Tuning parameter selection for penalized empirical likelihood with a diverging number of parameters

Chaowen Zheng

Department of Statistics, North Carolina State University, Raleigh, United States.

E-mail: czheng6@ncsu.edu

and Yichao Wu

Department of Mathematics, Statistics and Computer Science, University of Illinois at Chicago, Chicago, United States.

E-mail: yichaowu@uic.edu

Summary. Penalized likelihood methods have been a success in analyzing high dimensional data. Tang and Leng (2010) extended the penalization approach to the empirical likelihood scenario and showed that the penalized empirical likelihood estimator can identify the true predictors consistently in the linear regression models. However, this desired selection consistency property of the penalized empirical likelihood method relies heavily on the choice of tuning parameter. In this work, we propose a tuning parameter selection procedure for penalized empirical likelihood to guarantee that this selection consistency can be achieved. Specifically, we propose a generalized information criterion (GIC) for the penalized empirical likelihood in the linear regression case. We show that the tuning parameter selected by the GIC yields the true model consistently even when the number of predictors diverges to infinity with the sample size. We demonstrate the performance of our procedure by numerical simulations and a real data analysis.

1. Introduction

Empirical likelihood (EL) proposed by Owen (1991) has been a great success as a nonparametric likelihood approach. Not only does empirical likelihood enjoy the reliability of nonparametric methods, but also achieves the effectiveness of the likelihood methods. Especially, it turns out appealing in constructing confidence regions, formulating

goodness-of-fit tests and incorporating auxiliary information. For a comprehensive review of applications of EL method, interested readers can refer to Chen and Van Keilegom (2009).

Recently, penalized likelihood method has been extensively studied for dealing with high dimensional data (Fan and Li, 2001; Fan et al., 2004; Fan and Lv, 2011). Extending the regularization approach to the EL method, Tang and Leng (2010) proposed a penalized empirical likelihood (PEL). In Tang and Leng (2010), they established the oracle property of the PEL estimator for the high dimensional linear regression case that allows the dimensionality of the parameter p diverges with the sample size. And they applied the PEL approach to construct confidence regions and to facilitate hypothesis testing by showing that the profiled PEL ratio follows χ^2 distribution asymptotically. Furthermore, Leng and Tang (2012) extended the PEL approach to general estimating equation case with a diverging dimensionality. For more discussion on PEL approach, interested readers can refer to Lahiri et al. (2012), Chang et al. (2015) and Chang et al. (2018).

Obviously, the oracle property of the PEL estimator depends on the choice of tuning parameter. There is a rich body of literature for the tuning parameter selection in penalized likelihood methods. The most commonly used methods are cross-validation and information criterion, such as Akaike information criterion (AIC) (Akaike, 1973), and Bayes information criterion (BIC) (Schwarz et al., 1978). Wang et al. (2007) showed that the tuning parameter obtained by minimizing BIC can identify the true model with probability tending to 1. However, their results only applies to the fixed dimensionality. Wang et al. (2009) modified the BIC to deal with the high dimensional case, but their analysis can only deal with the penalized least square method. Wang and Zhu (2011) also proposed a family of high dimensional Bayesian Information Criterion, HBIC for tuning parameter selection in ultra-high dimensional situations. Recently, Fan and Tang (2013) proposed a generalized information criterion to select the tuning parameter in high dimensional penalized likelihood which is limited to parametric models.

The information criterion mentioned above can be summarized as follows,

$$\text{a measure of model fitting} + C_n \times \text{measure of model complexity},$$

where C_n is a positive sequence depending on sample size n that controls the balance between model fitting and model complexity. In parametric models, a common choice of measure of model fitting is the minus log-likelihood. Similarly, we propose a generalized information criterion (GIC) for PEL by choosing minus log of empirical likelihood as our measure of model fitting. And with a carefully selected C_n, we show that the tuning parameter selected by our GIC can identify the true model consistently. In this paper, we mainly focus on the high dimensional linear regression problems, that is, we allow the dimensionality to diverge to infinity as the sample size goes to infinity. Without any change, our proposed GIC can be extended to handle the tuning parameter selection in penalized empirical likelihood for general estimating equations with growing dimensionality. However, it is much more challenging to establish the selection coistency for this case and it merits further investigation.

The rest of this paper is organized as follows. Section 2 defines the GIC for tuning parameter selection in the PEL for linear models and presents the asymptotic properties of our proposed GIC. Numerical studies are conducted in Section 3 to demonstrate our theoretical findings. Some discussions are given in Section 4. All proofs are presented in Appendix.

2. GIC for penalized empirical likelihood

2.1. Penalized empirical likelihood estimator for linear model

We consider the following linear model:

$$Y_i = \mathbf{X}_i^T \boldsymbol{\beta} + \epsilon_i,\ i = 1, ..., n, \tag{1}$$

where $\mathbf{X}_i \in \mathbb{R}^p$ is the predictor vector with p denoting the dimensionality, $\beta \in \mathbb{R}^p$ is the regression coefficient, and ϵ_i is the error term with $\mathrm{E}(\epsilon_i) = 0$ and $\mathrm{var}(\epsilon_i) = \sigma^2$. We denote the true regression coefficient as $\boldsymbol{\beta}_0 = (\beta_{01}, \cdots, \beta_{0p})^T$. Without loss of generality, we assume $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{10}^T, 0, ..., 0)^T$ where $\boldsymbol{\beta}_{10} \in \mathbb{R}^d$ corresponds to the non-zero coefficients and d denotes the number of non-zero coefficients. In other words, we assume that only the first d predictors are included in the true model. In this paper, we allow the dimension of the predictors, p diverge to ∞ as $n \to \infty$. Also, we allow that the dimension of the true model, namely d, to diverge at the same rate of p as $n \to \infty$.

Following Tang and Leng (2010), we assume that $\{\mathbf{X}_i\}_{i=1}^n$ are independently and identically distributed(iid) random vectors from the following model,

$$\mathbf{X}_i = \mathbf{\Gamma}\mathbf{Z}_i, \tag{2}$$

where $\mathbf{\Gamma}$ is a $p \times m$ matrix with $m \geq p$ and $\mathbf{\Gamma}\mathbf{\Gamma}^T = \mathbf{\Sigma}$, and $\mathbf{Z}_i \in \mathbb{R}^m$ satisfies $\mathrm{E}(\mathbf{Z}_i) = \mathbf{0}, \mathrm{cov}(\mathbf{Z}_i) = \mathbf{I}_m, \mathrm{E}(Z_{il})^{4k} = m_{4k} \in (0, \infty)$ and $\mathrm{E}(Z_{il_1}^{\alpha_1} Z_{il_2}^{\alpha_2} \cdots Z_{il_q}^{\alpha_q}) = \mathrm{E}(Z_{il_1}^{\alpha_1})\mathrm{E}(Z_{il_2}^{\alpha_2}) \cdots \mathrm{E}(Z_{il_q}^{\alpha_q})$, $\sum_{l=1}^q \alpha_l \leq 4k$ for some positive integer k and $l_1 \neq l_2 \neq \cdots \neq l_q$. Here $\mathbf{I}_m$ denotes the m-dimensional identity matrix. This model is commonly used in high dimensional EL literature such as Chen et al. (2009) and Tang and Leng (2010) .

Owen (1991) proposed the empirical likelihood (EL) for linear models based on moment equations. Let $\mathbf{U}_i(\boldsymbol{\beta}) = \mathbf{X}_i(Y_i - \mathbf{X}_i^T\boldsymbol{\beta})$. The EL of $\boldsymbol{\beta}$ is defined as

$$L(\boldsymbol{\beta}) = \sup\{\prod_{i=1}^n w_i : w_i \geq 0, \sum_{i=1}^n w_i = 1, \sum_{i=1}^n w_i \mathbf{U}_i(\boldsymbol{\beta}) = \mathbf{0}\}.$$

Using Lagrange multiplier method, let $\boldsymbol{\lambda}_{\boldsymbol{\beta}}$ denote the solution to $\frac{1}{n}\sum_{i=1}^n \frac{\mathbf{U}_i(\boldsymbol{\beta})}{1+\boldsymbol{\lambda}_{\boldsymbol{\beta}}^T\mathbf{U}_i(\boldsymbol{\beta})} = \mathbf{0}$. Then we have that $w_i = \frac{1}{n}\frac{1}{1+\boldsymbol{\lambda}_{\boldsymbol{\beta}}^T\mathbf{U}_i(\boldsymbol{\beta})}$. As a result, we have the following expression for the log empirical likelihood,

$$\log(L(\boldsymbol{\beta})) = -n\log(n) - \sum_{i=1}^n \log(1 + \boldsymbol{\lambda}_{\boldsymbol{\beta}}^T\mathbf{U}_i(\boldsymbol{\beta})). \tag{3}$$

We define $\ell_c(\boldsymbol{\beta}) = -\log(L(\boldsymbol{\beta})) - n\log(n) = \sum_{i=1}^n \log(1 + \boldsymbol{\lambda}_{\boldsymbol{\beta}}^T\mathbf{U}_i(\boldsymbol{\beta}))$. Following Tang and Leng (2010), the penalized empirical likelihood(PEL) estimator $\hat{\boldsymbol{\beta}}$ is defined to be the minimizer of

$$\ell_p(\boldsymbol{\beta}) = \ell_c(\boldsymbol{\beta}) + n\sum_{i=1}^p p_\tau(|\beta_i|) = \sum_{i=1}^n \log\{1 + \boldsymbol{\lambda}_{\boldsymbol{\beta}}^T\mathbf{U}_i(\boldsymbol{\beta})\} + n\sum_{i=1}^p p_\tau(|\beta_i|), \tag{4}$$

where $p_\tau(t)$ is a penalty function with tuning parameter τ. In this paper, we adopt the smoothly clipped absolute deviation(SCAD) penalty function with tuning parameter τ which is symmetric with the first derivative $p_\tau'(t) = \tau\{I(t \leq \tau) + \frac{(a\tau - t)_+}{(a-1)\tau}I(t > \tau)\}$ for $t > 0$ with $a = 3.7$. The SCAD penalty function proposed in Fan and Li (2001) has been widely used in variable selection since it can shrink the estimates of unimportant coefficients to zero while yielding unbiased estimates for the important coefficients. For

a detailed explanation and application of SCAD penalty, interested readers can refer to Fan and Li (2001), Kim et al. (2008) and Fan and Li (2002).

Under appropriate regularity conditions, Tang and Leng (2010) showed that the PEL estimator $\hat{\boldsymbol{\beta}}$ can achieve selection consistency and asymptotic efficiency. Moreover, the PEL estimator enjoys the oracle properties which were first introduced by Fan and Li (2001). However, the oracle properties of the PEL estimator depend heavily on the choice of the tuning parameter. In order to guarantee the oracle properties of the PEL estimator, it is essential to propose an appropriate criterion to select suitable tuning parameter.

2.2. GIC tuning parameter selector

Before we propose our tuning parameter selector, we introduce the following notations first. Let $\alpha = \{j_1, ..., j_{p^*}\}$ denote a candidate model with predictors $X_{j_1}, ..., X_{j_{p^*}}$. For each candidate model α, we denote its model size as df_α. For each tuning parameter τ in the penalized empirical likelihood (3), we denote the PEL estimator for the coefficients as $\hat{\boldsymbol{\beta}}_\tau$. For each estimator $\hat{\boldsymbol{\beta}}_\tau$, we denote the corresponding model as $\alpha_\tau = \{j : (\hat{\boldsymbol{\beta}}_\tau)_j \neq 0\}$ and its model size as df_{α_τ}, where $(\hat{\boldsymbol{\beta}}_\tau)_j$ denotes the jth component of $\hat{\boldsymbol{\beta}}_\tau$. And we denote the full model as $\bar{\alpha} = \{1, \ldots, p\}$ and the true model as $\alpha_0 = \{1, \ldots, d\}$. In addition, we use $\mathscr{A}$ to denote the collection of all candidate models.

For each model α, we can obtain a non-penalized estimate for the regression coefficients, $\hat{\boldsymbol{\beta}}^*_\alpha$ by minimizing $\ell_c(\boldsymbol{\beta})$ subject to the the constraints $\beta_j = 0 \quad \forall j \notin \alpha$. Thus, for each selected model α_τ, we denote the corresponding non-penalized estimate for the regression coefficients as $\hat{\boldsymbol{\beta}}^\star_{\alpha_\tau}$.

Following Tang and Leng (2010), we propose the following Generalized Information Criterion(GIC),

$$\begin{aligned} \mathrm{GIC}(\hat{\boldsymbol{\beta}}) &= 2\ell_c(\hat{\boldsymbol{\beta}}) + C_n \cdot df_{\hat{\boldsymbol{\beta}}}, \\ \mathrm{GIC}(\tau) &= 2\ell_c(\hat{\boldsymbol{\beta}}_\tau) + C_n \cdot df_{\hat{\boldsymbol{\beta}}_\tau}, \end{aligned} \tag{5}$$

where $\hat{\boldsymbol{\beta}}$ is the parameter estimator and $df_{\hat{\boldsymbol{\beta}}}$ is the corresponding degrees of freedom associated with $\hat{\boldsymbol{\beta}}$. Let C_n be some positive constant to be discussed more carefully. Note

that $\hat{\boldsymbol{\beta}}_\tau$ is the penalized empirical likelihood estimator for the regression coefficients. The first term $2\ell_c(\hat{\boldsymbol{\beta}})$ in (5) evaluates the goodness of fit while the second term penalizes the model complexity. In other words, the GIC trades off between the model complexity and goodness of fit by using a proper C_n. Such criterion was first introduced by Akaike (1974) and Schwarz et al. (1978). Wang et al. (2009) modified the BIC for the high dimensional case. And Fan and Tang (2013) generalized such information criterion to the likelihood based method with diverging number of parameters. We shall justify the GIC in the case of penalized empirical likelihood in this paper.

2.3. Theoretical Property

In this section, we show that the above GIC can identify the true model consistently with an appropriately chosen C_n. First, we assume that the true model is unique and that the regression coefficients for the true model α_0 are nonzero. Thus, we say that each candidate model $\alpha \supset \alpha_0$ is an overfitted model while each candidate model $\alpha \not\supseteq \alpha_0$ is an underfitted model. As a result, we divide the tuning parameters into three separate sets as follows,

$$\Omega_- = \{\tau : \alpha_\tau \not\supseteq \alpha_0\}, \Omega_0 = \{\tau : \alpha_\tau = \alpha_0\}, \Omega_+ = \{\tau : \alpha_\tau \supset \alpha_0\}.$$

We introduce the following technical assumptions.

(E1) The $\{\mathbf{X}_i\}_{i=1}^n$ are *i.i.d.* from the model (2) for some $k \geq 3$. The errors $\{\epsilon_i\}$ are *i.i.d.* with mean 0 and $\mathrm{E}(\epsilon_i^{4k}) < \infty$ for the same k.

(E2) $\gamma_{min}(\boldsymbol{\Sigma}) \geq C_1$ and $\gamma_{max}(\boldsymbol{\Sigma}) \leq C_2$ for some $C_2 > C_1 > 0$, where γ_{min} denotes the minimum eigenvalue and γ_{max} denotes the maximum eigenvalue.

(E3) $p \to \infty$, $p^2/n^{1-1/(4k)} \to 0$, $p^{1-2\delta}/n^{1/2-2\delta} \to 0$ as $n \to \infty$ for the δ to be specified in statement (S1) in the proof of Lemma 2 in the Appendix.

(E4) There exits a constant h such that the penalty $p_\tau(\xi)$ satisfies $p_\tau^{'}(\xi) = 0$ for $\xi > h\tau$.

(E5) For the underfitted model $\alpha \not\supseteq \alpha_0$, there exits a constant $C_- > 0$ such that $\liminf_{n\to\infty}\{\min_{\alpha\not\supseteq\alpha_0} \frac{1}{n}\ell_c(\hat{\boldsymbol{\beta}}_\alpha^\star)\} \geq C_-$ with probability tending to 1.

(E6) For the $\{\mathbf{Z}_i\}_{i=1}^n$ in model (2), $\mathbf{B}_i = \mathbf{Z}_i\epsilon_i$ are sub-Gaussian random vectors with variance proxy $\sigma^2 > 0$. We briefly introduce the concept of sub-Gaussian here.

A random variable $T \in \mathbb{R}$ is said to be sub-Gaussian if $\mathrm{E}(T) = 0$ and $\mathrm{E}[\exp(sT)] \leq \exp(\sigma^2 s^2/2),\ \forall s \in \mathbb{R}$, for some $\sigma^2 > 0$. In this case, we wirte $T \sim \mathrm{subG}(\sigma^2)$. And we say a random vector $\mathbf{T} \in \mathbb{R}^m$ is sub-Gaussian with variance proxy σ^2, if $\mathrm{E}(\mathbf{T}) = \mathbf{0}$ and $\mathbf{u}^T\mathbf{T}$ is a sub-Gaussian random variable with variance proxy σ^2 for any vector $\mathbf{u} \in \mathbb{R}^m$ with $\mathbf{u}^T\mathbf{u} = 1$, which we denote as $\mathbf{T} \sim \mathrm{subG}_m(\sigma^2)$. For a comprehensive review of sub-Gaussian distribution, interested readers can refer to Rivasplata (2012).

Here, Assumption (E1)-(E3) are from Tang and Leng (2010). Briefly speaking, Assumption (E1) characterizes the tail probability behaviors of $\mathbf{X}_i$ which is satisfied by the elliptical contoured distribution and the Gaussian family. Assumption (E3) controls the rate at which the dimension of $\mathbf{X}_i$ is allowed to diverge. Assumption (E4) controls the effect of the penalty such that the penalized estimator is asymptotically unbiased. And it is easy to verify that the SCAD penalty satisfies Assumption (E4). Assumption (E5) assures that the underfitted model yields a larger model deviance than that of the true model. Assumption (E6) further controls the tail behavior of the estimating equation $\mathbf{U}_i(\boldsymbol{\beta}) = \mathbf{X}_i(Y_i - \mathbf{X}_i^T\boldsymbol{\beta})$. With these assumptions, we can now present our main result on asymptotic performance of GIC selector.

Theorem 1: Suppose that the data are generated according to the model (1) and that Assumptions (E1)-(E6) hold. If C_n satisfies $\frac{C_n}{\log(p)} \to \infty$ and $\frac{C_n}{\sqrt{n}} \to 0$ as $n \to \infty$, we have that the tuning parameter $\hat{\tau}$ obtained by minimizing $\mathrm{GIC}(\tau)$ satisfies $P(\alpha_{\hat{\tau}} = \alpha_0) \to 1$.

To better understand our main result, we give a sketch for the proof of Theorem 1 in the rest of this section. The details of the proof can be found in the Appendix.

First, we introduce the following lemma.

Lemma 1: Suppose that the data are generated according to the model (1) and that Assumptions (E1)-(E4) hold. Let τ_n be a sequence of tuning parameters satisfying that $\tau_n \to 0$, $\tau_n(n/p)^{1/2-\delta} \to \infty$ for the δ specified in Assumption (E3) and that $\min_{1\leq j\leq d}|\beta_{0j}|/\tau_n \to \infty$. Then we have that $\hat{\boldsymbol{\beta}}_{\tau_n} = \hat{\boldsymbol{\beta}}^{\star}_{\alpha_0}$ with probability tending to 1.

Lemma 1 assures that there exists a tuning parameter sequence τ_n such that $\alpha_{\tau_n} = \alpha_0$ with probability tending to 1. Thus, it remains to show that $P\{\inf_{\tau\in(\Omega_-\cup\Omega_+)}\mathrm{GIC}(\hat{\boldsymbol{\beta}}_\tau) > \mathrm{GIC}(\hat{\boldsymbol{\beta}}_{\tau_n})\} \to 1$.

By the definition of $\hat{\boldsymbol{\beta}}^*_{\alpha_\tau}$, we have that $\ell_c(\hat{\boldsymbol{\beta}}_\tau) \geq \ell_c(\hat{\boldsymbol{\beta}}^*_{\alpha_\tau})$. Therefore, we have that $\mathrm{GIC}(\hat{\boldsymbol{\beta}}_\tau) - \mathrm{GIC}(\hat{\boldsymbol{\beta}}_{\tau_n}) \geq 2\ell_c(\hat{\boldsymbol{\beta}}^\star_{\alpha_\tau}) - 2\ell_c(\hat{\boldsymbol{\beta}}^\star_{\alpha_0}) + C_n(df_{\alpha_\tau} - df_{\alpha_0})$ *as* $n \to \infty$ with probability tending to 1. Thus, the key to our problem is to characterize the asymptotic behavior of $\ell_c()$.

For the true and overfitted model $\alpha \supseteq \alpha_0$, we have the following lemma.

Lemma 2: Suppose that the data are generated according to the model (1) and that Assumptions (E1)-(E3) hold. Then, for any $\alpha \supseteq \alpha_0$, we have

$$2\ell_c(\hat{\boldsymbol{\beta}}^\star_\alpha) = n\bar{\mathbf{U}}^T\boldsymbol{\Sigma}^{-1}\mathbf{H}_\alpha^T(\mathbf{H}_\alpha\boldsymbol{\Sigma}^{-1}\mathbf{H}_\alpha^T)^{-1}\mathbf{H}_\alpha\boldsymbol{\Sigma}^{-1}\bar{\mathbf{U}} + o_p(1),$$

where $\mathbf{H}_\alpha$ is a $(p - df_\alpha) \times p$ matrix such that $(\mathbf{H}_\alpha)_{i,j_i} = 1$ for $i = 1, \ldots, p - df_\alpha$ with $\{j_i\}_{i=1,\ldots,p-df_\alpha} = \bar{\alpha} - \alpha$ and the other entries of $\mathbf{H}_\alpha$ are all zero, and $\bar{\mathbf{U}} = n^{-1}\sum_{i=1}^n \mathbf{X}_i\epsilon_i$.

Therefore, we have that $2\ell_c(\hat{\boldsymbol{\beta}}^\star_\alpha) - 2\ell_c(\hat{\boldsymbol{\beta}}^\star_{\alpha_0}) = n\bar{\mathbf{U}}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{A}_\alpha - \mathbf{A}_{\alpha_0})\boldsymbol{\Sigma}^{-\frac{1}{2}}\bar{\mathbf{U}} + o_p(1)$ for any $\alpha \supseteq \alpha_0$, where $\mathbf{A}_\alpha = \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{H}_\alpha^T(\mathbf{H}_\alpha\boldsymbol{\Sigma}^{-1}\mathbf{H}_\alpha^T)^{-1}\mathbf{H}_\alpha\boldsymbol{\Sigma}^{-\frac{1}{2}}$ and $\mathbf{A}_{\alpha_0} = \boldsymbol{\Sigma}^{-\frac{1}{2}}\mathbf{H}_{\alpha_0}^T(\mathbf{H}_{\alpha_0}\boldsymbol{\Sigma}^{-1}\mathbf{H}_{\alpha_0}^T)^{-1}\mathbf{H}_{\alpha_0}\boldsymbol{\Sigma}^{-\frac{1}{2}}$ are two projection matrices. In the Appendix, we show that $\min_{\alpha\supset\alpha_0} \frac{n\bar{\mathbf{U}}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{A}_\alpha - \mathbf{A}_{\alpha_0})\boldsymbol{\Sigma}^{-\frac{1}{2}}\bar{\mathbf{U}}}{df_\alpha - df_{\alpha_0}} = O_p(\log(p))$ which implies that with probability tending to 1, $\min_{\alpha\supset\alpha_0}\{2\ell_c(\hat{\boldsymbol{\beta}}^\star_\alpha) - 2\ell_c(\hat{\boldsymbol{\beta}}^\star_{\alpha_0}) + C_n(df_\alpha - df_{\alpha_0})\} > 0$ for C_n satisfying $\frac{C_n}{\log(p)} \to \infty$. As a result, we conclude that $P\{\inf_{\tau\in\Omega_+} \mathrm{GIC}(\hat{\boldsymbol{\beta}}_\tau) > \mathrm{GIC}(\hat{\boldsymbol{\beta}}_{\tau_n})\} \to 1$.

For any underfitted model, $\alpha \not\supseteq \alpha_0$, the Assumption (E5) ensures that $\min_{\alpha\not\supseteq\alpha_0} \frac{2\ell_c(\hat{\boldsymbol{\beta}}^\star_\alpha) - 2\ell_c(\hat{\boldsymbol{\beta}}^\star_{\alpha_0})}{n} \geq C_- - \frac{2\ell_c(\hat{\boldsymbol{\beta}}^\star_{\alpha_0})}{n}$ with probability tending to 1. We show in the Appendix that $\frac{-2\ell_c(\hat{\boldsymbol{\beta}}^\star_{\alpha_0})}{n} = o_p(1)$. Together with the fact that $C_n p/n \to 0$, we conclude that with probability tending to 1, $P\{\inf_{\tau\in\Omega_-} \mathrm{GIC}(\hat{\boldsymbol{\beta}}_\tau) > \mathrm{GIC}(\hat{\boldsymbol{\beta}}_{\tau_n})\} \geq P\{\min_{\alpha\not\supseteq\alpha_0} \frac{2\ell_c(\hat{\boldsymbol{\beta}}^\star_\alpha)}{n} - \frac{2\ell_c(\hat{\boldsymbol{\beta}}^\star_{\alpha_0})}{n} - \frac{C_n df_{\alpha_0}}{n} > 0\} \geq P\{C_- + o_p(1) > 0\} \to 1$.

In the above procedure, we give a sketch of the proof for the statement that $P\{\inf_{\tau\in(\Omega_-\cup\Omega_+)} \mathrm{GIC}(\hat{\boldsymbol{\beta}}_\tau) > \mathrm{GIC}(\hat{\boldsymbol{\beta}}_{\tau_n})\} \to 1$. Together with Lemma 1, Theorem 1 can be established easily as shown in the Appendix.

3. Numerical study

In this section, we conduct several simulation studies and analyze one real data problem to confirm our theoretical findings. For all computation issues in PEL, we follow the strategy in Tang and Leng (2010). For comparison purpose, we define an AIC-like

Table 1. Summary of results for Example 1

	Method	n=100	n=200	n=400	n=800	n=1600
MRME	AIC	1.00	0.76	0.65	0.71	0.63
	GIC	1.00	0.73	0.43	0.34	0.23
CM	AIC	0.33	0.46	0.52	0.35	0.32
	GIC	0.39	0.58	0.91	0.95	0.99
MS	AIC	5.39	5.74	6.22	7.06	7.67
	GIC	4.99	5.05	5.09	5.05	5.03

criterion, $\text{AIC}(\hat{\boldsymbol{\beta}}) = 2\ell_c(\hat{\boldsymbol{\beta}}) + 2df_{\hat{\boldsymbol{\beta}}}$. And we compare our GIC with AIC to demonstrate the superior performance of our tuning parameter selector. For all the simulation studies, we compute the median of the relative model error (MRME), the average model size (MS) and the percentage of the correctly identified true models (CM). Basically, a smaller MRME means more accurate prediction results. Interested readers can refer to Fan and Li (2001) and Wang et al. (2009) for more detailed explanation of MRME, MS and CM.

In our simulation studies, we directly adopted the settings in Wang et al. (2009). Specifically, the covariate $\mathbf{X}_i$ follows a multivariate normal distribution $N_p(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = (\sigma_{j_1,j_2}), \sigma_{j_1,j_2} = 1$ if $j_1 = j_2$ and $\sigma_{j_1,j_2} = 0.5$ if $j_1 \neq j_2$. The error term $\{\epsilon_i\}_{i=1}^n$ independently follows the standard Gaussian distribution $N(0,1)$. And we choose the C_n in our GIC criterion to be $\log(\log(p)) * \log(n)$ for all simulation studies. In addition, we repeat all the experiments 100 times.

3.1. Example 1.

In this example, we take $p = [4n^{1/4}] - 5$, $\boldsymbol{\beta} = (11/4, -23/6, 37/12, -13/9, 1/3, 0, \ldots, 0)^T \in \mathbb{R}^p$ where $[t]$ denotes the largest integer that is smaller or equal to t. It impies that the dimension of the true model is fixed to be 5.

The results are depicted in the left panels of Figure 1. For illustration purpose, we also present the results in Table 1. We can see that the GIC method approaches 100% CM quickly while AIC can not identify the true model consistently, which confirms our theoretical findings. Therefore, the MRME values corresponding to GIC method are smaller than those of AIC method.

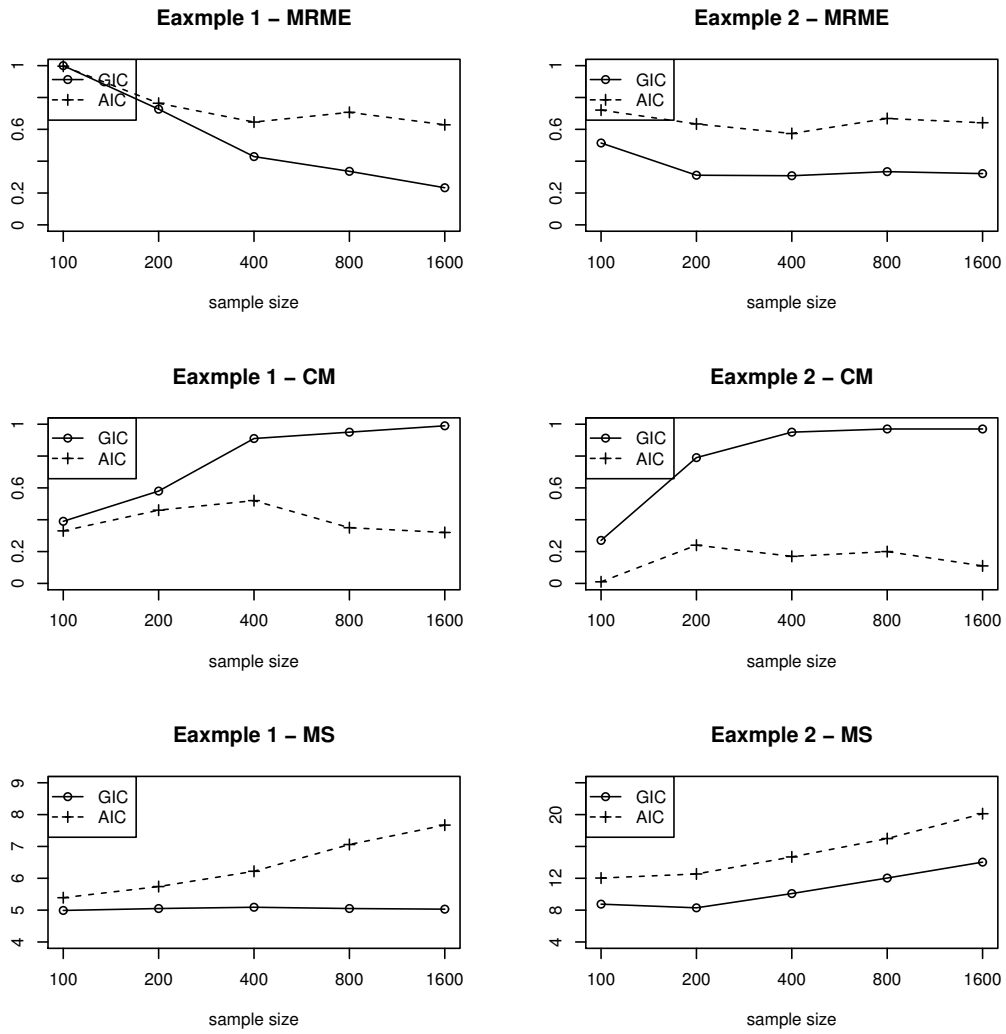

Fig. 1. Summary of simulation results for Example 1 and 2

Table 2. Summary of results for Example 2

	Method	n=100	n=200	n=400	n=800	n=1600
MRME	AIC	0.72	0.63	0.57	0.67	0.64
	GIC	0.51	0.31	0.31	0.33	0.32
CM	AIC	0.01	0.24	0.17	0.20	0.11
	GIC	0.27	0.79	0.95	0.97	0.97
MS	AIC	12.03	12.54	14.68	16.99	20.11
	GIC	8.75	8.29	10.08	12.03	14.03

3.2. Example 2.

In contrast to example 1, we are allowing the dimension of the true model to diverge. Specifically, we set $p = [7n^{1/4}]$ and $d = [p/3]$. And we generate the true regression coefficients from the uniform distribution over $[0.5, 1.5]$. The results are depicted in the right panels of Figure 1. For illustration purpose, we also present the results in Table 2. As one can see, the results are quite similar to Simulation 1 which confirms our theoretical findings again.

3.3. Real data analysis

In our real data example, we re-explore the Fifth National Bank's employee salary data analyzed by Fan et al. (2004) and Wang et al. (2009). The main goal here is to estimate the salary difference between male and female employees. Following Fan et al. (2004), we adopt their semiparametric model $\text{Salary} = \beta_0 + \beta_1 * \text{Female} + \beta_2 \text{PCJob} + \sum_{i=1}^{4} \beta_{2+i} \text{Edu}_i + \sum_{i=1}^{5} \beta_{6+i} \text{JobGrd}_i + f_1(\text{YrsExp}) + f_2(\text{Age}) + \epsilon$ where $f_i(x) = \alpha_{i1}x + \alpha_{i2}x^2 + \alpha_{i3}(x - x_{i1})_+^2 + \cdots + \alpha_{i7}(x - x_{i5})^2$ are parameterized continuous functions, $\{x_{i1}, x_{i2}, ..., x_{i5}\}$ denotes the $\{2/7, 3/7, ..., 6/7\}$ quantiles of the empirical distribution of the variables YrsExp ($i = 1$) and Age ($i = 2$). Thus, there are a total of 26 predictors whose corresponding coefficients are $\{\beta\}_{i=0}^{11}, \{\alpha_{1j}\}_{j=1}^{7}, \{\alpha_{2j}\}_{j=1}^{7}$ in this model. A detailed explanation of the variates and the parameterized functions f_1, f_2 in this model can be found in Fan et al. (2004). As suggested by Fan et al. (2004), we deleted the observations with working experience over 30 or age over 60, that results in a sample of size 199. Then we apply the PEL approach to estimate the coefficients with tuning parameters selected by our proposed GIC and

Table 3. Analysis result of the Bank Salary Dataset

Method	OLS	SCAD-AIC	SCAD-GIC
Female	-0.940	-0.885	0.000
PCJob	3.685	3.842	3.951
Edu1	-1.750	-1.426	0.000
Edu2	-3.134	-2.975	-2.040
Edu3	-2.277	-1.936	-1.189
Edu4	-2.112	-1.358	0.000
JobGrd1	-22.910	-22.995	-24.869
JobGrd2	-21.084	-21.180	-22.502
JobGrd3	-17.197	-17.460	-18.996
JobGrd4	-12.837	-12.985	-13.767
JobGrd5	-7.604	-7.808	-8.372

AIC. The detailed results are presented in Table 1. It is clear that the model selected by GIC is more sparse than that selected by AIC which is consistent with our theoretical findings. Consequently, only the tuning parameter selected by GIC identifies the gender as an irrelevant predictor which is accordance with the conclusion obtained by Fan et al. (2004) and Wang et al. (2009).

Acknowledgement

We thank two reviewers, an associate editor, and the editor for their most helpful comments that lead to substantial improvements in the paper. Wu is partially supported by NSF grants DMS-1821171 and CCF-1934915.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proceedings of the 2nd International Symposium on Information Theory*, pages 267–281. Akadémiai Kiadó.

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

Chang, J., Chen, S. X., and Chen, X. (2015). High dimensional generalized empirical likelihood for moment restrictions with dependent data. *Journal of Econometrics*, 185(1):283–304.

Chang, J., Tang, C. Y., Wu, T. T., et al. (2018). A new scope of penalized empirical likelihood with high-dimensional estimating equations. *The Annals of Statistics*, 46(6B):3185–3216.

Chen, S. X., Peng, L., and Qin, Y.-L. (2009). Effects of data dimension on empirical likelihood. *Biometrika*, 96(3):711–722.

Chen, S. X. and Van Keilegom, I. (2009). A review on empirical likelihood methods for regression. *Test*, 18(3):415–447.

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 Li, R. (2002). Variable selection for cox's proportional hazards model and frailty model. *Annals of Statistics*, pages 74–99.

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

Fan, J., Peng, H., et al. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics*, 32(3):928–961.

Fan, Y. and Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(3):531–552.

Kim, Y., Choi, H., and Oh, H.-S. (2008). Smoothly clipped absolute deviation on high dimensions. *Journal of the American Statistical Association*, 103(484):1665–1673.

Lahiri, S. N., Mukhopadhyay, S., et al. (2012). A penalized empirical likelihood method in high dimensions. *The Annals of Statistics*, 40(5):2511–2540.

Leng, C. and Tang, C. Y. (2012). Penalized empirical likelihood and growing dimensional general estimating equations. *Biometrika*, 99(3):703–716.

Owen, A. (1991). Empirical likelihood for linear models. *The Annals of Statistics*, pages 1725–1747.

Qin, J. and Lawless, J. (1995). Estimating equations, empirical likelihood and constraints on parameters. *Canadian Journal of Statistics*, 23(2):145–159.

Rivasplata, O. (2012). Subgaussian random variables: An expository note. *Internet publication, PDF*.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Tang, C. Y. and Leng, C. (2010). Penalized high-dimensional empirical likelihood. *Biometrika*, 97(4):905–920.

Wang, H., Li, B., and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):671–683.

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

Wang, T. and Zhu, L. (2011). Consistent tuning parameter selection in high dimensional sparse linear regression. *Journal of Multivariate Analysis*, 102(7):1141–1151.

4. Appendix

Proof of Lemma 1.

Under Assumption (E1)-(E4), by Theorem 3 in Tang and Leng (2010), it has been shown that with probability tending to 1,

$$\hat{\boldsymbol{\beta}}_{\tau_n\tilde{2}} = 0, \tag{6}$$

where $\hat{\boldsymbol{\beta}}_{\tau_n} = (\hat{\boldsymbol{\beta}}^T_{\tau_n\tilde{1}}, \hat{\boldsymbol{\beta}}^T_{\tau_n\tilde{2}})^T$ is the minimizer of (4), with $\hat{\boldsymbol{\beta}}_{\tau_n\tilde{1}} \in \mathbb{R}^d, \hat{\boldsymbol{\beta}}_{\tau_n\tilde{2}} \in \mathbb{R}^{p-d}$. As indicated by Lemma A4 in Tang and Leng (2010), we have that $\hat{\boldsymbol{\beta}}_{\tau_n}$ wil fall into $D_n = \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \leq ca_n\}$ with probability tending to 1, where a_n is defined as $(p/n)^{1/2-\sigma}$, c and σ are strictly positive constants satisfying $p^{1-\delta}/n^{1/2-\delta} \to 0$. Thus, we have that

$$\min_{j\in\alpha_0} \frac{|\hat{\beta}_{\tau_n j}|}{\tau_n} \geq \min_{j\in\alpha_0} \frac{|\beta_{0j}| - ca_n}{\tau_n},$$

with probability tending to 1, where $\hat{\beta}_{\tau_n j}$ denotes the jth component of $\hat{\boldsymbol{\beta}}_{\tau_n}$. By $\tau_n(n/p)^{1/2-\delta} \to \infty$, we have that $\frac{a_n}{\tau_n} \to 0$ and $\min_{j\in\alpha_0} \frac{|\beta_{0j}|}{\tau_n} \to \infty$. Therefore, we have that with probability tending to 1

$$\min_{j\in\alpha_0} \frac{|\hat{\beta}_{\tau_n j}|}{\tau_n} \to \infty. \tag{7}$$

Under Assumption (E4), there exits a constant h such that $p'_{\tau_n}(|\hat{\beta}_{\tau_n j}|) = 0$ for $\frac{|\hat{\beta}_{\tau_n j}|}{\tau_n} \geq h$ which implies, together with (7), we have that

$$b_n(\hat{\boldsymbol{\beta}}_{\tau_n}) = 0, \tag{8}$$

with probability tending to 1, where $b_n(\boldsymbol{\beta}) = (p'_{\tau_n}(\beta_1)\ \text{sign}\ (\beta_1), p'_{\tau_n}(\beta_2)\ \text{sign}\ (\beta_2), ..., p'_{\tau_n}(\beta_d)\ \text{sign}\ (\beta_d), \mathbf{0}^T)^T$.

By the definition of penalized empirical likelihood and the result (6), the estimator $\hat{\boldsymbol{\beta}}_{\tau_n}$ based on penalized empirical likelihood is the constrained minimizer of (4) subject to $\mathbf{H}_{\alpha_0}\boldsymbol{\beta} = 0$, with probability tending to 1, where $\mathbf{H}_{\alpha_0}$ are defined in Lemma 2. According to Qin and Lawless (1995), by the Lagrange multiplier method, obtaining the estimates is equivalent to minimizing a new objective function

$$\tilde{\ell}_{pen}(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\upsilon}) = n^{-1}\sum_{i=1}^{n} \log(1 + \boldsymbol{\lambda}^T\mathbf{U}_i(\boldsymbol{\beta})) + \sum_{j=1}^{p} p_{\tau_n}(|\beta_j|) + \boldsymbol{\upsilon}^T\mathbf{H}_{\alpha_0}\boldsymbol{\beta}, \tag{9}$$

where $\boldsymbol{\upsilon} \in \mathbb{R}^{p-df_{\alpha_0}}$ is the vector of extra Lagrange multipliers.

Define $\tilde{\mathbf{Q}}_{1n}(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\upsilon}) = \frac{1}{n}\sum_{i=1}^{n}\{1 + \boldsymbol{\lambda}^T\mathbf{U}_i(\boldsymbol{\beta})\}^{-1}\mathbf{U}_i(\boldsymbol{\beta})$, $\tilde{\mathbf{Q}}_{2n}(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\upsilon}) = -\frac{1}{n}\sum_{i=1}^{n}\{1 + \boldsymbol{\lambda}^T\mathbf{U}_i(\boldsymbol{\beta})\}^{-1}\boldsymbol{\lambda} + b_n(\boldsymbol{\beta}) + \mathbf{H}_{\alpha_0}^T\boldsymbol{\upsilon}$ and $\tilde{\mathbf{Q}}_{3n}(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\upsilon}) = \mathbf{H}_{\alpha_0}\boldsymbol{\upsilon}$. Denotes the minimizer of (9) as $(\hat{\boldsymbol{\beta}}_{pen}, \hat{\boldsymbol{\lambda}}_{pen}, \hat{\boldsymbol{\upsilon}}_{pen})$ which satisfies $\mathbf{0} = \tilde{\mathbf{Q}}_{jn}(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\upsilon}) \quad (j = 1, 2, 3)$. And since $\hat{\boldsymbol{\beta}}_{\tau_n}$ is the minimizer of (5), we have that $\hat{\boldsymbol{\beta}}_{pen} = \hat{\boldsymbol{\beta}}_{\tau_n}$.

By the definition of unpenalized empirical likelihood, the estimator $\hat{\boldsymbol{\beta}}_{\alpha_0}^{\star}$ is the constrained maximizer of $\ell_c(\boldsymbol{\beta})$ subject to $\mathbf{H}_{\alpha_0}\boldsymbol{\beta} = \mathbf{0}$. Similarly, obtaining the estimates is equivalent to minimizing a new objective function

$$\tilde{\ell}_c(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\upsilon}) = n^{-1}\sum_{i=1}^{n}\log(1 + \boldsymbol{\lambda}^T\mathbf{U}_i(\boldsymbol{\beta})) + \boldsymbol{\upsilon}^T\mathbf{H}_{\alpha_0}\boldsymbol{\beta}. \tag{10}$$

Define $\tilde{\mathbf{M}}_{1n}(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\upsilon}) = \frac{1}{n}\sum_{i=1}^{n}\{1+\boldsymbol{\lambda}^T\mathbf{U}_i(\boldsymbol{\beta})\}^{-1}\mathbf{U}_i(\boldsymbol{\beta})$, $\tilde{\mathbf{M}}_{2n}(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\upsilon}) = -\frac{1}{n}\sum_{i=1}^{n}\{1+ \boldsymbol{\lambda}^T\mathbf{U}_i(\boldsymbol{\beta})\}^{-1}\boldsymbol{\lambda} + \mathbf{H}_{\alpha}^T\boldsymbol{\upsilon}$ and $\tilde{\mathbf{M}}_{3n}(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\upsilon}) = \mathbf{H}_{\alpha}\boldsymbol{\upsilon}$. Denotes the minimizer of (10) as $(\hat{\boldsymbol{\beta}}_c, \hat{\boldsymbol{\lambda}}_c, \hat{\boldsymbol{\upsilon}}_c)$ which satisfies $\mathbf{0} = \tilde{\mathbf{M}}_{jn}(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\upsilon}) \quad (j = 1, 2, 3)$. And due to the fact that $\hat{\boldsymbol{\beta}}_{\alpha_0}^{\star}$ is the maximizer of $\ell_c(\boldsymbol{\beta})$, we have that $\hat{\boldsymbol{\beta}}_c = \hat{\boldsymbol{\beta}}_{\alpha_0}^{\star}$.

By (8), we have that with probability tending to 1, $(\hat{\boldsymbol{\beta}}_{\tau_n}, \hat{\boldsymbol{\lambda}}_{pen}, \hat{\boldsymbol{\upsilon}}_{pen})$ satisfies $\mathbf{0} = \tilde{\mathbf{M}}_{jn}(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\upsilon}) \quad (j = 1, 2, 3)$, which implies that $\hat{\boldsymbol{\beta}}_{\tau_n} = \hat{\boldsymbol{\beta}}_{\alpha_0}^{\star}$ with probability tending to 1.

Proof of Lemma 2.

From the definition, we know that $\hat{\boldsymbol{\beta}}_{\alpha}^{\star}$ is the solution to the following problem:

$$\min_{\boldsymbol{\beta}:\mathbf{H}_{\alpha}\boldsymbol{\beta}=0} \ell_c(\boldsymbol{\beta}) = \sum_{i=1}^{n}\log\{1 + \boldsymbol{\lambda}_{\boldsymbol{\beta}}^T\mathbf{U}_i(\boldsymbol{\beta})\}, \tag{11}$$

where $\boldsymbol{\lambda}_{\boldsymbol{\beta}}$ solves that $\dfrac{1}{n}\sum_{i=1}^{n}\dfrac{\mathbf{U}_i(\boldsymbol{\beta})}{1 + \boldsymbol{\lambda}_{\boldsymbol{\beta}}^T\mathbf{U}_i(\boldsymbol{\beta})} = 0.$

According to Qin and Lawless (1995), by the Lagrange multiplier method, obtaining the estimates $\hat{\boldsymbol{\beta}}_{\alpha}^{\star}$ is equivalent to minimizing a new objective function

$$\tilde{\ell}(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\upsilon}) = \frac{1}{n}\sum_{i=1}^{n}\log\{1 + \boldsymbol{\lambda}^T\mathbf{U}_i(\boldsymbol{\beta})\} + \boldsymbol{\upsilon}^T\mathbf{H}_{\alpha}\boldsymbol{\beta},$$

where $\boldsymbol{\upsilon} \in \mathbb{R}^{p-df_{\alpha}}$ is the vector of extra Lagrange multipliers.

Define $\tilde{\mathbf{Q}}_{1n}(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\upsilon}) = \frac{1}{n}\sum_{i=1}^{n}\{1+\boldsymbol{\lambda}^T\mathbf{U}_i(\boldsymbol{\beta})\}^{-1}\mathbf{U}_i(\boldsymbol{\beta})$, $\tilde{\mathbf{Q}}_{2n}(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\upsilon}) = -\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbf{X}_i\mathbf{X}_i^T\boldsymbol{\lambda}}{1+\boldsymbol{\lambda}^T\mathbf{U}_i(\boldsymbol{\beta})}^{-1} + \mathbf{H}_{\alpha}^T\boldsymbol{\upsilon}$ and $\tilde{\mathbf{Q}}_{3n}(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\upsilon}) = \mathbf{H}_{\alpha}\boldsymbol{\beta}$. The minimizer $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\lambda}}, \hat{\upsilon})$ satisfies $\mathbf{0} = \tilde{\mathbf{Q}}_{jn}(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\upsilon}) \quad (j =$

$1, 2, 3$). To ensure the expansions of $\tilde{\mathbf{Q}}_{jn}$ around the value $(\boldsymbol{\beta}_0, \mathbf{0}, \mathbf{0})$ $(j = 1, 2, 3)$, we first present the following two statements which are similar to the Lemma A3 and Lemma A4 in the Tang and Leng (2010).

(S1). Let $a_n = (p/n)^{1/2-\delta}, \tilde{D}_n = \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \leq ca_n, \mathbf{H}_\alpha \boldsymbol{\beta} = \mathbf{0}\}$ where $\delta, c > 0$ are strictly positive constants and δ satisfies $p^{1-\delta}/n^{1/2-\delta} \to 0$. Then $\|\boldsymbol{\lambda}_{\boldsymbol{\beta}}\| = O_p(a_n)$ for $\boldsymbol{\beta} \in \tilde{D}_n$.

(S2). As $n \to \infty$, with probability tending to 1, the problem (11) has a minimum in $\tilde{D}_n$.

Statement (S1) follows directly from Lemma A3 in Tang and Leng (2010) since $\tilde{D}_n \subseteq D_n$, where $D_n = \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \leq ca_n\}$. And according to the proof of Lemma A4 in Tang and Leng (2010), we have that for any given C, as $n \to \infty$, $P\{2\ell_c(\boldsymbol{\beta}) - 2\ell_c(\boldsymbol{\beta}_0) > C\} \to 1$ for $\boldsymbol{\beta} \in \partial D_n$ where ∂D_n denotes the boundary of D_n. Since $\partial \tilde{D}_n \subseteq \partial D_n$, we can see that as $n \to \infty$, $P\{2\ell_c(\boldsymbol{\beta}) - 2\ell_c(\boldsymbol{\beta}_0) > C\} \to 1$ for $\boldsymbol{\beta} \in \partial \tilde{D}_n$, which established Statement (S2).

Therefore, we have from (S2), $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p(a_n)$ and from (S1) that $\|\hat{\boldsymbol{\lambda}}\| = O_p(a_n)$ is stochastically small. Similar to the argument in Qin and Lawless (1995), from $\mathbf{0} = \tilde{\mathbf{Q}}_{2n}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\upsilon}})$, we conclude that $\|\hat{\boldsymbol{\upsilon}}\| = O_p(a_n)$. Hence, we can use the stochastic expansions of $\tilde{\mathbf{Q}}_{jn}$ $(j = 1, 2, 3)$ around the value $(\boldsymbol{\beta}_0, \mathbf{0}, \mathbf{0})$. This yields

$$\begin{pmatrix} -\tilde{\mathbf{Q}}_{1n}(\boldsymbol{\beta}_0, \mathbf{0}, \mathbf{0}) \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} -\boldsymbol{\Sigma} & -\boldsymbol{\Sigma} & \mathbf{0} \\ -\boldsymbol{\Sigma} & \mathbf{0} & \mathbf{H}_\alpha^T \\ \mathbf{0} & \mathbf{H}_\alpha & \mathbf{0} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\lambda}} - \mathbf{0} \\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \\ \hat{\boldsymbol{\upsilon}} - \mathbf{0} \end{pmatrix} + \mathbf{R}_n, \tag{12}$$

where $\|\mathbf{R}_n\| = o_p(n^{-1/2})$ according to the proof of Theorem 3 in Tang and Leng (2010).

Let

$$\mathbf{K} = \begin{pmatrix} -\boldsymbol{\Sigma} & -\boldsymbol{\Sigma} & \mathbf{0} \\ -\boldsymbol{\Sigma} & \mathbf{0} & \mathbf{H}_\alpha^T \\ \mathbf{0} & \mathbf{H}_\alpha & \mathbf{0} \end{pmatrix}.$$

By inverting (12), we have

$$\begin{pmatrix} \hat{\boldsymbol{\lambda}} - \mathbf{0} \\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \\ \hat{\boldsymbol{\upsilon}} - \mathbf{0} \end{pmatrix} = \mathbf{K}^{-1} \left(\begin{pmatrix} -\tilde{\mathbf{Q}}_{1n}(\boldsymbol{\beta}_0, \mathbf{0}, \mathbf{0}) \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix} - \mathbf{R}_n \right).$$

This implies that

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = \{\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^{-1}\mathbf{H}_\alpha^T(\mathbf{H}_\alpha\boldsymbol{\Sigma}^{-1}\mathbf{H}_\alpha^T)^{-1}\mathbf{H}_\alpha\boldsymbol{\Sigma}^{-1}\}(n^{-1}\sum_{i=1}^{n}\mathbf{X}_i\epsilon_i - \mathbf{R}_{2n}), \tag{13}$$

where $\mathbf{R}_{2n}$ is the corresponding component in vector $\mathbf{R}_n = (\mathbf{R}_{1n}^T, \mathbf{R}_{2n}^T, \mathbf{R}_{3n}^T)^T$ and $\|\mathbf{R}_{2n}\| = o_p(n^{-1/2})$.

Let $z_i = \boldsymbol{\lambda}_{\hat{\boldsymbol{\beta}}}^T\mathbf{U}_i(\hat{\boldsymbol{\beta}})$. As $\max_i |\boldsymbol{\lambda}_{\hat{\boldsymbol{\beta}}}^T\mathbf{U}_i(\hat{\boldsymbol{\beta}})| = o_p(1)$ indicated by the proof of Lemma A3 in Tang and Leng (2010), we have by Taylor's expansion,

$$\ell_c(\hat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{n} z_i - \sum_{i=1}^{n}\frac{z_i^2}{2} + \sum_{i=1}^{n}\frac{z_i^3}{3(1+\xi_i)^4}\right),$$

where $|\xi_i| < |\boldsymbol{\lambda}_{\hat{\boldsymbol{\beta}}}^T\mathbf{U}_i(\hat{\boldsymbol{\beta}})|$.

Following the proof of Lemma A4 in Tang and Leng (2010), we have the expansion of $\boldsymbol{\lambda}_{\boldsymbol{\beta}}$ for $\boldsymbol{\beta} \in \tilde{D}_n$ to be $\boldsymbol{\lambda}_{\boldsymbol{\beta}} = \mathbf{T}_n(\boldsymbol{\beta})^{-1}\bar{\mathbf{U}}(\boldsymbol{\beta}) + \mathbf{T}_n(\boldsymbol{\beta})^{-1}\mathbf{r}_n$, where $\mathbf{T}_n(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\mathbf{U}_i(\boldsymbol{\beta})\mathbf{U}_i^T(\boldsymbol{\beta})$, $\mathbf{r}_n = n^{-1}\sum_{i=1}^{n}[\mathbf{U}_i(\boldsymbol{\beta})\{\boldsymbol{\lambda}_{\boldsymbol{\beta}}^T\mathbf{U}_i(\boldsymbol{\beta})\}^2(1+\epsilon_i)^{-3}]$ and $|\epsilon_i| < |\boldsymbol{\lambda}_{\boldsymbol{\beta}}^T\mathbf{U}_i(\boldsymbol{\beta})|$. Let $\bar{\mathbf{U}} = n^{-1}\sum_{i=1}^{n}\mathbf{X}_i\epsilon_i$. Similarly to the establishment of (A5) in the proof of Theorem 2 in Tang and Leng (2010), by substituting the expansion of $\hat{\boldsymbol{\beta}}$ in (13) and $\boldsymbol{\lambda}_{\hat{\boldsymbol{\beta}}}$ into z_i, we have that

$$2\ell_c(\hat{\boldsymbol{\beta}}) = n\bar{\mathbf{U}}^T\boldsymbol{\Sigma}^{-1}\mathbf{H}_\alpha^T(\mathbf{H}_\alpha\boldsymbol{\Sigma}^{-1}\mathbf{H}_\alpha^T)^{-1}\mathbf{H}_\alpha\boldsymbol{\Sigma}^{-1}\bar{\mathbf{U}} + o_p(1),$$

which completes the proof of Lemma 2.

Proof of Theorem 1.

Let τ_n be the sequence of tuning parameters in Lemma 1. By Lemma 1, with probability tending to 1, we have

$$\ell_c(\hat{\boldsymbol{\beta}}_{\tau_n}) = \ell_c(\hat{\boldsymbol{\beta}}_{\alpha_0}^\star).$$

Thus we have that $df_{\alpha_{\tau_n}} = df_{\alpha_0}$ with probability tending to 1. Hence it follows that,

$$P\{\mathrm{GIC}(\hat{\boldsymbol{\beta}}_{\tau_n}) = \mathrm{GIC}(\hat{\boldsymbol{\beta}}_{\alpha_0}^\star)\} \to 1. \tag{14}$$

Next we want to show that $\mathrm{GIC}(\hat{\boldsymbol{\beta}}_\tau) > \mathrm{GIC}(\hat{\boldsymbol{\beta}}_{\tau_n})$ with probability tending to 1, for any τ that cannot result in the true model. First, we consider τ that would result in underfitting models, namely, $\tau \in \Omega_- = \{\tau : \alpha \not\supseteq \alpha_0\}$.

Recall that based on the selected model α_τ, we are able to obtain its corresponding non-penalized estimates $\hat{\boldsymbol{\beta}}^\star_{\alpha_\tau}$.Then we have that

$$\ell_c(\hat{\boldsymbol{\beta}}^\star_{\alpha_\tau}) \geq \ell_c(\hat{\boldsymbol{\beta}}_\tau),$$

and $2\ell_c(\hat{\boldsymbol{\beta}}_\tau) + C_n df_\tau > 2\ell_c(\hat{\boldsymbol{\beta}}^\star_{\alpha_\tau})$. Thus, we have

$$\mathrm{GIC}(\hat{\boldsymbol{\beta}}_\tau) > 2\ell_c(\hat{\boldsymbol{\beta}}^\star_{\alpha_\tau}). \tag{15}$$

By (15) and (14), with probability tending to 1, it follows that

$$\mathrm{GIC}(\hat{\boldsymbol{\beta}}_\tau) - \mathrm{GIC}(\hat{\boldsymbol{\beta}}_{\tau_n}) > 2\ell_c(\hat{\boldsymbol{\beta}}^\star_{\alpha_\tau}) - 2\ell_c(\hat{\boldsymbol{\beta}}^\star_{\alpha_0}) - C_n df_{\alpha_0}.$$

For any $\tau \in \Omega_- = \{\tau : \alpha \not\supseteq \alpha_0\}$, we can take $\inf_{\tau\in\Omega_-}$ over $\mathrm{GIC}(\hat{\boldsymbol{\beta}}_\tau)$. By Assumption (E5) and $\frac{C_n}{\sqrt{n}} \to 0, \frac{p}{\sqrt{n}} \to 0$, for any $\tau \in \Omega_-$, we have with probability tending to 1

$$\begin{aligned}
&\inf_{\tau\in\Omega_-} \mathrm{GIC}(\hat{\boldsymbol{\beta}}_\tau) - \mathrm{GIC}(\hat{\boldsymbol{\beta}}_{\tau_n}) \\
\geq\ &\inf_{\tau\in\Omega_-} \frac{2\ell_c(\hat{\boldsymbol{\beta}}^\star_{\alpha_\tau})}{n} - \frac{2\ell_c(\hat{\boldsymbol{\beta}}^\star_{\alpha_0})}{n} - \frac{C_n df_{\alpha_0}}{n} \\
\geq\ &\min_{\alpha\not\supseteq\alpha_0} \frac{\ell_c(\hat{\boldsymbol{\beta}}^\star_{\alpha})}{n} - \frac{\ell_c(\hat{\boldsymbol{\beta}}^\star_{\alpha_0})}{n} - \frac{C_n df_{\alpha_0}}{2n} \\
\geq\ &C_- + o_p(1),
\end{aligned} \tag{16}$$

as $n \to \infty$. (16) is due to Assumption (E5), $\frac{C_n p}{n} \to 0$ and the fact that $\frac{\ell_c(\hat{\boldsymbol{\beta}}^\star_{\alpha_0})}{n} = o_p(1)$, which we will show as follows.

By Lemma 2, we have that

$$\begin{aligned}
\frac{\ell_c(\hat{\boldsymbol{\beta}}^\star_{\alpha_0})}{n} &= \frac{1}{2}\frac{n\bar{\mathbf{U}}^T\mathbf{H}^T_{\alpha_0}(\mathbf{H}_{\alpha_0}\boldsymbol{\Sigma}^{-1}\mathbf{H}^T_{\alpha_0})^{-1}\mathbf{H}_{\alpha_0}\bar{\mathbf{U}} + o_p(1)}{n} \\
&= \frac{1}{2}\frac{\mathbf{W}_n^T\mathbf{P}_n\mathbf{W}_n}{n} + o_p(1),
\end{aligned}$$

where $\mathbf{W}_n = \sqrt{n}\boldsymbol{\Sigma}^{-1/2}\bar{\mathbf{U}}$ and $\mathbf{P}_n = \boldsymbol{\Sigma}^{-1/2}\mathbf{H}^T_{\alpha_0}(\mathbf{H}_{\alpha_0}\boldsymbol{\Sigma}^{-1}\mathbf{H}^T_{\alpha_0})^{-1}\mathbf{H}_{\alpha_0}\boldsymbol{\Sigma}^{-1/2}$. Since $\mathbf{P}_n$ is a projection matrix with rank $p - df_{\alpha_0}$, we have that $\mathbf{W}_n^T\mathbf{P}_n\mathbf{W}_n \leq \mathbf{W}_n^T\mathbf{W}_n$. Let $\mathbf{F}_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbf{Z}_i\epsilon_i$, where $\mathbf{Z}_i \in \mathbb{R}^m$ is defined in model (2). We have that $\mathbf{W}_n^T\mathbf{W}_n = \mathbf{F}_n^T\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Gamma}\mathbf{F}_n$ and that $\mathbf{F}_n \sim \mathrm{subG}_m(\sigma^2)$ by Assumption (E6). Since $\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T = \boldsymbol{\Sigma}$, we know that there exists vectors $\{\mathbf{a}_i\}_{i=1}^p$ such that $\Gamma^T\Sigma^{-1}\Gamma = \sum_{i=1}^p \mathbf{a}_i\mathbf{a}_i^T$ with $\mathbf{a}_i^T\mathbf{a}_i = 1$.

For the sub-Gaussian variables $\mathbf{a}_i^T\mathbf{F}_n \sim \text{subG}(\sigma^2)$, we have that $\text{E}\{(\mathbf{a}_i^T\mathbf{F}_n)^2\} \leq 4\sigma^2$ and $\text{E}\{(\mathbf{a}_i^T\mathbf{F}_n)^4\} \leq 16\sigma^4$. Thus, we have that

$$\text{E}(\frac{\mathbf{W}_n^T\mathbf{W}_n}{p}) = \text{E}(\frac{\sum_{i=1}^p \mathbf{F}_n^T\mathbf{a}_i\mathbf{a}_i^T\mathbf{F}_n}{p}) \leq 4\sigma^2.$$

For the second moment of $\frac{\mathbf{W}_n^T\mathbf{W}_n}{p}$, we have that

$$\text{E}\{(\frac{\mathbf{W}_n^T\mathbf{W}_n}{p})^2\} \leq p\text{E}(\frac{\sum_{i=1}^p(\mathbf{F}_n^T\mathbf{a}_i\mathbf{a}_i^T\mathbf{F}_n)^2}{p^2}) \leq 16\sigma^4.$$

Thus, we have that $\frac{\mathbf{W}_n^T\mathbf{P}_n\mathbf{W}_n}{p} = O_p(1)$. It follows that $\frac{\mathbf{W}_n^T\mathbf{P}_n\mathbf{W}_n}{n} = o_p(1)$ which completes the proof of (16). From (16), we have that $\inf_{\tau\in\Omega_-} \text{GIC}(\hat{\boldsymbol{\beta}}_\tau) - \text{GIC}(\hat{\boldsymbol{\beta}}_{\tau_n})$ must be positive asymptotically.

Therefore, we have that

$$P\{\inf_{\tau\in\Omega_-} \text{GIC}(\hat{\boldsymbol{\beta}}_\tau) > \text{GIC}(\hat{\boldsymbol{\beta}}_{\tau_n})\} \to 1. \tag{17}$$

Next, for any $\tau \in \Omega_+ = \{\tau : \alpha \supset \alpha_0\}$, we have with probability tending to 1

$$\begin{aligned} & \text{GIC}(\hat{\boldsymbol{\beta}}_\tau) - \text{GIC}(\hat{\boldsymbol{\beta}}_{\tau_n}) \\ = \;& 2\ell_c(\hat{\boldsymbol{\beta}}_\tau) - 2\ell_c(\hat{\boldsymbol{\beta}}_{\tau_n}) + C_n(df_{\alpha_\tau} - df_{\alpha_{\tau_n}}) \\ \geq \;& 2\ell_c(\hat{\boldsymbol{\beta}}^\star_{\alpha_\tau}) - 2\ell_c(\hat{\boldsymbol{\beta}}^\star_{\alpha_0}) + C_n(df_{\alpha_\tau} - df_{\alpha_0}), \end{aligned}$$

as $n \to \infty$.

We then take $\inf_{\tau\in\Omega_+}$ over $\text{GIC}(\hat{\boldsymbol{\beta}}_\tau)$. So we have with probability tending to 1

$$\begin{aligned} & \inf_{\tau\in\Omega_+} \text{GIC}(\hat{\boldsymbol{\beta}}_\tau) - \text{GIC}(\hat{\boldsymbol{\beta}}_{\tau_n}) \\ \geq \;& \min_{\alpha\supset\alpha_0} \left(2[\ell_c(\hat{\boldsymbol{\beta}}^\star_\alpha) - \ell_c(\hat{\boldsymbol{\beta}}^\star_{\alpha_0})] + C_n(df_\alpha - df_{\alpha_0})\right) \\ = \;& \min_{\alpha\supset\alpha_0} \left((df_\alpha - df_{\alpha_0})\{C_n - 2\frac{\ell_c(\hat{\boldsymbol{\beta}}^\star_{\alpha_0}) - \ell_c(\hat{\boldsymbol{\beta}}^\star_\alpha)}{df_\alpha - df_{\alpha_0}}\}\right). \end{aligned} \tag{18}$$

For any $\alpha \supset \alpha_0$, we have $df_\alpha - df_{\alpha_0} \geq 1$. Next, we are going to show that $\max_{\alpha\supset\alpha_0} \frac{\ell_c(\hat{\boldsymbol{\beta}}^\star_{\alpha_0}) - \ell_c(\hat{\boldsymbol{\beta}}^\star_\alpha)}{df_\alpha - df_{\alpha_0}} = O_p(\log(p))$.

By Lemma 2, we have

$$\begin{aligned} & \max_{\alpha\supset\alpha_0} \frac{\ell_c(\hat{\boldsymbol{\beta}}^\star_{\alpha_0}) - \ell_c(\hat{\boldsymbol{\beta}}^\star_\alpha)}{df_\alpha - df_{\alpha_0}} \\ = \;& \max_{\alpha\supset\alpha_0} \frac{n\bar{\mathbf{U}}^T\boldsymbol{\Sigma}^{-\frac{1}{2}}(\mathbf{A}_{\alpha_0} - \mathbf{A}_\alpha)\boldsymbol{\Sigma}^{-\frac{1}{2}}\bar{\mathbf{U}} + o_p(1)}{df_\alpha - df_{\alpha_0}}, \end{aligned} \tag{19}$$

where $\mathbf{A}_\alpha = \mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{H}_\alpha^T(\mathbf{H}_\alpha\mathbf{\Sigma}^{-1}\mathbf{H}_\alpha^T)^{-1}\mathbf{H}_\alpha\mathbf{\Sigma}^{-\frac{1}{2}}$, $\mathbf{A}_{\alpha_0} = \mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{H}_{\alpha_0}^T(\mathbf{H}_{\alpha_0}\mathbf{\Sigma}^{-1}\mathbf{H}_{\alpha_0}^T)^{-1}\mathbf{H}_{\alpha_0}\mathbf{\Sigma}^{-\frac{1}{2}}$ are two projection matrices. Let $\mathbf{W}_n = \sqrt{n}\mathbf{\Sigma}^{-\frac{1}{2}}\bar{\mathbf{U}}$, then (19) becomes

$$\max_{\alpha\supset\alpha_0} \frac{\mathbf{W}_n^T(\mathbf{A}_{\alpha_0} - \mathbf{A}_\alpha)\mathbf{W}_n + o_p(1)}{df_\alpha - df_{\alpha_0}}. \tag{20}$$

The numerator of (20) is dominated by $\mathbf{W}_n^T(\mathbf{A}_{\alpha_0} - \mathbf{A}_\alpha)\mathbf{W}_n$. Therefore, we only need to show that

$$\max_{\alpha\supset\alpha_0} \frac{\mathbf{W}_n^T(\mathbf{A}_{\alpha_0} - \mathbf{A}_\alpha)\mathbf{W}_n}{df_\alpha - df_{\alpha_0}} = O_p(\log(p)). \tag{21}$$

Before the derivation of (21), let's introduce the following notation. Let $\alpha^c = \alpha \setminus \alpha_0$, $\alpha_1 = \bar{\alpha}\setminus\alpha_0$, $\mathbf{B} = \mathbf{\Sigma}^{-1/2}$ and $\mathbf{B}_\alpha$ be the matrix composed by the columns of $\mathbf{B}$ indexed by α, that is, $\mathbf{\Sigma}^{-1/2} = \mathbf{B} = (\mathbf{B}_{\alpha_0}, \mathbf{B}_{\alpha_1})$. And let $\mathbf{P_A}$ denote $\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$, $\mathbf{Q_A} = \mathbf{I} - \mathbf{P_A}$ for any matrix $\mathbf{A}$. Denote $\tau_{min}(\mathbf{A})$ as the minimal eigenvalues of $\mathbf{A}$, for any arbitrary positive definite matrix $\mathbf{A}$. Then we have $\mathbf{W}_n^T(\mathbf{A}_{\alpha_0} - \mathbf{A}_\alpha)\mathbf{W}_n = \mathbf{W}_n^T(\mathbf{P}_{\mathbf{B}_\alpha} - \mathbf{P}_{\mathbf{B}_{\alpha_0}})\mathbf{W}_n$. Inspired by the idea used in the proof of Theorem 2 in Wang et al. (2009), we let $\tilde{\mathbf{B}}_{\alpha^c} = \mathbf{Q}_{\mathbf{B}_{\alpha_0}}\mathbf{B}_{\alpha^c}$. Then we have $\mathbf{P}_{\mathbf{B}_\alpha} - \mathbf{P}_{\mathbf{B}_{\alpha_0}} = \mathbf{P}_{\tilde{\mathbf{B}}_{\alpha^c}}$ and

$$\begin{aligned}
\frac{\mathbf{W}_n^T(\mathbf{A}_{\alpha_0} - \mathbf{A}_\alpha)\mathbf{W}_n}{df_\alpha - df_{\alpha_0}} &= \frac{\mathbf{W}_n^T\mathbf{P}_{\tilde{\mathbf{B}}_{\alpha^c}}\mathbf{W}_n}{df_\alpha - df_{\alpha_0}} \\
&= \frac{(\mathbf{W}_n^T\tilde{\mathbf{B}}_{\alpha^c})(\tilde{\mathbf{B}}_{\alpha^c}^T\tilde{\mathbf{B}}_{\alpha^c})^{-1}(\tilde{\mathbf{B}}_{\alpha^c}^T\mathbf{W}_n)}{|\alpha^c|} \\
&\le \frac{h_{max}^{\alpha^c}\|\mathbf{W}_n^T\tilde{\mathbf{B}}_{\alpha^c}\|^2}{|\alpha^c|} \\
&\le \frac{h_{max}^{\alpha^c}\sum_{j\in\alpha^c}(\mathbf{W}_n^T\tilde{\mathbf{B}}_j)^2}{|\alpha^c|} \\
&\le h_{max}^{\alpha^c}\max_{j\in\alpha^c}(\mathbf{W}_n^T\tilde{\mathbf{B}}_j)^2,
\end{aligned}$$

where $h_{max}^{\alpha^c} = \tau_{min}^{-1}(\tilde{\mathbf{B}}_{\alpha^c}^T\tilde{\mathbf{B}}_{\alpha^c})$, $\mathbf{B}_j$ is the jth column of $\mathbf{B}$ and $\tilde{\mathbf{B}}_j = \mathbf{Q}_{\mathbf{B}_{\alpha_0}}\mathbf{B}_j$. Note that $\alpha^c \subset \alpha_1$. Thus we have that $\tau_{min}^{-1}(\tilde{\mathbf{B}}_{\alpha^c}^T\tilde{\mathbf{B}}_{\alpha^c}) \le \tau_{min}^{-1}(\tilde{\mathbf{B}}_{\alpha_1}^T\tilde{\mathbf{B}}_{\alpha_1}) = (h_{max}^{\alpha_1})^{-1}$. We then must have

$$\begin{aligned}
\max_{\alpha\supset\alpha_0} \frac{\mathbf{W}_n^T(\mathbf{A}_{\alpha_0} - \mathbf{A}_\alpha)\mathbf{W}_n}{df_\alpha - df_{\alpha_0}} &= \max_{\alpha^c\subset\alpha_1} \frac{\mathbf{W}_n^T\mathbf{P}_{\tilde{\mathbf{B}}_{\alpha^c}}\mathbf{W}_n}{|\alpha^c|} \\
&\le h_{max}^{\alpha_1} \times \max_{j\in\alpha^c}(\mathbf{W}_n^T\tilde{\mathbf{B}}_j)^2.
\end{aligned} \tag{22}$$

We next examine the two terms of (22) respectively.

Firstly, let $\boldsymbol{\gamma}$ be the eigenvector associated with $\tau_{min}(\tilde{\mathbf{B}}_{\alpha_1}^T\tilde{\mathbf{B}}_{\alpha_1})$, i.e $\|\boldsymbol{\gamma}\|=1$ and

$$\tau_{min}(\tilde{\mathbf{B}}_{\alpha_1}^T\tilde{\mathbf{B}}_{\alpha_1})=\boldsymbol{\gamma}^T(\tilde{\mathbf{B}}_{\alpha_1}^T\tilde{\mathbf{B}}_{\alpha_1})\boldsymbol{\gamma}=\|\tilde{\mathbf{B}}_{\alpha_1}\boldsymbol{\gamma}\|^2.$$

By definition, we know that $\tilde{\mathbf{B}}_{\alpha_1}\boldsymbol{\gamma}=\mathbf{B}_{\alpha_1}\boldsymbol{\gamma}+\mathbf{B}_{\alpha_0}\boldsymbol{\gamma}^\star$ with $\boldsymbol{\gamma}^\star=-(\mathbf{B}_{\alpha_0}^T\mathbf{B}_{\alpha_0})^{-1}\mathbf{B}_{\alpha_0}^T\mathbf{B}_{\alpha_1}\boldsymbol{\gamma}$. Therefore, we have that

$$\begin{aligned}\tau_{min}(\tilde{\mathbf{B}}_{\alpha_1}^T\tilde{\mathbf{B}}_{\alpha_1}) &= \|\mathbf{B}_{\alpha_1}\boldsymbol{\gamma}+\mathbf{B}_{\alpha_0}\boldsymbol{\gamma}^\star\|^2=\|\mathbf{B}\boldsymbol{\gamma}_0\|^2=\boldsymbol{\gamma}_0^T\mathbf{B}^T\mathbf{B}\boldsymbol{\gamma}_0\\ &\geq \tau_{min}(\mathbf{B}^T\mathbf{B})\|\boldsymbol{\gamma}_0\|^2\geq\tau_{min}(\mathbf{B}^T\mathbf{B})\geq\tau_{min}(\boldsymbol{\Sigma})\geq C_1,\end{aligned}$$

where $\boldsymbol{\gamma}_0=(\boldsymbol{\gamma}^{\star T},\boldsymbol{\gamma}^T)^T$ satisfies that $\|\boldsymbol{\gamma}_0\|^2>1$ and C_1 is the constant in the Assumption (E2). This indicates that $h_{max}^{\alpha_1}\leq\frac{1}{C_1}$.

Secondly, under the sub-Gaussian Assumption (E6), we know that $\frac{\tilde{\mathbf{B}}_j^T\mathbf{W}_n}{\sigma_j}=\frac{\tilde{\mathbf{B}}_j^T\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Gamma}\mathbf{F}_n}{\sigma_j}\sim$ subG(σ^2), where $\mathbf{F}_n=\frac{1}{\sqrt{n}}\sum_{i=1}^n\mathbf{Z}_i\epsilon_i\sim\text{subG}_m(\sigma^2)$ and $\sigma_j^2=\|\tilde{\mathbf{B}}_j^T\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Gamma}\|^2=\|\tilde{\mathbf{B}}_j^T\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T\boldsymbol{\Sigma}^{-1/2}\tilde{\mathbf{B}}_j\|^2=\|\tilde{\mathbf{B}}_j\|^2\leq\|\mathbf{B}_j\|^2\leq\tau_{max}(\mathbf{B}_j^T\mathbf{B}_j)\leq\tau_{max}(\mathbf{B}^T\mathbf{B})\leq\tau_{max}(\boldsymbol{\Sigma})\leq C_2$, C_2 is the constant in the Assumption (E2). Thus, (22) can be further bounded by

$$\max_{\alpha^c\subset\alpha_1}\frac{\mathbf{W}_n^T\mathbf{P}_{\tilde{\mathbf{B}}_{\alpha^c}}\mathbf{W}_n}{|\alpha^c|}\leq\frac{C_2\sigma^2}{C_1}\times\max_{j\in\alpha_1}V_j^2,$$

where $V_j^2=\frac{(\mathbf{W}_n^T\tilde{\mathbf{B}}_j)^2}{\sigma_j^2\sigma^2}$ is the square of a sub-Gaussian variable with variance proxy 1. Even though these variables $V_j^2, j\in\alpha_1$ might be dependent, we can still proceed by using Bonferroni's inequality to obtain

$$\begin{aligned}P(\max_{j\in\alpha_1}V_j^2>3\log(p)) &\leq p\times P(V_j^2>3\log(p))\\ &\leq p\times P(|V_j|>\sqrt{3\log(p)})\\ &\leq 2p\exp^{-3\log(p)/2}\leq 2p^{-1/2},\end{aligned}$$

which implies that $\max_{j\in\alpha_1}V_j^2\leq 3\log(p)$ with probability tending to 1 as $p\to\infty$. Therefore, we have that

$$\max_{\alpha^c\subset\alpha_1}\frac{\mathbf{W}_n^T\mathbf{P}_{\tilde{\mathbf{B}}_{\alpha^c}}\mathbf{W}_n}{|\alpha^c|}\leq\frac{3\sigma C_2}{C_1}\log(p)$$

with probability tending to 1. So it follows that $\max_{\alpha\supset\alpha_0}\frac{\mathbf{W}_n^T(\mathbf{A}_{\alpha_0}-\mathbf{A}_\alpha)\mathbf{W}_n}{df_\alpha-df_{\alpha_0}}=O_p(\log(p))$ as $n\to\infty$.

Therefore, we have shown that with probability tending to 1,

$$\begin{aligned} & \min_{\alpha \supset \alpha_0} \{C_n - 2\frac{\ell_c(\hat{\boldsymbol{\beta}}^\star_{\alpha_0}) - \ell_c(\hat{\boldsymbol{\beta}}^\star_{\alpha})}{df_\alpha - df_{\alpha_0}}\} \\ \geq & \; C_n - 2\max_{\alpha \supset \alpha_0} \frac{\ell_c(\hat{\boldsymbol{\beta}}^\star_{\alpha_0}) - \ell_c(\hat{\boldsymbol{\beta}}^\star_{\alpha})}{df_\alpha - df_{\alpha_0}} \\ \geq & \; C_n - O_p(\log(p)), \end{aligned}$$

which is guaranteed to be positive asymptotically.

Now let's continue with (18). Together with the fact that $\min_{\alpha \supset \alpha_0}(df_\alpha - df_{\alpha_0}) > 0$, we have that with probability tending to 1,

$$\begin{aligned} & \inf_{\tau \in \Omega_+} \mathrm{GIC}(\hat{\boldsymbol{\beta}}_\tau) - \mathrm{GIC}(\hat{\boldsymbol{\beta}}_{\tau_n}) \\ \geq & \min_{\alpha \supset \alpha_0} \left(2[\ell_c(\hat{\boldsymbol{\beta}}^\star_{\alpha}) - \ell_c(\hat{\boldsymbol{\beta}}^\star_{\alpha_0})] + C_n(df_\alpha - df_{\alpha_0}) \right) \\ = & \min_{\alpha \supset \alpha_0} \left((df_\alpha - df_{\alpha_0})\{C_n - 2\frac{\ell_c(\hat{\boldsymbol{\beta}}^\star_{\alpha_0}) - \ell_c(\hat{\boldsymbol{\beta}}^\star_{\alpha})}{df_\alpha - df_{\alpha_0}}\} \right), \end{aligned}$$

which must be positive asymptotically.

Thus, we have that

$$P\left(\inf_{\tau \in \Omega_+} \mathrm{GIC}(\hat{\boldsymbol{\beta}}_\tau) > \mathrm{GIC}(\hat{\boldsymbol{\beta}}_{\tau_n}) \right) \to 1. \tag{23}$$

Based on (17) and (23) together, we have that $P\{\inf_{\tau \in (\Omega_- \cup \Omega_+)} \mathrm{GIC}(\hat{\boldsymbol{\beta}}_\tau) > \mathrm{GIC}(\hat{\boldsymbol{\beta}}_{\tau_n})\} \to 1$ which implies that $P\{\inf_{\tau \in (\Omega_- \cup \Omega_+)} \mathrm{GIC}(\hat{\boldsymbol{\beta}}_\tau) > \mathrm{GIC}(\hat{\boldsymbol{\beta}}_{\tau_n})\} \to 1$. Consequently, this completes the proof of theorem 1.