Nonnegative Garrote Component Selection in Functional ANOVA Models

Ming Yuan
School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332-0205
(Email: myuan@isye.gatech.edu)

Abstract

We consider the problem of component selection in a functional ANOVA model. A nonparametric extension of the nonnegative garrote (Breiman, 1996) is proposed. We show that the whole solution path of the proposed method can be efficiently computed, which, in turn , facilitates the selection of the tuning parameter. We also show that the final estimate enjoys nice theoretical properties given that the tuning parameter is appropriately chosen. Simulation and a real data example demonstrate promising performance of the new approach.

1 Introduction

Consider a multivariate nonparametric regression problem where we have n observations $(y_1, \mathbf{x}_1), \dots, (y_n, \mathbf{x}_n)$ on a dependent variable Y and p predictors $X = (X_1, X_2, ..., X_p)$, and we want to estimate the conditional mean function $f(X) = E(Y|X)$. In the functional ANOVA framework (Wahba, 1990), we write the regression function $f(X)$ as

$$f(X) = \mu + \sum_{j=1}^{p} f_j(X_j) + \sum_{1 \le j_1 < j_2 \le p} f_{j_1 j_2}(X_{j_1}, X_{j_2}) + \dots + f_{1\dots p}(X_1, \dots, X_p), \quad (1)$$

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. The series on the right hand side of (1) is usually truncated somewhere to enhance interpretability. The identifiability of the terms in (1) is assured by side conditions through averaging operators. The most popular example of functional ANOVA is the additive model proposed by Hastie and Tibshirani (1990) where only main effects are retained in (1).

Similar to variable selection in the multiple linear regression where

$$f(X) = \beta_0 + X_1\beta_1 + \dots + X_p\beta_p, \quad (2)$$

some of the components on the right hand side of (1) may not be significant and therefore an unknown but usually small subset of the the components are adequate to describe $f(X)$. By effectively identifying the subset of important components, we can improve estimation accuracy and enhance model interpretability.

Component selection for (1) is most commonly studied in the special case of the multiple linear regression where $f_i(X_i) = X_i\beta_i$ for $i = 1, \dots p$ and other components are zero. A number of variable selection methods have been introduced for this problem 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, Johnston, Hastie and Tibshirani, 2004; Yuan and Lin, 2005a; Zou and Hastie, 2005; Yuan and Lin, 2006). In particular, Breiman (1995) showed that the traditional subset selection methods are not satisfactory in terms of prediction accuracy and stability, and proposed the nonnegative garrote which is shown to be more accurate and stable.

In recent years, selection of important components in a nonparametric setting such as (1) has also attracted much attention. In a series of papers, Kohn and coauthors (Shively, Kohn and Wood, 1999; Wood, Kohn, Shively and Jiang, 2002; Yau, Kohn and Wood, 2003; Yau and Kohn, 2003) introduced a MCMC based Bayesian variable selection paradigm for additive models. Despite the elegance of their approach and promising performance reported, the implementation requires a great deal of expertise and it could be very computationally demanding for relatively large scale problems. Zhang, Wahba, Lin, Voelker, Ferris, Klein and Klein (2004) suggested a likelihood basis pursuit approach to model selection and estimation in the functional ANOVA for exponential families where after model fitting, a sequential Monte Carlo bootstrap test algorithm is applied for the purpose of model selection. Their approach potentially can also be expensive to compute and no theoretical properties are known about the resulting estimator so far. More recently, Lin and Zhang (2006) proposed the so-called COSSO estimator where model fitting and selection can be done simultaneously.

In this paper, we propose an alternative method for component selection in functional ANOVA models. The estimator can be seen as a generalization of a very successful estimator for the multiple linear regression, the nonnegative garrote estimator introduced by Breiman (1995). We show that the new estimator can be computed very efficiently and enjoys very nice theoretical properties.

The rest of the paper is organized as follows. We introduce the new method in the next section. Section 3 gives a fast algorithm for constructing the whole solution path for our estimate. The asymptotic properties of the proposed method are studied in Section 4. Section 5 presents some numerical examples. We conclude with some discussions in Section 6.

2 Method

Like other nonparametric settings, instead of assuming a parametric form of f, we allow it to reside in some Hilbert space

$$\mathcal{H} = \bigotimes_{j=1}^{p} \mathcal{H}_j \tag{3}$$

of smooth functions. Here $\mathcal{H}_j$ is a function space of univariate functions over $[0,1]$. A particular choice of $\mathcal{H}_j$ that is commonly used in practice is the Soblev-Hilbert space. For a nonnegative integer m, the Soblev-Hilbert space of order m is given as:

$$\{f : f^{(m)} \in \mathcal{L}_2[0,1], f, f^{(v)} (v = 1, \dots, m-1) \text{ are absolutely continuous}\} \tag{4}$$

Write

$$\mathcal{H}_j = \{1\} \bigoplus \bar{\mathcal{H}}_j \tag{5}$$

Similar to (1), the reproducing kernel Hilbert space $\mathcal{H}$ of functions on $[0,1]^d$ admits the following tensor-product decomposition

$$\mathcal{H} = \{1\} \bigoplus \bar{\mathcal{H}}_1 \bigoplus \dots \bigoplus \bar{\mathcal{H}}_p \bigoplus \left(\bar{\mathcal{H}}_1 \bigotimes \bar{\mathcal{H}}_2\right) \bigoplus \dots \bigoplus \left(\bigotimes_{j=1}^{p} \bar{\mathcal{H}}_j\right) \tag{6}$$

Therefore the ANOVA decomposition (1) can be uniquely determined with $f_j \in \bar{\mathcal{H}}_j$, $f_{jk} \in \bar{\mathcal{H}}_j \otimes \bar{\mathcal{H}}_k$ and so on.

The original nonnegative garrote estimator is developed for the multiple linear regression where $\bar{\mathcal{H}}_j = \{x_j\}$, and defined as a scaled version of the ordinary least square estimate. The shrinking factor $d(\lambda) = (d_1(\lambda), \dots, d_p(\lambda))'$ is given as the minimizer to

$$\frac{1}{2}\sum_{i=1}^{n} (y_i - \mathbf{z}_i d)^2 + n\lambda \sum_{j=1}^{p} d_j, \tag{7}$$

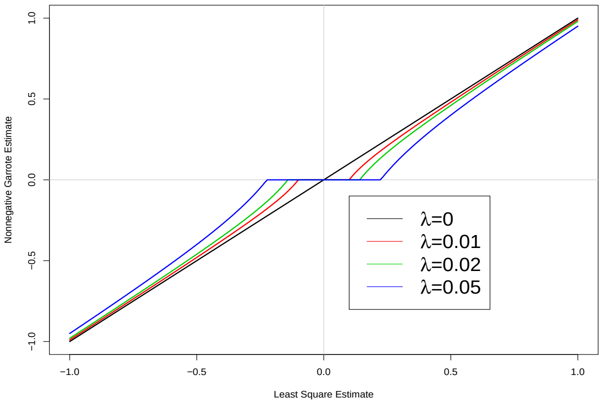
subject to $d_j \geq 0$ for all j,where $\mathbf{z}_i = (z_{i1}, \dots, z_{ip})$, $z_{ij} = x_{ij}\widehat{\beta}_j^{\text{LS}}$ and $\widehat{\beta}_j^{\text{LS}}$ is the ordinary least square estimate. Here $\lambda > 0$ is a tuning parameter. The nonnegative garrote estimate of the regression coefficient is subsequently defined as $\widehat{\beta}_j^{\text{NG}}(\lambda) = d_j(\lambda)\widehat{\beta}_j^{\text{LS}}$, $j = 1, ..., p$. 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 $\sum_i \mathbf{x}_i \mathbf{x}_i' = I_n$. In this case, the minimizer of (7) has an explicit form:

$$d_j(\lambda) = \left(1 - \frac{\lambda}{\left(\widehat{\beta}_j^{\text{LS}}\right)^2}\right)_+, \qquad j = 1, \dots, p. \tag{8}$$

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. This effect is illustrated in Figure 1.

Figure 1: Shrinkage Effect



The idea of the nonnegative garrote can be naturally generalized to the functional ANOVA models.

We begin with an initial estimate of the components $\tilde{f}_1^{\text{init}}, \ldots, \tilde{f}_{1\ldots p}^{\text{init}}$. Our generalized nonnegative garrote estimate is then given as $d_1(\lambda)\tilde{f}_1^{\text{init}}, \ldots, d_{1\ldots p}(\lambda)\tilde{f}_{1\ldots p}^{\text{init}}$ where d's are the minimizer of (7) with $\mathbf{z}_i = (\tilde{f}_1^{\text{init}}(x_{i1}), \ldots, \tilde{f}_{1\ldots p}^{\text{init}}(\mathbf{x}_i))'$.

One good choice of the initial estimate is the smoothing spline estimate which is given as the minimizer of

$$\sum_{i=1}^{n} [y_i - f(\mathbf{x}_i)]^2 + \tau_1 J_1(f_1) + \ldots + \tau_p J_p(f_p) + \ldots + \tau_{1\ldots p} J_{1\ldots p}(f_{1\ldots p}). \qquad (9)$$

where $\tau' s$ are tuning parameters and J's are squared norms defined over the subspace where the corresponding function comes from. The tuning parameters τ's are commonly selected by minimizing the GCV score. For more detailed discussions, the readers are referred to Wahba (1990).

To fix ideas, we shall focus on the additive model in the rest of paper. But the argument should easily be extended to the more general functional ANOVA model as well.

3 Computing the Solution Path

Similar to other methods of regularization, the nonnegative garrote estimation procedure proceeds in two steps once the initial estimate is chosen. First the solution path $d(\lambda)$ 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 (7) for a series of λ's, as done in Breiman (1995).

It can be shown that the solution path of (7) 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 (Osborne et al., 2000; Efron et al., 2004) for the LASSO (Tibshirani, 1996) in the multiple linear regression, with a complicating factor being the nonnegative constraints in (7).

Algorithm – Nonnegative Garrote

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

(2) Compute the current active set

$$\mathcal{C}_k = \arg\max_j \left(Z_j' r^{[k-1]} \right) \qquad (10)$$

(3) Compute the current direction γ, which is a p dimensional vector defined by $\gamma_{\mathcal{C}_k^c} = 0$ and

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

(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

$$Z_j' \left(r^{[k-1]} - \alpha_j Z\gamma \right) = Z_{j'}' \left(r^{[k-1]} - \alpha_j Z\gamma \right) \qquad (11)$$

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

The following theorem justifies the algorithm, and the sketch of its proof is given in the appendix. More detailed proof can be found in Yuan and Lin (2005b).

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).

4 Asymptotic Properties

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 $\widehat{f}$ is considered "desirable" if it is consistent in terms of both coefficient estimate and component 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. A sketch of the proof is relegated to the appendix. A more detailed proof can be found in Yuan and Lin (2005b).

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

$$E\left(f_j(X_j) - \tilde{f}_j^{\text{init}}(X_j)\right)^2 = O_p(\delta_n^2) \qquad (12)$$

for some $\delta_n \to 0$, 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 $P(\widehat{f}_j^{\text{NG}} = 0) \to 1$ for any j such that $f_j = 0$, and

$$\sup_j E\left(f_j(X_j) - \widehat{f}_j^{\text{NG}}(X_j)\right)^2 = O_p(\lambda^2). \qquad (13)$$

Theorem 2 shows that as long as the initial estimates $\tilde{f}_1^{\text{init}}, \ldots, \tilde{f}_p^{\text{init}}$ are consistent in terms of estimation, the nonnegative garrote estimate $\widehat{f}_1^{\text{NG}}, \ldots, \widehat{f}_p^{\text{NG}}$ are consistent in terms of both estimation and model selection given that the tuning parameter λ is appropriately chosen. In other words, 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.

In achieving the consistency in variable selection, we show in Theorem 2 that the nonnegative garrote estimate of a nonzero coefficient converges at a slower rate than its initial estimate. It is not clear to us whether this is the unavoidable price one has to pay for the purpose of variable selection in general. A simple remedy is to add another layer to the estimating procedure. After running our nonnegative garrote, we can run a nonparametric regression such as (9) only on those selected components to obtain a final estimate. We shall adopt this strategy in the numerical examples presented in the next section.

5 Numerical Examples

5.1 Simulations

The example setup is the same as Example 1 from Lin and Zhang (2006). Ten covariates were simulated in the following fashion. First $W_1, \ldots, W_{10}$ 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

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

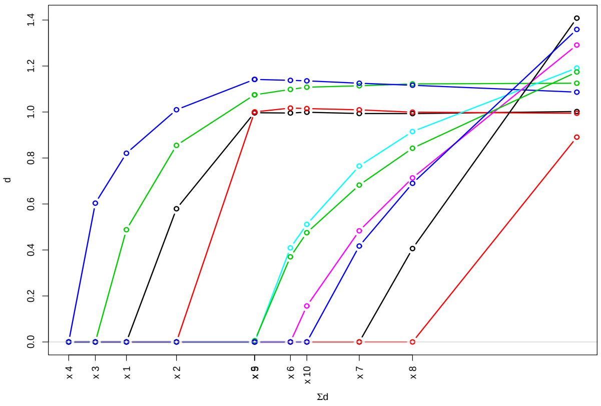
where

$$\begin{aligned} g_1(t) = t; \quad g_2(t) &= (2t-1)^2; \\ g_3(t) &= \frac{\sin(2\pi t)}{2 - \sin(2\pi t)}; \\ g_4(t) = 0.1\sin(2\pi t) + 0.2\cos(2\pi t) &+ 0.3\sin^2(2\pi t) \\ +0.4\cos^3(2\pi t) &+ 0.5\sin^3(2\pi t). \qquad (15) \end{aligned}$$

We chose the noise variance $\sigma^2 = 1.74$ to give a signal to noise ratio 3:1. We used smoothing spline estimate (9) as the initial estimate.

The solution path of the shrinkage factor $d(\lambda)$ is given in Figure 2. The true components $f_1, \ldots, f_4$ enter the model first, followed by the redundant components. Five fold cross-validation chooses a model with only the true component.

Figure 2: Solution Path of the Shrinkage Factor



The final estimate of the four components are also given in Figure 3.

To further investigate the performance of the proposed method, we estimate the MSE using a test set of size 10000. The sizes of the training and validation sets are 100. Table 1 documents the median MSE averaged over 200 runs. We also report its standard deviation estimated using 500 bootstrap samples (numbers in parentheses).

We also recorded the frequency of different model sizes in Table 2.

Figure 3: Estimated Nonzero Components

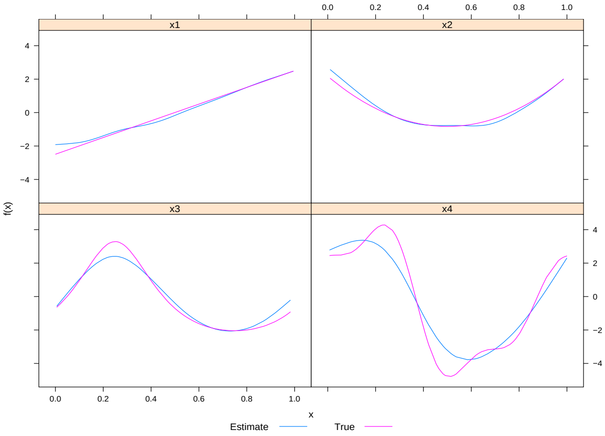


Table 1: Median MSE for the Additive Model Example

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

5.2 Real Data

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. This task has been carried for the multiple linear model previously (Tibshirani, 1996). We apply the nonparametric method on the data. As discussed in Hastie, Tibshirani and Friedman (2003), the dataset was split into training set with 67 observations and test set with 30 observations. We compute the estimate using the training data and the solution path of the shrinking factor is presented in Figure 4.

Five fold cross validation suggests a model with four components corresponding to *lcavol, lweight, lbph* and *svi*. The fitted components are given in Figure 5. We see clear nonlinearity for the effect of *lbph*, which suggests that our nonparametric approach might be more appropriate than a multiple linear model. To further demonstrate the predictive performance of the estimate, we compute the prediction error on the test set, which is 0.55. In contrast, if a multiple linear regression is applied, one of the best variable selection and estimation method, the LASSO, selects seven effects with only *gleason* being excluded. The corresponding prediction error is 0.57.

Table 2: Frequency of Model Sizes for the Additive Model Example

	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

Figure 4: Solution Path of the Shrinkage Factor

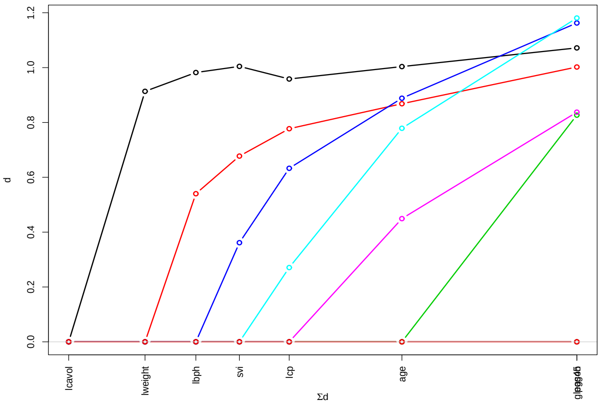
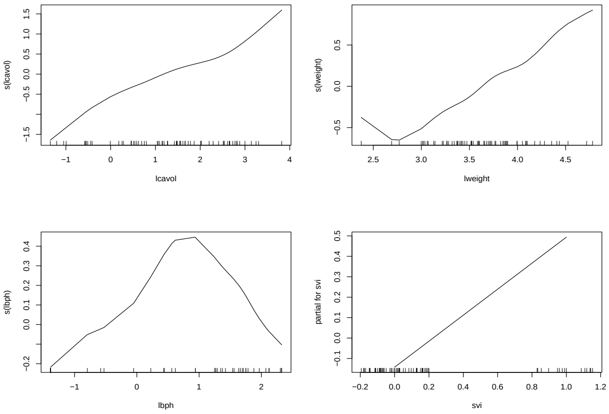


Figure 5: Effects of the Selected Variables



6 Conclusion

Variable selection in an additive model, or more generally component selection in a functional ANOVA model is an important problem in practice. Extending a successful variable selection and estimation in multiple linear regression models, we introduced an efficient method for this purpose. Although we have focused on mean regression problem, the proposed approach can naturally be extended to more general likelihood based learning problems. Given an initial estimate, we scale each component by a factor that minimizes a penalized likelihood. We leave such extension for future studies.

Acknowledgements

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

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. Scientific Computing*, **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. Amer. Statist. Assoc.*, **96** 1348-1360.

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

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. Amer. Statist. Assoc.*, **88**, 881-889.

Hastie, T. and Tibshirani, R. (1990), *Generalized Additive Models*, Chapman & Hall/CRC, London.

Hastie, T., Tibshirani, R. and Friedman, J. (2003), *The Elements of Statistical Learning*, Springer, New York

Lin, Y. and Zhang, H. H. (2006), Component selection and smoothing in smoothing spline analysis of variance models, *Ann. Statist.*, to appear.

Osborne, M.R., Presnell, B. and Turlach, B.A. (2000), On the LASSO and its dual, *J. Comput. Graph. Statist.*, **9**, 319-337.

Shen, X. and Ye, J. (2002), Adaptive model selection, *J. Amer. Statist. Assoc.*, **97**, 210-221.

Shively, T., Kohn, R. and Wood, S. (1999), Variable selection and function estimation in additive nonparametric regression using a data-based prior (with discussion), *J. Amer. Statist. Assoc.*, **94**, 777-806.

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. Royal. Statist. Soc. B.*, **58**, 267-288.

Wahba, G. (1990), *Spline Models for Observational Data*, SIAM, Philadelphia.

Wood, S., Kohn, R., Shively, T. and Jiang, W. (2002), Model selection in spline nonparametric regression, *J. R. Stat. Soc. Ser. B*, **64**, 119-139.

Yau, P. and Kohn, R. (2003), Estimation and variable selection in nonparametric heteroscedastic regression, *Stat. Comput.* **13**, 191-208.

Yau, P., Kohn, R. and Wood, S. (2003), Bayesian variable selection and model averaging in high-dimensional multinomial nonparametric regression, *J. Comput. Graph. Statist.*, **12**, 23-54.

Yuan, M. and Lin, Y. (2005a), Efficient empirical Bayes variable selection and estimation in linear models, *J. Amer. Statist. Assoc.*, **100**, 1215-1225.

Yuan, M. and Lin, Y. (2005b), On the nonnegative garrote estimator.

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

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

Appendix

Sketch of Proof for Theorem 1. Karush-Kuhn-Tucker Theorem suggests that a necessary and sufficient condition for a point d to be on the solution path of (7) is that there exists a $\lambda \geq 0$ such that for any $j = 1, \ldots, p$,

$$\begin{aligned}
\{-Z_j'(Y - Zd) + \lambda\} d_j &= 0 && (16)\\
-Z_j'(Y - Zd) + \lambda &\geq 0 && (17)\\
d_j &\geq 0 && (18)
\end{aligned}$$

We first show by induction that (16)-(18) are satisfied by any point on the solution path constructed by the algorithm. 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 (16)-(18) will continue to hold for any point between $d^{[k]}$ and $d^{[k+1]}$ if $\alpha = 1$. Now we consider the other two possibilities. When addition occurs, it can be shown that $d_{j^*}^{[k+1]} > 0$ and therefore the addition will preserve (16)-(18) until at least $d^{[k+1]}$. When deletion occurs, it can be shown that $Z_{j^*}' r^{[k+1]} < \lambda$ and therefore (16)-(18) are still satisfied. In summary, in all three situations, it can be shown that (16)-(18) are satisfied by any point until $d^{[k+1]}$. The readers are referred to Yuan and Lin (2005b) for a detailed proof.

Next, we need to show that for any $\lambda \geq 0$, the solution to (16)-(18) is on the solution path. By the continuity of the solution path and the uniqueness of the solution to (7), it is evident that for any $\lambda \in [0, \max_j Z_j'Y]$, the solution to (16)-(18) is on the path. The proof is now completed by the fact that for any $\lambda > \max Z_j'Y$, the solution to (16)-(18) is $\mathbf{0}$ which is also on the solution path. ▮

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

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

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 other vectors and matrices be defined in the same fashion unless otherwise indicated. Note that $d_{1\cdot}$ is also the unconstrained minimizer of

$$\frac{1}{2}||Y - Z_{1\cdot}\gamma||^2 + n\lambda \sum_j \gamma_j, \qquad (19)$$

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

$$\begin{pmatrix} d_{11} \\ d_{10} \end{pmatrix} = \begin{pmatrix} Z_{11}'Z_{11}/n & Z_{11}'Z_{10}/n \\ Z_{10}'Z_{11}/n & Z_{10}'Z_{10}/n \end{pmatrix}^- \times$$

$$\begin{pmatrix} Z_{11}'Y/n - \lambda \mathbf{1}_{p_{11}} \\ Z_{10}'Y/n - \lambda \mathbf{1}_{p_{10}} \end{pmatrix} \qquad (20)$$

Denote

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

Then

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

This implies that

$$\begin{aligned}
d_{10} = -A_{00.1}^- A_{01} A_{11}^- &\left(Z_{11}'Y/n - \lambda \mathbf{1}_{p_{11}}\right)\\
+A_{00.1}^- &\left(Z_{10}'Y/n - \lambda \mathbf{1}_{p_{10}}\right) \equiv A_{00.1}^- w \qquad (21)
\end{aligned}$$

Rewrite w as

$$\begin{aligned}
w = Z_{10}' \left[I_{p_{11}} - Z_{11}\left(Z_{11}'Z_{11}\right)^- Z_{11}'\right] Y/n - \lambda \mathbf{1}_{p_{10}}&\\
+\lambda A_{01} A_{11}^- \mathbf{1}_{p_{11}}.& \qquad (22)
\end{aligned}$$

Because $\tilde{f}^{\mathrm{init}}$ is δ_n consistent,

$$\begin{aligned}
w = Z_{10}' \left[I_{p_{11}} - Z_{11}\left(Z_{11}'Z_{11}\right)^- Z_{11}'\right] Y/n&\\
-\lambda\left(1 + O_p\left(\delta_n\right)\right) \mathbf{1}_{p_{10}}.& \qquad (23)
\end{aligned}$$

Now note that

$$\begin{aligned}
\left|\left|\left[I_{p_{11}} - Z_{11}\left(Z_{11}'Z_{11}\right)^- Z_{11}'\right] Y\right|\right|^2&\\
\leq Y'Y = O_p(n),& \qquad (24)
\end{aligned}$$

since $Z_{11}\left(Z_{11}'Z_{11}\right)^- Z_{11}'$ is a projection matrix. Thus by Cauchy-Schwartz inequality,

$$\begin{aligned}
&\left|\left|Z_{10}' \left[I_{p_{11}} - Z_{11}\left(Z_{11}'Z_{11}\right)^- Z_{11}'\right] Y\right|\right|\\
\leq\ & ||Z_{10}|| \left|\left|\left[I_{p_{11}} - Z_{11}\left(Z_{11}'Z_{11}\right)^- Z_{11}'\right] Y\right|\right|\\
=\ & O_p\left(\sqrt{n}\, ||Z_{10}||\right)\\
=\ & O_p(n\delta_n) = o_p(n\lambda).
\end{aligned}$$

This leads to $w = -\lambda(1 + o_p(1))1_{p_{10}}$. Since $d_j > 0$ for any $j \in \Lambda_{10}$, we have $w'd_{10} < 0$. This contradicts with (21) which implies that $w'd_{10} = w'A_{00.1}^- w \geq 0$. Thus, when $n \to \infty$, $P(\mathcal{A}) \to 0$.

Denote $\mathcal{B} = \{p_{01} = 0\}$. 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

$$\frac{1}{2}\, ||Y - Z_{\cdot 1}\gamma||^2 + n\lambda\gamma' \mathbf{1}_{p_{\cdot 1}}, \qquad (25)$$

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

$$d^u = \left(Z_{\cdot 1}'Z_{\cdot 1}/n\right)^- \left(Z_{\cdot 1}'Y/n - \lambda \mathbf{1}_{p_{\cdot 1}}\right). \qquad (26)$$

Similar as before, we have

$$d^u = \mathbf{1}_{p_{\cdot 1}}\left(1 + O_p(\lambda)\right). \qquad (27)$$

Thus, with probability tending to 1, $d^u \to \mathbf{1}_{p_{\cdot}}$. Now the proof is completed since $\tilde{f}_j^{\mathrm{init}}$ is δ_n consistent. ▮