Statistics and Its Interface Volume 3 (2010) 493–499

Dimension reduction and parameter estimation for additive index models*

Lingyan Ruan and Ming Yuan†

In this paper, we consider simultaneous model selection and estimation for the additive index model. The additive index model is a class of structured nonparametric models that can be expressed as additive models of a set of unknown linear transformation of the original predictor variables. We introduce a penalized least squares estimator and discuss how it can be efficiently computed in practice. Both theoretical and empirical properties of the estimate are presented to demonstrate its merits. Extensions to more general prediction framework are also discussed.

Keywords and phrases: Additive model, Index model, Model selection, Projection pursuit, Smoothing splines.

1. INTRODUCTION

In the additive index model, a response y is related to a predictor $\mathbf{x} \in \mathbf{R}^p$ through

$$f(\mathbf{x}) = h_1(\alpha_1'\mathbf{x}) + h_2(\alpha_2'\mathbf{x}) + \cdots + h_q(\alpha_q'\mathbf{x}), \tag{1}$$

with some projection indices $\alpha_1, \alpha_2, \ldots, \alpha_q$ and ridge functions $h_1, h_2, \ldots, h_q$. The goal is to infer f based on n independent copies $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ of $(\mathbf{x}, y)$ without assuming any parametric form for the ridge functions. The additive model has several popular structured nonparametric regression models, most notably additive models or single index models, as special cases. It is also known to be much more flexible than these more specialized examples. In particular, it has been shown that any square integrable function can be approximated to an arbitrary precision by a function of form (1) (Diaconis and Shahshahani, 1984).

To estimate f, the dimensionality q is often assumed to be known apriori. Various methods have been proposed to estimate f when q is given. See, e.g., Chen (1991) and Chiou and Müller (2004) among others. The assumption that q is known in advance, however, can be unrealistic. Its choice essentially amounts to a model selection or dimension reduction problem. It is clear that if q is too small, the additive index model cannot sufficiently capture the relationship between the response and the predictors. On the other hand, when q is too big, some of the components in the additive index model will be close to zero, which could cause identifiability problems (see, e.g., Yuan, 2008), and subsequently troubles in parameter estimation. The choice of q, albeit critical, is notoriously difficult in practice.

In this paper, we propose a penalized least squares method to simultaneously select the dimensionality q and estimate the regression function f for the additive index model. Our method is inspired by the COSSO recently proposed by Lin and Zhang (2006) for the purpose of variable selection in additive models. The penalty we employ encourages some of the ridge functions to be exactly zero instead of being estimated as close to zero, which avoids the potential problem of unidentifiability. We prove that the additive index model estimated using the proposed estimator achieves the same rate of convergence as the univariate nonparametric regression, which suggests that models of form (1) have the potential to mitigate the curse of dimensionality.

The rest of this paper is organized as follows. In Section 2, we introduce the proposed estimator for the additive index model and study its asymptotic properties. We show that given that all ridge functions come from a Sobolev space, the additive index model estimated in this fashion achieves the optimal convergence rate. Section 3 discusses an iterative algorithm that can be used to efficiently compute the proposed estimate in practice. Examples, both simulated and real data, are presented in Section 4 to demonstrate the merits of the proposed methodology. We conclude with some discussions in Section 5.

2. PENALIZED LEAST SQUARES ESTIMATE

To fix ideas, we shall begin by focusing on the usual mean regression model:

$$y_i = f_0(\mathbf{x}_i) + \epsilon_i, \qquad i = 1, 2, \ldots, n, \tag{2}$$

where the predictors are properly scaled so that $\mathbf{x}_i \in [0,1]^p$ and

$$f_0(\mathbf{x}) = \mu + h_{01}(\alpha_1'\mathbf{x}) + h_{02}(\alpha_2'\mathbf{x}) + \cdots + h_{0q}(\alpha_q'\mathbf{x}), \tag{3}$$

where the dimensionality q, although unknown, is assumed to be upper bounded by a known value M. Such an upper bound can be easily obtained in practice. For example, it

*This research was supported in part by NSF Grant DMS-0706724, DMS-0846234, and a grant from the Georgia Cancer Coalition.
†Corresponding author.

can be the final dimensionality obtained by the projection pursuit regression. To avoid ambiguity, we shall also assume that the projection indices α_js satisfy $\|\alpha_j\| = 1$. To model the ridge functions nonparametrically, we consider them as members of the usual Sobolev space of order m:

$$\mathcal{S}_m = \{h : h^{(s)} \text{ are absolutely continuous for } s = 0, \ldots, m-1, h^{(m)} \in \mathcal{L}_2[0,1]\}. \tag{4}$$

When equipped with a norm:

$$\|h\|^2 = \sum_{v=0}^{m-1} \left(\int h^{(v)} \right)^2 + \int_0^1 \left(h^{(m)} \right)^2, \tag{5}$$

$\mathcal{S}_m$ forms a reproducing kernel Hilbert space (see, e.g., Wahba, 1990). Let $\bar{\mathcal{S}}_m$ be the orthogonal complement of constant functions in $\mathcal{S}_m$ in that $\mathcal{S}_m = \{1\} \oplus \bar{\mathcal{S}}_m$. It then suffices to consider $h_j \in \bar{\mathcal{S}}_m$, i.e.,

$$f_0 \in \mathcal{F}_M := \{\mu + h_1(\alpha_1'\mathbf{x}) + h_2(\alpha_2'\mathbf{x}) + \cdots + h_M(\alpha_M'\mathbf{x}) : h_j(\cdot) \in \bar{\mathcal{S}}_m, \|\alpha_j\| = 1\}. \tag{6}$$

For any $f \in \mathcal{F}_M$, define

$$J(f) = \inf \{\|h_1\| + \|h_2\| + \cdots + \|h_M\|\} \tag{7}$$

where the infimum is taken over all possible h_js such that $f(\mathbf{x}) = \mu + h_1(\alpha_1'\mathbf{x}) + h_2(\alpha_2'\mathbf{x}) + \cdots + h_p(\alpha_p'\mathbf{x})$ and $h_j \in \bar{\mathcal{S}}_m$. It is not hard to check that J is a pseudo-norm on $\mathcal{F}_M$.

We now propose to estimate f_0 by the following penalized least squares estimator:

$$\hat{f}_\lambda = \arg\min_{f \in \mathcal{F}_M} \left\{ \frac{1}{n} \sum_{i=1}^n (y_i - f(\mathbf{x}_i))^2 + \lambda_n J(f) \right\}, \tag{8}$$

where $\lambda_n > 0$ is a tuning parameter. The following theorem states that our estimate achieves the same convergence rate as the usual additive model.

Theorem 1. *If f_0 is not a constant, and $\lambda_n^{-1} = O_p(n^{2m/(2m+1)}) J^{(2m-1)/(2m+1)}(f_0)$, then*

$$\frac{1}{n} \sum_{i=1}^n \left(\hat{f}_\lambda(\mathbf{x}_i) - f_0(\mathbf{x}_i) \right)^2 = O_p(\lambda_n) J(f_0). \tag{9}$$

If f_0 is a constant, then

$$\frac{1}{n} \sum_{i=1}^n \left(\hat{f}_\lambda(\mathbf{x}_i) - f_0(\mathbf{x}_i) \right)^2 = O_p\left(\max\{n^{-1}, \lambda_n^{-1/(2m-1)} n^{-2m/(2m-1)}\} \right). \tag{10}$$

Proof. Recall that $\hat{f}, f_0 \in \mathcal{F}_q$. We have $\hat{f} - f_0 \in \mathcal{F}_{2q} \subseteq \mathcal{F}_{2M}$. Therefore

$$\frac{\hat{f} - f_0}{J(\hat{f}) + J(f_0)} \in \{f \in \mathcal{F}_{2M} : \|f\| \le 1\} =: \mathcal{G}. \tag{11}$$

Denote the set on the right-hand side $\mathcal{G}$. In the light of Theorem 10.2 of van de Geer (2000), it suffices to show that

$$H_\infty(\delta, \mathcal{G}) \le C\delta^{-1/m}. \tag{12}$$

Hereafter, we use C to denote a generic positive constant that does not depend on δ.

For any $f_1, f_2 \in \mathcal{G}$, there exists h_js and g_js such that

$$f_1(\mathbf{x}) = h_1(\alpha_1'\mathbf{x}) + \cdots + h_{2M}(\alpha_{2M}'\mathbf{x}); \tag{13}$$

$$f_2(\mathbf{x}) = g_1(\beta_1'\mathbf{x}) + \cdots + g_{2M}(\beta_{2M}'\mathbf{x}), \tag{14}$$

where $h_j, g_j \in \bar{\mathcal{S}}_m$ and

$$\|h_1\|^2 + \cdots + \|h_{2M}\|^2 \le 1;$$
$$\|g_1\|^2 + \cdots + \|g_{2M}\|^2 \le 1.$$

Recall that

$$\|h_j\|^2 = \sum_{s=1}^{m-1} \left\{ h_j^{(s)}(1) - h_j^{(s)}(0) \right\}^2 + \int \left(h_j^{(m)} \right)^2, \quad j = 1, 2, \ldots, 2p. \tag{15}$$

Following the same argument as that of Lemma A.1 in Lin and Zhang (2006), one can show that $\|h_j'\|_\infty \equiv \max |h_j'(u)| \le 1$, $\|h_j\|_\infty \le 1$, $j = 1, 2, \ldots, 2M$ and the same holds true for g_js. Denote $\mathcal{B}_m$ the unit ball in $\mathcal{S}_m$. Then $h_j, g_j \in \mathcal{B}_m$. It is well known that the δ entropy of $\mathcal{B}_m$ for the supreme norm is bounded by

$$H_\infty(\delta, \mathcal{B}_m) \le C\delta^{-1/m}. \tag{16}$$

In other words, $\mathcal{B}_m$ can be covered by $N = \exp(C\delta^{-1/m})$ balls with radius $\delta/4M$. Note that

$$\begin{aligned} \|f_1 - f_2\|_\infty &\le \sum_{j=1}^{2M} \|g_j(\beta_j'\cdot) - h_j(\alpha_j'\cdot)\|_\infty \\ &\le \sum_{j=1}^{2M} \left(\|g_j(\beta_j'\cdot) - g_j(\alpha_j'\cdot)\|_\infty \right. \\ &\quad \left. + \|g_j(\alpha_j'\cdot) - h_j(\alpha_j'\cdot)\|_\infty \right) \\ &\le \sum_{j=1}^{2M} \left(\|g_j'\|_\infty \|\beta_j - \alpha_j\|_1 + \|g_j - h_j\|_\infty \right) \\ &\le \sum_{j=1}^{2M} \left(\|\beta_j - \alpha_j\|_1 + \|f_j - h_j\|_\infty \right) \end{aligned} \tag{17}$$

where $\|\beta_j - \alpha_j\|_1 = \sum_k |\beta_{jk} - \alpha_{jk}|$. Because a unit sphere in $\mathbf{R}^p$ can be covered by $C\delta^{-p}$ balls of radius $\delta/4p$, $\mathcal{G}$ can be covered by $CN^{2M}\delta^{-p}$, which implies the desired entropy condition (12). □

Note that additive models are also members of $\mathcal{F}$. According to Stone (1985), $n^{-2m/(2m+1)}$ is the optimal rate in

estimating f_0 if it follows an additive model. This suggests that the rate obtained in Theorem 1 is optimal and can not be improved. Allowing projection directions renders the additive index model more flexibility than the additive model. Theorem 1 shows that such flexibility can be gained without loss in terms of estimability asymptotically.

3. COMPUTATION

We now discuss the practical aspects in computing the penalized least squares estimator (8). The objective function on the right-hand side of (8) can be minimized in an iterative fashion. Similar to the COSSO, minimizing the objective function on the right-hand side of (8) is equivalent to minimizing

$$\frac{1}{n}\sum_{i=1}^{n}\{y_i - f(\mathbf{x}_i)\}^2 + \lambda_0\sum_{j=1}^{M}\tau_j^{-1}\|h_j\|^2 + \frac{\lambda^2}{4\lambda_0}\sum_{j=1}^{M}\tau_j \tag{18}$$

for any $\lambda_0 > 0$ with respect to both f and τ_js under the constraint that $\tau_j \geq 0$, $j = 1, \ldots, M$. Hereafter, we abbreviate the subscript of the tuning parameters that signifies their dependence on the sample size. Now we can minimize (18) with respect to the projection directions, ridge functions and τ_js iteratively.

Write $A = (\alpha_1, \ldots, \alpha_M)$ and $\mathbf{z} = A'\mathbf{x}$. When A and τ_js are known, (18) becomes

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \mu - \sum_{j=1}^{p}h_j(z_{ij})\right)^2 + \lambda_0\sum_{j=1}^{M}\tau_j^{-1}\|h_j\|^2, \tag{19}$$

the usual smoothing spline estimate of the additive model (Wahba, 1990). The representer Lemma (Wahba, 1990) implies that the minimizer of (19) can be given as

$$h_j(u) = \tau_j\sum_{i=1}^{n}c_iK(u; z_{ij}), \qquad j = 1, \ldots, M, \tag{20}$$

where $K(\cdot,\cdot)$ is the reproducing kernel associated with $\bar{\mathcal{S}}_m$ and z_{ij} is the jth factor of the ith observation, i.e., $z_{ij} = \alpha_j'\mathbf{x}_i$. Furthermore,

$$\|h_j\|^2 = \tau_j\sum_{i,k=1}^{n}c_ic_kK(z_{ij}, z_{kj}). \tag{21}$$

Plugging both formulae back to (19), we can get a closed form solution for μ and $\mathbf{c} = (c_1, \ldots, c_n)'$. Readers are referred to Wahba (1990) for details.

Now suppose that $\mathbf{c}$ and μ are known. Minimizing (18) with respect to A can be done using Newton iteration. We update A column by column. Consider, for example, updating α_j. Write

$$r_i = y_i - \left\{\mu + \sum_{k\neq j}h_k(\alpha_k'\mathbf{x}_i)\right\} \tag{22}$$

with α_js from the previous iteration and h_js being given by (20). Our goal is to solve

$$\min_{\alpha_j\in\mathbf{R}^p}\frac{1}{n}\sum_{i=1}^{n}\{r_i - h_j(\alpha_j'\mathbf{x}_i)\}^2, \qquad \text{subject to } \|\alpha_j\| = 1. \tag{23}$$

A Lagrange formulation leads to

$$\min_{\alpha_j\in\mathbf{R}^p}\left[\sum_{i=1}^{n}\{r_i - h_j(\alpha_j'\mathbf{x}_i)\}^2 + \theta\|\alpha_j\|^2\right], \tag{24}$$

for some constant θ. First order condition yields

$$\sum_{i=1}^{n}\left[\{r_i - h_j(\alpha_j'\mathbf{x}_i)\}\,h_j'(\alpha_j'\mathbf{x}_i)\mathbf{x}_i\right] = \theta\alpha_j. \tag{25}$$

Multiplying both sides by α_j,

$$\theta = \sum_{i=1}^{n}\left[\{r_i - h_j(\alpha_j'\mathbf{x})\}\,h_j'(\alpha_j'\mathbf{x}_i)\alpha_j'\mathbf{x}_i\right]. \tag{26}$$

To this end, we need to be able to compute the gradient and hessian of (24). Note that from (20)

$$h_j'(u) = \tau_j\sum_{i=1}^{n}c_i\frac{\partial K(u; z_{ij})}{\partial u}; \tag{27}$$

$$h_j''(u) = \tau_j\sum_{i=1}^{n}c_i\frac{\partial^2 K(u; z_{ij})}{\partial u^2}. \tag{28}$$

Both can be easily evaluated. The gradient and hessian of (24) can then be deduced from

$$G(\alpha_j) = -2\sum_{i=1}^{n}\left[\{r_i - h_j(\alpha_j'\mathbf{x})\}\,h_j'(\alpha_j'\mathbf{x}_i)\mathbf{x}_i\right] + 2\theta\alpha_j; \tag{29}$$

$$H(\alpha_j) = 2\sum_{i=1}^{n}\Big[\{h_j'(\alpha_j'\mathbf{x}_i)\}^2 - \{r_i - h_j(\alpha_j'\mathbf{x})\}\,h_j''(\alpha_j'\mathbf{x}_i)\Big]\mathbf{x}_i\mathbf{x}_i' + 2\theta I. \tag{30}$$

Now a Newton iteration would update the current estimate α_j by $\beta_j/\|\beta_j\|$ where

$$\beta_j = \alpha_j - H^{-1}(\alpha_j)\,G(\alpha_j). \tag{31}$$

Lastly, we update τ_js. This can be done in a similar fashion as the COSSO. Denote K_j an $n\times n$ matrix whose (i,l) entry is $K(\alpha_j'\mathbf{x}_i, \alpha_j'\mathbf{x}_l)$ and

$$F = (K_1\mathbf{c}, \ldots, K_p\mathbf{c}). \tag{32}$$

Define $\tilde{\mathbf{y}} = \mathbf{y} - n\lambda_0\mathbf{c}/2 - \mu\mathbf{1}$. It is not hard to show that the τ_js that minimize (18) are also the solution to

(33)
$$\min_{\tau_1,\ldots,\tau_p} \{\tilde{\mathbf{y}} - F(\tau_1,\ldots,\tau_p)'\}' \{\tilde{\mathbf{y}} - F(\tau_1,\ldots,\tau_p)'\} + \frac{\lambda^2}{4\lambda_0}\sum_{j=1}^{p}\tau_j$$
subject to $\tau_j \geq 0$.

This is the Lagrange formulation of

(34)
$$\min_{\tau_1,\ldots,\tau_p} \{\tilde{\mathbf{y}} - F(\tau_1,\ldots,\tau_p)'\}' \{\tilde{\mathbf{y}} - F(\tau_1,\ldots,\tau_p)'\}$$
$$\text{subject to } \tau_j \geq 0, \sum_{j=1}^{p}\tau_j \leq T$$

for some $T > 0$, which is in the same form as the so-called nonnegative garrote proposed for variable selection in linear regression (Breiman, 1995). This optimization problem is a quadratic program and can be solved efficiently using the standard quadratic program solver.

In summary, (18) can be minimized using the following algorithm.

Algorithm.

(I) Initialize A with a random $p \times M$ matrix normalized so that each column has norm one.
(II) Initialize $\tau_1 = \ldots = \tau_M = 1$.
(III) Choose λ_0 by the generalized cross validation and compute the initial estimate of μ and h_js using the smoothing spline algorithm (Wahba, 1990).
(IV) Compute the first and second derivative of h_js using (27) and (28).
(V) Update α_js by a one-step Newton iteration as described before.
(VI) Given A, update μ and h_js using the smoothing spline algorithm.
(VII) Update τ_js by solving (34).
(VIII) Go back to step (IV) until a certain convergence criterion is met.

Thus far, we have assumed that the tuning parameter λ, or equivalently T of (34) is fixed. In practice, it can be selected using cross-validation. In cross-validation, the full data set $\mathcal{D}$ is randomly split into K subsets of about the same size, denoted by $\mathcal{D}^{(1)},\ldots,\mathcal{D}^{(K)}$. For each $k = 1,\ldots,K$, we use the data in $\mathcal{D}-\mathcal{D}^{(k)}$ to estimate the model parameters and the data in $\mathcal{D}^{(k)}$ to validate. The squared prediction error could be used as the performance measure in our case. For each candidate value of M, the K-fold cross-validated score is defined as

(35)
$$\mathrm{CV}(T) = \sum_{k=1}^{K}\sum_{i\in I_k}(y_i - \hat{f}^{(-k)}(\mathbf{x}_i))^2$$

where I_k is the index set of the data in $\mathcal{D}^{(k)}$, and $\hat{f}^{(-k)}$ is the estimated regression function f using the training data set $\mathcal{D}-\mathcal{D}^{(k)}$. Typically, K is set to be five or ten and $\mathrm{CV}(T)$ is minimised over a grid of values of T. Let $\hat{T}$ be the minimiser of $\mathrm{CV}(T)$. Our final estimate of f is based on $\hat{T}$ and the full data set.

4. NUMERICAL EXAMPLES

We now illustrate the finite sample performance of the proposed method through several sets of numerical examples.

4.1 Partial single index model

We begin with a partial single index model from Carroll et al. (1997).

(36)
$$y = \sin\left[\frac{\pi(\alpha'\mathbf{x} - A)}{B - A}\right] + \beta z + \epsilon,$$

where z is a binary variable taking value 1 and 0 respectively for half of the observations, $\mathbf{x} = (x_1, x_2, x_3)'$ and x_i are independent observations from a uniform distribution $U[0,1]$, $A = \sqrt{3}/2 - 1.645/\sqrt{12}$, $B = \sqrt{3}/2 + 1.645/\sqrt{12}$, and $\epsilon \sim N(0, 0.1^2)$. The parameters used in this example are $\alpha = (1,1,1)'/\sqrt{3}$ and $\beta = 0.3$. Following Carroll et al. (1997), the sample size is $n = 200$.

The estimate we developed in Section 2 can be easily extended to accommodate the linear component βz. To estimate f and β, we minimize

(37)
$$\left\{\frac{1}{n}\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i) - \beta z)^2 + \lambda_n J(f)\right\},$$

with respect to both f and β. We apply this estimate to one hundred simulated data sets. The left panel of Figure 1 shows a typical simulated data set together with the estimated regression functions for both values of Z. The overall performance is also summarized in Table 1 and the box plots presented in the right panel of Figure 1. These results compare favorably with those reported by Carroll et al. (1997).

4.2 Additive index model

To gain further insight of predictive performance of the additive index model and the proposed penalized least squares estimate, we consider a more complex model. The simulated data consist of observations from the regression model $y = f_0(\mathbf{x}) + \epsilon$ where

$$f_0(\mathbf{x}) = 5\sin(2\pi\alpha_1'\mathbf{x}) + \frac{4\sin(2\pi\alpha_2'\mathbf{x})}{2 - \sin(2\pi\alpha_2'\mathbf{x})} + 6\times\big\{0.1\sin(2\pi\alpha_3'\mathbf{x}) + 0.2\cos(2\pi\alpha_3'\mathbf{x}) + 0.3\sin^2(2\pi\alpha_3'\mathbf{x}) + 0.4\cos^3(2\pi\alpha_3'\mathbf{x}) + 0.5\sin^3(2\pi\alpha_3'\mathbf{x})\big\}.$$

For each simulated data set, the predictor $\mathbf{x}$ is generated in two steps: we first sample $(\mathbf{z}, w)$ from a uniform distribution

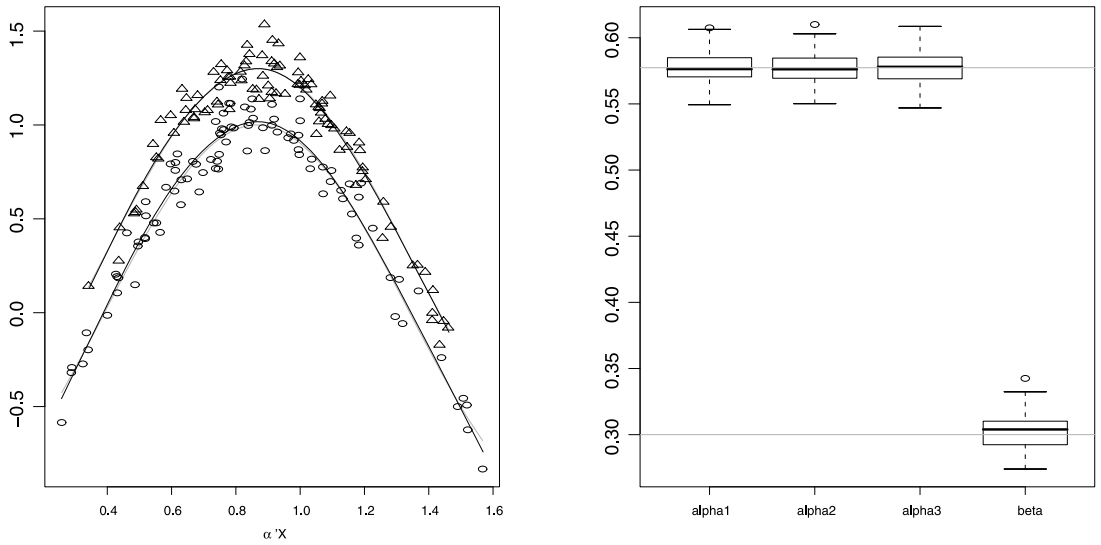

Figure 1. Simulation Example: left panel gives a typical simulated data set with size 200. The open circles correspond to observations with $z = 0$ and triangles correspond to observations with $z = 1$. The solid black lines are the estimated functions for $z = 0$ and 1 respectively. The grey lines, almost coincide with the estimated functions represent the true regression function. The right panel summarizes the parameter estimate for 100 repetitions. The grey horizontal lines correspond to the true values for α and β respectively.

Table 1. Mean values of the parameter estimation for the simulated example. The numbers in the parentheses are the standard errors

α_1	α_2	α_3	β
0.575 (0.003)	0.576 (0.001)	0.579 (0.002)	0.302 (0.001)

over $[0, 1]^{11}$ and then compute $x_j = (z_j + tu)/(1+t)$ for some t. We consider $t = 0$ and $t = 1$, corresponding to uncorrelated and correlated predictors respectively. The projection directions are also generated in two steps: a 10×10 random matrix was first generated and then we normalize each column so that it has norm one. We discard all the random matrices whose smallest eigenvalue is smaller than 0.1 to ensure that the projection directions are sufficiently linearly independent. The regression noise $\epsilon \sim N(0, (5/3)^2)$ to yield a signal to noise ratio 3:1. For each simulated data set, we generate 500 observations and estimate f using the proposed penalized least squares estimate. Note that we assumed that q is unknown in our estimation procedure. To measure the performance of an estimate of f_0, we use the mean squared error defined as

$$MSE(\hat{f}) = E_{\mathbf{x}} \left\{ \hat{f}(\mathbf{x}) - f_0(\mathbf{x}) \right\}^2 \tag{38}$$

where the expectation is taken with respect to $\mathbf{x}$. To evaluate the MSE, we generate an additional $10,000$ $\mathbf{x}'$s in the same fashion as before. We then estimate the MSE using its sample version:

$$\widehat{MSE}(\hat{f}) = \frac{1}{10000} \sum_{i=1}^{10000} \left\{ \hat{f}(\mathbf{x}_i^*) - f_0(\mathbf{x}_i^*) \right\}^2 \tag{39}$$

where $\mathbf{x}_i^*$s are the additionally generated $\mathbf{x}$s. For comparison, we also included the MARS (Friedman, 1991) and the projection pursuit regression of Friedman and Stuezle (1981). The MARS is a nonparametric regression technique with built-in variable selection features. For the projection pursuit regression, we use five fold cross validation to determine the number of components. The left panel of Figure 2 shows the boxplot of the MSE for the three methods when $t = 0$ summarized over 200 simulated data sets. The right panel corresponds to $t = 1$. From Figure 2, we observe that the proposed penalized least squares estimate enjoys superior performance in estimating the regression function f.

4.3 Generalized additive index model

Our last example illustrates how the the proposed method can be applied to more general regression settings. Although we have focused on mean regression, the additive index model can also be extended to more general regression settings such as the generalized linear model. In the generalized regression, the conditional mean $\mu(\mathbf{x}) = E(y|\mathbf{x})$ is related to a canonical function $\eta(\mathbf{x})$ via the so-called link function $\eta = g(\mu)$ which is known. For example the usual Gaussian regression amounts to the $g(\mu) = \mu$. A popular example

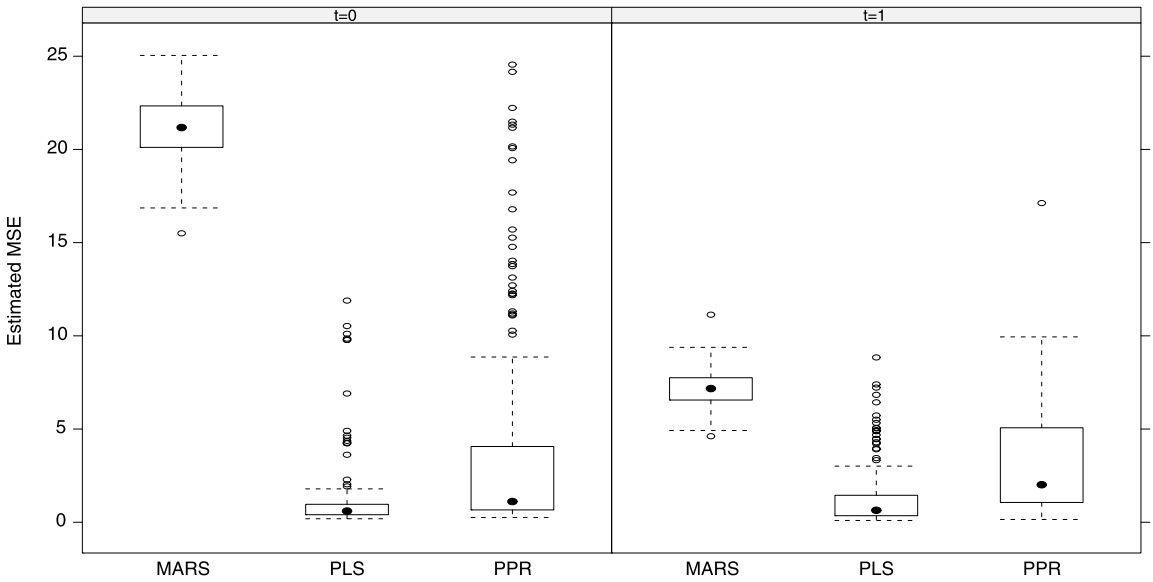

Figure 2. Simulation Example: boxplot of estimated MSE for three methods: MARS, the proposed penalized least squares estimate (PLS) and the original projection pursuit regression (PPR).

Table 2. Estimated projection directions for the Pima Indian Diabetes data

	term 1	term 2	term 3
pregnant	0.01	0.12	−0.15
glucose	−0.24	0.36	−0.05
pressure	0.45	0.26	−0.03
triceps	−0.16	−0.14	0.47
insulin	0.25	0.33	−0.30
mass	0.36	0.78	−0.07
pedigree	−0.32	0.21	−0.09
age	−0.65	−0.09	0.80

of the generalized regression is the logisitic regression for binary responses where $\eta(\mathbf{x})$ is the conditional log-odds ratio

$$\eta(\mathbf{x}) = \log \frac{\Pr(y=1|\mathbf{x})}{\Pr(y=0|\mathbf{x})}. \tag{40}$$

The idea of the additive index model can be generalized to these situations by assuming

$$\eta(\mathbf{x}) = \mu + h_1(\alpha_1'\mathbf{x}) + \cdots + h_q(\alpha_q'\mathbf{x}). \tag{41}$$

The generalized additive index model can be estimated in a similar manner as the usual additive index model. A deviance function such as the negative log-likelihood is used in place of the least squares of (8):

$$\mathcal{L}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell\left[\eta(\mathbf{x}_i); y_i\right] + \lambda J(f). \tag{42}$$

For example, in the case of logisitic regression, we estimate the generalized additive index model by minimizing

$$\mathcal{L}(f) = \frac{1}{n}\sum_{i=1}^{n}\left[y_i\eta(\mathbf{x}_i) - \log\left(1+e^{\eta(\mathbf{x}_i)}\right)\right] + \lambda J(f), \tag{43}$$

where $\eta(\cdot)$ is given by (41). The readers are referred to McCullagh and Nelder (1989) for more general discussion regarding the choice of the deviance. Similar to the generalized linear model $\mathcal{L}$ can be minimized by the iteratively reweighted least squares. In each iteration, the deviance function is approximated by a weighted least squares and can therefore be optimized using the algorithm given in Section 2. According to our experience, it suffices to run our algorithm only for one iteration at each step in the iteratively re-weighted least squares and the algorithm usually converges in a few iterations.

To illustrate the idea of the generalized additive index model, we consider here a real data example. The Pima Indians Diabetes data have 768 observations on nine variables. The purpose is to predict whether or not a particular subject has diabetes using eight remaining variables. It was often used as a benchmark data set for classification. An application of the additive index model and the penalized least squares estimate yields an additive model with three linear factors. The estimated projection directions are given in Table 2 and their corresponding ridge functions are given in Figure 3. The classification error of the additive index model is estimated as 18.3% using ten fold cross validation, which compares favorably with other popular classification tools (Newman et al., 1998).

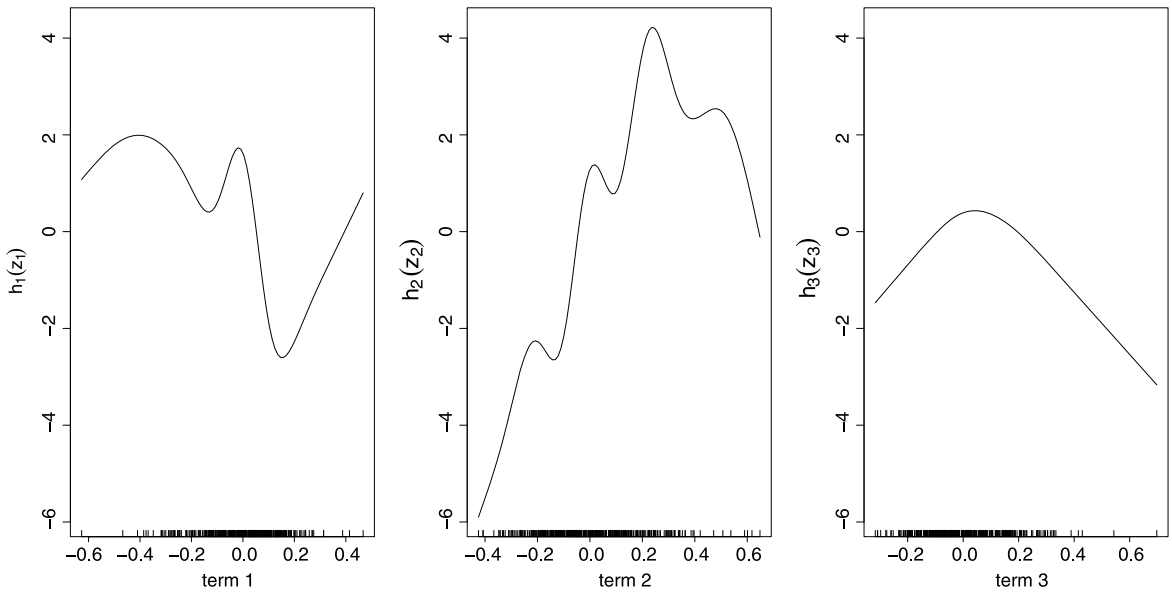

Figure 3. Estimated ridge functions for the Pima Indian Diabetes data.

5. CONCLUSION

In this paper, we propose a penalized least squares estimator to simultaneously determine the number of factors and estimate the regression function in the additive index model. The method becomes the recently proposed COSSO estimate if the factors are known in advance and the additive index model reduces to the additive model. We show the regression function estimate can achieve the optimal convergence rate. Thanks to the efficient algorithm for computing the smoothing spline estimate and their derivatives, the proposed estimate can be computed using an iterative algorithm.

Numerical experiments have been conducted to demonstrate the flexibility of the additive index model and the efficiency of the proposed estimate. We also used a real data example to show that the idea of the additive index model can be extended beyond the usual mean regression which is the primary focus of this paper.

Received 17 October 2010

REFERENCES

Breiman, L. (1995), Better subset regression using the nonnegative garrote, *Technometrics*, **37**, 373–384. MR1365720

Carroll, P., Fan, J., Gijbels, I. and Wand, M. (1997), Generalized partial linear single-index models, *Journal of the American Statistical Association*, **92(438)**, 477–489. MR1467842

Chen, H. (1991), Estimation of a projection-pursuit type regression model, *Annals of Statistics*, **19(1)**, 142–157. MR1091843

Chiou, J. and Müller, H. (2004), Quasi-likelihood regression with multiple indices and smooth link and variance functions, *Scandinavian J. Statistics*, **31**, 367–386. MR2087831

Diaconis, P. and Shahshahani, M. (1984), On nonlinear functions of linear combinations, *SIAM Journal of Scientific Computing*, **5(1)**, 175–191. MR0731890

Friedman, J. (1991), Multivariate adaptive regression splines, *Annals of Statistics*, **19**, 1–67. MR1091842

Friedman, J. and Stuetzle, W. (1981), Projection pursuit regression, *Journal of the American Statistical Association*, **76(376)**, 817–823. MR0650892

Lin, Y. and Zhang, H. (2006), Component selection and smoothing in smoothing spline analysis of variance models. *Annals of Statistics*, **34**, 2272–2297. MR2291500

McCullagh, P. and Nelder, J. (1989), *Generalized Linear Models*, London: Chapman and Hall. MR0727836

Newman, D., Hettich, S., Blake, C. and Merz, C. (1998), UCI repository of machine learning databaset [http://www.ics.uci.edu/˜mlearn/MLRepository.html], Irvine, CA: University of California, Department of Information and Computer Science.

Stone, C. (1985), Additive regression and other nonparametric models, *Annals of Statistics*, **13**, 689–705. MR0790566

van de Geer, S. (2000), *Empirical Processes in M-Estimation*, Cambridge: Cambridge University Press.

Wahba, G. (1990), *Spline Models for Observational Data*, Philadephia: SIAM. MR1045442

Yuan, M. (2008), On the identifiability of additive index model, Manuscript.

Lingyan Ruan
H. Milton Stewart School
of Industrial and Systems Engineering
Georgia Institute of Technology
USA

Ming Yuan
H. Milton Stewart School
of Industrial and Systems Engineering
Georgia Institute of Technology
USA
E-mail address: myuan@isye.gatech.edu