Communications for Statistical Applications and Methods
2018, Vol. 25, No. 5, 453–470

https://doi.org/10.29220/CSAM.2018.25.5.453
Print ISSN 2287-7843 / Online ISSN 2383-4757

Non-convex penalized estimation for the AR process

Okyoung Na[a], Sunghoon Kwon[1,b]

[a]Department of Applied Statistics, Kyonggi University, Korea;
[b]Department of Applied Statistics, Konkuk University, Korea

Abstract

We study how to distinguish the parameters of the sparse autoregressive (AR) process from zero using a non-convex penalized estimation. A class of non-convex penalties are considered that include the smoothly clipped absolute deviation and minimax concave penalties as special examples. We prove that the penalized estimators achieve some standard theoretical properties such as weak and strong oracle properties which have been proved in sparse linear regression framework. The results hold when the maximal order of the AR process increases to infinity and the minimal size of true non-zero parameters decreases toward zero as the sample size increases. Further, we construct a practical method to select tuning parameters using generalized information criterion, of which the minimizer asymptotically recovers the best theoretical non-penalized estimator of the sparse AR process. Simulation studies are given to confirm the theoretical results.

Keywords: autoregressive process, subset selection, non-convex penalty, oracle property, tuning parameter selection

1. Introduction

The autoregressive (AR) process is a basic and important processes for time series data analysis. The usual least square estimation may yield severe modeling biases when the AR model is sparse including zero parameters. Various information criteria (Akaike, 1969, 1973, 1979; Schwarz, 1978; Hannan and Quinn, 1979; Claeskens and Hjort, 2003) have been proposed to identify true non-zero parameters of the AR process, which we call subset selection problem in this paper. Theoretical properties such as asymptotic efficiency and selection consistency of the final sub-process from these information criteria have also been investigated (Shibata, 1976; Hannan and Quinn, 1979; Tsay, 1984; Claeskens and Hjort, 2003; Claeskens *et al.*, 2007). Recently, Na (2017) introduced the generalized information criterion (GIC) (Kim *et al.*, 2012) for the AR process that includes most of information criteria such as Akaike information criterion (AIC) (Akaike, 1973), Hannan-Quinn criterion (HQC) (Hannan, 1980) and Bayesian information criterion (BIC) (Schwarz, 1978). Na (2017) proved that there is a large class of GICs that is selection consistent, including the BIC as an example. However, these approaches suffer from computational complexity since it is almost impossible to compare all the candidate sub-processes when the maximal order is very large (McClave, 1975; Sarkar and Kanjilal, 1995; Chen, 1999; McLeod and Zhang, 2006).

For years, penalized estimation has been studied as an alternative for the subset selection problem (Nardi and Rinaldo, 2011; Schmidt and Makalic, 2013; Sang and Sun, 2015; Kwon *et al.*, 2017). The penalized estimation has nice asymptotic properties such as selection consistency and minimax

[1] Corresponding author: Department of Applied Statistics, Konkuk University, 120 Neungdong-ro, Gwangjin-gu, Seoul 05029, Korea. E-mail: shkwon0522@konkuk.ac.kr

Published 30 September 2018 / journal homepage: http://csam.or.kr

optimality for various statistical models that include the generalized linear regression model (Fan and Peng, 2004; Zou, 2006; Ye and Zhang, 2010; Kwon and Kim, 2012). However, the advantage of the penalized estimation comes from the efficiency of the computation since there exist many fast and efficient algorithms (Friedman *et al.*, 2007; Kim *et al.*, 2008; Lee *et al.*, 2016). Hence we need not to exhaustively search all the possible candidate sub-models when the AR process has very large model order.

There are many possible penalty functions for the penalized estimation such as the least absolute selection and shrinkage operator (LASSO) (Tibshirani, 1996) and smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001). There are some non-convex penalties including the SCAD that have very distinct advantages against others such as the bridge (Huang *et al.*, 2008; Kim *et al.*, 2016) and log (Zou and Li, 2008; Kwon *et al.*, 2016). First, they produce unbiased estimators of the parameters that help us understand the final model without considering the penalty effect. Second, they exactly select the non-zero parameters with probability tending to one which is impossible for other penalty functions such as the LASSO (Zhang, 2010b; Kim and Kwon, 2012; Zhang and Zhang, 2012) and ridge.

In this paper, we study the subset selection problem by using the non-convex penalized estimation used to identify non-zero parameters in various statistical models (Fan and Li, 2001; Zhang, 2010a; Kwon and Kim, 2012; Shen *et al.*, 2013). A large class of non-convex penalties is considered that includes the SCAD and minimax concave penalties (MCP) (Zhang, 2010a) as examples (Kim and Kwon, 2012; Zhang and Zhang, 2012). We first prove several asymptotic properties of the non-convex penalized estimators such as the weak and strong oracle properties that are standard in sparse linear regression framework (Kim *et al.*, 2016). Second, we prove that optimal tuning parameters in the penalty can be chosen by using an information criterion of which the minimizer exactly identifies true zero and non-zero parameters asymptotically.

The results hold when the candidate maximal order of the AR process increases to infinity and the minimal size of the true non-zero parameters decreases toward zero as the sample size increases. Further, we consider a class of error processes for the AR process that includes independently and identically distributed (iid), autoregressive conditional heteroscedastic (ARCH) and generalized ARCH (GARCH) processes, which is large enough to cover most of the recent and related literature (Nardi and Rinaldo, 2011; Schmidt and Makalic, 2013; Sang and Sun, 2015; Kwon *et al.*, 2017).

We introduce the non-convex penalized estimation for the AR process in Section 2. Asymptotic properties of the penalized estimator are presented in Section 3, introducing an information criterion for the tuning parameter selection. Simulation studies and proofs are given in Section 4 and Appendix, respectively.

2. Non-convex penalized estimation for the AR process

Consider the AR process $\{y_t, t \in \mathbb{Z}\}$,

$$y_t - \mu = \sum_{j=1}^{p} \beta_j (y_{t-j} - \mu) + \varepsilon_t, \quad t \in \mathbb{Z}, \tag{2.1}$$

where p is a positive integer, $\mu \in \mathbb{R}$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p$ are parameters of interest, $\{\varepsilon_t, t \in \mathbb{Z}\}$ is an error process and $\mathbb{R}$ and $\mathbb{Z}$ are the set of real numbers and integers, respectively. We assume that the process is sparse, that is, $\beta_j = 0$ for some $j \in \mathcal{S}_F$, where $\mathcal{S}_F = \{1, 2, \ldots, p\}$ denotes the full index set of regression parameters. In this case, we can estimate the true non-zero index set

$\mathcal{S}_T = \{j : \beta_j \neq 0\} \subset \mathcal{S}_F$ by minimizing the penalized sum of squared residuals

$$L_\lambda(u, \mathbf{b}) = \sum_{t=p+1}^{n} \left\{ (y_t - u) - \sum_{j=1}^{p} b_j (y_{t-j} - u) \right\}^2 + 2(n-p) \sum_{j=1}^{p} J_\lambda(|b_j|) \tag{2.2}$$

with respect to $u \in \mathbb{R}$ and $\mathbf{b} = (b_1, \ldots, b_p)^T \in \mathbb{R}^p$. Here, J_λ is a non-convex penalty with tuning parameter $\lambda > 0$ that is included in the penalty class defined in the next section.

Let $\bar{y}_j = \sum_{t=p+1}^{n} y_{t-j}/(n-p)$ then $L_\lambda(u, \mathbf{b})$ can be decomposed as:

$$L_\lambda(u, \mathbf{b}) = L_\lambda^A(u, \mathbf{b}) + L_\lambda^B(\mathbf{b}),$$

where $L_\lambda^A(u, \mathbf{b}) = (n-p)\{\bar{y}_0 - \sum_{j=1}^{p} b_j \bar{y}_j - (1 - \sum_{j=1}^{p} b_j)u\}^2$ and

$$L_\lambda^B(\mathbf{b}) = \sum_{t=p+1}^{n} \left\{ (y_t - \bar{y}_0) - \sum_{j=1}^{p} b_j (y_{t-j} - \bar{y}_j) \right\}^2 + 2(n-p) \sum_{j=1}^{p} J_\lambda(|b_j|).$$

Let $(\hat{\mu}^{\lambda,A}, \hat{\boldsymbol{\beta}}^{\lambda,B})$ be the minimizer of $L_\lambda(u, \mathbf{b})$ then it is easy to see that

$$\hat{\boldsymbol{\beta}}^{\lambda,B} = \left(\hat{\beta}_1^{\lambda,B}, \ldots, \hat{\beta}_p^{\lambda,B}\right)^T = \arg\min_{\mathbf{b} \in \mathbb{R}^p} L_\lambda^B(\mathbf{b}) \tag{2.3}$$

and $\hat{\mu}^{\lambda,A} = (\bar{y}_0 - \sum_{j=1}^{p} \hat{\beta}_j^{\lambda,B} \bar{y}_j)/(1 - \sum_{j=1}^{p} \hat{\beta}_j^{\lambda,B})$. Once $\hat{\boldsymbol{\beta}}^{\lambda,B}$ is obtained, we can estimate the true index set $\mathcal{S}_T$ by using the set $\{j : \hat{\beta}_j^{\lambda,B} \neq 0\}$. We often estimate μ by using the sample mean $\bar{y} = \sum_{t=1}^{n} y_t/n$ before estimating $\boldsymbol{\beta}$, which is the same as estimating $\boldsymbol{\beta}$ based on the centered samples $y_t - \bar{y}, t = 1, \ldots, n$. In this case, the penalized estimator, $\hat{\boldsymbol{\beta}}^{\lambda,C}$, can be defined as

$$\hat{\boldsymbol{\beta}}^{\lambda,C} = \left(\hat{\beta}_1^{\lambda,C}, \ldots, \hat{\beta}_p^{\lambda,C}\right)^T = \arg\min_{\mathbf{b} \in \mathbb{R}^p} L_\lambda^C(\mathbf{b}), \tag{2.4}$$

where $L_\lambda^C(\mathbf{b}) = L_\lambda(\bar{y}, \mathbf{b})$.

In this paper, we define the penalized estimator of the regression parameter $\boldsymbol{\beta}$ as

$$\hat{\boldsymbol{\beta}}^{\lambda} = \left(\hat{\beta}_1^{\lambda}, \ldots, \hat{\beta}_p^{\lambda}\right)^T = \arg\min_{\mathbf{b} \in \mathbb{R}^p} L_\lambda^G(\mathbf{b}; \mathbf{m}), \tag{2.5}$$

where

$$L_\lambda^G(\mathbf{b}; \mathbf{m}) = Q(\mathbf{b}; \mathbf{m}) + 2(n-p) \sum_{j=1}^{p} J_\lambda(|b_j|), \tag{2.6}$$

$\mathbf{m} = (m_0, m_1, \ldots, m_p)^T \in \mathbb{R}^{p+1}$ and

$$Q(\mathbf{b}; \mathbf{m}) = \sum_{t=p+1}^{n} \left\{ (y_t - m_0) - \sum_{j=1}^{p} b_j \left(y_{t-j} - m_j\right) \right\}^2.$$

The definition is general enough to include above two special cases. If we set $\mathbf{m} = (\bar{y}_0, \ldots, \bar{y}_p)^T$ and $\mathbf{m} = (\bar{y}, \ldots, \bar{y})^T$, then $L_\lambda^G(\mathbf{b}; \mathbf{m})$ becomes $L_\lambda^B(\mathbf{b})$ and $L_\lambda^C(\mathbf{b})$, respectively.

3. Asymptotic properties

In this section, we investigate some asymptotic properties such as weak and strong oracle properties that have been proved in sparse linear regression framework. The results hold when $p \to \infty$ and $\min_{j\in\mathcal{S}_T} |\beta_j| \to 0$ as $n \to \infty$. Further, we propose the GIC that asymptotically recovers the best theoretical non-penalized estimator of the AR process.

3.1. Model assumptions and penalty class

We assume the following conditions (E1)–(E5):

(E1) $\{y_t, t \in \mathbb{Z}\}$ is stationary and there exists an absolutely summable sequence $\{\psi_i, i \in \mathbb{N}\}$ such that

$$y_t - \mu = \varepsilon_t + \sum_{i=1}^{\infty} \psi_i \varepsilon_{t-i}, \quad t \in \mathbb{Z}, \tag{3.1}$$

where $\mathbb{N}$ is the set of natural numbers.

(E2) $\{\varepsilon_t, t \in \mathbb{Z}\}$ is a white noise with mean 0 and positive variance σ_ε^2.

(E3) $\{\varepsilon_t, t \in \mathbb{Z}\}$ is a sequence of martingale differences with respect to a filtration $\{\mathcal{F}_t, t \in \mathbb{Z}\}$.

(E4) $\{\varepsilon_t, t \in \mathbb{Z}\}$ takes the form

$$\varepsilon_t = g(\eta_t, \eta_{t-1}, \ldots),$$

where η_t, $t \in \mathbb{Z}$, are iid random variables and g is a measurable function.

(E5) $E(|\varepsilon_1|^\nu) < \infty$ and $\{\varepsilon_t, t \in \mathbb{Z}\}$ is ν-strong stable with $\nu \geq 2$ (Wu, 2005). Here, ν-strong stability means that

$$\Delta_\nu^\varepsilon = \sum_{i=0}^{\infty} \delta_\nu^\varepsilon(i) < \infty, \tag{3.2}$$

where $\delta_\nu^\varepsilon(i) = \{E(|\varepsilon_i - g(\eta_i, \ldots, \eta_1, \eta_0^*, \eta_{-1}, \ldots)|^\nu)\}^{1/\nu}$ and $\{\eta_t^*, t \in \mathbb{Z}\}$ is an iid copy of $\{\eta_t, t \in \mathbb{Z}\}$.

The error process satisfying conditions (E2)–(E5) includes iid, ARCH and GARCH processes. For example, if the errors are are iid with $E(|\varepsilon_1|^\nu) < \infty$, then $\Delta_\nu^\varepsilon = \delta_\nu^\varepsilon(0) \leq 2\{E(|\varepsilon_1|^\nu)\}^{1/\nu} < \infty$ and $E(\varepsilon_t|\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots) = 0$, and consequently (E2)–(E5) hold. From Bollerslev (1986) and Wu (2011), GARCH processes satisfy conditions (E2)–(E5) also. Let $\gamma : \mathbb{Z} \to \mathbb{R}$ be the autocovariance function of the process $\{y_t, t \in \mathbb{Z}\}$. From (E1)–(E3), $\gamma(0) = \sum_{i=0}^{\infty} \psi_i^2 \sigma_\varepsilon^2 \in (0, \infty)$ and $\gamma(h) = \sum_{i=0}^{\infty} \psi_i \psi_{i+h} \sigma_\varepsilon^2$ converges to 0 as h increases to ∞, where $\psi_0 = 1$. Therefore, the autocovariance matrix $\Gamma_p = (\gamma(i-j))_{1\leq i,j\leq p}$ is positive definite by Proposition 5.1.1 of Brockwell and Davis (2006). Also, γ is an absolutely summable autocovariance function because of the absolute summability of $\{\psi_i, i \in \mathbb{N}\}$.

We consider a class of non-convex penalties where the penalty J_λ satisfies (P1)–(P3):

(P1) There exists a decreasing function $\nabla J_\lambda : [0, \infty) \to [0, \infty)$ such that $J_\lambda(x) = \int_0^x \nabla J_\lambda(t)dt$ for all $x \geq 0$.

(P2) There exists a positive constant a such that $\nabla J_\lambda(x) = 0$ for all $x > a\lambda$.

(P3) $\lambda - x/a \leq \nabla J_\lambda(x) \leq \lambda$ for all $0 \leq x < a\lambda$.

The penalty class has been studied in high dimensional linear regression model (Kim and Kwon, 2012; Zhang and Zhang, 2012; Kim *et al.*, 2016). The class includes the MC and capped ℓ_1 penalties (Zhang and Zhang, 2012; Shen *et al.*, 2013) as special examples that are upper and lower bounds of the class that also include the SCAD penalty. (P1) implies J_λ is a continuous, increasing and concave function defined on $[0, \infty)$ and $J_\lambda(0) = 0$. If J_λ is differentiable, then ∇J_λ is merely the derivative function of J_λ. From conditions (P1)–(P3), we can see that J_λ has locally sub-differentiable at a point $x_0 \in (-\infty, -a\lambda) \cup \{0\} \cup (a\lambda, \infty)$ although it is not convex. By (P2), 0 is the unique local subgradient of J_λ at a point x_0 when $|x_0| > a\lambda$, which makes the non-zero elements of the penalized estimator to be unbiased with finite samples. It is easy to see that $\nabla J_\lambda(x) \geq \lambda/2$ for $0 \leq x \leq a\lambda/2$ from (P3) so that the set of local subgradients of J_λ at the origin includes $[-\lambda/2, \lambda/2]$, which makes the penalized estimator to be sparse. These properties of the subgradients play an important role in constructing the oracle properties.

3.2. Oracle properties

Theorem 1. *(Weak oracle property) Assume that (E1)–(E5) with $\nu \geq 4$ and (P1)–(P3) hold. Let $\kappa = \|\Gamma_p(\mathcal{S}_T)^{-1}\|_\infty$, where $\Gamma_p(\mathcal{S}_T) = (\gamma(i-j))_{i,j\in\mathcal{S}_T}$ is a $q \times q$ submatrix of Γ_p and $\|\mathbf{A}\|_\infty$ is the maximum absolute row sum of a matrix $\mathbf{A}$. If*

$$\lim_{n\to\infty} \frac{\left(1+\kappa^2\right)p^2}{n} = 0, \qquad \lim_{n\to\infty} \frac{\log p + \kappa^2 \log q}{n\lambda^2} = 0, \qquad \lim_{n\to\infty} \frac{\lambda}{\min_{j\in\mathcal{S}_T} |\beta_j|} = 0 \tag{3.3}$$

and $\mathbf{m}$ satisfies

$$\max_{0\leq j\leq p} |m_j - \mu| = O_P\left(\frac{1}{\sqrt{n}}\right), \tag{3.4}$$

then the oracle least squares estimator (LSE)

$$\hat{\boldsymbol{\beta}}^o = \left(\hat{\beta}_1^o, \ldots, \hat{\beta}_p^o\right)^T = \arg\min_{\mathbf{b}\in\mathbb{R}_o^p} Q(\mathbf{b}; \mathbf{m}) \tag{3.5}$$

is a local minimizer of $L_\lambda^G(\mathbf{b}; \mathbf{m})$ with probability tending to 1, where $\mathbb{R}_o^p = \{(x_1, \ldots, x_p)^T \in \mathbb{R}^p : x_j = 0, j \notin \mathcal{S}_T\}$.

Theorem 1 shows that the oracle LSE in (3.5) becomes one of local minimizers of (2.6), which often referred as the weak oracle property (Fan and Li, 2001; Kim *et al.*, 2016) in linear regression model. Note that the result holds for any m_j, $j \in \mathcal{S}_F$ that is $\sqrt{n}$-consistent estimator of μ. For example, we may use the trimmed mean instead of the sample mean when there are some outliers.

The objective function (2.6) is non-convex so that we need a stronger result than the weak oracle property to avoid bad local minimizers (Zhang, 2010a; Kim and Kwon, 2012). The next theorem shows that the oracle LSE is unique so that it becomes unique global minimizer of (2.6).

Theorem 2. *(Strong oracle property) Assume that the assumptions of Theorem 1 hold. Let $\rho = \lambda_{\min}(\Gamma_p)$ where $\lambda_{\min}(\mathbf{A})$ is the smallest eigenvalue of a matrix $\mathbf{A}$. If*

$$\lim_{n\to\infty} \frac{p^2}{n\rho^2} = 0, \qquad \lim_{n\to\infty} \frac{\log p + \kappa^2 \log q}{n\rho^2\lambda^2} = 0, \qquad \lim_{n\to\infty} \frac{\lambda}{\rho \min_{j\in\mathcal{S}_T} |\beta_j|} = 0, \tag{3.6}$$

then $\hat{\boldsymbol{\beta}}^o$ *is the unique local minimizer of* $L_\lambda^G(\mathbf{b};\mathbf{m})$ *with probability tending to 1.*

Remark 1. Note that κ and ρ often assumed to be fixed positive constants. In this case, the results in Theorems 1 and 2 hold when

$$\min_{j\in\mathcal{S}_T}|\beta_j| \gg \lambda \gg \sqrt{\log\frac{p}{n}} \quad \text{and} \quad n \gg p^2,$$

which are exactly the same as the results in linear regression model. Here, $a \gg b$ implies $b/a = o(1)$ as $n \to \infty$.

Remark 2. In the linear regression $\rho = \lambda_{\min}(\mathbf{X}^T\mathbf{X}/n)$ determines the size of possible minimum non-zero regression coefficient, where $\mathbf{X}$ is the design matrix since we need $\min_j|\beta_j| \geq \lambda \geq \sqrt{\log p/n}$ and $\min_j|\beta_j| \geq \lambda/\rho \geq \sqrt{\log p/n\rho^4}$ for the weak and strong oracle properties, respectively. This implies that the smaller ρ is the larger $\min_{j\in\mathcal{S}_T}\beta_j$ is required. In the AR process $\rho = \lambda_{\min}(\Gamma_p)$ and $\kappa = \|\Gamma_p(S_T)^{-1}\|_\infty$ take the same role since we need $\min_j|\beta_j| \geq \lambda \geq \sqrt{(\log p + \kappa^2\log q)/n}$ and $\min_j|\beta_j| \geq \lambda/\rho \geq \sqrt{(\log p + \kappa^2\log q)/n\rho^4}$.

Let $\hat{\boldsymbol{\beta}}^{o,B}$ and $\hat{\boldsymbol{\beta}}^{o,C}$ be the oracle LSEs that correspond to $\mathbf{m} = (\bar{y}_0,\ldots,\bar{y}_p)^T$ and $\mathbf{m} = (\bar{y},\ldots,\bar{y})^T$, respectively. By the functional central limit theorem,

$$\max_{0\leq j\leq p}|\bar{y}_j - \mu| \leq \frac{2}{n-p}\max_{1\leq k\leq n}\left|\sum_{t=1}^{k}(y_t-\mu)\right| = O_P\left(\frac{1}{\sqrt{n}}\right)$$

and $|\bar{y}-\mu| = O_P(1/\sqrt{n})$. Hence, from Theorem 2, we can see that two penalized estimators are exactly the same as the oracle LSEs, respectively, which is summarized in the next corollary.

Corollary 1. *Assume that the assumptions of Theorem 2 hold then*

$$\lim_{n\to\infty}\mathbf{P}\left(\hat{\boldsymbol{\beta}}^{\lambda,B} = \hat{\boldsymbol{\beta}}^{o,B}\right) = 1 \quad \textit{and} \quad \lim_{n\to\infty}\mathbf{P}\left(\hat{\boldsymbol{\beta}}^{\lambda,C} = \hat{\boldsymbol{\beta}}^{o,C}\right) = 1. \tag{3.7}$$

Remark 3. From Lemma 2 in Appendix,

$$\max_{j\in\mathcal{S}_F}\left|\hat{\beta}_j^{o,B} - \beta_j\right| = O_P\left(\kappa\sqrt{\frac{\log q}{n}}\right) \quad \text{and} \quad \max_{j\in\mathcal{S}_F}\left|\hat{\beta}_j^{o,C} - \beta_j\right| = O_P\left(\kappa\sqrt{\frac{\log q}{n}}\right),$$

which implies

$$\max_{j\in\mathcal{S}_F}\left|\hat{\beta}_j^{\lambda,B} - \hat{\beta}_j^{\lambda,C}\right| = O_P\left(\kappa\sqrt{\frac{\log q}{n}}\right).$$

Hence two oracle LSEs are asymptotically equivalent so that centering the samples does not affect regression parameter estimation.

3.3. Tuning parameter selection

The most practical and important issue for the penalized estimation is how to select tuning parameters (Nardi and Rinaldo, 2011; Kwon *et al.*, 2017). Some conventional ways can be applied by minimizing validation or cross-validation errors. However, selecting tuning parameters based on prediction errors may produce over-fitted sub-models (Wang *et al.*, 2007, 2009). We propose to use the GIC, which is easy to calculate without using extra independent samples, in order to select the tuning parameters. Given $\lambda > 0$, define the GIC as

$$\mathrm{GIC}(\lambda) = \log Q\left(\hat{\boldsymbol{\beta}}^{\lambda}; \mathbf{m}\right) + \alpha \left\|\hat{\boldsymbol{\beta}}^{\lambda}\right\|_0, \tag{3.8}$$

where $\|\hat{\boldsymbol{\beta}}^{\lambda}\|_0$ is the number of non-zero entries of $\hat{\boldsymbol{\beta}}^{\lambda}$. The next theorem proves that we can select optimal tuning parameter by minimizing the GIC.

Theorem 3. *Assume that (E1)–(E5) with $\nu \geq 4$, and (P1)–(P3) hold. If*

$$\lim_{n\to\infty} \frac{\left(1 + \kappa^2 + \rho^{-2}\right) p^2}{n} = 0, \tag{3.9}$$

$$\lim_{n\to\infty} \frac{\sqrt{\log p} + \kappa\sqrt{\log q} + \rho\sqrt{p}}{\rho^2 \sqrt{n} \min_{j\in\mathcal{S}_T} |\beta_j|} = 0, \tag{3.10}$$

$$\lim_{n\to\infty} p\alpha = 0, \quad \lim_{n\to\infty} \frac{p\alpha}{\rho \min_{j\in\mathcal{S}_T} \beta_j^2} = 0, \quad \lim_{n\to\infty} \frac{\log p + \kappa^2 \log q}{\rho\alpha n} = 0, \tag{3.11}$$

then there exists λ_0 such that

$$\lim_{n\to\infty} \mathbf{P}\left(\inf_{\lambda\in\Lambda_+\cup\Lambda_-} \mathrm{GIC}(\lambda) > \mathrm{GIC}(\lambda_0)\right) = 1,$$

where $\Lambda_+ = \{\lambda > 0 : \mathcal{S}_\lambda \supsetneq \mathcal{S}_T\}$, $\Lambda_- = \{\lambda > 0 : \mathcal{S}_\lambda \not\supset \mathcal{S}_T\}$ and $\mathcal{S}_\lambda = \{j \in \mathcal{S}_F : \hat{\beta}_j^{\lambda} \neq 0\}$ for $\lambda > 0$.

Remark 4. When κ and ρ are positive constants the result in Theorem 3 holds when

$$\min_{j\in\mathcal{S}_T} |\beta_j| \gg \sqrt{p\alpha} \gg \sqrt{\frac{p \log p}{n}} \quad \text{and} \quad n \gg p^2.$$

For example, $\alpha = \log(\log p) \log n/n$ satisfies the condition as suggested in Kwon *et al.* (2017) and Na (2017) for the adaptive LASSO and GIC, respectively.

4. Numerical studies

We consider two examples to show that the theoretical results hold with finite samples:

Example 1. $\mathcal{S}_T = \{1, 2, \ldots, q\}$, $\beta_j^* = (c_0/\sqrt{j})I(j \in \mathcal{S}_T)$, $\mu = 0$ and ε_t is iid samples from standard normal distribution, where $c_0 = 0.9/(\sum_{j=1}^{q} 1/\sqrt{j})$.

Example 2. $\mathcal{S}_T = \{j_1, j_2, \ldots, j_q\}$ is a random subset of $\mathcal{S}_F$ and $\beta_j^* = \sum_{k=1}^{q} (c_0/\sqrt{k})I(j = j_k)$.

Table 1: Simulation results from Example 1

	n	p	AIC	HQC	BIC	GIC6	ALASSO	SCAD	MCP
SEN	100	50	0.850	0.770	0.700	0.630	0.650	0.650	0.650
	200	60	0.898	0.805	0.715	0.655	0.670	0.668	0.668
	400	70	0.978	0.952	0.898	0.850	0.868	0.878	0.860
	800	90	0.994	0.972	0.934	0.872	0.908	0.908	0.908
	1600	120	0.998	0.992	0.973	0.925	0.955	0.955	0.955
	3200	150	1.000	0.995	0.974	0.925	0.971	0.971	0.970
	6400	190	1.000	1.000	0.996	0.971	0.993	0.993	0.993
	12800	230	1.000	1.000	1.000	0.992	0.997	0.997	0.997
SPC	100	50	0.784	0.909	0.968	0.986	0.982	0.982	0.982
	200	60	0.819	0.941	0.983	0.995	0.990	0.990	0.990
	400	70	0.837	0.947	0.992	0.998	0.996	0.996	0.996
	800	90	0.838	0.955	0.994	0.999	0.997	0.996	0.997
	1600	120	0.828	0.959	0.997	0.999	0.998	0.998	0.998
	3200	150	0.844	0.967	0.998	1.000	0.999	0.999	0.999
	6400	190	0.840	0.967	0.998	1.000	1.000	1.000	0.999
	12800	230	0.869	0.972	0.998	1.000	1.000	1.000	1.000
SA	100	50	0.000	0.040	0.070	0.050	0.080	0.080	0.080
	200	60	0.000	0.030	0.070	0.070	0.060	0.060	0.060
	400	70	0.000	0.120	0.400	0.370	0.420	0.430	0.410
	800	90	0.000	0.110	0.510	0.390	0.460	0.470	0.470
	1600	120	0.000	0.060	0.620	0.490	0.620	0.620	0.630
	3200	150	0.000	0.080	0.600	0.390	0.690	0.700	0.690
	6400	190	0.000	0.050	0.660	0.740	0.860	0.860	0.850
	12800	230	0.000	0.020	0.730	0.890	0.900	0.900	0.900
PE	100	50	1.591	1.388	1.312	1.201	1.260	1.265	1.267
	200	60	1.198	1.126	1.098	1.093	1.122	1.123	1.120
	400	70	1.116	1.074	1.047	1.044	1.056	1.054	1.056
	800	90	1.067	1.034	1.018	1.020	1.026	1.027	1.026
	1600	120	1.038	1.015	1.002	1.004	1.007	1.007	1.007
	3200	150	1.024	1.009	1.002	1.004	1.005	1.005	1.005
	6400	190	1.018	1.008	1.003	1.004	1.004	1.004	1.004
	12800	230	1.009	1.003	1.000	1.000	1.001	1.001	1.001

AIC = Akaike information criterion; HQC = Hannan-Quinn criterion; BIC = Bayesian information criterion; GIC = generalized information criterion; ALASSO = adaptive least absolute selection and shrinkage operator; SCAD = smoothly clipped absolute deviation MCP = minimax concave penalty; SEN = sensitivity; SPC = specificity; SA = selection accuracy; PE = prediction error.

In each example, we set $n = 100 \times 2^k, k \in \{0, 1, \ldots, 7\}$, $p = 10[n^{1/3}]$ and $q = [n^{1/4}]$, where $[x]$ is the closest integer from x.

We consider two non-convex penalties for the centered samples, the SCAD and MCP, and compare finite sample performance to some reference methods: four GICs (AIC, HQC, BIC, and GIC6) in Na (2017) and the adaptive LASSO (ALASSO) in Kwon *et al.* (2017). Tuning parameters are selected by the GIC in (3.8) with $\alpha = \log(\log p) \log n/n$ for all the penalized estimators. We report four measures: sensitivity (SEN), specificity (SPC), selection accuracy (SA), and prediction error (PE) from independent test samples $y_i^t, i \le n$. The measures are defined as $|\hat{\mathcal{S}}_T \cap \mathcal{S}_T|/|\mathcal{S}_T|$, $|\hat{\mathcal{S}}_T^c \cap \mathcal{S}_T^c|/|\mathcal{S}_T^c|$, $I(\hat{\mathcal{S}}_T = \mathcal{S}_T)$, and $\sum_{i=1}^n (y_i^t - \hat{y}_i^t)/n$ respectively, where $\hat{\mathcal{S}}_T$ is the index set of non-zero parameters estimated from the methods. We repeat each simulation 200 times and summarize the results in Tables 1, 2 and Figure 1, where all the standard errors are less than 0.03 and omitted.

The PEs are quite similar for all the methods but selection performance are significantly different. The AIC and HQC have better SEN but worse SPC than the others, which result in a very low SA. This shows that the AIC and HQC are not selection consistent (Na, 2017) but over-fit. The SEN, SPC,

Table 2: Simulation results from Example 2

	n	p	AIC	HQC	BIC	GIC6	ALASSO	SCAD	MCP
SEN	100	50	0.877	0.830	0.747	0.663	0.577	0.563	0.563
	200	60	0.910	0.835	0.718	0.655	0.602	0.598	0.598
	400	70	0.982	0.970	0.940	0.902	0.862	0.865	0.865
	800	90	0.996	0.986	0.966	0.920	0.928	0.924	0.918
	1600	120	0.998	0.993	0.978	0.948	0.967	0.968	0.967
	3200	150	1.000	1.000	0.991	0.958	0.976	0.978	0.976
	6400	190	1.000	1.000	1.000	0.990	0.997	0.997	0.997
	12800	230	1.000	1.000	1.000	0.993	0.998	0.998	0.998
SPC	100	50	0.800	0.897	0.961	0.980	0.985	0.986	0.985
	200	60	0.831	0.943	0.976	0.990	0.989	0.989	0.989
	400	70	0.850	0.946	0.984	0.997	0.993	0.992	0.992
	800	90	0.844	0.960	0.993	0.998	0.995	0.995	0.995
	1600	120	0.834	0.957	0.996	1.000	0.998	0.998	0.998
	3200	150	0.848	0.967	0.996	1.000	0.998	0.998	0.998
	6400	190	0.843	0.964	0.997	1.000	0.999	0.999	0.999
	12800	230	0.853	0.969	0.998	1.000	0.999	0.999	0.999
SA	100	50	0.000	0.010	0.070	0.130	0.130	0.130	0.130
	200	60	0.000	0.060	0.080	0.060	0.070	0.070	0.070
	400	70	0.000	0.070	0.310	0.550	0.330	0.310	0.310
	800	90	0.000	0.070	0.540	0.560	0.520	0.510	0.490
	1600	120	0.000	0.020	0.560	0.650	0.610	0.620	0.620
	3200	150	0.000	0.010	0.590	0.680	0.600	0.610	0.600
	6400	190	0.000	0.020	0.560	0.880	0.780	0.780	0.780
	12800	230	0.000	0.000	0.690	0.910	0.820	0.800	0.830
PE	100	50	1.402	1.282	1.210	1.191	1.210	1.218	1.214
	200	60	1.211	1.128	1.110	1.106	1.133	1.134	1.132
	400	70	1.096	1.055	1.034	1.026	1.050	1.048	1.047
	800	90	1.070	1.034	1.019	1.021	1.032	1.033	1.032
	1600	120	1.045	1.020	1.007	1.006	1.013	1.013	1.012
	3200	150	1.028	1.013	1.006	1.006	1.010	1.010	1.010
	6400	190	1.019	1.010	1.004	1.004	1.006	1.006	1.006
	12800	230	1.010	1.004	1.001	1.001	1.002	1.002	1.002

AIC = Akaike information criterion; HQC = Hannan-Quinn criterion; BIC = Bayesian information criterion; GIC = generalized information criterion; ALASSO = adaptive least absolute selection and shrinkage operator; SCAD = smoothly clipped absolute deviation MCP = minimax concave penalty; SEN = sensitivity; SPC = specificity; SA = selection accuracy; PE = prediction error.

and SA are increasing to 1 as the sample sizes increases for all the penalized estimators as well as the BIC and GIC6. The simulation results confirm that the theoretical properties provided hold for finite samples.

5. Concluding remarks

We present some asymptotic properties of the non-convex penalized estimators for the AR process, when the maximal order is large and minimal signal size is small. The results show that the non-convex penalized estimation can be used for parameter estimation and model identification simultaneously. This paper is intended to provide a theoretical starting point for future studies on other complex time series analysis.

A referee pointed out that assuming increasing p is unusual in real practice because the AR order does not necessarily increase with the sample size. First of all, we completely agree that many or almost all the AR process may have fixed or finite model orders so that p may not be assumed to be

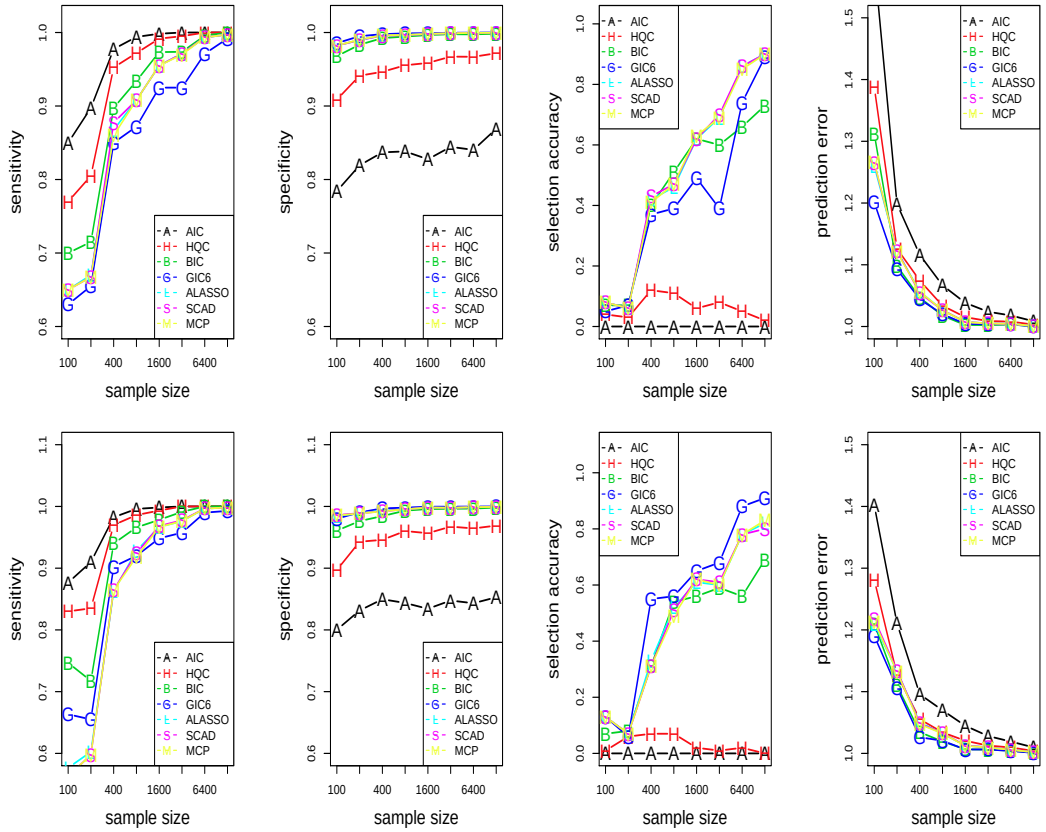

Figure 1: *Sensitivity, specificity, selection accuracy and prediction error in Example 1 (upper) and 2 (lower). AIC = Akaike information criterion; HQC = Hannan-Quinn criterion; BIC = Bayesian information criterion; GIC = generalized information criterion; ALASSO = adaptive least absolute selection and shrinkage operator; SCAD = smoothly clipped absolute deviation MCP = minimax concave penalty.*

increasing or varying, being independent of the sample size. However we think that the larger the sample size the more parameters become statistically significant, which implies that a fixed choice of small p may prevent us from finding important lags. In this sense, the expression "$p \to \infty$ as $n \to \infty$" does not imply the existence of such an order-increasing model but implies that we must try an order as large as possible, considering the sample size. Besides $p \to \infty$, the expression "$\min_j |\beta_j| \to 0$" can be understood in a similar way.

Acknowledgements

This work was supported by a Kyonggi University Research Grant 2015-099.

Appendix:

Let $x_{ij} = y_{i-j+p} - m_j$ and $z_{ij} = y_{i-j+p} - \mu$ for $i = 1, \ldots, n-p$ and $j = 1, \ldots, p$. Let $\mathbf{X} = (x_{ij})$, $\mathbf{Z} = (z_{ij})$, $\mathbf{y} = (y_{p+1}, \ldots, y_n)^T$, and $\boldsymbol{\varepsilon} = (\varepsilon_{p+1}, \ldots, \varepsilon_n)^T$. For $j = 1, \ldots, p$, let $\mathbf{X}_j$ and $\mathbf{Z}_j$ be the j^{th} column vectors

of $\mathbf{X}$ and $\mathbf{Z}$, respectively. For a set $\mathcal{S} = \{i_1, i_2, \ldots, i_k\}$ with $1 \le i_1 < i_2 < \cdots < i_k \le p$, $\mathbf{X}_\mathcal{S}$ denotes the $(n-p) \times k$ matrix with j^{th} column vector $\mathbf{X}_{i_j}$ and $\mathbf{Z}_\mathcal{S}$ is defined similarly. Given a p-dimensional vector $\mathbf{v} = (v_1, \ldots, v_p)^T$ and a set $\mathcal{S} = \{i_1, \ldots, i_k\}$, $\mathbf{v}_\mathcal{S}$ denotes a k-dimensional subvector $(v_{i_1}, \ldots, v_{i_k})^T$ of $\mathbf{v}$. For a matrix A, $\|A\|_2$ and $\|A\|_\infty$ are the induced matrix norms using the Euclidean and maximum norms for vectors, respectively. And for a set $\mathcal{S}$, $|\mathcal{S}|$ means the cardinality of $\mathcal{S}$.

Lemma 1. *Assume that* (2.1)*,* (3.1)*, (E1)–(E5) with* $\nu \ge 4$*, and* (3.4) *hold. If* $\lim_{n\to\infty} p^2/n = 0$*, then*

$$\max_{j\in\mathcal{S}_T} \left|\mathbf{X}_j^T \boldsymbol{\varepsilon}\right| = O_P\left(\sqrt{n \log q}\right), \tag{A.1}$$

$$\max_{j\in\mathcal{S}_F} \left|\mathbf{X}_j^T \boldsymbol{\varepsilon}\right| = O_P\left(\sqrt{n \log p}\right), \tag{A.2}$$

$$\max_{j\in\mathcal{S}_F} \left|\mathbf{X}_j^T \mathbf{u}\right| = O_P\left(\sqrt{n}\right), \tag{A.3}$$

$$\left\| \frac{\mathbf{X}^T\mathbf{X}}{n-p} - \Gamma_p \right\|_\infty = O_P\left(\frac{p}{\sqrt{n}}\right), \tag{A.4}$$

where $\mathbf{u}$ *is a* $(n-p)$*-dimensional vector with all entries 1.*

Proof: First, we can show that there exists a positive constant M such that

$$\lim_{n\to\infty} \mathbf{P}\left(\max_{j\in\mathcal{S}} \left|\mathbf{Z}_j \boldsymbol{\varepsilon}\right| > M\sqrt{n \log |\mathcal{S}|}\right) = 0$$

for all non-empty subset $\mathcal{S} \subset \mathcal{S}_F$ in a similar method to the proof of Lemma 1 in Kwon *et al.* (2017), because $\{z_{ij}\}$ is a stationary predictable process and $\{\varepsilon_i\}$ is a sequence of stationary martingale differences with respect to $\{\mathcal{F}_i\}$. Since $\mathbf{X}_j - \mathbf{Z}_j = (\mu - m_j)\mathbf{u}$ for all $j \in \mathcal{S}_F$ and $\mathbf{u}^T\boldsymbol{\varepsilon} = O_P(\sqrt{n})$, (3.4) implies that

$$\sup_{\emptyset\neq\mathcal{S}\subset\mathcal{S}_F} \left| \max_{j\in\mathcal{S}} \left|\mathbf{X}_j^T\boldsymbol{\varepsilon}\right| - \max_{j\in\mathcal{S}} \left|\mathbf{Z}_j^T\boldsymbol{\varepsilon}\right| \right| \le \max_{j\in\mathcal{S}_F} \left|\left(\mathbf{X}_j - \mathbf{Z}_j\right)^T \boldsymbol{\varepsilon}\right| \le \max_{j\in\mathcal{S}_F} \left|m_j - \mu\right| \times \left|\mathbf{u}^T\boldsymbol{\varepsilon}\right| = O_P(1).$$

Therefore, for a non-empty subset $\mathcal{S} \subset \mathcal{S}_F$

$$\max_{j\in\mathcal{S}} \left|\mathbf{X}_j^T\boldsymbol{\varepsilon}\right| = \max_{j\in\mathcal{S}} \left|\mathbf{Z}_j^T\boldsymbol{\varepsilon}\right| + O_P(1) = O_P\left(\sqrt{n \log |\mathcal{S}|}\right)$$

and hence (A.1) and (A.2) hold.

Next, note that $\{y_t, t \in \mathbb{Z}\}$ is ν-strong stable under assumptions (E1)–(E5). Applying the functional central limit theorem to $\{y_t, t \in \mathbb{Z}\}$ yields

$$\max_{j\in\mathcal{S}_F} \left|\mathbf{Z}_j^T\mathbf{u}\right| = O_P\left(\sqrt{n}\right) \tag{A.5}$$

and hence

$$\max_{j\in\mathcal{S}_F} \left|\mathbf{X}_j^T\mathbf{u}\right| = \max_{j\in\mathcal{S}_F} \left|\mathbf{Z}_j^T\mathbf{u} - \left(m_j - \mu\right)\mathbf{u}^T\mathbf{u}\right| \le \max_{j\in\mathcal{S}_F} \left|\mathbf{Z}_j^T\mathbf{u}\right| + (n-p) \times \max_{j\in\mathcal{S}_F} \left|m_j - \mu\right| = O_P\left(\sqrt{n}\right) \tag{A.6}$$

by (3.4).

Lastly, we can extend Lemma 1 in Kwon *et al.* (2017) to the case of $\mathbf{Z}$ and obtain

$$\left\| \frac{\mathbf{Z}^T\mathbf{Z}}{n-p} - \Gamma_p \right\|_\infty = O_P\left(\frac{p}{\sqrt{n}}\right) \tag{A.7}$$

due to the ν-strong stability of $\{y_t, t \in \mathbb{Z}\}$. Also, for any $i, j \in \mathcal{S}_F$

$$\begin{aligned} \left|\mathbf{X}_i^T\mathbf{X}_j - \mathbf{Z}_i^T\mathbf{Z}_j\right| &= \left|(\mu - m_i)(\mu - m_j)\mathbf{u}^T\mathbf{u} + (\mu - m_i)\mathbf{u}^T\mathbf{Z}_j + (\mu - m_j)\mathbf{u}^T\mathbf{Z}_i\right| \\ &\le (n-p) \times \left(\max_{j\in\mathcal{S}_F}\left|m_j - \mu\right|\right)^2 + 2\max_{j\in\mathcal{S}_F}\left|m_j - \mu\right| \times \max_{j\in\mathcal{S}_F}\left|\mathbf{u}^T\mathbf{Z}_j\right|. \end{aligned} \tag{A.8}$$

Therefore, from (A.5)–(A.8) and (3.4), we can deduce that

$$\left\| \frac{\mathbf{X}^T\mathbf{X}}{n-p} - \Gamma_p \right\|_\infty \le \left\| \frac{\mathbf{Z}^T\mathbf{Z}}{n-p} - \Gamma_p \right\|_\infty + \max_{1\le i\le p}\sum_{j=1}^{p}\frac{\left|\mathbf{X}_i^T\mathbf{X}_j - \mathbf{Z}_i^T\mathbf{Z}_j\right|}{n-p} = O_P\left(\frac{p}{\sqrt{n}}\right).$$

□

Lemma 2. *Assume that the assumptions of Lemma 1 hold and that* $\lim_{n\to\infty}\kappa^2 p^2/n = 0$*. Then,*

$$\max_{j\in\mathcal{S}_F}\left|\hat{\beta}_j^o - \beta_j\right| = O_P\left(\frac{\kappa\sqrt{\log q}}{\sqrt{n}}\right). \tag{A.9}$$

Proof: First, $\hat{\boldsymbol{\beta}}_{\mathcal{S}_T}^o$ can be written as follows

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\mathcal{S}_T}^o &= \left(\mathbf{X}_{\mathcal{S}_T}^T\mathbf{X}_{\mathcal{S}_T}\right)^{-1}\mathbf{X}_{\mathcal{S}_T}^T(\mathbf{y} - m_0\mathbf{u}) \\ &= \boldsymbol{\beta}_{\mathcal{S}_T} + \left(\mathbf{X}_{\mathcal{S}_T}^T\mathbf{X}_{\mathcal{S}_T}\right)^{-1}\mathbf{X}_{\mathcal{S}_T}^T\boldsymbol{\varepsilon} + \left(\mathbf{X}_{\mathcal{S}_T}^T\mathbf{X}_{\mathcal{S}_T}\right)^{-1}\mathbf{X}_{\mathcal{S}_T}^T\mathbf{u} \times \left\{\sum_{j\in\mathcal{S}_T}(m_j - \mu)\beta_j - (m_0 - \mu)\right\}, \end{aligned}$$

provided that $\mathbf{X}_{\mathcal{S}_T}^T\mathbf{X}_{\mathcal{S}_T}$ is non-singular. Therefore,

$$\begin{aligned} \max_{j\in\mathcal{S}_F}\left|\hat{\beta}_j^o - \beta_j\right| &= \max_{j\in\mathcal{S}_T}\left|\hat{\beta}_j^o - \beta_j\right| \\ &\le \left\|\left(\mathbf{X}_{\mathcal{S}_T}^T\mathbf{X}_{\mathcal{S}_T}\right)^{-1}\right\|_\infty \times \left\{\max_{j\in\mathcal{S}_T}\left|\mathbf{X}_j^T\boldsymbol{\varepsilon}\right| + \max_{j\in\mathcal{S}_F}\left|\mathbf{X}_j^T\mathbf{u}\right| \times \max_{0\le j\le p}\left|m_j - \mu\right| \times \left(1 + \sum_{j\in\mathcal{S}_T}\left|\beta_j\right|\right)\right\}. \end{aligned}$$

If $\mathbf{X}_{\mathcal{S}_T}$ satisfies that

$$\left\| \frac{\mathbf{X}_{\mathcal{S}_T}^T\mathbf{X}_{\mathcal{S}_T}}{n-p} - \Gamma_p(\mathcal{S}_T) \right\|_\infty \times \left\|\Gamma_p(\mathcal{S}_T)^{-1}\right\|_\infty < \frac{1}{2}, \tag{A.10}$$

then $\mathbf{X}_{\mathcal{S}_T}^T\mathbf{X}_{\mathcal{S}_T}/(n-p)$ is non-singular and

$$\begin{aligned} \left\|\left(\frac{\mathbf{X}_{\mathcal{S}_T}^T\mathbf{X}_{\mathcal{S}_T}}{n-p}\right)^{-1}\right\|_\infty &\le \left\|\Gamma_p(\mathcal{S}_T)^{-1}\right\|_\infty + \left\|\left(\frac{\mathbf{X}_{\mathcal{S}_T}^T\mathbf{X}_{\mathcal{S}_T}}{n-p}\right)^{-1} - \Gamma_p(\mathcal{S}_T)^{-1}\right\|_\infty \\ &\le \left\|\Gamma_p(\mathcal{S}_T)^{-1}\right\|_\infty + \frac{\left\|\mathbf{X}_{\mathcal{S}_T}^T\mathbf{X}_{\mathcal{S}_T}/(n-p) - \Gamma_p(\mathcal{S}_T)\right\|_\infty \left\|\Gamma_p(\mathcal{S}_T)^{-1}\right\|_\infty^2}{1 - \left\|\mathbf{X}_{\mathcal{S}_T}^T\mathbf{X}_{\mathcal{S}_T}/(n-p) - \Gamma_p(\mathcal{S}_T)\right\|_\infty \times \left\|\Gamma_p(\mathcal{S}_T)^{-1}\right\|_\infty} \\ &\le 2\left\|\Gamma_p(\mathcal{S}_T)^{-1}\right\|_\infty. \end{aligned}$$

Since $\lim_{n\to\infty} \kappa^2 p^2/n = 0$, the probability that (A.10) holds tends to 1 as $n \to \infty$ by (A.4). Consequently,

$$\left\|\left(\mathbf{X}_{\mathcal{S}_T}^T \mathbf{X}_{\mathcal{S}_T}\right)^{-1}\right\|_\infty = O_P\left(\frac{\kappa}{n}\right).$$

According to (A.1), (A.3), and (3.4),

$$\max_{j\in\mathcal{S}_T} \left|\mathbf{X}_j^T \boldsymbol{\varepsilon}\right| + \max_{j\in\mathcal{S}_F} \left|\mathbf{X}_j^T \mathbf{u}\right| \times \max_{0\le j\le p} \left|m_j - \mu\right| \times \left(1 + \sum_{j\in\mathcal{S}_T} |\beta_j|\right) = O_P\left(\sqrt{n\log q}\right) + O_P(q)$$

and this completes the proof. □

Now, let us prove the main theorems in Section 3.

Proof of Theorem 1: Since $\hat{\boldsymbol{\beta}}_{\mathcal{S}_T}^o$ satisfies the normal equations $\mathbf{X}_{\mathcal{S}_T}^T \mathbf{X}_{\mathcal{S}_T} \hat{\boldsymbol{\beta}}_{\mathcal{S}_T}^o = \mathbf{X}_{\mathcal{S}_T}^T (\mathbf{y} - m_0\mathbf{u})$ and $\hat{\boldsymbol{\beta}}_{\mathcal{S}_F\cap\mathcal{S}_T^c} = \mathbf{0}$, we have

$$\mathbf{X}_{\mathcal{S}_T}\left(\mathbf{y} - m_0\mathbf{u} - \mathbf{X}\hat{\boldsymbol{\beta}}^o\right) = \mathbf{0}.$$

Therefore, if $\min_{j\in\mathcal{S}_T} |\hat{\beta}_j^o| > a\lambda$ and $\max_{j\notin\mathcal{S}_T} |\mathbf{X}_j^T(\mathbf{y} - m_0\mathbf{u} - \mathbf{X}\hat{\boldsymbol{\beta}}^o)| \le n\lambda/2$, then $\hat{\boldsymbol{\beta}}^o$ becomes a local minimizer of $L_\lambda^G(\mathbf{b};\mathbf{m})$ by the KKT conditions and (P1)–(P3).

From (3.3) and Lemma 2, we have

$$\lim_{n\to\infty} \mathbf{P}\left(\min_{j\in\mathcal{S}_T} \left|\hat{\beta}_j^o\right| > a\lambda\right) \ge \lim_{n\to\infty} \mathbf{P}\left(\min_{j\in\mathcal{S}_T} \left|\beta_j\right| - \max_{j\in\mathcal{S}_F} \left|\hat{\beta}_j^o - \beta_j\right| > a\lambda\right) = 1.$$

By using the results of Lemmas 1 and 2 and the boundedness of $\|\Gamma_p\|_\infty$, we obtain that

$$\begin{aligned}
&\max_{j\notin\mathcal{S}_T} \left|\mathbf{X}_j^T\left(\mathbf{y} - m_0\mathbf{u} - \mathbf{X}\hat{\boldsymbol{\beta}}^o\right)\right| \\
&\le \max_{j\notin\mathcal{S}_T} \left|\mathbf{X}_j^T \mathbf{X}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^o\right)\right| + \max_{j\notin\mathcal{S}_T} \left|\mathbf{X}_j^T\boldsymbol{\varepsilon}\right| + \max_{j\notin\mathcal{S}_T} \left|\mathbf{X}_j^T\left\{(\mathbf{Z} - \mathbf{X})\boldsymbol{\beta} + (\mu - m_0)\mathbf{u}\right\}\right| \\
&\le \left\|\mathbf{X}^T\mathbf{X}\right\|_\infty \max_{j\in\mathcal{S}_F} \left|\hat{\beta}_j^o - \beta_j\right| + \max_{j\in\mathcal{S}_F} \left|\mathbf{X}_j^T\boldsymbol{\varepsilon}\right| + \max_{j\in\mathcal{S}_F} \left|\mathbf{X}_j^T\mathbf{u}\right| \max_{0\le j\le p} \left|m_j - \mu\right| \left(1 + \sum_{j\in\mathcal{S}_T} |\beta_j|\right) \\
&= O_P\left(\kappa\sqrt{n\log q}\right) + O_P\left(\sqrt{n\log p}\right) + O_P(q) \qquad \text{(A.11)}
\end{aligned}$$

and hence (3.3) implies that

$$\lim_{n\to\infty} \mathbf{P}\left(\max_{j\notin\mathcal{S}_T} \left|\mathbf{X}_j^T\left(\mathbf{y} - m_0 - \mathbf{X}\hat{\boldsymbol{\beta}}^o\right)\right| \le \frac{n\lambda}{2}\right) = 1.$$

Therefore

$$\lim_{n\to\infty} \mathbf{P}\left(\hat{\boldsymbol{\beta}}^o \in \mathcal{A}_\lambda\right) \ge \lim_{n\to\infty} \mathbf{P}\left(\min_{j\in\mathcal{S}_T} \left|\hat{\beta}_j^o\right| > a\lambda\right) + \lim_{n\to\infty} \mathbf{P}\left(\max_{j\notin\mathcal{S}_T} \left|\mathbf{X}_j^T\left(\mathbf{y} - m_0 - \mathbf{X}\hat{\boldsymbol{\beta}}^o\right)\right| \le \frac{n\lambda}{2}\right) - 1 = 1,$$

where $\mathcal{A}_\lambda$ is the set of all local minimizers of $L_\lambda^G(\mathbf{b};\mathbf{m})$. □

Proof of Theorem 2: For given $\lambda > 0$ and $c > 0$, let $\mathcal{U}_c$ be the set of $\mathbf{b} \in \mathbb{R}^p$ such that

$$\frac{\max_{b_j=0} \left|\mathbf{X}_j^T (\mathbf{y} - m_0\mathbf{u} - \mathbf{X}\mathbf{b})\right|}{n-p} < \inf_{0<x<a\lambda} (cx + \nabla J_\lambda(x)) \le \sup_{0<x<a\lambda} (cx + \nabla J_\lambda(x)) < c \min_{b_j \neq 0} |b_j|.$$

Let $\hat{\rho}$ be the smallest eigenvalue of $\mathbf{X}^T\mathbf{X}/(n-p)$. By Theorem 1 of Kim and Kwon (2012), $\hat{\beta}^o \in \mathcal{A}_\lambda \cap \mathcal{U}_{\hat{\rho}}$ with $\hat{\rho} > 0$ is a sufficient condition to be a unique local minimizer of $L_\lambda^G(\mathbf{b}; \mathbf{m})$. Also, since $\lim_{n\to\infty} \mathbf{P}(\hat{\boldsymbol{\beta}}^o \in \mathcal{A}_\lambda) = 1$ by Theorem 1, it is enough to show

$$\lim_{n\to\infty} \mathbf{P}\left(\hat{\boldsymbol{\beta}}^o \in \mathcal{U}_{\hat{\rho}}, \hat{\rho} > 0\right) = 1 \tag{A.12}$$

in order to obtain the result of Theorem 2.

Using the condition $\lim_{n\to\infty} p^2/(n\rho^2) = 1$, (A.4) and the symmetry of $\mathbf{X}^T\mathbf{X}/(n-p) - \Gamma_p$,

$$\lim_{n\to\infty} \mathbf{P}\left(\left\|\frac{\mathbf{X}^T\mathbf{X}}{n-p} - \Gamma_p\right\|_2 > \frac{\rho}{2}\right) \le \lim_{n\to\infty} \mathbf{P}\left(\left\|\frac{\mathbf{X}^T\mathbf{X}}{n-p} - \Gamma_p\right\|_\infty > \frac{\rho}{2}\right) = 0. \tag{A.13}$$

And if $\|\mathbf{X}^T\mathbf{X}/(n-p) - \Gamma_p\|_2 \le \rho/2$, then $\hat{\rho} \ge \rho/2 > 0$ and hence $\mathcal{U}_{\rho/2} \subset \mathcal{U}_{\hat{\rho}}$. Therefore, we have

$$\begin{aligned}
\lim_{n\to\infty} \mathbf{P}\left(\hat{\boldsymbol{\beta}}^o \in \mathcal{U}_{\hat{\rho}}, \hat{\rho} > 0\right) &\ge \lim_{n\to\infty} \mathbf{P}\left(\hat{\boldsymbol{\beta}}^o \in \mathcal{U}_{\hat{\rho}}, \hat{\rho} > 0, \left\|\frac{\mathbf{X}^T\mathbf{X}}{n-p} - \Gamma_p\right\|_2 \le \frac{\rho}{2}\right) \\
&\ge \lim_{n\to\infty} \mathbf{P}\left(\hat{\boldsymbol{\beta}}^o \in \mathcal{U}_{\frac{\rho}{2}}, \left\|\frac{\mathbf{X}^T\mathbf{X}}{n-p} - \Gamma_p\right\|_2 \le \frac{\rho}{2}\right) \\
&\ge \lim_{n\to\infty} \mathbf{P}\left(\hat{\boldsymbol{\beta}}^o \in \mathcal{U}_{\frac{\rho}{2}}\right).
\end{aligned}$$

Note that $\lim_{n\to\infty} \mathbf{P}(\min_{j\in\mathcal{S}_T} |\beta_j| > 2\max_{j\in\mathcal{S}_F} |\hat{\beta}_j^o - \beta_j|) = 1$ by (3.3) and Lemma 2. In this case, $\min_{j\in\mathcal{S}_T} |\hat{\beta}_j| > \min_{j\in\mathcal{S}_T} |\beta_j|/2$ and hence $\{j \in \mathcal{S}_F : \hat{\beta}_j^o \neq 0\} = \mathcal{S}_T$. Since $\inf_{0<x<a\lambda}(\rho x/2 + \nabla J_\lambda(x)) \ge \min(\lambda, a\rho\lambda/2)$ and $\sup_{0<x<a\lambda}(x + 2\nabla J_\lambda(x)/\rho) \le a\lambda + 2\lambda/\rho$ by (P3), we have

$$\begin{aligned}
\lim_{n\to\infty} \mathbf{P}\left(\hat{\boldsymbol{\beta}}^o \in \mathcal{U}_{\frac{\rho}{2}}\right) \ge \lim_{n\to\infty} \mathbf{P}&\left(\frac{\max_{j\notin\mathcal{S}_T} \left|\mathbf{X}_j^T\left(\mathbf{y} - m_0\mathbf{u} - \mathbf{X}\hat{\boldsymbol{\beta}}^o\right)\right|}{n-p} < \min\left(\lambda, \frac{a\rho\lambda}{2}\right)\right) \\
&+ \lim_{n\to\infty} \mathbf{P}\left(\min_{j\in\mathcal{S}_T} |\hat{\beta}_j^o| > a\lambda + \frac{2\lambda}{\rho}, \min_{j\in\mathcal{S}_T} |\beta_j| > 2\max_{j\in\mathcal{S}_F} |\hat{\beta}_j^o - \beta_j|\right) - 1 \\
&= 1
\end{aligned}$$

by (3.6) and (A.11). Hence, (A.12) holds. □

Lemma 3. *For a given $\mathcal{S} \subset \mathcal{S}_F$, let $\tilde{\boldsymbol{\beta}}^{\mathcal{S}} = \arg\min_{\mathbf{b}\in\mathbb{R}^p_{\mathcal{S}}} Q(\mathbf{b}; \mathbf{m})$ and define*

$$\widetilde{\mathrm{GIC}}(\mathcal{S}) = \log Q\left(\tilde{\boldsymbol{\beta}}^{\mathcal{S}}; \mathbf{m}\right) + \alpha|\mathcal{S}|,$$

where $\mathbb{R}^p_{\mathcal{S}} = \{(x_1, \ldots, x_p)^T \in \mathbb{R}^p : x_j = 0, j \notin \mathcal{S}\}$. Under the assumptions of Theorem 3,

$$\lim_{n\to\infty} \mathbf{P}\left(\inf_{\mathcal{S}\neq\mathcal{S}_T} \widetilde{\mathrm{GIC}}(\mathcal{S}) > \widetilde{\mathrm{GIC}}(\mathcal{S}_T)\right) = 1. \tag{A.14}$$

Proof: First, note that ρ is positive, $\|\Gamma_p\|_2$ is bounded, and

$$\lim_{n\to\infty} \mathbf{P}\left(\frac{\rho}{2} \le \hat{\rho} \le \left\|\frac{\mathbf{X}^T\mathbf{X}}{n-p}\right\|_2 \le 2\left\|\Gamma_p\right\|_2\right) \ge \lim_{n\to\infty} \mathbf{P}\left(\left\|\frac{\mathbf{X}^T\mathbf{X}}{n-p} - \Gamma_p\right\| \le \frac{\rho}{2}\right) = 1 \tag{A.15}$$

by (A.13). Given a set $\mathcal{S} \subset \mathcal{S}_F$, we can show that $\|\mathbf{Z}_{\mathcal{S}}^T\boldsymbol{\varepsilon}\|_2 = O_P(\sqrt{n|\mathcal{S}|})$ by extending Lemma 1 of Na (2017). Since $\max_{j\in\mathcal{S}_F} |\mathbf{X}_j^T\boldsymbol{\varepsilon} - \mathbf{Z}_j^T\boldsymbol{\varepsilon}| = \max_{j\in\mathcal{S}_F} |m_j - \mu|\|\mathbf{u}^T\boldsymbol{\varepsilon}| = O_P(1)$,

$$\left\|\mathbf{X}_{\mathcal{S}}^T\boldsymbol{\varepsilon}\right\|_2 \le \left\|\mathbf{Z}_{\mathcal{S}}^T\boldsymbol{\varepsilon}\right\|_2 + \sqrt{|\mathcal{S}|}\max_{j\in\mathcal{S}_F}\left|\mathbf{X}_j^T\boldsymbol{\varepsilon} - \mathbf{Z}_j^T\boldsymbol{\varepsilon}\right| = O_P\left(\sqrt{n|\mathcal{S}|}\right). \tag{A.16}$$

From (A.6), we have

$$\left\|\mathbf{X}_{\mathcal{S}}^T\mathbf{u}\right\|_2 \le \sqrt{|\mathcal{S}|}\max_{j\in\mathcal{S}_F}\left|\mathbf{X}_j^T\mathbf{u}\right| = O_P\left(\sqrt{n|\mathcal{S}|}\right). \tag{A.17}$$

Therefore, combining (A.15)–(A.17) and (3.4) gives that

$$\begin{aligned}\left\|\tilde{\boldsymbol{\beta}}^{\mathcal{S}_F} - \boldsymbol{\beta}\right\|_2 I\left(\hat{\rho} \ge \frac{\rho}{2}\right) &\le \left\|\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right\|_2 \left\{\left\|\mathbf{X}^T\boldsymbol{\varepsilon}\right\|_2 + \left\|\mathbf{X}^T\mathbf{u}\right\|_2 \max_{0\le j\le p}|m_j - \mu|\left(1 + \sum_{j\in\mathcal{S}_T}|\beta_j|\right)\right\} \\ &= O_P\left(\frac{\sqrt{p}}{\sqrt{n\rho^2}}\right)\end{aligned}$$

and consequently, we have

$$\left\|\tilde{\boldsymbol{\beta}}^{\mathcal{S}_F} - \boldsymbol{\beta}\right\|_2 = O_P\left(\frac{\sqrt{p}}{\sqrt{n\rho^2}}\right) \quad \text{and} \quad \frac{Q\left(\tilde{\boldsymbol{\beta}}^{\mathcal{S}_F};\mathbf{m}\right)}{n-p} = \sigma_\epsilon^2 + o_P(1). \tag{A.18}$$

For a while, we assume that $\min_{j\in\mathcal{S}_T}|\beta_j| \ge 2\|\tilde{\boldsymbol{\beta}}^{\mathcal{S}_F} - \boldsymbol{\beta}\|_2$, $\hat{\rho} \ge \rho/2$, and $Q(\tilde{\boldsymbol{\beta}}^{\mathcal{S}_F};\mathbf{m}) \ge (n-p)\sigma_\epsilon^2/2$. Then for any $\mathcal{S} \supsetneq \mathcal{S}_T$,

$$\begin{aligned}\log Q\left(\tilde{\boldsymbol{\beta}}^{\mathcal{S}};\mathbf{m}\right) - \log Q\left(\tilde{\boldsymbol{\beta}}^{\mathcal{S}_F};\mathbf{m}\right) &\ge \min\left\{\frac{\left(Q\left(\tilde{\boldsymbol{\beta}}^{\mathcal{S}};\mathbf{m}\right) - Q\left(\tilde{\boldsymbol{\beta}}^{\mathcal{S}_F};\mathbf{m}\right)\right)}{\left(2Q\left(\tilde{\boldsymbol{\beta}}^{\mathcal{S}_F};\mathbf{m}\right)\right)}, \frac{1}{2}\right\} \\ &\ge \min\left\{\frac{\hat{\rho}\left\|\tilde{\boldsymbol{\beta}}^{\mathcal{S}_F} - \tilde{\boldsymbol{\beta}}^{\mathcal{S}}\right\|_2^2}{\sigma_\epsilon^2}, \frac{1}{2}\right\} \\ &\ge \min\left\{\frac{\hat{\rho}\min_{j\in\mathcal{S}_T}\left|\tilde{\beta}_j^{\mathcal{S}_F}\right|^2}{\sigma_\epsilon^2}, \frac{1}{2}\right\} \\ &\ge \min\left\{\frac{\rho\min_{j\in\mathcal{S}_F}|\beta_j|^2}{4\sigma_\epsilon^2}, \frac{1}{2}\right\},\end{aligned}$$

because $\log(1+x) \ge \min(x/2, 1/2)$ for all $x > 0$. Therefore, we obtain that

$$\lim_{n\to\infty} \mathbf{P}\left(\inf_{\mathcal{S}\supsetneq\mathcal{S}_T} \widetilde{\mathrm{GIC}}(\mathcal{S}) > \widetilde{\mathrm{GIC}}(\mathcal{S}_F)\right) \ge \lim_{n\to\infty} \mathbf{P}\left(\min\left\{\frac{\rho\min_{j\in\mathcal{S}_F}|\beta_j|^2}{4\sigma_\epsilon^2}, \frac{1}{2}\right\} > p\alpha\right) = 1 \tag{A.19}$$

by using (3.10), (3.11), (A.15), and (A.18).

Since $\log(1+x) \le x$ for all $x > -1$ and $Q(\tilde{\boldsymbol{\beta}}^{\mathcal{S}}; \mathbf{m}) \ge Q(\tilde{\boldsymbol{\beta}}^{\mathcal{S}_F}; \mathbf{m})$ for all $\mathcal{S} \subset \mathcal{S}_F$, we have

$$\begin{aligned}\log Q\left(\tilde{\boldsymbol{\beta}}^{\mathcal{S}_T}; \mathbf{m}\right) - \log Q\left(\tilde{\boldsymbol{\beta}}^{\mathcal{S}}; \mathbf{m}\right) &\le \frac{\left(Q\left(\tilde{\boldsymbol{\beta}}^{\mathcal{S}_T}; \mathbf{m}\right) - Q\left(\tilde{\boldsymbol{\beta}}^{\mathcal{S}}; \mathbf{m}\right)\right)}{Q\left(\tilde{\boldsymbol{\beta}}^{\mathcal{S}}; \mathbf{m}\right)} \\ &\le \frac{\left\|\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right\|_2 \left\|\mathbf{X}^T_{\mathcal{S}\cap\mathcal{S}_T^c}\left(I_{n-p} - \mathbf{X}_{\mathcal{S}_T}\left(\mathbf{X}^T_{\mathcal{S}_T}\mathbf{X}_{\mathcal{S}_T}\right)^{-1}\mathbf{X}^T_{\mathcal{S}_T}\right)(\mathbf{y} - m_0\mathbf{u})\right\|_2^2}{Q\left(\tilde{\boldsymbol{\beta}}^{\mathcal{S}_T}; \mathbf{m}\right)}\end{aligned}$$

for all set $\mathcal{S} \supsetneq \mathcal{S}_T$, where I_{n-p} is a $(n-p)$-dimensional identity matrix. Also,

$$\begin{aligned}&\sup_{\mathcal{S}\supsetneq\mathcal{S}_T} \left\|\mathbf{X}^T_{\mathcal{S}\cap\mathcal{S}_T^c}\left(I_{n-p} - \mathbf{X}_{\mathcal{S}_T}\left(\mathbf{X}^T_{\mathcal{S}_T}\mathbf{X}_{\mathcal{S}_T}\right)^{-1}\mathbf{X}^T_{\mathcal{S}_T}\right)(\mathbf{y} - m_0\mathbf{u})\right\|_2 \Big/ \sqrt{\left|\mathcal{S}\cap\mathcal{S}_T^c\right|} \\ &\le \sup_{\mathcal{S}\supsetneq\mathcal{S}_T} \left\|\mathbf{X}^T_{\mathcal{S}\cap\mathcal{S}_T^c}\left(I_{n-p} - \mathbf{X}_{\mathcal{S}_T}\left(\mathbf{X}^T_{\mathcal{S}_T}\mathbf{X}_{\mathcal{S}_T}\right)^{-1}\mathbf{X}^T_{\mathcal{S}_T}\right)\left(\mathbf{y} - m_0\mathbf{u} - \mathbf{X}_{\mathcal{S}_T}\boldsymbol{\beta}_{\mathcal{S}_T}\right)\right\|_\infty \\ &\le \left\|\mathbf{X}^T\boldsymbol{\varepsilon}\right\|_\infty + \left\|\mathbf{X}^T\mathbf{X}_{\mathcal{S}_T}\right\|_\infty \left\|\left(\mathbf{X}^T_{\mathcal{S}_T}\mathbf{X}_{\mathcal{S}_T}\right)^{-1}\right\|_\infty \left\|\mathbf{X}^T_{\mathcal{S}_T}\boldsymbol{\varepsilon}\right\|_\infty \\ &\quad + \left(\left\|\mathbf{X}^T\mathbf{u}\right\|_\infty + \left\|\mathbf{X}^T\mathbf{X}_{\mathcal{S}_T}\right\|_\infty \left\|\left(\mathbf{X}^T_{\mathcal{S}_T}\mathbf{X}_{\mathcal{S}_T}\right)^{-1}\right\|_\infty \left\|\mathbf{X}^T_{\mathcal{S}_T}\mathbf{u}\right\|_\infty\right) \max_{0\le j\le p}|m_j - \mu|\left(1 + \sum_{j\in\mathcal{S}_T}|\beta_j|\right) \\ &= O_P\left(\sqrt{n\log p}\right) + O_P\left(\kappa\sqrt{n\log q}\right) + O_P(q\kappa).\end{aligned}$$

Hence, (3.11), (A.15), and (A.18) imply that

$$\lim_{n\to\infty} \mathbf{P}\left(\inf_{\mathcal{S}\supsetneq\mathcal{S}_T} \frac{\widetilde{\mathrm{GIC}}(\mathcal{S}) - \widetilde{\mathrm{GIC}}(\mathcal{S}_T)}{\left|\mathcal{S}\cap\mathcal{S}_T^c\right|} > 0\right) = 1. \tag{A.20}$$

Therefore, we obtain the result (A.14) by (A.19) and (A.20). □

Proof of Theorem 3: By Theorems 1 and 2, there exists a sequence λ_0 such that

$$\lim_{n\to\infty} \mathbf{P}\left(\hat{\boldsymbol{\beta}}^{\lambda_0} = \hat{\boldsymbol{\beta}}^o, \mathcal{S}_{\lambda_0} = \mathcal{S}_T\right) = 1.$$

Since $\mathrm{GIC}(\lambda) \ge \widetilde{\mathrm{GIC}}(\mathcal{S}_\lambda)$ for all $\lambda > 0$ and $\mathrm{GIC}(\lambda_0) = \widetilde{\mathrm{GIC}}(\mathcal{S}_T)$ when $\hat{\boldsymbol{\beta}}^{\lambda_0} = \hat{\boldsymbol{\beta}}^o$ and $\mathcal{S}_{\lambda_0} = \mathcal{S}_T$,

$$\begin{aligned}\lim_{n\to\infty} \mathbf{P}\left(\inf_{\lambda\in\Lambda_+\cup\Lambda_-} \mathrm{GIC}(\lambda) > \mathrm{GIC}(\lambda_0)\right) &\ge \lim_{n\to\infty} \mathbf{P}\left(\inf_{\lambda\in\Lambda_+\cup\Lambda_-} \mathrm{GIC}(\lambda) > \mathrm{GIC}(\lambda_0), \hat{\boldsymbol{\beta}}^{\lambda_0} = \hat{\boldsymbol{\beta}}^o, \mathcal{S}_{\lambda_0} = \mathcal{S}_T\right) \\ &\ge \lim_{n\to\infty} \mathbf{P}\left(\inf_{\lambda\in\Lambda_+\cup\Lambda_-} \widetilde{\mathrm{GIC}}(\mathcal{S}_\lambda) > \widetilde{\mathrm{GIC}}(\mathcal{S}_T), \hat{\boldsymbol{\beta}}^{\lambda_0} = \hat{\boldsymbol{\beta}}^o, \mathcal{S}_{\lambda_0} = \mathcal{S}_T\right) \\ &\ge \lim_{n\to\infty} \mathbf{P}\left(\inf_{\mathcal{S}\ne\mathcal{S}_T} \widetilde{\mathrm{GIC}}(\mathcal{S}) > \widetilde{\mathrm{GIC}}(\mathcal{S}_T)\right) - \lim_{n\to\infty} \mathbf{P}\left(\hat{\boldsymbol{\beta}}^{\lambda_0} \ne \hat{\boldsymbol{\beta}}^o \text{ or } \mathcal{S}_{\lambda_0} \ne \mathcal{S}_T\right) \\ &= 1\end{aligned}$$

by Lemma 3. □

References

Akaike H (1969). Fitting autoregressive models for prediction, *Annals of the Institute of Statistical Mathematics*, **21**, 243–247.

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

Akaike H (1979). A Bayesian extension of the minimum AIC procedure of autoregressive model fitting, *Biometrika*, **66**, 237–242.

Bollerslev T (1986). Generalized autoregressive conditional heteroskedasticity, *Journal of Econometrics*, **31**, 307–327.

Brockwell PJ and Davis RA (2006). *Time Series: Theory and Methods* (2nd ed), Springer, New York.

Chen C (1999). Subset selection of autoregressive time series models, *Journal of Forecasting*, **18** 505–516.

Claeskens G, Croux C, and Van Kerckhoven J (2007). Prediction focused model selection for autoregressive models, *The Australian and New Zealand Journal of Statistics*, **49**, 359–379.

Claeskens G and Hjort NL (2003). The focussed information criterion, *Journal of the American Statistical Association*, **98**, 900–916.

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

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

Friedman J, Hastie T, Hofling H, and Tibshirani R (2007). Pathwise coordinate optimization, *The Annals of Applied Statistics*, **1**, 302–332.

Hannan EJ (1980). The estimation of the order of an ARMA process, *The Annals of Statistics*, **8**, 1071–1081.

Hannan EJ and Quinn BG (1979). The determination of the order of an autoregression, *Journal of Royal Statistical Society*, **41**, 190–195.

Huang J, Horowitz JL, and Ma S (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models, *The Annals of Statistics*, **36**, 587–613.

Kim Y, Choi H, and Oh H (2008). Smoothly clipped absolute deviation on high dimensions, *Journal of the American Statistical Association*, **103**, 1656–1673.

Kim Y, Jeon JJ, and Han S (2016). A necessary condition for the strong oracle property, *Scandinavian Journal of Statistics*, **43**, 610–624.

Kim Y and Kwon S (2012). Global optimality of nonconvex penalized estimators, *Biometrika*, **99**, 315–325.

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

Kwon S and Kim Y (2012). Large sample properties of the SCAD-penalized maximum likelihood estimation on high dimensions, *Statistica Sinica*, **22**, 629–653.

Kwon S, Lee S, and Na O (2017). Tuning parameter selection for the adaptive lasso in the autoregressive model, *Journal of the Korean Statistical Society*, **46**, 285–297.

Kwon S, Oh S, and Lee Y (2016). The use of random-effect models for high-dimensional variable selection problems, *Computational Statistics & Data Analysis*, **103**, 401–412.

Lee S, Kwon S, and Kim Y (2016). A modified local quadratic approximation algorithm for penalized optimization problems, *Computational Statistics & Data Analysis*, **94**, 275–286.

McClave J (1975). Subset autoregression, *Technometrics*, **17**, 213–220.

McLeod AI and Zhang Y (2006). Partial autocorrelation parametrization for subset autoregression, *Journal of Time Series Analysis*, **27**, 599–612.

Na O (2017). Generalized information criterion for the ar model, *Journal of the Korean Statistical Society*, **46**, 146–160.

Nardi Y and Rinaldo A (2011). Autoregressive process modeling via the lasso procedure, *Journal of Multivariate Analysis*, **102**, 528–549.

Sang H and Sun Y (2015). Simultaneous sparse model selection and coefficient estimation for heavy-tailed autoregressive processes, *Statistics*, **49**, 187–208.

Sarkar A and Kanjilal PP (1995). On a method of identification of best subset model from full ar-model, *Communications in Statistics-Theory and Methods*, **24**, 1551–1567.

Schmidt DF and Makalic E (2013). Estimation of stationary autoregressive models with the Bayesian LASSO, *Journal of Time Series Analysis*, **34**, 517–531.

Schwarz G (1978). Estimating the dimension of a model, *The Annals of Statistics*, **6**, 461–464.

Shen X, Pan W, Zhu Y, and Zhou H (2013). On constrained and regularized high-dimensional regression, *Annals of the Institute of Statistical Mathematics*, **1**, 1–26.

Shibata R (1976). Selection of the order of an autoregressive model by Akaike's information criterion, *Biometrika*, **63**, 117–126.

Tibshirani RJ (1996). Regression shrinkage and selection via the LASSO, *Journal of the Royal Statistical Society, Series B*, **58**, 267–288.

Tsay RS (1984). Order selection in nonstationary autoregressive models, *The Annals of Statistics*, **12**, 1425–1433.

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

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

Wu WB (2005). Nonlinear system theory: another look at dependence, *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 14150–14154.

Wu WB (2011). Asymptotic theory for stationary processes, *Statistics and Its Interface*, **4**, 207–226.

Ye F and Zhang CH (2010). Rate Minimaxity of the Lasso and Dantzig selector for the lq loss in lr balls, *Journal of Machine Learning Research*, **11**, 3519–3540.

Zhang CH (2010a). Nearly unbiased variable selection under minimax concave penalty, *The Annals of Statistics*, **38**, 894–942.

Zhang CH and Zhang T (2012). A general theory of concave regularization for high-dimensional sparse estimation problems, *Statistical Science*, **27**, 576–593.

Zhang T (2010b). Analysis of multi-stage convex relaxation for sparse regularization, *Journal of Machine Learning Research*, **11**, 1081–1107.

Zou H (2006). The adaptive lasso and its oracle properties, *Journal of the American Statistical Association*, **101**, 1418–1429.

Zou H and Li R (2008). One-step sparse estimates in nonconcave penalized likelihood models, *The Annals of Statistics*, **36**, 1509–1533.

Received January 22, 2018; Revised April 3, 2018; Accepted April 4, 2018