Published in final edited form as:
Stat Anal Data Min. 2010 October 1; 3(5): 350–358. doi:10.1002/sam.10088.

Model selection procedure for high-dimensional data

Yongli Zhang[Assistant Professor] and
Lundquist College of Business, University of Oregon, 1208 University Ave, Eugene, OR 97403 (yongli@uoregon.edu).

Xiaotong Shen[Professor]
School of Statistics, University of Minnesota, 224 Church Street S.E., Minneapolis, MN 55455 (xshen@stat.umn.edu).

Summary

For high-dimensional regression, the number of predictors may greatly exceed the sample size but only a small fraction of them are related to the response. Therefore, variable selection is inevitable, where consistent model selection is the primary concern. However, conventional consistent model selection criteria like BIC may be inadequate due to their nonadaptivity to the model space and infeasibility of exhaustive search. To address these two issues, we establish a probability lower bound of selecting the smallest true model by an information criterion, based on which we propose a model selection criterion, what we call RIC$_c$, which adapts to the model space. Furthermore, we develop a computationally feasible method combining the computational power of least angle regression (LAR) with of RIC$_c$. Both theoretical and simulation studies show that this method identifies the smallest true model with probability converging to one if the smallest true model is selected by LAR. The proposed method is applied to real data from the power market and outperforms the backward variable selection in terms of price forecasting accuracy.

Keywords

Model selection; information criterion; large *p* but small *n*; RIC; power market

1 Introduction

With the advent of computers, data are exploding in size and complexity. How to extract information from mountains of data imposes many new challenges to statisticians. In the power market, for example, identifying relevant nodes is important from hundreds of candidates for the sake of price forecasting and risk hedging, where the sample size is smaller than the number of candidate nodes. In such a situation, is to identify the smallest true model M_0 from p_n predictors $x_1, ..., x_{p_n}$ based on a sample of size n. The large model space resulted from high dimension brings two challenges to most existing model selection criteria, namely, inadequate penalization coefficient and huge computing workload. The goal of this article is to develop a model selection criterion or procedure to deal with these two issues.

In the literature many of model selection criteria have been proposed, most of which are information criteria in the form of

$$\|\widehat{\mu}_{\mathbf{n}} - \boldsymbol{Y}_{\mathbf{n}}\|_2^2 + \lambda |M| \sigma^2 \tag{1}$$

where $\mathbf{Y_n}$ is a vector of n observations on the response Y , $\widehat{\mu}_{\mathbf{n}}$ is the least square estimate of $\mu_{\mathbf{n}} = E(\mathbf{Y_n})$, $|M|$ is the number of predictors in model M, $\| \ \|_2$ is the Euclidean norm and $\lambda > 0$

is the penalization coefficient. In (1), $\|\widehat{\mu}_{\mathbf{n}} - \boldsymbol{Y}_{\mathbf{n}}\|^2$ and $|M|\sigma^2$ measure the goodness-of-fit and a model's complexity, respectively. The penalization coefficient λ controls the balance between goodness-of-fit and a model's complexity.

The information criterion includes Akaike's information criterion (AIC, Akaike, 1973) with $\lambda = 2$, the Bayesian information criterion (BIC, Schwarz, 1978) with $\lambda = \log n$, ϕ criterion (Hannan and Quinn, 1979) with $\lambda = c \log\log(n)$ where $c > 2$, the risk inflation criterion (RIC, Foster and George, 1994) with $\lambda = 2 \log p_n$, the modified risk inflation criterion (MRIC, George and Foster, 2000) with $\lambda = \frac{2}{|M|}\left(\sum_{j=1}^{|M|} \log(p_n/j)\right)$, and covariance inflation criterion (CIC, Tibshirani and Knight, 1999) with $\lambda = \frac{4}{|M|}\left(\sum_{j=1}^{|M|} \log(p_n/j)\right)$. The asymptotic properties of information criteria have been studied extensively and the readers may refer to Shao (1997) and references therein for more details. Consistency and asymptotic loss efficiency are two major aspects assessing asymptotic properties of a model selection criterion. Consistent model selection that is the probability of selecting the smallest true model converging to 1, is our major objective in this article when the true model is one of the candidate models. A well-known fact is that BIC is consistent in the parametric case (Stone, 1979). However, there are two issues in applying BIC to high-dimensional data.

1. *The penalization coefficient of BIC,* log *n is nonadaptive to the model space (Chen and Chen, 2008);*
2. *Exhaustive search is infeasible*.

Some progress has been made with regard to the aforementioned two issues. Chen and Chen (2008) suggested a criterion based on BIC in which an p_n-dependent term is added to log n. For computation, the stepwise search method such as LASSO (Tibshirani, 1996) offers a feasible approach of performing model selection for high-dimensional data. The entire solution path of LASSO, generated by the Least Angle Regression (LAR) algorithm (Efron et al., 2004), is a sequence of models, which may include the smallest true model (Zhao and Yu, 2006). Zhao and Yu (2006) proved that if the irrepresentable condition is satisfied and a suitable penalization coefficient is selected LASSO will be consistent for model selection.

Concerning the difficulty of selecting a suitable penalization coefficient in LASSO, Zou et al. (2007) suggested a criterion based on BIC in which the least square estimate of β is replaced by the LASSO estimate.

In this article we propose a novel approach to overcome the two hindrances. The relationship between the size of model space, the penalization coefficient and the probability of selecting the smallest true model by (1) is derived. Next, a model selection procedure for high-dimensional data is constructed and studied from theoretical and empirical perspectives.

The rest of this article is organized as follows. Section 2 establishes a probability lower bound of selecting the smallest true model by (1), based on which a consistent information criterion is proposed as well as a procedure combining LAR with the proposed information criterion. Section 3 presents simulations and a real data example in which our approach compares favorably with other competing methods. Section 4 concludes this article and all technical proofs are deferred to Section 5.

2 Consistent model selection by RIC$_c$

Let $\mathbf{Y_n} = \mu_{\mathbf{n}} + \epsilon_{\mathbf{n}}$, where $\epsilon_{\mathbf{n}}$ is a vector of n *iid* errors, $\epsilon_i \sim N(0, \sigma^2)$; $i = 1, ..., n$. Linear regression is considered, so all subsets of predictors constitute the model space $\mathcal{M}^n$ and each element in $\mathcal{M}^n$ defines a model M. Let $\mathbf{X}_M$ be a submatrix composed of columns of $\mathbf{X}$ corresponding to predictors in M. The least square estimator of $\mu_\mathbf{n}$ for a given model is $\widehat{\mu}_\mathbf{n}(M)=P(M)\,\mathbf{Y_n}$, where $P(M)$ is the projection matrix and $P(M)=\mathbf{X}_M\left(\mathbf{X}'_M\mathbf{X}_M\right)^{-1}\mathbf{X}'_M$. Let $Q(M) = \mathbf{I_{n \times n}} - P(M)$ denote the corresponding orthogonal projection matrix, where $\mathbf{I_{n \times n}}$ is an n-dimension identity matrix and $Tr(P(M)) \leq |M|$.

The true model satisfies $\mu_\mathbf{n} = P(M)\mu_\mathbf{n}$. Let $\mathcal{M}_t^n$ denote the set of all true models, and $\mathcal{M}_f^n$ denote the set of all wrong models. The size of $\mathcal{M}_t^n$ in a typical multiple linear regression setting. The smallest true model is supposed to be unique, fixed, full-rank and in the model space. Assume that $p_n \to \infty$ as $n \to \infty$ and $|M_0| < n$.

The definition of consistency and asymptotic loss efficiency is as follows:

DEFINITION 1 *A model selection criterion is consistent if* $P(\hat{M} = M_0) \to 1$ *as* $n \to \infty$, *where* $\hat{M}$ *is a model selected by this model selection criterion over* $\mathcal{M}^n$.

DEFINITION 2 *A model selection criterion is asymptotically loss efficient if* $\frac{L_n(\widehat{M})}{\inf_{\mathcal{M} \in \mathcal{M}^n} L_n(M)} \to 1$ *in probability as* $n \to \infty$, *where* $\hat{M}$ *is a model selected by this model selection criterion over* $\mathcal{M}^n$ *and* $L_n(M)=\frac{\|\widehat{\mu}_\mathbf{n}(M)-\mu_\mathbf{n}\|_2^2}{n}$ *is the loss of model* M.

Note that consistency and asymptotic loss efficiency imply each other in the parametric case (Zhang, 2009).

2.1 Motivation

It has been observed that BIC is inconsistent when n is small compared to the size of $\mathcal{M}_t^n=2^{(p_n-|M_0|)}$ (Chen and Chen, 2008). To investigate the relationship between model selection consistency of (1) and λ, the probability lower bound of selecting the smallest true model by (1) is established in Theorem 1.

THEOREM 1 *Suppose that the smallest true model* M_0 *is unique, fixed, full-rank and in the model space. Let* $\hat{M}_\lambda$ *denote model selected by (1). Then*

$$P\left(\widehat{M}_\lambda=M_0|\widehat{M}_\lambda \in \mathcal{M}_t^n\right) \geq 2-\left(1+\exp\left(-\frac{(\lambda-1)}{2}\right)\lambda^{\frac{1}{2}}\right)^{p_n-|M_0|} \tag{2}$$

In Theorem 1, any relationship between p_n and n is not assumed, so p_n may be greater or less than n. Obviously, the lower bound $2-\left(1+\exp\left(-\frac{(\lambda-1)}{2}\right)\lambda^{\frac{1}{2}}\right)^{p_n-|M_0|}$ is a function of λ and $p_n - |M_0|$. The performance of an information criterion depends on p_n and λ, so we may construct consistent model selection criteria by choosing a suitable λ as a function of p_n such that $\left(1+\exp\left(-\frac{(\lambda-1)}{2}\right)\lambda^{\frac{1}{2}}\right)^{p_n-|M_0|} \to 1$ as $p_n \to \infty$. The following three corollaries are established.

COROLLARY 2.1 *Let* $\lambda = 2(1 + \gamma) \log p_n$ *in (1), where* γ *is a given positive real number. Then* $P\left(\widehat{M}_\lambda=M_0|\widehat{M}_\lambda \in \mathcal{M}_t^n\right) \to 1$ *as* $p_n \to \infty$.

COROLLARY 2.2 *Let* $\lambda = 2 \log p_n + 2 \log \log p_n$ *in (1). Then* $P\left(\widehat{M}_\lambda=M_0|\widehat{M}_\lambda \in \mathcal{M}_t^n\right) \to 1$ *as* $p_n \to \infty$.

COROLLARY 2.3 *Let* $\lambda = 2 \log p_n + (1 + \gamma) \log \log p_n$ *in (1), where* γ *is a given positive real number. Then* $P\left(\widehat{M}_\lambda = M_0 | \widehat{M}_\lambda \in \mathcal{M}_t^n\right) \to 1$ *as* $p_n \to \infty$.

The probability lower bound (2) reveals why a p_n-dependent penalization coefficient outperforms an n−dependent penalization coefficient with respect to consistent model selection, which will be examined in simulations in Section 3.1.

2.2 RIC$_c$

Based on Corollary 2.2, we propose the following criterion for model selection by minimizing

$$\|Q(M)\,\boldsymbol{Y}_\mathbf{n}\|_2^2 + 2\left[\log(p_n) + \log\log(p_n)\right]|M|\sigma^2. \tag{3}$$

This criterion is the information criterion (1) with $\lambda = 2 \log p_n + 2 \log \log p_n$, and can be viewed as RIC plus a correctional term $2 \log \log p_n |M| \sigma^2$ (RIC$_c$). The RIC$_c$ criterion is consistent as shown in Corollary 2.2. The influence of the correctional term, $2 \log \log p_n |M| \sigma^2$ will be examined by comparing RIC$_c$ with RIC through simulations in Section 3.1.

However, exhaustive search required by RIC$_c$ becomes infeasible as p_n is moderately large, say 50, so a stepwise method is considered. Next we will propose a procedure through combining LAR with RIC$_c$.

The LASSO estimate, $\widehat{\beta}^l(\lambda)$ is defined as the minimizer of

$$\|\boldsymbol{Y}_\mathbf{n} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_1 \tag{4}$$

where β is a p_n-dimension vector of regression coefficients β_j; $j = 1, ..., p_n$ and $\|\beta\|_1 = \sum_{j=1}^{p_n} |\beta_j|$ is the L_1 norm. Let k, a positive integer, denote the k-th step in the LAR algorithm and $\mathcal{A}_k$ denote the model selected at the k-th step. Let $\widehat{\beta}^l_{\mathcal{A}_k}$ denote the LASSO estimates corresponding to model $\mathcal{A}_k$, where $\widehat{\beta}^l_{\mathcal{A}_k}$ is obtained at step $k + 1$ (Efron et al., 2004).

A feasible model selection strategy is to select a model minimizing

$$\|\boldsymbol{Y}_\mathbf{n} - \mathbf{X}_\mathbf{n}\widehat{\beta}^l_{\mathcal{A}_k}\|_2^2 + (2\log(p_n) + 2\log\log(p_n))\,|\mathcal{A}_k|\sigma^2, \tag{5}$$

where possible candidate models are the sequence of models generated by LAR. However, the bias of the LASSO estimate needs be treated. If $\|\boldsymbol{Y}_\mathbf{n} - \mathbf{X}_\mathbf{n}\widehat{\beta}^l_{\mathcal{A}_k}\|_2^2$ is used as a measure of the goodness-of-fit and $|\mathcal{A}_k|$ as a measure of a model's complexity, it results in overpenalization on sparse models making them unlikely to be selected. Therefore, we need to construct a new complexity measure to reduce the effect of the bias of LASSO estimates.

Since the bias of LASSO estimate increases in λ, we suggest $\frac{2n}{\lambda^k}\|\widehat{\beta}^l_{\mathcal{A}_k}\|_1$ for a model's complexity, where λ^k is calculated through

$$\left(\mathbf{X}_{\mathcal{A}_k}\right)'\left(\boldsymbol{Y}_{\mathbf{n}}-\mathbf{X}_{\mathcal{A}_k}\widehat{\beta}^{l}_{\mathcal{A}_k}\right)=\frac{\lambda^k}{2}sign\left(\widehat{\beta}^{l}_{\mathcal{A}_k}\right), \tag{6}$$

which is derived by setting the derivative of (4) over β to be 0.

Replacing $|\mathcal{A}_k|$ by $\frac{2n}{\lambda^k}\|\widehat{\beta}^{l}_{\mathcal{A}_k}\|_1$ in (5), we obtain the LAR-RIC$_c$ criterion,

$$C(k)=\|\mathbf{X}\widehat{\beta}^{l}_{\mathcal{A}_k}-\boldsymbol{Y}_{\mathbf{n}}\|_2^2+\frac{4n\sigma^2(\log p_n+\log\log p_n)}{\lambda^k}\|\widehat{\beta}^{l}_{\mathcal{A}_k}\|_1. \tag{7}$$

Therefore, model selection by LAR-RIC$_c$ proceeds in two steps: (1) Generate a sequence of models by LAR, (2) Select the best model by minimizing (7). The LAR-RIC$_c$ criterion can achieve consistent model selection if LAR selects the smallest true model at its certain step; see Theorem 2.

Some additional assumptions are assumed about the design matrix **X** in addition to those introduced in Section 1. We assume that the design matrix has been normalized such that $\bar{\mathbf{X}}_j = 0$ and $\|\boldsymbol{X}_j\|^2 = n$ for each column vector of **X**, $\mathbf{X}_j$; $j = 1, ..., p_n$. Assume also that $\frac{\mathbf{x}'_i\mathbf{x}_j}{n}\rightarrow c_{ij}$ as $n \rightarrow \infty$ for two column vector $\mathbf{X}_i$ and $\mathbf{X}_j$, and obviously $c_{ij} \le 1$ and $c_{ij} = 1$ if $i = j$. Assume that $\frac{\mathbf{X}'_{\mathcal{A}^0}\mathbf{X}_{\mathcal{A}^0}}{n}\rightarrow C_{qq}$ as $n \rightarrow \infty$ and the smallest and largest eigenvalue of C_{qq} are $\phi_1(C_{qq})$ and $\phi_q(C_{qq})$, respectively with $0 < \phi_1(C_{qq}) < \phi_q(C_{qq}) < \infty$. Let u_k and u_{k+1} be unit equiangular vectors as defined in Efron et al. (2004), where $u'_k u_{k+1}$ is the inner product of u_k and u_{k+1} and less than 1.

In addition to assumptions about the design matrix, we assume that:

1. $\frac{\log(p_n)}{\sqrt{n}}\rightarrow 0$ as $n \rightarrow \infty$;
2. $\liminf_{n\rightarrow\infty}\frac{\|\mu_{\mathbf{n}}-\mathbf{X}_{\mathcal{A}_k}\widehat{\beta}^{l}_{\mathcal{A}_k}\|_2^2}{n}>0$ in probability for all finite-dimensional wrong models;
3. the smallest true model $\mathcal{A}_{k_0}$ is selected at step k_0 of LAR and $\mathcal{A}_{k_0}\subset\mathcal{A}_{k_0+1}$;
4. $\limsup_{n\rightarrow\infty}\left(u'_{k_0}u_{k_0+1}\right)\le\delta<1$ in probability as $n \rightarrow \infty$;
5. $\frac{2\|\widehat{\beta}^{l}_{\mathcal{A}_{k_0+1}}\|_1}{\lambda^{k_0+1}}-\frac{2\|\widehat{\beta}^{l}_{\mathcal{A}_{k_0}}\|_1}{\lambda^{k_0}}$ is bounded below by a positive number in probability as $n \rightarrow \infty$;
6. the irrepresentable condition (Zhao and Yu, 2006) is met.

The utility of these assumptions will be explained in the proof of consistency of LAR-RIC$_c$, which is summarized in the following theorem.

THEOREM 2 *If assumptions 1-6 hold, then LAR-RIC$_c$ is consistent.*

Next we perform simulations to examine the performance of RIC and LAR-RIC$_c$.

3 Data examples

3.1 Simulations: RIC

Simulations are performed to compare AIC, BIC, RIC, MRIC, CIC with RIC$_c$, with $n = 50$, 5000 and $p_n = 30$, based on 200 replications. The design matrix $\mathbf{X}$ is $n \times p_n$ and each row of $\mathbf{X}$ is generated by $MVN(\mathbf{0}, \Sigma_{p_n \times p_n})$ independently, where the ijth element of the covariance matrix $\Sigma_{p_n \times p_n}$ is 1 if $i = j$, and $\rho^{|i-j|}$ if $i \neq j$. Three values of ρ, -0.5, 0, 0.5, are examined.

The smallest true models is generated by $\mu_{\mathbf{n}} = \mathbf{X}\eta$, where η is a 30-dimension vector of regression coefficients. Only $|M_0|$ entries of η are assigned value 5 and the rest are 0. Let $\epsilon_{\mathbf{n}} = (\epsilon_1, .., \epsilon_{\mathbf{n}})'$ be generated from the standard normal distribution and $\mathbf{Y}_{\mathbf{n}} = \mu_{\mathbf{n}} + \epsilon_n$. The proportion of selecting M_0 is shown in the following tables.

As indicated in Tables 1-2, the performance of AIC and BIC worsens as $|\mathcal{M}_t^n|$ increases, which demonstrates their non-adaptivity to the model space. Though BIC selects the smallest true model with proportion close to 1 when $n = 5000$, its poor performance in the low sample size case ($n = 50$) makes it inadequate for high-dimensional data. In contrast to BIC, the performance RIC$_c$ is stable across different settings. Through a comparison of RIC with RIC$_c$, we note that the correctional term plays a key role. The performance of CIC and MRIC is the worst when $|M_0|$ is moderately large.

Overall, RIC$_c$ outperforms its competitors in terms of accuracy of identifying the smallest true model, especially when the sample size is small compared with the model space size. Therefore, RIC$_c$ is expected to be a competitor in model selection for high-dimensional data. The performance of RIC$_c$ testifies Corollary 2.2, too.

In the above simulations, the error variance, σ^2 is assumed known; see Shen and Ye (2002) and references therein for some estimating methods.

3.2 Simulations: LAR-RIC$_c$

The simulations are to test the performance of the LAR-RIC$_c$ criterion with $n = 50$, and p_n=25, 50, 75, 100, 250, 500, 750, 1000, 1500, 2000, 2500 based on 200 replications. The design matrix and the smallest true model are generated by the same method as in Section 3.1. The simulating output is shown in the following three tables.

The above simulations show that the smallest true model is not necessarily selected by LAR (Zhao and Yu, 2006), but LAR-RIC$_c$ is able to identify the smallest true model whenever LAR is able to do that. This confirms Theorem 2.

We note that the proportion of selecting over-fitting true models is greater than that of selecting the smallest true model and the median model size selected by LAR-RIC$_c$ is much less than p_n. Therefore, a reasonable strategy in empirical data analysis and modeling is to remove most redundant variables by LAR-RIC$_c$, then perform an exhaustive search through RIC$_c$. Osborne et al (1998) suggested a similar strategy. These simulations also show that the OLS estimate performs better than the LASSO estimate with respect to estimation accuracy due to the potential bias of the LASSO estimate. Thus we suggest a three-step procedure:

1. Remove redundant variables by LAR-RIC$_c$,
2. Select the best model by exhaustive subset selection using RIC$_c$,
3. Estimate parameters by the ordinary least square method based on the model selected in step (2).

We will use this procedure to analyze real data from the regional wholesale power market of the state of New York.

3.3 A real data example

Since 1998, the wholesale electricity market in USA has gone through a transition from vertically regulated markets to free and competitive markets. Some regional wholesale markets have taken into shape including New York (NY). Electricity is the most volatile commodity that brings much risk to the market participants. The high volatility of the electricity price makes price forecast challenging and crucial for market participants. The electricity price in these regional wholesale markets is determined by the the cost of supplying the next megawatt of electricity demand at a specific location, which is called a *node* in the power market. For example, there are more than four hundred pricing nodes in the regional power market of New York, and the nodal price differs with each other due to different local supply and demand conditions. If two nodes are geographically close to each other and connected by a transmission line that makes transmission and congestion loss low, they tend to have similar prices. We could predict the price of a node through other nodes whose prices are easier to predict. The data set represents 422 price observations on each of the 423 nodes, one of which works as the response. For our data $p_n = 422$ is greater than $n = 281$, but only a small fraction of nodes are expected useful in forecasting the price. Therefore it is highly desirable to construct a sparse model to achieve prediction accuracy.

The cross-validation method is used to compare two groups of model selection procedures. The first group of three procedures does exhaustive search by AIC, BIC and RIC$_c$ (respectively) preceded by LAR-RIC$_c$ (denoted by LAR.AIC, LAR.BIC and LAR.RIC$_c$). The second group of three procedures does backward stepwise variable selection (Cook and Weisberg, 1999) by AIC, BIC and RIC$_c$, respectively (denoted by STEP.AIC, STEP.BIC and STEP.RIC$_c$). The 422 observations are divided into a training set including the first 281 observations and a validation set including the remaining 141 observations. Let Y_i and $\widehat{Y}_i^0$; $i = 1, ..., 141$ denote the observed and predicted prices, respectively. The prediction error (PE) is defined as $|Y_i - \widehat{Y}_i^0|$; $i = 1, ..., 141$.

The cross-validation result based on Y_i versus $\widehat{Y}_i^0$; $i = 1, ..., 141$ of all six procedures is displayed in Figure 1, where the straight line passes (0,0) with slope 1.

As indicated in Table 10, the sparsest model yields the best prediction accuracy in both groups of procedures. The out-performance of RIC$_c$ over AIC and BIC can be attributed to the aggressiveness of RIC$_c$ in removing redundant variables. We note that though the model selected by STEP.RIC$_c$ is sparser than that selected by LAR.AIC or LAR.BIC, the former is worse than the latter with respect to prediction accuracy. A reasonable explanation is that STEP.AIC and STEP.BIC may miss some most relevant nodes.

Overall, Table 10 and Figure 1 show that the three-step procedure outperforms the backward stepwise method in terms of prediction accuracy. Another advantage of the three-step procedure is the computational efficiency.

4 Summary

This article contributes to model selection for high-dimensional data in two aspects. First, this article establishes a probability lower bound of selecting the smallest true model, which lays foundations for constructing model selection criteria as $p_n \to \infty$. Secondly, a model selection criterion RIC$_c$ and a procedure LAR-RIC$_c$ are proposed and their properties are studied. Simulating and real data examples show that LAR-RIC$_c$ is a potent tool in doing model selection for high-dimensional data. Admittedly, the consistency of LAR-RIC$_c$ is subject to the design matrix satisfying irrepresentable conditions, which is worth further study.

Acknowledgments

The research is supported in part by National Science Foundation Grant DMS-0906616 and National Institute of Health Grant 1R01GM081535. The authors thank the editor, the associate editor and three referees for their helpful comments and suggestions.

5 Technical proofs

5.1 Appendix A

Proof of Theorem 1: Suppose M is an over-fitting true model. Since $\mathbf{X}_M\beta_M = \mathbf{X}_{M_0}\beta_{M_0} = \mu_{\mathbf{n}}$, where all components of β_M and β_{M0} are nonzero, $\frac{\|P(M)\epsilon_{\mathbf{n}}\|_2^2-\|P(M_0)\epsilon_{\mathbf{n}}\|_2^2}{\sigma^2}$ follows a chisquare distribution with degree-of-freedom $Tr(P(M)) - |M_0| \le |M| - |M_0|$. To prove

$$P\left(\inf_{M\neq M_0;M\in\mathcal{M}_t^n}\left(\|Q(M)\boldsymbol{Y}_{\mathbf{n}}\|^2+\lambda|M|\sigma^2\right)<\|Q(M_0)\boldsymbol{Y}_{\mathbf{n}}\|^2+\lambda|M_0|\sigma^2\right)\le\left(1+\exp\left(-\frac{(\lambda-1)}{2}\right)\lambda^{\frac{1}{2}}\right)^{p_n-|M_0|}-1, \tag{8}$$

consider

$$P\left(\|P(M)\epsilon_{\mathbf{n}}\|^2-\|P(M_0)\epsilon_{\mathbf{n}}\|^2\right)\ge\ \lambda\left(|M|-|M_0|\right)\sigma^2\Big) \tag{9}$$

$$\le\quad P\left(\|P(M)\epsilon_{\mathbf{n}}\|^2-\|P(M_0)\epsilon_{\mathbf{n}}\|^2\right)\ge\ \lambda\left(Tr\left(P(M)\right)-|M_0|\right)\sigma^2\Big) \tag{10}$$

$$\le\quad \exp\left(-\delta/2\right)\left(1+\frac{\delta}{k}\right)^{k/2}=\left(\lambda\exp\left(1-\lambda\right)\right)^{k/2} \tag{11}$$

where $\delta = (\lambda - 1)k$ and $k = Tr(P(M)) - |M_0|$. Since $Tr(P(M)) \le |M|$,

$$P\left(\|P(M)\epsilon_{\mathbf{n}}\|^2-\|P(M_0)\epsilon_{\mathbf{n}}\|^2\right)\ge\ \lambda\left(|M|-|M_0|\right)\sigma^2\Big)\le\left(\lambda\exp\left(1-\lambda\right)\right)^{(|M|-|M_0|)/2} \tag{12}$$

Let $m = |M| - |M_0|$, then $|M| = |M_0| + m$. Then by Markov's inequality,

$$\begin{aligned}
& P\left(\inf_{M\neq M_0; M\in\mathcal{M}_t^n}\left(\|Q(M)\boldsymbol{Y}_{\mathbf{n}}\|^2+\lambda|M|\sigma^2\right)<\|Q(M_0)\boldsymbol{Y}_{\mathbf{n}}\|^2+\lambda|M_0|\sigma^2\right)\\
\leq & \sum_{M\neq M_0; M\in\mathcal{M}_t^n} P\left(\|P(M)\epsilon_{\mathbf{n}}\|^2-\|P(M_0)\epsilon_{\mathbf{n}}\|^2\right)\geq\ \lambda(|M|-|M_0|)\sigma^2\right)\\
\leq & \sum_{|M|=|M_0|+1}^{p_n}\begin{pmatrix} p_n-|M_0| \\ |M|-|M_0| \end{pmatrix}\lambda^{\frac{|M|-|M_0|}{2}}\exp\left(-\frac{(\lambda-1)(|M|-|M_0|)}{2}\right)\\
= & \sum_{m=1}^{p_n-|M_0|}\begin{pmatrix} p_n-|M_0| \\ m \end{pmatrix}\exp\left(-\frac{(\lambda-1)m}{2}\right)\lambda^{\frac{m}{2}}\\
= & \sum_{m=0}^{p_n-|M_0|}\begin{pmatrix} p_n-|M_0| \\ m \end{pmatrix}\exp\left(-\frac{(\lambda-1)m}{2}\right)\lambda^{\frac{m}{2}}-1=\left(1+\exp\left(-\frac{(\lambda-1)}{2}\right)\lambda^{\frac{1}{2}}\right)^{p_n-|M_0|}-1.
\end{aligned}$$

This completes the proof.

Proof of Corollary 1: The result follows from

$$\left(1+\exp\left(-\frac{(\lambda-1)}{2}\right)\lambda^{\frac{1}{2}}\right)^{p_n-|M_0|}=\left(1+\frac{\sqrt{2(1+\gamma)(\log p_n)e}}{p_n^{1+\gamma}}\right)^{p_n-|M_0|}\to 1 \tag{13}$$

as $p_n \to \infty$. This completes the proof.

Proof of Corollary 2: The result follows from

$$\left(1+\exp\left(-\frac{(\lambda-1)}{2}\right)\lambda^{\frac{1}{2}}\right)^{p_n-|M_0|}=\left(1+\frac{\sqrt{2\{\log(p_n)+\log(\log(p_n))\}e}}{p_n\log(p_n)}\right)^{p_n-|M_0|}\to 1 \tag{14}$$

as $p_n \to \infty$. This completes the proof.

Proof of Corollary 3: The result follows from

$$\left(1+\exp\left(-\frac{(\lambda-1)}{2}\right)\lambda^{\frac{1}{2}}\right)^{p_n-|M_0|}=\left(1+\frac{\sqrt{\{2\log p_n+(1+\gamma)\log\log p_n\}e}}{p_n\log^{\frac{1+\gamma}{2}}(p_n)}\right)^{p_n-|M_0|}\to 1 \tag{15}$$

as $p_n \to \infty$. This completes the proof.

In the last step of the proofs, we make use of the following fact: if a positive and increasing sequence $y_n \to \infty$ and a positive and decreasing sequence $x_n \to 0$, respectively as $n \to \infty$, then $\left(1+\frac{x_n}{y_n}\right)^{y_n} \to 1$ as $n \to \infty$. The proof is omitted.

5.2 Appendix B: Proof of Theorem 2

Set $\sigma^2 = 1$. We need the following seven lemmas to prove Theorem 2.

LEMMA 1 (Efron et al., 2004)

$$ctg_k\|P(\mathcal{A}_{k+1})\boldsymbol{Y}_{\mathbf{n}}-P(\mathcal{A}_k)\boldsymbol{Y}_{\mathbf{n}}\|_2=\|P(\mathcal{A}_k)\boldsymbol{Y}_{\mathbf{n}}-\mathbf{X}_{\mathcal{A}_k}\widehat{\beta}^l_{\mathcal{A}_k}\|_2 \tag{16}$$

where $P(\mathcal{A}_k) = \mathbf{X}_{\mathcal{A}_k}\left(\mathbf{X}'_{\mathcal{A}_k}\mathbf{X}_{\mathcal{A}_k}\right)^{-1}\mathbf{X}'_{\mathcal{A}_k}$ *and* $ctg_k = \frac{u'_k u_{k+1}}{\sqrt{1-\left(u'_k u_{k+1}\right)^2}}$.

LEMMA 2 *If* $\mathcal{A}_{k_0} \subset \mathcal{A}_k$, *then*
$\|Y - \mathbf{X}_{\mathcal{A}_{k_0}}\widehat{\beta}^l_{\mathcal{A}_{k_0}}\|_2^2 - \|Y - \mathbf{X}_{\mathcal{A}_k}\widehat{\beta}^l_{\mathcal{A}_k}\|_2^2 \le \|\left[P(\mathcal{A}_k) - P(\mathcal{A}_{k_0})\right]\epsilon_{\mathbf{n}}\|_2^2 + (ctg_{k_0})^2\|\left[P(\mathcal{A}_{k_0+1}) - P(\mathcal{A}_{k_0})\right]\epsilon_{\mathbf{n}}\|_2^2$
.

Proof: Note that

$$\begin{aligned}
& \|\boldsymbol{Y}_{\mathbf{n}} - \mathbf{X}_{\mathcal{A}_{k_0}}\widehat{\beta}^l_{\mathcal{A}_{k_0}}\|_2^2 - \|\boldsymbol{Y}_{\mathbf{n}} - \mathbf{X}_{\mathcal{A}_k}\widehat{\beta}^l_{\mathcal{A}_k}\|_2^2 \\
= \quad & \|P(\mathcal{A}_{k_0})\boldsymbol{Y}_{\mathbf{n}} + Q(\mathcal{A}_{k_0})\boldsymbol{Y}_{\mathbf{n}} - \mathbf{X}_{\mathcal{A}_{k_0}}\widehat{\beta}^l_{\mathcal{A}_{k_0}}\|_2^2 - \|P(\mathcal{A}_k)\boldsymbol{Y}_{\mathbf{n}} + Q(\mathcal{A}_k)\boldsymbol{Y}_{\mathbf{n}} - \mathbf{X}_{\mathcal{A}_k}\widehat{\beta}^l_{\mathcal{A}_k}\|_2^2 \\
= \quad & \|\left[P(\mathcal{A}_k) - P(\mathcal{A}_{k_0})\right]\epsilon_{\mathbf{n}}\|_2^2 + (ctg_{k_0})^2\|\left[P(\mathcal{A}_{k_0+1}) - P(\mathcal{A}_{k_0})\right]\epsilon_{\mathbf{n}}\|_2^2 - (ctg_k)^2\|\left[P(\mathcal{A}_{k+1}) - P(\mathcal{A}_k)\right]\epsilon_{\mathbf{n}}\|_2^2 \\
\le \quad & \|\left[P(\mathcal{A}_k) - P(\mathcal{A}_{k_0})\right]\epsilon_{\mathbf{n}}\|_2^2 + (ctg_{k_0})^2\|\left[P(\mathcal{A}_{k_0+1}) - P(\mathcal{A}_{k_0})\right]\epsilon_{\mathbf{n}}\|_2^2
\end{aligned}$$

This completes the proof.

LEMMA 3 $\widehat{\beta}^n_{\mathcal{A}_{k_0}} \to \beta^0_{\mathcal{A}_{k_0}}$ *in probability as* $n \to \infty$. *Both* $\widehat{\beta}^l_{\mathcal{A}_{k_0}}$ *and* $\beta^0_{\mathcal{A}_{k_0}}$ *are q-dimensional vectors*.

Proof: By Lemma 1, we have

$$\|P(\mathcal{A}_{k_0})\boldsymbol{Y}_{\mathbf{n}} - \mathbf{X}_{\mathcal{A}_{k_0}}\widehat{\beta}^l_{\mathcal{A}_{k_0}}\|_2^2 = ctg^2_{k_0}\|P(\mathcal{A}_{k_0+1})\boldsymbol{Y}_{\mathbf{n}} - P(\mathcal{A}_{k_0})\boldsymbol{Y}_{\mathbf{n}}\|_2^2 \quad (17)$$

By Assumptions 3 and 4,

$$\frac{\|P(\mathcal{A}_{k_0})\boldsymbol{Y}_{\mathbf{n}} - \mathbf{X}_{\mathcal{A}_{k_0}}\widehat{\beta}^l_{\mathcal{A}_{k_0}}\|_2^2}{n} \to 0 \quad (18)$$

in probability, which implies

$$\frac{\|\mathbf{X}_{\mathcal{A}_{k_0}}\left(\widehat{\beta}^l_{\mathcal{A}_{k_0}} - \beta^0_{\mathcal{A}_{k_0}}\right)\|_2^2 + 2\left(\epsilon'_n\mathbf{X}_{\mathcal{A}_{k_0}}\left(\widehat{\beta}^l_{\mathcal{A}_{k_0}} - \beta^0_{\mathcal{A}_{k_0}}\right)\right) + \|P(\mathcal{A}_{k_0})\epsilon_{\mathbf{n}}\|_2^2}{n} \to 0 \quad (19)$$

in probability as $n \to \infty$. If $\frac{\|\mathbf{X}_{\mathcal{A}_{k_0}}\left(\widehat{\beta}^l_{\mathcal{A}_{k_0}} - \beta^0_{\mathcal{A}_{k_0}}\right)\|_2^2}{n}$ did not converge to 0 in probability, then $\epsilon'_n\mathbf{X}_{\mathcal{A}_{k_0}}\left(\widehat{\beta}^l_{\mathcal{A}_{k_0}} - \beta^0_{\mathcal{A}_{k_0}}\right)$ would be negligible compared with $\|\mathbf{X}_{\mathcal{A}_{k_0}}\left(\widehat{\beta}^l_{\mathcal{A}_{k_0}} - \beta^0_{\mathcal{A}_{k_0}}\right)\|_2^2$ and (18) would not hold. Contradiction. This implies the desired result.

LEMMA 4 $\lambda_{\mathcal{A}_{k_0}} \sim O_p\sqrt{n\left(1 + ctg^2_{k_0}\right)}$.

Proof: It follows from Efron et al. (2004) that

$$\|P\left(\mathcal{A}_{k_0+1}\right)\boldsymbol{Y}_\mathbf{n}-\mathbf{X}_{\mathcal{A}_{k_0}}\widehat{\beta}^l_{\mathcal{A}_{k_0}}\|_2^2=\left(\frac{\lambda_{\mathcal{A}_{k_0}}}{2\alpha}\right)^2 \tag{20}$$

where $\alpha=\left(s'_{\mathcal{A}_{k_0}}\left(\mathbf{X}_{\mathcal{A}_{k_0}}\mathbf{X}_{\mathcal{A}_{k_0}}\right)^{-1}s_{\mathcal{A}_{k_0}}\right)^{-1/2}$ and $s'_{\mathcal{A}_{k_0}}$ is a $|\mathcal{A}_{k_0}|$ dimension vector of 1 and -1.

$$\alpha/\sqrt{n}=\left(s'_{\mathcal{A}_{k_0}}\left(\mathbf{X}_{\mathcal{A}_{k_0}}\mathbf{X}_{\mathcal{A}_{k_0}}/n\right)^{-1}s_{\mathcal{A}_{k_0}}\right)^{-0.5} \tag{21}$$

We have known that $\mathbf{X}_{\mathcal{A}_{k_0}}\mathbf{X}_{\mathcal{A}_{k_0}}/n\to C_{qq}$ and $\phi_1(C_{qq})$ and $\phi_q(C_{qq})$ are the smallest and largest eigenvalues of C_{qq}, respectively. Hence,

$$\frac{q}{\phi_q\left(C_{qq}\right)}\le\liminf_{n\to\infty} ns'_{\mathcal{A}_{k_0}}\left(\mathbf{X}_{\mathcal{A}_{k_0}}\mathbf{X}_{\mathcal{A}_{k_0}}\right)^{-1}s_{\mathcal{A}_{k_0}}\le\limsup_{n\to\infty} ns'_{\mathcal{A}_{k_0}}\left(\mathbf{X}_{\mathcal{A}_{k_0}}\mathbf{X}_{\mathcal{A}_{k_0}}\right)^{-1}s_{\mathcal{A}_{k_0}}\le\frac{q}{\phi_1\left(C_{qq}\right)} \tag{22}$$

So

$$\sqrt{\frac{\phi_1\left(C_{qq}\right)}{q}}\le\liminf_{n\to\infty}\alpha/\sqrt{n}\le\limsup_{n\to\infty}\alpha/\sqrt{n}\le\sqrt{\frac{\phi_q\left(C_{qq}\right)}{q}}, \tag{23}$$

implying

$$\left(\frac{\lambda_{\mathcal{A}_{k_0}}}{2\alpha}\right)^2=\left(1+ctg^2_{k_0}\right)\|\left(P\left(\mathcal{A}_{k_0+1}\right)-P\left(\mathcal{A}_{k_0}\right)\right)\epsilon_\mathbf{n}\|_2^2 \tag{24}$$

This completes the proof.

LEMMA 5 *For $k \ge k_0$, $\frac{\lambda^k}{n}\to 0$ in probability.*

Proof: By Lemma 4 and Assumption 4, we know that $\frac{\lambda^{k_0}}{n}\to 0$ in probability. And since $\lambda^k \le \lambda^{k_0}$ (Efron et al., 2004), $\frac{\lambda^k}{n}\to 0$ in probability. This completes the proof.

LEMMA 6 $\|\widehat{\beta}^l_{\mathcal{A}_k}\|_1$ *is an increasing function of k.*

Proof: Suppose $k_2 > k_1$, then $\lambda^{k_2} \le \lambda^{k_1}$ (Efron et al., 2004)

$$\|\boldsymbol{Y}_\mathbf{n}-\mathbf{X}_{\mathcal{A}_{k_2}}\widehat{\beta}^l_{\mathcal{A}_{k_2}}\|_2^2+\lambda^{k_2}\|\widehat{\beta}^l_{\mathcal{A}_{k_2}}\|_1\le\|\boldsymbol{Y}_\mathbf{n}-\mathbf{X}_{\mathcal{A}_{k_1}}\widehat{\beta}^l_{\mathcal{A}_{k_1}}\|_2^2+\lambda^{k_2}\|\widehat{\beta}^l_{\mathcal{A}_{k_1}}\|_1 \tag{25}$$

$$\|\boldsymbol{Y}_\mathbf{n}-\mathbf{X}_{\mathcal{A}_{k_1}}\widehat{\beta}^l_{\mathcal{A}_{k_1}}\|_2^2+\lambda^{k_1}\|\widehat{\beta}^l_{\mathcal{A}_{k_1}}\|_1\le\|\boldsymbol{Y}_\mathbf{n}-\mathbf{X}_{\mathcal{A}_{k_2}}\widehat{\beta}^l_{\mathcal{A}_{k_2}}\|_2^2+\lambda^{k_1}\|\widehat{\beta}^l_{\mathcal{A}_{k_2}}\|_1 \tag{26}$$

Therefore, $\left(\lambda^{k_2} - \lambda^{k_1}\right)\left(\|\widehat{\beta}^{l}_{\mathcal{A}_{k_2}}\|_1 - \|\widehat{\beta}^{l}_{\mathcal{A}_{k_1}}\|_1\right) \le 0$. This completes the proof.

LEMMA 7 *If $\mathcal{A}_{k_0} \subset \mathcal{A}_k$, then $\frac{|\mathcal{A}_k|-q}{2n\|\widehat{\beta}^{l}_{\mathcal{A}_k}\|_1/\lambda^{k}-2n\|\widehat{\beta}^{l}_{\mathcal{A}_{k_0}}\|_1/\lambda^{k_0}}$ is bounded above by a positive real number in probability as $n \to \infty$.*

Proof: By assumption 5

$$\frac{|\mathcal{A}_k| - q}{2n\|\widehat{\beta}^{l}_{\mathcal{A}_k}\|_1/\lambda^{k} - 2n\|\widehat{\beta}^{l}_{\mathcal{A}_{k_0}}\|_1/\lambda^{k_0}} \le \frac{1}{2\|\widehat{\beta}^{l}_{\mathcal{A}_{k_0+1}}\|_1/\lambda^{k_0+1} - 2\|\widehat{\beta}^{l}_{\mathcal{A}_{k_0}}\|_1/\lambda^{k_0}} \tag{27}$$

This completes the proof.

Proof of Theorem 2:

We consider two cases: $\mathcal{A}_k \not\supseteq \mathcal{A}_{k_0}$ and $\mathcal{A}_k \supset \mathcal{A}_{k_0}$.

1. We have known for a wrong model $\mathcal{A}_k$, $\liminf_{n\to\infty}\frac{\|Y-\mathbf{X}_{\mathcal{A}_k}\|_2^2}{n}>0$ in probability by Assumption 2, whereas for the true model $\mathcal{A}_{k_0}$, $C(k_0)/n \to 0$ in probability as $n \to \infty$ by Assumptions 1 and 4.

2. For the second case ($\mathcal{A}_{k_0} \subset \mathcal{A}_k$), we want to bound the probability $P(C(k_0) \ge \inf_k C(k))$.

$$\begin{aligned}
& P\left(\|Y-\mathbf{X}_{\mathcal{A}_{k_0}}\widehat{\beta}^{l}_{\mathcal{A}_{k_0}}\|_2^2+2\{\log(p_n)+\log(\log(p_n))\}\frac{2n\|\widehat{\beta}^{l}_{\mathcal{A}_{k_0}}\|_1}{\lambda_{k_0}} \ge \|Y-\mathbf{X}_{\mathcal{A}_k}\widehat{\beta}^{l}_{\mathcal{A}_k}\|_2^2+2\{\log(p_n)+\log(\log(p_n))\}\frac{2n\|\widehat{\beta}^{l}_{\mathcal{A}_k}\|_1}{\lambda^{k}}\right) \\
\le\ & P\left(\|[P(\mathcal{A}_k)-P(\mathcal{A}_{k_0})]\,\epsilon_{\mathbf{n}}\|_2^2+(ctg_{k_0})^2\|[P(\mathcal{A}_{k_0+1})-P(\mathcal{A}_{k_0})]\,\epsilon_{\mathbf{n}}\|_2^2 \ge 2\{\log(p_n)+\log(\log(p_n))\}\left\{\frac{2n\|\widehat{\beta}^{l}_{\mathcal{A}_k}\|_1}{\lambda^{k}}-\frac{2n\|\widehat{\beta}^{l}_{\mathcal{A}_{k_0}}\|_1}{\lambda^{k_0}}\right\}\right) \\
\le\ & P\left(\left(1+(ctg_{k_0})^2\right)\|[P(\mathcal{A}_k)-P(\mathcal{A}_{k_0})]\,\epsilon_{\mathbf{n}}\|_2^2 \ge 2\{\log(p_n)+\log(\log(p_n))\}\left\{\frac{2n\|\widehat{\beta}^{l}_{\mathcal{A}_k}\|_1}{\lambda^{k}}-\frac{2n\|\widehat{\beta}^{l}_{\mathcal{A}_{k_0}}\|_1}{\lambda^{k_0}}\right\}\right) \\
\le\ & P\left(\frac{\left(1+\left(ctg_{k_0}\right)^2\right)(|\mathcal{A}_k|-q)}{\left\{\frac{2n\|\widehat{\beta}^{l}_{\mathcal{A}_k}\|_1}{\lambda^{k}}-\frac{2n\|\widehat{\beta}^{l}_{\mathcal{A}_{k_0}}\|_1}{\lambda^{k}{}_{0}}\right\}}\|[P(\mathcal{A}_k)-P(\mathcal{A}_{k_0})]\,\epsilon_{\mathbf{n}}\|_2^2 \ge 2\{\log(p_n)+\log(\log(p_n))\}(|\mathcal{A}_k|-q)\right)
\end{aligned}$$

We can prove $\sum_{\mathcal{A}_{k_0}\subset\mathcal{A}_k} P(C(k_0) \ge \inf_k C(k)) \to 0$ using the method in the proof of Theorem 1. This completes the proof.

References

1. Akaike, H. Information theory and an extension of the maximum likelihood principle. In: Petrov, B.; Csaki, F., editors. Second International Symposium on Information Theory. Akademiai Kiado; Budapest: 1973. p. 267-281.
2. Chen J, Chen Z. Extended Bayesian information criteria for model selection with large model spaces. Biometrika. 2008; 95:759–771.
3. Cook, RD.; Weisberg, S. Applied Regression Including Computing and Graphics. Wiley; 1999.
4. Efron B, Hastie T, Johnstone I, Tibshirani R. Least Angle Regression. Annals of Statistics. 2004; 32:407–499.

5. Foster DP, George EI. The risk inflation criterion for multiple regression. The Annals of Statistics. 1994; 22:1947–1975.

6. George EI, Foster DP. Calibration and Empirical Bayes Variable Selection. Biometrika. 2000; 87:731–747.

7. Hannan EJ, Quinn BG. The determination of the order of an autoregression. Journal of the Royal Statistical Society, Series B. 1979; 41:190–195.

8. Osborne MR, Presnell B, Turlach BA. Knot selection for regression splines via the LASSO. Computing Science and Statistics. 1998; 30:44–49.

9. Schwarz G. Estimating the dimension of a model. The Annals of Statistics. 1978; 6:2, 461–464.

10. Shao J. An asymptotic theory for linear model selection (with discussion). Statistica Sinica. 1997; 7:221–264.

11. Shen X, Ye J. Adaptive model selection. Journal of the American Statistical Association. 2002; 97:210–221.

12. Shibata R. An optimal selection of regression variables. Biometrika. 1981; 68:1, 45–54.

13. Stone M. Comments on model selection criteria of Akaike and Schwarz. Journal of the Royal Statistical Society, Series B. 1979; 41:276–278.

14. Tibshirani R. Regression shrinkage and selection via the LASSO. Journal of the Royal Statistical Society, B. 1996; 58:267–288.

15. Tibshirani R, Knight K. The covariance inflation criterion for adaptive model selection. Journal of the Royal Statistical Society, B. 1999; 61:529–546.

16. Zhang Y. Model selection: A Lagrange optimization approach. Journal of Statistical Planning and Inference. 2009; 139:3142–3159.

17. Zhao P, Yu B. On model selection consistency of Lasso. Journal of Machine Learning Research. 2006; 7:2541–2563.

18. Zou H, Hastie T, Tibshirani R. On the “degrees of freedom” of the LASSO. The Annals of Statistics. 2007; 35:2173–2192.

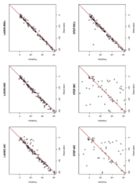

Figure 1.

Predicted hourly price by six procedures vs the corresponding observed hourly price.

Table 1

The first three columns list the value of ρ, $|M_0|$ and $|\mathcal{M}_t^n|$, respectively. The 4th-9th columns show the proportion of selecting the smallest true model by AIC, BIC, RIC, MRIC, CIC and RIC$_c$ based on 200 simulation replications with n = 50, p_n = 30 and σ^2 = 1.

ρ	$\lvert M_0 \rvert$	$\lvert M_t^n \rvert$	AIC	BIC	RIC	MRIC	CIC	RIC$_c$
–0.5	3	2^{27}	0.010	0.340	0.790	0.245	0.825	0.945
	12	2^{18}	0.055	0.485	0.875	0.015	0.275	0.955
	21	2^{9}	0.230	0.655	0.905	0	0.04	0.975
	30	2^{1}	1	1	1	1	1	1
0	3	2^{27}	0.010	0.325	0.825	0.215	0.830	0.935
	12	2^{18}	0.060	0.470	0.890	0.005	0.235	0.965
	21	2^{9}	0.265	0.660	0.920	0	0.025	0.985
	30	2^{0}	1	1	1	1	1	1
0.5	3	2^{27}	0.015	0.305	0.825	0.220	0.855	0.965
	12	2^{18}	0.105	0.505	0.895	0.020	0.290	0.960
	21	2^{9}	0.230	0.680	0.925	0	0.030	0.985
	30	2^{0}	1	1	1	1	1	1

Table 2

The first three columns list the value of ρ, $|M_0|$ and $|\mathcal{M}_t^n|$, respectively. The 4th-9th columns show the proportion of selecting the smallest true model by AIC, BIC, RIC, MRIC, CIC and RIC$_c$ based on 200 simulation replications with n = 5000, p_n = 30 and σ^2 = 1.

ρ	$\|M_0\|$	$\|\mathrm{M}_t^n\|$	AIC	BIC	RIC	MRIC	CIC	RIC$_c$
–0.5	3	2^{27}	0.015	0.920	0.770	0.165	0.810	0.950
	12	2^{18}	0.040	0.940	0.875	0.015	0.190	0.975
	21	2^{9}	0.180	0.980	0.950	0	0.030	0.980
	30	2^{0}	1	1	1	1	1	1
0	3	2^{27}	0.010	0.935	0.785	0.170	0.825	0.950
	12	2^{18}	0.045	0.940	0.820	0.010	0.230	0.945
	21	2^{9}	0.200	0.980	0.910	0	0.045	0.990
	30	2^{0}	1	1	1	1	1	1
0.5	3	2^{27}	0.005	0.695	0.765	0.190	0.805	0.930
	12	2^{18}	0.030	0.780	0.840	0.005	0.245	0.955
	21	2^{9}	0.235	0.900	0.910	0	0.020	0.975
	30	2^{0}	1	1	1	1	1	1

Table 3

prop$_1$: proportion of including the smallest true model by LAR; *prop*$_2$: proportion of selecting true models by LAR-RIC$_c$; *prop*$_3$: proportion of selecting the smallest true model by LAR-RIC$_c$; *median*: median model size selected by LAR-RIC$_c$; *rmse*$_1$: root MSE computed by the least square estimate of the model selected by LAR-RIC$_c$; *rmse*$_2$: root MSE computed by the LASSO estimate of the model selected by LAR-RIC$_c$ based on 200 replications with $\sigma^2 = 1$, $\rho = -0.5$, $n = 50$ and $|M_0| = 5$.

p_n	$prop_1$	$prop_2$	$prop_3$	median	$rmse_1$	$rmse_2$
25	1.00	1.00	1.00	5.00	2.31	4.39
50	1.00	1.00	1.00	5.00	2.34	4.87
75	1.00	1.00	1.00	5.00	2.33	5.38
100	1.00	1.00	1.00	5.00	2.29	5.48
250	1.00	1.00	1.00	5.00	2.35	6.55
500	0.99	1.00	0.98	5.00	2.38	7.09
750	1.00	1.00	0.99	5.00	2.33	7.70
1000	0.99	1.00	0.98	5.00	2.33	8.02
1500	1.00	1.00	0.99	5.00	2.35	8.33
2000	1.00	1.00	0.99	5.00	2.34	8.48
2500	0.99	1.00	0.95	5.00	2.28	8.72

Table 4

prop$_1$: proportion of including the smallest true model by LAR; *prop*$_2$: proportion of selecting true models by LAR-RIC$_c$; *prop*$_3$: proportion of selecting the smallest true model by LAR-RIC$_c$; *median*: median model size selected by LAR-RIC$_c$; *rmse*$_1$: root MSE computed by the least square estimate of the model selected by LAR-RIC$_c$; *rmse*$_2$: root MSE computed by the LASSO estimate of the model selected by LAR-RIC$_c$ based on 200 replications with $\sigma^2 = 1$, $\rho = 0$, $n = 50$ and $|M_0| = 5$.

p_n	prop$_1$	prop$_2$	prop$_3$	median	rmse$_1$	rmse$_2$
25	0.95	1.00	0.91	5.00	2.34	7.74
50	0.90	1.00	0.84	5.00	2.43	9.22
75	0.87	1.00	0.74	5.00	2.48	10.02
100	0.84	1.00	0.71	5.00	2.57	10.58
250	0.68	1.00	0.50	6.00	2.81	12.94
500	0.45	1.00	0.28	7.00	3.30	13.99
750	0.34	0.98	0.18	8.00	3.58	14.39
1000	0.32	0.98	0.16	8.00	3.68	14.89
1500	0.25	0.95	0.09	10.00	4.25	15.58
2000	0.21	0.94	0.05	11.00	4.54	15.72
2500	0.13	0.90	0.03	13.50	4.95	16.26

Table 5

prop$_1$: proportion of including the smallest true model by LAR; *prop*$_2$: proportion of selecting true models by LAR-RIC$_c$; *prop*$_3$: proportion of selecting the smallest true model by LAR-RIC$_c$; *median*: median model size selected by LAR-RIC$_c$; *rmse*$_1$: root MSE computed by the least square estimate of the model selected by LAR-RIC$_c$; *rmse*$_2$: root MSE computed by the LASSO estimate of the model selected by LAR-RIC$_c$ based on 200 replications with $\sigma^2 = 1$, $\rho = 0.5$, $n = 50$ and $|M_0| = 5$.

p_n	prop$_1$	prop$_2$	prop$_3$	median	rmse$_1$	rmse$_2$
25	0.35	1.00	0.27	6.00	2.76	12.50
50	0.10	0.98	0.08	8.00	3.17	14.70
75	0.05	0.99	0.02	10.00	3.56	15.52
100	0.02	0.97	0.01	11.00	4.05	15.93
250	0.00	0.65	0.00	22.00	6.87	17.67
500	0.00	0.22	0.00	28.00	8.63	17.70
750	0.00	0.09	0.00	31.00	8.84	17.37
1000	0.00	0.04	0.00	31.00	8.61	17.06
1500	0.00	0.01	0.00	33.00	8.35	16.76
2000	0.00	0.00	0.00	33.00	8.30	16.58
2500	0.00	0.00	0.00	33.00	8.19	16.44

Table 6

Median prediction error (MPE) and the number of selected nodes ($|\hat{M}|$)

Procedure	LAR.AIC	LAR.BIC	LAR.RICc	STEP.AIC	STEP.BIC	STEP.RICc
MPE	1.806	1.671	1.328	49.470	38.117	2.290
$\|\hat{M}\|$	9	7	4	101	95	5