Screening and Selection Methods in High-Dimensional Linear Regression Model

Shinpei Imori[1]*, Shota, Katayama[2], and Hirofumi, Wakaki[1].

[1] *Graduate School of Science, Hiroshima University*

[2] *Graduate School of Engineering Science, Osaka University*

Abstract

In the present paper, we propose a new variable selection procedure for a high-dimensional linear model from two perspectives of the true and risk minimizing model selection. The proposed method consists of two factors: screening and selection. Both parts are based on the residual sum of squares, which can be easily understood. Our objective is to select a model consisting of indices of all nonzero regression coefficients, which is known as the true model when the true mean structure is included in the full model. Moreover, it minimizes the risk function under a restriction of explanatory variables. Even when the space of the target model is large, our selection method is consistent under mild conditions, i.e., the selection probability of the objective model goes to 1. Additionally, we reveal that consistency is retained when the true mean structure cannot be constructed from all available explanatory variables. Through simulation studies, we illustrate that our screening and selection methods are more effective than previous methods in various situations.

AMS 2010 *subject classifications*: Primary 62H12; Secondary 62J05.

Key words: High-dimensional analysis; linear regression model; screening and selection method; variable selection.

*Corresponding author, *E-mail address*: imori.stat@gmail.com

1 Introduction

By enhancing of technique for obtaining data and the speed of processing data, we can acquire many relevant factors to obtain a response variable of interest. In applied fields, we often encounter high-dimensional data in which the number of explanatory variable p is as bigger as the sample size n, although $n > p$ (Bühlmann & Geer, 2011). To determine the relationship between the response and the explanatory variables, selection of effective variables is important. The analysis of such high-dimensional data also involves the problem of selecting the best subset of explanatory variables; thus, the variable selection problem in a large parameter space is a topic of extensive study.

In model selection research, two classes of criteria are constructed in different perspectives that include predictive model selection and true model selection. We regard the risk minimizing model as the best model for predictive model selection, whereas the model constructed with all nonzero regression parameters is the best model for true model selection. The Akaike information criterion (AIC) (Akaike, 1973, 1974), Mallows C_p (Mallows, 1973), and cross-validation (CV) choice (Stone, 1974) are well known criteria that are suitable for prediction (Shibata, 1981, 1983; Li, 1987; Shao, 1997). On the other hand, the Bayesian information criterion (BIC) (Schwarz, 1978) is a famous criterion used for selecting the true model since it has a widely accepted consistency property (Nishii, 1984) such that the true model selection probability by BIC convergences to 1. For high-dimensional data analysis, other model selection criteria are useful for true model selection (Chen & Chen, 2008; Wang et al., 2009; Kim et al., 2012). However, when p is large, these methods are not computationally feasible since the number of considerable candidate models significantly increases.

A method based on penalized estimation such as least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996) is preferred for model selection in high-dimensional data because it has a consistency property and its computational costs are inexpensive. However, its consistency depends on strict assumptions such as irrepresentable conditions

(IRC) (Zhao & Yu, 2006) or sparse Riesz condition (SRC) (Zhang & Huang, 2008). These assumptions limit the structure of the explanatory variable matrix and the dimension of true model spaces. Moreover, the consistency of LASSO is sensitive to small variations in its tuning parameter. Many recent studies have proposed a method by combining solution paths of penalized estimation and model selection criteria (Zou et al., 2007; Wang et al., 2009; Wang & Zhu, 2011; Fan & Tang, 2013).

In the present study, we consider the model selection problem in high-dimensional data by dividing the model selection into two parts. We first attempt to screen extra candidate models, and then identify the best model by using the selection method. Both methods are based on the residual sum of squares, and focus on the gap between noncentral and central chi-square distributions. A crucial objective of this study is to derive the properties that can be selected for the true or risk minimizing model under mild conditions even when the model space is large. Furthermore, the method is not lengthy in comparison with selection among all models.

The remainder of the present paper is organized as follows: In Section 2, we introduce the notation for various quantities and propose the screening and variable selection method. In Section 3, we show the consistency properties of our proposed method. Simulation studies are reported in Section 4, and in Section 5, we conclude our paper. Technical details are provided in the Appendices.

2 Model Frameworks and Proposed Method

2.1 High-dimensional linear regression model

Let the data consist of a sequence $\{(y_i, \boldsymbol{x}_i); i = 1, \dots, n\}$, where $y_1, \dots, y_n$ are independent response variables distributed with normal distribution, $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ are p-dimensional nonstochastic vectors referred to as explanatory variables. $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)'$ denotes an $n \times p$

explanatory variable matrix. We can divide the true mean structure of $\boldsymbol{y} = (y_1, \ldots, y_n)'$ into

$$\begin{aligned} \mathrm{E}[\boldsymbol{y}] = \boldsymbol{\mu}_* &= \boldsymbol{P}_{\boldsymbol{X}}^{\perp}\boldsymbol{\mu}_* + \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{\mu}_* \\ &= \boldsymbol{P}_{\boldsymbol{X}}^{\perp}\boldsymbol{\mu}_* + \boldsymbol{X}\boldsymbol{\beta}, \end{aligned} \tag{1}$$

where $\boldsymbol{P}_{\boldsymbol{A}} = \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'$, $\boldsymbol{P}_{\boldsymbol{A}}^{\perp} = \boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{A}}$ for all full rank matrix $\boldsymbol{A}$, and $\boldsymbol{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\mu}_*$. Usually, $\boldsymbol{\beta}$ is regarded as a p-dimensional regression parameter. By supposing the first term of (1) as zero, i.e.,

(C0) $\boldsymbol{\mu}_* = \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{\mu}_* = \boldsymbol{X}\boldsymbol{\beta}$,

we can obtain the following linear regression model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \sigma_*\boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}_n, \boldsymbol{I}_n)$, $\boldsymbol{0}_n$ is an n-dimensional vector of zeros, and σ_*^2 is an unknown variance parameter. Suppose $\sigma_*^2 < \infty$. In this paper, we do not need to assume (C0).

Our main purpose is to select the best combination of explanatory variables. We define the full model as $j_F = \{1, \ldots, p\}$ whose element means the order of columns, and the set of considerable candidate models as $\mathcal{J}_n$. The goodness of fit for the candidate model j $(\subset j_F)$ is defined by the following risk function based on the predictive mean squared error:

$$Risk(j) = \mathrm{E}[||\boldsymbol{\mu}_* - \hat{\boldsymbol{\mu}}_j||^2] = \boldsymbol{\mu}_*'\boldsymbol{P}_{\boldsymbol{X}_j}^{\perp}\boldsymbol{\mu}_* + p_j\sigma_*^2,$$

where $\hat{\boldsymbol{\mu}}_j = \boldsymbol{P}_{\boldsymbol{X}_j}\boldsymbol{y}$, $\boldsymbol{X}_j$ is the submatrix of $\boldsymbol{X}$ corresponding to the elements of j, and p_j is the number of elements in model j; e.g., if $j = \{1, 2, 4\}$, $\boldsymbol{X}_j$ means the first, second, and fourth columns of $\boldsymbol{X}$, and $p_j = 3$. Let j_* be the risk minimization model, i.e., $j_* = \underset{j \in \mathcal{J}_n}{\mathrm{argmin}}\{Risk(j)\}$. From the perspective of predictive model selection, we can regard j_* as the best model. In contrast, from the perspective of clarifying the mean structure, $j_0 = \{\ell \in j_F | \beta_\ell \neq 0\}$ can be regarded as the best model, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$. It should be noted that many papers refer to j_0 as the true model by assuming (C0).

In the present paper, we define the high-dimensional framework as

$$\text{(C1)}\ \lim_{n\to\infty} \frac{p}{n} = c \in [0, 1),\ \lim_{n\to\infty} \frac{p_0}{n} = c_0 \in [0, c],$$

where $p_{j_0} = p_0$. Then, we allow p and p_0 to diverge infinity with n, although p and p_0 are less than n. The reason why we consider this situation is to improve the accuracy of approximations even when n, p, and p_0 are close values.

2.2 Proposed method

In this subsection, we give the screening and selection methods. For an element $\ell_k \in j_F$, we define a set $[-\ell_k] = j_F \setminus \{\ell_k\}$. Furthermore, the residual sum of squares for the model $j \in \mathcal{J}_n$ denotes $RSS(j) = \boldsymbol{y}'\boldsymbol{P}^{\perp}_{\boldsymbol{X}_j}\boldsymbol{y}$. Thereby, the nested candidate model set we propose is given as $\hat{\mathcal{J}}_n = \{\hat{j}_1, \ldots, \hat{j}_p\}$, where $\hat{j}_k = \{\ell_1, \ldots, \ell_k\}$ and $\ell_1, \ldots, \ell_p \in j_F$ satisfy

$$RSS([-\ell_1]) \geq \cdots \geq RSS([-\ell_p]).$$

Note that $\hat{j}_p = j_F$, and $RSS([-\ell_k])$ is regarded as an indicator in order to measure the significance of the kth element ℓ_k.

After screening, we consider a selection by using the nested model criterion (NC) defined as

$$\text{NC}(c_n) = \hat{j}_m,\quad m = 1 + \max_{1\leq k\leq p-1}\{kI(F_k > c_n)\},\quad F_k = \frac{RSS(\hat{j}_k) - RSS(\hat{j}_{k+1})}{RSS(j_F)/N}$$

where c_n is a positive monotonically increasing sequence, and $N = n - p$. The NC is derived for a nested candidate model set. If the condition (C0) holds, and $\hat{j}_k$ includes j_0, F_k is distributed as the F distribution with 1 and N degrees of freedom. Hence, a selection by using the above criterion is equivalent to the multiple F tests, and c_n means a threshold point of multiple F tests.

3 Properties of the Proposed Method

In this section, we derive three theorems on the screening and selection methods. First, we consider the consistency of the screening method. Let $\delta_j^2 = \boldsymbol{\mu}_*'(\boldsymbol{P}_{\boldsymbol{X}} - \boldsymbol{P}_{\boldsymbol{X}_j})\boldsymbol{\mu}_*$. In order to identify j_0 among $\mathcal{J}_n$, we assume the next condition:

(C2) $\lim_{n\to\infty} \delta_{\min}^2/\alpha_n = \infty$, where $\delta_{\min}^2 = \min_{j \not\supset j_0} \delta_j^2$, and α_n is a positive sequence.

A similar condition in which the denominator is $\log n$ instead of α_n is considered in Chen & Chen (2008). When this condition holds, we can show the following screening consistency.

Theorem 1. Under (C1) and (C2) with $\alpha_n \geq \log p$, $\hat{\mathcal{J}}_n$ includes j_0 with probability tending to 1, i.e.,

$$\lim_{n\to\infty} Pr(j_0 \in \hat{\mathcal{J}}_n) = 1.$$

A proof of the theorem 1 is given in Appendix 1. By applying the screening method to $\mathcal{J}_n$, we can reduce the number of candidate models from 2^p to p.

Furthermore, we can select j_0 after screening. We assume the following:

(C3) $\lim_{n\to\infty} \delta_F^2/n < \infty$, where $\delta_F^2 = \boldsymbol{\mu}_*'\boldsymbol{P}_{\boldsymbol{X}}^{\perp}\boldsymbol{\mu}_*$.

Note that $\delta_F^2 = 0$ if the condition (C0) holds, and $\delta_F^2/n = O(1)$ if all elements of $\boldsymbol{\mu}_*$ are bounded. Then, the assumption (C3) is established in these situations. If $\delta_F^2/n \to \infty$, $Risk(j_F) = \delta_F^2 + p\sigma_*^2 = \delta_F^2\{1 + o(1)\}$ and $Risk(j_*) = \boldsymbol{\mu}_*'\boldsymbol{P}_{\boldsymbol{X}_{j_*}}^{\perp}\boldsymbol{\mu}_* + p_{j_*}\sigma_*^2 \geq \delta_F^2 + p_{j_*}\sigma_*^2 = \delta_F^2\{1 + o(1)\}$ since $p_{j_*}/n \leq p/n \to c \in [0, 1)$ under (C1). Hence, the ratio of $Risk(j_F)$ to

$Risk(j_*)$ is

$$1 \leq \frac{Risk(j_F)}{Risk(j_*)} \leq \frac{\delta_F^2\{1+o(1)\}}{\delta_F^2\{1+o(1)\}} \to 1.$$

It turns out that the violation of (C3) induces j_F to be an efficient model without model selection.

We state the consistency of the NC method.

Theorem 2. Under (C1)-(C3) with $\alpha_n \geq \log p$, for all positive sequence γ_n satisfying $\alpha_n/\gamma_n = o(1)$ and $\gamma_n/\delta_{\min}^2 = o(1)$,

$$\lim_{n\to\infty} Pr\{\mathrm{NC}(\gamma_n) = j_0\} = 1.$$

A proof of theorem 2 is given in Appendix 2.

Both screening and selection procedures distinguish j_0 by using the gap between noncentral and central chi-square distributions. Such determination by the gap works for selection of the risk minimization model. We consider the following situation:

(C4) $\liminf\limits_{n\to\infty} \dfrac{\alpha_n \lambda_{\min}(\boldsymbol{X}'\boldsymbol{X})}{\lambda_{\max}(\boldsymbol{X}'\boldsymbol{X})} > 0$, where α_n is defined in the assumption (C2).

Theorem 3. Under (C1), (C2) and (C4) with $\alpha_n \geq 1$, for sufficient large n, $j_0 = j_*$.

A proof of the theorem 3 is given in Appendix 3. A similar result is given in Yanagihara et al. (2013) by assuming that $\lim_{n\to\infty} \boldsymbol{X}'\boldsymbol{X}/n$ exists and is positive definite. This theorem implies that the selection of j_0 is appropriate regardless of holding the condition (C0). By combining the theorems 2 and 3, we can select the risk minimizing model by using our method.

Table 1: Screening frequencies

				Frequency			
CASE	p_0	τ	$\delta^2_{\min}$	Proposed	LASSO	SCAD	MCP
A	20	0.3	69.331	989	513	934	982
B	20	0.6	60.038	772	0	141	317
C	40	0.3	62.782	994	35	774	908
D	40	0.6	56.496	819	0	172	290

Corollary 1. Under (C1)-(C4) with $\alpha_n \geq \log p$, for all positive sequence γ_n satisfying $\alpha_n/\gamma_n = o(1)$ and $\gamma_n/\delta^2_{\min} = o(1)$,

$$\lim_{n\to\infty} Pr\{\mathrm{NC}(\gamma_n) = j_*\} = 1.$$

4 Numerical Studies

We conduct two simulations by using "R" (R Core Team, 2013) to demonstrate the performance of the screening and selection methods in high-dimensional frameworks. The data is constructed in n explanatory variables $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \overset{i.i.d.}{\sim} N_p(\mathbf{0}_p, \boldsymbol{\Sigma}_\tau)$, $\boldsymbol{\Sigma}_\tau$ is given as

$$(\boldsymbol{\Sigma}_\tau)_{ij} = (\boldsymbol{\Sigma}_\tau)_{ji} = \tau^{|i-j|}, \quad 1 \leq i \leq j \leq p,$$

and $\tau = 0.3$ or 0.6. Let y_i be independently distributed as $N(\tilde{\boldsymbol{x}}_i'\boldsymbol{\beta}, 1)$ where $\tilde{\boldsymbol{x}}_i = (1, \boldsymbol{x}_i')'$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$, $\beta_\ell = \mathrm{sign}(U_\ell)I(\ell \leq p_0)$, and $U_0, U_1, \ldots, U_p \overset{\text{i.i.d.}}{\sim} U(-1, 1)$. In this simulation study, we assume all candidate models to include an intercept. This indicates that the true model is expressed as $j_0 = \{1, \ldots, p_0\}$, and the condition (C0) is established. We consider the situation $n = 100$, $p = 50$, and $p_0 = 20$ or 40, which corresponds to our high-dimensional framework with $c_0 > 0$. We present four cases as $(p_0, \tau) = (20, 0.3)$, $(20, 0.6)$, $(40, 0.3)$, and $(40, 0.6)$ referred to as case A, B, C, and D, respectively. We verified that the IRC is only established in case A.

Firstly, we compare the performance of screening procedure with the solution paths of LASSO and its improvements, SCAD in Fan & Li (2001), and MCP in Zhang (2010). The solution paths are provided by the "`R`" package, '`lars`' and '`ncvreg`' with a default setting. Table 1 lists $\delta^2_{\min}$ and frequencies of including the true model by each screening method in 1000 iterations. From table 1, we can see that the theorem 1 holds, and that the performance of the screening is better than the other methods in all cases. The MCP solution path often includes the true model in the case of $\tau = 0.3$; however, this value is far behind our screening method for the case of $\tau = 0.6$. We have made simulations in other settings and can confirm that our screening method includes the true model with high frequency. Next, we consider selection of the true model j_0 among the nested model set after screening. We iterate 1000 times, which satisfies $j_0 \in \hat{\mathcal{J}}_n$, in order to compare the performance of the NC with the GIC introduced in Nishii (1984). The GIC in model j is obtained as

$$\mathrm{GIC}(j) = n \log \{\boldsymbol{y}'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}_j})\boldsymbol{y}/n\} + p_j b_n,$$

where b_n is a hyper-parameter. In the GIC selection procedure, we selected the best model that minimizes the GIC. Both criteria require selection of a hyper-parameter, and the best value of each parameter depends on the situation, i.e., the magnitude of $\delta_{\min}$ and p_{j_*}. In figure 1, frequency transitions created by changing the penalty terms are shown. These frequencies are calculated by the points $b_n, c_n = k/25$, $k = 1, \ldots, 1000$. From figure 1, we can see that the peaks of the selection probability by the NC are close to that by the GIC. Therefore, we can choose a hyper-parameter of the NC in a similar manner as that for the GIC procedure. However, the NC selects the true model with high frequency for a wider range of penalty than the GIC. In particular, in cases A and C, the NC is almost completely superior to the GIC. These results likely occurred because the consistency of NC depends on only the value of $\delta^2_{\min}$, whereas the GIC selection must consider δ^2_j for all $j \not\supset j_0$ in $\hat{J}_n$ in order to maintain

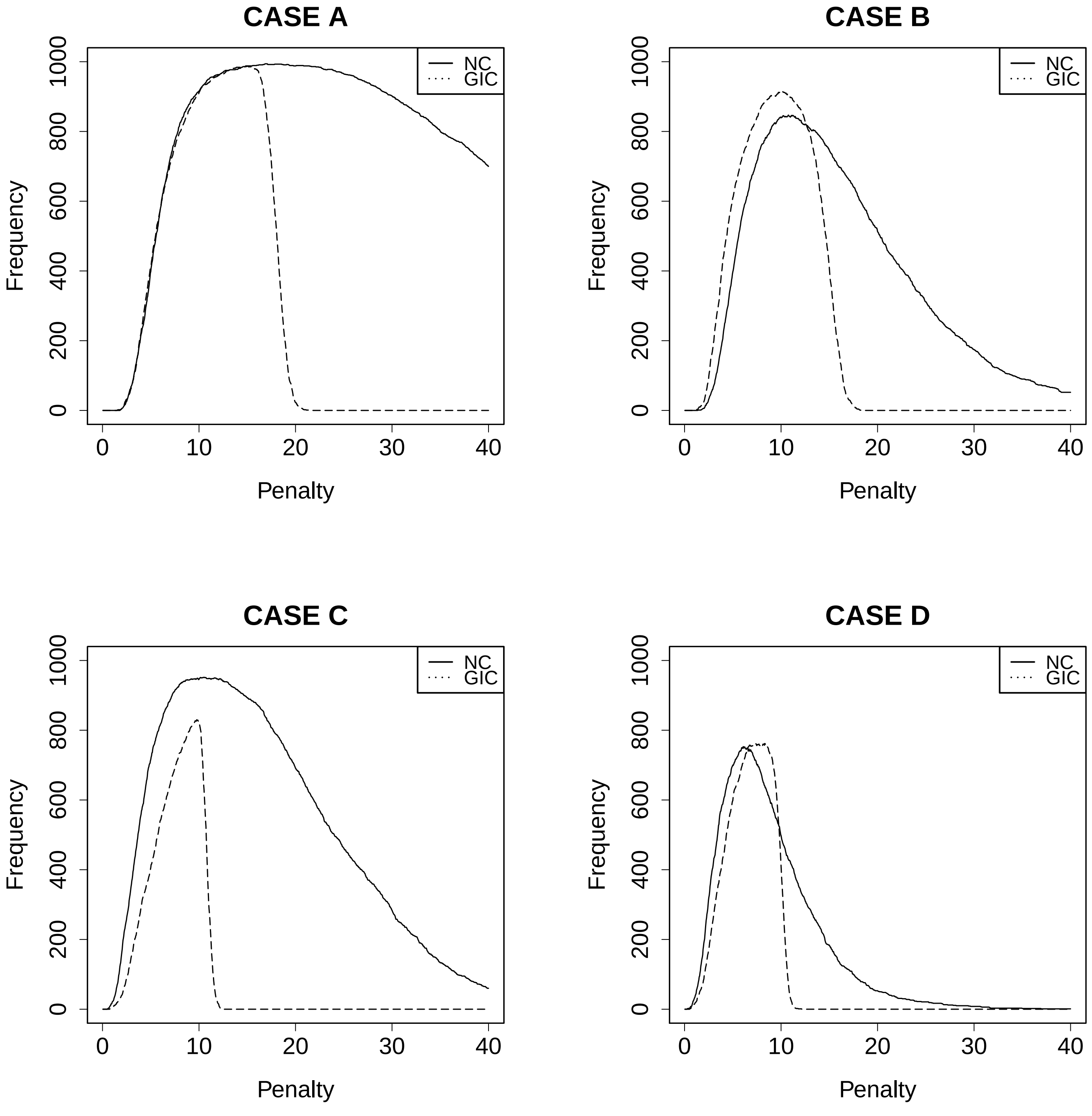

Figure 1: Transitions of frequency

consistency. The above results may imply that we can choose a hyper-parameter of the NC, robustly.

5 Conclusions

In this paper, we proposed a new variable selection procedure, which is constructed by screening and selection methods. Both methods are based on the residual sum of squares, which is essentially statistics, and these methods are fairly simple forms. The three theorems indicate that the screening and selection methods are suitable for the $n > p$ high-dimensional situation. Through simulation studies, we can confirm that our proposed method has higher frequencies in various conditions. Furthermore, we can select a hyper-parameter depending on the situation, such as that for GIC, and expect that the NC enables a more robust choice of the hyper-parameter with respect to selecting the best model. The screening method constructs a nested model set, and we can adjust the best model in stages based on the knowledge of experts, which is a good point for actual usage. Since the proof of the theorems is shown by the asymptotic theory of $N = n - p$, our proposed method works well if the difference between a sample size n and dimension of full model p is not small.

For theorems 1 and 2, the assumption (C2) is the most important, which assumes the existence of a gap between j_0 and the model $j \not\supset j_0$. Even when this assumption is not established, which may arise in real data analysis, we expect that j_* includes the higher priority order of explanatory variables decided by $RSS(j)$ under some assumptions.

Occasionally, we obtained $p > n$ high-dimensional or ultra-high dimensional ($p \gg n$) data in a situation such as gene data analysis. Recently, there have been many attempts to screen extra candidate models or variables in such high-dimensional cases (Fan & Lv, 2008; Wang, 2009; Wasserman & Roeder, 2009; Ing & Lai, 2011). By combining these methods, we will be able to apply our procedure to $p > n$ and $p \gg n$ high-dimensional data and select j_0.

Appendix

Firstly, we derive some inequalities to prove the theorems. Let z, χ^2_m and $\chi^2_m(\delta^2)$ be random variables distributed as $N(0,1)$, chi-square distribution with m degrees of freedom, and noncentral χ^2_m with the noncentral parameter δ^2, respectively. The following facts have been documented in previous papers (Shibata, 1981; Yang, 1999):

Fact 1: $Pr(|z| \geq t) \leq \exp(-t^2/2)$,

Fact 2: $Pr\{\chi^2_m \geq (1+t)m\} \leq \exp[-\{t-\log(1+t)\}m/2], \quad t \geq 0$,

Fact 3: $Pr\{\chi^2_m \leq (1-t)m\} \leq \begin{cases} \exp(-tm/4), & t \in [0,1), \\ 0 & t \in [1,\infty), \end{cases}$

Let $C = (1-\log 2)/2$. By applying the following inequality to fact 2:

$$t - \log(1+t) \geq 2Ct, \quad t \in [1,\infty),$$

then, we can obtain

$$Pr\{\chi^2_m \geq (1+t)m\} \leq \exp(-Ctm), \quad t \in [1,\infty). \tag{A.1}$$

By expanding $\chi^2_m(\delta^2)$ to $\chi^2_m + 2\delta z + \delta^2$, for all $s \geq 0$, it follows

$$\begin{aligned} Pr\{\chi^2_m(\delta^2) \leq t\} &= Pr(\{\chi^2_m + 2\delta z + \delta^2 \leq t\} \cap \{|z| > s\}) \\ &\quad + Pr(\{\chi^2_m + 2\delta z + \delta^2 \leq t\} \cap \{|z| \leq s\}) \\ &\leq Pr(|z| > s) + Pr(\chi^2_m \leq t - \delta^2 + 2\delta s). \end{aligned} \tag{A.2}$$

Furthermore, by using fact 1 and (A.1),

$$\begin{aligned}
&Pr\{\chi^2_m(\delta^2) \geq m(2+\delta/\sqrt{m}+\delta^2/m)\}\\
&= Pr(\{\chi^2_m + 2\delta z \geq m(2+\delta/\sqrt{m})\} \cap \{|z| > \sqrt{m}/2\})\\
&\quad + Pr(\{\chi^2_m + 2\delta z \geq m(2+\delta/\sqrt{m})\} \cap \{|z| \leq \sqrt{m}/2\})\\
&\leq Pr(|z| > \sqrt{m}/2) + Pr(\chi^2_m \geq 2m)\\
&= \exp(-m/8) + \exp(-Cm). \qquad\qquad (A.3)
\end{aligned}$$

A.1 Proof of theorem 1

It is clear that theorem 1 is established when $p_0 = p$. Therefore, we consider $p_0 < p$. Without loss of generality, we assume $\sigma^2_* = 1$. Let $RSS^*(j) = RSS(j) - RSS(j_F)$. Since

$$RSS^*([-\ell_1]) \geq \cdots \geq RSS^*([-\ell_p]),$$

by considering the definition of the screening method, we can show that

$$\begin{aligned}
Pr(j_0 \in \hat{\mathcal{J}}_n) &= Pr(\hat{j}_{p_0} = j_0)\\
&= Pr(\min_{\ell\in j_0} RSS^*([-\ell]) \geq \max_{\ell\notin j_0} RSS^*([-\ell])). \qquad (A.4)
\end{aligned}$$

$$\begin{aligned}
Pr(j_0 \in \hat{\mathcal{J}}_n) &= 1 - Pr\{\min_{\ell\in j_0} RSS^*([-\ell]) < \max_{\ell\notin j_0} RSS^*([-\ell])\}\\
&= 1 - Pr(\{\min_{\ell\in j_0} RSS^*([-\ell]) < \max_{\ell\notin j_0} RSS^*([-\ell])\} \cap \{\max_{\ell\notin j_0} RSS^*([-\ell]) > \delta^2_{\min}/2\})\\
&\quad - Pr(\{\min_{\ell\in j_0} RSS^*([-\ell]) < \max_{\ell\notin j_0} RSS^*([-\ell])\} \cap \{\max_{\ell\notin j_0} RSS^*([-\ell]) \leq \delta^2_{\min}/2\})\\
&\geq 1 - Pr\{\max_{\ell\notin j_0} RSS^*([-\ell]) > \delta^2_{\min}/2\} - Pr\{\min_{\ell\in j_0} RSS^*([-\ell]) < \delta^2_{\min}/2\}. \quad (A.5)
\end{aligned}$$

Note that

$$RSS^*([-\ell]) \sim \begin{cases} \chi_1^2, & \forall \ell \notin j_0, \\ \chi_1^2(\delta^2_{[-\ell]}), & \forall \ell \in j_0, \end{cases}$$

where $\delta^2_{[-\ell]} = \boldsymbol{\mu}'_*(\boldsymbol{P_X} - \boldsymbol{P_{X_{[-\ell]}}})\boldsymbol{\mu}_*$. From Caraux & Gascuel (1992),

$$\begin{aligned} Pr\{\max_{\ell \notin j_0} RSS^*([-\ell]) > \delta^2_{\min}/2\} &\leq (p - p_0) Pr(\chi_1^2 > \delta^2_{\min}/2), \\ Pr\{\min_{\ell \in j_0} RSS^*([-\ell]) < \delta^2_{\min}/2\} &\leq \sum_{\ell \in j_0} Pr\{\chi_1^2(\delta^2_{[-\ell]}) < \delta^2_{\min}/2\}. \end{aligned} \tag{A.6}$$

Since $\delta^2_{\min}/2 - 1 > 1$ for a sufficient large n, from (A.1),

$$Pr(\chi_1^2 > \delta^2_{\min}/2) = Pr(\chi_1^2 > 1 + \delta^2_{\min}/2 - 1) \leq \exp\{-C(\delta^2_{\min}/2 - 1)\}. \tag{A.7}$$

On the contrary, from (A.2) with $s = \delta_{[-\ell]}/4$, facts 1 and 3, and the result $\delta^2_{\min} \leq \delta^2_{[-\ell]}$, we can show that

$$\begin{aligned} Pr\{\chi_1^2(\delta^2_{[-\ell]}) < \delta^2_{\min}/2\} &\leq Pr(|z| > \delta_{[-\ell]}/4) + Pr(\chi_1^2 \leq \delta^2_{\min}/2 - \delta^2_{[-\ell]} + \delta_{\min}\delta_{[-\ell]}/2) \\ &\leq \exp(-\delta^2_{[-\ell]}/32). \end{aligned} \tag{A.8}$$

By substituting (A.6), (A.7), and (A.8) into (A.5), we get

$$\begin{aligned} Pr(j_0 \in \hat{\mathcal{J}}_n) &\geq 1 - (p - p_0)\exp\{-C(\delta^2_{\min}/2 - 1)\} - \sum_{\ell \in j_0} \exp(-\delta^2_{[-\ell]}/32) \\ &\geq 1 - (p - p_0)\exp\{-C(\delta^2_{\min}/2 - 1)\} - p_0 \exp(-\delta^2_{\min}/32) \to 1. \end{aligned}$$

The last convergence follows from the assumption (C2) with $\alpha_n \geq \log p$. □

A.2 Proof of theorem 2

Without loss of generality, we assume $\sigma_*^2 = 1$. The probability of $\text{NC}(\gamma_n) = j_0$ can be evaluated as

$$\begin{aligned} Pr\{\text{NC}(\gamma_n) = j_0\} &= 1 - Pr\{\text{NC}(\gamma_n) \neq j_0\} \\ &\geq 1 - Pr(\{\text{NC}(\gamma_n) \neq j_0\} \cap \{j_0 \in \hat{\mathcal{J}}_n\}) - Pr(j_0 \notin \hat{\mathcal{J}}_n). \end{aligned} \tag{A.9}$$

From the theorem 1 the last term of (A.9) is $o(1)$. Hence, we show $Pr(\{\text{NC}(\gamma_n) \neq j_0\} \cap \{j_0 \in \hat{\mathcal{J}}_n\}) = o(1)$. Hereafter, we assume $j_0 \in \hat{\mathcal{J}}_n$, i.e, $\hat{j}_k \supset j_0$ for all $k \geq p_0$. From the definition of NC, it follows that

$$\begin{aligned} Pr\{\text{NC}(\gamma_n) \neq j_0\} &\leq Pr(\cup_{k=p_0}^{p-1}\{F_k > \gamma_n\} \cup \{F_{p_0-1} \leq \gamma_n\}) \\ &\leq \sum_{k=p_0}^{p-1} Pr(F_k > \gamma_n) + Pr(F_{p_0-1} \leq \gamma_n). \end{aligned} \tag{A.10}$$

First, we attempt to show that $Pr(F_k > \gamma_n) = o((p - p_0)^{-1})$. Note that the numerator of F_k is distributed χ_1^2 for all $k \geq p_0$. Therefore,

$$\begin{aligned} Pr(F_k > \gamma_n) &= Pr\{\chi_1^2 > \gamma_n RSS(j_F)/N\} \\ &= Pr(\{\chi_1^2 > \gamma_n RSS(j_F)/N\} \cap \{RSS(j_F)/N \geq 1/2\}) \\ &\quad + Pr(\{\chi_1^2 > \gamma_n RSS(j_F)/N\} \cap \{RSS(j_F)/N < 1/2\}) \\ &\leq Pr(\chi_1^2 > \gamma_n/2) + Pr\{RSS(j_F)/N < 1/2\}. \end{aligned} \tag{A.11}$$

From the result that $\gamma_n/2 \geq 1 + \gamma_n/4$ for a sufficient large n and (A.1), we can show that

$$Pr(\chi_1^2 > \gamma_n/2) \leq Pr(\chi_1^2 > 1 + \gamma_n/4) \leq \exp(-C\gamma_n/4). \tag{A.12}$$

On the other hand, since $RSS(j_F)$ is distributed as $\chi^2_N(\delta^2_F)$, by applying (A.2) with $s = \max\{\sqrt{N}/4, \delta_F/2\}$ to the following probability, we obtain

$$\begin{aligned} Pr\{RSS(j_F)/N < 1/2\} &= Pr\{\chi^2_N(\delta^2_F) < N/2\} \\ &\leq Pr(|z| > s) + Pr(\chi^2_N \leq N/2 - \delta^2_F + 2\delta_F s) \\ &\leq Pr(|z| > \sqrt{N}/4) + Pr(\chi^2_N \leq 3N/4) \\ &\leq \exp(-N/32) + \exp(-N/16). \end{aligned} \tag{A.13}$$

The last inequality follows from facts 1 and 3. By substituting (A.12) and (A.13) into (A.11), we obtain

$$Pr(F_k > \gamma_n) \leq \exp(-C\gamma_n/4) + \exp(-N/32) + \exp(-N/16) = o((p-p_0)^{-1}). \tag{A.14}$$

Next, we evaluate the probability of $\{F_{p_0-1} \leq \gamma_n\}$. Denote $j_{0-} = \hat{j}_{p_0-1}$, $\delta^2_{0-} = \boldsymbol{\mu}'_*(\boldsymbol{P}_{\boldsymbol{X}_{j_0}} - \boldsymbol{P}_{\boldsymbol{X}_{j_{0-}}})\boldsymbol{\mu}_*$, and $C_1 = 2 + \delta_F/\sqrt{N} + \delta^2_F/N$. Therefore,

$$\begin{aligned} Pr(F_{p_0-1} \leq \gamma_n) &= Pr\{\chi^2_1(\delta^2_{0-}) \leq \gamma_n RSS(j_F)/N\} \\ &= Pr(\{\chi^2_1(\delta^2_{0-}) \leq \gamma_n RSS(j_F)/N\} \cap \{RSS(j_F)/N > C_1\}) \\ &\quad + Pr(\{\chi^2_1(\delta^2_{0-}) \leq \gamma_n RSS(j_F)/N\} \cap \{RSS(j_F)/N \leq C_1\}) \\ &\leq Pr\{RSS(j_F)/N > C_1\} + Pr\{\chi^2_1(\delta^2_{0-}) \leq C_1\gamma_n\} \\ &\leq Pr\{\chi^2_N(\delta^2_F) > C_1 N\} + Pr\{\chi^2_1(\delta^2_{0-}) \leq C_1\gamma_n\}. \end{aligned} \tag{A.15}$$

From (A.3),

$$Pr\{\chi^2_N(\delta^2_F) > C_1 N\} \leq \exp(-N/8) + \exp(-CN). \tag{A.16}$$

Note that $C_1\gamma_n < \delta^2_{0-}/2$ for a sufficient large n because C_1 is bounded from the assumption

(C3), and $\delta_{0-}^2 \geq \delta_{\min}^2$, $\gamma_n/\delta_{0-}^2 = o(1)$ holds from the assumption of theorem 2. Hence, from the above results, fact 1, and (A.2) with $s = \delta_{0-}/4$,

$$\begin{aligned} Pr\{\chi_1^2(\delta_{0-}^2) \leq C_1\gamma_n\} &\leq Pr(|z| > \delta_{0-}/4) + Pr(\chi_1^2 \leq C_1\gamma_n - \delta_{0-}^2/2) \\ &\leq \exp(-\delta_{0-}^2/2). \end{aligned} \tag{A.17}$$

Substituting (A.16) and (A.17) into (A.15), we show that

$$Pr(F_{p_0-1} \leq \gamma_n) \leq \exp(-N/8) + \exp(-CN) + \exp(-\delta_{0-}^2/2) = o(1). \tag{A.18}$$

By substituting (A.10), (A.14) and (A.18) into (A.9), we see that

$$Pr\{\mathrm{NC}(\gamma_n) = j_0\} \geq 1 + o(1) \rightarrow 1.$$

□

A.3 Proof of theorem 3

For all models $j \in \mathcal{J}_n$ $(p_j \geq p_0)$, since $\boldsymbol{\mu}_*'\boldsymbol{P}_{\boldsymbol{X}_j}^{\perp}\boldsymbol{\mu}_* \geq \boldsymbol{\mu}_*'\boldsymbol{P}_{\boldsymbol{X}_{j_0}}^{\perp}\boldsymbol{\mu}_* = \boldsymbol{\mu}_*'\boldsymbol{P}_{\boldsymbol{X}}^{\perp}\boldsymbol{\mu}_*$, we see that $Risk(j) \geq Risk(j_0)$. For all models $j \in \mathcal{J}_n$ $(p_j < p_0)$,

$$Risk(j) - Risk(j_0) = \boldsymbol{\mu}_*'(\boldsymbol{P}_{\boldsymbol{X}_{j_0}} - \boldsymbol{P}_{\boldsymbol{X}_j})\boldsymbol{\mu}_* - (p_0 - p_j)\sigma_*^2.$$

Note that $\boldsymbol{\mu}_*'(\boldsymbol{P}_{\boldsymbol{X}} - \boldsymbol{P}_{\boldsymbol{X}_{[-\ell]}})\boldsymbol{\mu}_* = \boldsymbol{\beta}'\boldsymbol{X}'(\boldsymbol{P}_{\boldsymbol{X}} - \boldsymbol{P}_{\boldsymbol{X}_{[-\ell]}})\boldsymbol{X}\boldsymbol{\beta} = \beta_\ell^2\boldsymbol{w}_\ell'(\boldsymbol{P}_{\boldsymbol{X}} - \boldsymbol{P}_{\boldsymbol{X}_{[-\ell]}})\boldsymbol{w}_\ell \leq \beta_\ell^2\boldsymbol{w}_\ell'\boldsymbol{w}_\ell$, and $\lambda_{\min}(\boldsymbol{X}'\boldsymbol{X}) \leq \boldsymbol{w}_\ell'\boldsymbol{w}_\ell \leq \lambda_{\max}(\boldsymbol{X}'\boldsymbol{X})$, where $\boldsymbol{w}_\ell$ is the ℓth column of $\boldsymbol{X}$, i.e., $\boldsymbol{X} =$

$(\boldsymbol{w}_1,\dots,\boldsymbol{w}_p)$. Then, we can evaluate $\boldsymbol{\mu}_*'(\boldsymbol{P}_{\boldsymbol{X}_{j_0}}-\boldsymbol{P}_{\boldsymbol{X}_j})\boldsymbol{\mu}_*$ as follows:

$$\begin{aligned}\boldsymbol{\mu}_*'(\boldsymbol{P}_{\boldsymbol{X}_{j_0}}-\boldsymbol{P}_{\boldsymbol{X}_j})\boldsymbol{\mu}_* &\geq \sum_{\ell\in j_0\cap j^c}\lambda_{\min}(\boldsymbol{X}'\boldsymbol{X})\beta_\ell^2\\ &\geq \sum_{\ell\in j_0\cap j^c}\frac{\lambda_{\min}(\boldsymbol{X}'\boldsymbol{X})}{\lambda_{\max}(\boldsymbol{X}'\boldsymbol{X})}\beta_\ell^2\boldsymbol{w}_\ell'\boldsymbol{w}_\ell\\ &\geq \frac{\alpha_n\lambda_{\min}(\boldsymbol{X}'\boldsymbol{X})}{\lambda_{\max}(\boldsymbol{X}'\boldsymbol{X})}\sum_{\ell\in j_0\cap j^c}\boldsymbol{\mu}_*'(\boldsymbol{P}_{\boldsymbol{X}}-\boldsymbol{P}_{\boldsymbol{X}_{[-\ell]}})\boldsymbol{\mu}_*/\alpha_n.\end{aligned}$$

The first inequality follows from the result of Chen & Chen (2008). Note that $\#(j_0\cap j^c)\geq p-p_j$. From the conditions (C2) and (C4), $\boldsymbol{\mu}_*'(\boldsymbol{P}_{\boldsymbol{X}}-\boldsymbol{P}_{\boldsymbol{X}_{[-\ell]}})\boldsymbol{\mu}_*/\alpha_n$ diverges to infinity and $\alpha_n\lambda_{\min}(\boldsymbol{X}'\boldsymbol{X})/\lambda_{\max}(\boldsymbol{X}'\boldsymbol{X})>0$, respectively. These results indicate that $Risk(j)-Risk(j_0)\geq 0$ for all $j\in\mathcal{J}_n$. □

Acknowledgement

This work is supported by Research Fellowships of Japan Society for the Promotion of Science for Young Scientists.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory* (eds. Petrov, B. N. & Csáki, F.), 267–281, Akadémiai Kiadó, Budapest.

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

Bühlmann, P. & Geer, S. (2011). *Statistics for high-dimensional data, methods, theory and applications.* Springer-verlag, Berlin.

Caraux, G & Gascuel, O. (1992). Bounds on distribution functions of order statistics for dependent variates. *Statist. Probab. Lett.*, **14**, 103–105.

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

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

Fan, J. & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Statist. Soc. B.*, **70**, 849–911.

Fan, Y. & Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *J. R. Statist. Soc. B.*, **75**, 531–552.

Ing, C.-K. & Lai, T. L. (2011). A stepwise regression method and consistent model selection for high-dimensional sparse linear models. *Statist. Sinica*, **21**, 1473–1513.

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

Li, K.-C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set. *Ann. Statist.*, **15**, 958–975.

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

Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.*, **12**, 758–765.

R Core Team (2013). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statist. Sinica*, **7**, 221–264.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika*, **68**, 45–54.

Shibata, R. (1983). Asymptotic mean efficiency of a selection of regression variables. *Ann. Inst. Statist. Math.*, **35**, 415–423.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. R. Statist. Soc. B.*, **36**, 111–147.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. R. Statist. Soc. B.*, **58**, 267–288.

Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *J. Amer. Statist. Assoc.*, **104**, 1512–1524.

Wang, H., Li, B. & Leng, C. (2009) Shrinkage tuning parameter selection with a diverging number of parameters. *J. R. Statist. Soc. B.*, **71**, 671–683.

Wang, T. & Zhu, L. (2011). Consistent tuning parameter selection in high dimensional sparse linear regression. *J. Multivariate Anal.*, **102**, 1141–1151.

Wasserman, L. & Roeder, K. (2009). High-dimensional variable selection. *Ann. Statist.*, **37**, 2178–2201.

Yang, Y. (1999). Model selection for nonparametric regression. *Statist. Sinica*, **9**, 475–499.

Yanagihara, H., Kamo, K., Imori, S., & Yamamura, M. (2013). A Study on the Bias-Correction Effect of the AIC for Selecting Variables in Normal Multivariate Linear Regression Models under Model Misspecification *TR 13-08, Statistical Research Group, Hiroshima University*, Hiroshima.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, **38**, 894–942.

Zhang, C.-H. & Huang, J. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Ann. Statist.*, **36**, 1567–1594.

Zhao, P. & Yu, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.*, **7**, 2541–2563.

Zou, H., Hastie, T. & Tibshirani, R. (2007). On the "degrees of freedom" of the Lasso. *Ann. Statist.*, **35**, 2173–2192.