A Cross-Validation Method for Linear Regression Model Selection

Jingwei Xiong Junfeng Shang
Department of Mathematics and Statistics
Bowling Green State University
Bowling Green, OH 43403, USA

Abstract: In linear regression model setting, motivated by Wasserman and Roeder (2009), we develop a cross-validation procedure for selecting an appropriate model which can best fit the data. In the procedure, we make use of adaptive Lasso method to select the most appropriate model. In the selection of the suitable tuning parameter, the Bayesian Information Criterion (BIC, Schwarz, 1978) is utilized. We conduct the hypothesis testings for the significance of nonzero coefficients of fixed effects to further select the model. A simulation study investigates the effectiveness of performance for the proposed procedure. The simulation results demonstrate that BIC and the adaptive Lasso method can both lower Type I error and false positive rate; they can also improve the test power and the rate of selecting an overfitted model.

Key words: Model selection, adaptive Lasso, linear regression, cross-validation, tuning parameter

1 Introduction

Model selection is to select a subset of candidate variables that will contribute to the prediction of the response variable. The traditional subset method and ridge regression have drawbacks. For the subset method, because we add or drop one variable in the model, the procedure is discontinuous. While losing the unbiasedness, the ridge regression may decrease the mean square error (MSE). Further, the predictors whose coefficients are close to zero are still stand in the model, making the model quite complicated and then it is not easily interpreted.

To avoid the problems in traditional methods, penalized model selection in linear regression has caught enough attention. Following the Lasso method, by Tibshirani in 1996, the SCAD (Fan and Li, 2001) and the adaptive Lasso (Zou, 2006) modified the Lasso L-1 penalty term by using adaptive weights, to give consistent estimator of non-zero sets under mild regularity conditions. Candes and Tao (2007) modified the penalty term and developed the Dantzig selector to achieve the same goal. The sparsity pattern lying in these approaches makes the selected model simpler.

A number of other papers also tackle the model selection problem using the penalized method. In this pool, they are Meinshausen and Bühlmann (2006), Wainwright (2006), Zhao and Yu (2006), Fan and Lv (2008), Meinshausen and Yu (2009), Tropp (2004, 2006), Donoho (2006) and Zhang and Huang (2006).

We wish to employ a penalized method to conduct model selection for achieving a better selection result. The paper of Wasserman and Roeder (2009) has investigated a penalized method for high-dimension variable selection, which inspires us to improve its method. We therefore develop a cross-validation procedure to select nonzero effects in linear regression models, whose goal is to obtain a better model selection result using cross-validation, adaptive Lasso and BIC.

In what follows, the model and assumptions are first introduced. The procedure is depicted in Section 3. Section 4 presents a simulation study. The last section concludes and discusses.

2 The model and assumptions

The fundamental settings are first presented and then the assumptions are interpreted.

2.1 The model

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d observations from the regression model

$$Y_i = X_i^T \beta + \varepsilon_i, \ \ i = 1, \cdots, n, \tag{2.1}$$

where $\varepsilon_i \sim N(0, \sigma^2), X_i = (X_{i1}, \ldots, X_{ip})^T \in R^p$. Here we have n observations and p potential factors affecting the response variable. Let X represent the design matrix for the model and we denote the jth column of the design matrix using $X_{\bullet}j = (X_{1j}, \ldots, X_{nj})^T$, and the response vector Y is denoted by $Y = (Y_1, \ldots, Y_n)^T$. Let $D = \{j : \beta_j \neq 0\}$ represent the set of covariates with nonzero coefficients. We assume that X and Y are given and known prior to analysis, but D is unknown.

2.2 The assumptions

Throughout the paper, we have the following assumptions:

(A1) $Y_i = X_i^T \beta + \varepsilon_i$ where $\varepsilon_i \sim N(0, \sigma^2)$, for $i = 1, \ldots, n$.

(A2) The covariates are standardized: $E(X_{ij}) = 0$ and $E(X_{ij}^2) = 1$. Also there exists $0 < B < \infty$ such that $P(|X_{jk}| \leq B) = 1$.

(A1) is required since we are dealing the problems under the normal settings.

(A2) means the data are standardized before proceeding to analysis.

We also require that (1) the number of columns of design matrix X is not far beyond the number of rows; (2) the number of non-zero covariates is bounded by a constant regardless of n, and at least we have one non-zero covariates. (3) We require that the largest eigenvalue of p by p matrix $\frac{1}{n} X^T X$, when n approaches to infinity, is almost bounded. The smallest eigenvalue of $C_1 \log n$ by $C_1 \log n$ matrix of $\frac{1}{n} X_k^T X_k$ is almost greater than a positive value for some $C_1 > 0$. Here X_k denotes a design matrix with k predictors in the model. This condition is added due to some proof requirements.

3 Cross-validation procedure

3.1 The procedure

First, we split the data into three parts, D_1, D_2, D_3. The D_1 is to select the most appropriate model through a penalized method, Lasso or adaptive Lasso; the D_2 is to choose the best tuning parameter for D_1 by means of MSE or BIC; that is, the D_1 and D_2 are utilized in one iteration, and the determination of the tuning parameter will affect the selection of final model. The D_3 is to refine the selected model from D_1 and D_2 by removing the insignificant covariates. Theoretically, the D_3 may not be needed, yet it will fortify the selected model.

For the first part D_1, we fit a series of candidate models, and each model depends on a tuning parameter λ. The whole set of candidate models is denoted as $S = \{\widehat{S}_n(\lambda) : \lambda \in \Lambda\}$. Let us call the best candidate model $\widehat{S}_n$, and it can be shown that $P(D \subset \widehat{S}_n) \to 1$ and $|\widehat{S}_n| = o_P(n)$, where $|\widehat{S}_n|$ represents the number of elements in the set S_n. That means the candidate models are overfitted ones which contain the unknown true model and therefore correspond to a sequence of λ values which can determine such overfitted models. This fact will also be discussed afterwards. We address that this fact that D_1 will choose the overfitted models provides the foundation for the entire procedure and guarantees the validity of tests for D_3 and the test power will not be too small. Moreover, it shows the selected power approaches to 1 as the sample size n goes to infinity, thus we can almost reject all the non-zero coefficients when the data size is large enough.

A penalized method will be applied to D_1 for selecting the overfitted models. Of course, a large sample size will result in a large probability of containing the true model. The Lasso penalized target function to be utilized in D_1 is defined as:

$$\sum_{i=1}^{n} (Y_i - X_i^T \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{3.1}$$

where Y_i, X_i are the corresponding vector and matrix from D_1. To minimize this function, the Lasso estimator of the β will be obtained, and the model will therefore be selected. In the selection procedure of using D_1, all the non-zero coefficients are included in the selected models.

We adopt the adaptive Lasso penalized method and it follows a quite similar way as the function in expression 3.1, and has been changed to:

$$\sum_{i=1}^{n}(Y_i - X_i^T\beta)^2 + \lambda\sum_{j=1}^{p} w_j|\beta_j|, \tag{3.2}$$

where $w_j = \frac{1}{|\beta_j|}$. Firstly, we compute the β_{ols} from D_1, and take their reciprocals to have the weight for each β_j as starting values. Analogously, minimizing this function will result in the adaptive Lasso estimator of the β. For each λ value, we will obtain a different selected model.

For functions (3.1) or (3.2), we assign multiple different values of λ in a reasonable interval, and each λ outputs a selected model and then the selected model outputs a criterion. The best model is selected corresponding to the smallest criterion, as shown in what follows. With respect to the reasonable interval of λ, it can be investigated in the simulations and then is determined.

For the second part D_2, the target is to select the best λ. The first selection criterion is the mean square error (MSE) and is also called the loss.

Generally, for an estimator, we want to measure how close it was to the true parameter, thus the concept of loss was brought in. The loss of any estimator $\widehat{\beta}$ is defined as:

$$L(\widehat{\beta}) = \frac{1}{n}(\widehat{\beta} - \beta)^T X^T X(\widehat{\beta} - \beta).$$

The loss measures the "distance" between the statistic and the true parameter, and a smaller loss means our estimate statistic is closer to the parameter. Therefore, we want to have a $\widehat{\beta}$, which can minimize the loss. Meanwhile we can see that each tuning parameter λ exports a $\widehat{\beta}$, and $\widehat{\beta}$ determines the loss, so the loss can be treated as a function of λ. We will denote the loss as $L(\lambda)$. Again and more importantly, we have $P(D \subset \widehat{S}_n(\widehat{\lambda})) \to 1$ where $\widehat{\lambda} = argmin_{\lambda\in\Lambda_n} L(\lambda)$, indicating that by the loss function, the overfit or overspecified models are selected.

However, due to the unknown β, we plug in the $\widehat{\beta}(\lambda)$ to a substitute estimable formula for loss, we then have

$$\widehat{L}(\lambda) = \frac{1}{n}\sum_{X_i\in D_2}(Y_i - X_i^T\widehat{\beta}(\lambda))^2.$$

Corresponding to the "best" selected λ, the estimated loss $\widehat{L}(\lambda)$ is minimized. Note that the estimated loss $\widehat{L}(\lambda)$ behaves similarly as the true loss. This optimal property is shown by Theorem 3.2 in Wasserman and Roeder (2009). Suppose that the $max_{\lambda\in\lambda_n}|\widehat{S_n}(\lambda)| \leq k_n$. Then there exists a sequence of random variables $\delta_n = O_P(1)$ that do not depend on λ or X, such that with probability tending to 1,

$$\sup_{\lambda\in\Lambda_n}|L(\lambda) - \widehat{L}(\lambda) - \delta_n| = O_P(\frac{k_n}{n^{1-c_2}}) + O_P(\frac{k_n}{\sqrt{n}})$$

Note that if k_n is a good sequence which makes $\frac{k_n}{n^{1-c_2}}, \frac{k_n}{\sqrt{n}}$ converge to 0, the difference between $\widehat{L}(\lambda)$ and $L(\lambda)$ can be approximated by a random variable δ, which is bounded in probability. Based on this theorem, the estimated loss $\widehat{L}(\lambda)$ performs as the true loss. For the true loss, the ovefitted models are selected, which is identical to the selecting procedure of D_1 in that the penalized method will also choose the overfitted models which include the true parameters or true model.

Now we remark on the procedure using D_1 and D_2. In D_1, the penalized Lasso method will choose the overfitted models; in D_2, the loss function MSE will also choose the best λ which agrees with an overfitted model. So it is convincing to develop the cross-validation for D_1 and D_2, and we select the λ which minimizes the $\widehat{L}(\lambda)$ for the data D_2. After the cross-validation procedure using D_1 and D_2, the "best" model which is overfitted will be selected.

The criterion we take for selecting the best λ is BIC, and it is written as

$$\text{BIC} = n\log\hat{\sigma}^2 + k\log n,$$

where $\hat{\sigma}^2$ is computed by $\frac{1}{n}(Y - X\widehat{\beta}_{ols})^T(Y - X\widehat{\beta}_{ols})$, and k is the number of the estimated parameters in the corresponding model. Note that the constant in BIC is ignored here. Based on the model selected in D_1 for each λ, we can estimate the least squares estimate of β, $\widehat{\beta}_{ols}$ in D_2, and n is the sample size in D_2, BIC can therefore be calculated in D_2. We select the λ which minimizes the $\widehat{L}(\lambda)$ or BIC for the data D_2. After the best λ is selected by minimizing the criterion MSE or BIC, by functions in (3.1) or (3.2), the best model will be selected.

Note that as a model selection criterion, MSE is efficient yet not consistent; whereas BIC is consistent, which means that BIC will select the true model with probability 1 as the sample size increases to infinity. We therefore expect that BIC can perform better in selecting the tuning parameter λ.

For high-dimensional model selection, it requires the property that $|\widehat{S_n}| = o_p(n)$, which means when n approaches to ∞, the number of selected non-zero regression coefficients is much smaller than n in probability. This property ensures that the subsequent steps of the procedure can function well even in high-dimensional model selection.

We remark that in the previous procedure, we adopt the adaptive Lasso and BIC for the choice of proper model and tuning parameter. Although the method in Wasserman and Roeder (2009) is effective in high-dimensional model selection, the adaptive Lasso can perform well only if the matrix X^TX is of full rank, so we can assume $p < n$ to facilitate the implementation of adaptive Lasso.

For the third part D_3, we will finalize the selected model. In fact, using D_3 is one step which is not required for model selection. However, this step can always improve the rate of selecting the correct model. We thus consider it as an effective supplementary step to the cross-validation procedure.

For the best model from the previous steps, we can find its corresponding least squares estimate $\widehat{\beta}$ by using the data D_3. Then for each coefficient in the $\widehat{\beta}$, we use the t-test to decide which coefficients will be included, and the final non-zero set is given by:

$$\widehat{D}_n = \{j \in \widehat{S}_n : |T_j| > c_n\},$$

where T_j is the t-statistic, $c_n = t_{\alpha/2m,n-m}$ or $c_n = z_{\alpha/2m}$ and $m = |\widehat{S}_n|$.

To be more specific, the estimated covariance matrix can be computed by $\widehat{\sigma}^2(X_M^TX_M)^{-1}$, where $\widehat{\sigma}^2 = \frac{SSE}{n^*-p^*}$. Here, X_M represents the design matrix for the best model, n^*, p^* represent its numbers of rows and columns respectively in D_3. The α is divided by $2m$ due to the Bonferroni correction, and m is the number of t tests.

Theorem 4.1 in Wasserman and Roeder (2009) states $P(D \subset \widehat{S}_n(\widehat{\lambda})) \rightarrow 1$ where $\widehat{\lambda} = argmin_{\lambda\in\Lambda_n}\widehat{L}(\lambda)$. This theorem makes sure that the subset selected by the cross-validation procedure can cover all the non-zero coefficients with probability 1. Consequently, the best model selected from this procedure is overfitting, and the hypothesis testings in D_3 for t-tests are not biased. When the criterion BIC is utilized, as the sample size n goes to infinity, we can have $P(D = \widehat{S}_n(\widehat{\lambda})) \rightarrow 1$ where $\widehat{\lambda} = argmin_{\lambda\in\Lambda_n}\mathrm{BIC}(\lambda)$. As mentioned earlier, because of the consistency of BIC in selecting models, it may be expected that using BIC instead of $\widehat{L}(\lambda)$, the model selection will result in a better implementation.

Theorem 4.2 asserts the consistency of the procedure. In particular, let $\alpha_n \rightarrow 0$, and$\sqrt{n}\alpha_n \rightarrow \infty$, then we have $P(\widehat{D}_n = D) \rightarrow 1$, which indicates that as the sample size n goes to infinity, this cross-validation procedure will choose the true model with probability 1, and we can achieve a consistent estimate of the non-zero subset by utilizing this cross-validation procedure.

3.2 Type I error and test power

The goal of the paper is to derive a procedure $\widehat{D}_n$, such that:

$$\limsup_{n\rightarrow\infty} P(\widehat{D}_n \subset D) \geq 1 - \alpha.$$

The formula above shows that the probability of selecting "false positive" variable is controlled by α. In other words, the probability of rejecting zero covariates is less than α, thus the Type I error is controlled. When n is large enough, the inequality is satisfied for whatever positive α. In a sense, we can let the α approach to 0, ensuring that the selected subset does not involve any zero covariates. Note that as n goes to infinity, the asymptotic results will be used in the derivation.

On the other hand, we wish the procedure can take nontrivial power, thus we still have chance to reject the nonzero covariates. We will compare the rejecting power of this procedure utilizing adaptive Lasso and BIC with that in Wasserman and Roeder (2009) utilizing Lasso and MSE in a simulation study in the next section.

For any test, we have two major concerns, the error and the power. The type I error rate regarding the estimated non-zero subset is defined as:

$$q(\widehat{D}_n) = P(\widehat{D}_n \cap D^c \neq \emptyset).$$

The power is defined as:

$$\pi(\widehat{D}_n) = P(D \subset \widehat{D}_n).$$

We can see the Type I error actually means the probability of rejecting any zero-coefficients. The power actually means the probability of rejecting all the nonzero coefficients. Our target is to keep a low error rate and meanwhile to maximize the power. Regarding the power presented in the tables of next section, we use the average power which is defined as:

$$\pi_{av} = \frac{1}{s} \sum_{j \in D} P(j \in \widehat{D}_n),$$

where s is the number of non-zero coefficients.

4 A simulation study

4.1 The simulation settings

First, the data are generated from the true models in equation (2.1). For the design matrix X, each element x_{ij} is generated from $N(0, 2)$, the normal distribution with mean 0 and variance 2. For the error term ϵ, each element ϵ_j is simulated from $N(0, 1)$. For the regression coefficients vector β, we have two true models.

Model (a): $\beta = \text{rep}(0, 20)$, which contains 20 predictors with zero coefficients.

Model (b): $\beta = \text{c}\ (9 : 1, \text{rep}(0, 11))$, which contains 9 non-zero coefficients and 11 zero coefficients.

For the true models, we have dimension $p = 20$. Utilizing the generated X, β and ϵ, we can compute the response vector Y under the true model.

For each of the two true models, we generate 1000 replicates, and for each replicate, we implement the cross-validation procedure. As the result, we will have the coefficient selection for 1000 replicates and $p = 20$ parameters either zero or nonzero.

Prior to analysis, the covariates are scaled with mean 0 and variance 1. The tests are performed using one third of the data for each of the 3 stages of the procedure. The level α is set as 0.05, and we want to compare the 3 approaches mentioned in stage one and detect which level $\alpha = 0.05$ test gives the greatest power.

As introduced for the procedure, we split the data (one replicate) into 3 parts, D_1, D_2 and D_3.

For D_1, we choose a sequence values of λ from 0.01 to 0.4 (reasonable interval for λ). In fact, there are 70 values with equal width. For each λ, since the Y_i, X_i are known, this is a 20 dimensional function of β, we use the "optim" function in R to minimize the target function (3.1) or (3.2). The problem is that the numerical solution can never be 0 for β_j, thus we set the boundary value as 0.03. If β_j is less than 0.03, then it could be treated as 0. In this way, for each λ, a candidate model is selected from D_1.

For D_2, we use the candidate model from D_1 to compute the criterion (MSE or BIC). Computing MSE or BIC, we record the candidate model corresponding to the minimum MSE or BIC value, which is considered as the best model.

For the model selected from D_2, we compute the β_{ols} regarding D_3, the covariance matrix of $\hat{\beta}$ is $\hat{\sigma}^2(X^T X)^{-1}$. Taking the diagonal elements of the covariance matrix, we have the estimated variance regarding each $\hat{\beta}_j$. Then we compute the t-test statistic of each coefficient by $\frac{\hat{\beta}_j}{sd(\hat{\beta})}$, if it is greater than $T_{\alpha/2m,n-m}$, we reject and consider that coefficient as non-zero in the final model. Note that m is the number of coefficients in the best model from the previous steps. Since $n - m$ is very large, we may approximate the t-value by z-value.

4.2 The simulation results

We repeat the procedure for 1000 times and organize the results for a replicate in the format of a vector from the simulations, and all the results are stored in a 1000×20 matrix. For each replicate, we use an indicated vector (length is 20) to record the

Table 1: Selection results for model (a)

Model (a)	(c)	(d)	(e)
Lasso. MSE	0.0023	0.0360	1.0
Lasso. BIC	0.0017	0.0320	1.0
A.Lasso. MSE	0.0021	0.0340	1.0
A.Lasso. BIC	0.0016	0.0320	1.0

Note: (c) = FPR, (d)=Type I error, (e)=Overspecified.

selection result. For instance, if an indicated vector is (0,0,0,1,0...,0), it suggests only the 4th coefficient are kept in the final model.

Since true model (a) is a null model, the expected vector of estimated parameters is $(0, 0, \cdots, 0)$. If the ith predictor is not selected in the model, the ith element is denoted as 0; otherwise, keep it as 1. We compute the column means, which is the rejection rate of each regression coefficient, which is also the false positive rate. Since there are no non-zero coefficients, the power is always 0. The type I error occurs when a result vector contains not all 0.

For true model (b), the expected vector of estimated parameters is $(1, ...1, 0, 0, \cdots, 0)$, where the first 9 elements are 1 and the rest are 0. Similarly, we can compute the false positive rate of each regression coefficient. The average power is to average over the rejection rates in the first 9 columns, which are the first 9 column means. Type I error appears when the last 11 elements of a result vector are not all 0.

For model (a), Type I error occurs when at least one of the coefficients are not zero, the false positive rate is computed as the rate of 1 occurring in the result matrix. The rate of overspecified models is always 1 and the power does not exist for the model (a). Similarly, we can output everything for the model (b). The results are listed in Tables 1 and 2.

Note that "FPR" represents the false positive rate; "overspecified" column inputs the rate that the selected model contains the true model; "AV-Power" represents the average power for all the t-tests in D_3.

From the numbers in Table 1, we can observe that for the model with all zero coefficients, employing BIC for selecting the appropriate tuning parameter can lower the false positive rate and the probability of Type I error. The adoption of the adaptive Lasso for selecting the appropriate model can decrease the above two rates as well.

Table 2: Selection results for model (b)

(b)	(c)	(d)	(e)	(f)
(1)	0.0027	0.0300	0.9550	0.9950
(2)	0.0027	0.0275	0.9825	0.9980
(3)	0.0016	0.0175	0.9650	0.9960
(4)	0.0020	0.0225	0.9750	0.9972

Note: (b) = Model (b), (c) = FPR, (d)=Type I error, (e)=Overspecified, (f)=AV-Power, (1)= Lasso with MSE, (2)= Lasso with BIC, (3)= adaptive Lasso with MSE, (4)= adaptive Lasso with BIC.

Table 2 shows that BIC and the adaptive Lasso can not only lower the false positive rate and the probability of Type I error, but also increase the rate of selecting a model containing the true model and test powers.

The simulation study shows that BIC and the adaptive Lasso are an optimal choice for the cross-validation method in model selection.

5 Concluding remarks

Based on the method in Wasserman and Roeder (2009), we adopt the adaptive Lasso and BIC for selecting a model and for selecting the tuning parameter respectively to develop a cross-validation model selection procedure in linear regression model setting.

The simulation results demonstrate that BIC and the adaptive Lasso method can both lower Type I error and false positive rate, and meanwhile they can also increase both the test power and the rate of selecting a model containing the true model.

This paper is the launch of further research. Starting from the procedure in this paper, we will extend similar methodology to generalized linear, linear mixed, and generalized linear mixed models with solidified proofs. Of course, the future research is confronted with quite a few challenges. For instance, in the step of selecting a model, handling the random effects using the penalized method is appealing. Regarding random effects selection, we may consider using the EM algorithm. Treating the random effects as unobserved data, we may compute the conditional expectation of likelihood given the random effects, which is considered as E-step. Other than that, different papers propose different penalty terms, which makes the target expectation different, then M-step is to maximize the target function.

The other idea is to estimate the mixed effects and random effects separately. Moreover, we can replace the unknown covariance matrix by some simple matrix, such as $\log n \times I$, which significantly simplify the penalized likelihood.

In addition, we will consider the model selection with missing values. Missing values occur commonly, effectively coping with missing values therefore plays an important role in model selection literature. We may explore imputation and bootstrapping for missing values.

It is expected that the bootstrap method will improve the test power. However, the necessary bootstrapping theories must be steadily justified.

We can also investigate leave-one-out cross-validation to make this procedure more effective.

References

[1] Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, pp. 2313-2351.

[2] Donoho, D. (2006). For most large underdetermined systems of linear equations, the minimal l1-norm near-solution approximates the sparsest near-solution. *Comm. Pure Appl. Math.* **59** 797-829.

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

[4] Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high-dimensional feature space. *J. Roy. Statist. Soc. Ser. B* **70** 849-911.

[5] Meinshausen, N. and Bühlmann, P. (2006). High dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436-1462.

[6] Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations of high dimensional data. *Ann. Statist.* bf 37 246-270.

[7] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 461-464.

[8] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, Series B **58**, 268-288.

[9] Tropp, J. A. (2004). Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. Inform. Theory* **50** 2231-2242.

[10] Tropp, J. A. (2006). Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Trans. Inform. Theory* **52** 1030-1051.

[11] Zhang, C. H. and Huang, J. (2006). Model selection consistency of the lasso in high-dimensional linear regression. *Ann. Statist.* **36** 1567-1594.

[12] Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *J. Mach. Learn. Res.* **7** 2541-2563.

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

[14] Wainwright, M. (2006). Sharp thresholds for high-dimensional and noisy recovery of sparsity. Available at arxiv.org/math.ST/0605740.

[15] Wasserman, L and Roeder, K. (2009). High-Dimensional variable selection. *The Annals of Statistics* **37, No 5A** pp. 2178-2201.