A Note on the LASSO and Related Procedures in Model Selection

Chenlei Leng, Yi Lin and Grace Wahba

chenlei,yilin,wahba@stat.wisc.edu

http://www.stat.wisc.edu/~chenlei, ~yilin, ~wahba

1Chenlei Leng is graduate student, Department of Statistics, University of Wisconsin, Madison, WI 53706 (E-mail: chenlei@stat.wisc.edu). 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 and Bascom 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. Lin’s research was supported in part by NSF Grant DMS 0134987. Wahba’s research was supported in part by NSF Grant DMS 0072292.
A Note on the LASSO and Related Procedures in Model Selection

Chenlei Leng, Yi Lin and Grace Wahba
Department of Statistics
University of Wisconsin
Madison, WI 53706-1685
chenlei, yilin, wahba@stat.wisc.edu

April 15, 2004

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. We show, however, that the dual goal of accurate estimation and consistent variable selection can not be achieved simultaneously: when the tuning parameter is chosen to minimize the prediction error, as is commonly done in practice, in general these procedures are not consistent in terms of variable selection. That is, the sets of variable selected by the procedures 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 $n$.

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 the 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

$$y = X\beta + \epsilon,$$

where $y = (y_1, ..., y_n)^T$ are the responses, $\beta = (\beta_1, ..., \beta_d)^T$ are the regression coefficients, $X = (x_1, ..., x_d)$ 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,

$$1^T y = 0, \quad 1^T x_j = 0, \quad \text{and} \quad x_j^T x_j = 1 \text{ for } j = 1, ..., d.$$

In many practical situations, some covariates are superfluous. That is, 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) and Shao (1997) for some earlier studies on the consistent variable selection problem.

It is of interest to investigate whether the Lasso and related methods are consistent 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. This implies the inconsistency for model selection but is actually much stronger. It is a finite sample result, since it is true for any sample size n.

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 give a simple example to illustrate the ideas and demonstrate that the three methods fail to find the right model with certain probability. The general results are given in section 4. We present some simulation results in section 5 and a summary is given in section 6.

2 The Lasso, the Lars and the Forward Stagewise regression

The Lasso estimate is the solution to

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

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 estimation accuracy. Equivalently, the Lasso estimate can be obtained as the solution to the penalized likelihood problem

$$\min_{\beta} \frac{1}{n}(y - X\beta)^T(y - X\beta) + \lambda \sum_{j=1}^{d} |\beta_j|, \quad (2.1)$$

where there is a one to one correspondence between $t$ and $\lambda$.

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
\[ \eta = X\beta, \text{ and beginning with } \hat{\eta}_0 = 0, \text{ if } \hat{\eta} \text{ is the current estimate, the next step is taken in the direction of the greatest correlation between covariate } x_j \text{ and the current residual. That is, writing } \hat{c} = X^T(y - \hat{\eta}) \text{ and } \hat{j} = \text{argmax}(|\hat{c}_j|), \text{ the update is} \\
\hat{\eta} \leftarrow \hat{\eta} + \epsilon \cdot \text{sign}(\hat{c}_j) \cdot x_j, \]

where \( \epsilon > 0 \) is some small constant. It is readily seen that \( \epsilon = |\hat{c}_j| \) yields the familiar standard forward selection. Smaller \( \epsilon \) 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 \( \hat{c} = X^T(y - \hat{\eta}) \). Define the active set \( \mathcal{A} \) as the set of the indices corresponding to the covariates with the largest absolute correlations,

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

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

\[ X_\mathcal{A} = (s_j x_j)_{j \in \mathcal{A}}, \text{ where } s_j = \text{sign}(\hat{c}_j). \]

Let

\[ G_\mathcal{A} = X_\mathcal{A}^T X_\mathcal{A} \quad \text{and} \quad A_\mathcal{A} = (1_{|\mathcal{A}|} G_\mathcal{A}^{-1} 1_{\mathcal{A}})^{-1/2}, \]

where \( 1_{\mathcal{A}} \) 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

\[ u_\mathcal{A} = X_\mathcal{A} w_\mathcal{A}, \text{ where } w_\mathcal{A} = A_\mathcal{A} G_\mathcal{A}^{-1} 1_{\mathcal{A}}, \]

so that

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

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

\[ \hat{\gamma} = \min_{j \in \mathcal{A}^c} \left\{ \frac{\hat{C} - \hat{c}_j}{A_\mathcal{A} - a_j}, \frac{\hat{C} + \hat{c}_j}{A_\mathcal{A} + a_j} \right\}, \]

where \( \mathcal{A}^c \) denotes the complement of \( \mathcal{A}. \)
where \( \min^+ \) means the minimum is taken over only positive components and \( a_j \) is the \( j \)th component of the vector \( a = X_{A \cup A} \).

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. Therefore we concentrate on the Lasso in the following, and all the results apply to the Lars and the FSW as well. In the orthogonal design matrix case, the Lasso solution has the form

\[
\hat{\beta}_j = \text{sign}(\beta_j^0)(|\beta_j^0| - \gamma)^+, \quad j = 1, \ldots, d, \tag{2.2}
\]

where \( \gamma = \lambda/2 \) for the \( \lambda \) in (2.1); 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.

The implementation of the Lasso, the Lars and the FSW attempts to find a model with the smallest estimation error among the sequence of candidate models built by these procedures. The estimation error is typically 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 \( \beta \) 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 the true coefficient vector is \( \beta^0 = (\beta_1^0, 0)^T \) with \( \beta_1^0 > 0 \), and the design matrix \( X \) is orthonormal. Therefore the model has one true component \( x_1 \) and one noisy component \( x_2 \), and \( X^TX = I_2 \). Denote the ordinary least squares solution by \( \hat{\beta}^0 \). In this case the solution to the Lasso problem (2.1) is

\[
\hat{\beta}_j = \text{sign}(\beta_j^0)(|\beta_j^0| - \gamma)^+, \quad j = 1, 2.
\]  

(3.1)

Figure 3.1 shows the Lasso estimate versus the OLS estimate. The Lasso estimate is shifted towards zero by some constant. For completeness, the estimate by subset selection versus the OLS estimate is included in the figure.

Figure 3.1: In the left plot, the red solid line indicates the Lasso estimate versus the OLS estimate; the right plot shows the subset estimate against the OLS estimate. For comparison, the 45 degree lines are drawn.

Let \( \hat{\delta} = (\hat{\delta}_1, \hat{\delta}_2)^T = \hat{\beta}^0 - \beta^0 \). Since \( \epsilon \) is a normal variate and \( X^TX = I_2 \), we have

\[
\hat{\delta} \sim N(0, \sigma^2 I_2),
\]

(3.2)

where \( \sigma^2 \) is the noise variance.
Define
\[ R_1 = \{(\delta_1, \delta_2)^T : \delta_1 < 0\}, \]
\[ R_2 = \{(\delta_1, \delta_2)^T : \delta_1 > 0, \delta_1 < |\delta_2|\}, \]
\[ R_3 = \{(\delta_1, \delta_2)^T : \delta_1 > 0, \delta_1 > |\delta_2|\}. \]
We will show that when \( \hat{\delta} \in R_1 \) or \( \hat{\delta} \in R_2 \), the Lasso does not select the right model. The only region where the Lasso selects the right model is \( R_3 \). Thus from (3.2), the probability of the Lasso selecting the correct model is 1/4.

It is clear that when \( |\hat{\beta}^0_1| \leq |\hat{\beta}^0_2| \), the Lasso can not select the correct variables by (3.1). As a result, we only need to consider the situation where \( |\hat{\beta}^0_1| > |\hat{\beta}^0_2| \) in the following.

For \( \hat{\delta} \in R_1 \), we consider the situations \( \hat{\beta}^0_1 \leq 0 \) and \( \hat{\beta}^0_1 > 0 \) separately. For \( \hat{\beta}^0_1 \leq 0 \), when \( |\hat{\beta}^0_1| > |\hat{\beta}^0_2| \), it is easy to see that a naive estimate \( \hat{\eta} = 0 \) with \( \hat{\gamma} = |\hat{\beta}^0_1| \) yields the Lasso estimate with the smallest squared loss.

For \( \hat{\beta}^0_1 > 0 \), the Lasso solution minimizes
\[ SL(\gamma) = (\hat{\beta}_1 - \beta_1^0)^2 + (\hat{\beta}_2 - \beta_2^0)^2 = (\beta_1^0 + \hat{\delta}_1 - \gamma)^+ - \beta_1^0)^2 + [(|\hat{\delta}_2| - \gamma)^+]^2. \]
For \( \gamma \in [|\hat{\beta}_1^0|, \hat{\beta}_1^0] \), \( SL(\gamma) = (\hat{\delta}_1 - \gamma)^2 \). Since \( \hat{\delta}_1 < 0 \), we have
\[ SL(\gamma) = (\hat{\delta}_1 - \gamma)^2 > \hat{\delta}_1^2 + \hat{\delta}_2^2 = SL(0), \]
where \( SL(0) \) is the \( SL \) of the OLS estimate. Therefore, the optimal \( \gamma \) that minimizes \( SL(\gamma) \) is not in the interval \( [|\hat{\beta}_1^0|, \hat{\beta}_1^0] \). We then see from (3.1) that the optimal \( \gamma \) does not yield the correct model. The claim is proved for \( \hat{\delta} \in R_1 \).

For \( \hat{\delta} \in R_2 \) and \( \gamma \in [|\hat{\beta}_2^0|, \hat{\beta}_2^0] \), \( SL(\gamma) = (\hat{\delta}_1 - \gamma)^2 \). Since \( \hat{\delta}_1 < |\hat{\delta}_2| \), the minimum is obtained at \( \gamma_1 = |\hat{\delta}_2| \) on this interval and \( SL(\gamma_1) = (\hat{\delta}_1 - |\hat{\delta}_2|)^2 \). However, when \( \gamma_2 = (\hat{\delta}_1 + |\hat{\delta}_2|)/2 < |\hat{\delta}_2| \),
\[ SL(\gamma_2) = (\hat{\delta}_1 - |\hat{\delta}_2|)^2/2 < SL(\gamma_1). \]
The estimated coefficients corresponding to \( \gamma_2 \) are
\[ \hat{\beta}_1 = \hat{\beta}_1^0 + (\hat{\delta}_1 - |\hat{\delta}_2|)/2 \quad \text{and} \quad \hat{\beta}_2 = \text{sign}(\hat{\delta}_2)(|\hat{\delta}_2| - \hat{\delta}_1)/2 
eq 0. \]
Again, the optimal \( \gamma \) that minimizes \( SL(\gamma) \) is not in the interval \( [|\hat{\beta}_1^0|, \hat{\beta}_1^0] \). Therefore, the Lasso does not select the right model for \( \hat{\delta} \in R_2 \) either.

The Lasso, however, selects the right model when \( \hat{\delta} \in R_3 \). It is easy to see that \( SL(\hat{\gamma}) = 0 \) and \( \eta(\hat{\gamma}) = \beta_1^0 x_1 \) for the Lasso solution \( \hat{\gamma} = \hat{\delta}_1 \). Therefore, we have shown that the Lasso selects the right model only when \( \hat{\delta} \in R_3 \). The probability associated with \( R_3 \) is 1/4.
The argument above is valid for any finite sample size, and shows that with probability 3/4, the Lasso with the tuning parameter selected to minimize the estimation error does not select the correct model. The argument can be generalized to the following lemma.

**Lemma 3.1.** When \( \beta^0 = (\beta^0_1, 0, \cdots, 0)^T \) with \((d-1) > 0\) zero components and \(X^TX = I_d\), the Lasso selects the right model only when \( \delta = \beta^0 - \beta^0 \in \mathcal{R} \), where

\[
\mathcal{R} = \{ \delta : \delta_1 \delta^0_1 > 0, \ |\delta_1| \geq \max(|\delta_2|, \cdots, |\delta_d|) \},
\]
that is, the probability of the Lasso selecting the right model is \( 1/(2d) \).

**Proof.** The Lasso solution has the form (2.2) when \( X^TX = I \). Without loss of generality, assume \( \beta^0_1 > 0 \) and \( |\beta^0_2| > |\beta^0_3| > \cdots > |\beta^0_d| \). We will show for \( \delta \) not in \( \mathcal{R} \), the Lasso does not select the right model. It is clear that when \( |\beta^0_1| \leq |\beta^0_2| \), the Lasso cannot select the correct variables by (2.2). Therefore, we concentrate on the situation where \( |\beta^0_1| > |\beta^0_2| \) in the following.

1. For \( \hat{\delta}_1 \leq 0 \) and \( \hat{\beta}^0 \leq 0 \), a naive estimate \( \hat{\eta} = 0 \) yields the Lasso estimate.

2. For \( \hat{\delta}_1 \leq 0 \) and \( \hat{\beta}^0 > 0 \), for the Lasso to select the right model, \( \gamma \) must satisfy \( \gamma \in [|\hat{\beta}^0_2|, \hat{\beta}^0) \) and thus \( SL(\gamma) = (\hat{\delta}_1 - \gamma)^2 \). It is easy to see that the minimum is obtained at \( \gamma_1 = |\hat{\delta}_2| \) and \( SL(\gamma_1) = (\hat{\delta}_1 - |\hat{\delta}_2|)^2 \). But when \( \gamma_2 = |\hat{\delta}_3| \),

\[
SL(\gamma_1) = (\hat{\delta}_1 - |\hat{\delta}_2|)^2 = (\hat{\delta}_1 - |\hat{\delta}_3| + |\hat{\delta}_3| - |\hat{\delta}_2|)^2
\]
\[
= (\hat{\delta}_1 - |\hat{\delta}_3|)^2 + (|\hat{\delta}_2| - |\hat{\delta}_3|)^2 + 2(\hat{\delta}_1 - |\hat{\delta}_3|)(|\hat{\delta}_3| - |\hat{\delta}_2|)
\]
\[
>(\hat{\delta}_1 - |\hat{\delta}_3|)^2 + (|\hat{\delta}_2| - |\hat{\delta}_3|)^2 = SL(\gamma_2).
\]

The estimated model corresponding to \( \gamma_2 \) is

\[
\eta(\gamma_2) = (\hat{\beta}^0_1 - |\hat{\beta}^0_2|)x_1 + \text{sign}(\hat{\beta}^0_2)(|\hat{\beta}^0_2| - |\hat{\beta}^0_3|)x_2,
\]

which is not the right model.

3. For \( 0 < \hat{\delta}_1 < |\hat{\delta}_2| \), the \( \gamma \) which minimizes \( SL(\gamma) \) on the interval \([|\hat{\beta}^0_2|, \hat{\beta}^0) \) 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 \( SL(\gamma_3) < SL(\gamma_1) \). The estimated models corresponding to \( \gamma_2 \) and \( \gamma_3 \) both include \( x_1 \) and \( x_2 \).

Therefore, when \( \delta \in \mathcal{R}^C \), the Lasso selects a wrong model. For \( \delta \in \mathcal{R} \), the Lasso solution \( \hat{\gamma} = \hat{\delta}_1 \) yields the correct model \( \eta(\hat{\gamma}) = \beta^0_1 x_1 \) 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. \( \square \)
Now we return to our two dimensional example considered at the beginning of this section. Figure 3.2 is a schematic sketch of the Lars algorithm in this situation. The ordinary least squares estimate is shown in the figure as point B. The initial Lars estimate is simply \( \tilde{\eta}_0 = 0 \), corresponding to point A. The Lars estimate after step one is shown in the figure as point D, which has the property that the angle between the line DB and the axes is 45 degrees. The step two estimate is simply the ordinary least squares estimate corresponding to point B in the figure. If we consider the whole path of solutions (line segments AD and DB), and choose the estimate along the path with the smallest squared error, in our example with orthonormal design matrix, the Lars is exactly equivalent to the Lasso and the above results for the Lasso applies to the Lars directly. In practical implementation of the Lars, however, the final solution is often chosen only among the models after each complete step, that is, points A, D, and B in the figure, while in subset selection, the final solution is chosen among points A, C and B, where point C is the projection of point B to \( x_1 \). Thus we consider this situation in the following, and study the probability of choosing the correct model (point D in this example) when the squared error is used as the criterion. In the following we show that when \( \hat{\delta} \in \mathcal{R}_1 \), which has probability 1/2, the Lars does not select the correct model. It is clear from the Lars algorithm that the Lars does not yield the correct model when \( \hat{\beta}^0_1 \leq 0 \) or \( |\hat{\beta}^0_1| \leq |\hat{\beta}^0_2| \). We only need to consider the situation when \( \hat{\beta}^0_1 > 0 \) and \( |\hat{\beta}^0_1| \geq |\hat{\beta}^0_2| \). In this case, the Lars estimate can be written as \( \hat{\eta}_0 = 0, \hat{\eta}_1 = (\beta^0_1 + \delta_1 - |\delta_2|) x_1 \) and \( \hat{\eta}_2 = (\beta^0_1 + \delta_1) x_1 + \delta_2 x_2 \). It follows

\[
SL(\hat{\eta}_0) = (\beta^0_1)^2, \quad SL(\hat{\eta}_1) = (\delta_1 - |\delta_2|)^2, \quad \text{and} \quad SL(\hat{\eta}_2) = \delta_1^2 + \delta_2^2.
\]

We immediately see \( SL(\hat{\eta}_1) > SL(\hat{\eta}_2) \) when \( \delta_1 < 0 \). Therefore, we have shown that the Lars does not select the right model when the OLS estimate satisfies \( \hat{\beta}^0_1 < \beta^0_1 \). The probability of this region is \( \Pr(\mathcal{R}) = 1/2 \) since \( \hat{\delta} \) is a normal variate. The overall probability that the Lars selects the right model is no larger than 1/2.

We prove a more general theorem in the following section.

4 More general situations

**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 \( X^T X = I_d \), we have

\[
\Pr(\text{the Lasso selects the right model}) \leq C,
\]

with respect to any sample size, where \( C < 1 \) is a constant depending only on \( \sigma^2 \) and \( d_1 \).
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 $|\hat{\delta}_{d_{i+1}}| > |\hat{\delta}_{d_{i+2}}| > ... > |\hat{\delta}_d|$ 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 does not select the right model.

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

$$
\left\{ |\hat{\beta}_{j}^0| > |\hat{\beta}_{k}^0|, \text{ for } j \in \{1, ..., d_1\} \text{ and } k \in \{d_1 + 1, ..., d\}\right\},
$$
(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 to select the right model, the solution must satisfy

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

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

$$
\eta(\gamma) = (\hat{\beta}_1^0 - \gamma)x_1 + ... + (\hat{\beta}_{d_1}^0 - \gamma)x_{d_1}
= (\alpha_1 + \hat{\delta}_1 - \gamma)x_1 + ... + (\alpha_{d_1} + \delta_{d_1} - \gamma)x_{d_1}.
$$
On the other hand, the estimate with $\gamma = |\hat{\beta}_{d_i+2}^0|$ has the form

$$
\eta(\gamma) = (\beta_1^0 - |\beta_{d_i+2}^0|)x_1 + \cdots + (\beta_{d_i}^0 - |\beta_{d_i+2}^0|)x_{d_i} + \text{sgn}(\beta_{d_i+1}^0)(|\beta_{d_i+1}^0| - |\beta_{d_i+2}^0|)x_{d_i+1}
$$

$$
= (\alpha_1 + \hat{\delta}_1 - |\hat{\delta}_{d_i+2}|)x_1 + \cdots + (\alpha_{d_i} + \hat{\delta}_d - |\hat{\delta}_{d_i+2}|)x_{d_i} + \text{sgn}(\hat{\delta}_{d_i+1})(|\hat{\delta}_{d_i+1}| - |\hat{\delta}_{d_i+2}|)x_{d_i+1}.
$$

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

$$
SL(\gamma) = \sum_{i=1}^{d_i} (\hat{\delta}_i - \gamma)^2;
$$

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

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

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

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

$$
= SL(\gamma_1) - (|\hat{\delta}_{d_i+1}| - |\hat{\delta}_{d_i+2}|)^2
$$

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

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

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

It follows

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

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

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

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

$$
Pr(\mathcal{R}) > Pr\{ (\hat{\delta}_1, \ldots, \hat{\delta}_d) : 0 > \hat{\delta}_j > -\alpha_j, j = 1, \ldots, d_1 \} = C.
$$
where $C$ is a constant strictly less than 1 depending on $\sigma^2$ and $d_1$ but not on the sample size $n$. We have proved that with a positive probability not depending on $n$, the Lasso algorithm 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 $\gamma$ by $|\hat{\delta}_{d+1}|$ in the preceding proof. When the design matrix satisfies $X^TX = nI_d$, following the same argument in theorem 4.1, we can prove that 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 shown in the simulation in the next section.

5 Simulation

We conduct some simple simulations in general design matrix cases to demonstrate that the Lasso is not consistent in terms of model selection, when the prediction error is to be minimized. All simulations were conducted using MATLAB code. We used the algorithm as suggested in Tibshirani (1996). Each $\beta_j$ is rewritten as $\beta_j^+ - \beta_j^-$, where $\beta_j^+$ and $\beta_j^-$ 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

$$ y = X\beta + \epsilon, $$

where

- Model 1: $\beta = (1,0)^T$,
- Model 2: $\beta = (3,1.5,0,0)^T$;

$\epsilon$ follows standard normal distribution and $x_j$ has marginal distribution $N(0,1)$. The pairwise correlation between $x_i$ and $x_j, i \neq j$, is $\rho$ with $\rho = 0, 0.5, 0.9$. We simulate data with sample size $n = 40, 400, 4000$. For each $\rho$ and each sample size, we simulate 100 data sets and apply the Lasso method. 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 see that the Lasso misses the right model a large fraction of the time, and this is independent of the sample size and the correlation. The results of the experiment are consistent with the previous conclusion.
<table>
<thead>
<tr>
<th>n</th>
<th>( \rho )</th>
<th>Model 1 PCM (%)</th>
<th>Model 2 PCM (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>26</td>
<td>15</td>
</tr>
<tr>
<td>40</td>
<td>0.5</td>
<td>16</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>22</td>
<td>16</td>
</tr>
<tr>
<td>400</td>
<td>0.5</td>
<td>23</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>25</td>
<td>13</td>
</tr>
<tr>
<td>4000</td>
<td>0.5</td>
<td>21</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>24</td>
<td>20</td>
</tr>
</tbody>
</table>

Table 5.1: Simulation results for the Lasso.

6 Summary

We have showed in this paper that the Lasso, the Lars and the FSW are not consistent in terms of model selection when a prediction based criterion is used to select the tuning parameters, and there are superfluous variables in the model.

We remark that our results should not be taken to imply that the Lasso and related methods can not be used as variable selection tools. What our results imply is that the dual goal of accurate estimation and consistent variable selection can not be achieved simultaneously by these methods. The common practice in applying these methods is to choose the tuning parameter to minimize the prediction error, our results state that in this case the procedures are not consistent in terms of variable selection. It is possible that some other criteria of choosing the tuning parameter can yield consistent variable selection for these methods.

References


