Variable Selection by Lasso-Type Methods

Sohail Chand
College of Statistical and Actuarial Sciences
University of the Punjab
Lahore, Pakistan
sohail.stat@pu.edu.pk

Shahid Kamal
College of Statistical and Actuarial Sciences
University of the Punjab
Lahore, Pakistan
kamal_shahid@yahoo.com

Abstract

Variable selection is an important property of shrinkage methods. The adaptive lasso is an oracle procedure and can do consistent variable selection. In this paper, we provide an explanation that how use of adaptive weights makes it possible for the adaptive lasso to satisfy the necessary and almost sufficient condition for consistent variable selection. We suggest a novel algorithm and give an important result that for the adaptive lasso normalisation of predictors after the introduction of adaptive weights makes the adaptive lasso performance identical to the lasso.

Keywords: Lasso, Adaptive lasso, Variable Selection, LARS.

1. Introduction

Tibshirani(1996) proposed a new shrinkage method named least absolute shrinkage and selection operator, abbreviated as lasso. The theoretical properties of lasso-type methods are well studied in the past decade. For example, Fan and Li (2001) have discussed the relationship between the penalized least squares and subset selection and also studied the variable selection properties for lasso-type methods. The lasso can do consistent model selection if it satisfies a necessary condition on the covariance matrix of predictors (Zhao and Yu, 2006). This same condition is also independently derived by Zou (2006).

As discussed by Fan and Li (2001), penalised regression methods such as the lasso, ideally, possess two oracle properties:

- the zero components (and only the zero components) are estimated as exactly zero with probability approaches 1 as $n \to \infty$, where n is the sample size; and
- the non-zero parameters are estimated as efficiently well as when the correct submodel is known.

The oracle properties of these procedures are studied for different models and under various conditions e.g. the necessary condition for consistent selection discussed in Zhao and Yu (2006) and Zou (2006). We will demonstrate

numerically that when this condition fails the adaptive lasso can still do correct variable selection while the lasso cannot.

The rest of the paper is organised as follows: Section 2 gives a review of some shrinkage methods and popular algorithm LARS proposed by Efron et al. (2004) to obtain solution for these methods. Section 3 discusses an important asymptotic condition for consistent variable selection by lasso-type methods. In this section we also provide an explanation that why the adaptive lasso is an oracle procedure while the lasso is not. Section 4 gives an important result about the normalisation of predictors. We also show that in the situations when the necessary condition for the consistent variable selection fails for the lasso and if for the adaptive lasso predictors are normalised after the introduction of adaptive weights, then the adaptive lasso performs identical to the lasso. In this section we also suggest a novel algorithm to attain consistent variable selection for the adaptive lasso. Concluding remarks can be found in Section 5.

2. Shrinkage Methods

The ready availability of fast and powerful computers, combined with rapid technological advances in methods of automated data collection, have led to the routine production of massive datasets, e.g. in bioinformatics. There are many real-life examples where we are dealing with a very large number of predictors, and this naturally leads to consideration of high-dimensional settings.

Traditional statistical estimation procedures such as ordinary least squares (OLS) tend to perform poorly in high-dimensional problems. In particular, although OLS estimators typically have low bias, they tend to have high prediction variance, and may be difficult to interpret (Brown, 1993). In such situations it is often beneficial to use shrinkage i.e. shrink the estimator towards the zero vector, which in effect involves introducing some bias so as to decrease the prediction variance, with the net result of reducing the mean squared error of prediction.

There are several shrinkage methods suggested in the literature including ridge regression (Hoerl and Kennard, 1970). The paper by Tibshirani (1996), in which he suggested the lasso, is a big breakthrough in the field of sparse model estimation which performs the variable selection and coefficient shrinkage simultaneously. Other shrinkage methods include non-negative garotte (Breiman, 1995), smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001), elastic net (Zou and Hostile, 2005), adaptive lasso (Zou, 2006), Dantzig selector (Candes and Tao, 2007), relaxed lasso (Meinshausen, 2007) variable inclusion and selection algorithm (VISA) (Radchenko, 2008) and the double Dantzig selector (James and Radchenoko, 2009). Many other methods have been suggested in the literature but lasso-type methods are currently popular among researchers (Knight and Fu, 2000; Fan and Li, 2001; Wang and Leng, 2007; Hsu et al., 2008). The group lasso was originally suggested by Bakin (1999). This technique selects a group of variables; rather than individual variables, for more details see e.g. Yuan and Lin (2006), Zhao and Kulasekera (2006).

James et al. (2009) proposed an algorithm DASSO (Dantzig selector with sequential optimization) to obtain the entire coefficient path for the Dantzig selector and they also investigated the relationship between the lasso and Dantzig selector. Hesterberg et al. (2008) have given a good survey of L_1 penalised regression. Very recent papers by Fan and Lv (2008,2009), and Lv and Fan (2009) are good references for variable selection especially in high dimension setting. Very recently, the applications of lasso-type methods and their oracle properties are studied by Chand (2011) for regression and multivariate time series models. Leng (2010) in his recent paper suggested a shrinkage method based on the rank regression. He also proposed a score based information criterion for tuning parameter selection. To optimize the variable selection of the lasso and Forward selection method, Radchenko and James (2011) suggested an adjustment in level of shrinkage at every step in his proposed method called forward lasso adaptive shrinkage (FLASH).

In the following paragraphs we will define the linear model and some notations used and referred to frequently in the later sections.

Let $(\mathbf{x}_1^T, y_1), \ldots, (\mathbf{x}_n^T, y_n)$ be n independent and identically distributed random vectors, assumed to satisfy the linear model

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \tag{2.1}$$

such that $y_i \in \mathbb{R}$ is the response variable, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T \in \mathbb{R}^p$ is the p-dimensional set of predictors, the ε_i's are independently and identically distributed with mean 0 and variance σ^2 and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ is the set of parameters.

We define $\mathcal{A} = \{j : \beta_j \neq 0\}$ and $\mathcal{A}^c = \{j : \beta_j = 0\}$. Assume that only p_0 $(p_0 < p)$ parameters are non-zero i.e. $\beta_j \neq 0$ for $j \in \mathcal{A}$ where $|\mathcal{A}| = p_0$ and $|.|$ stands for the number of elements in the set i.e. the cardinality of the set. Thus we can define $\boldsymbol{\beta}_{\mathcal{A}} = \{\beta_j : j \in \mathcal{A}\}$ and $\boldsymbol{\beta}_{\mathcal{A}^c} = \{\beta_j : j \in \mathcal{A}^c\}$. Also assume that $\frac{1}{n}\mathbf{X}^T\mathbf{X} \xrightarrow{p} \mathbf{C}$, where $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ is the design matrix and $\mathbf{C}$ is a positive definite matrix. We can define a partition of the matrix $\mathbf{C}$ as

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix} \tag{2.2}$$

where $\mathbf{C}_{11}$ is the $p_0 \times p_0$ submatrix corresponding to the active predictors $\{\mathbf{x}_j : j \in \mathcal{A}\}$. The least squares estimator estimates the zero coefficients as non-zero in the model defined above. We would like a method which is consistent in variable selection i.e. which correctly classifies the active (i.e. non-zero

coefficients) and non-active (i.e. zero coefficients) predictors. This is an important property of lasso-type methods as mentioned by Knight and Fu (2000).

2.1 The Lasso

The lasso shrinks some coefficients while setting others exactly to zero, and thus theoretical properties suggest that the lasso potentially enjoys the good features of both subset selection and ridge regression (Tibshirani, 1996). The lasso estimator of $\mathbf{b}$ is defined by

$$\widehat{\boldsymbol{\beta}}^{*} = argmin \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}\beta_j x_{ij} \right)^2 \quad subjectto \sum_{j=1}^{p} | \beta_j | \leq t,$$

or equivalently,

$$\widehat{\boldsymbol{\beta}}^{*} = argmin \left\{ \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}\beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} | \beta_j | \right\},$$

where t and λ are user-defined tuning parameters and control the amount of shrinkage. Smaller values of t and larger values of λ result in a higher amount of shrinkage.

2.2 LARS Algorithm

Efron et al. (2004) developed an efficient algorithm known as least angle regression (LARS) algorithm for finding the solution path of the lasso method, where the solution path is the set of values of $\hat{\mathbf{b}}^{*}(\lambda)$ as λ varies. Efron et al. (2004) also showed that both forward stagewise linear regression and the lasso are variants of the LARS. ("L" for least, "A" for angle, "R" for regression and "S" suggests "Lasso" and "Stagewise"). LARS cleverly organizes the calculations and thus the computational cost of the entire p steps is of the same order as that required for the usual least squares solution for the full model, though LARS modified for the lasso solution requires some additional steps (Efron et al., 2004). LARS, like classic forward selection, starts with all coefficients equal to zero.

Hastie et al. (2007) described the LARS algorithm to obtain the lasso solution as follows:

Algorithm 1: LARS alogorithm

Step 1 Standardise the predictors $\{\mathbf{x}_j : j = 1,\ldots,p\}$ to have zero mean and unit variance. Start with the residual $\mathbf{r} = \mathbf{y} - \bar{\mathbf{y}}$, $\beta_1,\ldots,\beta_p = 0$.

Step 2 Find the predictor $\mathbf{x}_j$ most correlated with $\mathbf{r}$.

Step 3 Move β_j from 0 towards its least squares coefficient $(\mathbf{x}_j, \mathbf{r})$, until some other competitor $\mathbf{x}_k$ has as much correlation with the current residual as does $\mathbf{x}_j$.

Step 4 Move (β_j, β_k) in the direction defined by their joint least squares coefficient of the current residual on $(\mathbf{x}_j, \mathbf{x}_k)$, until some other competitor $\mathbf{x}_l$ has as much correlation with the current residual.

Step 5 If a non-zero coefficient hits zero, drop it from $\mathcal{A}$ and recompute the current joint least squares direction.

Step 6 Continue in this way until all p predictors have been entered in the model and we arrive at the full least squares solution. LARS algorithm

2.3 The Adaptive Lasso

Zou (2006) proposed a new version of the lasso, named the adaptive lasso, by using adaptive weights which result in different penalisation for the coefficients appearing in the L_1 penalty term. The adaptive lasso can be defined as

$$\hat{\boldsymbol{\beta}}^{**} = argmin\left\{\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p} w_j \mid \beta_j \mid\right\},$$

where $\mathbf{w} = (w_1, \ldots, w_p)$ are the adaptive weights. Zou has shown that if the weights are efficiently chosen in a data-dependent way then the adaptive lasso can achieve the oracle properties. He suggested the use of estimated weights, $\hat{w}_j = \mid b_j \mid^{-\gamma}$, where $\mathbf{b} = \{b_j : j = 1, \ldots, p\}$ is a root-n-consistent estimator of $\boldsymbol{\beta}$ and $\gamma > 0$ is a user-chosen constant.

The choice of $\hat{w}_j$ is very important and Zou (2006) suggested using ordinary least squares estimates while γ can be chosen by k-fold cross-validation. Zou (2006) has also noted that the adaptive lasso, like the lasso, is a convex optimisation problem and so does not suffer from having more than one local minimum, and its global minimum can be obtained by the LARS algorithm (Efron et al., 2004) after a simple modification given in Algorithm 2.

Algorithm 2: Zou algorithm to obtain the adaptive lasso solution.

Step 1 Define $\mathbf{x}_j^* = \mathbf{x}_j / \hat{w}_j$, $j = 1, \ldots, p$.

Step 2 Solve the lasso problem for all λ

$$\hat{\boldsymbol{\beta}}^{*} = argmin\left\{\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}\mathbf{x}_j^{*}\beta_j\right)^2 + \lambda\sum_{j=1}^{p} \mid \beta_j \mid\right\}.$$

Step 3 Output $\hat{\beta}_j^{**} = \hat{\beta}_j^{*} / \hat{w}_j$.

Zou (2006) has studied whether the standard lasso has the oracle properties discussed by Fan and Li (2001). He showed that there are some scenarios e.g. when condition (2.3) given below does not hold, the lasso variable selection is not consistent. The oracle properties of other shrinkage methods are also studied

in the literature. Fan and Li (2001) have studied the asymptotic properties of the SCAD and showed that penalized likelihood methods have some local maximisers for which the oracle properties hold.

Zou (2006) also gave a necessary and almost sufficient condition for the consistency of lasso variable selection. This condition, named as the irrepresentable condition, was also found independently by Zhao and Yu (2006). We will call this condition the Zhao-Yu-Zou condition (ZYZ condition). Assuming $\mathbf{C}_{11}$ is invertible, the ZYZ condition can be stated as

$$\left|\left[\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{s}_{\beta(\mathcal{A})}\right]_r\right| \le 1, \quad r = 1,\ldots,p-p_0, \tag{2.3}$$

where $\mathbf{C}_{11}$, $\mathbf{C}_{21}$ are the partitions of $\mathbf{C}$ defined in (2.2), $\mathbf{s}_{\beta(\mathcal{A})} = \{\text{sgn}(\beta_j) : j \in \mathcal{A}\}$ and p_0 is the number of elements in $\mathcal{A}$.

In general, lasso-type methods are more effective than conventional methods, e.g. ordinary least squares, when the true model is sparse. If sparsity is not known to be present then there are not many advantages of using lasso-type methods as the shrinkage results in biased estimates for the nonzero components (Hsu et al., 2008).

3. ZYZ Condition and Variable Selection

The ZYZ condition (2.3) discussed by Zhao and Yu (2006) and Zou (2006) is a necessary condition on the matrix $\mathbf{C}$ defined in (2.2) for consistent variable selection. The ZYZ condition is always satisfied for an orthogonal design but there are some scenarios where this condition fails. Zhao and Yu (2006) and Zou (2006) have presented some examples where this condition fails, in which case, the lasso is inconsistent in variable selection. However, Zou (2006) has shown that the adaptive lasso has the oracle properties for the linear regression model, so that variable selection is consistent.

An important point to note is that the ZYZ condition is an asymptotic condition. The condition requires $\lambda \xrightarrow{p} 0$, which refers to large sample sizes $(n \to \infty)$. For finite sample sizes, the ZYZ condition does not always guarantee good variable selection.

When using the *lars* package in R for the implementation of the adaptive lasso, we notice that the theoretical properties are not shown in the simulated examples. As shown by Zou (2006) the adaptive lasso is consistent in variable selection even where the ZYZ condition fails for the standard lasso, but we failed to approach the variable selection oracle property of the adaptive lasso in the numerical example when the sample size becomes large. These strange results for the adaptive lasso led us to look in depth to the LARS algorithm. We noticed that if we implement LARS algorithm to obtain the adaptive lasso solution as mentioned in Step 2 of Zou's Algorithm 2 and as predictors are rescaled by adaptive weights in Step 1 of this algorithm thus normalisation at Step 2 in LARS algorithm nullifies the effect of adaptive weights. See Section 4 for details.

Normalisation of predictors is a common practice so it is important to know that at which stage we should perform this normalisation especially in the case of the adaptive lasso.

Now we show how the use of adaptive weights makes the ZYZ condition hold even when it originally fails. Assume $n^{-1}\mathbf{X}^T\mathbf{X} = \mathbf{C}^{(n)} \xrightarrow{p} \mathbf{C}$ and is partitioned as indicated in (2.2). The adaptive lasso rescales the design matrix $\mathbf{X}$ using some data-driven adaptive weights $\{w_j : j = 1, \ldots, p\}$. We can rearrange and partition the weight matrix, $\mathbf{W}$, as

$$\mathbf{W} = \begin{pmatrix} \mathbf{W}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_{22} \end{pmatrix},$$

where $\mathbf{W}_{11} = diag(w_j^{-1}; j \in \mathcal{A})$ and $\mathbf{W}_{22} = diag(w_j^{-1}; j \in \mathcal{A}^c)$.

Writing $\widetilde{\mathbf{X}} = \mathbf{X}\mathbf{W}$, we can define $\widetilde{\mathbf{C}}^{(n)} = n^{-1}\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}} \xrightarrow{p} \widetilde{\mathbf{C}}$ $\widetilde{\mathbf{C}}^{(n)}$ and can be partitioned as

$$\widetilde{\mathbf{C}}^{(n)} = \begin{pmatrix} \widetilde{\mathbf{C}}_{11}^{(n)} & \widetilde{\mathbf{C}}_{12}^{(n)} \\ \widetilde{\mathbf{C}}_{21}^{(n)} & \widetilde{\mathbf{C}}_{22}^{(n)} \end{pmatrix}.$$

Now,

$$\begin{aligned} \widetilde{\mathbf{C}}^{(n)} &= \frac{1}{n}\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}} \\ &= \mathbf{W}^T\mathbf{C}^{(n)}\mathbf{W} \\ &= \begin{pmatrix} \mathbf{W}_{11}\mathbf{C}_{11}^{(n)}\mathbf{W}_{11} & \mathbf{W}_{11}\mathbf{C}_{12}^{(n)}\mathbf{W}_{22} \\ \mathbf{W}_{22}\mathbf{C}_{21}^{(n)}\mathbf{W}_{11} & \mathbf{W}_{22}\mathbf{C}_{22}^{(n)}\mathbf{W}_{22} \end{pmatrix}. \end{aligned} \tag{3.1}$$

Take

$$\begin{aligned} \widetilde{\mathbf{C}}_{21}^{(n)}\widetilde{\mathbf{C}}_{11}^{(n)^{-1}}\mathbf{s}_{\beta(\mathcal{A})} &= \left(\mathbf{W}_{22}\mathbf{C}_{21}^{(n)}\mathbf{W}_{11}\right)\left(\mathbf{W}_{11}\mathbf{C}_{11}^{(n)}\mathbf{W}_{11}\right)^{-1}\mathbf{s}_{\beta(\mathcal{A})} \\ &= \mathbf{W}_{22}\mathbf{C}_{21}^{(n)}(\mathbf{C}_{11}^{(n)})^{-1}\mathbf{W}_{11}^{-1}\mathbf{s}_{\beta(\mathcal{A})}, \end{aligned}$$

where $\mathbf{s}_{\beta(\mathcal{A})}$ is defined in (2.3). If the weights $\{w_j\}$ are chosen appropriately (typical choices are inverse powers of absolute values of least squares estimates or ridge estimates or lasso estimates) then,

$$w_j = \frac{1}{|\hat{\beta}_j|^\gamma} \xrightarrow{p} \begin{cases} 1/|\beta_j|^\gamma, & j \in \mathcal{A} \\ \infty, & j \notin \mathcal{A}. \end{cases} \tag{3.2}$$

As $\mathbf{W}_{11} = diag\left(w_j^{-1} : j \in \mathcal{A}\right)$, $\mathbf{W}_{11}^{-1} = diag\left(w_j : j \in \mathcal{A}\right)$, so we can say, when in general $|\beta_j| >> 1$ for $j \in \mathcal{A}$, the elements of $\mathbf{W}_{11}^{-1}$ will be bounded by some finite value say k^*. Moreover, since $\mathbf{W}_{22} = diag\left(w_j^{-1} : j \notin \mathcal{A}\right)$, it can be easily concluded from (3.2) that $\mathbf{W}_{22} \xrightarrow{p} \mathbf{0}_{p-p_0}$, the $(p-p_0)\times(p-p_0)$ matrix of zeros. So for an appropriate choice of the adaptive lasso weights, we can say that componentwise

$$\left|\left[\mathbf{W}_{22}\mathbf{C}_{21}^{(n)}(\mathbf{C}_{11}^{(n)})^{-1}\mathbf{W}_{11}\mathbf{s}_{\beta(\mathcal{A})}\right]_r\right| \to 0, \quad r=1,\ldots,p-p_0$$

thus we can conclude that componentwise

$$\left|\left[\mathbf{W}_{22}\mathbf{C}_{21}^{(n)}(\mathbf{C}_{11}^{(n)})^{-1}\mathbf{W}_{11}\mathbf{s}_{\beta(\mathcal{A})}\right]_r\right| \le 1, \quad r=1,\ldots,p-p_0 \tag{3.3}$$

always holds, at least asymptotically. So the adaptive lasso always satisfies the ZYZ condition asymptotically.

4. Normalisation after Rescaling by the Adaptive Weights

We have mentioned earlier that, under certain conditions, normalisation of the design matrix often improves the performance of the lasso. As penalized least squares methods are not scale equivariant, it is recommended to normalize the predictors so that each variable has unit L_2 norm. Such a scaling is also the default option of the *lars* package in R.

To provide insight into the effect of normalisation, we consider a simple case. Suppose we have p predictors $\{\mathbf{x}_j : j=1,\ldots,p\}$ for the model defined in (2.1) such that $n^{-1}\mathbf{X}^T\mathbf{X} \xrightarrow{p} \mathbf{C}$. LARS uses $\mathbf{x}_j / h_j$ to normalise the predictors, where $h_j = \sqrt{\sum_{i=1}^{n} x_{ij}^2}; j=1,\ldots,p$.

Let $\widetilde{\mathbf{Z}}$ be the normalised design matrix of $\widetilde{\mathbf{X}}$, which can be defined as

$$\widetilde{\mathbf{Z}} = \widetilde{\mathbf{X}}\mathbf{D}, \tag{4.1}$$

where $\mathbf{D} = diag\left(1/h_1,\ldots,1/h_p\right)$. For illustrative purposes we consider

$$h_j = \begin{cases} h_1^* & for\ all\ j \in \mathcal{A} \\ h_2^* & for\ all\ j \in \mathcal{A}^c \end{cases}.$$

Thus $\mathbf{D}$ can be partitioned as $\mathbf{D} = \begin{pmatrix} \mathbf{D}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_{22} \end{pmatrix}$, where $\mathbf{D}_{11} = h_1^{*-1}\mathbf{I}_{p_0}$ and $\mathbf{D}_{22} = h_2^{*-1}\mathbf{I}_{p-p_0}$. We can write the covariance matrix for the normalised predictors defined in (4.1) as $\widetilde{\mathbf{C}}_Z^{(n)} = n^{-1}\widetilde{\mathbf{Z}}^T\widetilde{\mathbf{Z}}$ as follows:

$$\begin{aligned} \widetilde{\mathbf{C}}_Z^{(n)} &= \begin{pmatrix} \mathbf{D}_{11}\widetilde{\mathbf{C}}_{11}^{(n)}\mathbf{D}_{11} & \mathbf{D}_{11}\widetilde{\mathbf{C}}_{12}^{(n)}\mathbf{D}_{22} \\ \mathbf{D}_{22}\widetilde{\mathbf{C}}_{21}^{(n)}\widetilde{\mathbf{C}}_{11}^{(n)} & \mathbf{D}_{22}\widetilde{\mathbf{C}}_{22}^{(n)}\mathbf{D}_{22} \end{pmatrix} \\ &= \begin{pmatrix} \widetilde{\mathbf{C}}_{11(Z)}^{(n)} & \widetilde{\mathbf{C}}_{12(Z)}^{(n)} \\ \widetilde{\mathbf{C}}_{21(Z)}^{(n)} & \widetilde{\mathbf{C}}_{22(Z)}^{(n)} \end{pmatrix}, say, \end{aligned}$$

where $\widetilde{\mathbf{C}}_{ij(Z)}^{(n)} = (h_i^* h_j^*)^{-1}\widetilde{\mathbf{C}}_{ij}^{(n)}$, $i,j=1,2$.

Now take

$$\widetilde{\mathbf{C}}_{21(Z)}^{(n)}(\widetilde{\mathbf{C}}_{11(Z)}^{(n)})^{-1}\mathbf{s}_{\beta(\mathcal{A})} = \left(\mathbf{D}_{22}\widetilde{\mathbf{C}}_{21}^{(n)}\mathbf{D}_{11}\right)\left(\mathbf{D}_{11}\widetilde{\mathbf{C}}_{11}^{(n)}\mathbf{D}_{11}\right)^{-1}\mathbf{s}_{\beta(\mathcal{A})}$$

$$= \mathbf{D}_{22}\widetilde{\mathbf{C}}_{21}^{(n)}(\widetilde{\mathbf{C}}_{11}^{(n)})^{-1}\mathbf{D}_{11}^{-1}\mathbf{s}_{\beta(\mathcal{A})}$$

$$= h_1^* h_2^{*-1}\widetilde{\mathbf{C}}_{21}^{(n)}(\widetilde{\mathbf{C}}_{11}^{(n)})^{-1}\mathbf{s}_{\beta(\mathcal{A})}.$$

Using $\widetilde{\mathbf{C}}_{11}^{(n)} = \mathbf{W}_{11}\mathbf{C}_{11}^{(n)}\mathbf{W}_{11}$ and $\widetilde{\mathbf{C}}_{21}^{(n)} = \mathbf{W}_{22}\mathbf{C}_{21}^{(n)}\mathbf{W}_{11}$, we get

$$\widetilde{\mathbf{C}}_{21(Z)}^{(n)}(\widetilde{\mathbf{C}}_{11(Z)}^{(n)})^{-1}\mathbf{s}_{\beta(\mathcal{A})} = h_1^* h_2^{*-1}\left(\mathbf{W}_{22}\mathbf{C}_{21}^{(n)}(\mathbf{C}_{11}^{(n)})^{-1}\mathbf{W}_{11}\right)\mathbf{s}_{\beta(\mathcal{A})}.$$

For the necessary condition for consistent variable selection to hold, we require

$$\left|\left[h_1^* h_2^{*-1}\left(\mathbf{W}_{22}\mathbf{C}_{21}^{(n)}(\mathbf{C}_{11}^{(n)})^{-1}\mathbf{W}_{11}\right)\mathbf{s}_{\beta(\mathcal{A})}\right]_r\right| \le 1, \quad r = 1,\ldots,p-p_0.$$

Using the result in (3.3), this will lead to two different scenarios:

- if $h_1^* \le h_2^*$, then the ZYZ condition always holds;
- if $h_1^* > h_2^*$, then normalisation can lead to failure of the ZYZ condition thus making the variable selection inconsistent.

4.1 Numerical Results

Consider the $y_i = \mathbf{x}_i^{'}{}_0 + \varepsilon_i$, where $\boldsymbol{\beta}_0 = (5.6, 5.6, 5.6, 0)^T$, $\varepsilon_i \sim N(0,\sigma^2)$ and $\mathbf{X} = (\mathbf{x}_1,\ldots,\mathbf{x}_4)^T \sim N_4(\mathbf{0},\mathbf{C})$ where

$$\mathbf{C} = \begin{bmatrix} 1 & -0.39 & -0.39 & 0.23 \\ -0.39 & 1 & -0.39 & 0.23 \\ -0.39 & -0.39 & 1 & 0.23 \\ 0.23 & 0.23 & 0.23 & 1 \end{bmatrix}.$$

This is the same model as Model 0 studied by (Zou, 2006). We will also refer this model as Model 0. For this above choice of $\mathbf{C}$, we have $\left|\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{s}_{\beta(\mathcal{A})}\right| = 3.1363 > 1$, thus the ZYZ condition fails. Suppose $\mathbf{C}^{(n)}$ be the covariance matrix of the simulated set of predictors, $\mathbf{x}_i$:

$$\mathbf{C}^{(n)} = \begin{bmatrix} 1.0428311 & -0.4203259 & -0.3738564 & 0.2409415 \\ -0.4203259 & 0.9585507 & -0.3345396 & 0.2163182 \\ -0.3738564 & -0.3345396 & 0.9631588 & 0.2878907 \\ 0.2409415 & 0.2163182 & 0.2878907 & 1.0548909 \end{bmatrix}.$$

We observed that $\left|\mathbf{C}_{21}^{(n)}(\mathbf{C}_{11}^{(n)})^{-1}\mathbf{s}_{\beta(\mathcal{A})}\right| = 3.161134 > 1$. Thus the ZYZ condition fails so the lasso variable selection will be inconsistent.

Now if we apply the adaptive lasso, we need to rescale the predictors $\tilde{\mathbf{x}}_j = \mathbf{x}_j / w_j$ using the adaptive weights, w_j. Here, for example, we use estimated weights $\hat{w}_j = |\hat{\beta}_j|^{-1}$, for $j = 1, \ldots, p$, where $\hat{\beta}_j$ is the least squares estimate of β_j, i.e. we choose the tuning parameter γ to be 1. Now the covariance matrix, $\widetilde{\mathbf{C}}^{(n)}$, of the $\tilde{\mathbf{x}}_j$'s is given by

$$\widetilde{\mathbf{C}}^{(n)} = \begin{bmatrix} 0.169194693 & -0.080898724 & -0.067396782 & 0.004310662 \\ -0.080898724 & 0.218853480 & -0.071542606 & 0.004591009 \\ -0.067396782 & -0.071542606 & 0.192927391 & 0.005722972 \\ 0.004310662 & 0.004591009 & 0.005722972 & 0.002081131 \end{bmatrix}$$

and we observed that $\left|\widetilde{\mathbf{C}}_{21}^{(n)} (\widetilde{\mathbf{C}}_{11}^{(n)})^{-1} \mathbf{s}_{\beta(\mathcal{A})}\right| = 0.3185492 < 1$. Thus the ZYZ condition holds and leads the adaptive lasso to consistent variable selection. Now if we normalise the predictors after rescaling by the adaptive weights, the effect of the adaptive weights is nullified and the resulting covariance matrix after normalisation is given as

$$\widetilde{\mathbf{C}}_z^{(n)} = \begin{bmatrix} 6.9640450 \times 10^{-3} & -2.068320 \times 10^{-4} & -1.109767 \times 10^{-3} & 8.745434 \times 10^{-4} \\ -2.068320 \times 10^{-4} & 3.475609 \times 10^{-5} & -7.317432 \times 10^{-5} & 5.785579 \times 10^{-5} \\ -1.1097671 \times 10^{-3} & -7.317432 \times 10^{-5} & 1.270880 \times 10^{-3} & 4.644906 \times 10^{-4} \\ 8.745434 \times 10^{-4} & 5.785579 \times 10^{-5} & 4.644906 \times 10^{-4} & 2.081131 \times 10^{-3} \end{bmatrix}$$

and we observe that $\left|\widetilde{\mathbf{C}}_{21(z)} (\widetilde{\mathbf{C}}_{11(z)}^{(n)})^{-1} \mathbf{s}_{\beta(\mathcal{A})}\right| = 9.645727 > 1$. Hence if predictors are normalised after introducing adaptive weights the adaptive lasso will result into the standard lasso.

The general case is less transparent but, even so, this illustrative example throws some light in to the effect of normalisation. The use of adaptive weights makes the adaptive lasso an oracle procedure. Therefore it is crucial to determine at which stage we should normalise the predictors, if required. We observe that normalisation nullifies the effect of adaptive weights if it is done after the introduction of adaptive weights. In Algorithm 3, we elaborate the Zou (2006) Algorithm 1 to obtain the adaptive lasso estimates for normalised predictors.

Algorithm 3: New algorithm to obtain the adaptive lasso solution

Step 1 Standardise the predictors $\mathbf{x}_1, \ldots, \mathbf{x}_p$ so that each has mean 0 and variance 1.

Step 2 Estimate the weights $\hat{w}_j, j = 1, \ldots, p$ using the normalised predictors obtained in Step 1 above.

Step 3 Define $\mathbf{x}_j^* = \mathbf{x}_j / \hat{w}_j$, $j = 1, \ldots, p$.

Step 4 Solve the lasso problem for all λ

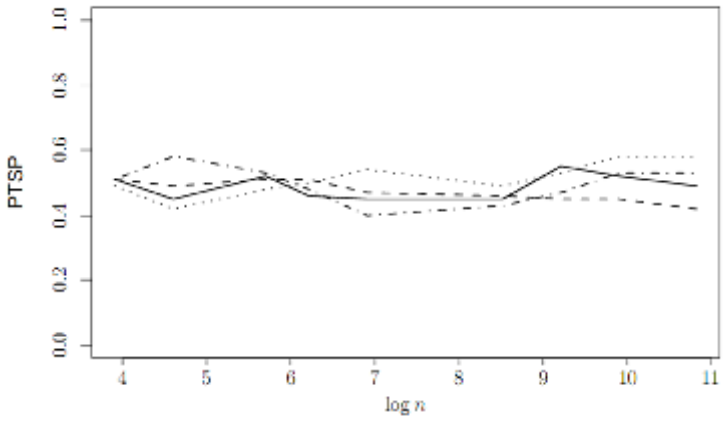
$$\widehat{\boldsymbol{\beta}}^{*} = argmin\left\{\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}\mathbf{x}_j^{*}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right\}.$$

Step 5 Output $\hat{\beta}_j^{**} = \hat{\beta}_j^{*} / \hat{w}_j$. New algorithm to obtain the adaptive lasso solution.

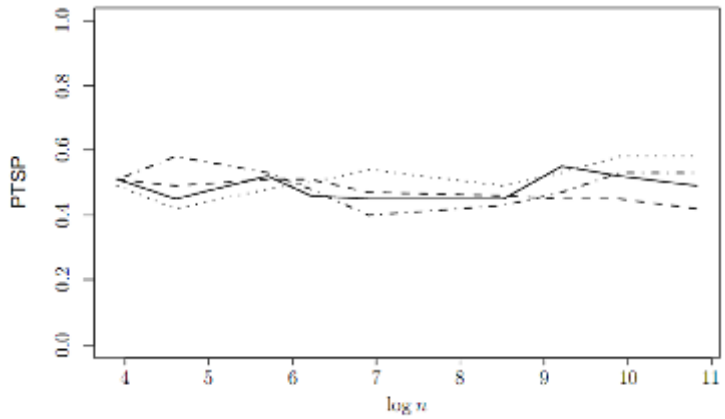
Now we compare the lasso and adaptive lasso variable selection when ZYZ condition fails. We obtain the lasso solution path using Algorithm 1 while for the adaptive lasso we use Algorithm 2 and our suggested novel Algorithm 3.

We study the same Model 0 and use 100 Monte Carlo runs to study the empirical probability of containing the true model on the solution path (PTSP). The PTSP is the proportion of Monte Carlo runs for which the true model lies on the lasso (adaptive lasso) solution path. In practice, to pick up the correct model from the entire solution path also depends on the efficiency of tuning parameter selector. The advantage of studying this PTSP is that it does not require a single choice of tuning parameter rather we are looking at the entire solution path. The lower the PTSP, more challenging will be for tuning parameter selectors to pick up the correct model. We consider various choices of sample size ranging from $n = 50$ to $n = 50000$ and for error standard deviation $\sigma = 1, 3, 6, 9$. The results are shown in Figure 1.

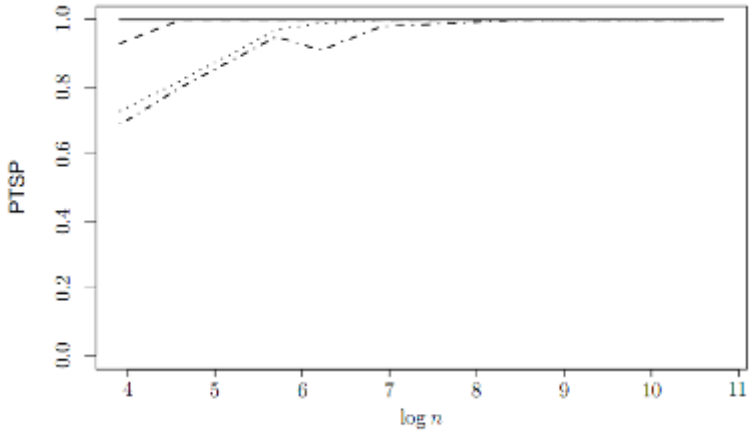
Figure 1 shows the plots of PTSP for the lasso and adaptive lasso based. It can be noticed that PTSP for the lasso (Figure 1(a)) lies around 0.5 and does not converge to 1 even for the sample size as large as $n = 50000$. The adaptive lasso performance is identical to the lasso when solution is obtained using Algorithm 2. While for the adaptive lasso with the new Algorithm 3 shows that the PTSP is converging to 1 thus resulting in consistent variable selection. As expected, the larger error variance makes the variable selection harder.



(a) Lasso; LARS algorithm



(b) Adaptive lasso; Zou algorithm



(c) Adaptive lasso; New Algorithm

Figure 1: Probability, based on 100 runs, that solution paths of the lasso $(\gamma = 0)$ and adaptive lasso $(\gamma = 1)$ for the Model 0: $\mathbf{b}_0 = (5.6, 5.6, 5.6, 0)^T$. The error distribution is $\varepsilon_i \sim N(0, \sigma^2)$ where ([.5ex]10mm0.3pt $\sigma = 1$); ($---$ $\sigma = 3$); ($\cdots\cdots$ $\sigma = 6$); ($-\cdot-\cdot-\cdot$ $\sigma = 9$).

5. Conclusion

The adaptive lasso is an oracle procedure and able to do consistent variable selection. This good property of the adaptive lasso is due to the rescaling of the predictors by the adaptive weights. While using LARS to obtain the solution for the adaptive lasso it is important to do normalisation before this recscaling otherwise the adaptive lasso performs identical to the lasso.

References

1. Bakin, S. (1999). Adaptive regression and model selection in data mining problems. *PhD Thesis, School of Mathematical Sciences, The Australian National University, Canberra*.

2. Breiman, L. (1995). Better subset selection using the non-negative garotte. *Technometrics*, 37(4):373–384.
3. Brown, J. (1993). *Measurement, Regression and Calibration*. Oxford University Press: Oxford, UK.
4. Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. Annals of Statistics, 35(6):2313–2351.
5. Chand, S. (2011). Diagnsotic checking and lasso variable selection in time series analysis. *PhD Thesis, School of Mathematical Sciences, The University of Nottingham, United Kingdom*.
6. Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, 32(2):407–499.
7. Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.
8. Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society*: Series B, 70(5):849–911.
9. Fan, J. and Lv, J. (2009). Non-concave penalized likelihood with NP-dimensionality. *Arxiv preprint arXiv:0910.1119*.
10. Hastie, T., Taylor, J., Tibshirani, R., Walther, G., et al. (2007). Forward stagewise regression and the monotone lasso. *Electronic Journal of Statistics*, 1(1):1–29.
11. Hesterberg, T., Choi, N., Meier, L., and Fraley, C. (2008). Least angle and L1 penalized regression: A review. *Statistics Surveys*, 2:61–93.
12. Hoerl, A. and Kennard, R. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.
13. Hsu, N., Hung, H., and Chang, Y. (2008). Subset selection for vector autoregressive processes using Lasso. *Computational Statistics and Data Analysis*, 52(7):3645–3657.
14. James, G. and Radchenko, P. (2009). A generalized dantzig selector with shrinkage tuning. *Biometrika*, 96(2):323.
15. James, G., Radchenko, P., and Lv, J. (2009). DASSO: Connections between the dantzig selector and lasso. *Journal of Royal Statistical Society*, Series B, 71(1):121–142.
16. Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *Annals of Statistics*, 28(5):1356–1378.
17. Leng, C. (2010). Variable selection and coefficient estimation via regularized rank regression. *Statist. Sinica*, 20:167–181.
18. Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Annals of Statistics*, 37(6A):3498–3528.
19. Meinshausen, N. (2007). Relaxed lasso. *Computational Statistics & Data Analysis*, 52(1):374–393.

20. Radchenko, P. and James, G. (2008). Variable inclusion and shrinkage algorithms. *Journal of the American Statistical Association*, 103(483):1304–1315.
21. Radchenko, P. and James, G. (2011). Improved variable selection with forwardlasso adaptive shrinkage. *The Annals of Applied Statistics*, 5(1):427–448.
22. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, Series B, 58(1):267–288.
23. Wang, H. and Leng, C. (2007). Unified lasso estimation via least squares approximation. *Journal of the American Statistical Association*, 102(479):1039–1048.
24. Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society*: Series B, 68(1):49– 67.
25. Zhao, M. and Kulasekera, K. (2006). Consistent linear model selection. *Statistics & Probability Letters*, 76(5):520–530.
26. Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563.
27. Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.
28. Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B*, 67(2):301–320.