SpAM: Sparse Additive Models

Pradeep Ravikumar[†] **Han Liu**[†‡] **John Lafferty**[*†] **Larry Wasserman**[‡†]

[†]Machine Learning Department
[‡]Department of Statistics
[*]Computer Science Department

Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

We present a new class of models for high-dimensional nonparametric regression and classification called sparse additive models (SpAM). Our methods combine ideas from sparse linear modeling and additive nonparametric regression. We derive a method for fitting the models that is effective even when the number of covariates is larger than the sample size. A statistical analysis of the properties of SpAM is given together with empirical results on synthetic and real data, showing that SpAM can be effective in fitting sparse nonparametric models in high dimensional data.

1 Introduction

Substantial progress has been made recently on the problem of fitting high dimensional linear regression models of the form $Y_i = X_i^T\beta + \epsilon_i$, for $i = 1, \ldots, n$. Here Y_i is a real-valued response, X_i is a p-dimensional predictor and ϵ_i is a mean zero error term. Finding an estimate of β when $p > n$ that is both statistically well-behaved and computationally efficient has proved challenging; however, the lasso estimator (Tibshirani (1996)) has been remarkably successful. The lasso estimator $\widehat{\beta}$ minimizes the ℓ_1-penalized sums of squares

$$\sum_i (Y_i - X_i^T\beta) + \lambda \sum_{j=1}^{p} |\beta_j| \tag{1}$$

with the ℓ_1 penalty $\|\beta\|_1$ encouraging sparse solutions, where many components $\widehat{\beta}_j$ are zero. The good empirical success of this estimator has been recently backed up by results confirming that it has strong theoretical properties; see (Greenshtein and Ritov, 2004; Zhao and Yu, 2007; Meinshausen and Yu, 2006; Wainwright, 2006).

The nonparametric regression model $Y_i = m(X_i)+\epsilon_i$, where m is a general smooth function, relaxes the strong assumptions made by a linear model, but is much more challenging in high dimensions. Hastie and Tibshirani (1999) introduced the class of additive models of the form

$$Y_i = \sum_{j=1}^{p} m_j(X_{ij}) + \epsilon_i \tag{2}$$

which is less general, but can be more interpretable and easier to fit; in particular, an additive model can be estimated using a coordinate descent Gauss-Seidel procedure called backfitting. An extension of the additive model is the functional ANOVA model

$$Y_i = \sum_{1\le j\le p} m_j(X_{ij}) + \sum_{j<k} m_{j,k}(X_{ij}, X_{ik}) + \sum_{j<k<\ell} m_{j,k,\ell}(X_{ij}, X_{ik}, X_{i\ell}) + \cdots + \epsilon_i \tag{3}$$

which allows interactions among the variables. Unfortunately, additive models only have good statistical and computational behavior when the number of variables p is not large relative to the sample size n.

In this paper we introduce sparse additive models (SpAM) that extend the advantages of sparse linear models to the additive, nonparametric setting. The underlying model is the same as in (2), but constraints are placed on the component functions $\{m_j\}_{1\le j\le p}$ to simultaneously encourage smoothness of each component and sparsity across components; the penalty is similar to that used by the COSSO of Lin and Zhang (2006). The SpAM estimation procedure we introduce allows the use of arbitrary nonparametric smoothing techniques, and in the case where the underlying component functions are linear, it reduces to the lasso. It naturally extends to classification problems using generalized additive models. The main results of the paper are (i) the formulation of a convex optimization problem for estimating a sparse additive model, (ii) an efficient backfitting algorithm for constructing the estimator, (iii) simulations showing the estimator has excellent behavior on some simulated and real data, even when p is large, and (iv) a statistical analysis of the theoretical properties of the estimator that support its good empirical performance.

2 The SpAM Optimization Problem

In this section we describe the key idea underlying SpAM. We first present a population version of the procedure that intuitively suggests how sparsity is achieved. We then present an equivalent convex optimization problem. In the following section we derive a backfitting procedure for solving this optimization problem in the finite sample setting.

To motivate our approach, we first consider a formulation that scales each component function g_j by a scalar β_j, and then imposes an ℓ_1 constraint on $\beta = (\beta_1, \ldots, \beta_p)^T$. For $j \in \{1, \ldots, p\}$, let $\mathcal{H}_j$ denote the Hilbert space of measurable functions $f_j(x_j)$ of the single scalar variable x_j, such that $\mathbb{E}(f_j(X_j)) = 0$ and $\mathbb{E}(f_j(X_j)^2) < \infty$, furnished with the inner product

$$\left\langle f_j,\, f_j' \right\rangle = \mathbb{E}\left(f_j(X_j)\, f_j'(X_j) \right). \tag{4}$$

Let $\mathcal{H}^{\text{add}} = \mathcal{H}_1 + \mathcal{H}_2 + \ldots, \mathcal{H}_p$ denote the Hilbert space of functions of $(x_1, \ldots, x_p)$ that have an additive form: $f(x) = \sum_j f_j(x_j)$. The standard additive model optimization problem, in the population setting, is

$$\min_{f_j \in \mathcal{H}_j,\, 1\le j\le p} \mathbb{E}\left(Y - \sum_{j=1}^p f_j(X_j) \right)^2 \tag{5}$$

and $m(x) = \mathbb{E}(Y \mid X = x)$ is the unknown regression function. Now consider the following modification of this problem that imposes additional constraints:

$$(P) \qquad \min_{\beta \in \mathbb{R}^p,\, g_j \in \mathcal{H}_j} \quad \mathbb{E}\left(Y - \sum_{j=1}^p \beta_j g_j(X_j) \right)^2 \tag{6a}$$

$$\text{subject to} \quad \sum_{j=1}^p |\beta_j| \le L \tag{6b}$$

$$\mathbb{E}\left(g_j^2\right) = 1,\ \ j = 1, \ldots, p \tag{6c}$$

$$\mathbb{E}\left(g_j\right) = 0,\ \ j = 1, \ldots, p \tag{6d}$$

noting that g_j is a function while β is a vector. Intuitively, the constraint that β lies in the ℓ_1-ball $\{\beta : \|\beta\|_1 \le L\}$ encourages sparsity of the estimated β, just as for the parametric lasso. When β is sparse, the estimated additive function $f(x) = \sum_{j=1}^p f_j(x_j) = \sum_{j=1}^p \beta_j g_j(x_j)$ will also be sparse, meaning that many of the component functions $f_j(\cdot) = \beta_j g_j(\cdot)$ are identically zero. The constraints (6c) and (6c) are imposed for identifiability; without (6c), for example, one could always satisfy (6a) by rescaling.

While this optimization problem makes plain the role ℓ_1 regularization of β to achieve sparsity, it has the unfortunate drawback of not being convex. More specifically, while the optimization problem is convex in β and $\{g_j\}$ separately, it is not convex in β and $\{g_j\}$ jointly.

However, consider the following related optimization problem:

$$(Q) \qquad \min_{f_j \in \mathcal{H}_j} \quad \mathbb{E}\left(Y - \sum_{j=1}^p f_j(X_j)\right)^2 \tag{7a}$$

$$\text{subject to} \quad \sum_{j=1}^p \sqrt{\mathbb{E}(f_j^2(X_j))} \;\leq\; L \tag{7b}$$

$$\mathbb{E}(f_j) = 0, \;\; j = 1, \ldots, p. \tag{7c}$$

This problem is convex in $\{f_j\}$. Moreover, the solutions to problems (P) and (Q) are equivalent:

$$\left(\left\{\beta_j^*\right\}, \left\{g_j^*\right\}\right) \text{ optimizes } (P) \text{ implies } \left\{f_j^* = \beta_j^* g_j^*\right\} \text{ optimizes } (Q);$$

$$\left\{f_j^*\right\} \text{ optimizes } (Q) \text{ implies } \left(\left\{\beta_j^* = (\|f_j\|_2)^T\right\}, \left\{g_j^* = f_j^*/\|f_j^*\|_2\right\}\right) \text{ optimizes } (P).$$

While optimization problem (Q) has the important virtue of being convex, the way it encourages sparsity is not intuitive; the following observation provides some insight. Consider the set $C \subset \mathbb{R}^4$ defined by $C = \left\{(f_{11}, f_{12}, f_{21}, f_{22})^T \in \mathbb{R}^4 : \sqrt{f_{11}^2 + f_{12}^2} + \sqrt{f_{21}^2 + f_{22}^2} \leq L\right\}$. Then the projection $\pi_{12}C$ onto the first two components is an ℓ_2 ball. However, the projection $\pi_{13}C$ onto the first and third components is an ℓ_1 ball. In this way, it can be seen that the constraint $\sum_j \|f_j\|_2 \leq L$ acts as an ℓ_1 constraint across components to encourage sparsity, while it acts as an ℓ_2 constraint within components to encourage smoothness, as in a ridge regression penalty. It is thus crucial that the norm $\|f_j\|_2$ appears in the constraint, and not its square $\|f_j\|_2^2$. For the purposes of sparsity, this constraint could be replaced by $\sum_j \|f_j\|_q \leq L$ for any $q \geq 1$. In case each f_j is linear, $(f_j(x_{1j}), \ldots, f(x_{nj})) = \beta_j(x_{1j}, \ldots, x_{nj})$, the optimization problem reduces to the lasso.

The use of scaling coefficients together with a nonnegative garrote penalty, similar to our problem (P), is considered by Yuan (2007). However, the component functions g_j are fixed, so that the procedure is not asymptotically consistent. The form of the optimization problem (Q) is similar to that of the COSSO for smoothing spline ANOVA models (Lin and Zhang, 2006); however, our method differs significantly from the COSSO, as discussed below. In particular, our method is scalable and easy to implement even when p is much larger than n.

3 A Backfitting Algorithm for SpAM

We now derive a coordinate descent algorithm for fitting a sparse additive model. We assume that we observe $Y = m(X) + \epsilon$, where ϵ is mean zero Gaussian noise. We write the Lagrangian for the optimization problem (Q) as

$$\mathcal{L}(f, \lambda, \mu) = \frac{1}{2}\mathbb{E}\left(Y - \sum_{j=1}^p f_j(X_j)\right)^2 + \lambda \sum_{j=1}^p \sqrt{\mathbb{E}(f_j^2(X_j))} + \sum_j \mu_j \mathbb{E}(f_j). \tag{8}$$

Let $R_j = Y - \sum_{k \neq j} f_k(X_k)$ be the jth residual. The stationary condition for minimizing $\mathcal{L}$ as a function of f_j, holding the other components f_k fixed for $k \neq j$, is expressed in terms of the Frechet derivative $\delta\mathcal{L}$ as

$$\delta\mathcal{L}(f, \lambda, \mu; \delta f_j) = \mathbb{E}\left[(f_j - R_j + \lambda v_j)\delta f_j\right] = 0 \tag{9}$$

for any $\delta f_j \in \mathcal{H}_j$ satisfying $\mathbb{E}(\delta f_j) = 0$, where $v_j \in \partial\sqrt{\mathbb{E}(f_j^2)}$ is an element of the subgradient, satisfying $\sqrt{\mathbb{E}v_j^2} \leq 1$ and $v_j = f_j \Big/ \sqrt{\mathbb{E}(f_j^2)}$ if $\mathbb{E}(f_j^2) \neq 0$. Therefore, conditioning on X_j, the stationary condition (9) implies

$$f_j + \lambda v_j = \mathbb{E}(R_j \mid X_j). \tag{10}$$

Letting $P_j = \mathbb{E}[R_j \mid X_j]$ denote the projection of the residual onto $\mathcal{H}_j$, the solution satisfies

$$\left(1 + \frac{\lambda}{\sqrt{\mathbb{E}(f_j^2)}}\right) f_j \;=\; P_j \;\text{ if }\; \mathbb{E}(P_j^2) > \lambda \tag{11}$$

Input: Data (X_i, Y_i), regularization parameter λ.

Initialize $f_j = f_j^{(0)}$, for $j = 1, \ldots, p$.

Iterate until convergence:

For each $j = 1, \ldots, p$:

Compute the residual: $R_j = Y - \sum_{k \neq j} f_k(X_k)$;

Estimate the projection $P_j = \mathbb{E}[R_j \mid X_j]$ by smoothing: $\widehat{P}_j = \mathcal{S}_j R_j$;

Estimate the norm $s_j = \sqrt{\mathbb{E}[P_j]^2}$ using, for example, (15) or (35);

Soft-threshold: $f_j = \left[1 - \frac{\lambda}{\widehat{s}_j}\right]_+ \widehat{P}_j$;

Center: $f_j \leftarrow f_j - \text{mean}(f_j)$.

Output: Component functions f_j and estimator $\widehat{m}(X_i) = \sum_j f_j(X_{ij})$.

Figure 1: THE SPAM BACKFITTING ALGORITHM

and $f_j = 0$ otherwise. Condition (11), in turn, implies

$$\left(1 + \frac{\lambda}{\sqrt{\mathbb{E}(f_j^2)}}\right) \sqrt{\mathbb{E}(f_j^2)} = \sqrt{\mathbb{E}(P_j^2)} \quad \text{or} \quad \sqrt{\mathbb{E}(f_j^2)} = \sqrt{\mathbb{E}(P_j^2)} - \lambda. \tag{12}$$

Thus, we arrive at the following multiplicative soft-thresholding update for f_j:

$$f_j = \left[1 - \frac{\lambda}{\sqrt{\mathbb{E}(P_j^2)}}\right]_+ P_j \tag{13}$$

where $[\cdot]_+$ denotes the positive part. In the finite sample case, as in standard backfitting (Hastie and Tibshirani, 1999), we estimate the projection $\mathbb{E}[R_j \mid X_j]$ by a smooth of the residuals:

$$\widehat{P}_j = \mathcal{S}_j R_j \tag{14}$$

where $\mathcal{S}_j$ is a linear smoother, such as a local linear or kernel smoother. Let $\widehat{s}_j$ be an estimate of $\sqrt{\mathbb{E}[P_j^2]}$. A simple but biased estimate is

$$\widehat{s}_j = \frac{1}{\sqrt{n}} \|\widehat{P}_j\|_2 = \sqrt{\text{mean}(\widehat{P}_j^2)}. \tag{15}$$

More accurate estimators are possible; an example is given in the appendix. We have thus derived the SpAM backfitting algorithm given in Figure 1.

While the motivating optimization problem (Q) is similar to that considered in the COSSO (Lin and Zhang, 2006) for smoothing splines, the SpAM backfitting algorithm decouples smoothing and sparsity, through a combination of soft-thresholding and smoothing. In particular, SpAM backfitting can be carried out with any nonparametric smoother; it is not restricted to splines. Moreover, by iteratively estimating over the components and using soft thresholding, our procedure is simple to implement and scales to high dimensions.

3.1 SpAM for Nonparametric Logistic Regression

The SpAM backfitting procedure can be extended to nonparametric logistic regression for classification. The additive logistic model is

$$\mathbb{P}(Y = 1 \mid X) \equiv p(X; f) = \frac{\exp\left(\sum_{j=1}^p f_j(X_j)\right)}{1 + \exp\left(\sum_{j=1}^p f_j(X_j)\right)} \tag{16}$$

where $Y \in \{0, 1\}$, and the population log-likelihood is $\ell(f) = \mathbb{E}\left[Yf(X) - \log(1 + \exp f(X))\right]$. Recall that in the local scoring algorithm for generalized additive models (Hastie and Tibshirani, 1999) in the logistic case, one runs the backfitting procedure within Newton's method. Here one iteratively computes the transformed response for the current estimate f_0

$$Z_i = f_0(X_i) + \frac{Y_i - p(X_i;\, f_0)}{p(X_i;\, f_0)(1 - p(X_i;\, f_0))} \tag{17}$$

and weights $w(X_i) = p(X_i;\, f_0)(1 - p(X_i;\, f_0)$, and carries out a weighted backfitting of (Z, X) with weights w. The weighted smooth is given by

$$\widehat{P}_j = \frac{\mathcal{S}_j(wR_j)}{\mathcal{S}_j w}. \tag{18}$$

To incorporate the sparsity penalty, we first note that the Lagrangian is given by

$$\mathcal{L}(f, \lambda, \mu) = \mathbb{E}\left[\log(1 + \exp f(X)) - Yf(X)\right] + \lambda \sum_{j=1}^{p} \sqrt{\mathbb{E}(f_j^2(X_j))} + \sum_j \mu_j \mathbb{E}(f_j) \tag{19}$$

and the stationary condition for component function f_j is $\mathbb{E}\left(p - Y \mid X_j\right) + \lambda v_j = 0$ where v_j is an element of the subgradient $\partial\sqrt{\mathbb{E}(f_j^2)}$. As in the unregularized case, this condition is nonlinear in f, and so we linearize the gradient of the log-likelihood around f_0. This yields the linearized condition $\mathbb{E}\left[w(X)(f(X) - Z) \mid X_j\right] + \lambda v_j = 0$. When $\mathbb{E}(f_j^2) \neq 0$, this implies the condition

$$\left(\mathbb{E}\left(w \mid X_j\right) + \frac{\lambda}{\sqrt{\mathbb{E}(f_j)^2}}\right) f_j(X_j) = \mathbb{E}(wR_j \mid X_j). \tag{20}$$

In the finite sample case, in terms of the smoothing matrix $\mathcal{S}_j$, this becomes

$$f_j = \frac{\mathcal{S}_j(wR_j)}{\mathcal{S}_j w + \lambda \Big/ \sqrt{\mathbb{E}(f_j^2)}}. \tag{21}$$

If $\|\mathcal{S}_j(wR_j)\|_2 < \lambda$, then $f_j = 0$. Otherwise, this implicit, nonlinear equation for f_j cannot be solved explicitly, so we propose to iterate until convergence:

$$f_j \leftarrow \frac{\mathcal{S}_j(wR_j)}{\mathcal{S}_j w + \lambda\sqrt{n} \,/\, \|f_j\|_2}. \tag{22}$$

When $\lambda = 0$, this yields the standard local scoring update (18). An example of logistic SpAM is given in Section 5.

4 Properties of SpAM

4.1 SpAM is Persistent

The notion of risk consistency, or persistence, was studied by Juditsky and Nemirovski (2000) and Greenshtein and Ritov (2004) in the context of linear models. Let (X, Y) denote a new pair (independent of the observed data) and define the predictive risk when predicting Y with $f(X)$ by

$$R(f) = \mathbb{E}(Y - f(X))^2. \tag{23}$$

Since we consider predictors of the form $f(x) = \sum_j \beta_j g_j(x_j)$ we also write the risk as $R(\beta, g)$ where $\beta = (\beta_1, \ldots, \beta_p)$ and $g = (g_1, \ldots, g_p)$. Following Greenshtein and Ritov (2004), we say that an estimator $\widehat{m}_n$ is *persistent* relative to a class of functions $\mathcal{M}_n$ if

$$R(\widehat{m}_n) - R(m_n^*) \xrightarrow{P} 0 \tag{24}$$

where $m_n^* = \operatorname{argmin}_{f \in \mathcal{M}_n} R(f)$ is the predictive oracle. Greenshtein and Ritov (2004) showed that the lasso is persistent for the class of linear models $\mathcal{M}_n = \{f(x) = x^T\beta : \|\beta\|_1 \leq L_n\}$ if $L_n = o((n/\log n)^{1/4})$. We show a similar result for SpAM.

Theorem 4.1. *Suppose that $p_n \leq e^{n^\xi}$ for some $\xi < 1$. Then SpAM is persistent relative to the class of additive models $\mathcal{M}_n = \left\{ f(x) = \sum_{j=1}^p \beta_j g_j(x_j) : \|\beta\|_1 \leq L_n \right\}$ if $L_n = o\left(n^{(1-\xi)/4}\right)$.*

4.2 SpAM is Sparsistent

In the case of linear regression, with $m_j(X_j) = \beta_j^T X_j$, Wainwright (2006) shows that under certain conditions on n, p, $s = |\text{supp}(\beta)|$, and the design matrix X, the lasso recovers the sparsity pattern asymptotically; that is, the lasso estimator $\widehat{\beta}_n$ is *sparsistent*: $\mathbb{P}\left(\text{supp}(\beta) = \text{supp}(\widehat{\beta}_n)\right) \to 1$. We show a similar result for SpAM with the sparse backfitting procedure.

For the purpose of analysis, we use orthogonal function regression as the smoothing procedure. For each $j = 1, \ldots, p$ let ψ_j be an orthogonal basis for $\mathcal{H}_j$. We truncate the basis to finite dimension d_n, and let $d_n \to \infty$ such that $d_n/n \to 0$. Let Ψ_j denote the $n \times d$ matrix $\Psi_j(i,k) = \psi_{jk}(X_{ij})$. If $A \subset \{1, \ldots, p\}$, we denote by Ψ_A the $n \times d|A|$ matrix where for each $i \in A$, Ψ_i appears as a submatrix in the natural way. The SpAM optimization problem can then be written as

$$\min_{\beta} \frac{1}{2n} \left(Y - \textstyle\sum_{j=1}^p \Psi_j \beta_j\right)^2 + \lambda_n \sum_{j=1}^p \sqrt{\frac{1}{n} \beta_j^T \Psi_j^T \Psi_j \beta_j} \tag{25}$$

where each β_j is a d-dimensional vector. Let S denote the true set of variables $\{j : m_j \neq 0\}$, with $s = |S|$, and let S^c denote its complement. Let $\widehat{S}_n = \{j : \widehat{\beta}_j \neq 0\}$ denote the estimated set of variables from the minimizer $\widehat{\beta}_n$ of (25).

Theorem 4.2. *Suppose that Ψ satisfies the conditions*

$$\Lambda_{max}\left(\frac{1}{n}\Psi_S^T \Psi_S\right) \leq C_{max} < \infty \quad \textit{and} \quad \Lambda_{min}\left(\frac{1}{n}\Psi_S^T \Psi_S\right) \geq C_{min} > 0 \tag{26}$$

$$\left\| \left(\tfrac{1}{n}\Psi_{S^c}^T \Psi_S\right) \left(\tfrac{1}{n}\Psi_S^T \Psi_S\right)^{-1} \right\|_2^2 \leq \sqrt{\frac{C_{\min}}{C_{\max}}} \frac{1-\delta}{\sqrt{s}}, \quad \textit{for some } 0 < \delta \leq 1 \tag{27}$$

Let the regularization parameter $\lambda_n \to 0$ be chosen to satisfy

$$\lambda_n \sqrt{s d_n} \to 0, \quad \frac{s}{d_n \lambda_n} \to 0, \quad \textit{and} \quad \frac{d_n(\log d_n + \log(p-s))}{n\lambda_n^2} \to 0. \tag{28}$$

Then SpAM is sparsistent: $\mathbb{P}\left(\widehat{S}_n = S\right) \longrightarrow 1$.

5 Experiments

In this section we present experimental results for SpAM applied to both synthetic and real data, including regression and classification examples that illustrate the behavior of the algorithm in various conditions. We first use simulated data to investigate the performance of the SpAM backfitting algorithm, where the true sparsity pattern is known. We then apply SpAM to some real data. If not explicitly stated otherwise, the data are always rescaled to lie in a d-dimensional cube $[0,1]^d$, and a kernel smoother with Gaussian kernel is used. To tune the penalization parameter λ, we use a C_p statistic, which is defined as

$$C_p(\widehat{f}) = \frac{1}{n} \sum_{i=1}^n \left(Y_i - \textstyle\sum_{j=1}^p \widehat{f}_j(X_j)\right)^2 + \frac{2\widehat{\sigma}^2}{n} \sum_{j=1}^p \text{trace}(\mathcal{S}_j)\, \mathbf{1}[\widehat{f}_j \neq 0] \tag{29}$$

where $\mathcal{S}_j$ is the smoothing matrix for the j-th dimension and $\widehat{\sigma}^2$ is the estimated variance.

5.1 Simulations

We first apply SpAM to an example from (Härdle et al., 2004). A dataset with sample size $n = 150$ is generated from the following 200-dimensional additive model:

$$Y_i = f_1(x_{i1}) + f_2(x_{i2}) + f_3(x_{i3}) + f_4(x_{i4}) + \epsilon_i \tag{30}$$

$$f_1(x) = -2\sin(2x), \;\; f_2(x) = x^2 - \tfrac{1}{3}, \;\; f_3(x) = x - \tfrac{1}{2}, \;\; f_4(x) = e^{-x} + e^{-1} - 1 \tag{31}$$

and $f_j(x) = 0$ for $j \geq 5$ with noise $\epsilon_i \sim \mathcal{N}(0,1)$. These data therefore have 196 irrelevant dimensions. The results of applying SpAM with the plug-in bandwidths are summarized in Figure 2.

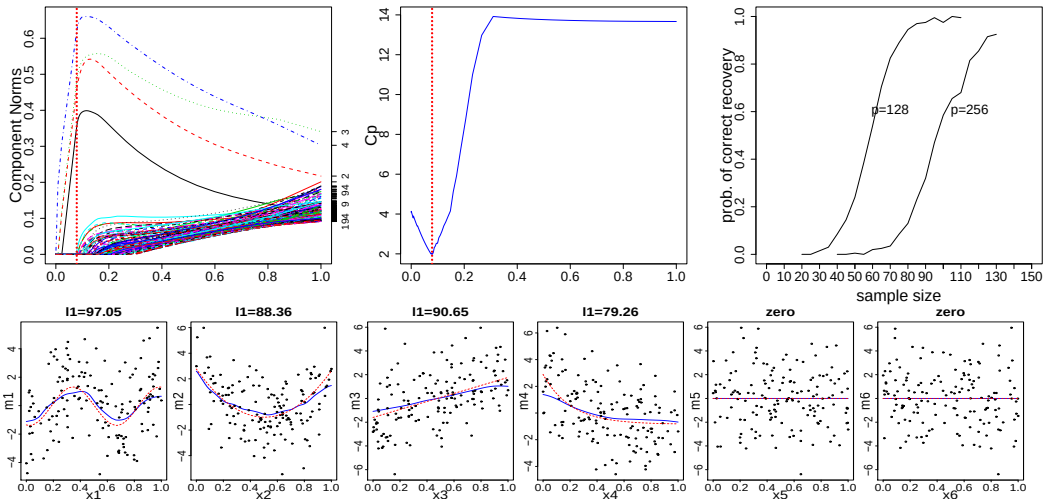

Figure 2: (Simulated data) Upper left: The empirical ℓ_2 norm of the estimated components as plotted against the tuning parameter λ; the value on the x-axis is proportional to $\sum_j \|\widehat{f}_j\|_2$. Upper center: The C_p scores against the tuning parameter λ; the dashed vertical line corresponds to the value of λ which has the smallest C_p score. Upper right: The proportion of 200 trials where the correct relevant variables are selected, as a function of sample size n. Lower (from left to right): Estimated (solid lines) versus true additive component functions (dashed lines) for the first 6 dimensions; the remaining components are zero.

5.2 Boston Housing

The Boston housing data was collected to study house values in the suburbs of Boston; there are altogether 506 observations with 10 covariates. The dataset has been studied by many other authors (Härdle et al., 2004; Lin and Zhang, 2006), with various transformations proposed for different covariates. To explore the sparsistency properties of our method, we add 20 irrelevant variables. Ten of them are randomly drawn from Uniform(0, 1), the remaining ten are a random permutation of the original ten covariates, so that they have the same empirical densities.

The full model (containing all 10 chosen covariates) for the Boston Housing data is:

$$\begin{aligned} \texttt{medv} \;=\; & \alpha + f_1(\texttt{crim}) + f_2(\texttt{indus}) + f_3(\texttt{nox}) + f_4(\texttt{rm}) + f_5(\texttt{age}) \\ & + f_6(\texttt{dis}) + f_7(\texttt{tax}) + f_8(\texttt{ptratio}) + f_9(\texttt{b}) + f_{10}(\texttt{lstat}) \end{aligned} \tag{32}$$

The result of applying SpAM to this 30 dimensional dataset is shown in Figure 3. SpAM identifies 6 nonzero components. It correctly zeros out both types of irrelevant variables. From the full solution path, the important variables are seen to be `rm`, `lstat`, `ptratio`, and `crim`. The importance of variables `nox` and `b` are borderline. These results are basically consistent with those obtained by other authors (Härdle et al., 2004). However, using C_p as the selection criterion, the variables `indux`, `age`, `dist`, and `tax` are estimated to be irrelevant, a result not seen in other studies.

5.3 SpAM for Spam

Here we consider an email spam classification problem, using the logistic SpAM backfitting algorithm from Section 3.1. This dataset has been studied by Hastie et al. (2001), using a set of 3,065 emails as a training set, and conducting hypothesis tests to choose significant variables; there are a total of 4,601 observations with $p = 57$ attributes, all numeric. The attributes measure the percentage of specific words or characters in the email, the average and maximum run lengths of upper case letters, and the total number of such letters. To demonstrate how SpAM performs well with sparse data, we only sample $n = 300$ emails as the training set, with the remaining 4301 data points used as the test set. We also use the test data as the hold-out set to tune the penalization parameter λ. The results of a typical run of logistic SpAM are summarized in Figure 4, using plug-in bandwidths.

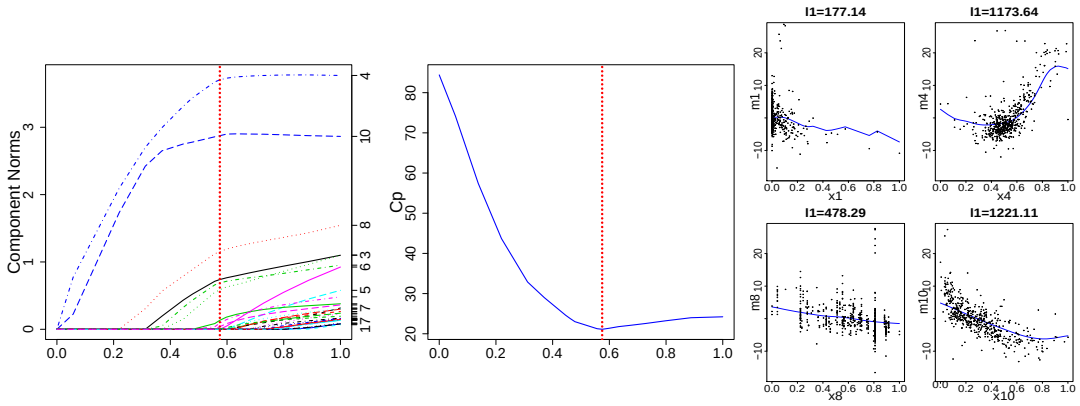

Figure 3: (Boston housing) Left: The empirical ℓ_2 norm of the estimated components versus the regularization parameter λ. Center: The C_p scores against λ; the dashed vertical line corresponds to best C_p score. Right: Additive fits for four relevant variables.

$\lambda(\times 10^{-3})$	ERROR	# ZEROS	SELECTED VARIABLES
5.5	0.2009	55	{ 8,54}
5.0	0.1725	51	{ 8, 9, 27, 53, 54, 57}
4.5	0.1354	46	{7, 8, 9, 17, 18, 27, 53, 54, 57, 58}
4.0	0.1083 ($\surd$)	20	{4, 6–10, 14–22, 26, 27, 38, 53–58}
3.5	0.1117	0	ALL
3.0	0.1174	0	ALL
2.5	0.1251	0	ALL
2.0	0.1259	0	ALL

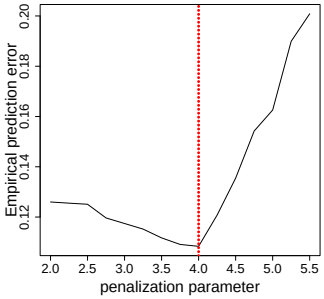

Figure 4: (Email spam) Classification accuracies and variable selection for logistic SpAM.

6 Acknowlegments

This research was supported in part by NSF grant CCF-0625879 and a Siebel Scholarship to PR.

References

GREENSHTEIN, E. and RITOV, Y. (2004). Persistency in high dimensional linear predictor-selection and the virtue of over-parametrization. *Journal of Bernoulli* **10** 971–988.

HÄRDLE, W., MÜLLER, M., SPERLICH, S. and WERWATZ, A. (2004). *Nonparametric and Semiparametric Models*. Springer-Verlag Inc.

HASTIE, T. and TIBSHIRANI, R. (1999). *Generalized additive models*. Chapman & Hall Ltd.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. H. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag.

JUDITSKY, A. and NEMIROVSKI, A. (2000). Functional aggregation for nonparametric regression. *Ann. Statist.* **28** 681–712.

LIN, Y. and ZHANG, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression. *Ann. Statist.* **34** 2272–2297.

MEINSHAUSEN, N. and YU, B. (2006). Lasso-type recovery of sparse representations for high-dimensional data. Tech. Rep. 720, Department of Statistics, UC Berkeley.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B, Methodological* **58** 267–288.

WAINWRIGHT, M. (2006). Sharp thresholds for high-dimensional and noisy recovery of sparsity. Tech. Rep. 709, Department of Statistics, UC Berkeley.

YUAN, M. (2007). Nonnegative garrote component selection in functional ANOVA models. In *Proceedings of AI and Statistics, AISTATS*.

ZHAO, P. and YU, B. (2007). On model selection consistency of lasso. *J. of Mach. Learn. Res.* **7** 2541–2567.