

Consistent Linear Model Selection

Meng Zhao

and

K.B. Kulasekera

Department of Mathematical Sciences

Clemson University

Clemson, SC 29634

Abstract

We examine the penalty term in linear model selection using penalized least squares. The rate of divergence of the penalty term for consistent model selection is discussed under a general error structure.

Key Words: Design variables; Linear Model; Squared Error Loss.

1 Introduction

Consider the regression model

$$\mathbf{y}_n = \boldsymbol{\mu}_n + \mathbf{e}_n, \quad (1)$$

where $\mathbf{y}_n = (y_1, y_2, \dots, y_n)'$ is a vector of n independent responses, with unknown mean vector $\boldsymbol{\mu}_n = (\mu_1, \dots, \mu_n)'$, and $\mathbf{e}_n = (e_1, \dots, e_n)'$ is a vector of n independent, identically distributed errors with common mean 0 and variance σ^2 . Suppose that associated with y_i there is a p_n vector of covariates $\mathbf{x}_i = (x_{i1}, \dots, x_{ip_n})'$, and let $\mathbf{X}_n = (\mathbf{x}_1, \dots, \mathbf{x}_n)'$ be the $n \times p_n$ design matrix, which for simplicity, is assumed to have full rank. Generally, p_n is assumed to be less than n . To estimate $\boldsymbol{\mu}_n$ we propose the linear model $\boldsymbol{\mu}_n = \mathbf{X}_n \boldsymbol{\beta}_n$ where $\boldsymbol{\beta}_n = (\beta_1, \dots, \beta_{p_n})$ is a p_n vector of real valued parameters. It is possible that some of the components of $\boldsymbol{\beta}_n$ are 0. Thus, we may consider submodels of the form

$$\boldsymbol{\mu}_n = \mathbf{X}_n(\alpha) \boldsymbol{\beta}(\alpha), \quad (2)$$

where α is a subset of $\{1, \dots, p_n\}$ and $\boldsymbol{\beta}(\alpha)$ ($\mathbf{X}_n(\alpha)$) is the subvector (sub-matrix) containing the components of $\boldsymbol{\beta}_n$ (columns of \mathbf{X}_n) that are indexed by the integers in α .

Instead of taking all of the $2^{p_n} - 1$ models into account, we assume that there is a class \mathcal{A}_n of subsets of $\{1, \dots, p_n\}$ and only models that correspond to members of \mathcal{A}_n are considered to be candidate models. Let α_0 be the true model, that is, $\beta_i \neq 0$ for $i \in \alpha_0$ and $\beta_i = 0$ for $i \notin \alpha_0$. A model $\alpha \in \mathcal{A}_n$ is said to be correct if $\boldsymbol{\mu}_n = \mathbf{X}_n(\alpha)\boldsymbol{\beta}_n(\alpha)$ holds. This is equivalent to $\alpha_0 \subset \alpha$. Denote the set of all correct models in \mathcal{A}_n by \mathcal{AC}_n , and define $\mathcal{AI}_n = \mathcal{A}_n \setminus \mathcal{AC}_n$. For each model in \mathcal{A}_n , we use least squares method to estimate $\boldsymbol{\mu}_n$. Let $\mathbf{M}_n(\alpha) = \mathbf{X}_n(\alpha)(\mathbf{X}_n(\alpha)'\mathbf{X}_n(\alpha))^{-1}\mathbf{X}_n(\alpha)'$ be the hat matrix. Then $\hat{\boldsymbol{\mu}}_n(\alpha) = \mathbf{M}_n(\alpha)\mathbf{y}_n$ is the least square estimator (LSE) of $\boldsymbol{\mu}_n$ under model α .

The dimension of a model α is defined to be $|\alpha|$, the number of elements in α . We are interested in procedures for finding correct model(s) that have the smallest dimension (we assume that such “smallest” correct model(s) exist). Many of the widely used procedures including AIC (Akaike, 1970), BIC (Schwarz, 1978), C_p (Mallows, 1973) and GIC (Rao and Wu, 1989) choose the minimizer $\hat{\alpha}_n$ of a criterion that has the form

$$T_{n,\lambda_n}(\alpha) = \frac{1}{n}RSS_n(\alpha) + \frac{\lambda_n |\mathbf{M}_n(\alpha)| \hat{\sigma}^2}{n}, \quad (3)$$

with respect to all α in \mathcal{A}_n , where $RSS_n(\alpha) = (\mathbf{y}_n - \hat{\boldsymbol{\mu}}_n(\alpha))'(\mathbf{y}_n - \hat{\boldsymbol{\mu}}_n(\alpha))$ is the residual sum of squares, $|\mathbf{M}_n(\alpha)|$ is the rank of the matrix $\mathbf{M}_n(\alpha)$, $\hat{\sigma}^2$ is a consistent estimator of the error variance using the full model and λ_n controls the degree of penalization. The term λ_n has a major role in the consistency of the selection where we say that a model selection procedure is to be consistent if

$$P(\hat{\alpha}_n = \alpha_n^c) \rightarrow 1, \quad (4)$$

where $\hat{\alpha}_n$ is the index set of the model chosen by the procedure and $\alpha_n^c \in \mathcal{A}_n$ is a smallest correct model.

In this paper we discuss the role of λ_n in consistent model selection. The literature on model selection is very extensive. Shao (1997), Eubank (1998), Fan and Li (2001), Shi and Tsai (2002) and the citations therein give a good account on the available methods and properties of these methods. Some of these authors discuss the conditions on λ_n for consistent model selection. In particular, Shao (1997) shows that the selection is consistent if $\lambda_n \rightarrow \infty$

when the error distribution has at least eight finite moments. Zheng and Loh (1997) discuss model selection for errors with finite variance under the assumptions that $p_n/n \rightarrow 0$ and the penalty term diverges faster than p_n where p_n is the size of the full model. Shi and Tsai (2002) establish that for consistent model selection, λ_n should grow at a rate of $\log(n)$ under the condition that the error distribution has at least four finite moments and other restrictions on the model (see discussion after Theorem 4 below).

In this article, we establish conditions on the penalty term λ_n in (3) in order to satisfy (4) for general model settings that cover as special cases the cases that have been discussed in the literature. We will establish the conditions on the penalty term for a wide class of models that can include past responses as predictors for current responses and models that are more general than the classical hierarchical models. In particular, for models with finite error variance, we determine the rate of divergence of the penalty term for consistent model selection with mild restrictions (compared with those in Zheng and loh (1997)) on the model size p_n and the class of candidate models. We also show that λ_n can diverge to ∞ at any rate if the error distribution has a finite fourth moment, a more flexible choice on the sequence λ_n than that of Shao (1997) and Shi and Tsai (2002). and verifying choices like $\lambda_n = \ln(n)$ that has been used in the literature with normal error structures. We also give examples applying the technical results that are developed.

Fan and Li (2001) deal with the model selection using a penalized likelihood where the distribution of the responses is assumed to have regularity properties (Lehmann and Casella, 1998). The penalty term in their approach is in the form of a sum of suitable functions. We show that their results agree with those in this article for normal log-likelihood and an appropriate class of penalty functions. In the process we show that for the results in their paper to hold, they actually need a slightly stronger differentiability condition than what was assumed.

In the remainder of this paper, we shall use $\hat{\alpha}_n$ to denote the model selected by minimizing (3). We will discuss the consistency of $\hat{\alpha}_n$ in the next section.

2 Main Results

We first list several assumptions that are used in the sequel. Most of these have been used in Shao (1997) among others. The error distribution is as-

summed to be unknown with zero mean and finite variance.

Assumption 1. The smallest correct model denoted by α_n^c exists and for every model α in \mathcal{AC}_n , we have $\alpha_n^c \subset \alpha$. Let $r_n = |\alpha_n^c|$

Assumption 2. $\sup_n r_n < \infty$.

For a set of models \mathcal{A} , let $\mathcal{A}^{(k)} = \{\alpha | \alpha \in \mathcal{A}, |\alpha| \leq k\}$, and $\mathcal{A}_n^{[k]} = \mathcal{A}^{(k)} \setminus \mathcal{A}^{(k-1)}$ and $\alpha_{\mathcal{A}} = \cup_{\alpha \in \mathcal{A}} \alpha$.

Assumption 3. There exists a sequence of positive numbers $\{d_1, d_2, \dots\}$ such that $|\alpha_{\mathcal{A}_n^{(k)}}| \leq d_n k$ for all k and n .

Note that this allows for a larger class of models than the hierarchical model assumption in Zheng and Loh (1997). With this requirement \mathcal{A}_n can be created without any preordering of covariates. This is a very desirable feature on the class of candidate models. Now we give a series of Lemmas and Theorems establishing the consistency of model selection under various conditions on λ_n and the error distribution. The proofs of these statements are deferred to an Appendix.

Lemma 1. *Suppose assumptions 1 and 2 hold and*

$$\lim_{n \rightarrow \infty} \frac{\sum_{1 \leq k} |\alpha_{\mathcal{A}_n^{(k)}} \setminus \alpha_{\mathcal{A}_n^{(k-1)}}| / k}{\lambda_n} = 0. \quad (5)$$

Let $\mathcal{D}_n = \mathcal{A}_n \setminus \mathcal{A}_n^{(r_n)}$. Then $P(\hat{\alpha}_n \in \mathcal{D}_n) \rightarrow 0$.

The term $\sum_{1 \leq k} |\alpha_{\mathcal{A}_n^{(k)}} \setminus \alpha_{\mathcal{A}_n^{(k-1)}}| / k$ in condition (5) determines how fast the number of variables included in the models grows with k and how much the models overlap. ****give an intuitive reasoning where this is useful**** If assumption 3 is satisfied and the penalty goes to infinity faster than $d_n \log(p_n)$, then, it can be shown that (5) is satisfied. Then, we have the following lemma.

Lemma 2. *Under Assumptions 1, 2, and 3, and under the condition that $d_n \log p_n / \lambda_n \rightarrow 0$ we have $P(\hat{\alpha}_n \in \mathcal{D}_n) \rightarrow 0$.*

For a collection of models \mathcal{A} , let $\bar{\mathcal{A}}$ denote the set of all maximum elements in \mathcal{A} (an element α in \mathcal{A} is said to be maximum if for any $\beta \in \mathcal{A}$ such that $\alpha \subset \beta$, we have $\beta = \alpha$). Let $\Delta_n(\alpha) = \boldsymbol{\mu}'_n (\mathbf{I} - \mathbf{M}_n(\alpha)) \boldsymbol{\mu}_n / n$. Then we can prove

Lemma 3. *Suppose that Assumption 1 holds. Let $\mathcal{B}_n \subset \mathcal{AT}_n, n = 1, \dots$. If*

$$\liminf_n \min_{\alpha \in \mathcal{B}_n} n\Delta(\alpha)/\lambda_n = \infty, \quad (6)$$

and

$$\sum_{\alpha \in \overline{\mathcal{B}_n}} |\alpha|/n\Delta_n(\alpha) \rightarrow 0, \quad (7)$$

then, $P(\hat{\alpha}_n \in \mathcal{B}_n) \rightarrow 0$.

If Assumptions 2 and 3 hold, and the \mathcal{B}_n in Lemma 3 is taken to be $\mathcal{AT}_n^{(r_n)}$, then, $|\alpha_{\mathcal{AT}_n^{(r_n)}}| \leq d_n r_n$. Thus, $\max_{\alpha \in \overline{\mathcal{AT}_n^{(r_n)}}} |\alpha| \leq r_n$ and $|\overline{\mathcal{AT}_n^{(r_n)}}| = O(d_n^{r_n})$. Since r_n is bounded we see that (7) with $\mathcal{B}_n = \mathcal{AT}_n^{(r_n)}$ is satisfied if

$$\min_{\alpha \in \mathcal{AT}_n^{(r_n)}} n\Delta_n(\alpha)/d_n^{r_n} \rightarrow \infty. \quad (8)$$

Note that if d_n is of order $\log n$ or n^α for some $\alpha < \inf_n 1/r_n$, then (8) is a weaker constraint than (2.5) in Shao (1997). With this observation and Lemma 2 we have

Theorem 4. *Suppose Assumptions 1, 2, and 3 hold, and (8) is satisfied. Also suppose that*

$$\min_{\alpha \in \mathcal{AT}_n^{(r_n)}} n\Delta_n(\alpha)/\lambda_n \rightarrow \infty, \quad (9)$$

and

$$d_n \log p_n/\lambda_n \rightarrow 0 \quad (10)$$

Then, $P(\hat{\alpha}_n = \alpha_n^c) \rightarrow 1$.

Remark 1. Zheng and Loh (1997) use a penalty term in the form $h_n(k)$, where k is the size of the model in a hierarchical model selection with random covariates. Their $h_n(k)$ plays the same role as $\lambda_n k$ in this note. If we allow a general form $\lambda_n(k)$, where $\lambda_n(k)$ is increasing for every n , by using the same techniques in the proofs of Lemmas 1, 2, 3 and Theorem 4, we can show the following.

Corollary 5. *Suppose assumptions 1, 2 are met and assumption 3 is satisfied with d_n bounded. Furthermore, suppose*

$$\liminf_n \min_{\alpha \in \mathcal{AT}_n^{(r_n)}} n\Delta_n(\alpha)/\lambda_n(r_n) = \infty, \quad (11)$$

and

$$\sum_{r_n+1 \leq k \leq p_n} 1/(\lambda_n(k) - \lambda_n(r_n)) \rightarrow 0. \quad (12)$$

Then $P(\hat{\alpha}_n = \alpha_n^c) \rightarrow 1$.

Note that (12) is a weaker condition on the penalty terms than the combination of conditions B2 and B3 in Zheng and Loh (1997). Also, $p_n/n \rightarrow 0$ is not required.

Remark 2. Comparing Theorem 4 with Theorem 1 of Shi and Tsai(2002) (ST hereafter), we find that the assumptions here are weaker. The assumption 1 of ST requires the errors to have at least four finite moments, while we only require finite variance. We do not need a constraint like the assumption 2 of ST. Also, we only need $\min_{\alpha \in \mathcal{AT}_n^{(r_n)}} n\Delta_n(\alpha)/\lambda_n \rightarrow \infty$. Since λ_n is allowed to go to infinity faster than $\log n$ for a reasonably chosen λ_n , this is a weaker condition on $\Delta_n(\alpha)$ than that of assumption 3 in ST, which required $\min_{\alpha \in \mathcal{AT}_n^{(r_n)}} n\Delta_n(\alpha)/n \rightarrow \infty$. If the size of the models, p_n , is bounded, then it is easy to check that (5) is satisfied, and then, by Lemma 1 above, consistency is achieved as long as $\lambda_n \rightarrow \infty$, a more general result than Theorem 1 of ST. Even with stronger assumptions they only showed that the selection avoids over estimation in the weak sense. However, these stronger assumptions were used to show that the procedure will not under-estimate in the strong sense.

Example 1. Suppose that for each n , the p_n explanatory variables are arranged in decreasing order of importance, and the true model is $\alpha_0 = \{1, \dots, p\}$ for some positive integer p . We are interested in models that have the form $\alpha_i = \{1, \dots, i\}$ for $i \leq p_n$. Let $\mathcal{A}_n = \{\alpha_1, \dots, \alpha_{p_n}\}$, $n = 1, 2, \dots$. Then $\mathcal{AC}_n = \{\alpha_1, \dots, \alpha_{p_n}\}$ for $p_n \geq p$, $r_n = p$, and $\mathcal{AT}_n = \{\alpha_1, \dots, \alpha_{p-1}\}$. It is easy to see that assumptions 1, 2 are satisfied and 3 is satisfied with $d_n = 1$, $n = 1, 2, \dots$. So, if conditions (9) and (10) are satisfied, then $\hat{\alpha}_n$ is consistent in the sense of (4). It should be noted that Shao's technique gives a similar result if we assume that e_1 has finite eighth moment.

Example 2. Suppose that the explanatory variables are indexed by two integers, which means that the set of variables can be represented as

$$\{x_1, \dots, x_{p_n}\} = \{z_{1,1}, \dots, z_{1,s_n}, \dots, z_{t_n,1}, \dots, z_{t_n,s_n}\}.$$

This is the case when we need not only to select the right number of variables, but also the right ‘‘order’’ for each variable. In this setting, again we assume

that the variables are arranged in decreasing order of importance, and if we include a certain order of a variable in a model, we also include all the lower orders of that variable. Thus, a model in this case has the form

$$\alpha_{j:i_1, \dots, i_j} = \{z_{1,1}, \dots, z_{1,i_1}, \dots, z_{j,1}, \dots, z_{j,i_j}\},$$

and

$$\mathcal{A}_n = \{a_{j:i_1, \dots, i_j} : j \leq t_n \text{ and } i_k \leq s_n \text{ for } k = 1, \dots, t_n\}$$

It can be shown that

$$\left| \alpha_{\mathcal{A}_n^{(k)}} \right| \leq \begin{cases} k(k+1), & \text{for } k \leq t_n; \\ t_n(t_n+1) + (k-t_n)t_n & \text{for } k > t_n. \end{cases}$$

and

$$\left| \alpha_{\mathcal{A}_n^{(k)}} \setminus \alpha_{\mathcal{A}_n^{(k-1)}} \right| \leq \begin{cases} k, & \text{for } k \leq t_n; \\ t_n, & \text{for } t_n < k \leq t_n + s_n. \end{cases}$$

Hence,

$$\begin{aligned} \sum_{k=1}^{p_n} \left| \alpha_{\mathcal{A}_n^{(k)}} \setminus \alpha_{\mathcal{A}_n^{(k-1)}} \right| / k &\leq t_n + t_n \sum_{k=t_n+1}^{t_n+s_n} 1/k \\ &\leq t_n + t_n \log(t_n + s_n). \end{aligned}$$

In particular, if we take s_n to be of order $\log n$ and t_n to be of order $\log n / \log \log n$, then (5) is satisfied if we have $\lambda_n / \log n \rightarrow \infty$. If the true model α_0 is finite, that is $|\alpha_0| < \infty$, it is clear that assumptions 1 and 2 are satisfied. If in addition (6) is true, then it can be shown that (7) with $\mathcal{B}_n = \mathcal{AI}_n^{(r_n)}$ is also true. And thus, by Lemma 1, Lemma 3 and the subsequent discussion, consistency in the sense of (4) is achieved.

Note that for the above results it is only necessary for the errors to have finite variance. Now suppose that we know that the errors also have finite fourth moment. We will show that this information will enable us to choose a λ_n diverging to infinity at any rate for consistent model selection.

Assumption 4. The iid errors e_1, \dots, e_n in (1) satisfy $E(e_1^4) = \tau < \infty$.

Lemma 6. *Suppose Assumptions 1, 2, 4 are satisfied. Also suppose that there exist constants a, b, c and a strictly increasing sequence of nonnegative numbers $u_{n,1} = 0, u_{n,2}, \dots$ for every n such that*

$$\sum_{u_{n,k} < i < u_{n,k+1}} \frac{|\alpha_{\mathcal{A}_n^{(i)}} \setminus \alpha_{\mathcal{A}_n^{(i-1)}}|}{i} \leq a, \text{ for } k = 1, \dots, n = 1, \dots, \quad (13)$$

$$|\alpha_{\mathcal{A}_n^{(u_{n,k})}}|/u_{n,k} \leq b, \text{ for } k = 2, \dots, \quad (14)$$

and

$$\sum_{1 < k} 1/u_{n,k} \leq c \text{ for } n = 1, \dots, \quad (15)$$

Then, $P(\hat{\alpha}_n \in \mathcal{D}_n) \rightarrow 0$ if $\lambda_n \rightarrow \infty$.

Now suppose that we are given assumptions 2 and 3. Also suppose that the sequence $\{d_n\}$ in assumption 3 is bounded by d . Let $u_{n,k} = 2^{k-1} - 1$ for $k = 1, 2, \dots$ give an intuitive reasoning for the choice of u 's in terms of the number of model terms. It is not hard to check that conditions (13), (14), and (15) of Lemma 6 are all satisfied if we set $a = 2d, b = d$, and $c = 2$. Notice also that the discussion after Lemma 3 is valid here. Then, by Lemma 6 we have the following theorem.

Theorem 7. *Suppose that assumptions 1, 2, 3, 4, and (9) hold. Moreover, let the sequence $\{d_n\}$ in assumption 3 be bounded and $\lambda_n \rightarrow \infty$. Then, $P(\hat{\alpha}_n = \alpha_n^c) \rightarrow 1$.*

Now reconsider the model selection problem described in Example 1. If $E(e_1^4) < \infty$, then, by Theorem 7, it is only necessary for the penalty to go to infinity to get consistency.

Remark 3. In Fan and Li (2001) model selection using the penalized likelihood function $L(\boldsymbol{\beta}) - n \sum_{j=1}^d p_{\eta_n}(|\beta_j|)$, where p_{η_n} are suitable penalty functions, is discussed. We can relate their penalty term $\sum p_{\eta_n}(|\beta_j|)$ and the penalty term in this paper if we assume that the errors are normal. Then the log likelihood function will be proportional to the $-RSS_n$, and it will be maximized by the least squares estimator. If in addition, the value of the function p_{η_n} is constant outside an interval $(-\delta_n, \delta_n)$, where $\delta_n \rightarrow 0$, then the $n \sum p_{\eta_n}(|\beta_j|)$ term is comparable to the λ_n in this paper.

We note that for the argument in the proof of Lemma 1 in Fan and Li (2001) to hold, one needs a slightly stronger condition on $p_{\eta_n}(\theta)$ than was assumed. The following example shows that the argument about the sign of the derivative of Q in the proof of Lemma 1 in Fan and Li (2001) fails under the assumption $\liminf_{n \rightarrow \infty} \liminf_{\theta \rightarrow 0^+} \eta_n^{-1} p'_{\eta_n}(\theta) > 0$. Let $\eta_n = u_n n^{-1/2}$, where $u_n \rightarrow \infty$ and

$$p_{\eta_n}(\theta) = \begin{cases} \sin(n^{2/3}\theta), & \text{for } |\theta| \leq \pi/2n^{2/3}; \\ 1, & \text{otherwise.} \end{cases}$$

We see that $\liminf_{n \rightarrow \infty} \liminf_{\theta \rightarrow 0^+} \eta_n^{-1} p'_{\eta_n}(\theta) > 0$, $n^{-1/2}/\eta_n \rightarrow 0$, and $\max\{|p''_{\eta_n}(|\beta_{j0}|) : \beta_{j0} \neq 0\} \rightarrow 0$. However, for $\beta_j = cn^{-1/2}$, we have $\eta_n^{-1} p'_{\eta_n}(|\beta_j|) \rightarrow 0$ as $n \rightarrow \infty$. Thus a condition like

$$\liminf_{n \rightarrow \infty} \inf_{0 < \theta < cn^{-1/2}} \eta_n^{-1} p'_{\eta_n}(\theta) > 0$$

for some $c > 0$ has to be used. This requires $p'_{\eta_n}(\theta)$ to be at least of the same order as η_n in the interval $(0, cn^{-1/2})$, and this will force the value of $p_{\eta_n}(cn^{-1/2})$ to be of order $\eta_n n^{-1/2} = u_n/n$. This means that $np_{\eta_n}(cn^{-1/2}) \rightarrow \infty$, which coincides with the result in this paper, that $\lambda_n \rightarrow \infty$.

3 Appendix

In these proofs, for two matrices A and B , $A \geq B$, means that $A - B$ is positive semi-definite.

Proof of Lemma 1

Note that $\hat{\alpha}_n \in \mathcal{D}_n$ implies that $\min_{\alpha \in \mathcal{D}_n} T_{n,\lambda_n}(\alpha) - T_{n,\lambda_n}(\alpha_n^c) < 0$. Let

$\mathbf{A}_n(\alpha) = \mathbf{I} - \mathbf{M}_n(\alpha)$. For $\alpha \in \mathcal{D}_n^{[k]}$, $r_n < k \leq p_n$, we have

$$\begin{aligned}
& T_{n,\lambda_n}(\alpha) - T_{n,\lambda_n}(\alpha_n^c) \\
&= \frac{1}{n} (\mathbf{y}'_n (\mathbf{M}_n(\alpha_n^c) - \mathbf{M}_n(\alpha)) \mathbf{y}_n - \lambda_n (|\alpha_n^c| - |\alpha|) \hat{\sigma}^2) \\
&\geq \frac{1}{n} \left(\mathbf{y}'_n \left(\mathbf{M}_n(\alpha_n^c) - \mathbf{M}_n(\alpha_{\mathcal{A}_n^{(k)}}) \right) \mathbf{y}_n + \lambda_n (k - r_n) \hat{\sigma}^2 \right) \\
&= \Delta_n(\alpha_{\mathcal{A}_n^{(k)}}) + \frac{2}{n} \boldsymbol{\mu}'_n \mathbf{A}_n(\alpha_{\mathcal{A}_n^{(k)}}) \mathbf{e} \\
&+ \frac{1}{n} \mathbf{e}' \left(\mathbf{M}_n(\alpha_n^c) - \mathbf{M}_n(\alpha_{\mathcal{A}_n^{(k)}}) \right) \mathbf{e} + \lambda_n (k - r_n) \hat{\sigma}^2 / n \\
&= \frac{1}{n} \mathbf{e}' \left(\mathbf{M}_n(\alpha_n^c) - \mathbf{M}_n(\alpha_{\mathcal{A}_n^{(k)}}) \right) \mathbf{e} + \lambda_n (k - r_n) \hat{\sigma}^2 / n \\
&\geq -\frac{1}{n} \mathbf{e}' \mathbf{M}_n(\alpha_{\mathcal{A}_n^{(k)}}) \mathbf{e} + \lambda_n (k - r_n) \hat{\sigma}^2 / n
\end{aligned} \tag{16}$$

Thus

$$\begin{aligned}
& \min_{\alpha \in \mathcal{D}_n} T_{n,\lambda_n}(\alpha) < T_{n,\lambda_n}(\alpha_n^c) \Rightarrow \\
& \max_{r_n < k \leq p_n} \frac{\mathbf{M}_n(\alpha_{\mathcal{A}_n^{(k)}})}{k - r_n} > \lambda_n \hat{\sigma}^2
\end{aligned} \tag{17}$$

Define $\alpha_{\mathcal{A}_n^{(0)}}$ to be the empty set. Set $\mathbf{M}(\alpha_{\mathcal{A}_n^{(0)}}) = \mathbf{0}$ and

$$\mathcal{M}_n = \sum_{1 \leq k \leq p_n} \left(\mathbf{M}_n(\alpha_{\mathcal{A}_n^{(k)}}) - \mathbf{M}_n(\alpha_{\mathcal{A}_n^{(k-1)}}) \right) / k.$$

For any integer $r_n < j \leq p_n$ we have

$$\begin{aligned}
(r_n + 1) \mathcal{M}_n &\geq (r_n + 1) \sum_{r_n < k \leq j} \left(\mathbf{M}_n(\alpha_{\mathcal{A}_n^{(k)}}) - \mathbf{M}_n(\alpha_{\mathcal{A}_n^{(k-1)}}) \right) / k \\
&\geq (r_n + 1) \sum_{r_n < k \leq j} \left(\mathbf{M}_n(\alpha_{\mathcal{A}_n^{(k)}}) - \mathbf{M}_n(\alpha_{\mathcal{A}_n^{(k-1)}}) \right) / j \\
&= (r_n + 1) \mathbf{M}_n(\alpha_{\mathcal{A}_n^{(j)}}) / j \\
&\geq \mathbf{M}_n(\alpha_{\mathcal{A}_n^{(j)}}) / (j - r_n).
\end{aligned}$$

Thus,

$$(r_n + 1) \mathbf{e}' \mathcal{M}_n \mathbf{e} \geq \max_{r_n < k \leq p_n} \mathbf{e}' \mathbf{M}_n(\alpha_{\mathcal{A}_n^{(k)}}) \mathbf{e} / (k - r_n). \tag{18}$$

Since

$$\begin{aligned} E[\mathbf{e}'\mathcal{M}_n\mathbf{e}] &= \sigma^2 \sum_{1 \leq k \leq p_n} \frac{\text{tr} \left(\mathbf{M}_n \left(\alpha_{\mathcal{A}_n^{(k)}} \right) - \mathbf{M}_n \left(\alpha_{\mathcal{A}_n^{(k-1)}} \right) \right)}{k} \\ &= \sigma^2 \sum_{1 \leq k \leq p_n} \frac{\left| \alpha_{\mathcal{A}_n^{(k)}} \setminus \alpha_{\mathcal{A}_n^{(k-1)}} \right|}{k}, \end{aligned}$$

we have

$$P \left((r_n + 1) \mathbf{e}'\mathcal{M}_n\mathbf{e} > \lambda_n \sigma^2 / 2 \right) \leq \frac{2(r_n + 1) \sum_{1 \leq k \leq p_n} \left| \alpha_{\mathcal{A}_n^{(k)}} \setminus \alpha_{\mathcal{A}_n^{(k-1)}} \right| / k}{\lambda_n} \rightarrow 0,$$

and

$$P \left(\hat{\sigma}^2 < \sigma^2 / 2 \right) \rightarrow 0.$$

Thus,

$$\begin{aligned} P \left((r_n + 1) \mathbf{e}'\mathcal{M}_n\mathbf{e} > \lambda_n \hat{\sigma}^2 \right) &\leq P \left((r_n + 1) \mathbf{e}'\mathcal{M}_n\mathbf{e} > \lambda_n \sigma^2 / 2 \right) \\ &\quad + P \left(\hat{\sigma}^2 < \sigma^2 / 2 \right) \\ &\rightarrow 0. \end{aligned}$$

This, combined with (18) gives

$$\begin{aligned} P \left(\hat{\alpha}_n \in \mathcal{AC}_n \setminus \mathcal{AC}_n^{(r_n)} \right) &\leq P \left(\min_{\alpha \in \mathcal{D}_n} T_{n, \lambda_n}(\alpha) \leq T_{n, \lambda_n}(\alpha_n^c) \right) \\ &\leq P \left((r_n + 1) \mathbf{e}'\mathcal{M}_n\mathbf{e} \geq \lambda_n \hat{\sigma}^2 \right) \rightarrow 0, \end{aligned}$$

giving the required result. \square

Proof of Lemma 2

For simplicity we write $\left| \alpha_{\mathcal{A}_n^{(i)}} \setminus \alpha_{\mathcal{A}_n^{(i-1)}} \right|$ as a_i for $i = 1, 2, \dots$. It is easy to see that $a_i = 0$ for $i > p_n$. First we will prove that

$$\sum_{i=1}^{p_n} a_i / i \leq d_n \sum_{i=1}^{p_n} 1 / i$$

In fact, by assumption 3 we have

$$\begin{aligned}
\sum_{i=1}^{p_n} a_i/i &= a_1 \left(1 - \frac{1}{2}\right) + (a_1 + a_2) \left(\frac{1}{2} - \frac{1}{3}\right) + \dots \\
&+ (a_1 + \dots + a_{p_n-1}) \left(\frac{1}{p_n-1} - \frac{1}{p_n}\right) \\
&+ (a_1 + \dots + a_{p_n}) \left(\frac{1}{p_n}\right) \\
&= \sum_{j=1}^{p_n-1} |\alpha_{\mathcal{A}_n^{(j)}}| \left(\frac{1}{j} - \frac{1}{j+1}\right) + |\alpha_{\mathcal{A}_n^{(p_n)}}| \frac{1}{p_n} \\
&\leq \sum_{j=1}^{p_n-1} d_n j \left(\frac{1}{j} - \frac{1}{j+1}\right) + d_n p_n \frac{1}{p_n} \\
&= d_n \sum_{i=1}^{p_n} 1/i.
\end{aligned}$$

Since $\sum_{i=1}^{p_n} 1/i = O(\log p_n)$. Thus (5) in Lemma 1 is satisfied. \square

Proof of Lemma 3

We know that $\hat{\alpha}_n \in \mathcal{B}_n$ implies that

$$\min_{\alpha \in \mathcal{B}_n} T_{n,\lambda_n}(\alpha) \leq T_{n,\lambda_n}(\alpha_n^c) \quad (19)$$

For $\alpha \in \mathcal{B}_n$, we can find a $\beta \in \overline{\mathcal{B}_n}$ such that $\alpha \subset \beta$. We have

$$\begin{aligned}
&T_{n,\lambda_n}(\alpha) - T_{n,\lambda_n}(\alpha_n^c) \\
&= \frac{1}{n} (\mathbf{y}'_n (\mathbf{M}_n(\alpha_n^c) - \mathbf{M}_n(\alpha)) \mathbf{y}_n - \lambda_n (|\alpha_n^c| - |\alpha|) \hat{\sigma}^2) \\
&\geq \frac{1}{n} (\mathbf{y}'_n (\mathbf{M}_n(\alpha_n^c) - \mathbf{M}_n(\beta)) \mathbf{y}_n - \lambda_n r_n \hat{\sigma}^2) \\
&= \frac{1}{n} \boldsymbol{\mu}'_n (\mathbf{M}_n(\alpha_n^c) - \mathbf{M}_n(\beta)) \boldsymbol{\mu}_n + \frac{2}{n} \boldsymbol{\mu}'_n (\mathbf{M}_n(\alpha_n^c) - \mathbf{M}_n(\beta)) \mathbf{e} \\
&+ \frac{1}{n} \mathbf{e}' (\mathbf{M}_n(\alpha_n^c) - \mathbf{M}_n(\beta)) \mathbf{e} - \frac{1}{n} \lambda_n r_n \hat{\sigma}^2 \\
&\geq \Delta_n(\beta) + \frac{2}{n} \boldsymbol{\mu}'_n \mathbf{A}_n(\beta) \mathbf{e} - \lambda_n r_n \hat{\sigma}^2/n - \frac{1}{n} \mathbf{e}' (\mathbf{M}_n(\beta)) \mathbf{e}
\end{aligned}$$

(19) implies that at least one of the three inequalities:

$$\max_{\beta \in \overline{\mathcal{B}_n}} \left| \frac{\boldsymbol{\mu}'_n \mathbf{A}_n(\beta) \mathbf{e}}{n \Delta_n(\beta)} \right| \geq 1/3$$

$$\begin{aligned} \max_{\beta \in \overline{\mathcal{B}_n}} \frac{\mathbf{e}'(\mathbf{M}_n(\beta))\mathbf{e}}{n\Delta_n(\beta)} &\geq 1/3 \\ \max_{\beta \in \overline{\mathcal{B}_n}} \lambda_n r_n \hat{\sigma}^2 / n\Delta_n(\beta) &\geq 1/3 \end{aligned}$$

holds. Since $Var(\boldsymbol{\mu}'_n \mathbf{A}_n(\beta)\mathbf{e}) = n\sigma^2\Delta_n(\beta)$, we have for any positive δ

$$P\left(\max_{\beta \in \overline{\mathcal{B}_n}} \left| \frac{\boldsymbol{\mu}'_n \mathbf{A}_n(\beta)\mathbf{e}}{n\Delta_n(\beta)} \right| \geq \delta\right) \leq \sum_{\beta \in \overline{\mathcal{B}_n}} \frac{\sigma^2}{\delta^2 n\Delta_n(\beta)} \rightarrow 0. \quad (20)$$

Also, $E[\mathbf{e}'(\mathbf{M}_n(\beta))\mathbf{e}] = \sigma^2|\beta|$, so that

$$P\left(\max_{\beta \in \overline{\mathcal{B}_n}} \frac{\mathbf{e}'(\mathbf{M}_n(\beta))\mathbf{e}}{n\Delta_n(\beta)} \geq \delta\right) \leq \sum_{\beta \in \overline{\mathcal{B}_n}} \frac{\sigma^2|\beta|}{\delta n\Delta_n(\beta)} \rightarrow 0. \quad (21)$$

By condition (6), we also have $\lambda_n r_n \hat{\sigma}^2 / n\Delta_n(\beta) \rightarrow 0$. Thus

$$P(\min_{\alpha \in \overline{\mathcal{B}_n}} T_{n,\lambda_n}(\alpha) > T_{n,\lambda_n}(\alpha_n^c)) \rightarrow 1.$$

□

Proof of Lemma 6

For $k = 1, 2, \dots$, define

$$\mathbf{N}_k = \frac{\mathbf{M}_n(\alpha_{\mathcal{A}_n^{(u_{n,k})}})}{u_{n,k}} + \sum_{i=u_{n,k}+1}^{u_{n,k+1}-1} \frac{\mathbf{M}_n(\alpha_{\mathcal{A}_n^{(i)}}) - \mathbf{M}_n(\alpha_{\mathcal{A}_n^{(i-1)}})}{i}$$

Define $0/0=0$. We have

$$\begin{aligned} E(\mathbf{e}'\mathbf{N}_k\mathbf{e}) &= \sigma^2 \left(\frac{|\alpha_{\mathcal{A}_n^{(u_{n,k})}}|}{u_{n,k}} + \sum_{u_{n,k} < i < u_{n,k+1}} \frac{|\alpha_{\mathcal{A}_n^{(i)}} \setminus \alpha_{\mathcal{A}_n^{(i-1)}}|}{i} \right) \\ &\leq \sigma^2(a+b). \end{aligned}$$

For $\delta > 2\sigma^2(a+b)$, we have

$$\begin{aligned} P(\mathbf{e}'\mathbf{N}_k\mathbf{e} > \delta) &\leq P(|\mathbf{e}'\mathbf{N}_k\mathbf{e} - E(\mathbf{e}'\mathbf{N}_k\mathbf{e})| > \delta/2) \\ &= P(|\mathbf{e}'\mathbf{N}_k\mathbf{e} - E(\mathbf{e}'\mathbf{N}_k\mathbf{e})|^2 > \delta^2/4) \\ &\leq \frac{E(|\mathbf{e}'\mathbf{N}_k\mathbf{e} - E(\mathbf{e}'\mathbf{N}_k\mathbf{e})|^2)}{\delta^2/4}. \end{aligned}$$

By Theorem 2 of Whittle (1960), we have

$$E \left(|\mathbf{e}'\mathbf{N}_k\mathbf{e} - E(\mathbf{e}'\mathbf{N}_k\mathbf{e})|^2 \right) \leq C (\text{tr}(\mathbf{N}'_k\mathbf{N}_k) \tau).$$

It is easy to see that

$$\mathbf{N}'_k\mathbf{N}_k = \frac{\mathbf{M}_n \left(\alpha_{\mathcal{A}_n^{(u_{n,k})}} \right)}{u_{n,k}^2} + \sum_{i=u_{n,k}+1}^{u_{n,k}+1-1} \frac{\mathbf{M}_n \left(\alpha_{\mathcal{A}_n^{(i)}} \right) - \mathbf{M}_n \left(\alpha_{\mathcal{A}_n^{(i-1)}} \right)}{i^2}.$$

Thus,

$$\begin{aligned} \text{tr}(\mathbf{N}'_k\mathbf{N}_k) &= \frac{|\alpha_{\mathcal{A}_n^{(u_{n,k})}}|}{u_{n,k}^2} + \sum_{i=u_{n,k}+1}^{u_{n,k}+1-1} \frac{|\alpha_{\mathcal{A}_n^{(i)}} \setminus \alpha_{\mathcal{A}_n^{(i-1)}}|}{i^2} \\ &\leq \begin{cases} a, & \text{for } k = 1; \\ \frac{a+b}{u_{n,k}}, & \text{for } 2 \leq k. \end{cases} \end{aligned}$$

Hence

$$P(\mathbf{e}'\mathbf{N}_k\mathbf{e} > \delta) \leq \begin{cases} \frac{aC\tau}{\delta^2}, & \text{for } k = 1; \\ \frac{(a+b)C\tau}{\delta^2 u_{n,k}}, & \text{for } k > 1, \end{cases}$$

for some constant C . Notice that for any $u_{n,k} \leq i < u_{n,k+1}$ we have

$$\begin{aligned} \mathbf{N}_k &\geq \frac{\mathbf{M}_n \left(\alpha_{\mathcal{A}_n^{(u_{n,k})}} \right)}{u_k} + \dots + \frac{\mathbf{M}_n \left(\alpha_{\mathcal{A}_n^{(i)}} \right) - \mathbf{M}_n \left(\alpha_{\mathcal{A}_n^{(i-1)}} \right)}{i} \\ &\geq \frac{\mathbf{M}_n \left(\alpha_{\mathcal{A}_n^{(u_{n,k})}} \right)}{i} + \dots + \frac{\mathbf{M}_n \left(\alpha_{\mathcal{A}_n^{(i)}} \right) - \mathbf{M}_n \left(\alpha_{\mathcal{A}_n^{(i-1)}} \right)}{i} \\ &= \frac{\mathbf{M}_n \left(\alpha_{\mathcal{A}_n^{(i)}} \right)}{i} \geq \frac{\mathbf{M}_n(\alpha)}{i} \text{ for any } \alpha \in \mathcal{A}_n^{[i]}. \end{aligned}$$

For $i > r_n$, suppose that $u_{n,k} \leq i < u_{n,k+1}$. Then,

$$(r_n + 1)\mathbf{e}'\mathbf{N}_k\mathbf{e} \geq \max_{\alpha \in \mathcal{A}_n^{[i]}} \frac{(r_n + 1)\mathbf{e}'\mathbf{M}_n(\alpha)\mathbf{e}}{i} \geq \max_{\alpha \in \mathcal{A}_n^{[i]}} \frac{\mathbf{e}'\mathbf{M}_n(\alpha)\mathbf{e}}{i - r_n}.$$

From this we get

$$\max_{r_n < i \leq p_n} \max_{\alpha \in \mathcal{A}_n^{[i]}} \frac{\mathbf{e}'\mathbf{M}_n(\alpha)\mathbf{e}}{i - r_n} \leq (r_n + 1) \max_{1 \leq k} \mathbf{e}'\mathbf{N}_k\mathbf{e}.$$

Noting $\mathcal{D}_n^{[k]} = \mathcal{A}_n^{[k]}$ for $r_n < k \leq p_n$ we have

$$\max_{r_n < k \leq p_n} \max_{\alpha \in \mathcal{D}_n^{[k]}} \frac{\mathbf{e}'\mathbf{M}_n(\alpha)\mathbf{e}}{k - r_n} \leq (r_n + 1) \max_{1 \leq k} \mathbf{e}'\mathbf{N}_k\mathbf{e}.$$

Thus, when λ_n is large enough, (17) gives

$$\begin{aligned} P(\hat{\alpha}_n \in \mathcal{D}_n) &\leq P\left((r_n + 1) \max_{1 \leq k} \mathbf{e}'\mathbf{N}_k\mathbf{e} > \lambda_n \hat{\sigma}^2\right) \\ &\leq P\left((r_n + 1) \max_{1 \leq k} \mathbf{e}'\mathbf{N}_k\mathbf{e} > \lambda_n \sigma^2/2\right) + P(\hat{\sigma}^2 < \sigma^2/2) \\ &\leq \sum_{1 \leq k} P\left((r_n + 1) \mathbf{e}'\mathbf{N}_k\mathbf{e} > \lambda_n \sigma^2/2\right) + P(\hat{\sigma}^2 < \sigma^2/2) \\ &\leq \frac{4(r_n + 1)^2 C \tau}{\lambda_n^2 \sigma^4} \left(a + \sum_{2 \leq k} \frac{a + b}{u_{n,k}}\right) + P(\hat{\sigma}^2 < \sigma^2/2) \\ &\leq \frac{(r_n + 1)^2 C(a + (a + b)c)\tau}{\lambda_n^2 \sigma^4} + P(\hat{\sigma}^2 < \sigma^2/2) \\ &\rightarrow 0. \quad \square \end{aligned}$$

Reference

- AKAIKE, H. (1970). Statistical predictor identification. *Ann. Inst. Statist. Math.* **22** 203-217.
- EUBANK, R. L. (1999). *Spline smoothing and non parametric regression*. Dekker, NY.
- FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348-1360.
- LEHMANN, E. L. and CAESLLA, G. (1998). *Theory of point estimation* (2nd ed.) Springer, NY.
- MALLOWS, C. L. (1973). Some comments on C_p . *Technometrics* **15** 661-675.
- RAO, C. R. and WU, Y. (1989). A strongly consistent procedure for model selection in a regression problem. *Biometrika* **76** 369-374.
- SCHWARZ, G. (1978). Estimating the dimensions of a model. *Ann. Statist.* **6** 461-464.
- SHAO, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica* **7** 221-264.
- SHI, P. and TSAI, C. (2002). Regression model selection —a residual likelihood approach. *J. Roy. Statist. Soc. Ser. B* **64** 237-252.

WHITTLE, P. (1960). Bounds for the moments of linear and quadratic forms in independent variables. *Theory Probab.Appl.* **5** 302-305.

ZHENG, X. and LOH, W. (1997). A consistent variable selection criterion for linear models with high-dimensional covariates. *Statistica Sinica* **7** 311-325