

On the Nonnegative Garrote Estimator¹

Ming Yuan and Yi Lin

(October 24, 2005)

Abstract

We study the nonnegative garrote estimator from three different aspects: computation, consistency and flexibility. We show that the nonnegative garrote estimate has a piecewise linear solution path. Using this fact, we propose an efficient algorithm for computing the whole solution path for the nonnegative garrote estimate. We also show that the nonnegative garrote has the nice property that with probability tending to one, the solution path contains an estimate that correctly identifies the set of important variables and is consistent for the coefficients of the important variables. Such property is valid for another popular variable selection method, LASSO, only under restrictive conditions. We propose a slight modification that retains the attractive properties of the original nonnegative garrote, but is more widely applicable. To demonstrate the flexibility of the proposed estimator, we consider an extension to the nonparametric regression setup. Simulations and a real example show that the proposed method is very competitive in terms of variable selection and estimation accuracy when compared with other variable selection and estimation methods.

Keywords: Nonnegative garrote; path consistency; piecewise linear solution path; LASSO.

¹Ming Yuan is Assistant Professor, School of Industrial and Systems Engineering, Georgia Institute of Technology, 755 Ferst Drive NW, Atlanta, GA 30332 (E-mail: myuan@isye.gatech.edu). Yi Lin is Associate Professor, Department of Statistics, University of Wisconsin, Madison, WI 53706 (E-mail: yilin@stat.wisc.edu). Lin's research was supported in part by National Science Foundation grant DMS-0134987.

1 Introduction

Consider a multiple linear regression problem where we have n observations on a dependent variable Y and p predictors $X = (X_1, X_2, \dots, X_p)$, and

$$Y = X\beta + \epsilon, \tag{1}$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, and $\beta = (\beta_1, \dots, \beta_p)'$. Throughout this paper, we center each input variable so that the observed mean is zero, and scale each predictor so that the sample standard deviation is one. The underlying notion behind variable selection is that some of the predictors are redundant and therefore only an unknown subset of the β coefficients are nonzero. By effectively identifying the subset of important predictors, variable selection can improve estimation accuracy and enhance model interpretability.

Classical variable selection methods, such as C_p , AIC, and BIC, choose among possible models using penalized sum of squares criteria, with the penalty being an increasing function of the model dimension. These methods, however, are computationally infeasible for even moderate number of predictors since the number of candidate models increases exponentially as the number of predictors increases. In practice, this type of method is implemented in stepwise fashion, through forward selection or backward elimination. Because of the myopic nature of the stepwise algorithm, these implementations are known to be suboptimal in many applications (Chen, Donoho and Saunders, 1998). A number of other variable selection methods have been introduced in recent years (George and McCulloch, 1993; Foster and George, 1994; Breiman, 1995; Tibshirani, 1996; George and Foster, 2000; Fan and Li, 2001; Shen and Ye, 2002; Efron, Johnstone, Hastie and Tibshirani, 2004; Yuan and Lin, 2003; and Zou and Hastie, 2005). In particular, Breiman (1995, 1996) proposed the nonnegative garrote estimator, which he showed to be a stable variable selection method that often outperforms its competitors including subset regression and ridge regression.

The original nonnegative garrote estimator introduced by Breiman (1995) is a scaled version of the least square estimate. The shrinking factor $d(\lambda) = (d_1(\lambda), \dots, d_p(\lambda))'$

is given as the minimizer to

$$\frac{1}{2} \|Y - Zd\|^2 + n\lambda \sum_{j=1}^p d_j, \quad \text{subject to } d_j > 0, \forall j, \quad (2)$$

where $Z = (Z_1, \dots, Z_p)$, $Z_j = X_j \hat{\beta}_j^{\text{LS}}$ and $\hat{\beta}_j^{\text{LS}}$ is the least square estimate based on (1). Here $\lambda > 0$ is a tuning parameter. The nonnegative garrote estimate of the regression coefficient is subsequently defined as $\hat{\beta}_j^{\text{NG}}(\lambda) = d_j(\lambda) \hat{\beta}_j^{\text{LS}}$, $j = 1, \dots, p$. Hereafter, we omit subscript or/and superscript n if no confusion occurs.

The mechanism of the nonnegative garrote can be illustrated under orthogonal designs, where $X'X = I_n$. In this case, the minimizer of (2) has an explicit form:

$$d_j(\lambda) = \left(1 - \frac{\lambda}{(\hat{\beta}_j^{\text{LS}})^2} \right)_+, \quad j = 1, \dots, p. \quad (3)$$

Therefore, for those coefficients whose full least square estimate is large in magnitude, the shrinking factor will be close to 1. But for a redundant predictor, the least square estimate is likely to be small and consequently the shrinking factor will have a good chance to be exactly zero.

A drawback of the nonnegative garrote is its explicit reliance on the full least square estimate. Obviously, with a small sample size, the least squares may perform poorly, and the nonnegative garrote is expected to suffer as well. In particular, the original nonnegative garrote, as proposed by Breiman (1995) can not be applied when the sample size is smaller than the number of predictors. We propose a simple modification to the original nonnegative garrote. We suggest to use the ridge regression as the initial estimate in defining the nonnegative garrote estimate, instead of the least square estimate. The ridge estimate for (1) is given as $\beta^{\text{init}} \equiv \hat{\beta}^{\text{ridge}}(\tau) = (X'X + \tau I_p)^{-1} X'Y$, where $\tau > 0$ is a pre-specified tuning parameter. With slight abuse of notation, write $Z_j = X_j \beta_j^{\text{init}}$. Similar to the original nonnegative garrote estimator, we compute the shrinking factor by minimizing (2) and define the final estimate as $\hat{\beta}_j^{\text{NG}}(\lambda) = d_j(\lambda) \beta_j^{\text{init}}$ for $j = 1, \dots, p$. Such modification mitigates the problems caused by the least squares.

Another potential hurdle when using the nonnegative garrote estimator for large

scale problems is the computational cost. The nonnegative garrote is so far computed using the standard quadratic programming technique for a given tuning parameter, which can be computationally demanding for high dimensional problems, especially if a fine grid of tuning parameters are to be considered. In this paper, we show that the solution path of the nonnegative garrote is piecewise linear, regardless of the initial estimate, and use this to construct a more efficient algorithm of building the nonnegative garrote solution path. The proposed algorithm computes the whole solution path of the nonnegative garrote with the computational load in the same magnitude as the ordinary least squares.

Furthermore, we prove that the nonnegative garrote is a consistent variable selection and estimation procedure in the sense that it is consistent in terms of both estimation and model selection given that the tuning parameter λ is appropriately chosen, while such property is oftentimes not shared by another popular variable selection method, LASSO. More generally, we argue that the nonnegative garrote has the ability to turn a consistent estimate into an estimate that is not only consistent in terms of estimation but also in terms of variable selection. For illustration purpose, we propose an extension of the method to variable selection in nonparametric regression within the functional analysis of variance (ANOVA) framework.

The rest of the paper is organized as follows. In the next section, we introduce an efficient algorithm for computing the nonnegative garrote solution path. The path consistency of the nonnegative garrote estimator is addressed in Section 3. Section 4 extends the method to the nonparametric additive model. Sections 5 and 6 present some simulations and a real example to support the theoretical results. We conclude the paper with a summary in Section 7. All technical proofs are relegated to the Appendix.

2 Algorithm

Similar to other methods of regularization, the nonnegative garrote estimation procedure proceeds in two steps in practice. First the solution path indexed by the tuning parameter λ is constructed. The second step, oftentimes referred to as tuning, selects the final estimate on the solution path. For most methods of regularization, it

is very expensive to compute the exact solution path. One has to approximate the solution path by evaluating the estimate for a fine grid of tuning parameters and there is a tradeoff between the approximation accuracy and the computational cost in determining how fine a grid of tuning parameters to be considered. In particular, the nonnegative garrote solution path can be approximated by solving the quadratic programming problem (2) for a series of λ 's, as done in Breiman (1995).

We show that the solution path of the nonnegative garrote is piecewise linear, and use this to construct an efficient algorithm of building the exact nonnegative garrote solution path. The following algorithm is quite similar to the modified LARS algorithm (Efron et al., 2004) for the LASSO, with a complicating factor being the nonnegative constraints in (2).

Algorithm – Nonnegative Garrote

(1) Start from $d^{[0]} = 0$, $k = 1$ and $r^{[0]} = Y$

(2) Compute the current active set

$$\mathcal{C}_1 = \arg \max_j Z'_j r^{[k-1]}$$

(3) Compute the current direction γ , which is a p dimensional vector defined by

$$\gamma_{\mathcal{C}_k^c} = 0 \text{ and}$$

$$\gamma_{\mathcal{C}_k} = \left(Z'_{\mathcal{C}_k} Z_{\mathcal{C}_k} \right)^{-1} Z'_{\mathcal{C}_k} r^{[k-1]}$$

(4) For every $j \notin \mathcal{C}_k$, compute how far the group nonnegative garrote will progress in direction γ before X_j enters the active set. This can be measured by a α_j such that

$$Z'_j \left(r^{[k-1]} - \alpha_j Z \gamma \right) = Z'_{j'} \left(r^{[k-1]} - \alpha_j Z \gamma \right) \quad (4)$$

where j' is arbitrarily chosen from \mathcal{C}_k .

(5) For every $j \in \mathcal{C}_k$, compute $\alpha_j = \min(\beta_j, 1)$ where $\beta_j = -d_j^{[k-1]}/\gamma_j$, if nonnegative, measures how far the group nonnegative garrote will progress before d_j becomes zero.

(6) If $\alpha_j \leq 0$, $\forall j$ or $\min_{j:\alpha_j>0}\{\alpha_j\} > 1$, set $\alpha = 1$. Otherwise, denote $\alpha =$

$\min_{j:\alpha_j>0}\{\alpha_j\} \equiv \alpha_{j^*}$. Set $d^{[k]} = d^{[k-1]} + \alpha\gamma$. If $j^* \notin \mathcal{C}_k$, update $\mathcal{C}_{k+1} = \mathcal{C}_k \cup \{j^*\}$; else update $\mathcal{C}_{k+1} = \mathcal{C}_k - \{j^*\}$.

(7) Set $r^{[k]} = Y - Z d^{[k]}$ and $k = k + 1$. Go back to step (3) until $\alpha = 1$.

Theorem 1 *Under the “one at a time” condition discussed below, the trajectory of this algorithm coincides with the nonnegative garrote solution path.*

The same condition as we assumed in Theorem 1, referred to as “one at a time”, was used in deriving the connection between the LASSO and the LARS by Efron et al. (2004). With the current notation, the condition states that j^* in Step (6) is uniquely defined. This assumption basically means that (i) the addition occurs only for one variable a time at any stage of the above algorithm; (ii) no variable vanishes at the time of addition; and (iii) no two variables vanish simultaneously. This is generally true in practice and can always be enforced by slightly perturbing the response. For more detailed discussions, the readers are referred to Efron et al. (2004).

Since

$$\sum_j Z'_j Y = (\beta^{\text{init}})' X' Y = Y' X (X' X + \tau I_p)^{-1} X' Y > 0,$$

we have $\max_j Z'_j r^{[k-1]} > 0$ in Step (2). It is not hard to see that $\max_j Z'_j r^{[k-1]} > 0$ is monotonically decreasing as the algorithm progresses and \mathcal{C}_k maintains the collection of predictors which maximize $Z'_j r^{[k-1]}$. The stopping rule in Step (7) makes sure that the algorithm ends when $\max_j Z'_j r^{[k-1]} = 0$.

Breiman (1995) conjectured that the models produced by the nonnegative garrote are nested in that the model corresponding to a smaller λ always contains the model corresponding to a larger λ . This amounts to stating that $j^* \in \mathcal{C}_k$ never takes place in Step (6). However, we found this conjecture not true although $j^* \in \mathcal{C}_k$ happens only very rarely in our simulation. A numerical counterexample can be obtained from the authors.

3 Consistency

Since the final estimate comes from the solution path, it is of great importance to make sure that the solution path indeed contains at least one “desirable” candidate

estimate. In our context, an estimate $\widehat{\beta}$ is considered “desirable” if it is consistent in terms of both coefficient estimate and variable selection. We call a solution path “path consistent” if it contains at least one such “desirable” estimate. The following theorem states that such consistency holds for the nonnegative garrote solution path.

Theorem 2 *Assume that (i) the initial estimate is δ_n consistent, i.e., $\max_j |\widehat{\beta}_j^{\text{init}}(\tau) - \beta_j| = O_p(\delta_n)$ for some $\delta_n \rightarrow 0$; (ii) the design matrix is nondegenerate, i.e., the smallest eigenvalue of $X'X/n$ is bounded from below by a positive constant with probability tending to one. If λ tends to zero in a fashion such that $\delta_n = o(\lambda)$, then $P(\widehat{\beta}_j^{\text{NG}}(\lambda) = 0) \rightarrow 1$ for any $j \notin \mathcal{I}$, and $\widehat{\beta}_j^{\text{NG}}(\lambda) = \beta_j + O_p(\lambda)$ for any $j \in \mathcal{I}$ where $\mathcal{I} = \{j : \beta_j \neq 0\}$.*

Theorem 2 provides a sufficient condition for an estimate on the nonnegative garrote solution path to be consistent. However, it is worth pointing out that the condition can be weakened in special cases. In particular, we have

Lemma 1 *Assume that the design matrix satisfies $X'_{\mathcal{I}^c}X_{\mathcal{I}} = 0$. Under the conditions of Theorem 2, if λ goes to zero in a fashion such that $\delta_n^2 = o(\lambda)$, then $P(\widehat{\beta}_j^{\text{NG}}(\lambda) = 0) \rightarrow 1$ for any $j \notin \mathcal{I}$, and $\widehat{\beta}_j^{\text{NG}}(\lambda) = \beta_j + O_p(\max\{\delta_n, \lambda\})$ for any $j \in \mathcal{I}$.*

Such path consistency property of the nonnegative garrote is to be contrast with the following result for the LASSO.

Theorem 3 *The sufficient and necessary condition for the LASSO to be path consistent is*

$$\max_{j \notin \mathcal{I}} \text{Cov}(X_j, X_{\mathcal{I}}) [\text{Cov}(X_{\mathcal{I}})]^{-1} s_{\mathcal{I}} < 1, \quad (5)$$

where s is a p dimensional vector with the j th element being $\text{sign}(\beta_j)$.

The fact that the LASSO is not always path consistent has also been independently discovered by other authors (Peng Zhao, personal communication; Hui Zou, personal communication).

Of course the condition given in Theorem 3 can not be checked in practice since it involves the true regression coefficient β . Without the prior knowledge of β , a

stronger condition is required in order to ensure the path consistency of the LASSO:

$$\max_{j \notin \mathcal{I}} \left\| \text{Cov}(X_j, X_{\mathcal{I}}) [\text{Cov}(X_{\mathcal{I}})]^{-1} \right\|_{\ell_1} < 1 \quad (6)$$

In fact, following the same proof as that of Theorem 3, one can show that (6) is the sufficient condition that the LASSO solution path contains an estimate $\hat{\beta}$ such that $\hat{\beta}_j \neq 0$ if and only if $j \in \mathcal{I}$. On the other hand, it is easy to see that if (6) is violated, then there always exists a β such that (5) is not satisfied. By Theorem 3, the LASSO is not path consistent at least for such β .

4 Extension

Although we have focused on variable selection for usual multiple linear regression, the aforementioned results for the nonnegative garrote also apply for more general setup. For example, Yuan and Lin (2004) have discussed the so-called grouped variable selection where the predictors of (1) are naturally grouped. Such situations occur in the case of categorical predictors or additive models with polynomial components. Along with other methods, they proposed the group nonnegative garrote for variable selection in such cases. It is trivial to show that both the piecewise linear solution path property and the path consistency hold for the group nonnegative garrote.

In this section, we consider the application of the proposed method to a more general nonparametric setting, within the functional ANOVA framework (Wahba, 1990; Gu, 2002). In the functional ANOVA, we rewrite the regression function $f(X) = E(Y|X)$ as

$$f(X) = \mu + \sum_{j=1}^p f_j(x_j) + \sum_{1 \leq j_1 < j_2 \leq p} f_{j_1 j_2}(x_{j_1}, x_{j_2}) + \dots + f_{1 \dots p}(x_1, \dots, x_p), \quad (7)$$

where μ is a constant, f_j 's are the main effects, $f'_{j_1 j_2}$'s are the two way interactions, and so on. The functional ANOVA provides a general framework for nonparametric multivariate function estimation and the series on the right hand side of (7) is usually truncated somewhere to enhance interpretability. The identifiability of the terms in (7) is assured by side conditions through averaging operators. The most popular ex-

ample of functional ANOVA is the additive model proposed by Hastie and Tibshirani (1990) where only main effects are retained in (7). Of interest here is to select and estimate the active components on the right hand side of (7). For illustration purpose, we shall restrict our attention to the additive model.

Again, we start with an initial estimate of the components $\hat{f}_1^{\text{init}}, \dots, \hat{f}_p^{\text{init}}$. One good choice is the smoothing spline estimate which is the minimizer to

$$\sum_{i=1}^n (Y_i - f(X_i))^2 + \sum_{j=1}^p \tau_j J_j(f_j) \quad (8)$$

where τ 's are tuning parameters and J 's are squared norms defined over the subspace where the corresponding function comes from. For more detailed discussion, the readers are referred to Wahba (1990) or Gu (2002).

Our final estimate is then defined by scaling each component. Similar to before, the shrinking factor is obtained by minimizing

$$\frac{1}{2} \|Y - Zd\|^2 + n\lambda \sum_{j=1}^p d_j \quad (9)$$

subject to the constraint that every component of d is nonnegative. Here Z is a matrix whose j th column is \hat{f}_j^{init} evaluated on the sample points. Similar to Theorem 2, we have

Theorem 4 *Assume that the initial estimate is δ_n^2 consistent in ℓ_2 norm, i.e.,*

$$\int \left(f_j(x) - \hat{f}_j^{\text{init}}(x) \right)^2 p_j(x) dx = O_p(\delta_n^2) \quad \text{for some } \delta_n \rightarrow 0, \quad (10)$$

where $p_j(\cdot)$ is the density of X_j . If λ tends to zero in a fashion such that $\delta_n = o(\lambda)$, then $P(\hat{f}_j^{\text{NG}} = 0) \rightarrow 1$ for any j such that $f_j = 0$, and

$$\int \left(f_j(x) - \hat{f}_j^{\text{NG}}(x) \right)^2 p_j(x) dx = O_p(\lambda^2). \quad (11)$$

The proof follows in the same line as that of Theorem 2 and is therefore omitted.

5 Simulation

In this section, we investigate the finite sample properties of the nonnegative garrote estimator. In all simulations, we choose to use $\widehat{\beta}^{\text{ridge}}(\tau)$ as the initial estimate with τ being the minimizer of the GCV score (Golub, Heath and Wahba, 1979):

$$\text{GCV}(\tau) = \frac{1}{n} \frac{\|Y - X\widehat{\beta}^{\text{ridge}}(\tau)\|^2}{(1 - \text{df}(\tau)/n)^2}, \quad (12)$$

where $\text{df}(\tau) = \text{tr}(X(X'X + \tau I_p)^{-1}X')$ and construct the nonnegative garrote solution path using the algorithm presented in Section 2.

In the first set of simulation, we demonstrate the path consistency of the nonnegative garrote procedure in contrast to the LASSO. We consider a simple model:

$$Y = X_1 + X_2 + 0 \cdot X_3 + \epsilon, \quad (13)$$

where $\epsilon \sim \mathcal{N}(0, 1)$. The two active predictors X_1 and X_2 were independently simulated from a standard normal distribution. An additional noisy variable X_3 was also included in the analysis. Conditional on X_1 and X_2 , X_3 was generated from a normal distribution with mean $\alpha(X_1 + X_2)$ and variance $1 - 2\alpha^2$. Therefore, the marginal distribution of X_3 is also $\mathcal{N}(0, 1)$. We consider four different α 's: 0.35, 0.45, 0.55 and 0.65. For each α value, we consider 20 equally spaced sample sizes, 25, 50, \dots , 500. For each combination of α and sample size, one hundred datasets were simulated and we counted how many times, the nonnegative garrote and the LASSO solution paths cover the true model, i.e., how many times the path contains at least one estimate $\widehat{\beta}$ such that $\widehat{\beta}_1 \neq 0$, $\widehat{\beta}_2 \neq 0$ and $\widehat{\beta}_3 = 0$. Figure 1 depicts the frequency for each method to cover the true model.

When $\alpha = 0.35$, both estimating procedures are consistent. But the nonnegative garrote selects the correct model more often than the LASSO for smaller sample sizes. When α increases, the convergence of the coverage probabilities for both the nonnegative garrote and the LASSO slows down. For $\alpha = 0.55$ or 0.65, the LASSO does not seem to be consistent any more. In contrast, the nonnegative garrote is still very capable of selecting the right model for α as large as 0.65. It is worth pointing

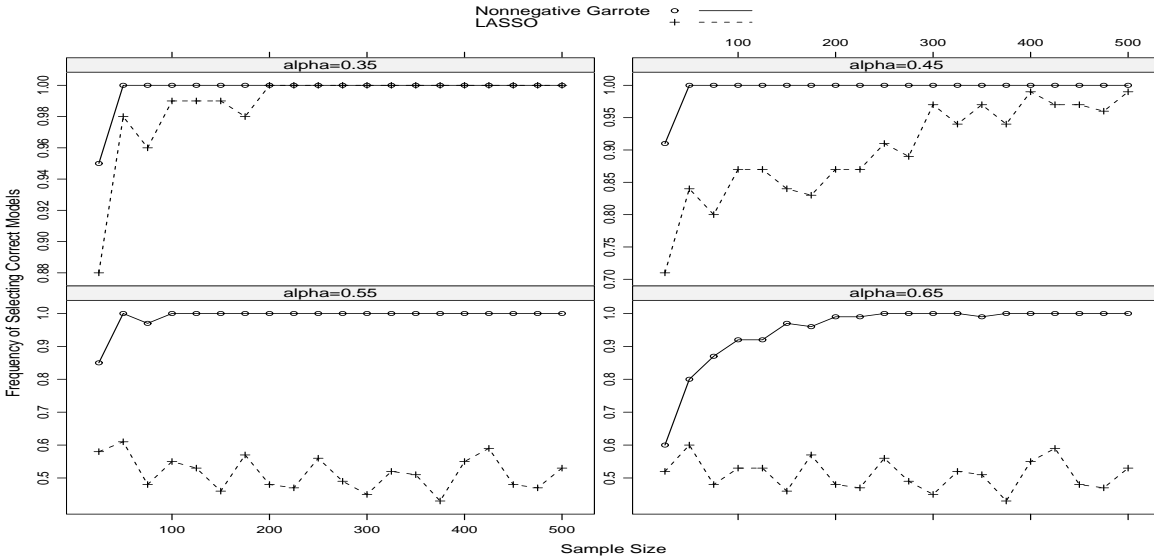


Figure 1: Consistency of the Nonnegative Garrote in contrast to the LASSO

out that such empirical evidence agrees with our theoretical result presented before. According to Theorem 3, the LASSO is path consistent only if $\alpha < 0.5$.

In the second set of simulation, we consider the four models used in the original LASSO paper (Tibshirani, 1996). The models are

- I. $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)'$ and $\sigma = 3$. The correlation between X_i and X_j is $\rho^{|i-j|}$ with $\rho = 0.5$.
- II. Same as (I) except that $\beta_j = 0.85$ for all j .
- III. Same set-up as before, but with $\beta = (5, 0, 0, 0, 0, 0, 0, 0)'$ and $\sigma = 2$.
- IV. Forty correlated predictors are considered. $x_{ij} = z_{ij} + w_i$, where z_{ij}, w_i are independent standard normal random variables. The true regression coefficients are 2 for the first 20 predictors and 0 for the other predictors. σ is set to 15.

We repeated each of the four examples 200 times. For each run, we simulated a training set, a validation set and a test set. We use the training set to construct the solution path; the validation set to tune λ , and the test set to evaluate the prediction error of the final estimate. For all models, the size of the test set is 10000. We use the

notation \cdot/\cdot to denote the sizes of the training and validation sets respectively. We compare the proposed nonnegative garrote estimate (GARROTE) with several other popular methods including the LASSO, the elastic net (ENET; Zou and Hastie, 2005), the original nonnegative garrote (OGARROTE) and the ridge regression (RIDGE). Table 1 reports the median mean squared errors (MSE) together with their standard deviation estimated using bootstrap with 500 re-samplings (numbers in parentheses).

Several observations can be made from Table 1. Comparing with the original nonnegative garrote, the proposed estimate enjoys better performance in almost all settings. In relatively sparse models, Models I and III, the proposed estimate also outperforms the LASSO. For other models, the LASSO and the elastic net perform better than the proposed estimate in the case of low sample sizes. But the performance of the proposed method improves quickly as the sample size increases. For medium or large sample sizes, it often outperforms both the LASSO and the elastic net estimate. The median model sizes reported in Table 2 also indicate that the new estimate is the most successful in variable selection.

The last set of simulation concerns the variable selection for additive models. The example setup is the same as Example 1 from Lin and Zhang (2003). Ten covariates were simulated in the following fashion. First W_1, \dots, W_{10} and U were independently simulated from $U[0, 1]$. Then $X_j = (W_j + tU)/(1 + t)$, where parameter t controls the amount of correlation among predictors. We consider $t = 0, 1, 3$ in our simulation. The true model is

$$y = 5g_1(x_1) + 3g_2(x_2) + 4g_3(x_3) + 6g_4(x_4) + \epsilon \quad (14)$$

where

$$g_1(t) = t; \quad g_2(t) = (2t - 1)^2; \quad g_3(t) = \frac{\sin(2\pi t)}{2 - \sin(2\pi t)}; \quad (15)$$

$$g_4(t) = \left(0.1 \sin(2\pi t) + 0.2 \cos(2\pi t) + 0.3 \sin^2(2\pi t) + 0.4 \cos^3(2\pi t) + 0.5 \sin^3(2\pi t)\right). \quad (16)$$

We chose the noise variance $\sigma^2 = 1.74$ to give a signal to noise ratio 3:1. We used smoothing spline estimate (8) as the initial estimate. Similar to before, we chose τ 's using GCV (Wahba, 1990) and estimate the MSE using a test set of size 10000 and

	Model I			Model II			Model III			Model IV		
	20/20	50/50	100/100	20/20	50/50	100/100	20/20	50/50	100/100	100/100	200/200	400/400
LASSO	3.24 (0.19)	1.29 (0.07)	0.56 (0.03)	3.70 (0.18)	1.70 (0.07)	0.82 (0.02)	0.77 (0.05)	0.29 (0.03)	0.16 (0.01)	61.49 (1.12)	36.02 (0.79)	18.04 (0.35)
OGARROTE	3.64 (0.26)	1.20 (0.08)	0.43 (0.03)	4.70 (0.15)	1.83 (0.08)	0.85 (0.04)	0.49 (0.04)	0.16 (0.02)	0.08 (0.01)	98.68 (1.78)	49.82 (1.45)	21.59 (0.34)
GARROTE	3.19 (0.32)	1.17 (0.06)	0.44 (0.03)	4.09 (0.20)	1.75 (0.08)	0.85 (0.04)	0.48 (0.04)	0.15 (0.02)	0.08 (0.01)	69.89 (1.91)	42.00 (1.07)	19.69 (0.41)
ENET	3.08 (0.15)	1.18 (0.06)	0.54 (0.03)	3.10 (0.11)	1.35 (0.07)	0.67 (0.04)	0.92 (0.06)	0.32 (0.03)	0.17 (0.01)	52.05 (1.20)	31.48 (0.55)	16.64 (0.44)
RIDGE	3.90 (0.18)	1.67 (0.08)	0.80 (0.03)	2.62 (0.16)	1.19 (0.07)	0.62 (0.03)	2.35 (0.08)	0.85 (0.03)	0.39 (0.02)	37.14 (0.67)	27.65 (0.54)	16.39 (0.32)

Table 1: Median MSE for the four models considered in the simulation.

	Model I			Model II			Model III			Model IV		
	20/20	50/50	100/100	20/20	50/50	100/100	20/20	50/50	100/100	100/100	200/200	400/400
LASSO	5.00	6.00	6.00	6.00	8.00	8.00	3.00	3.00	3.00	24.00	28.00	30.00
OGARROTE	4.00	5.00	4.00	5.00	8.00	8.00	2.00	2.00	2.00	20.00	25.00	28.00
GARROTE	4.00	5.00	4.00	6.00	7.00	8.00	2.00	2.00	2.00	23.00	26.00	28.00
ENET	6.00	6.00	6.00	7.00	8.00	8.00	4.00	3.00	4.00	27.00	29.00	30.00
RIDGE	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	8.00	40.00	40.00	40.00

Table 2: Median model sizes for the four models considered in the simulation.

the sizes of the training and validation sets are 100. Table 3 documents the median MSE averaged over 200 runs. We also report its standard deviation estimated using 500 bootstrap samples (numbers in parentheses).

t=0	t=1	t=3
0.57 (0.02)	0.62 (0.03)	0.63 (0.02)

Table 3: Median MSE for the Additive Model Example

We also recorded the frequency of different model sizes in Table 4. Compared with the results from Lin and Zhang (2003), the nonnegative garrote enjoys better predictive performance but their COSSO estimate is more successful in variable selection.

	2	3	4	5	6	7	8	9	10
t=0	0	0	40	91	47	14	2	4	2
t=1	1	11	39	65	49	19	10	4	2
t=3	1	16	41	36	44	33	16	9	4

Table 4: Frequency of Model Sizes for the Additive Model Example

6 Real Example

To further illustrate our results, we re-analyze the prostate cancer dataset from the study by Stamey *et. al.* (1989). This dataset, previously used in Tibshirani (1996), consists the medical records of 97 male patients who were about to receive a radical prostatectomy. The response variable is the level of prostate specific antigen. The predictors are eight clinical measures: log(cancer volume) (lcavol), log(prostate weight) (lweight), age, log(benign prostatic hyperplasia amount) (lbph), seminal vesicle invasion (svi), log(capsular penetration) (lcp), Gleason score (gleason) and percentage Gleason scores 4 or 5 (pgg45).

One of the main interests here is to identify which predictors are more important in predicting the response. The first row of Figure 2 gives the solution paths of

LASSO, GARROTE and ENET. There are two tuning parameters for ENET. As suggested in Zou and Hastie (2005), we fix one tuning parameter at 1000 and the solution path corresponds to different values of another one. In each panel, the gray vertical line indicates the tuning parameter chosen by ten-fold cross-validation. All methods indicate that **gleason** may be an unimportant predictor whereas **lcavol** is the most important predictor. To demonstrate the path consistency results from Section 3, we replace **gleason** with an artificial variable $2 \text{lcavol} + \text{gleason}$. This new variable again contains little extra information for predicting the response and a path consistent method should be able to recognize this fact. The solution paths of the three methods on the new dataset are given in the second row of Figure 2. Comparing with the original solution path, GARROTE is the least disturbed by such change. Both LASSO and ENET select the artificial variable as an important predictor. This observation confirms the theory from Section 3 that the path consistency of LASSO depends on the correlation of the design matrix whereas GARROTE is always path consistent.

7 Conclusion

In this paper we show that the solution path of the nonnegative garrote is piecewise linear, and the whole path can be computed quickly at the same time. The nonnegative garrote estimator is also path consistent given an appropriate initial estimate. We have also shown by simulations and a real example that the proposed method enjoys competitive performance when compared with other popular variable selection and estimation methods. Several practical issues in using the nonnegative garrote in a wider range of applications are worth further investigation. Instead of the ordinary least square estimate, we suggest to use the ridge regression with the ridge parameter chosen by GCV as the initial estimate. Our theoretical development indicates that other consistent estimating procedure can also be used. It is of interest to see if other initial estimate could lead to improved performance. In our simulation, we tune λ using a validation set. In practice, cross-validation is often used for the lack of such a validation set. Cross-validation can be computationally expensive for large scale problems. It is therefore of great practical importance to devise new criteria

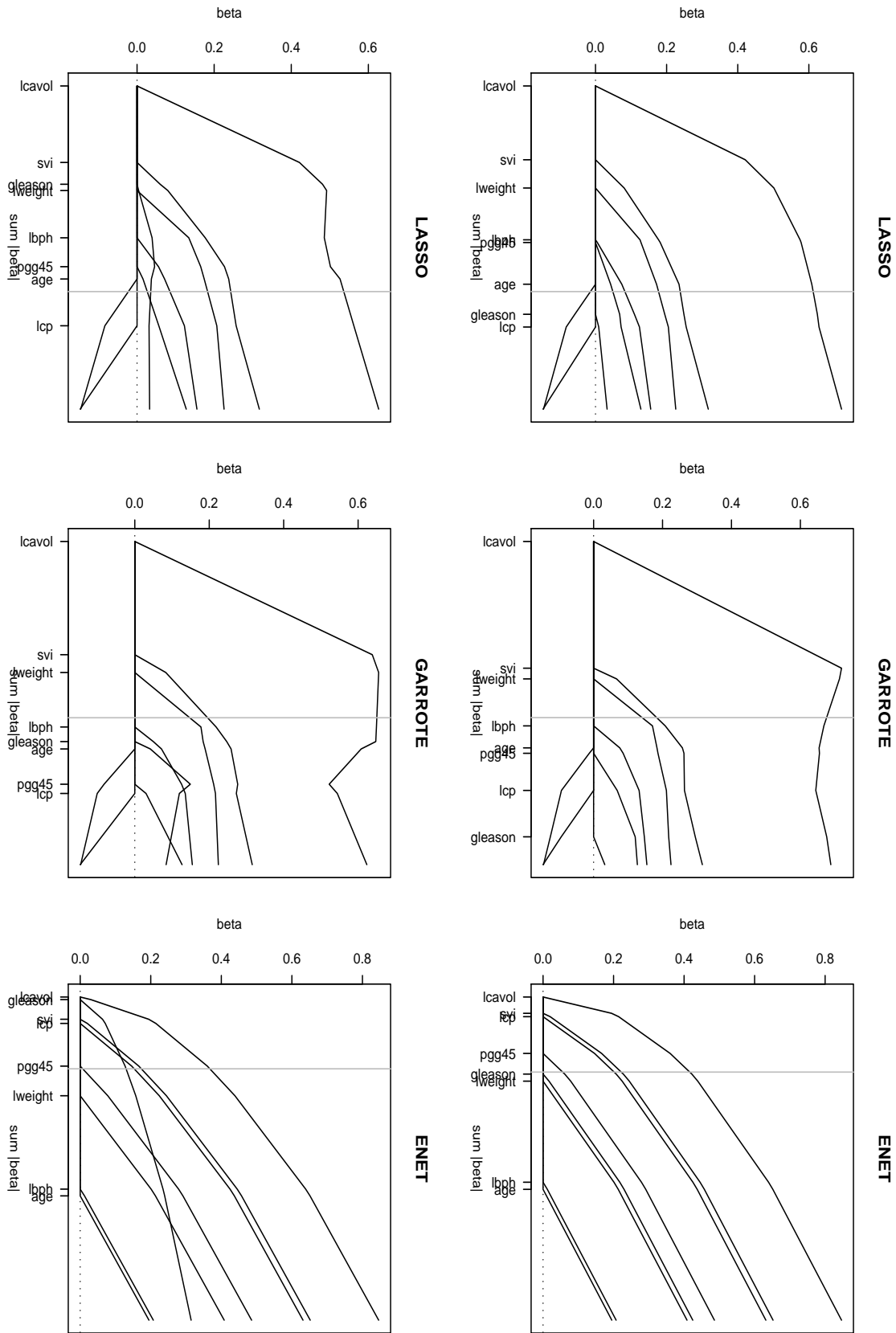


Figure 2: Solution Paths for the Prostate Cancer Example

for selecting final estimate on the nonnegative garrote solution path. We leave these questions for future studies.

Appendix

Proof of Theorem 1 Karush-Kuhn-Tucker Theorem suggests that a necessary and sufficient condition for a point d to be on the solution path of (2) is that there exists a $\lambda \geq 0$ such that for any $j = 1, \dots, p$,

$$\{-Z'_j(Y - Zd) + \lambda\}d_j = 0 \quad (17)$$

$$-Z'_j(Y - Zd) + \lambda \geq 0 \quad (18)$$

$$d_j \geq 0 \quad (19)$$

In the following we show that (17)-(19) are satisfied by any point on the solution path constructed by the algorithm; and any solution to (17)-(19) for certain $\lambda \geq 0$ is also on the constructed solution path.

We verify (17)-(19) for the solution path by induction. Obviously, they are satisfied by $d^{[0]}$. Now suppose that they hold for any point prior to $d^{[k]}$. It suffices to show that they are also true for any point between $d^{[k]}$ and $d^{[k+1]}$. There are three possible actions at step k : (i) a variable is added to active set: $j^* \notin \mathcal{C}_k$; (ii) a variable is deleted from the active set: $j^* \in \mathcal{C}_k$; and (iii) $\alpha = 1$. It is easy to see that (17)-(19) will continue to hold for any point between $d^{[k]}$ and $d^{[k+1]}$ if $\alpha = 1$. Now we consider the other two possibilities.

First consider additions. Without loss of generality, assume that $\mathcal{C}_k - \mathcal{C}_{k-1} = \{1\}$. Note that a point between $d^{[k]}$ and $d^{[k+1]}$ can be expressed as $d^\alpha \equiv d^{[k]} + \alpha\gamma$, where $\alpha \in (0, \alpha_1]$ and γ is a vector defined by $\gamma_{\mathcal{C}_k^c} = \mathbf{0}$ and

$$\gamma_{\mathcal{C}_k} = (Z'_{\mathcal{C}_k} Z_{\mathcal{C}_k})^{-1} Z'_{\mathcal{C}_k} r^{[k]}. \quad (20)$$

It is not hard to show that (17) and (18) are true for d^α . It now suffices to check (19). By the construction of the algorithm, it boils down to verify that $\gamma_1 > 0$.

By the definition of \mathcal{C}_k and \mathcal{C}_{k-1} , we know that for any $j \in \mathcal{C}_{k-1}$,

$$Z'_j r^{[k-1]} > Z'_1 r^{[k-1]} \quad (21)$$

$$Z'_j r^{[k]} = Z'_1 r^{[k]} \quad (22)$$

Therefore,

$$Z'_1 (r^{[k-1]} - r^{[k]}) < Z'_j (r^{[k-1]} - r^{[k]}).$$

Because there exists a positive constant b such that $r^{[k-1]} - r^{[k]} = b Z_{\mathcal{C}_{k-1}} (Z'_{\mathcal{C}_{k-1}} Z_{\mathcal{C}_{k-1}})^{-1} Z'_{\mathcal{C}_{k-1}} r^{[k-1]}$, one concludes that

$$Z'_1 Z_{\mathcal{C}_{k-1}} (Z'_{\mathcal{C}_{k-1}} Z_{\mathcal{C}_{k-1}})^{-1} Z'_{\mathcal{C}_{k-1}} r^{[k-1]} < Z'_j Z_{\mathcal{C}_{k-1}} (Z'_{\mathcal{C}_{k-1}} Z_{\mathcal{C}_{k-1}})^{-1} Z'_{\mathcal{C}_{k-1}} r^{[k-1]}.$$

Write $s = \mathbf{1}_p$. Since $Z'_{\mathcal{C}_{k-1}} r^{[k-1]} = (Z'_j r^{[k-1]}) s_{\mathcal{C}_{k-1}}$, we have

$$Z'_1 Z_{\mathcal{C}_{k-1}} (Z'_{\mathcal{C}_{k-1}} Z_{\mathcal{C}_{k-1}})^{-1} s_{\mathcal{C}_{k-1}} < 1. \quad (23)$$

Together with (20),

$$\gamma_1 = \frac{\left\{ 1 - Z'_1 Z_{\mathcal{C}_{k-1}} (Z'_{\mathcal{C}_{k-1}} Z_{\mathcal{C}_{k-1}})^{-1} s_{\mathcal{C}_{k-1}} \right\} Z'_j r^{[k]}}{Z'_1 \left\{ I_n - Z_{\mathcal{C}_{k-1}} (Z'_{\mathcal{C}_{k-1}} Z_{\mathcal{C}_{k-1}})^{-1} Z'_{\mathcal{C}_{k-1}} \right\} Z_1} > 0, \quad (24)$$

Now let us consider the case of deletion. Without loss of generality, assume that $\mathcal{C}_{k-1} - \mathcal{C}_k = \{1\}$. In this case, a point between $d^{[k]}$ and $d^{[k+1]}$ can still be expressed as $d^\alpha \equiv d^{[k]} + \alpha \gamma$, where $\alpha \in (0, \alpha_1]$ and γ is still defined by (20). It is readily to show that (17) and (19) are true with $\lambda = Z'_j (Y - Z d^\alpha)$ where j is arbitrarily chosen from \mathcal{C}_k . It suffices to verify (18). By the construction of the solution path, it suffices to show that (18) holds for $j = 1$.

Note that any point between $d^{[k-1]}$ and $d^{[k]}$ can be written as $d^{[k-1]} + c \tilde{\gamma}$, where $c > 0$ and $\tilde{\gamma}$ is given by $\tilde{\gamma}_{\mathcal{C}_{k-1}^c} = \mathbf{0}$ and

$$\tilde{\gamma}_{\mathcal{C}_{k-1}} = (Z'_{\mathcal{C}_{k-1}} Z_{\mathcal{C}_{k-1}})^{-1} Z'_{\mathcal{C}_{k-1}} r^{[k-1]}. \quad (25)$$

Clearly, $\tilde{\gamma}_1 < 0$. Similar to (24), we have

$$\tilde{\gamma}_1 = \frac{\left\{1 - Z_1' Z_{C_k} \left(Z_{C_k}' Z_{C_k}\right)^{-1} s_{C_k}\right\} Z_j' r^{[k]}}{Z_1 \left\{I_n - Z_{C_k} \left(Z_{C_k}' Z_{C_k}\right)^{-1} Z_{C_k}'\right\} Z_1} \quad (26)$$

where j is arbitrarily chosen from C_k . Therefore,

$$Z_1' Z_{C_k} \left(Z_{C_k}' Z_{C_k}\right)^{-1} s_{C_k} = (p_j / Z_j' r^{[k]}) Z_1' Z \gamma < 1.$$

In other words, $Z_1' Z \gamma < Z_j' r^{[k]} = Z_j' Z \gamma$. Since $Z_1' r^{[k]} = Z_j' r^{[k]}$, we conclude that $Z_1'(Y - Z d^\alpha) < Z_j'(Y - Z d^\alpha) = \lambda$.

Next, we need to show that for any $\lambda \geq 0$, the solution to (17)-(19) is on the solution path. By the continuity of the solution path and the uniqueness of the solution to (2), it is evident that for any $\lambda \in [0, \max_j Z_j' Y]$, the solution to (17)-(19) is on the path. The proof is now completed by the fact that for any $\lambda > \max_j Z_j' Y$, the solution to (17)-(19) is $\mathbf{0}$ which is also on the solution path. ■

Proof of Theorem 2 For brevity, we suppress the dependence on λ in the proof. Let

$$\begin{aligned} \Lambda_{01} &= \{j : d_j = 0, \beta_j \neq 0\}, \\ \Lambda_{00} &= \{j : d_j = 0, \beta_j = 0\}, \\ \Lambda_{11} &= \{j : d_j > 0, \beta_j \neq 0\}, \\ \Lambda_{10} &= \{j : d_j > 0, \beta_j = 0\}, \end{aligned}$$

and $p_{ij} = \#(\Lambda_{ij})$. Denote event $\mathcal{A} = \{p_{10} > 0\}$. First we show that $P(\mathcal{A}) \rightarrow 0$ as $n \rightarrow \infty$. Write $d_{ij} = d_{\Lambda_{ij}}$, $i, j = 0, 1$ and other vectors and matrices be defined in the same fashion unless otherwise indicated. Note that d_1 is also the unconstrained minimizer of

$$\frac{1}{2} \|Y - Z_1 \gamma\|^2 + n \lambda \sum_j \gamma_j, \quad (27)$$

where $\gamma \in R^{p_1}$. Therefore

$$\begin{pmatrix} d_{11} \\ d_{10} \end{pmatrix} = \begin{pmatrix} Z'_{11}Z_{11}/n & Z'_{11}Z_{10}/n \\ Z'_{10}Z_{11}/n & Z'_{10}Z_{10}/n \end{pmatrix}^{-} \begin{pmatrix} Z'_{11}Y/n - \lambda \mathbf{1}_{p_{11}} \\ Z'_{10}Y/n - \lambda \mathbf{1}_{p_{10}} \end{pmatrix}$$

Denote

$$\begin{aligned} A &= Z'_1 Z_1, \\ A_{ij} &= Z'_{1i} Z_{1j}, \quad i, j = 0, 1, \\ A_{00.1} &= A_{00} - A_{01} A_{11}^{-} A_{10}. \end{aligned}$$

Then

$$A^{-} = \begin{pmatrix} * & * \\ -A_{00.1}^{-} A_{01} A_{11}^{-} & A_{00.1}^{-} \end{pmatrix}.$$

This implies that

$$d_{10} = -A_{00.1}^{-} A_{01} A_{11}^{-} (Z'_{11}Y/n - \lambda \mathbf{1}_{p_{11}}) + A_{00.1}^{-} (Z'_{10}Y/n - \lambda \mathbf{1}_{p_{10}}) \equiv A_{00.1}^{-} w \quad (28)$$

Rewrite w as

$$w = Z'_{10} \left[I_{p_{11}} - Z_{11} (Z'_{11} Z_{11})^{-} Z'_{11} \right] Y/n - \lambda \mathbf{1}_{p_{10}} + \lambda A_{01} A_{11}^{-} \mathbf{1}_{p_{11}}. \quad (29)$$

Because $\widehat{\beta}^{\text{init}}$ is δ_n consistent, for any $i, j \in \{1, \dots, p\}$,

$$\begin{aligned} \left| \widehat{\beta}_i^{\text{init}} \widehat{\beta}_j^{\text{init}} - \beta_i \beta_j \right| &= \left| \widehat{\beta}_i^{\text{init}} (\widehat{\beta}_j^{\text{init}} - \beta_j) + \beta_j (\widehat{\beta}_i^{\text{init}} - \beta_i) \right| \\ &\leq \left(\left| \widehat{\beta}_i^{\text{init}} \right| + |\beta_j| \right) \left| \widehat{\beta}_j^{\text{init}} - \beta_j \right| \\ &= O_p(\delta_n). \end{aligned} \quad (30)$$

This entails

$$A_{11} = \frac{1}{n} \Delta_{11} X'_{11} X_{11} \Delta_{11} + O_p(\delta_n), \quad (31)$$

$$A_{01} = O_p(\delta_n), \quad (32)$$

where Δ is a diagonal matrix with diagonal elements β . Consequently,

$$w = Z'_{10} \left[I_{p_{11}} - Z_{11} (Z'_{11} Z_{11})^{-1} Z'_{11} \right] Y/n - \lambda (1 + O_p(\delta_n)) \mathbf{1}_{p_{10}}. \quad (33)$$

Now note that

$$\left\| \left[I_{p_{11}} - Z_{11} (Z'_{11} Z_{11})^{-1} Z'_{11} \right] Y \right\|^2 \leq Y'Y = O_p(n), \quad (34)$$

since $Z_{11} (Z'_{11} Z_{11})^{-1} Z'_{11}$ is a projection matrix. Thus by Cauchy-Schwartz inequality,

$$\begin{aligned} & \left\| Z'_{10} \left[I_{p_{11}} - Z_{11} (Z'_{11} Z_{11})^{-1} Z'_{11} \right] Y \right\| \\ & \leq \|Z_{10}\| \left\| \left[I_{p_{11}} - Z_{11} (Z'_{11} Z_{11})^{-1} Z'_{11} \right] Y \right\| \\ & = O_p(\sqrt{n} \|Z_{10}\|) \\ & = O_p\left(n \max_{j \notin \mathcal{I}} |\hat{\beta}_j^{\text{init}}|\right) \\ & = O_p(n\delta_n) = o_p(n\lambda). \end{aligned} \quad (35)$$

This leads to $w = -\lambda(1 + o_p(1))\mathbf{1}_{p_{10}}$. Since $d_j > 0$ for any $j \in \Lambda_{10}$, we have $w'd_{10} < 0$. This contradicts with (28) which implies that $w'd_{10} = w'A_{00.1}^- w \geq 0$. Thus, when $n \rightarrow \infty$, $P(\mathcal{A}) \rightarrow 0$.

Denote $\mathcal{B} = \{p_{10} = 0\}$. It now suffices to show that $P(\mathcal{B}|\mathcal{A}^c) \rightarrow 1$. Assume that $p_{10} = 0$. Let d^u be the unconstrained minimizer of

$$\frac{1}{2} \|Y - Z_{\cdot 1} \gamma\|^2 + n\lambda \gamma' \mathbf{1}_{p_{\cdot 1}}, \quad (36)$$

where $\gamma \in R^{p_{\cdot 1}}$. Note that

$$d^u = (Z'_{\cdot 1} Z_{\cdot 1} / n)^{-1} (Z'_{\cdot 1} Y / n - \lambda \mathbf{1}_{p_{\cdot 1}}). \quad (37)$$

Following the same argument as (31), we have

$$\frac{1}{n} Z'_{\cdot 1} Z_{\cdot 1} = \frac{1}{n} \Delta_{\cdot 1} X'_{\cdot 1} X_{\cdot 1} \Delta_{\cdot 1} + O_p(\delta_n). \quad (38)$$

Consequently,

$$d^u = (\Delta_{\cdot 1} X'_{\cdot 1} X_{\cdot 1} \Delta_{\cdot 1} / n)^{-} (Z'_{\cdot 1} Y / n - \lambda \mathbf{1}_{p \cdot 1}) (1 + O_p(\delta_n)). \quad (39)$$

Furthermore, for any $j \in \Lambda_{\cdot 1}$,

$$\left| \frac{1}{n} \left((Z_{\cdot 1} - X_{\cdot 1} \Delta_{\cdot 1})' Y \right)_j \right| = O \left(\left| \left(\widehat{\beta}_{\cdot 1}^{\text{init}} - \beta_{\cdot 1} \right)_j \right| \right) = O_p(\delta_n). \quad (40)$$

Thus,

$$d^u = (\Delta_{\cdot 1} X'_{\cdot 1} X_{\cdot 1} \Delta_{\cdot 1} / n)^{-} (\Delta_{\cdot 1} X'_{\cdot 1} Y / n - \lambda \mathbf{1}_{p \cdot 1}) (1 + O_p(\delta_n)). \quad (41)$$

Combining (41) and the fact that $(\Delta_{\cdot 1} X'_{\cdot 1} X_{\cdot 1} \Delta_{\cdot 1} / n)^{-} \Delta_{\cdot 1} X'_{\cdot 1} Y / n = \mathbf{1}_{p \cdot 1}$, we get

$$d^u = \mathbf{1}_{p \cdot 1} - \lambda (\Delta_{\cdot 1} X'_{\cdot 1} X_{\cdot 1} \Delta_{\cdot 1} / n)^{-} \mathbf{1}_{p \cdot 1} + O_p(\delta_n) = \mathbf{1}_{p \cdot 1} (1 + O_p(\lambda)). \quad (42)$$

Thus, with probability tending to 1, $d^u \rightarrow \mathbf{1}_{p \cdot}$. In other words $\widehat{\beta}_j^{\text{NG}}(\lambda) = \widehat{\beta}_j^{\text{init}}(1 + O_p(\lambda))$ for $j \in \mathcal{I}$ as $n \rightarrow \infty$. Now the proof is completed since $\widehat{\beta}_j^{\text{init}} \rightarrow_p \beta_j$. ■

Proof of Lemma 1 In this case, the first term on the left hand side of (33) can be expressed as $Z'_{10} Y = \Delta_{10} (X'_{10} X_{10}) \Delta_{10} = O_p(\delta_n^2) = o_p(\lambda)$. Therefore, $w = -\lambda(1 + o_p(1))$. The rest of the proof is exactly the same as the proof of Theorem 2. ■

Proof of Theorem 3 Recall that the LASSO with tuning parameter λ is given as the minimizer to

$$\frac{1}{2} \|Y - X\gamma\|^2 + n\lambda \sum_{j=1}^p |\gamma_j|. \quad (43)$$

Karush-Kuhn-Tucker Theorem suggests that a necessary and sufficient condition for any p dimensional vector $\tilde{\beta}$ to be on the LASSO solution path is

$$\frac{1}{n} X'_j (Y - X\tilde{\beta}) = \lambda \text{sign}(\tilde{\beta}_j), \quad \text{if } \tilde{\beta}_j \neq 0 \quad (44)$$

$$\left| \frac{1}{n} X'_j (Y - X\tilde{\beta}_1) \right| \leq \lambda, \quad \text{if } \tilde{\beta}_j = 0 \quad (45)$$

Now suppose that (5) holds. Let $\tilde{\beta}_{\mathcal{I}}$ be the minimizer to

$$\frac{1}{2} \|Y - X_{\mathcal{I}}\gamma\|^2 + n\lambda \sum_j |\gamma_j|, \quad (46)$$

where $\lambda = 1/\ln n$. It is easy to see that $\tilde{\beta}_j \rightarrow_p \beta_j$ for any $j \in \mathcal{I}$ and therefore with probability tending to one, $\tilde{\beta}_j \neq 0$ for any $j \in \mathcal{I}$. Let $\tilde{\beta}_{\mathcal{I}^c} = \mathbf{0}$. It now suffices to show that with probability tending to one, such $\tilde{\beta}$ is also on the solution path of (43). Note that from (46),

$$\frac{1}{n} X'_{\mathcal{I}}(Y - X_{\mathcal{I}}\tilde{\beta}_{\mathcal{I}}) = \lambda \text{sign}(\tilde{\beta}_{\mathcal{I}}). \quad (47)$$

On the other hand, because $X'X/n = \text{Cov}(X) + O_p(1/\sqrt{n})$ and $\hat{\beta}^{\text{LS}} = \beta + O_p(1/\sqrt{n})$,

$$\frac{1}{n} X'_{\mathcal{I}}(Y - X_{\mathcal{I}}\tilde{\beta}_{\mathcal{I}}) = \frac{1}{n} X'_{\mathcal{I}}X_{\mathcal{I}}(\hat{\beta}_{\mathcal{I}}^{\text{LS}} - \tilde{\beta}_{\mathcal{I}}) + \frac{1}{n} X'_{\mathcal{I}}X_{\mathcal{I}^c}\hat{\beta}_{\mathcal{I}^c}^{\text{LS}} = \text{Cov}(X_{\mathcal{I}})(\beta_{\mathcal{I}} - \tilde{\beta}_{\mathcal{I}}) + O_p(n^{-1/2}). \quad (48)$$

Combining (47) and (48),

$$\tilde{\beta}_{\mathcal{I}} = \beta_{\mathcal{I}} - \lambda [\text{Cov}(X_{\mathcal{I}})]^{-1} \text{sign}(\tilde{\beta}_{\mathcal{I}}) + O_p(n^{-1/2}). \quad (49)$$

Therefore,

$$\begin{aligned} \frac{1}{n} X'_{\mathcal{I}^c}(Y - X_{\mathcal{I}}\tilde{\beta}_{\mathcal{I}}) &= \frac{1}{n} X'_{\mathcal{I}^c}X_{\mathcal{I}}(\beta_{\mathcal{I}}^{\text{LS}} - \tilde{\beta}_{\mathcal{I}}) + \frac{1}{n} X'_{\mathcal{I}^c}X_{\mathcal{I}^c}\beta_{\mathcal{I}^c}^{\text{LS}} \\ &= \text{Cov}(X_{\mathcal{I}^c}, X_{\mathcal{I}})(\beta_{\mathcal{I}} - \tilde{\beta}_{\mathcal{I}}) + O_p(n^{-1/2}) \\ &= \lambda \text{Cov}(X_{\mathcal{I}^c}, X_{\mathcal{I}}) [\text{Cov}(X_{\mathcal{I}})]^{-1} \text{sign}(\tilde{\beta}_{\mathcal{I}}) + O_p(n^{-1/2}), \end{aligned} \quad (50)$$

From (50), for any positive constant ε and $\forall j \notin \mathcal{I}$, then with probability tending to one

$$\left| \frac{1}{n} X'_j(Y - X_{\mathcal{I}}\tilde{\beta}_{\mathcal{I}}) \right| \leq c\lambda + \varepsilon \quad (51)$$

where $c < 1$ is the quantity on the left hand side of (6). By choosing $\varepsilon < (1 - c)\lambda$ in (51), together with (47), we have with probability tending to one, $\tilde{\beta}$ satisfies (44) and (45). Hence it is on the LASSO solution path.

Now consider the case when (5) does not hold. Without loss of generality, assume that $\beta_1 = 0$ and $\text{Cov}(X_1, X_{\mathcal{I}}) [\text{Cov}(X_{\mathcal{I}})]^{-1} s_{\mathcal{I}} \geq 1$. Assume that contrary. With

probability tending to one, we can find a “desirable” estimate on the LASSO solution path. Denote $\tilde{\beta}$ a “desirable” estimate. Then with probability tending to one, $\text{sign}(\tilde{\beta}_j) = \text{sign}(\beta_j)$ for any $j \in \mathcal{I}$. Therefore, with probability tending to one,

$$\text{Cov}(X_1, X_{\mathcal{I}}) [\text{Cov}(X_{\mathcal{I}})]^{-1} \text{sign}(\tilde{\beta}_{\mathcal{I}}) \geq 1 \quad (52)$$

Similar to (48), we conclude that with probability tending to one,

$$\frac{1}{n} X_1' (Y - X_{\mathcal{I}} \tilde{\beta}_{\mathcal{I}}) = \lambda \text{Cov}(X_1, X_{\mathcal{I}}) [\text{Cov}(X_{\mathcal{I}})]^{-1} \text{sign}(\tilde{\beta}_{\mathcal{I}}) + \xi \geq \lambda + \xi \quad (53)$$

where

$$\begin{aligned} \xi &= \left(\frac{1}{n} X_1' X_{\mathcal{I}} - \text{Cov}(X_1, X_{\mathcal{I}}) \right) (\beta_{\mathcal{I}} - \tilde{\beta}_{\mathcal{I}}) + \frac{1}{n} X_1' X_{\mathcal{I}} (\hat{\beta}_{\mathcal{I}}^{\text{LS}} - \beta_{\mathcal{I}}) + \frac{1}{n} X_1' X_{\mathcal{I}^c} \hat{\beta}_{\mathcal{I}^c}^{\text{LS}} \\ &= \text{Cov}(X) (\hat{\beta}^{\text{LS}} - \beta) + o_p(n^{-1/2}). \end{aligned} \quad (54)$$

It is not hard to see from the asymptotic normality of $\hat{\beta}^{\text{LS}}$ that $P(\xi > 0)$ is bounded below by a positive constant. This implies that with a nonvanishing probability, $\tilde{\beta}$ cannot satisfy (45), which contradicts the construction of $\tilde{\beta}$. ■

References

- [1] Breiman, L. (1995), Better Subset Regression Using the Nonnegative Garrote, *Technometrics*, **37**, 373-384.
- [2] Breiman, L. (1996), Heuristics of instability and stabilization in model selection, *Ann. Statist.*, **24**, 2350–2383.
- [3] Chen, S. S., Donoho, D. L. and Saunders, M. A. (1998), Atomic Decomposition by Basis Pursuit, *SIAM J. Scientific Computing*, **20**, 33-61.
- [4] Efron, B., Johnstone, I., Hastie, T. and Tibshirani, R. (2004), Least angle regression, *Ann. Statist.*, **32** 407-499.
- [5] Fan, J. and Li, R. (2001), Variable selection via nonconcave penalized likelihood and its oracle properties, *J. Amer. Statist. Assoc.*, **96** 1348-1360.

- [6] Foster, D. P. and George, E. I. (1994), The risk inflation criterion for multiple regression, *Ann. Statist.*, **22**, 1947-1975.
- [7] George, E. I. and Foster, D. P. (2000), Calibration and empirical Bayes variable selection, *Biometrika*, **87**, 731-747.
- [8] George, E. I. and McCulloch, R. E. (1993), Variable selection via Gibbs sampling, *J. Amer. Statist. Assoc.*, **88**, 881-889.
- [9] Golub, G., Heath, M. and Wahba, G. (1979), Generalized cross validation as a method for choosing a good ridge parameter, *Technometrics*, **21**, 215-224.
- [10] Gu, C. (2002), *Smoothing Spline ANOVA Models*, Springer, New York.
- [11] Hastie, T. and Tibshirani, R. (1990), *Generalized Additive Models*, Chapman & Hall/CRC, London.
- [12] Lin, Y. and Zhang, H. H. (2003), Component selection and smoothing in smoothing spline analysis of variance models, *Technical Report 1072*, Department of Statistics, University of Wisconsin. (available at <http://www.stat.wisc.edu/~yilin/>)
- [13] Shen, X. and Ye, J. (2002), Adaptive model selection, *J. Amer. Statist. Assoc.*, **97**, 210-221.
- [14] Stamey, T., Kabalin, J., McNeal, J., Johnston, I., Freiha, F., Redwine, E. and Yang, N. (1989), Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate, ii: Radical prostatectomy treated patients, *J. Urol.*, **16**, 1076-1083.
- [15] Tibshirani, R. (1996), Regression shrinkage and selection via the lasso, *J. Royal. Statist. Soc. B.*, **58**, 267-288.
- [16] Wahba, G. (1990), *Spline Models for Observational Data*, SIAM, Philadelphia.
- [17] Yuan, M. and Lin, Y. (2003), Efficient empirical Bayes variable selection and estimation in linear models, to appear in *J. Amer. Statist. Assoc.* (available at <http://www.isye.gatech.edu/~myuan/>)

- [18] Yuan, M. and Lin, Y. (2004), Model selection and estimation in regression with grouped variables, to appear in *J. Royal. Statist. Soc. B.* (available at <http://www.isye.gatech.edu/~myuan/>)
- [19] Zou, H. and Hastie, T. (2005), Regularization and variable selection via the elastic net, *J. Royal. Statist. Soc. B.*, **67**, 301-320.