widehat{\beta}_s) - Q(\beta_s^*)| = O_p(p_n^2 \log p_n);$$

$$\sup_{s\in S_-} |Q(\widehat{\beta}_s) - Q(\beta_s^*)| = O_p\{(np_n^3 \log p_n)^{1/2}\}.$$

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 $n^{-1}U(\beta_s^*)^T \Omega(\beta_s^*)^{-1}U(\beta_s^*)$. According to Lemma 4, we have $2Q(\widehat{\beta}_s) - 2Q(\widehat{\beta}_T) = -\Delta_{s/T}$ $\{1 + o_p(1)\}$,with $\Delta_{s/T} = n^{-1}U(\beta_s^*)^T M_{s/T} U(\beta_s^*)$, where $M_{s/T}$ denotes the difference matrix $\Omega(\beta_s^*)^{-1} - D_s^T \Omega(\beta_T^*)^{-1} D_s$.

Lemma 6. *Under Assumptions 1 - 4, for overfitting model $s \in S_+$, $M_{s/T} = \Omega(\beta_s^*)^{-1} - D_s^T \Omega(\beta_T^*)^{-1} D_s$ 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 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 1 - 4,

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

Theorem 2. *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 1 - 4, as $n \to \infty$,*

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

In practice, the effective degree of freedom $d_s^* = \mathrm{Tr}\{W_s(\beta_s^*)\Omega_s^{-1}(\beta_s^*)\}$ is not known and we estimate it by $\widehat{d}_s = \mathrm{Tr}\{W_s(\widehat{\beta}_s)\Omega_s^{-1}(\widehat{\beta}_s)\}$. In the following lemma, we show the estimator is consistent for the unknown effective degrees of freedom.

Lemma 8. *Under Assumptions 1 - 4, as $n \to \infty$,*

$$|d_s^* - \widehat{d}_s| = O_p\{(p_n^5 \log p_n/n)^{1/2}\},$$

and the consistency result holds uniformly over the model space.

In light of this new Lemma, in Equation 2.4, if the effective degrees of freedom is replaced by its estimate, the model selection consistency of the criterion still holds true.

Corollary 1. *Under Assumptions 1 - 4, as* $n \to \infty$,

$$\Pr\{\min_{s \in S, s \neq T} GIC(s) > GIC(T)\} \to 1,$$

with $GIC(s) = 2Q(\widehat{\beta}_s) + \hat{d}_s \gamma_n$.

Through all the asymptotic discussions above, we rely on the full model with size p_n to obtain the consistent variance estimate $\hat{V}_i$. Alternatively, we can constrain the competing models to be bounded by size s_n and assume $s_n \ll p_n$. 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. However, with $p_n > n$, we cannot obtain the variance estimate under the full model. This is a common problem for model selection in high dimensional regression problem (Kim et al., 2012). If we can identify a set of s_n variables that includes all relevant variables with probability 1 asymptotically, we can obtain a consistent variance estimate at this model. This is the additional requirement for the relaxation of p_n to s_n.

Theorem 2 provides the asymptotic order for $\gamma_n = 6\omega(1+\log\log p_n)\log p_n$ to guarantee the model selection consistency. Given that ω is usually unknown and $\log\log p_n$ is rather small compared to $\log p_n$, we could choose different $\gamma_n = c \log p_n$, where c is a constant. Via empirical studies in Section 3 and Supplementary File, it is observed that $c = 1$ or $c = 2$ generates

the most satisfactory model selection results in the cases examined in our simulations, whereas for $c \geq 3$, the GIC tends to have lower Positive Selection Rate (PSR). In practice, we suggest to use penalty term as $\gamma = c \log p_n$, where $c = 1$ or 2.

When p_n is large, a full exhaustive search among all 2^{p_n} candidate models is computationally infeasible. Zhao and Yu (2006) established LASSO's (Tibshirani, 1996) variable selection consistency under the irrepresentable condition for linear regression. Wang et al. (2012) proposed a penalized GEE method using SCAD (Fan and Li, 2001) penalty and established its variable selection consistency. Thus the penalized GEE with LASSO or SCAD penalty can be used to generate different candidate models under a sequence of shrinkage parameter λ_n. However, the penalized methods depend on model selection criteria to choose the optimum penalty size. Given a specific penalty size, the penalized method can be used to generate a subset model. Using the proposed model selection criterion to evaluate different subset models, one can choose the subset model for which the criterion is minimized. Cross validation (Wang et al., 2007) can be used as an alternative model selection criterion, however, it is more computational intensive as it requires seperate steps of training and cross validation.

Section 2.5 illustrates that the generalized information criterion is selec-

tion consistent with the working correlation matrix R_i being any arbitrary positive definite matrix. Hence the selection consistency is robust against mis-specification of the working correlation. This matrix R_i needs to be fixed when we compare the generalized information criterion across different competing models. In practice, the choice of working correlation matrix R_i used in the criterion could impact its model selection efficiency. In our simulation, we compare different choices of R_i including independence, AR-1, compound symmetry, and unstructured working correlation. When the cluster size does not depends on i, Balan and Schiopu-Kratina (2005) suggested using the formula below to estimate the unstructured working correlation matrix

$$\widehat{R}_B = \frac{1}{n}\sum_{i=1}^{n} A_i^{-1/2}(\widetilde{\beta}_F)\{Y_i - \mu_i(\widetilde{\beta}_F)\}\{Y_i - \mu_i(\widetilde{\beta}_F)\}^T A_i^{-1/2}(\widetilde{\beta}_F), \qquad (2.7)$$

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.

For simplicity, we have assumed $\phi = 1$ in the paper. When ϕ is unknown, we can estimate it at the full model denoted as $\hat{\phi}_F$ (Pan, 2001) .

The residual quadratic form of equation 2.2 is rescaled as follows:

$$Q(\widehat{\beta}_s) = \frac{1}{2\hat{\phi}_F} \sum_{i=1}^{n} \{Y_i - \mu_i(\widehat{\beta}_s)\}^T A_i(\widehat{\beta}_F)^{-1/2} R^{-1} A_i(\widehat{\beta}_F)^{-1/2} \{Y_i - \mu_i(\widehat{\beta}_s)\}. \tag{2.8}$$

The proof of model selection consistency remains the same given $\hat{\phi}_F$ remains a constant across all candidate models.

3. Numerical Analysis

3.1 Simulations

We conduct simulations on clustered binary responses and clustered Gaussian responses. We consider different settings with sample size $n = 500$ or 1000, the number of covariates $p_n = 500$ or 1000, and the cluster size $m = 10$ or 20. The number of true covariates d_T is set be 50. For $j = 1, 2, \ldots, d_T$, 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 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 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 distribution or Gaussian distribution with mean μ_i and an unstructured correlation matrix. For each dataset, a common unstructured correlation matrix is used for all the clusters, whereas different datasets are simulated under different correlation matrices. The r package "SimCorMultRes" (Touloumis, 2019) is used to simulate the correlated multivariate binary distribution. We use the LASSO(Friedman et al., 2009) 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 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 c is varied 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 GIC.

In Table 1, we compare the PSR and FDR of different methods on multivariate normal responses. It is shown that our proposed method has high PSR and low FDR similar to those of the cross validation method . The advantage of our method is the computational simplicity whereas the cross validation is more computationally intensive and requires data partition

and separate steps of training and validation. We also compare our method with QIC, which has much higher FDR compared to our method and the cross validation method. This demonstrates that with large p_n, QIC tends to select overfitting models. This is due to the fact that QIC uses the AIC type of penalty, which is too small to control the error rate. Table 2 provides the comparison of the proposed GIC with other methods for multivariate binary responses, and the result is similar to the comparison on multivariate normal responses.

We vary the multiplicative factor of c from 1 to 4 and examine how the sensitivity and selectivity of our method changes. It is observed that when $c = 1$, or 2, the proposed GIC achieves high PSR and low FDR. When c increases, the GIC tends to have lower PSR and FDR as shown by Tables 1 and 2 in the Supplementary File. The PSR and FDR decrease faster with the increase of c in binary data than in normal data.

For the proposed GIC method, as the true correlation matrix is unstructured, the choice of unstructured working correlation matrix using the formula from Balan and Schiopu-Kratina (2005) outperforms the independent, exchangeable, and autoregression correlation matrices. As shown in Table 3 in the Supplementary File, the performance of the proposed GIC improves with higher PSR and lower FDR with increasing number of clus-

ters n or increasing number of cluster size m.

3.2 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 surveyed 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. 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 use the proposed model selection method to choose important predictors of the depression status of seniors. The missing value is imputed using median value for numerical variables and mode value for categorical variables. The LASSO method is used for generating the regularization path. We randomly split the data into two parts with 80% as the training set and 20% as the test set for five times. We use the GIC, QIC and cross validation methods to determine the best subset model. The QIC method chooses 71 variables with AUC 0.9114, the GIC method with $c = 1$ chooses 18 variables with AUC 0.9135, and the cross validation method chooses 20 variables with AUC 0.9135.

Table 1: The PSR and FDR of different methods for Normal Response

		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	Independent	1.0000	0.0000	0.6937	0.0326	1.0000	0.0000	0.5401	0.0607
	Exchangeable	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
	Unstructured	1.0000	0.0000	0.7281	0.0136	1.0000	0.0000	0.7109	0.0324
GIC	Independent	1.0000	0.0000	0.1077	0.0460	1.0000	0.0000	0.0871	0.0449
(c=1)	Exchangeable	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
	Unstructured	1.0000	0.0000	0.0226	0.0295	1.0000	0.0000	0.0272	0.0355
GIC	Independent	1.0000	0.0000	0.0012	0.0047	1.0000	0.0000	0.0008	0.0038
(c=2)	Exchangeable	1.0000	0.0000	0.0028	0.0095	1.0000	0.0000	0.0015	0.0065
	AR1	1.0000	0.0000	0.0012	0.0047	1.0000	0.0000	0.0008	0.0038
	Unstructured	1.0000	0.0000	0.0012	0.0047	1.0000	0.0000	0.0004	0.0027
GIC	Independent	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
(c=3)	Exchangeable	1.0000	0.0000	0.0004	0.0028	1.0000	0.0000	0.0000	0.0000
	AR1	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
	Unstructured	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
GIC	Independent	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
(c=4)	Exchangeable	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
	AR1	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
	Unstructured	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
CV	Independent	1.0000	0.0000	0.0034	0.0127	1.0000	0.0000	0.0041	0.0160
	Exchangeable	1.0000	0.0000	0.0034	0.0127	1.0000	0.0000	0.0049	0.0167
	AR1	1.0000	0.0000	0.0034	0.0127	1.0000	0.0000	0.0041	0.0159
	Unstructured	1.0000	0.0000	0.0139	0.0438	1.0000	0.0000	0.0217	0.0474

The true parameters size d_T is 50 and the cluster size m is 10. The free multiplicative constant c for the penalty is set to be 1, 2, 3 or 4. QIC denotes quasi-likelihood information criteria, GIC denotes generalized information criteria, and CV denotes cross validation.

Table 2: The PSR and FDR of different methods for Binary Response

		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	Independent	1.0000	0.0000	0.7093	0.0241	0.9974	0.0084	0.5677	0.0735
	Exchangeable	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
	Unstructured	1.0000	0.0000	0.7234	0.0179	0.9976	0.0082	0.6163	0.0677
GIC	Independent	0.9982	0.0081	0.0596	0.0476	0.9182	0.0710	0.0250	0.0548
	Exchangeable	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
	Unstructured	0.9990	0.0066	0.0445	0.0399	0.9498	0.0538	0.0242	0.0381

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. QIC denotes quasi-likelihood information criteria and GIC denotes generalized information criteria.

In comparison, GIC tends to select fewer variables than QIC with similar predictive power, and GIC has similar performance as CV in this data set.

4. Conclusion

We propose a generalized information criterion to select important covariates for GEE with diverging number of covariates. The proposed generalized information criterion is based on a goodness-of-fit measure which takes a quadratic form of the fitted residuals. The variable selection for

the mean model of GEE is robust to the mis-specification of the underlying correlation structure. This approach of constructing quadratic form as model fitting measure and using its large deviation properties to determine the appropriate penalty can be extended to other high dimensional model selection problems.

Our method is focused on the selection of mean models with fixed working correlation structure. Future research is needed to develop methods for the joint selection of mean and covariance structure with divergent number of covariates.

APPENDIX

In the following proofs, we assume $m_i = m$ for simplicity.

Proof of Theorem 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,

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

Let $\beta_s - \beta_s^* = \Delta(p_n^2 \log p_n/n)^{1/2} v$, 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 2, $-n^{-1}\mathrm{E}[U(\beta_s^*)^{(1)}] = \Omega(\beta_s^*)$, which is a positive definite matrix with bounded eigenvalues. From Lemma 1, the $(r,k)th$ entry of the matrix $n^{-1}[U(\beta_s^*)^{(1)} - \mathrm{E}\{U(\beta_s^*)^{(1)}\}]_{[rk]} = O_p\{(p_n \log p_n/n)^{1/2}\}$. 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 $U(\check{\beta}_s)^{(2)}_{[rk]} = \{U(\check{\beta}_s)^{(2)}_{[rk1]}, U(\check{\beta}_s)^{(2)}_{[rk2]}, \ldots, U(\check{\beta}_s)^{(2)}_{[rkd_s]}\}^T$ is a $d_s \times 1$ vector. Since $\check{\beta}_s$ is between $\widetilde{\beta}_s$ and β_s^*, $||\widetilde{\beta}_s - \beta_s^*|| = O_p\{(p_n^2 \log p_n/n)^{1/2}\}$. 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 1,$n^{-1}[U(\check{\beta}_s)^{(2)}_{[rkl]} - \mathrm{E}\{U(\check{\beta}_s)^{(2)}_{[rkl]}\}] = O_p\{(p_n \log p_n/n)^{1/2}\}$. This entails $n^{-1}||U(\check{\beta}_s)^{(2)}_{[rk]} - \mathrm{E}\{U(\check{\beta}_s)^{(2)}_{[rk]}\}|| = O_p\{(p_n^2 \log p_n/n)^{1/2}\}$. From Assumption 4, $\mathrm{E}\{U_i^{(2)}(\check{\beta}_s)_{[rkl]}\}$ is bounded. Then $n^{-1}||\mathrm{E}\{U(\check{\beta}_s)^{(2)}_{[rk]}\}|| = O_p(p_n^{1/2})$. This implies $n^{-1}\{U(\widetilde{\beta}_s)^{(1)} - U(\beta_s^*)^{(1)}\}_{[rk]} = O_p\{(p_n^3 \log p_n/n)^{1/2}\}$. Thus, $U(\widetilde{\beta}_s)^{(1)} = n\{\Omega(\beta_s^*) + Res\}$, and each element in the residual matrix Res is $O_p\{(p_n^3 \log p_n/n)^{1/2}\}$. For true and overfitting models, $\mathrm{E}\{U(\beta_s^*)\} = 0$. For underfitting models, based on the definition of β_s^*, it can be shown that

$\mathrm{E}\{U(\beta_s^*)\} = \mathrm{E}[\sum_{i=1}^n D_i(\beta_s^*)^T V_i(\beta_s^*)\{Y_i - \mu_i(\beta_s^*)\}] = \mathrm{E}[\sum_{i=1}^n D_i(\beta_s^*)^T V_i(\beta_s^*)$ $\{Y_i - \mu_i(\beta_T^*)\}] + \sum_{i=1}^n D_i(\beta_s^*)^T V_i(\beta_s^*)\{\mu_i(\beta_T^*) - \mu_i(\beta_s^*)\} = 0$ as well. From Lemma 1, we have $||U(\beta_s^*)|| = ||U(\beta_s^*) - \mathrm{E}\{U(\beta_s^*)\}|| = O_p\{(np_n^2 \log p_n)^{1/2}\}$. Thus there exists a constant number b_u such that $||U(\beta_s^*)|| \leq b_u(np_n^2 \log p_n)^{1/2}$ for n sufficiently large. In addition, we have

$$|v^T Res v| = |\sum_{kr} v_k v_r Res_{kr}| \leq \max_{kr} |Res_{kr}| \times p_n \times ||v||^2 = O_p\{(p_n^5 \log p_n/n)^{1/2}\} = o_p(1).$$

Combining the results above, we have

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

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

□

Proof of Theorem 2. First for overfitting models $s \in S_+$, we have

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

According to Lemma 7, $\Pr\{\max_{s\in S_+, s\neq T} \Delta_{s/T}/(d_s^* - d_T^*) > \gamma_n\} = o(1)$. Therefore $\Pr\{\min_{s\in S_+, s\neq T} \mathrm{GIC}(s) > \mathrm{GIC}(T)\} \to 1$. Next for the underfitting models, we have $\min_{s\in S_-} \mathrm{GIC}(s) - \mathrm{GIC}(T) = 2\{\min_{s\in S_-} Q(\widehat{\beta}_s) - Q(\widehat{\beta}_T)\} + (d_s^* - d_T^*)\gamma_n$. 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 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 5 implies $Q(\widehat{\beta}_T) - Q(\beta_T^*) = O_p(p_n^2 \log p_n)$ 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 S2.6, $\mathrm{E}\{n^{-1}\sum_{i=1}^{n}|Y_{ij} - \mu_{ij}(\beta_T^*)|\}$ is bounded. Based on Lemma S2.2 and Lemma 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.

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

Lemma S2.2 implies that $A_{ij}(\widehat{\beta}_F)$ is uniformly bounded from zero and infinity for all i and therefore $\lambda_{\min_i}(\widehat{V}_i^{-1})$ is a positive value bounded away from zero. Furthermore based on Assumption 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, $|d_s^* - d_T^*| = \omega|d_s - d_T| = O(p_n)$. So $\mathrm{E}\{Q(\beta_s^*) - Q(\beta_T^*)\}$ is the leading term in the difference between the two information criteria. Thus $\Pr\{\min_{s\in S, s\neq T} \mathrm{GIC}(s) > \mathrm{GIC}(T)\} \to 1$.

□

Supplementary Materials The online supplementary file contains the proofs of Lemmas 1 - 8 in the main paper and several technical lemmas. The file also provides some additional simulation results.

Acknowledgements

We are grateful to the referees, the associate editor and the editor for their insightful comments. Gao's research was supported by the Natural Sciences and Engineering Research Council of Canada. Carroll's research was supported by the National Cancer Institute.

References

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

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

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

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

Chen, J. and Z. Chen (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 R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc. 96*(456), 1348–1360.

Fan, J. and J. Lv (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B Stat. Methodol. 70*(5), 849–911.

Fan, J. and J. Lv (2011). Nonconcave penalized likelihood with np-dimensionality. *IEEE Trans. Inform. Theory 57*(8), 5467–5484.

Fang, E. X., Y. Ning, and R. Li (2020). Test of significance for high-dimensional longitudinal data. *The Annals of Statistics (in press)*.

Friedman, J., T. Hastie, and R. Tibshirani (2009). glmnet: Lasso and elastic-net regularized generalized linear models. *R package version 1*(4).

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

Gao, X. and P. X.-K. Song (2010). Composite likelihood bayesian information criteria for model selection in high-dimensional data. *J. Amer. Statist. Assoc. 105*(492), 1531–1540.

Kim, Y., S. Kwon, and H. Choi (2012). Consistent model selection criteria on high dimensions. *J. Mach. Learn. Res. 13*(Apr), 1037–1057.

Li, B. (1997). On the consistency of generalized estimating equations. *IMS Lecture Notes Monogr. Ser.*, 115–136.

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

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

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

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

Ning, Y. and H. Liu (2017). A general theory of hypothesis tests and confidence regions for

sparse high dimensional models. *The Annals of Statistics 45*(1), 158–195.

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. *Ann. Statist. 16*(1), 356–366.

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

Spokoiny, V. and M. Zhilova (2013). Sharp deviation bounds for quadratic forms. *Math. Methods Statist. 22*(2), 100–113.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*(1), 267–288.

Touloumis, A. (2019). Simcormultres: Simulates correlated binary responses assuming a regression model for the marginal probabilities. *R package version*.

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

Wang, H., R. Li, and C.-L. Tsai (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika 94*(3), 553–568.

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

Wang, L. and A. Qu (2009). Consistent model selection and data-driven smooth tests for longitudinal data in the estimating equations approach. *J. R. Stat. Soc. Ser. B Stat. Methodol. 71*(1), 177–190.

Wang, L., J. Zhou, and A. Qu (2012). Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics 68*(2), 353–360.

White, H. (1981). Consequences and detection of misspecified nonlinear regression models. *J. Amer. Statist. Assoc. 76*(374), 419–433.

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

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

Zhang, Y. and X. Shen (2010). Model selection procedure for high-dimensional data. *Statistical Analysis and Data Mining: The ASA Data Science Journal 3*(5), 350–358.

Zhao, P. and B. Yu (2006). On model selection consistency of lasso. *Journal of Machine learning research 7*(Nov), 2541–2563.

Department of Mathematics and Statistics York University, Toronto, ON M3J 1P3, Canada

E-mail: (scheng.wu@gmail.com)

Department of Mathematics and Statistics York University, Toronto, ON M3J 1P3, Canada

E-mail: (xingao@mathstat.yorku.ca)

Department of Statistics, Texas A&M University, College Station, Texas, 77843-3143 USA

E-mail: (carroll@stat.tamu.edu)