DEPARTMENT OF STATISTICS
University of Wisconsin
1210 West Dayton St.
Madison, WI 53706

TECHNICAL REPORT NO. 1091r

April 2004, Revised December 2004

A Note on the Lasso and Related Procedures in Model Selection

Chenlei Leng, Yi Lin and Grace Wahba [1]

National University of Singapore and Univeristy of Wisconsin-Madison

[1]Chenlei Leng is Assistant Professor, Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546. (E-mail: stalc@nus.edu.sg). Yi Lin is Associate Professor, Department of Statistics, University of Wisconsin, Madison, WI 53706 (E-mail: yilin@stat.wisc.edu). Grace Wahba is IJ Schoenberg-Hilldale Professor, Department of Statistics, University of Wisconsin, Madison, WI 53706 (E-mail: wahba@stat.wisc.edu). Leng's research was supported in part by NSF Grant DMS 0072292 and NIH Grant EYO9946. Lin's research was supported in part by NSF Grant DMS 0134987. Wahba's research was supported in part by NSF Grant DMS 0072292 and NIH Grant EYO9946.

A Note on the Lasso and Related Procedures in Model Selection

December 15, 2004

Chenlei Leng, Yi Lin and Grace Wahba

Abstract

The Lasso, the Forward Stagewise regression and the Lars are closely related procedures recently proposed for linear regression problems. Each of them can produce sparse models and can be used both for estimation and variable selection. In practical implementations these algorithms are typically tuned to achieve optimal prediction accuracy. We show that, when the prediction accuracy is used as the criterion to choose the tuning parameter, in general these procedures are not consistent in terms of variable selection. That is, the sets of variables selected are not consistent at finding the true set of important variables. In particular, we show that for any sample size n, when there are superfluous variables in the linear regression model and the design matrix is orthogonal, the probability of the procedures correctly identifying the true set of important variables is less than a constant (smaller than one) not depending on n. This result is also shown to hold for two dimensional problems with general correlated design matrices. The results indicate that in problems where the main goal is variable selection, prediction accuracy based criteria alone are not sufficient for this purpose. Adjustments will be discussed to make the Lasso and related procedures useful/consistent for variable selection.

Keyword: consistent model selection, Forward Stagewise regression, Lars, Lasso, variable selection

1 Introduction

The Least Absolute Shrinkage and Selection Operator (the Lasso) proposed by Tibshirani (1996) is a popular technique for model selection and estimation in linear regression models. It employs an L_1 type penalty on the regression coefficients which tends to produce sparse models, and thus is often used as a variable selection tool as in Tibshirani (1997), Osborne, Presnell & Turlach (2000). Knight & Fu (2000) studied the asymptotic properties of Lasso-type estimators. They showed that under appropriate conditions, the Lasso estimators are consistent for estimating the regression coefficients, and the limit distribution of the Lasso estimators can have positive probability mass at 0 when the true value of the parameter is 0. It has been demonstrated in Tibshirani (1996) that the Lasso is more stable and accurate than traditional variable selection methods such as the best subset selection. Efron, Hastie, Johnstone & Tibshirani (2004) proposed the Least Angle Regression (the Lars), and showed that there is a close connection between the Lars, the Lasso, and another model selection procedure called the Forward Stagewise regression. Each of these procedures involves a tuning parameter that is chosen to minimize the prediction error. This paper is concerned with the properties of the resulting estimators in terms of variable selection.

Consider the common Gaussian linear regression model

$$\mathbf{y} = X\beta + \epsilon,$$

where $\mathbf{y} = (y_1, ..., y_n)^T$ are the responses, $\beta = (\beta_1, ..., \beta_d)^T$ are the regression coefficients, $X = (\mathbf{x}_1, ..., \mathbf{x}_d)$ is the covariate matrix, and $\epsilon = (\epsilon_1, ..., \epsilon_n) \sim N(0, \sigma^2 I_n)$ are the normal noises. Without loss of generality, throughout this paper we assume that the covariates have been standardized to mean 0 and variance 1, and the response has mean 0. That is,

$$\mathbf{1}^T\mathbf{y} = 0,\ \mathbf{1}^T\mathbf{x}_j = 0,\ \text{and}\ \mathbf{x}_j^T\mathbf{x}_j = 1 \text{ for } j = 1, ..., d.$$

In many practical situations, some covariates are superfluous. That is, conditional on a subset of the covariates, the response does not depend on the other covariates. In other words, only a proper subset of the regression coefficients are nonzero. The

problem of variable selection is to identify this set of important covariates. A variable selection procedure is said to be consistent, if the probability that the procedure correctly identifies the set of important covariates approaches one when the sample size n goes to infinity. See, for example, Rao & Wu (1989), Shao (1997), Broman & Speed (2002) for some studies on the consistent variable selection problem.

It is of interest to investigate the consistency properties of the Lasso and related methods in terms of variable selection as they are often used as variable selectors. Tibshirani (1996) noted in one of the simulation examples, that in the majority of the runs the Lasso chose models that contain the true model, but only in a small fraction of runs did the Lasso pick the correct model. Fan and Li (2001) studied the penalized likelihood methods in linear regression, of which the Lasso is a special case. They proposed a nonconcave penalized likelihood method that enjoys the oracle property when the tuning parameter is appropriately chosen. The nonconcave penalized likelihood method is consistent in terms of variable selection, and it estimates the nonzero regression coefficients as well as when the correct submodel is known. They conjectured that the Lasso does not enjoy the oracle property. In this paper we show that when the tuning parameter is chosen to minimize the prediction error, as is commonly done in practice, in general the Lasso and related procedures are not consistent variable selectors. In particular, we show that when there are superfluous variables in the linear regression model and the design matrix is orthogonal, the probability of the procedures correctly identifying the true set of important variables is less than a constant (smaller than one) not depending on n. The result is also shown to hold for certain correlated design matrices. These are actually finite sample results, since they are true for any sample size n. These results indicate that in variable selection problems, prediction accuracy based criteria are not suitable for these procedures. Simple adjustments do exist to make the Lasso and related procedures consistent for variable selection, and we discuss possible ways to achieve this.

The remaining part of this article is organized as follows. In section 2, we review the Lasso, the Lars and the Forward Stagewise regression. In section 3, we consider a two dimensional problem with general correlation structure and demonstrate that the three methods fail to find the right model with certain probability, when the tuning

parameters are chosen to minimize the prediction error. The results concerning higher dimension problems are given in section 4 for orthogonal designs. We present some simulation results in section 5 and discussions are given in section 6.

2 The Lasso, the Lars and the Forward Stagewise regression

The Lasso estimate is the solution to

$$\min_{\beta} (\mathbf{y} - X\beta)^T(\mathbf{y} - X\beta), \quad \text{s.t.} \sum_{j=1}^{d} |\beta_j| \leq t. \tag{2.1}$$

Here $t \geq 0$ is a tuning parameter. Let $\hat{\beta}^0$ be the ordinary least square (OLS) estimate and $t_0 = \sum |\hat{\beta}_j^0|$. Values of $t < t_0$ will shrink the solutions toward 0. As shown in Tibshirani (1996), the Lasso gives sparse interpretable models and has excellent prediction accuracy. An alternative formulation of the Lasso is to solve the penalized likelihood problem

$$\min_{\beta} \frac{1}{n}(\mathbf{y} - X\beta)^T(\mathbf{y} - X\beta) + \lambda \sum_{j=1}^{d} |\beta_j|. \tag{2.2}$$

(2.1) and (2.2) are equivalent in the sense that for any given $\lambda \in [0, \infty)$, there exists a $t \geq 0$ such that the two problems have the same solution, and vice versa.

The Forward Stagewise regression, which will be called the FSW hereafter, is an iterative procedure, where successive estimates are built via a series of small steps. Letting $\eta = X\beta$, and beginning with $\hat{\eta}_0$ =0, if $\hat{\eta}$ is the current estimate, the next step is taken in the direction of the greatest correlation between covariate $\mathbf{x}_j$ and the current residual. That is, writing $\hat{\mathbf{c}} = X^T(\mathbf{y} - \hat{\eta})$ and $\hat{j} = \text{argmax}|\hat{c}_j|$, the update is

$$\hat{\eta} \leftarrow \hat{\eta} + \epsilon \cdot \text{sign}(\hat{c}_{\hat{j}}) \cdot \mathbf{x}_{\hat{j}},$$

where $\epsilon > 0$ is some constant. Smaller ϵ yields less greedy algorithm for the FSW and is recommended.

The Lars is a newly proposed model selection tool. We briefly describe the procedure in the following. For a detailed account of the procedure, the readers are referred to Efron, Hastie, Johnstone & Tibshirani (2004). The algorithm begins at $\hat{\eta}_0 = 0$. Suppose $\hat{\eta}$ is the current estimate and write $\hat{\mathbf{c}} = X^T(\mathbf{y} - \hat{\eta})$. Define the active set $\mathcal{A}$ as the set of the indices corresponding to the covariates with the largest absolute correlations,

$$\hat{C} = \max_j\{|\hat{c}_j|\} \text{ and } \mathcal{A} = \{j : |\hat{c}_j| = |\hat{C}|\}.$$

Define the active matrix corresponding to $\mathcal{A}$ as

$$X_{\mathcal{A}} = (s_j \mathbf{x}_j)_{j\in\mathcal{A}}, \text{ where } s_j = \text{sign}(\hat{c}_j).$$

Let

$$G_{\mathcal{A}} = X_{\mathcal{A}}^T X_{\mathcal{A}} \text{ and } A_{\mathcal{A}} = (\mathbf{1}_{\mathcal{A}}^T G_{\mathcal{A}}^{-1} \mathbf{1}_{\mathcal{A}})^{-1/2},$$

where $\mathbf{1}_{\mathcal{A}}$ is a vector of ones of length being $|\mathcal{A}|$, the size of $\mathcal{A}$. A unit equiangular vector with columns of the active set matrix $X_{\mathcal{A}}$ can be defined as

$$u_{\mathcal{A}} = X_{\mathcal{A}} w_{\mathcal{A}}, \quad \text{where } w_{\mathcal{A}} = A_{\mathcal{A}} G_{\mathcal{A}}^{-1} \mathbf{1}_{\mathcal{A}},$$

so that

$$X_{\mathcal{A}}^T u_{\mathcal{A}} = A_{\mathcal{A}} \mathbf{1}_{\mathcal{A}} \text{ and } ||u_{\mathcal{A}}||^2 = 1.$$

The next step of the Lars estimate gives the update

$$\hat{\eta} \leftarrow \hat{\eta} + \hat{\gamma} u_{\mathcal{A}},$$

where $\hat{\gamma}$ is the smallest positive number such that one and only one new index joins the active set $\mathcal{A}$. It can be shown that

$$\hat{\gamma} = \min^+_{j\in\mathcal{A}^C}\{\frac{\hat{C} - \hat{c}_j}{A_{\mathcal{A}} - a_j}, \frac{\hat{C} + \hat{c}_j}{A_{\mathcal{A}} + a_j}\},$$

where $\min^+$ means the minimum is taken over only positive components and a_j is the jth component of the vector $\mathbf{a} = X_{\mathcal{A}} u_{\mathcal{A}}$.

The Lasso, the FSW and the Lars all build a sequence of candidate models, from which the final model is chosen. In the Lasso, the sequence is controlled by t and

in the FSW, it is controlled by the number of steps (the step size in the procedure is taken to be a small constant arbitrarily close to zero). The Lars builds $(d+1)$ models with the number of variables ranging from 0 to d. Efron, Hastie, Johnstone & Tibshirani (2004) showed that there is a close relationship among these procedures in that they give almost identical solution paths. That is, if the candidate models are connected in each of these procedures, the resulting graphs are very similar. The solution path of the Lars is formed by connecting the $(d+1)$ models with linear segments. They noted that in the special case of orthogonal design matrix, the solution paths of the procedures are identical. In this case, Tibshirani (1996) showed that the Lasso solution has the form

$$\hat{\beta}_j = \text{sign}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)^+, \ j = 1, ..., d, \tag{2.3}$$

where $\gamma = \lambda/2$ for the λ in (2.2); and $(\pi)^+ = \pi, \ \pi > 0; \ 0, \ \pi \leq 0$. It coincides with the soft thresholding solution of Donoho & Johnstone (1994), where it is applied to wavelet coefficients.

In the implementation of the Lars, it is often the case that only the $(d+1)$ models at the end of the steps are considered as candidate models. The final model is chosen among the $(d+1)$ models, not the whole solution path. In this case the Lars is slightly different from the Lasso or the FSW, even in the orthogonal design matrix case. We will treat this case separately in this article.

Typical implementations of the Lasso, the Lars and the FSW attempt to find a model with the smallest prediction error among the sequence of candidate models built by these procedures. The prediction error is in terms of the squared loss (SL). For an estimate $\hat{\eta} = X\hat{\beta}$, the squared loss is

$$SL(\hat{\eta}) = (\hat{\eta} - \eta)^T(\hat{\eta} - \eta) = (\hat{\beta} - \beta)^T X^T X(\hat{\beta} - \beta).$$

In practice, since β is unknown, several methods, such as generalized cross validation (Craven & Wahba 1979), k-fold cross validation or Stein's unbiased estimate of risk (Stein 1981), can be used for the purpose of minimizing the squared error.

3 A simple example

In this section we give a simple example to demonstrate that the Lasso, the FSW and the Lars when tuned to minimize the squared error (as people usually attempt to do), miss the right model with a certain probability.

Consider a linear regression model with two predictors. Suppose that the true coefficient vector is $\beta^0 = (\beta_1^0, 0)^T$ with $\beta_1^0 > 0$, and the standardized design matrix X is

$$X^T X = \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

where $|\rho| < 1$. Therefore the model has one true component $\mathbf{x}_1$ and one noisy component $\mathbf{x}_2$, and the two covariates may be correlated. Denote the ordinary least squares solution by $\hat{\beta}^0$. The solution to the Lasso problem (2.2) when $d = 2$ can be easily seen as

$$\hat{\beta}_j = \text{sign}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)^+, \ j = 1, 2. \tag{3.1}$$

This formula holds even if the predictors are correlated, see Tibshirani (1996). Let $\hat{\delta} = (\hat{\delta}_1, \hat{\delta}_2)^T = \hat{\beta}^0 - \beta^0$. Since $\epsilon \sim N(0, \sigma^2 I_n)$ and $X^T X = \Sigma$, we have

$$\hat{\delta} \sim N(0, \sigma^2 \Sigma^{-1}). \tag{3.2}$$

Define

$$\mathcal{R} = \{(\delta_1, \delta_2)^T : \ \delta_1 \geq |\delta_2|\}.$$

We will show that the Lasso tuned with prediction accuracy selects the right model only for $\hat{\delta} \in \mathcal{R}$ and the probability of this event happening is 1/4.

It is clear that when $|\hat{\beta}_1^0| \leq |\hat{\beta}_2^0|$, the Lasso can not select the correct variables. When $\hat{\beta}_1^0 < -|\hat{\beta}_2^0|$, from (3.1) it is clear that the null model (the model with no predictor) is the best in terms of the squared loss, and therefore the Lasso tuned with prediction accuracy does not select the correct model. Now we only need to consider the case of $\hat{\beta}_1^0 - |\hat{\beta}_2^0| > 0$.

When the Lasso solution is tuned to correspond to the γ that minimizes the squared loss

$$SL(\gamma) = (\hat{\beta} - \beta^0)' \Sigma (\hat{\beta} - \beta^0) = (\hat{\beta}_1 - \beta_1^0, \hat{\beta}_2) \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 - \beta_1^0 \\ \hat{\beta}_2 \end{pmatrix}, \tag{3.3}$$

it chooses the correct model if and only if the minimizer of (3.3) lies in $[|\hat{\beta}_2^0|, \hat{\beta}_1^0)$. Denote $a = \hat{\beta}_1^0 - |\hat{\beta}_2^0| - \beta_1^0$. We consider two cases of $\hat{\beta}_1^0 - |\hat{\beta}_2^0| > 0$ separately.

(i) $\hat{\beta}_1^0 - |\hat{\beta}_2^0| > 0$, but $a < 0$. Write $s = \text{sign}(\hat{\beta}_2^0)$. By (3.1) and (3.3) we get

$$\begin{aligned} SL(\gamma) &= \left(a + |\hat{\beta}_2^0| - \gamma, s(|\hat{\beta}_2^0| - \gamma)\right)^T \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \left(a + |\hat{\beta}_2^0| - \gamma, s(|\hat{\beta}_2^0| - \gamma)\right) \\ &= a^2 + 2(1 + s\rho)[a(|\hat{\beta}_2^0| - \gamma) + (|\hat{\beta}_2^0| - \gamma)^2]. \end{aligned}$$

Since $1 + s\rho > 0$ and $a < 0$, we have that the minimum of the second term above over $\gamma \in [0, |\hat{\beta}_2^0|)$ is negative. Therefore the minimum of $SL(\gamma)$ is smaller than a^2. On the other hand, for any $\gamma \in [|\hat{\beta}_2^0|, \hat{\beta}_1^0)$, from (3.1) we get,

$$SL(\gamma) = (\hat{\beta}_1^0 - \gamma - \beta_1^0)^2 \geq a^2.$$

Hence the minimizer of $SL(\gamma)$ is not in $[|\hat{\beta}_2^0|, \hat{\beta}_1^0)$, and the Lasso tuned with prediction accuracy does not select the correct model.

(ii) $a \geq 0$. In this case it is easy to see that $\gamma = |\hat{\delta}_1|$ obtains the minimum of $SL(\gamma)$, and the Lasso selects the right model.

The above arguments show that the Lasso tuned with prediction accuracy selects the right model if and only if $\hat{\delta} \in R$. By (3.2), the probability associated with region R is 1/4.

Intuitively, the existence of the correlation makes model selection more difficult. However, the result states that the probability of the Lasso selecting the correct model is independent of correlation in this particular two dimensional example.

The argument above is valid for any finite sample size, and can also be applied to the following lemma when the design matrix is orthonormal.

Lemma 3.1. *When $\beta^0 = (\beta_1^0, 0, \cdots, 0)^T$ with $(d-1) > 0$ zero components and $X^T X = I_d$, the Lasso tuned with prediction accuracy selects the right model only when $\hat{\delta} = \hat{\beta}^0 - \beta^0 \in \mathcal{R}$, where*

$$\mathcal{R} = \left\{\delta : \ \delta_1 \beta_1^0 > 0, \ |\delta_1| \geq max\{|\delta_2|, \cdots, |\delta_d|\}\right\},$$

that is, the probability of the right model being selected is $1/(2d)$.

Proof. The Lasso solution has the form (2.3) when $X^TX = I$. Without loss of generality, assume $\beta_1^0 > 0$ and $|\hat{\beta}_2^0| > |\hat{\beta}_3^0| > \cdots > |\hat{\beta}_d^0|$. We will show for $\hat{\delta}$ not in $\mathcal{R}$, the Lasso tuned with prediction accuracy does not select the right model. Again, we only need to consider the situation where $\hat{\beta}_1^0 > |\hat{\beta}_2^0|$ in the following. The Lasso tuned with prediction accuracy selects the right model if and only if the minimizer of $SL(\gamma)$ is in $[|\hat{\beta}_2^0|, \hat{\beta}_1^0)$.

1. When $\hat{\delta}_1 \leq 0$, for any $\gamma \in [|\hat{\beta}_2^0|, \hat{\beta}_1^0)$, we have

$$\begin{aligned} SL(\gamma) &= (\hat{\delta}_1 - \gamma)^2 \\ &\geq (\hat{\delta}_1 - |\hat{\delta}_2|)^2 = (\hat{\delta}_1 - |\hat{\delta}_3| + |\hat{\delta}_3| - |\hat{\delta}_2|)^2 \\ &= (\hat{\delta}_1 - |\hat{\delta}_3|)^2 + (|\hat{\delta}_2| - |\hat{\delta}_3|)^2 + 2(\hat{\delta}_1 - |\hat{\delta}_3|)(|\hat{\delta}_3| - |\hat{\delta}_2|) \\ &> (\hat{\delta}_1 - |\hat{\delta}_3|)^2 + (|\hat{\delta}_2| - |\hat{\delta}_3|)^2 = SL(|\hat{\delta}_3|). \end{aligned}$$

   Since $|\hat{\delta}_3| \notin [|\hat{\beta}_2^0|, \hat{\beta}_1^0)$, the minimizer of $SL(\gamma)$ is not in $[|\hat{\beta}_2^0|, \hat{\beta}_1^0)$, and the right model is not selected.

2. For $0 < \hat{\delta}_1 < |\hat{\delta}_2|$, the minimum of $SL(\gamma)$ on the interval $[|\hat{\beta}_2^0|, \hat{\beta}_1^0)$ is obtained at $\gamma_1 = |\hat{\delta}_2|$ and $SL(\gamma_1) = (\hat{\delta}_1 - |\hat{\delta}_2|)^2$. However, when $|\hat{\delta}_3| < (\hat{\delta}_1 + |\hat{\delta}_2|)/2$, if we let $\gamma_2 = (\hat{\delta}_1 + |\hat{\delta}_2|)/2$, we have $SL(\gamma_2) = (\hat{\delta}_1 - |\hat{\delta}_2|)^2/2 < SL(\gamma_1)$. When $|\hat{\delta}_3| \geq (\hat{\delta}_1 + |\hat{\delta}_2|)/2$, if we let $\gamma_3 = |\hat{\delta}_3|$, we have $SL(\gamma_3) < SL(\gamma_1)$. Therefore the minimizer of $SL(\gamma)$ is outside $[|\hat{\beta}_2^0|, \hat{\beta}_1^0)$, and the right model is not selected.

Therefore, when $\hat{\delta} \in \mathcal{R}^C$, the Lasso tuned with prediction accuracy does not select the right model. For $\hat{\delta} \in \mathcal{R}$, the Lasso solution $\hat{\gamma} = \hat{\delta}_1$ yields the correct model $\eta(\hat{\gamma}) = \beta_1^0 \mathbf{x}_1$ with $SL(\hat{\gamma}) = 0$. Since $\hat{\delta} \sim N(0, \sigma^2 I_d)$, we have $Pr(\hat{\delta} \in \mathcal{R}) = 1/(2d)$. This completes the proof. □

Now we return to our two dimensional example considered at the beginning of this section. The solution path of the Lars and that of the Lasso are identical in this two dimensional problem. Therefore our results about the Lasso apply directly to the Lars if the final Lars estimate is chosen from the whole solution path of the Lars. In practical implementations of the Lars, however, the final solution is often chosen only among the models after each complete step. Thus we consider this

situation in the following. It is clear from the Lars algorithm that the Lars tuned with prediction accuracy does not yield the correct model when $\hat{\beta}_1^0 \leq 0$ or $|\hat{\beta}_1^0| \leq |\hat{\beta}_2^0|$. In the situation where $\hat{\beta}_1^0 > |\hat{\beta}_2^0|$, the three step Lars estimates can be written as $\hat{\eta}_0 = 0,\ \hat{\eta}_1 = (\hat{\beta}_1^0 - |\hat{\beta}_2^0|)\mathbf{x}_1,\ \hat{\eta}_2 = \hat{\beta}_1^0\mathbf{x}_1 + \hat{\beta}_2^0\mathbf{x}_2$. The corresponding square losses are

$$SL(\hat{\eta}_0) = (\beta_1^0)^2,\ SL(\hat{\eta}_1) = (\hat{\delta}_1 - |\hat{\delta}_2|)^2, \text{ and } SL(\hat{\eta}_2) = \hat{\delta}_1^2 + \hat{\delta}_2^2.$$

We immediately see $SL(\hat{\eta}_1) > SL(\hat{\eta}_2)$ when $\hat{\delta}_1 < 0$. Putting these together, we get that the Lars tuned with prediction accuracy does not select the right model when the OLS estimate satisfies $\hat{\beta}_1^0 < \beta_1^0$, which happens with probability 1/2 by (3.2). The overall probability that the Lars tuned with prediction accuracy selects the right model is no larger than 1/2.

4 Higher dimensional problems with orthogonal designs

Theorem 4.1. *When the true coefficient vector is $\beta^0 = (\alpha_1, ..., \alpha_{d_1}, 0, ..., 0)^T$ with $d_2 = (d - d_1) > 0$ zero coefficients and $X^T X = I_d$, if the Lasso is tuned according to prediction accuracy, then it selects the right model with a probability less than a constant $C < 1$, where C depends only on σ^2 and d_1, and not on the sample size n.*

Proof. Let the OLS estimate be $\hat{\beta}^0$ and denote $\hat{\beta}^0 - \beta^0 = (\hat{\delta}_1, ..., \hat{\delta}_d)^T$. Without loss of generality we assume $|\hat{\delta}_{d_1+1}| > |\hat{\delta}_{d_1+2}| > ... > |\hat{\delta}_d|$ and $\alpha_i > 0,\ i = 1, ..., d_1$. We will show for the region

$$\mathcal{R} = \{(\delta_1, ..., \delta_d)^T :\ \delta_j > -\alpha_j,\ j = 1, ..., d_1 \text{ and } \sum_{i=1}^{d_1} \delta_i < 0\},$$

the Lasso tuned with prediction accuracy does not select the right model.

If $\hat{\beta}^0$ does not satisfy

$$\left\{ |\hat{\beta}_j^0| > |\hat{\beta}_k^0|,\ \text{for } j \in \{1, ..., d_1\} \text{ and } k \in \{d_1 + 1, .., d\} \right\}, \tag{4.1}$$

obviously the Lasso does not select the right model. So we can concentrate on the situation where (4.1) is satisfied. For the Lasso tuned with prediction accuracy to

select the right model, the solution must satisfy

$$\min\{|\hat{\beta}_1^0|, ..., |\hat{\beta}_{d_1}^0|\} > \gamma \geq |\hat{\beta}_{d_1+1}^0|. \tag{4.2}$$

Since $\hat{\delta} \in \mathcal{R}$, we have $\hat{\beta}_j^0 > 0,\ j = 1, ..., d_1$. The estimate corresponding to any γ satisfying (4.2) is

$$\begin{aligned}\eta(\gamma) &= (\hat{\beta}_1^0 - \gamma)\mathbf{x}_1 + ... + (\hat{\beta}_{d_1}^0 - \gamma)\mathbf{x}_{d_1} \\ &= (\alpha_1 + \hat{\delta}_1 - \gamma)\mathbf{x}_1 + ... + (\alpha_{d_1} + \hat{\delta}_{d_1} - \gamma)\mathbf{x}_{d_1}.\end{aligned}$$

On the other hand, the estimate with $\gamma_1 = |\hat{\beta}_{d_1+2}^0|$ has the form

$$\begin{aligned}\eta(\gamma_1) =&(\hat{\beta}_1^0 - |\hat{\beta}_{d_1+2}^0|)\mathbf{x}_1 + \cdots + (\hat{\beta}_{d_1}^0 - |\hat{\beta}_{d_1+2}^0|)\mathbf{x}_{d_1} + \text{sign}(\hat{\beta}_{d_1+1}^0)(|\hat{\beta}_{d_1+1}^0| - |\hat{\beta}_{d_1+2}^0|)\mathbf{x}_{d_1+1} \\ =&(\alpha_1 + \hat{\delta}_1 - |\hat{\delta}_{d_1+2}|)\mathbf{x}_1 + \cdots + (\alpha_{d_1} + \hat{\delta}_{d_1} - |\hat{\delta}_{d_1+2}|)\mathbf{x}_{d_1} \\ &+ \text{sign}(\hat{\delta}_{d_1+1})(|\hat{\delta}_{d_1+1}| - |\hat{\delta}_{d_1+2}|)\mathbf{x}_{d_1+1}.\end{aligned}$$

It is easy to see the squared losses for the two estimates are

$$\begin{aligned}SL(\gamma) &= \sum_{i=1}^{d_1}(\hat{\delta}_i - \gamma)^2; \\ SL(\gamma_1) &= \sum_{i=1}^{d_1}(\hat{\delta}_i - |\hat{\delta}_{d_1+2}|)^2 + (|\hat{\delta}_{d_1+1}| - |\hat{\delta}_{d_1+2}|)^2.\end{aligned}$$

We show for any γ satisfying (4.2), $SL(\gamma) > SL(\gamma_1)$. Simple algebra yields

$$\begin{aligned}SL(\gamma) = \sum_{i=1}^{d_1}(\hat{\delta}_i - \gamma)^2 &= \sum_{i=1}^{d_1}(\hat{\delta}_i - |\hat{\delta}_{d_1+2}| + |\hat{\delta}_{d_1+2}| - \gamma)^2 \\ =&\sum_{i=1}^{d_1}(\hat{\delta}_i - |\hat{\delta}_{d_1+2}|)^2 + d_1(\gamma - |\hat{\delta}_{d_1+2}|)^2 + 2(\gamma - |\hat{\delta}_{d_1+2}|)\sum_{i=1}^{d_1}(|\hat{\delta}_{d_1+2}| - \hat{\delta}_i) \\ =&SL(\gamma_1) - (|\hat{\delta}_{d_1+1}| - |\hat{\delta}_{d_1+2}|)^2 \\ &+ d_1(\gamma - |\hat{\delta}_{d_1+2}|)^2 + 2(\gamma - |\hat{\delta}_{d_1+2}|)\sum_{i=1}^{d_1}(|\hat{\delta}_{d_1+2}| - \hat{\delta}_i).\end{aligned}$$

Since $\gamma \geq |\hat{\delta}_{d_1+1}|$, we have

$$d_1(\gamma - |\hat{\delta}_{d_1+2}|)^2 - (|\hat{\delta}_{d_1+1}| - |\hat{\delta}_{d_1+2}|)^2 \geq (d_1 - 1)(\gamma - |\hat{\delta}_{d_1+2}|)^2.$$

It follows

$$SL(\gamma) \geq SL(\gamma_1) + (d_1 - 1)(\gamma - |\hat{\delta}_{d_1+2}|)^2 + 2(\gamma - |\hat{\delta}_{d_1+2}|) \sum_{i=1}^{d_1} (|\hat{\delta}_{d_1+2}| - \hat{\delta}_i).$$

It is easy to see when $\sum_{i=1}^{d_1} \hat{\delta}_i < 0$, the following satisfies

$$(d_1 - 1)(\gamma - |\hat{\delta}_{d_1+2}|) + 2\sum_{i=1}^{d_1} (|\hat{\delta}_{d_1+2}| - \hat{\delta}_i) = (d_1 + 1)|\hat{\delta}_{d_1+2}| + (d_1 - 1)\gamma - 2\sum_{i=1}^{d_1} \hat{\delta}_i > 0.$$

Therefore, we have $SL(\gamma) > SL(\gamma_1)$ when $\hat{\delta} \in \mathcal{R}$. The optimal γ that minimizes $SL(\gamma)$ does not satisfy (4.2), that is, the optimal γ does not yield the correct model. Since $(\hat{\delta}_1, \cdots, \hat{\delta}_d)^T$ follows a multivariate normal distribution $N(0, I_d)$, it is readily seen

$$Pr(\hat{\delta} \in \mathcal{R}) > Pr(\hat{\delta} \in \{(\delta : 0 > \delta_j > -\alpha_j, j = 1, ..., d_1\}) = C,$$

where C is a constant strictly less than 1 depending on σ^2 and d_1 but not on the sample size n. We have proved that with a positive probability not depending on n, the Lasso tuned with prediction accuracy does not select the right model. □

The conclusion holds for the Lars and the FSW due to the equivalence of the three procedures, if the whole solution path of the Lars is considered. When only $(d + 1)$ candidate models in the Lars are considered, the conclusion follows by replacing γ by $|\hat{\delta}_{d_1+1}|$ in the preceding proof. When the design matrix satisfies $X^T X = nI_d$, following the same argument in theorem 4.1, we can prove that when the Lasso is tuned according to prediction accuracy, the probability of the Lasso selecting the wrong model is larger than a strictly positive constant not depending on n.

Although the conclusion of the theorem is proved with the design matrix being orthonormal, it is expected to hold for general design matrix cases, as is the case for $d = 2$. We demonstrate this point via simulations in the next section.

5 Simulation

We conduct some simple simulations in general design matrix cases to demonstrate our results. All simulations were conducted using MATLAB code. We used the

algorithm as suggested in Tibshirani (1996). Each β_j is rewritten as $\beta_j^+ - \beta_j^-$, where β_j^+ and β_j^- are nonnegative. We then used the quadratic programming module quadprog in MATLAB to find the Lasso solution.

We generate data from two models which have the form

$$\mathbf{y} = X\beta + \epsilon,$$

where

$$\begin{aligned} &\text{Model 1}: \ \beta = (1,0)^T, \\ &\text{Model 2}: \ \beta = (3,1.5,0,0)^T; \end{aligned}$$

ϵ follows standard normal distribution and $\mathbf{x}_j$ has marginal distribution $N(0,1)$. The pairwise correlation between $\mathbf{x}_i$ and $\mathbf{x}_j$, $i \neq j$, is $\rho^{|i-j|}$ with $\rho = 0,\ 0.5,\ 0.9$. We simulate data with sample size $n = 40,\ 400,\ 4000$. For each ρ and each sample size, we simulate 100 data sets and apply the Lasso method. The tuning parameter is chosen to minimize the prediction accuracy. We summarize the result for various sample sizes and correlations in Table 5.1. The percentage of correctly selected models is summarized in the PCM column. We divide the SL of the Lasso by the SL of corresponding OLS estimate and report the mean of these relative squared losses in the MRSL column. When it is tuned according to the prediction accuracy, the Lasso can achieve much improved accuracy, but misses the right model a large fraction of the time. This seems to be true independent of the sample size and the correlation among the covariates. The results of the experiment are consistent with our analytical results above.

6 Discussions

We wish to make it extra clear that the results in this paper should not be interpreted as a criticism of the Lasso and related methods or taken to imply that the Lasso and related methods can not be used as variable selection tools. In most practical applications, prediction accuracy is the golden standard. As shown by many authors and also in our simulations, the Lasso can improve greatly over the ordinary

n	ρ	Model 1 PCM (%)	MRSL	Model 2 PCM (%)	MRSL
	0	26	0.464	15	0.632
40	0.5	16	0.573	20	0.573
	0.9	22	0.487	16	0.569
	0	27	0.482	9	0.655
400	0.5	23	0.513	15	0.594
	0.9	25	0.546	13	0.583
	0	22	0.492	15	0.667
4000	0.5	21	0.500	18	0.597
	0.9	24	0.523	20	0.539

Table 5.1: Simulation results for the Lasso.

least estimate in terms of accuracy. In problems where the primary goal is selecting the set of true variables, our results simply imply that prediction based methods for tuning are not sufficient for the Lasso and related methods. It is possible that some other criteria of choosing the tuning parameter can yield consistent variable selection for the Lasso and related methods, and there are also simple adjustments to the solution of the Lasso and related methods that make the estimators consistent for variable selection. For example, Meinshausen & Bühlmann (2004) proposed doing neighborhood selection with the Lasso in a multivariate Gaussian graphical model. They allow the number of variables to grow (even rapidly) with the number of observations. They noted that the optimal tuning parameter for prediction accuracy does not lead to a consistent neighborhood estimate by giving an example in which the number of variables grows to infinity and in which the tuning parameter chosen according to prediction accuracy leads to the wrong model with probability tending to one. They proposed to choose the tuning parameter by controlling the probability of falsely joining some distinct connectivity components of the graph instead of minimizing prediction accuracy, and showed that their approach delivers consistent model selection. In fact, they showed that if the tuning parameter λ has a rate $n^{-1/2+\epsilon}$ with some $\epsilon > 0$, then the Lasso can achieve consistent model selection. Another possibility in our setup is to use a λ that achieves $\sqrt{n}$ consistency in terms

of estimating the coefficients (Knight and Fu, 2000), and then threshold the Lasso solution by a quantity of the order n^{α} with $-1/2 < \alpha < 0$. Since the coefficient estimates are $\sqrt{n}$ consistent, it is easy to see that with probability tending to one, the estimates of coefficients of important variables will not be thresholded to zero, whereas the estimates of the coefficients of redundant variables will be thresholded to zero. Thus consistent variable selection can be achieved. A third possibility arises upon considering the intuitively clear argument that tuning for prediction rather than variable selection tends to create non-zero coefficients when the true coefficient is zero, but not the other way around. Simulation results (not shown) support this argument, and this is also noted in Meinshausen & Bühlmann (2004). Thus, this possibility entails tuning for prediction and then creating a test for estimating when a non-zero coefficient is 'small enough' to be set to 0 that is compatible with this tuning. A similar approach was taken in a somewhat different context in Zhang, Wahba, Lin, Voelker, Ferris, Klein & Klein (2004).

ACKNOWLEDGEMENTS

The authors thank the editor, the associate editor, and two anonymous referees for constructive comments. We acknowledge the suggestion from one of the referees, which extends our original exposition for orthonormal covariates to correlated covariates when the dimension is two.

References

Broman, K. W. & Speed, T. P. (2002), 'A model selection approach for the identification of quantitative trait loci in experimental crosses', *Journal of the Royal Statistical Society, Series B* **64**(4), 641–656.

Craven, P. & Wahba, G. (1979), 'Smoothing noisy data with spline functions', *Numerische Mathematik* **31**, 377–403.

Donoho, D. L. & Johnstone, I. M. (1994), 'Ideal spatial adaptation by wavelet shrinkage', *Biometrika* **81**, 425–455.

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004), 'Least angle regression (Disc: P452-499)', *The Annals of Statistics* **32**, 407–451.

Fan, J. & Li, R. (2001), 'Variable selection via nonconcave penalized likelihood and its oracle properties', *Journal of the American Statistical Association* **96**(456), 1348–1360.

Knight, K. & Fu, W. (2000), 'Asymptotics for lasso-type estimators', *The Annals of Statistics* **28**(5), 1356–1378.

Meinshausen, N. & Bühlmann, P. (2004), Consistent neighbourhood selection for high-dimensional graphs with the lasso, Technical Report 123, Seminar für Statistik, ETH, Zürich.

Osborne, M. R., Presnell, B. & Turlach, B. A. (2000), 'A new approach to variable selection in least squares problems', *IMA Journal of Numerical Analysis* **20**(3), 389–404.

Rao, C. R. & Wu, Y. (1989), 'A strongly consistent procedure for model selection in a regression problem', *Biometrika* **76**, 369–374.

Shao, J. (1997), 'An asymptotic theory for linear model selection (Disc: P243-264)', *Statistica Sinica* **7**, 221–242.

Stein, C. M. (1981), 'Estimation of the mean of a multivariate normal distribution', *The Annals of Statistics* **9**, 1135–1151.

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

Tibshirani, R. (1997), 'The lasso method for variable selection in the Cox model', *Statistics in Medicine* **16**, 385–395.

Zhang, H., Wahba, G., Lin, Y., Voelker, M., Ferris, M., Klein, R. & Klein, B. (2004), 'Variable selection and model building via likelihood basis pursuit', *Journal of the American Statistical Association* **99**, 659–672.