On Tuning Parameter Selection of Lasso-Type Methods - A Monte Carlo Study

Sohail Chand
College of Statistical and Actuarial Sciences, University of the Punjab, Lahore-Pakistan.
email: sohail.stat@gmail.com

Abstract*—In regression analysis, variable selection is a challenging task. Over the last decade, the lasso-type methods have become popular method for variable selection due to their property of shrinking some of the model coefficients to exactly zero. Theory says that lasso-type methods are able to do consistent variable selection but it is hard to achieve this property in practice. This consistent variable selection highly depends on the right choice of the tuning parameter. In this paper, we show that selection of tuning parameter by cross validation almost always fail to achieve consistent variable selection. We have also shown that lasso-type methods with a BIC-type tuning parameter selector, under certain conditions, can do the consistent variable selection. We have also made a novel suggestion for choosing the value of C_n, a weight on estimated model size, in BIC. Our results show that with this choice of C_n, the lasso-type methods can do consistent variable selection.*In regression analysis, variable selection is a challenging task. Over the last decade, the lasso-type methods have become popular method for variable selection due to their property of shrinking some of the model coefficients to exactly zero. Theory says that lasso-type methods are able to do consistent variable selection but it is hard to achieve this property in practice. This consistent variable selection highly depends on the right choice of the tuning parameter. In this paper, we show that selection of tuning parameter by cross validation almost always fail to achieve consistent variable selection. We have also shown that lasso-type methods with a BIC-type tuning parameter selector, under certain conditions, can do the consistent variable selection. We have also made a novel suggestion for choosing the value of C_n, a weight on estimated model size, in BIC. Our results show that with this choice of C_n, the lasso-type methods can do consistent variable selection.**

I. INTRODUCTION

Shrinkage methods are popular among the researchers for their theoretical properties e.g. variable selection. 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 [3]. 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.

The usefulness of shrinkage methods is widely discussed in high-dimensional settings but not limited to only this case. These methods can also be effectively applied to a modest number of dimensions when sparsity in the estimated model is desired. In this paper, we have studied the selection of tuning parameter in this setting.

The paper by [21] opens a new horizon of research in variable selection. The lasso, suggested by Tibshirani, is a big breakthrough in the field of sparse model estimation which performs the variable selection and coefficient shrinkage simultaneously. The Other shrinkage methods include non-negative garotte [2], smoothly clipped absolute deviation (SCAD) [8], elastic net [29], adaptive lasso [28], Dantzig selector [4], variable inclusion and selection algorithm (VISA) [19]. Reference [7] suggested a very efficient algorithm, the LARS, which provides the solution path of the lasso at the computational cost of least squares estimates. The solution of the adaptive lasso can also be obtained using the LARS after a simple modification as discussed by [28]. Reference [6] has elaborated the LARS algorithm for the adaptive lasso.

Many other methods have been suggested in the literature but lasso-type methods are currently popular among researchers (16; 8; 22; 14). The group lasso was originally suggested by [1] in his PhD Thesis from The Australian National University, Canberra. This technique selects a group of variables; rather than individual variables, for more details see e.g. Reference [25], [26].

Most recently, [15] 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. Reference [13] have given a good survey of L_1 penalised regression. Very recent papers by [9], [10] and [18] are good reference for variable selection especially in high dimension setting.

The theoretical properties of lasso-type methods are very appealing but there are still a number of unanswered questions including some issues in their practical application, e.g. the selection of the tuning parameter. As discussed by [8], penalised regression methods such as the lasso, ideally, possess two oracle properties:

1) 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

978-1-4577-1929-5/12/$26.00 ©2011 IEEE

size; and

2) the non-zero parameters are estimated as efficiently well as when the correct submodel is known.

Reference [5] has proved these oracle properties of the adaptive lasso for multivariate time series models. The tuning parameter plays a vital role in consistent variable selection. It controls the degree of shrinkage of the estimator. We compare the performance of lasso-type methods using different tuning parameter selectors suggested in the literature.

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

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

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \tag{I.1}$$

such that $y_i \in \mathbb{R}$ is the response variable, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T \in \mathbb{R}^p$ 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 $\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X} \xrightarrow{p} \boldsymbol{C}$, where $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ is the design matrix and $\boldsymbol{C}$ is a positive definite matrix. We can define a partition of the matrix $\boldsymbol{C}$ as

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{C}_{11} & \boldsymbol{C}_{12} \\ \boldsymbol{C}_{21} & \boldsymbol{C}_{22} \end{bmatrix} \tag{I.2}$$

where $\boldsymbol{C}_{11}$ is the $p_0 \times p_0$ submatrix corresponding to the active predictors $\{\boldsymbol{x}_j : j \in \mathcal{A}\}$. 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 [16].

Reference [28] has studied whether the standard lasso has the oracle properties discussed by [8]. He showed that there are some scenarios e.g. when condition (1.3) 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. Reference [8] has studied the asymptotic properties of the SCAD and showed that penalized likelihood methods have some local maximisers for which the oracle properties hold.

Reference [28] 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 [27]. We will call this condition the Zhao-Yu-Zou condition (ZYZ condition). Assuming $\boldsymbol{C}_{11}$ is invertible, the ZYZ condition can be stated as

$$\left|\left[\boldsymbol{C}_{21}\boldsymbol{C}_{11}^{-1}\boldsymbol{s}_{\beta(\mathcal{A})}\right]_r\right| \leq 1, \quad r = 1, \ldots, p - p_0, \tag{I.3}$$

where $\boldsymbol{C}_{11}$, $\boldsymbol{C}_{21}$ are the partitions of $\boldsymbol{C}$ defined in (1.2), $\boldsymbol{s}_{\beta(\mathcal{A})} = \{\text{sgn}(\beta_j) : j \in \mathcal{A}\}$ and p_0 is the number of elements in $\mathcal{A}$. Reference [21] proposed a new shrinkage method named least absolute shrinkage and selection operator, abbreviated as 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 [21]. The lasso estimator of $\boldsymbol{\beta}$ is defined by

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

$$\text{subject to} \sum_{j=1}^{p} |\beta_j| \leq t,$$

or equivalently,

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

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. For detailed discussion on the selection of the tuning parameter see Section 2.

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 [27] and [28]. Reference [6] has demonstrated numerically that when this condition fails the adaptive lasso can still do correct variable selection while the lasso cannot.

Section 2 discusses various methods for choosing the appropriate value of the tuning parameter and its effect on the performance of lasso-type procedures. Section 3 gives some numerical results on the performance of lasso methods for regression models. Finally, Section 4 gives discussion and conclusions about the performance of these lasso-type methods under various conditions.

II. Selection of Tuning Parameter

Selection of the tuning parameter is very important as it can have a big influence on the performance of the estimator. Cross-validation is considered the simplest and most widely used method for minimisation the prediction error [12]. In the literature, cross-validation (CV) is commonly used for estimating the tuning parameter. It is defined later in this section. The most common forms of cross-validation are k-fold, leave-one-out and the generalized cross-validation. The *lars* package uses k-fold cross-validation. We can describe the k-fold cross-validation as below:

1) Data consisting of n observations are divided at random into k mutually exclusive subsamples, known as k-folds.
2) The entire solution path is obtained as a function of the standardized tuning parameter $s \in [0, 1]$ using the LARS algorithm, while omitting the ith fold, where $i = 1, \ldots, k$.
3) The fitted model is then used for prediction of the omitted ith subsample and the prediction error is obtained against each choice of the tuning parameter $s \in [0, 1]$.
4) The value of s which minimizes the prediction error is considered the optimal choice of the tuning parameter.

Typical choices of k are 5 or 10. The choice $k = n$ is known as leave-one-out cross-validation, in this case we have n subsamples and for the ith subsample the fit is computed using all the data after omitting ith observation. Leave-one-out cross-validation is computationally very expensive. Generalized cross-validation provides an approximation to leave-one-out cross-validation. It is used when a linear model is fitted under a squared-error loss function. See [12] for more details.

The theory suggests that consistent variable selection depends very much on the selection of the tuning parameters. We will show and discuss later in Section 3 how the choice of tuning parameter affects the performance of the lasso and adaptive lasso. Our results (Section 3) show that when the tuning parameter is selected using cross-validation, the lasso and adaptive lasso do not appear to be consistent in variable selection, as independently showed by [23]. Reference [17] have shown that if the prediction accuracy criterion is used to select the tuning parameter then lasso-type procedures cannot be consistent in variable selection.

We have noticed that the oracle performance of the lasso can be achieved if a reliable method of tuning parameter selection is used. Recently, papers by [23] and [11] confirmed our conclusions about the poor performance of cross-validation based on numerical results. Reference [23] suggested a Bayesian information criterion (BIC) type criterion to choose the value of the tuning parameter.

The BIC has previously been used as a model selection tool. As in model building, we have several candidate models and adding new parameters to a model will increase the likelihood, but by including more parameters in the model, the model becomes more complex and the estimates also tend to have greater variance. In order to address this problem, [20] suggested a Bayesian information criterion (BIC) for the selection of a better model which achieves a suitable trade-off between simplicity (fewer parameters) and goodness of fit (greater likelihood). In the Gaussian case this takes the form given as

$$BIC = \log(\hat{\sigma}^2) + p \times \frac{\log(n)}{n},$$

where $\hat{\sigma}^2$ is the residual variance and p is the number of parameters. The candidate model which minimizes the BIC is selected. Note that $\log(\hat{\sigma}^2)$ is proportional to a maximised Gaussian likelihood. Reference [24] defined a BIC as follows:

$$BIC_{\mathcal{S}} = \log(\hat{\sigma}^2_{\mathcal{S}}) + |\mathcal{S}| \times \frac{\log(n)}{n} \times C_n,$$

where $|\mathcal{S}|$ is the size of the model i.e. the number of non-zero parameters in the model, $\hat{\sigma}^2_{\mathcal{S}} = SSE_{\mathcal{S}}/n$, $C_n > 0$ and $SSE_{\mathcal{S}}$ is the sum of squares of error for the non-zero component of model. For $C_n = 1$ the modified BIC of [24] reduces to the traditional BIC of [20].

Suppose p_0 is the size of the true model, i.e. the number of non-zero parameters in the true model and $|\mathcal{S}|$ is the size of an arbitrary overfitted model i.e. $\mathcal{S}_T \subset \mathcal{S}$ and $|\mathcal{S}| > p_0$. Under a condtion on the size of non-zero coefficients and standard conditions of finite fourth order moments, [23] showed that $P(BIC_{\mathcal{S}} > BIC_{\mathcal{S}_T}) \longrightarrow 1$ for any overfitted model, $\mathcal{S}$. Thus, the BIC is consistent in differentiating the true model from every overfitted model. Using this property of the BIC, [23] defined a modified BIC for the selection of the optimal value of the tuning parameter λ:

$$BIC_\lambda = \log(\hat{\sigma}^2_\lambda) + |\mathcal{S}_\lambda| \times \frac{\log n}{n} \times C_n, \qquad \text{(II.1)}$$

where $\hat{\sigma}^2_\lambda = SSE_\lambda/n$, $SSE_\lambda = \sum_{i=1}^{n} \sum_{j=1}^{p} \left(y_{ij} - \sum_{j=1}^{p} \boldsymbol{x}_j^T \hat{\beta}_\lambda \right)^2$ is the sum of squared error, $\mathcal{S}_\lambda = \{j : \hat{\beta}_{j,\lambda} \neq 0\}$, $\hat{\beta}_{j,\lambda}$ is the estimate for some chosen value of λ. Importantly, $C_n > 0$ is a constant, which must be very carefully chosen. Wang and Leng used $C_n = \log\log p$ in their simulation study when the number of parameters diverge with sample size.

In our study, we have tried several choices, for more discussion, see Section 3.2.

III. NUMERICAL RESULTS

In this section we look at the oracle properties (see Section 1) of the lasso [21] and adaptive lasso [28]. The theoretical properties of the lasso and adaptive lasso suggest that these methods are consistent in variable selection under some conditions see e.g. Reference [28], [27]. We compare the performance of these two shrinkage methods looking at the following properties:

(1) consistency in variable selection, and
(2) prediction performance.

For (1), we look at the probability of containing the true model on the solution path ($PTSP$) of these shrinkage methods. This measure has been used by [28]. The solution path is the entire set of estimates corresponding to various choices of the tuning parameter. We obtain this solution path using the *lars* package in R. The solution path is said to contain the true model if it results in a correct estimated model (CM) for some choice of the tuning parameter, measure CM is defined more precisely later in this section. We define $PTSP$ as the proportion of times we get the CM out of N Monte Carlo runs. For an oracle performance, $PTSP \xrightarrow{p} 1$ as $n \to \infty$.

Convergence of $PTSP$ to 1 in probability supports theoretical consistent variable selection but to achieve it in practice requires the right choice of the tuning parameter. Selection of the appropriate value of the tuning parameter is very challenging as there is no precise theoretical answer to this question yet. In this study, we compare two methods, k-fold cross-validation and the BIC, in their selection of the value of the tuning parameter. We define two measures we will use to assess and compare the tuning parameters selectors' performance.

Model size (MS)

As we have defined earlier, model size, in the linear regression context, is the number of non-zero components in the model. For the simplicity of presentation, we assume that model (1.1) has $p_0 < p$, say, non-zero components i.e. $\{\beta_j \neq 0 : j \in \mathcal{A}\}$ then $|\mathcal{A}| = p_0$ while $|\mathcal{S}_F| = p$, where $\mathcal{A}$ and $|\mathcal{S}_F|$ are model size for true model and full model respectively. An oracle procedure, say μ, should have the model size $|\mathcal{S}_\mu| = |\mathcal{A}| = p_0$. Thus this measure guarantees that the prediction procedure is shrinking exactly the same number of estimates to zero as in the true model. In our results, we present the median MS (MMS) for the prediction procedure resulting from the M replicates. For an oracle procedure $MMS \xrightarrow{p} p_0$.

Correct model (CM)

The correct model is the measure we use to determine if the procedure is correctly shrinking the zero and non-zero components of the model. For oracle performance, the estimated model should have $\{\hat{\beta}_j = 0$ for $j \in \mathcal{A}\}$ and $\{\hat{\beta}_j \neq 0$ for $j \in \mathcal{A}^c\}$ i.e.

$$CM = \{\hat{\beta}_j = 0 : j \in \mathcal{A}, \hat{\beta}_j \neq 0 : j \in \mathcal{A}^c\}. \qquad \text{(III.1)}$$

In our Monte Carlo study for each of these two methods, we compute and compare the percent of correct models (PCM). For an oracle procedure $MMS \xrightarrow{p} p_0$ and $PCM \xrightarrow{p} 100$. The measures MMS and PCM are also used by [23].

For (2), we compute the median of relative model error ($MRME$) of the lasso and adaptive lasso estimates, when the tuning parameter is selected by k-fold cross-validation and the BIC. The measure $MRME$ is used by [24]. We define the measure $MRME$ as follows.

Median of relative model error ($MRME$)

As defined in [8], if $\{(\boldsymbol{x}_i, y_i) : i = 1, \ldots, n\}$ are assumed to be a random sample from the distribution $(\boldsymbol{X}, \boldsymbol{y})$. For the model (1.1), the model error can be defined as

$$ME(\hat{\mu}) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T E(\boldsymbol{x}\boldsymbol{x}^T)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}), \qquad \text{(III.2)}$$

where $\hat{\boldsymbol{\beta}}$ are the estimates used in the prediction procedure $\hat{\mu}(\boldsymbol{x})$. Now we can define the relative model error as the ratio of the model error for any prediction procedure $\hat{\mu}(\boldsymbol{x})$ to the model error for least squares. The median of the relative model error ($MRME$) for N Monte Carlo runs is obtained to assess the average lack of fit in the prediction procedure.

Ideally, a model should have a low $MRME$. In order to have a standard reference for the comparison, we define the oracle relative model error ($ORME$) as a ratio of oracle model error, where we have knowledge of the zero components of the model and the non-zero components have been replaced by the least square estimates, to the model error of least squares estimates. The $MRME$ for each model was compared to the $ORME$ and the model with $MRME$ closest to the $ORME$ is considered as the best prediction procedure.

We study the following three examples:

Model 0:

Suppose $p = 4$ and $\beta_0 = (5.6, 5.6, 5.6, 0)^T$, we consider this example to observe the effect on the lasso and adaptive lasso consistency in variable selection when the ZYZ condition does not hold. Using the partitioning of $\boldsymbol{C}$ defined in (1.2), we consider $\boldsymbol{C}_{11} = (1 - \rho_1)\,\boldsymbol{I} + \rho_1 \boldsymbol{J}_1$, where $\boldsymbol{I}$ is the identity matrix, $\boldsymbol{J}_1$ is the matrix of 1's and $\boldsymbol{C}_{12} = \rho_2 \mathbf{1}$, where $\mathbf{1}$ is the vector of 1's. In this model, we chose $\rho_1 = -0.39$ and $\rho_2 = 0.23$. This model is the same as that studied in [28] to illustrate the inconsistent lasso solution path.

Model 1:
Suppose $\beta_0 = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$ and $\boldsymbol{C} = \left\{(0.5^{|i-j|}); i, j = 1, \ldots, p\right\}$. The ZYZ condition holds for this choice of $\boldsymbol{C}$. This model was also studied by [8], [28] and [11].

Model 2:
Suppose $\beta_0 = (0.85, 0.85, 0.85, 0)^T$ and $\boldsymbol{C}$ is the same as for Model 0. We have considered this example to compare with the results obtained in Model 0, where we have relatively large effects.

For all of the three examples, we designed a Monte Carlo study of 100 runs. For each Monte Carlo run, we simulate the linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon$ for the fixed set of parameters given above, where $\boldsymbol{X} \sim N_p(\boldsymbol{0}, \boldsymbol{C})$. In the Gaussian case, $\varepsilon_i \sim N(0, \sigma^2)$, we have considered the choices $\sigma = 1, 3, 6$ and 9.

In the next section, we will see if the numerical results support the conclusion that the lasso and adaptive lasso are consistent in variable selection. We will give results for *PTSP* to compare variable selection done by these lasso-type methods, without involving tuning parameter selection. In the second part of the next section, we will give results for *PCM*, *MMS* and *MRME*, which are obtained after selecting the tuning parameter. We will use k-fold cross-validation and BIC for the selection of tuning parameter. These results will also throw some light on how possible it is, in practice, to achieve these oracle properties.

A. *VARIABLE SELECTION*

To be consistent in variable selection is an important property of the shrinkage methods. The consistency or otherwise of the lasso selection depends on some model features e.g. the ZYZ condition (1.3).

In this section, we give results for the probability that it contains the true model.Now in the rest of this section, we will give results on the basis of 100 Monte Carlo runs and will look at some empirical results for the performance measures defined earlier at the start of this section.

We now consider a selection of sample sizes ranging from $n = 50$ to $n = 50000$ to study the performance of these methods for small sizes and also for their asymptotic behaviour. We assume $\varepsilon_i \sim N(0, \sigma^2)$, where $\sigma = 1$,3,6 and 9 are the choices of error standard deviation.

Figure 1 gives the plots for the lasso and adaptive lasso showing the empirical probability of containing the true model for each of the three models defined earlier. In these plots, the horizontal axis corresponds to sample size on a logarithmic scale and the vertical axis corresponds to the empirical probability that the true model lies on the solution path.

Figure 1(a) shows the empirical probability of containing the true model for the standard lasso, which shows that the lasso cannot be consistent in variable selection for Model 0 as the ZYZ condition fails for this model. We can see that this probability varies between 0.4 and 0.6 and does not converge to 1 even for sample sizes as large as $n = 50000$. The results do not differ much for different choices of error variance.

For the adaptive lasso, Figures 1(b)-(d), show that the probability is converging to 1 and the larger the value of γ, the smaller the sample size is required to be to get the probability exactly one. This shows that the adaptive lasso can be consistent in variable selection if an appropriate value of the tuning parameter is selected. However, the result that the adaptive lasso is doing well for larger values of γ should be interpreted with caution. We have noticed that with an increase in γ, the range of values of s which correspond to the true model decreases thereby making it harder for the tuning parameter selector to pick an appropriate value of the tuning parameter. We will discuss this in detail later in this section.

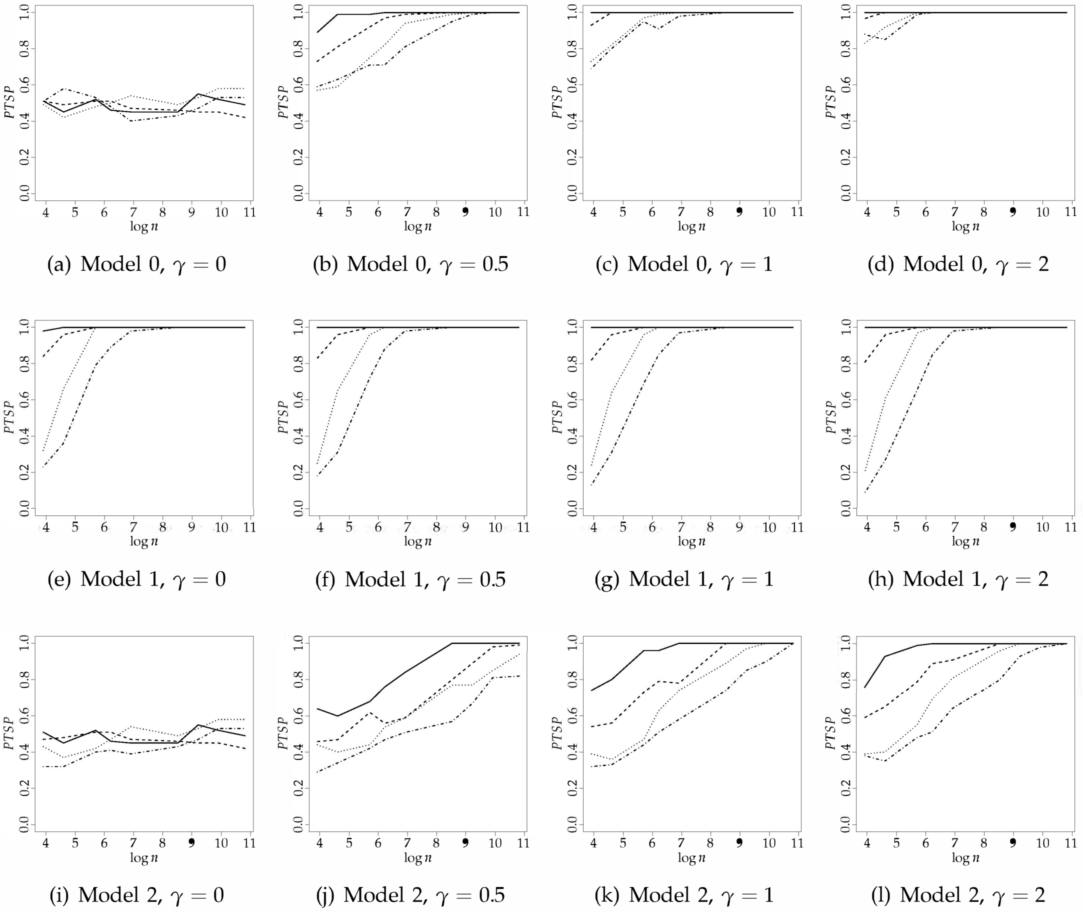
In the case of Model 1, the lasso and adaptive lasso for all choices of γ does not differ much and the probability for all of them is converging to one. These results suggest that it is sometimes easier to select the correct value of the tuning parameter for the lasso as compared to the adaptive lasso.

For Model 2, we have small effects and the ZYZ condition also fails. It is obvious that this situation becomes more challenging. Now it can be seen from the results shown in Figure 1(j),(k), that, in general, the probability for the adaptive lasso when $\gamma = 0.5$ and 1 converges to one at a rate slower than in the case of Model 0. But the results for the adaptive lasso when $\gamma = 2$ do not differ much for the two models.

In the next section, we will compare the tuning parameter selectors and will also see if oracle properties of lasso-type methods can be achieved in practice.

B. *ESTIMATION OF TUNING PAREMETER*

As we have discussed earlier in Section 3.1, when the ZYZ condition fails, the lasso is not consistent in variable selection but the adaptive lasso is. In cases where the ZYZ condition holds, the lasso and adaptive lasso theoretically do consistent variable selection but



(a) Model 0, $\gamma = 0$ (b) Model 0, $\gamma = 0.5$ (c) Model 0, $\gamma = 1$ (d) Model 0, $\gamma = 2$

(e) Model 1, $\gamma = 0$ (f) Model 1, $\gamma = 0.5$ (g) Model 1, $\gamma = 1$ (h) Model 1, $\gamma = 2$

(i) Model 2, $\gamma = 0$ (j) Model 2, $\gamma = 0.5$ (k) Model 2, $\gamma = 1$ (l) Model 2, $\gamma = 2$

Fig. 1: Probability, based on 100 runs, that solution paths of the lasso ($\gamma = 0$) and adaptive lasso ($\gamma = 0.5, 1, \text{and} 2$) for the three models defined in Section 3. Model 0: $\beta_0 = (5.6, 5.6, 5.6, 0)^T$. Model 1: $\beta_0 = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$. Model 2: $\beta_0 = (0.85, 0.85, 0.85, 0)^T$. The error distribution is $\varepsilon_i \sim N(0, \sigma^2)$ where (——— $\sigma = 1$); (− − − $\sigma = 3$); (······ $\sigma = 6$); (− · − · − · $\sigma = 9$).

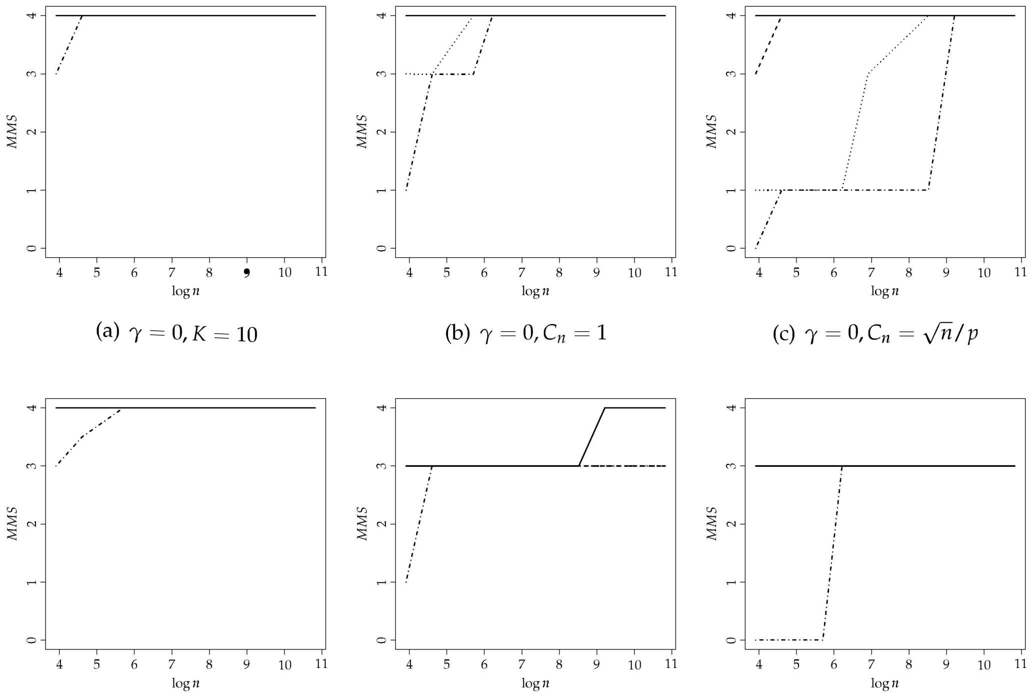
consistency can only be achieved if we have a method which can select an appropriate value of the tuning parameter. If the tuning parameter is not appropriately selected, even though the solution path contains the true model, it is likely we will select an incorrect model. Reference [21] also noted in a simulation example that though the lasso solution path contains the true model, only for a small fraction of possible choices of tuning parameter $s \in [a, b] \subset [0, 1]$ the lasso does pick the correct model. So it becomes very challenging for any tuning parameter selector to pick the right choice of tuning parameter that corresponds to the oracle variable selection.

The discussion above shows the importance of tuning parameter selection. In this section, we compare two methods used for tuning parameter selection: (1) k-fold cross-validation and (2) the Bayesian information criterion (BIC). These methods are defined in Section 2. We use 5-fold and 10-fold cross-validation as tuning parameter selector but, to save the space, the results only for 10-fold cross validation are shown as it performed well.

For the BIC, defined in (2.1), we have used several values for C_n, e.g. $C_n = 1, 5$, and 10. We noticed in our numerical study that all of these considered choices of C_n fail to work as n increases. This may be due to failure of [23], condition 4, given below:

$$\sqrt{\frac{n}{C_n p \log(n)}} \lim_{n\to\infty} \left(min_{j\in\mathcal{A}}|\beta_{0,j}|\right) \to \infty, \quad \text{and } C_n p \log(n)/n \to 0. \tag{III.3}$$

We also observe from our numerical results that each of the considered fixed choice of C_n leads to consistent variable selection up to a certain sample size, say n_0, and then variable selection becomes inconsistent for $n > n_0$. We notice that, for oracle variable selection, the larger the sample size the larger the value of C_n is required and vice versa.



(a) $\gamma = 0, K = 10$ (b) $\gamma = 0, C_n = 1$ (c) $\gamma = 0, C_n = \sqrt{n}/p$

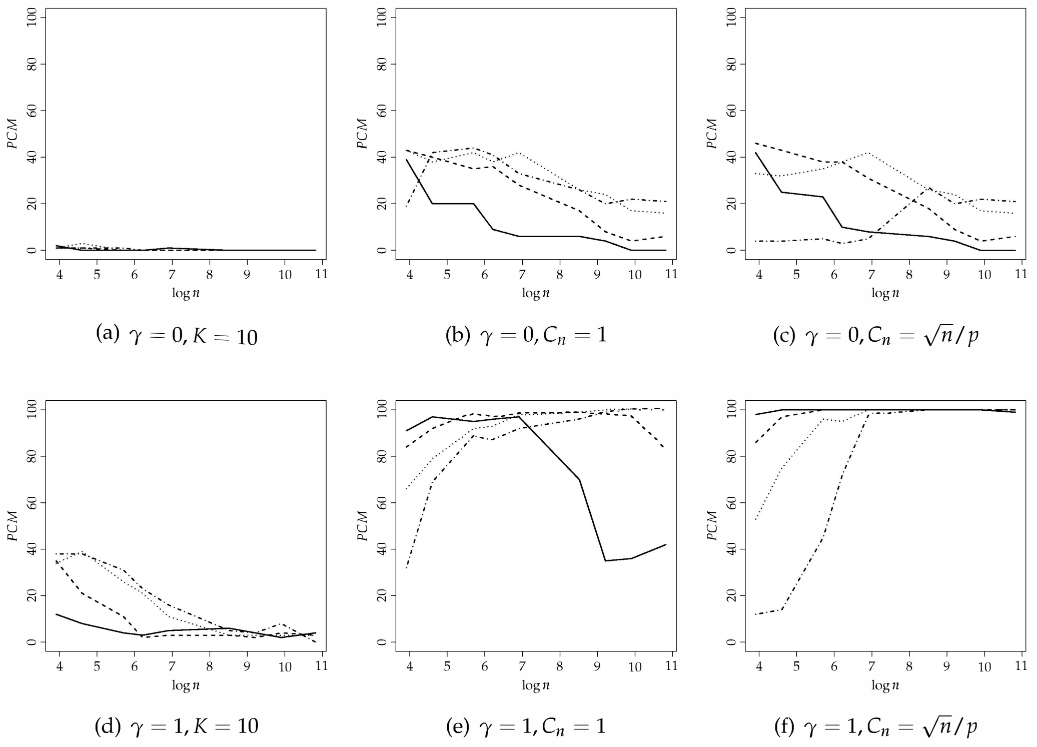
(d) $\gamma = 1, K = 10$ (e) $\gamma = 1, C_n = 1$ (f) $\gamma = 1, C_n = \sqrt{n}/p$

Fig. 2: MMS: Median model size, based on 100 Monte Carlo runs, for the lasso ($\gamma = 0$) and adaptive lasso ($\gamma = 1$). Tuning parameter is selected using $10 - fold$ cross-validation and BIC ($C_n = 1$ and $C_n = \sqrt{n}/p$) for Model-0 defined in Section 3. Model 0: $\beta_0 = (5.6, 5.6, 5.6, 0)^T$. The error distribution is $\varepsilon_i \sim N(0, \sigma^2)$ where (——— $\sigma = 1$); (− − − $\sigma = 3$); (· · · · · · $\sigma = 6$); (− · − · −· $\sigma = 9$).

These results lead us to the conclusion that the performance of the BIC approach is highly dependent on the value of C_n and we need a value of C_n which increases at a certain rate with n. The results for these fixed values of C_n intuitively guided us to the use of $C_n = \sqrt{n}/p$, where n is the sample size and p is the number of predictors. In the rest of the section, we give the results for this choice of C_n.

From now onwards we will show and discuss the results on the performance of tuning parameter selectors for Model 0 only given in Section 3. We will focus on the cases shown in Figure 1(a) and (c). As these are the cases, in which for the lasso we have a moderate probability for containing the true model on the solution path and a probability approaching to one in the case of the adaptive lasso. Thus, we expect the behaviour of performance measures, MMS and PCM stated in Section 3, parallel to PTSP.

As we have discussed earlier, smaller values of the tuning parameter, s, lead to a greater amount of shrinkage and this results in underfitted models. To see if the tuning parameter selector is shrinking the right number of estimates towards zero, we look at the median model size (MMS) for the value of s chosen by each of the tuning parameter selectors viz cross-validation and the BIC. Figure 2 shows the median model size for the lasso and adaptive lasso obtained for the choices of tuning parameters selected by cross validation and the BIC. For this model, we have The oracle model size is $p_0 = 3$. We can see that for the lasso, both cross validation and the BIC, the $MMS = 4$. Thus, in general, resulting in overfitted models. Figure 3 show the plots of percentage of correct model identified. We give the results for the lasso $(\gamma = 0)$ and adaptive lasso $(\gamma = 1)$ when the tuning parameter is selected by 5-fold cross-validation and the BIC $(C_n = \sqrt{n}/p)$. The horizontal axis corresponds to the sample size on a logarithmic scale and the vertical axis corresponds to the percent of correct models. Ideally, these plots should match with the corresponding plots of probability of containing the true model on the solution path shown in Figure 1. For example, Figure 1(c) shows that for the adaptive lasso $(\gamma = 1)$, the probability of containing the true model on the solution path converges to one for each choice of the error variance. Now if we compare this with Figure 3(c), which shows the percentage of correct models identified using cross-validation, we can see this percentage is very low and even decreases to zero



(a) $\gamma = 0, K = 10$

(b) $\gamma = 0, C_n = 1$

(c) $\gamma = 0, C_n = \sqrt{n}/p$

(d) $\gamma = 1, K = 10$

(e) $\gamma = 1, C_n = 1$

(f) $\gamma = 1, C_n = \sqrt{n}/p$

Fig. 3: PCM: Percent correct models, based on 100 Monte Carlo runs, for the lasso ($\gamma = 0$) and adaptive lasso ($\gamma = 1$). Tuning parameter is selected using $10 - fold$ cross-validation and BIC ($C_n = 1$ and $C_n = \sqrt{n}/p$) for Model-0 defined in Section 3. Model 0: $\beta_0 = (5.6, 5.6, 5.6, 0)^T$. The error distribution is $\varepsilon_i \sim N(0, \sigma^2)$ where (——— $\sigma = 1$); (– – – $\sigma = 3$); (······ $\sigma = 6$); (– · – · –· $\sigma = 9$).

as sample size increases. In contrast, a comparison of Figure 1(c) with Figure 3(d) shows that for the BIC we can do consistent variable selection with probability approaching to 1.

It is found that cross-validation fails to select the appropriate value of the tuning parameter thus resulting in the selection of an incorrect model from the lasso and adaptive lasso solution path. Our results show that the oracle property of consistent variable selection can be achieved for the lasso if the ZYZ condition holds, while the adaptive lasso can do the consistent variable selection even if the ZYZ condition does not hold in the standard lasso. We also found that an appropriate value of the tuning parameter can be selected if a tuning parameter selector based on the BIC is used.

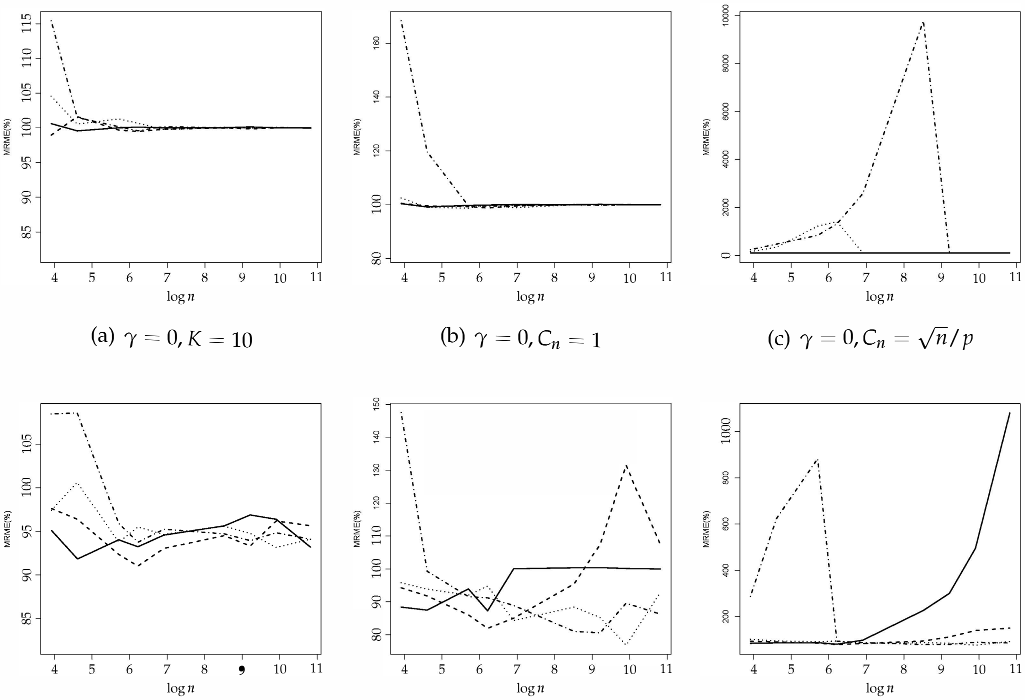
In the following paragraphs we will give some results on the performance measure Median of Relative Model error (MRME) defined earlier at the start of Section 3 along with the definition of Model error (ME) given in (3.2).

Figure 4 gives the plots of MRME for the lasso and adaptive lasso. These plots are for the models corresponding to the value of the tuning parameter selected by cross-validation and the BIC. It can be noticed that MRME is higher in the cases where we achieve consistent variable selection e.g. in the case of the adaptive lasso when tuning parameter is selected using BIC with $C_n = \sqrt{n}/k$. As we have seen in Figure 2, that in all the cases, except Figure 2(f), the estimated model size is greater than the true model size, i.e. $p_0 = 3$. So our results show that shrinkage can lead to increased relative model error at least in the situations considered in this paper.

IV. CONCLUSION

The ZYZ condition is an important condition for consistent variable selection for the lasso and adaptive lasso. The lasso can be consistent in variable selection when the ZYZ condition holds provided that an appropriate value of the tuning parameter is selected. It should be noted that the ZYZ condition always holds for the adaptive lasso due to the use of adaptive weights and thus it showed consistent variable selection in all the cases.

Our numerical results suggest that cross-validation is not a reliable method especially if the primary objective is variable selection. In all situations considered, cross validation as tuning parameter selector leads the lasso



(a) $\gamma = 0, K = 10$ (b) $\gamma = 0, C_n = 1$ (c) $\gamma = 0, C_n = \sqrt{n}/p$

(d) $\gamma = 1, K = 10$ (e) $\gamma = 1, C_n = 1$ (f) $\gamma = 1, C_n = \sqrt{n}/p$

Fig. 4: MRME: Median model size, based on 100 Monte Carlo runs, for the lasso ($\gamma = 0$) and adaptive lasso ($\gamma = 1, 2$). Tuning parameter is selected using $10 - fold$ cross-validation and BIC ($C_n = 1$ and $C_n = \sqrt{n}/p$) for Model-0 defined in Section 3. Model 0: $\beta_0 = (5.6, 5.6, 5.6, 0)^T$. The error distribution is $\varepsilon_i \sim t_\nu$. The error distribution is $\varepsilon_i \sim N(0, \sigma^2)$ where (——— $\sigma = 1$); (– – – $\sigma = 3$); ($\cdot \cdots\cdots$ $\sigma = 6$); ($- \cdot - \cdot - \cdot$ $\sigma = 9$).

and adaptive lasso to inconsistent variable selection. In contrast, the BIC with our suggested value of C_n has shown its capability to choose a value for the tuning parameter which correctly shrinks the coefficients of non-active predictors to zero. Though in some cases especially for very large sample sizes, this consistent variable selection results in an increased relative model error, but the results are comparable for small and moderate choices of sample size.

REFERENCES

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

[2] L. Breiman. Better subset selection using the non-negative garotte. *Technometrics*, 37(4):373–384, 1995.

[3] J.M. Brown. *Measurement, Regression and Calibration*. Oxford University Press: Oxford, UK., 1993.

[4] E. Candes and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, 35(6):2313–2351, 2007.

[5] S. Chand. Diagnsotic checking and lasso variable selection in time series analysis. *PhD Thesis, School of Mathematical Sciences, The University of Nottingham, United Kingdom*, 2011.

[6] S. Chand and S. Kamal. Variable selection by lasso-type methods. *Pakistan Journal of Statistics and Operation Research*, 7(2, Special issue on variable selection):451–464, 2011.

[7] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.

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

[9] J. Fan and J. Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

[10] J. Fan and J. Lv. Non-concave penalized likelihood with NP-dimensionality. *Arxiv preprint arXiv:0910.1119*, 2009.

[11] P. Hall, E.R. Lee, and B.U. Park. Bootstrap-based penalty choice for the lasso achieving oracle performance. *Statistica Sinica*, 19:449–471, 2009.

[12] T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, J. Friedman, and R. Tibshirani. *The Elements of Statistical Learning*. Springer-Verlag: New York, USA., 2001.

[13] T. Hesterberg, N. Choi, L. Meier, and C. Fraley. Least angle and L1 penalized regression: A review. *Statistics Surveys*, 2:61–93, 2008.

[14] N.J. Hsu, H.L. Hung, and Y.M. Chang. Subset selection for vector autoregressive processes using Lasso. *Computational Statistics and Data Analysis*, 52(7):3645–3657, 2008.

[15] G.M. James, P. Radchenko, and J. Lv. DASSO: Connections between the dantzig selector and lasso. *Journal of Royal Statistical Society, Series B*, 71(1):121–142, 2009.

[16] K. Knight and W. Fu. Asymptotics for lasso-type estimators. *Annals of Statistics*, 28(5):1356–1378, 2000.

[17] C. Leng, Y. Lin, and G. Wahba. A note on the lasso and related procedures in model selection. *Statistica Sinica*, 16(4):1273–1284, 2006.

[18] J. Lv and Y. Fan. A unified approach to model selection and sparse recovery using regularized least squares. *Annals of Statistics*, 37(6A):3498–3528, 2009.

[19] P. Radchenko and G.M. James. Variable inclusion and shrinkage algorithms. *Journal of the American Statistical Association*, 103(483):1304–1315, 2008.

[20] G. Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[21] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society, Series B*, 58(1):267–288, 1996.

[22] H. Wang and C. Leng. Unified lasso estimation via least squares approximation. *Journal of the American Statistical Association*, 102(479):1039–1048, 2007.

[23] H. Wang and C. Leng. Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of Royal Statistical Society, Series B*, 71(3):671–683, 2009.

[24] H. Wang, R. Li, and C.L. Tsai. Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94(3):553–568, 2007.

[25] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[26] M. Zhao and KB Kulasekera. Consistent linear model selection. *Statistics & Probability Letters*, 76(5):520–530, 2006.

[27] P. Zhao and B. Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

[28] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

[29] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B*, 67(2):301–320, 2005.