J. R. Statist. Soc. B (2007)
69, *Part* 2, *pp.* 143–161

On the non-negative garrotte estimator

Ming Yuan

Georgia Institute of Technology, Atlanta, USA

and Yi Lin

University of Wisconsin—Madison, USA

[Received February 2006. Revised October 2006]

Summary. We study the non-negative garrotte estimator from three different aspects: consistency, computation and flexibility. We argue that the non-negative garrotte is a general procedure that can be used in combination with estimators other than the original least squares estimator as in its original form. In particular, we consider using the lasso, the elastic net and ridge regression along with ordinary least squares as the initial estimate in the non-negative garrotte. We prove that the non-negative garrotte has the nice property that, with probability tending to 1, the solution path contains an estimate that correctly identifies the set of important variables and is consistent for the coefficients of the important variables, whereas such a property may not be valid for the initial estimators. In general, we show that the non-negative garrotte can turn a consistent estimate into an estimate that is not only consistent in terms of estimation but also in terms of variable selection. We also show that the non-negative garrotte has a piecewise linear solution path. Using this fact, we propose an efficient algorithm for computing the whole solution path for the non-negative garrotte. Simulations and a real example demonstrate that the non-negative garrotte is very effective in improving on the initial estimator in terms of variable selection and estimation accuracy.

Keywords: Elastic net; Lasso; Least angle regression selection; Non-negative garrotte; Path consistency; Piecewise linear solution path

1. Introduction

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

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

where $X = (X_1, X_2, \ldots, X_p)$, $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ and $\beta = (\beta_1, \ldots, \beta_p)'$. Throughout this paper, we centre each input variable so that the observed mean is 0, and scale each predictor so that the sample standard deviation is 1. 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 non-zero. By effectively identifying the subset of important predictors, variable selection can improve the accuracy of estimation and enhance model interpretability.

Classical variable selection methods, such as C_p, Akaike's information criterion AIC and the Bayes information criterion BIC, choose between possible models by using penalized sum-of-squares criteria, with the penalty being an increasing function of the dimension of the model. These methods, however, are computationally infeasible for even moderate

Address for correspondence: Ming Yuan, School of Industrial and Systems Engineering, Georgia Institute of Technology, 755 Ferst Drive North-West, Atlanta, GA 30332, USA.
E-mail: myuan@isye.gatech.edu

1369–7412/07/69143

numbers of predictors since the number of candidate models increases exponentially as the numbers 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 *et al.*, 1998). Various 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 *et al.*, 2004; Yuan and Lin, 2005; Zou and Hastie, 2005). In particular, Breiman (1995, 1996) proposed the non-negative garrotte estimator, which he showed to be a stable variable selection method that often outperforms its competitors including subset regression and ridge regression.

The original non-negative garrotte estimator that was introduced by Breiman (1995) is a scaled version of the least square estimate. The shrinking factor $d(\lambda)=(d_1(\lambda),\ldots,d_p(\lambda))'$ is given as the minimizer to

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

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

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

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

Therefore, for those coefficients whose full least square estimate is large, 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 of being exactly 0.

A drawback of the original non-negative garrotte is its explicit reliance on the full least squares estimate. Obviously, with a small sample size, least squares may perform poorly, and the non-negative garrotte is expected to suffer as well. In particular, the original non-negative garrotte, as proposed by Breiman (1995) cannot be applied when the sample size is smaller than the number of predictors. However, we argue that the idea of the non-negative garrotte can also be used in combination with estimators other than least squares. We consider, in particular, using the lasso (Tibshirani, 1996), ridge regression and the elastic net (Zou and Hastie, 2005) as alternative initial estimates for the non-negative garrotte. We prove that, as long as the initial estimate is consistent in terms of estimation, the non-negative garrotte estimate is consistent in terms of both estimation and model selection given that the tuning parameter λ is appropriately chosen. In other words, the non-negative garrotte can turn a consistent estimate into an estimate that is not only consistent in terms of estimation but also in terms of variable selection. In contrast, such a path consistency property does not always hold for the initial estimators.

A potential hurdle when using the non-negative garrotte estimator for large scale problems is the computational cost. The non-negative garrotte is so far computed by 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 is to be considered. In this paper, we show that the solution path of the non-negative garrotte is piecewise linear, regardless of the initial estimate, and use this to construct a more efficient algorithm for building the non-negative garrotte solution path. The algorithm proposed computes the whole solution path of the non-negative garrotte with the computational load of the same magnitude as ordinary least squares.

The rest of the paper is organized as follows. In the next section, we investigate the path consistency of the non-negative garrotte estimator as well as several other popular variable selection and estimation methods. An efficient algorithm for computing the non-negative garrotte solution path is introduced in Section 3. Sections 4 and 5 present some simulations and a real data example to support the theoretical results. We conclude the paper with a summary in Section 6. All technical proofs are relegated to Appendix A.

2. Path consistency

After an initial estimate has been obtained, the non-negative garrotte proceeds in two steps in practice. First the solution path that is indexed by the tuning parameter λ is constructed. The second step, which is often referred to as tuning, selects the final estimate on the solution path. 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 $\hat{\beta}$ 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. In this section, we investigate the path consistency property for the non-negative garrotte together with several other popular variable selection methods, namely, the lasso, least angle regression selection (LARS) (Efron *et al.*, 2004) and the elastic net.

2.1. Non-negative garrotte

The following theorem states that the non-negative garrotte is always path consistent given that the initial estimate is consistent in estimation.

Theorem 1. Assume that

(a) the initial estimate is δ_n consistent, i.e. $\max_j|\hat{\beta}_j^{\text{init}} - \beta_j| = O_p(\delta_n)$ for some $\delta_n \to 0$ and
(b) the design matrix is non-degenerate, i.e. the smallest eigenvalue of $X'X/n$ is bounded from below by a positive constant with probability tending to 1.

If $\lambda \to 0$ in a fashion such that $\delta_n = o(\lambda)$, then $P\{\hat{\beta}_j^{\text{NG}}(\lambda) = 0\} \to 1$ for any $j \notin \mathcal{I}$, and $\hat{\beta}_j^{\text{NG}}(\lambda) = \beta_j + O_p(\lambda)$ for any $j \in \mathcal{I}$ where $\mathcal{I} = \{j : \beta_j \neq 0\}$.

During the preparation of this paper, it was brought to our attention that Zou (2006) also independently obtained that the original non-negative garrotte with the least squares estimate as the initial estimate is consistent in variable selection if p is held fixed as $n \to \infty$. His result can be viewed as a special case of theorem 1 with the choice $\delta_n = \surd n$. Theorem 1 is more general because it also indicates that the non-negative garrotte estimator is consistent in estimation. More importantly, it is worth pointing out that we do not require p to be held fixed and allow for more general initial estimates in theorem 1.

In achieving the consistency in variable selection, we show in theorem 1 that the non-negative garrotte estimate of a non-zero coefficient converges at a slower rate than its initial estimate. It is not clear to us whether this is the unavoidable price that we must pay for variable selection in general. The numerical results that are presented in Section 5 clearly suggest otherwise but theoretical justification has so far eluded us. As a partial answer, the following lemma demonstrates that sharper convergence rates may be available for the coefficient estimate at least in some special cases.

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

This path consistency property of the non-negative garrotte is to be contrasted with the following results for several other modern variable selection methods.

2.2. Lasso

The popular lasso that was proposed by Tibshirani (1996) is defined as

$$\hat{\beta}^{\text{LASSO}}(\lambda) = \arg\min_{\beta} \left(\tfrac{1}{2} \|Y - X\beta\|^2 + n\lambda \|\beta\|_{l_1} \right), \tag{4}$$

where λ is a tuning parameter and $\|\cdot\|_{l_1}$ stands for the vector l_1-norm. By using the l_1-penalty, minimizing equation (4) yields a sparse estimate of β if λ is chosen appropriately. Consequently, a submodel of equation (1) which contains only the covariates corresponding to the non-zero components in $\hat{\beta}^{\text{LASSO}}(\lambda)$ is selected as the final model. The lasso has exploded in popularity since its introduction because of its great success in a wide range of applications.

Despite its superior performance in prediction, the following theorem suggests that the lasso must be used with caution as a variable selection method. The path consistency property holds for the lasso only under restrictive conditions of the design matrix.

Theorem 2. The sufficient and necessary condition for the lasso to be path consistent is

$$\max_{j \notin \mathcal{I}} \{\text{cov}(X_j, X_{\mathcal{I}}) \text{ cov } (X_{\mathcal{I}})^{-1}\} s_{\mathcal{I}} < 1, \tag{5}$$

where s is a p-dimensional vector with the jth element being $\text{sgn}(\beta_j)$.

The fact that the lasso may not be consistent in variable selection was first noted in Meinshausen and Bühlmann (2006) who, in the context of Gaussian graphical model selection, argued that a condition similar to inequality (5) is required to ensure the consistency in variable selection for a lasso-type procedure. Several other researchers have also independently discovered results that are similar to theorem 2 during the preparation of this paper. In particular, Zou (2006) showed that a necessary condition for the lasso to be consistent in variable selection is

$$\max_{j \notin \mathcal{I}} \{\text{cov}(X_j, X_{\mathcal{I}}) \text{ cov}(X_{\mathcal{I}})^{-1}\} s^*_{\mathcal{I}} \leqslant 1, \tag{6}$$

for some sign vector s^*. The necessity result that is reported here is stronger in that it implies condition (6). Zhao and Yu (2006) also considered conditions that were similar to inequality (5), but their focus was on sign consistency, i.e. $\text{sgn}(\hat{\beta}_j)$ agrees with the sign of the true regression coefficient. The necessity of condition (5) in our theorem 2 follows directly from their theorem 2 because sign consistency is weaker than consistency in both variable selection and estimation. Similarly, their sufficiency result for sign consistency also follows immediately from the sufficiency of condition (5) in our theorem 2. For this reason, we omit the proof of the necessity and only present that of the sufficiency in Appendix A.

Theorem 2 indicates that, if condition (5) is not satisfied, we cannot use the lasso to select the right variables even with the freedom of choosing the tuning parameter λ. It is therefore of clear importance to be able to determine in practice when the lasso can be used for variable selection. Of course the condition that is given in theorem 2 cannot be checked since it involves the true regression coefficient β. For this purpose, a stronger condition can be enforced to ensure the path consistency of the lasso:

$$\max_{j\notin\mathcal{I}} \|\mathrm{cov}(X_j, X_{\mathcal{I}})\,\mathrm{cov}(X_{\mathcal{I}})^{-1}\|_{l_1} < 1. \tag{7}$$

In fact, following the same proof as that of theorem 2, one can show that inequality (7) 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 $j \in \mathcal{I}$. In contrast, it is easy to see that, if condition (7) is violated, then there is always a β such that condition (5) is not satisfied. By theorem 2, the lasso is not path consistent at least for such β.

2.3. *Least angle regression selection*

The LARS method that was proposed by Efron *et al.* (2004) is a method which is closely related to the lasso. The LARS method uses a variable selection strategy which is similar to forward selection. Starting with all coefficients equal to 0, the algorithm finds the predictor that is most correlated with the response variable and proceeds in this direction. Instead of taking a full step towards the projection of Y on the variable, as would be done in forward selection, the LARS algorithm only takes the largest step possible in this direction until some other variable has as much correlation with the current residual. Then this new predictor is entered and the process is continued. Readers are referred to Efron *et al.* (2004) and Osborne *et al.* (2000) for the detailed LARS algorithm. The great computational advantage of LARS comes from the fact that the LARS path is piecewise linear and all that we need to do is to locate the changepoints. Once a variable enters the LARS solution path, it will stay on the solution path. Therefore, the LARS method cannot be path consistent if a redundant variable is the first to be selected.

Theorem 3. If

$$\max_{j\notin\mathcal{I}} |\mathrm{cov}(X_j, X_{\mathcal{I}})\beta| \geqslant \max_{j\in\mathcal{I}} |\mathrm{cov}(X_j, X_{\mathcal{I}})\beta|, \tag{8}$$

then the LARS method is not path consistent with a non-vanishing probability.

2.4. *Elastic net*

The elastic net was recently proposed by Zou and Hastie (2005) to combine the strength of the lasso and ridge regression. The elastic net estimate is defined as

$$\hat\beta^{\mathrm{ENET}}(\lambda) = \arg\min_{\beta}\left(\frac{1}{2}\|Y - X\beta\|^2 + n\lambda\|\beta\|_{l_1} + \frac{n}{2}\tau\|\beta\|_{l_2}^2\right), \tag{9}$$

where λ and τ are tuning parameters and $\|\cdot\|_{l_2}$ stands for the vector l_2-norm. Clearly, the elastic net has both the lasso ($\tau=0$) and ridge regression ($\lambda=0$) as special cases. Similar to the lasso, the l_1-penalty encourages sparse estimates of β, and the squared l_2-penalty encourages highly correlated predictors to have similar coefficient estimates. It has been demonstrated in Zou and Hastie (2005) that the elastic net often enjoys better prediction performance than both the lasso and ridge regression in simulations. Like the lasso and LARS, the elastic net is not always path consistent.

Theorem 4. A necessary and sufficient condition for the elastic net to be path consistent is

$$\max_{j\notin\mathcal{I}}\left(\liminf_{c_1,c_2\to 0^+}\left[\mathrm{cov}(X_j, X_{\mathcal{I}})\,\{\mathrm{cov}(X_{\mathcal{I}}) + c_1 I\}^{-1}\left(s_{\mathcal{I}} + \frac{c_1}{c_2}\beta_{\mathcal{I}}\right)\right]\right) < 1. \tag{10}$$

To gain further insight into conditions (10), consider the special case when $\mathrm{cov}(X_{\mathcal{I}}) = I$ and $\beta_{\mathcal{I}} = b s_{\mathcal{I}}$ for some scalar $b > 0$. The left-hand side of inequality (10) becomes

$$\max_{j\notin\mathcal{I}}\left(\liminf_{c_1,c_2\to 0^+}\left[\operatorname{cov}(X_j,X_{\mathcal{I}})\,\{\operatorname{cov}(X_{\mathcal{I}})+c_1 I\}^{-1}\left(s_{\mathcal{I}}+\frac{c_1}{c_2}\beta_{\mathcal{I}}\right)\right]\right)$$
$$=\max_{j\notin\mathcal{I}}\left\{\liminf_{c_1,c_2\to 0^+}\left(\frac{1+bc_1/c_2}{1+c_2}\right)\operatorname{cov}(X_j,X_{\mathcal{I}})\operatorname{cov}(X_{\mathcal{I}})^{-1}s_{\mathcal{I}}\right\}$$
$$=\max_{j\notin\mathcal{I}}\{\operatorname{cov}(X_j,X_{\mathcal{I}})\operatorname{cov}(X_{\mathcal{I}})^{-1}\}s_{\mathcal{I}}. \tag{11}$$

Now condition (10) is the same as condition (5).

3. Algorithm

For most methods of regularization, it is very expensive, if not impossible, to compute the exact solution path. We must approximate the solution path by evaluating the estimate for a fine grid of tuning parameters and there is a trade-off between the accuracy of the approximation and the computational cost in determining how fine a grid of tuning parameters to be considered. In particular, given the initial estimate, the non-negative garrotte solution path can be approximated by solving the quadratic programming problem (2) for a series of λs, as done in Breiman (1995). We show that, similarly to the lasso, the solution path of the non-negative garrotte is piecewise linear, and we use this to construct an efficient algorithm of building the exact non-negative garrotte solution path.

Let $\hat{\beta}^{\text{NG}}(\lambda)=(d_1(\lambda)\beta_1^{\text{init}},d_2(\lambda)\beta_2^{\text{init}},\ldots,d_p(\lambda)\beta_p^{\text{init}})$ be the solution of equation (2) where β^{init} is the initial estimate. A simple application of the Karush–Kuhn–Tucker condition yields

$$\frac{1}{n}\beta_j^{\text{init}\prime}X_j'\{Y-X\hat{\beta}^{\text{NG}}(\lambda)\}=\lambda, \qquad \text{if } \hat{\beta}_j^{\text{NG}}(\lambda)\neq 0, \tag{12}$$

$$\frac{1}{n}\beta_j^{\text{init}\prime}X_j'\{Y-X\hat{\beta}^{\text{NG}}(\lambda)\}<\lambda, \qquad \text{if } \hat{\beta}_j^{\text{NG}}(\lambda)=0. \tag{13}$$

Such characteristics of the solution path are similar to those of LARS and can be used to build the solution path. Starting with all coefficients equal to 0, the algorithm finds the predictor such that the covariance between $X_j\beta_j^{\text{init}}$ and the response variable is maximized and proceeds in this direction. Then, we can take the largest step possible in this direction until one of the following situations occurs:

(a) some other variable enters the model because it also maximizes the covariance between $X_j\beta_j^{\text{init}}$ and the current residual;
(b) a variable should be dropped because the non-negativity constraint $d_j\geqslant 0$ would be violated if we continue in this direction.

It turns out that both situations can be handled with ease. The former can be dealt with by adding this variable to the model and recomputing the direction with the updated set of variables so that conditions (12) and (13) continue to hold. The latter occurs if a non-zero coefficient reaches 0. In this case, we can simply drop the variable and again recompute the direction with the updated set of variables. To sum up, we propose the following algorithm to compute the non-negative garrotte solution path.

3.1. Algorithm—non-negative garrotte

Step 1: start from $d^{[0]}=0$, $k=1$ and $r^{[0]}=Y$.
Step 2: compute the current active set

$$\mathcal{C}_1 = \arg\max_j (Z_j' r^{[k-1]}),$$

where $Z_j = X_j \beta_j^{\text{init}}$.

Step 3: compute the current direction γ, which is a p-dimensional vector defined by $\gamma_{\mathcal{C}_k^c} = 0$ and

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

Step 4: for every $j \notin \mathcal{C}_k$, compute how far the group non-negative garrotte will progress in direction γ before X_j enters the active set. This can be measured by an α_j such that

$$Z_j'(r^{[k-1]} - \alpha_j Z\gamma) = Z_{j'}'(r^{[k-1]} - \alpha_j Z\gamma) \tag{14}$$

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

Step 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 non-negative, measures how far the group non-negative garrotte will progress before d_j becomes 0.

Step 6: if $\alpha_j \leqslant 0$, $\forall j$ 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 $\mathcal{C}_{k+1} = \mathcal{C}_k \cup \{j^*\}$; otherwise update $\mathcal{C}_{k+1} = \mathcal{C}_k - \{j^*\}$.

Step 7: set $r^{[k]} = Y - Zd^{[k]}$ and $k = k + 1$. Go back to step 3 until $\alpha = 1$.

Such an algorithm is quite similar to the LARS or the modified LARS algorithm (Efron *et al.*, 2004) for the LASSO and has a computational cost that is of the same magnitude as ordinary least squares. A complicating factor for the non-negative garrotte is the non-negativity constraints in model (2). We shall show in the next theorem that these constraints are automatically enforced and the whole solution path of the non-negative garrotte indeed can be constructed by using the procedure that was described above.

Theorem 5. Under the 'one at a time' condition that is discussed below, the trajectory of this algorithm coincides with the non-negative garrotte 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 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

(a) the addition occurs only for one variable a time at any stage of the above algorithm,
(b) no variable vanishes at the time of addition and
(c) 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, readers are referred to Efron *et al.* (2004).

4. Simulation

In this section, we investigate the finite sample properties of the non-negative garrotte estimator. Our discussion in the previous sections applies to any consistent estimate as the initial estimate. In practice, the accuracy of the non-negative garrotte estimate depends on the choice of the initial estimate. We consider four choices of the initial estimate in this paper: the ordinary least squares estimate, the ridge estimate, the lasso and the elastic net. Except for the ordinary least squares estimate, the other initial estimates all involve tuning parameters, which are chosen by tenfold cross-validation in our numerical examples.

4.1. Example 1

In the first set of simulations, we demonstrate the path consistency of the non-negative garrotte procedure in contrast with the lasso, LARS and the elastic net. We consider a simple model:

$$Y = X_1 + X_2 + 0 \cdot X_3 + \varepsilon, \tag{15}$$

where $\varepsilon \sim \mathcal{N}(0,1)$. 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. Conditionally 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 αs: 0.35, 0.45, 0.55 and 0.65. For each α-value, we consider 20 equally spaced sample sizes: $25, 50, \ldots, 500$. For each combination of α and sample size, 100 data sets were simulated, and we counted how many times different solution paths cover the true model, i.e. how many times the path contains at least one estimate $\hat{\beta}$ such that $\hat{\beta}_1 \neq 0$, $\hat{\beta}_2 \neq 0$ and $\hat{\beta}_3 = 0$. To fix ideas, we consider only the ordinary least squares estimate as the initial estimate of the non-negative garrotte in this example. Fig. 1 depicts the frequency for each method to cover the true model. It is worth noting that, in this example, the elastic net and lasso have indistinguishable performance, which can be expected from the equivalence between conditions (5) and (10) in this case.

When $\alpha = 0.35$, all estimating procedures are consistent in variable selection. But the non-negative garrotte selects the correct model more often than the others for smaller sample sizes. When α increases, the convergence of the coverage probabilities for both the non-negative garrotte and the other methods slows down. For $\alpha = 0.55$ or $\alpha = 0.65$, the lasso, LARS and the elastic net do not seem to be consistent in variable selection any more. In contrast, the non-negative garrotte is still very capable of selecting the right model for α as large as 0.65. It is worth pointing out that such empirical evidence agrees with our theoretical results that were presented before. According to theorems 2–4, the lasso, LARS and the elastic net can be path consistent only if $\alpha < 0.5$.

4.2. Example 2

In the second set of simulations, we consider a model that was used in Tibshirani (1996). 20, 50 or 100 observations were simulated from model (1) where $p = 8$, $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)'$ and $\sigma = 3$. The correlation between X_i and X_j is $\rho^{|i-j|}$ with $\rho = 0.5$. For the non-negative garrotte, we use the algorithm that was presented in Section 3 to construct the non-negative garrotte solution path and, following Yuan and Lin (2006), we use the following C_p-type criterion to determine λ:

$$C_p(\hat{\mu}) = \frac{\|Y - \hat{\mu}\|^2}{\sigma^2} - n + 2\tilde{\mathrm{df}}_{\mu,\sigma^2}, \tag{16}$$

where

$$\tilde{\mathrm{df}} = 2 \sum_j I(d_j > 0) - \sum_j d_j.$$

For the lasso, LARS and the elastic net, tenfold cross-validation was used to determine the corresponding tuning parameters. For each sample size, we repeat the experiment 200 times and compare different methods in terms of model error, model size and false positive and false negative results in variable selection. The model error of an estimate $\hat{\beta}$ is given by

$$\mathrm{ME}(\hat{\beta}) = (\hat{\beta} - \beta)' V (\hat{\beta} - \beta),$$

where $V = E(X'X)$ is the population covariance matrix of X.

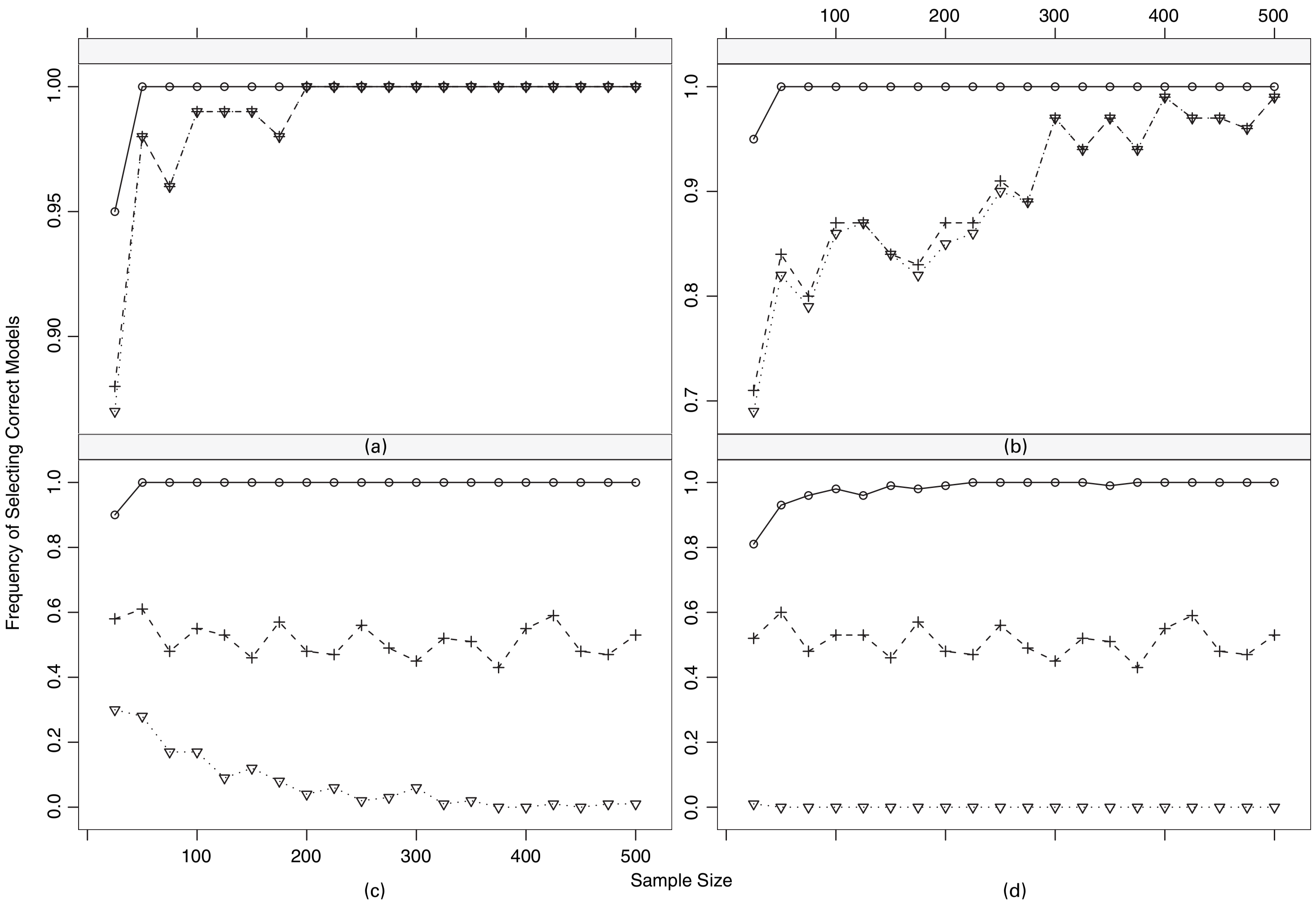

Fig. 1. Path consistency of the non-negative garrotte in contrast with other methods (∘, non-negative garrotte; +, lasso or elastic net; ∇, LARS): (a) $\alpha = 0.35$; (b) $\alpha = 0.45$; (c) $\alpha = 0.55$; (d) $\alpha = 0.65$

Table 1. Simulation example 2—averaged model error ME, model size SIZE, false positive results FP and false negative results FN based on 200 runs†

Method	*Results for the following values of n:*											
	$n=20$				$n=50$				$n=100$			
	ME	*SIZE*	*FP*	*FN*	*ME*	*SIZE*	*FP*	*FN*	*ME*	*SIZE*	*FP*	*FN*
LASSO	4.41	4.22	1.56	0.34	1.52	4.92	1.92	0.00	0.67	5.24	2.24	0.00
	(0.21)	(0.12)	(0.10)	(0.04)	(0.06)	(0.11)	(0.11)	(0.00)	(0.03)	(0.11)	(0.11)	(0.00)
GLASSO	4.07	2.98	0.64	0.66	1.21	3.48	0.64	0.16	0.55	3.64	0.64	0.00
	(0.21)	(0.09)	(0.07)	(0.05)	(0.06)	(0.06)	(0.06)	(0.03)	(0.03)	(0.07)	(0.07)	(0.00)
RIDGE	5.76	8.00	5.00	0.00	1.83	8.00	5.00	0.00	0.88	8.00	5.00	0.00
	(0.29)	(0.00)	(0.00)	(0.00)	(0.06)	(0.00)	(0.00)	(0.00)	(0.03)	(0.00)	(0.00)	(0.00)
GRIDGE	5.09	4.26	1.74	0.48	1.36	4.10	1.22	0.12	0.61	4.02	1.02	0.00
	(0.31)	(0.14)	(0.13)	(0.05)	(0.07)	(0.11)	(0.11)	(0.02)	(0.02)	(0.09)	(0.09)	(0.00)
ENET	4.05	4.70	1.98	0.28	1.60	5.18	2.18	0.00	0.74	5.20	2.20	0.00
	(0.19)	(0.12)	(0.11)	(0.04)	(0.08)	(0.11)	(0.11)	(0.00)	(0.04)	(0.11)	(0.11)	(0.00)
GENET	3.94	3.34	0.90	0.56	1.22	3.92	1.02	0.10	0.56	3.80	0.80	0.00
	(0.19)	(0.10)	(0.08)	(0.05)	(0.06)	(0.10)	(0.10)	(0.02)	(0.03)	(0.08)	(0.08)	(0.00)
OLS	5.83	8.00	5.00	0.00	1.83	8.00	5.00	0.00	0.88	8.00	5.00	0.00
	(0.29)	(0.00)	(0.00)	(0.00)	(0.06)	(0.00)	(0.00)	(0.00)	(0.03)	(0.00)	(0.00)	(0.00)
GOLS	5.07	4.24	1.72	0.48	1.36	4.10	1.22	0.12	0.61	4.02	1.02	0.00
	(0.31)	(0.15)	(0.13)	(0.05)	(0.07)	(0.11)	(0.11)	(0.02)	(0.02)	(0.09)	(0.09)	(0.00)

†OLS, ordinary least squares.

Table 1 summarizes the results from the simulation. We use ENET to denote the elastic net and prefix G to indicate the non-negative garrotte estimate with certain initial estimates. The figures in parentheses are the standard errors. A few observations can be made from Table 1. The true model contains a moderate number of moderate size effects, and the signal-to-noise ratio is approximately 5.7. In terms of the accuracy of estimation, all versions of the non-negative garrotte improve on the initial estimate. It is also clear from Table 1 that the non-negative garrotte is more effective in reducing both the false positive and the false negative results.

4.3. Example 3

The set-up of the third example is the same as for example 2 except that the true regression coefficients are $\beta_j = 0.85$, $j = 1, 2, \ldots, 8$. The true model contains all variables each with a small effect, and the signal-to-noise ratio is approximately 1.7. Table 2 documents the results from the simulation. On the basis of Table 2, the non-negative garrotte tends to be less accurate than the initial estimates because it often selects models with sizes that are too small. It is worth pointing out that such suboptimal performance is not in contradiction to our theoretical results and may be partially attributed to the inefficiency of the tuning criterion.

4.4. Example 4

The set-up of the last example is also the same as for example 2 except that the true regression coefficient is $\beta = (5, 0.5, 0.5, 0.5, 0, 0, 0, 0)'$. The true model contains one large effect and several small effects. Table 3 gives a summary of the simulation results. A clear advantage of the non-negative garrotte over its initial estimate can be observed. Note that, as the sample size increases, the number of false negative results from the non-negative garrotte reduces.

Table 2. Simulation example 3—averaged model error ME, model size SIZE, false positive results FP and false negative results FN based on 200 runs†

Method	*Results for the following values of n:*											
	n = 20				*n = 50*				*n = 100*			
	ME	*SIZE*	*FP*	*FN*	*ME*	*SIZE*	*FP*	*FN*	*ME*	*SIZE*	*FP*	*FN*
LASSO	5.35	4.98	0.00	3.02	1.82	7.14	0.00	0.86	0.86	7.90	0.00	0.10
	(0.24)	(0.15)	(0.00)	(0.15)	(0.07)	(0.07)	(0.00)	(0.07)	(0.03)	(0.02)	(0.00)	(0.02)
GLASSO	5.95	3.16	0.00	4.84	2.42	5.34	0.00	2.66	1.06	7.08	0.00	0.92
	(0.21)	(0.12)	(0.00)	(0.12)	(0.07)	(0.09)	(0.00)	(0.09)	(0.05)	(0.08)	(0.00)	(0.08)
RIDGE	5.67	8.00	0.00	0.00	1.75	8.00	0.00	0.00	0.83	8.00	0.00	0.00
	(0.27)	(0.00)	(0.00)	(0.00)	(0.06)	(0.00)	(0.00)	(0.00)	(0.03)	(0.00)	(0.00)	(0.00)
GRIDGE	5.69	4.16	0.00	3.84	2.38	5.60	0.00	2.40	1.03	7.12	0.00	0.88
	(0.19)	(0.14)	(0.00)	(0.14)	(0.08)	(0.09)	(0.00)	(0.09)	(0.04)	(0.07)	(0.00)	(0.07)
ENET	4.57	5.70	0.00	2.30	1.73	7.26	0.00	0.74	0.88	7.86	0.00	0.14
	(0.20)	(0.15)	(0.00)	(0.15)	(0.07)	(0.06)	(0.00)	(0.06)	(0.03)	(0.02)	(0.00)	(0.02)
GENET	5.86	3.74	0.00	4.26	2.22	5.74	0.00	2.26	1.04	7.20	0.00	0.80
	(0.20)	(0.14)	(0.00)	(0.14)	(0.07)	(0.09)	(0.00)	(0.09)	(0.04)	(0.07)	(0.00)	(0.07)
OLS	5.74	8.00	0.00	0.00	1.75	8.00	0.00	0.00	0.83	8.00	0.00	0.00
	(0.27)	(0.00)	(0.00)	(0.00)	(0.06)	(0.00)	(0.00)	(0.00)	(0.03)	(0.00)	(0.00)	(0.00)
GOLS	5.69	4.14	0.00	3.86	2.38	5.60	0.00	2.40	1.03	7.12	0.00	0.88
	(0.19)	(0.14)	(0.00)	(0.14)	(0.08)	(0.09)	(0.00)	(0.09)	(0.04)	(0.07)	(0.00)	(0.07)

†OLS, ordinary least squares.

Table 3. Simulation example 4—averaged model error ME, model size SIZE, false positive results FP and false negative results FN based on 200 runs†

Method	*Results for the following values of n:*											
	n = 20				*n = 50*				*n = 100*			
	ME	*SIZE*	*FP*	*FN*	*ME*	*SIZE*	*FP*	*FN*	*ME*	*SIZE*	*FP*	*FN*
LASSO	4.08	3.34	1.04	1.70	1.62	4.04	1.28	1.24	0.72	4.98	1.50	0.52
	(0.20)	(0.13)	(0.09)	(0.06)	(0.08)	(0.13)	(0.09)	(0.06)	(0.03)	(0.10)	(0.08)	(0.04)
GLASSO	2.72	2.12	0.44	2.32	1.32	2.54	0.56	2.02	0.65	2.94	0.40	1.46
	(0.15)	(0.10)	(0.06)	(0.05)	(0.06)	(0.09)	(0.06)	(0.05)	(0.02)	(0.08)	(0.05)	(0.05)
RIDGE	5.96	8.00	4.00	0.00	1.96	8.00	4.00	0.00	0.89	8.00	4.00	0.00
	(0.28)	(0.00)	(0.00)	(0.00)	(0.07)	(0.00)	(0.00)	(0.00)	(0.03)	(0.00)	(0.00)	(0.00)
GRIDGE	3.53	3.54	1.42	1.88	1.34	3.08	0.90	1.82	0.69	3.48	0.78	1.30
	(0.23)	(0.18)	(0.11)	(0.08)	(0.06)	(0.13)	(0.09)	(0.06)	(0.02)	(0.11)	(0.08)	(0.06)
ENET	4.35	3.74	1.32	1.58	1.70	4.00	1.14	1.14	0.80	4.94	1.40	0.46
	(0.22)	(0.14)	(0.10)	(0.07)	(0.09)	(0.11)	(0.08)	(0.06)	(0.03)	(0.12)	(0.10)	(0.04)
GENET	2.70	2.24	0.54	2.30	1.29	2.64	0.54	1.90	0.66	3.04	0.40	1.36
	(0.16)	(0.10)	(0.07)	(0.05)	(0.05)	(0.10)	(0.06)	(0.05)	(0.02)	(0.09)	(0.05)	(0.05)
OLS	6.02	8.00	4.00	0.00	1.96	8.00	4.00	0.00	0.89	8.00	4.00	0.00
	(0.29)	(0.00)	(0.00)	(0.00)	(0.07)	(0.00)	(0.00)	(0.00)	(0.03)	(0.00)	(0.00)	(0.00)
GOLS	3.47	3.52	1.42	1.90	1.34	3.08	0.90	1.82	0.69	3.48	0.78	1.30
	(0.23)	(0.17)	(0.11)	(0.07)	(0.06)	(0.13)	(0.09)	(0.06)	(0.02)	(0.11)	(0.08)	(0.06)

†OLS, ordinary least squares.

In summary, we found from examples 2–4 that the non-negative garrotte performs very well when the true model is relatively sparse, and it should be favourable in many applications on the basis of the bet-on-sparsity principle that was advocated by Friedman *et al.* (2004).

5. Real data

To illustrate our results further, we reanalyse the prostate cancer data set from the study by Stamey *et al.* (1989). This data set, which was previously used in Tibshirani (1996), consists of 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. Figs 2(a), 2(c) and 2(e) give the solution paths of the lasso and LARS (the lasso and LARS share the same solution path in this example), the non-negative garrotte and the elastic net. For illustration, we used the ordinary least squares estimate as the initial estimate for the non-negative garrotte. For the elastic net, as suggested in Zou and Hastie (2005), we fix $n\tau$ at 1000 and the solution path corresponds to different values of λ. In each panel, the dotted vertical line indicates the tuning parameter that was chosen by tenfold 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 2, we replace gleason with an artificial variable 2lcavol + 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 four methods on the new data set are given in Figs 2(b), 2(d) and 2(f). Comparing with the original solution path, the non-negative garrotte is the least disturbed by such change. Both the lasso and LARS and the elastic net select the artificial variable as an important predictor. This observation supports the theory from Section 2 that the path consistency of the lasso and LARS and the elastic net depends on the correlation of the design matrix whereas the non-negative garrotte is always path consistent.

To gain further insights, we estimate the prediction error of each method on both the original data set and the perturbed data set by using fivefold cross-validation. On the original data, the prediction error is 0.571, 0.558 and 0.623 respectively for the lasso and LARS, the non-negative garrotte and the elastic net. After modifying the gleason variable, the prediction error becomes 0.579, 0.560 and 0.666 respectively for the lasso and LARS, the non-negative garrotte and the elastic net. This agrees with our findings from Fig. 2.

6. Conclusion

In this paper we proved that the non-negative garrotte estimator is path consistent given an appropriate initial estimate. It can turn a consistent estimate into an estimate that is consistent in terms of both variable selection and coefficient estimation. We showed that the solution path of the non-negative garrotte is piecewise linear, and the whole path can be computed quickly. We have also shown by simulations and a real example that the non-negative garrotte is an effective tool to improve the variable selection and estimation accuracy of a given estimator. The encouraging results that are presented here suggest that the idea of the non-negative garrotte might be useful in a wider range of applications. For example, one can consider an extension to multivariate nonparametric regression and devise a variable selection and estimation procedure using the non-negative garrotte. Further studies are needed to explore this and other possibilities.

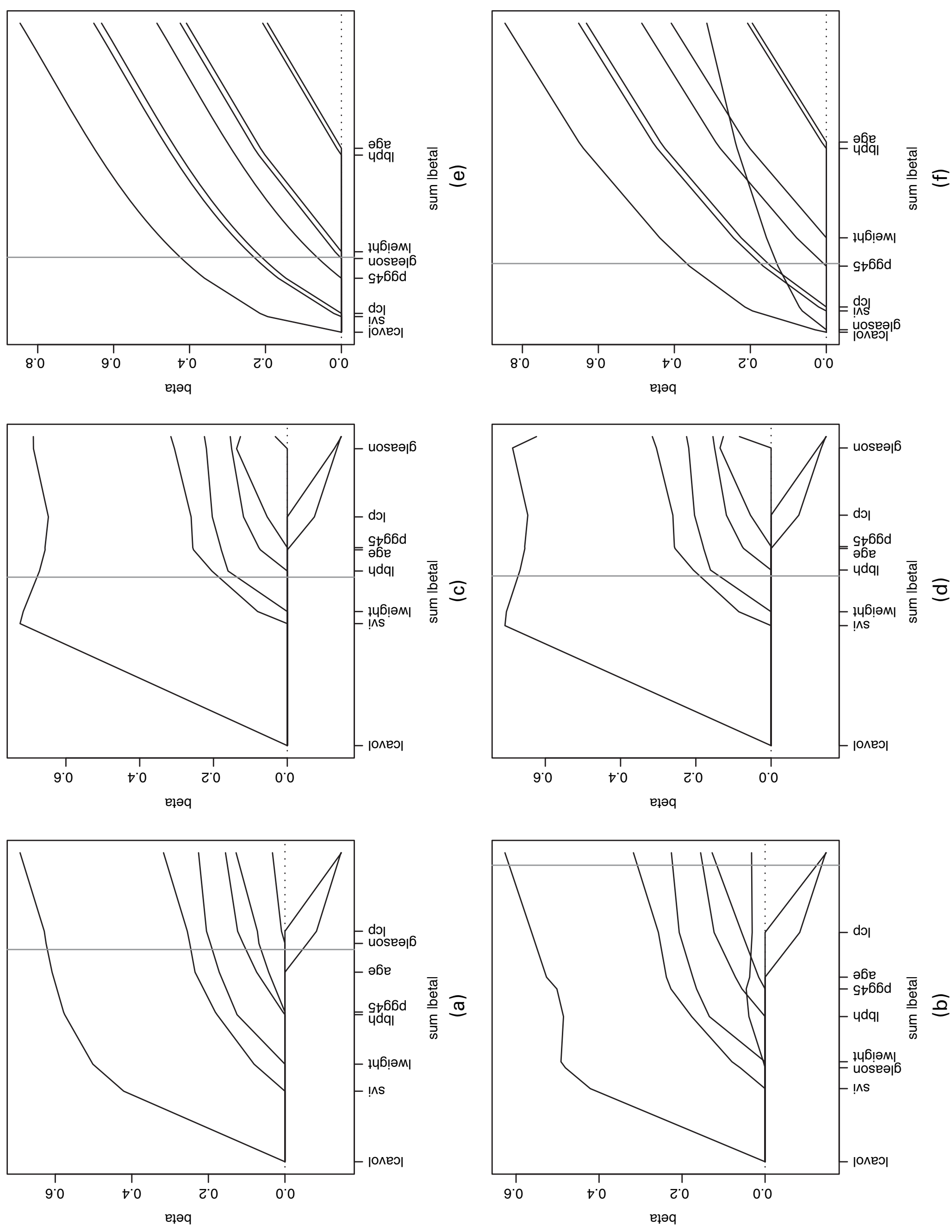

Fig. 2. Solution paths for the prostate cancer example: (a), (b) the lasso; (c), (d) the non-negative garrotte; (e), (f) the elastic net

Acknowledgements

Yuan's research was supported in part by National Science Foundation grant DMS-0624841. Lin's research was supported in part by National Science Foundation grant DMS-0134987.

Appendix A

A.1. Proof of theorem 1

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 $P(\mathcal{A})\to 0$ as $n\to\infty$. Write $d_{ij}=d_{\Lambda_{ij}}$, $i,j=0,1$, and let other vectors and matrices be defined in the same fashion unless otherwise indicated. Note that $d_{1.}$ is also the unconstrained minimizer of

$$\frac{1}{2}\|Y-Z_{1.}\gamma\|^2+n\lambda\sum_j\gamma_j, \tag{17}$$

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

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

Then

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

This implies that

$$d_{10}=-A^{-}_{00.1}A_{01}A^{-}_{11}(Z'_{11}Y/n-\lambda\mathbf{1}_{p_{11}})+A^{-}_{00.1}(Z'_{10}Y/n-\lambda\mathbf{1}_{p_{10}})\equiv A^{-}_{00.1}w. \tag{18}$$

Rewrite w as

$$w=Z'_{10}\{I_{p_{11}}-Z_{11}(Z'_{11}Z_{11})^{-}Z'_{11}\}Y/n-\lambda\mathbf{1}_{p_{10}}+\lambda A_{01}A^{-}_{11}\mathbf{1}_{p_{11}}. \tag{19}$$

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

$$\begin{aligned}|\hat{\beta}^{\text{init}}_i\hat{\beta}^{\text{init}}_j-\beta_i\beta_j|&=|\hat{\beta}^{\text{init}}_i(\hat{\beta}^{\text{init}}_j-\beta_j)+\beta_j(\hat{\beta}^{\text{init}}_i-\beta_i)|\\ &\leqslant(|\hat{\beta}^{\text{init}}_i|+|\beta_j|)|\hat{\beta}^{\text{init}}_j-\beta_j|\\ &=O_p(\delta_n).\end{aligned} \tag{20}$$

This entails

$$A_{11}=\frac{1}{n}\Delta_{11}X'_{11}X_{11}\Delta_{11}+O_p(\delta_n), \tag{21}$$

$$A_{01}=O_p(\delta_n), \tag{22}$$

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

$$w=Z'_{10}\{I_{p_{11}}-Z_{11}(Z'_{11}Z_{11})^{-}Z'_{11}\}Y/n-\lambda\{1+O_p(\delta_n)\}\mathbf{1}_{p_{10}}. \tag{23}$$

Now note that

$$\|\{I_{p_{11}} - Z_{11}(Z'_{11}Z_{11})^- Z'_{11}\}Y\|^2 \leqslant Y'Y = O_p(n), \tag{24}$$

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

$$\begin{aligned} \|Z'_{10}\{I_{p_{11}} - Z_{11}(Z'_{11}Z_{11})^- Z'_{11}\}Y\| &\leqslant \|Z_{10}\| \|\{I_{p_{11}} - Z_{11}(Z'_{11}Z_{11})^- Z'_{11}\}Y\| \\ &= O_p(\sqrt{n}\|Z_{10}\|) \\ &= O_p(n \max_{j \notin \mathcal{I}} |\hat{\beta}_j^{\text{init}}|) \\ &= O_p(n\delta_n) \\ &= o_p(n\lambda). \end{aligned} \tag{25}$$

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 is in contradiction with equation (18), which implies that $w'd_{10} = w'A_{00.1}^- w \geqslant 0$. Thus, when $n \to \infty$, $P(\mathcal{A}) \to 0$.

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

$$\tfrac{1}{2}\|Y - Z_{.1}\gamma\|^2 + n\lambda\gamma'\mathbf{1}_{p_{.1}}, \tag{26}$$

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

$$d^u = (Z'_{.1}Z_{.1}/n)^-(Z'_{.1}Y/n - \lambda\mathbf{1}_{p_{.1}}). \tag{27}$$

Following the same argument as for equation (21), we have

$$\frac{1}{n}Z'_{.1}Z_{.1} = \frac{1}{n}\Delta_{.1}X'_{.1}X_{.1}\Delta_{.1} + O_p(\delta_n). \tag{28}$$

Consequently,

$$d^u = (\Delta_{.1}X'_{.1}X_{.1}\Delta_{.1}/n)^-(Z'_{.1}Y/n - \lambda\mathbf{1}_{p_{.1}})\{1 + O_p(\delta_n)\}. \tag{29}$$

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

$$\left|\frac{1}{n}((Z_{.1} - X_{.1}\Delta_{.1})'Y)_j\right| = O\{|(\hat{\beta}_{.1}^{\text{init}} - \beta_{.1})_j|\} = O_p(\delta_n). \tag{30}$$

Thus,

$$d^u = (\Delta_{.1}X'_{.1}X_{.1}\Delta_{.1}/n)^-(\Delta_{.1}X'_{.1}Y/n - \lambda\mathbf{1}_{p_{.1}})\{1 + O_p(\delta_n)\}. \tag{31}$$

Combining equation (31) and the fact that

$$(\Delta_{.1}X'_{.1}X_{.1}\Delta_{.1}/n)^-\Delta_{.1}X'_{.1}Y/n = \mathbf{1}_{p_{.1}},$$

we obtain

$$\begin{aligned} d^u &= \mathbf{1}_{p_{.1}} - \lambda(\Delta_{.1}X'_{.1}X_{.1}\Delta_{.1}/n)^-\mathbf{1}_{p_{.1}} + O_p(\delta_n) \\ &= \mathbf{1}_{p_{.1}}\{1 + O_p(\lambda)\}. \end{aligned} \tag{32}$$

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

A.2. Proof of lemma 1

In the proof of lemma 1, the first term on the left-hand side of equation (23) 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 for the proof of theorem 1.

A.3. Proof of theorem 2

Recall that the lasso with tuning parameter λ is given as the minimizer to

$$\frac{1}{2}\|Y - X\gamma\|^2 + n\lambda\sum_{j=1}^{p}|\gamma_j|. \tag{33}$$

The 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

$$\frac{1}{n}X_j'(Y - X\tilde{\beta}) = \lambda\,\mathrm{sgn}(\tilde{\beta}_j), \qquad \text{if } \tilde{\beta}_j \neq 0, \tag{34}$$

$$\left|\frac{1}{n}X_j'(Y - X\tilde{\beta})\right| \leqslant \lambda, \qquad \text{if } \tilde{\beta}_j = 0. \tag{35}$$

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

$$\frac{1}{2}\|Y - X_{\mathcal{I}}\gamma\|^2 + n\lambda\sum_j |\gamma_j|, \tag{36}$$

where $\lambda = 1/\ln(n)$. It is easy to see that $\tilde{\beta}_j \rightarrow_{\mathrm{p}} \beta_j$ for any $j \in \mathcal{I}$ and therefore, with probability tending to 1, $\tilde{\beta}_j \neq 0$ for any $j \in \mathcal{I}$. Let $\tilde{\beta}_{\mathcal{I}^c} = \mathbf{0}$. It now suffices to show that, with probability tending to 1, such a $\tilde{\beta}$ is also on the solution path of expression (33). Note that, from expression (36),

$$\frac{1}{n}X_{\mathcal{I}}'(Y - X_{\mathcal{I}}\tilde{\beta}_{\mathcal{I}}) = \lambda\,\mathrm{sgn}(\tilde{\beta}_{\mathcal{I}}). \tag{37}$$

However, because $X'X/n = \mathrm{cov}(X) + O_p(1/\sqrt{n})$ and $\hat{\beta}^{\mathrm{LS}} = \beta + O_p(1/\sqrt{n})$,

$$\begin{aligned}\frac{1}{n}X_{\mathcal{I}}'(Y - X_{\mathcal{I}}\tilde{\beta}_{\mathcal{I}}) &= \frac{1}{n}X_{\mathcal{I}}'X_{\mathcal{I}}(\hat{\beta}_{\mathcal{I}}^{\mathrm{LS}} - \tilde{\beta}_{\mathcal{I}}) + \frac{1}{n}X_{\mathcal{I}}'X_{\mathcal{I}^c}\hat{\beta}_{\mathcal{I}^c}^{\mathrm{LS}} \\ &= \mathrm{cov}(X_{\mathcal{I}})(\beta_{\mathcal{I}} - \tilde{\beta}_{\mathcal{I}}) + O_p(n^{-1/2}).\end{aligned} \tag{38}$$

Combining equations (37) and (38),

$$\tilde{\beta}_{\mathcal{I}} = \beta_{\mathcal{I}} - \lambda\,\mathrm{cov}(X_{\mathcal{I}})^{-1}\,\mathrm{sgn}(\tilde{\beta}_{\mathcal{I}}) + O_p(n^{-1/2}). \tag{39}$$

Therefore,

$$\begin{aligned}\frac{1}{n}X_{\mathcal{I}^c}'(Y - X_{\mathcal{I}}\tilde{\beta}_{\mathcal{I}}) &= \frac{1}{n}X_{\mathcal{I}^c}'X_{\mathcal{I}}(\beta_{\mathcal{I}}^{\mathrm{LS}} - \tilde{\beta}_{\mathcal{I}}) + \frac{1}{n}X_{\mathcal{I}^c}'X_{\mathcal{I}^c}\beta_{\mathcal{I}^c}^{\mathrm{LS}} \\ &= \mathrm{cov}(X_{\mathcal{I}^c}, X_{\mathcal{I}})(\beta_{\mathcal{I}} - \tilde{\beta}_{\mathcal{I}}) + O_p(n^{-1/2}) \\ &= \lambda\,\mathrm{cov}(X_{\mathcal{I}^c}, X_{\mathcal{I}})\,\mathrm{cov}(X_{\mathcal{I}})^{-1}\,\mathrm{sgn}(\tilde{\beta}_{\mathcal{I}}) + O_p(n^{-1/2}).\end{aligned} \tag{40}$$

From equation (40), for any positive constant ε and $\forall j \notin \mathcal{I}$, then with probability tending to 1

$$\left|\frac{1}{n}X_j'(Y - X_{\mathcal{I}}\tilde{\beta}_{\mathcal{I}})\right| \leqslant c\lambda + \varepsilon \tag{41}$$

where $c < 1$ is the quantity on the left-hand side of equation (7). By choosing $\varepsilon < (1 - c)\lambda$ in inequality (41), together with equation (37), we have, with probability tending to 1, that $\tilde{\beta}$ satisfies conditions (34) and (35). Hence it is on the lasso solution path.

The necessity of condition (5) follows immediately from theorem 2 of Zhao and Yu (2006) and is therefore omitted here.

A.4. Proof of theorem 3

The proof of theorem 3 is obvious from the fact that

$$\frac{1}{n}X_j'Y \rightarrow \mathrm{cov}(X_j, X_{\mathcal{I}})\beta.$$

A.5. Proof of theorem 4

The proof of theorem 4 proceeds in the same fashion as that of theorem 2. The Karush–Kuhn–Tucker theorem suggests that a necessary and sufficient condition for any p-dimensional vector $\tilde{\beta}$ to be on the elastic net solution path is

$$\frac{1}{n}X_j'(Y - X\tilde{\beta}) - \tau\tilde{\beta}_j = \lambda\,\mathrm{sgn}(\tilde{\beta}_j), \qquad \text{if } \tilde{\beta}_j \neq 0, \tag{42}$$

$$\left|\frac{1}{n}X'_j(Y-X\tilde{\beta})\right| \leqslant \lambda, \qquad \text{if } \tilde{\beta}_j=0. \tag{43}$$

We first show that inequality (10) is a sufficient condition for the elastic net to be path consistent. For this, define $\tilde{\beta}_{\mathcal{I}^c}=\mathbf{0}$ and $\tilde{\beta}_{\mathcal{I}}$ as the minimizer of

$$\frac{1}{2}\|Y-X_{\mathcal{I}}\gamma\|^2+n\lambda\sum_j|\gamma_j|+n\tau\sum_j\gamma_j^2, \tag{44}$$

where $\lambda,\tau\rightarrow 0$ satisfy $n^{1/2}\lambda\rightarrow\infty$ and $c_1=\tau$ and $c_2=\lambda$ are such that

$$\max_{j\notin\mathcal{I}}\left(\lim_n\left[\operatorname{cov}(X_j,X_{\mathcal{I}})\{\operatorname{cov}(X_{\mathcal{I}})+\tau I\}^{-1}\left(s_{\mathcal{I}}+\frac{\tau}{\lambda}\beta_{\mathcal{I}}\right)\right]\right)<1. \tag{45}$$

It is not difficult to check that

$$\tilde{\beta}_{\mathcal{I}}=\beta_{\mathcal{I}}-\operatorname{cov}(X_{\mathcal{I}^c},X_{\mathcal{I}})\left\{\operatorname{cov}(X_{\mathcal{I}})+\tau I\right\}^{-1}\{\lambda\operatorname{sgn}(\tilde{\beta}_{\mathcal{I}})+\tau\beta_{\mathcal{I}}\}+O_p(n^{-1/2}). \tag{46}$$

It now suffices to show that $\tilde{\beta}$ satisfies condition (43). Similarly to equation (40), we have

$$\frac{1}{n}X'_{\mathcal{I}^c}(Y-X_{\mathcal{I}}\tilde{\beta}_{\mathcal{I}})=\lambda\operatorname{cov}(X'_{\mathcal{I}^c})\left\{\operatorname{cov}(X_{\mathcal{I}})+\tau I\right\}^{-1}\{\operatorname{sgn}(\tilde{\beta}_{\mathcal{I}})+\beta_{\mathcal{I}}\}+O_p(n^{-1/2}), \tag{47}$$

which is smaller than λ with probability tending to 1.

Next we show that the elastic net is not path consistent if condition (10) is violated. Without loss of generality, assume that $\beta_1=0$ and

$$\liminf_{c_1,c_2\rightarrow 0^+}\left[\operatorname{cov}(X_1,X_{\mathcal{I}})\left\{\operatorname{cov}(X_{\mathcal{I}})+c_1 I\right\}^{-1}\left(s_{\mathcal{I}}+\frac{c_1}{c_2}\beta_{\mathcal{I}}\right)\right]\geqslant 1. \tag{48}$$

Assume the contrary, that, with probability tending to 1, we can find a desirable estimate on the elastic net solution path. Denote $\tilde{\beta}$ a desirable estimate that satisfies conditions (42) and (43). Then, with probability tending to 1, $\operatorname{sgn}(\tilde{\beta}_j)=\operatorname{sgn}(\beta_j)$ for any $j\in\mathcal{I}$. From condition (42),

$$\tilde{\beta}_{\mathcal{I}}=\beta_{\mathcal{I}}-\operatorname{cov}(X_{\mathcal{I}^c},X_{\mathcal{I}})\left\{\operatorname{cov}(X_{\mathcal{I}})+\tau I\right\}^{-1}\{\lambda\operatorname{sgn}(\tilde{\beta}_{\mathcal{I}})+\tau\beta_{\mathcal{I}}\}+O_p(n^{-1/2}). \tag{49}$$

Together with the fact that $\tilde{\beta}_{\mathcal{I}^c}=\mathbf{0}$, we have, with probability tending to 1,

$$\frac{1}{n}X'_1(Y-X_{\mathcal{I}}\tilde{\beta}_{\mathcal{I}})=\lambda\operatorname{cov}(X_1,X_{\mathcal{I}})\left\{\operatorname{cov}(X_{\mathcal{I}})+\tau I\right\}^{-1}\left\{\operatorname{sgn}(\tilde{\beta}_{\mathcal{I}})+\frac{\tau}{\lambda}\beta_{\mathcal{I}}\right\}+\xi \tag{50}$$

where $P(\xi>0)$ is bounded below by a positive constant. This implies that, with a non-vanishing probability, $\tilde{\beta}$ cannot satisfy condition (43), which contradicts the construction of $\tilde{\beta}$.

A.6. Proof of theorem 5

The Karush–Kuhn–Tucker theorem suggests that a necessary and sufficient condition for a point d to be on the solution path of model (2) is that there is a $\lambda\geqslant 0$ such that, for any $j=1,\ldots,p$,

$$\{-Z'_j(Y-Zd)+\lambda\}d_j=0, \tag{51}$$

$$-Z'_j(Y-Zd)+\lambda\geqslant 0, \tag{52}$$

$$d_j\geqslant 0. \tag{53}$$

In what follows we show that conditions (51)–(53) are satisfied by any point on the solution path constructed by the algorithm, and any solution to conditions (51)–(53) for certain $\lambda\geqslant 0$ is also on the solution path constructed.

We verify conditions (51)–(53) for the solution path by induction. Obviously, they are satisfied by $d^{[0]}$. Now suppose that they hold for any point before $d^{[k]}$. 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:

(a) a variable is added to the active set, $j^*\notin\mathcal{C}_k$;
(b) a variable is deleted from the active set, $j^*\in\mathcal{C}_k$;
(c) $\alpha=1$.

It is easy to see that conditions (51)–(53) 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

$$\gamma_{\mathcal{C}_k}=(Z'_{\mathcal{C}_k}Z_{\mathcal{C}_k})^{-1}Z'_{\mathcal{C}_k}r^{[k]}. \tag{54}$$

It is not difficult to show that conditions (51) and (52) are true for d^{α}. It now suffices to check condition (53). 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 $j\in\mathcal{C}_{k-1}$,

$$Z'_jr^{[k-1]}>Z'_1r^{[k-1]}, \tag{55}$$

$$Z'_jr^{[k]}=Z'_1r^{[k]}. \tag{56}$$

Therefore,

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

Because there is a positive constant b such that

$$r^{[k-1]}-r^{[k]}=bZ_{\mathcal{C}_{k-1}}(Z'_{\mathcal{C}_{k-1}}Z_{\mathcal{C}_{k-1}})^{-1}Z'_{\mathcal{C}_{k-1}}r^{[k-1]},$$

we conclude that

$$Z'_1Z_{\mathcal{C}_{k-1}}(Z'_{\mathcal{C}_{k-1}}Z_{\mathcal{C}_{k-1}})^{-1}Z'_{\mathcal{C}_{k-1}}r^{[k-1]}<Z'_jZ_{\mathcal{C}_{k-1}}(Z'_{\mathcal{C}_{k-1}}Z_{\mathcal{C}_{k-1}})^{-1}Z'_{\mathcal{C}_{k-1}}r^{[k-1]}.$$

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

$$Z'_1Z_{\mathcal{C}_{k-1}}(Z'_{\mathcal{C}_{k-1}}Z_{\mathcal{C}_{k-1}})^{-1}s_{\mathcal{C}_{k-1}}<1. \tag{57}$$

Together with equation (54),

$$\gamma_1=\frac{\{1-Z'_1Z_{\mathcal{C}_{k-1}}(Z'_{\mathcal{C}_{k-1}}Z_{\mathcal{C}_{k-1}})^{-1}s_{\mathcal{C}_{k-1}}\}Z'_jr^{[k]}}{Z'_1\{I_n-Z_{\mathcal{C}_{k-1}}(Z'_{\mathcal{C}_{k-1}}Z_{\mathcal{C}_{k-1}})^{-1}Z'_{\mathcal{C}_{k-1}}\}Z_1}>0. \tag{58}$$

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 equation (54). It is readily possible to show that conditions (51) and (53) are true with $\lambda=Z'_j(Y-Zd^{\alpha})$ where j is arbitrarily chosen from $\mathcal{C}_k$. It suffices to verify condition (52). By the construction of the solution path, it suffices to show that condition (52) 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

$$\tilde{\gamma}_{\mathcal{C}_{k-1}}=(Z'_{\mathcal{C}_{k-1}}Z_{\mathcal{C}_{k-1}})^{-1}Z'_{\mathcal{C}_{k-1}}r^{[k-1]}. \tag{59}$$

Clearly, $\tilde{\gamma}_1<0$. Similarly to equation (58), we have

$$\tilde{\gamma}_1=\frac{\{1-Z'_1Z_{\mathcal{C}_k}(Z'_{\mathcal{C}_k}Z_{\mathcal{C}_k})^{-1}s_{\mathcal{C}_k}\}Z'_jr^{[k]}}{Z_1\{I_n-Z_{\mathcal{C}_k}(Z'_{\mathcal{C}_k}Z_{\mathcal{C}_k})^{-1}Z'_{\mathcal{C}_k}\}Z_1}, \tag{60}$$

where j is arbitrarily chosen from $\mathcal{C}_k$. Therefore,

$$Z'_1Z_{\mathcal{C}_k}(Z'_{\mathcal{C}_k}Z_{\mathcal{C}_k})^{-1}s_{\mathcal{C}_k}=(p_j/Z'_jr^{[k]})Z'_1Z\gamma<1.$$

In other words, $Z'_1Z\gamma<Z'_jr^{[k]}=Z'_jZ\gamma$. Since $Z'_1r^{[k]}=Z'_jr^{[k]}$, we conclude that $Z'_1(Y-Zd^{\alpha})<Z'_j(Y-Zd^{\alpha})=\lambda$.

Next, we need to show that, for any $\lambda\geqslant0$, the solution to conditions (51)–(53) is on the solution path. By the continuity of the solution path and the uniqueness of the solution to equation (2), it is evident that, for any $\lambda\in[0,\max_j(Z'_jY)]$, the solution to conditions (51)–(53) is on the path. The proof is now completed by the fact that, for any $\lambda>\max(Z'_jY)$, the solution to conditions (51)–(53) is $\mathbf{0}$, which is also on the solution path.

References

Breiman, L. (1995) Better subset regression using the nonnegative garrote. *Technometrics*, **37**, 373–384.

Breiman, L. (1996) Heuristics of instability and stabilization in model selection. *Ann. Statist.*, **24**, 2350–2383.

Chen, S. S., Donoho, D. L. and Saunders, M. A. (1998) Atomic decomposition by basis pursuit. *SIAM J. Scient. Comput.*, **20**, 33–61.

Efron, B., Johnstone, I., Hastie, T. and Tibshirani, R. (2004) Least angle regression. *Ann. Statist.*, **32**, 407–499.

Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Statist. Ass.*, **96**, 1348–1360.

Foster, D. P. and George, E. I. (1994) The risk inflation criterion for multiple regression. *Ann. Statist.*, **22**, 1947–1975.

Friedman, J., Hastie, T., Rosset, S., Tibshirani, R. and Zhu, J. (2004) Discussion of boosting papers. *Ann. Statist.*, **32**, 102–107.

George, E. I. and Foster, D. P. (2000) Calibration and empirical Bayes variable selection. *Biometrika*, **87**, 731–747.

George, E. I. and McCulloch, R. E. (1993) Variable selection via Gibbs sampling. *J. Am. Statist. Ass.*, **88**, 881–889.

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

Osborne, M., Presnell, B. and Turlach, B. (2000) On the LASSO and its dual. *J. Computnl Graph. Statist.*, **9**, 319–337.

Shen, X. and Ye, J. (2002) Adaptive model selection. *J. Am. Statist. Ass.*, **97**, 210–221.

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.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B, **58**, 267–288.

Yuan, M. and Lin, Y. (2005) Efficient empirical Bayes variable selection and estimation in linear models. *J. Am. Statist. Ass.*, **100**, 1215–1225.

Yuan, M. and Lin, Y. (2006) Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc.* B, **68**, 49–67.

Zhao, P. and Yu, B. (2006) On model selection consistency of Lasso. *J. Mach. Learn. Res.*, **7**, 2481–2514.

Zou, H. (2006) The adaptive Lasso and its oracle properties. *J. Am. Statist. Ass.*, **101**, 1418–1429.

Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. *J. R. Statist. Soc.* B, **67**, 301–320.