Statistica Sinica **21** (2011), 679-705

SURFACE ESTIMATION, VARIABLE SELECTION, AND THE NONPARAMETRIC ORACLE PROPERTY

Curtis B. Storlie[1], Howard D. Bondell[2], Brian J. Reich[2]
and Hao Helen Zhang[2]

[1] *University of New Mexico and* [2] *North Carolina State University*

Abstract: Variable selection for multivariate nonparametric regression is an important, yet challenging, problem due, in part, to the infinite dimensionality of the function space. An ideal selection procedure would be automatic, stable, easy to use, and have desirable asymptotic properties. In particular, we define a selection procedure to be nonparametric oracle (np-oracle) if it consistently selects the correct subset of predictors and, at the same time, estimates the smooth surface at the optimal nonparametric rate, as the sample size goes to infinity. In this paper, we propose a model selection procedure for nonparametric models, and explore the conditions under which the new method enjoys the aforementioned properties. Developed in the framework of smoothing spline ANOVA, our estimator is obtained via solving a regularization problem with a novel adaptive penalty on the sum of functional component norms. Theoretical properties of the new estimator are established. Additionally, numerous simulations and examples suggest that the new approach substantially outperforms other existing methods in the finite sample setting.

Key words and phrases: Adaptive LASSO, nonparametric regression, regularization method, smoothing spline ANOVA, variable selection.

1. Introduction

We consider the multiple predictor nonparametric regression model $y_i = f(\boldsymbol{x}_i) + \varepsilon_i$, $i = 1, \ldots, n$, where f is the unknown regression function, $\boldsymbol{x}_i = (x_{1,i}, \ldots, x_{p,i})$ is a p-dimensional vector of predictors, and the ε_i's are independent noise terms with mean 0 and variances σ_i^2. Many approaches to this problem have been proposed, such as kernel regression (Nadaraya (1964) and others) and locally weighted polynomial regression (LOESS), (Cleveland (1979)). See Schimek (2000) for a detailed list of references. When there are multiple predictors, these procedures suffer from the well known curse of dimensionality. Additive models (GAM's) (Hastie and Tibshirani (1990)) avoid some of the problems with high dimensionality and have been shown to be quite useful in cases when the true surface is nearly additive. A generalization of additive modeling is the Smoothing Spline ANOVA (SS-ANOVA) approach (Wahba (1990), Stone, Buja, and Hastie

(1994), Wahba et al. (1995), Lin (2000), and Gu (2002)). In SS-ANOVA, the function f is decomposed into several orthogonal functional components.

We are interested in the variable selection problem in the context of multiple predictor nonparametric regression. For example, it might be thought that the function f only depends on a subset of the p predictors. Traditionally this problem has been solved in a stepwise or best subset approach. The MARS procedure (Friedman (1991)) and variations thereof (Stone et al. (1997)) build an estimate of f by adding and deleting individual basis functions in a stepwise manner so that the omission of entire variables occurs as a side effect. However, stepwise variable selection is known to be unstable due to its inherent discreteness (Breiman (1995)). COmponent Selection Shrinkage Operator (COSSO; Lin and Zhang (2006)) performs variable selection via continuous shrinkage in SS-ANOVA models by penalizing the sum of norms of the functional components. Since each of the components is continuously shrunk toward zero, the resulting estimate is more stable than in subset or stepwise regression.

What are the desired properties of a variable selection procedure? For the parametric linear model Fan and Li (2001) discuss the *oracle* property. A method is said to possess the oracle property if it selects the correct subset of predictors with probability tending to one, and estimates the non-zero parameters as efficiently as would be possible if we knew which variables were uninformative ahead of time. Parametric models with the oracle property include Fan and Li (2001) and Zou (2006). In the context of nonparametric regression, we extend this notion by saying a nonparametric regression estimator has the nonparametric *(np)-oracle* property if it selects the correct subset of predictors with probability tending to one, and estimates the regression surface f at the optimal nonparametric rate.

None of the aforementioned nonparametric regression methods have been demonstrated to possess the np-oracle property. In particular, COSSO has a tendency to over-smooth the nonzero functional components in order to set the unimportant functional components to zero. We propose the adaptive COSSO (ACOSSO) to alleviate this major stumbling block. The intuition behind the ACOSSO is to penalize each component differently so that more flexibility is given to estimate functional components with more trend and/or curvature, while penalizing unimportant components more heavily. Hence, it is easier to shrink uninformative components to zero without much degradation to the overall model fit. This is motivated by the adaptive LASSO procedure for linear models of Zou (2006). We explore a special case where the ACOSSO possesses the np-oracle property. The practical benefit is demonstrated on several simulated and data examples where the ACOSSO substantially outperforms existing methods.

In Section 2 we review the necessary literature on smoothing spline ANOVA. The ACOSSO is introduced in Section 3, and its asymptotic properties are presented in Section 4. In Section 5 we discuss the computational details of the estimate. Its superior performance to existing methods is shown in simulations and on data in Sections 6 and 7, respectively. Section 8 concludes. Proofs are given in an appendix.

2. Smoothing Splines and the COSSO

In this section we review only the necessary concepts of SS-ANOVA needed for development. For a more detailed overview of Smoothing Splines and SS-ANOVA see Wahba (1990), Wahba et al. (1995), Schimek (2000), Gu (2002), and Berlinet and Thomas-Agnan (2004).

In the smoothing spline literature it is typically assumed that $f \in \mathcal{F}$ where $\mathcal{F}$ is a reproducing kernel Hilbert space (RKHS). Denote the reproducing kernel (r.k.), inner product, and norm of $\mathcal{F}$ as $K_{\mathcal{F}}$, $\langle \cdot, \cdot \rangle_{\mathcal{F}}$, and $\| \cdot \|_{\mathcal{F}}$ respectively. Often $\mathcal{F}$ is a space of functions with a certain degree of smoothness, for example, the second order Sobolev space, $\mathcal{S}^2 = \{g : g, g'$ are absolutely continuous and $g'' \in \mathcal{L}^2[0,1]\}$.

Without loss of generality, take $\boldsymbol{x} \in \mathcal{X} = [0,1]^p$. In smoothing spline (SS)-ANOVA, $\mathcal{F}$ is constructed by first taking a tensor product of p one-dimensional RKHS's. For example, let $\mathcal{H}_j$ be a RKHS on $[0,1]$ such that $\mathcal{H}_j = \{1\} \oplus \bar{\mathcal{H}}_j$ where $\{1\}$ is the RKHS consisting of only constant functions and $\bar{\mathcal{H}}_j$ is the RKHS consisting of functions $f_j \in \mathcal{H}_j$ such that $< f_j, 1 >_{\mathcal{H}_j} = 0$. If $\mathcal{F}$ is the tensor product of the $\mathcal{H}_j$, $j = 1, \ldots, p$, then

$$\mathcal{F} = \bigotimes_{j=1}^{p} \mathcal{H}_j = \{1\} \oplus \left\{ \bigoplus_{j=1}^{p} \bar{\mathcal{H}}_j \right\} \oplus \left\{ \bigoplus_{j<k} (\bar{\mathcal{H}}_j \otimes \bar{\mathcal{H}}_k) \right\} \oplus \cdots . \tag{2.1}$$

The right side of (2.1) has decomposed $\mathcal{F}$ into the constant space, the main effect spaces, the two-way interaction spaces, etc., which gives rise to SS-ANOVA. Typically (2.1) is truncated so that $\mathcal{F}$ includes only lower order interactions for better estimation and ease of interpretation. Regardless of the order of the interactions involved, $\mathcal{F}$ can be written in general as

$$\mathcal{F} = \{1\} \oplus \left\{ \bigoplus_{j=1}^{q} \mathcal{F}_j \right\}, \tag{2.2}$$

where $\{1\}, \mathcal{F}_1 \ldots \mathcal{F}_q$ is an orthogonal decomposition of the space and each of the $\mathcal{F}_j$ is itself a RKHS. We focus on two special cases, the additive model $f(\boldsymbol{x}) = b + \sum_{j=1}^{p} f_j(x_j)$, and the two-way interaction model $f(\boldsymbol{x}) = b + \sum_{j=1}^{p} f_j(x_j) + \sum_{j<k}^{p} f_{jk}(x_j, x_k)$, where $b \in \{1\}$, $f_j \in \bar{\mathcal{H}}_j$ and $f_{jk} \in \bar{\mathcal{H}}_j \otimes \bar{\mathcal{H}}_k$.

A smoothing spline estimate, $\hat{f}$, is the function $f \in \mathcal{F}$ that minimizes

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(\boldsymbol{x}_i))^2 + \lambda_0 \sum_{j=1}^{q} \frac{1}{\theta_j}\|P^j f\|_{\mathcal{F}}^2, \tag{2.3}$$

where $P^j f$ is the orthogonal projection of f onto the $\mathcal{F}_j$, $j = 1, \ldots, q$, that form an orthogonal partition of the space as in (2.2). We use the convention $0/0 = 0$ so that when $\theta_j = 0$ the minimizer satisfies $\|P^j f\|_{\mathcal{F}} = 0$.

The COSSO (Lin and Zhang (2006)) penalizes the sum of the norms instead of the squared norms, as in (2.3), and achieves (sparse) solutions in which some of the functional components are estimated to be exactly zero. Specifically, the COSSO estimate, $\hat{f}$, is the function $f \in \mathcal{F}$ that minimizes

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(\boldsymbol{x}_i))^2 + \lambda \sum_{j=1}^{q} \|P^j f\|_{\mathcal{F}}. \tag{2.4}$$

In Lin and Zhang (2006), $\mathcal{F}$ was formed using $\mathcal{S}^2$ with squared norm

$$\|f\|^2 = \left(\int_0^1 f(x)dx\right)^2 + \left(\int_0^1 f'(x)dx\right)^2 + \int_0^1 \left(f''(x)\right)^2 dx \tag{2.5}$$

for each of the $\mathcal{H}_j$ of (2.1). The reproducing kernel can be found in Wahba (1990).

3. An Adaptive Proposal

Although the COSSO is a significant improvement over classical stepwise procedures, it tends to oversmooth functional components. This seemingly prevents COSSO from achieving a nonparametric version of the oracle property. We propose an adaptive approach that uses individually weighted norms to smooth each of the components. Specifically, we select as our estimate the function $f \in \mathcal{F}$ that minimizes

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(\boldsymbol{x}_i))^2 + \lambda \sum_{j=1}^{q} w_j\|P^j f\|_{\mathcal{F}}, \tag{3.1}$$

where the $0 < w_j \leq \infty$ are weights that can depend on an initial estimate of f, denoted $\tilde{f}$. For example we could initially estimate f via the smoothing spline (2.3) with all $\theta_j = 1$ and λ_0 chosen by the generalized cross validation (GCV) criterion (Craven and Wahba (1979)). Note that there is only one tuning parameter, λ, in (3.1); the w_j's are not tuning parameters, rather they are weights to be estimated from the data in a manner described below.

3.1. Choosing adaptive weights

Given an initial estimate $\tilde{f}$, we wish to construct weights so that the prominent functional components enjoy the benefit of a smaller penalty relative to less important ones. One possible scheme is to make use of the L_2 norm of $P^j\tilde{f}$, $\|P^j\tilde{f}\|_{L_2} = (\int_{\mathcal{X}}(P^j\tilde{f}(\boldsymbol{x}))^2 d\boldsymbol{x})^{1/2}$. For a reasonable initial estimator, this quantity is a consistent estimate of $\|P^j f\|_{L_2}^2$, a term often used to quantify the importance of functional components. This suggests the choice

$$w_j = \|P^j\tilde{f}\|_{L_2}^{-\gamma}. \tag{3.2}$$

In Section 4, the use of these weights is shown to result in favorable theoretical properties.

There are other possibilities for the weights. As an extension of the adaptive LASSO for linear models, it may seem more natural to set $w_j = \|P^j\tilde{f}\|_{\mathcal{F}}^{-\gamma}$, but these weights do not provide an estimator with sound theoretical properties. Building an additive model using RKHS's with norm given by (2.5) leads to w_j that essentially require estimation of the functionals $\int_0^1(f_j''(x_j))^2 dx_j$, known to be a harder problem than estimating $\int_0^1 f_j^2(x)dx$ (Efromovich and Samarov (2000)). In light of the results of our empirical studies, we recommend the use of the weights in (3.2).

4. Asymptotic Properties

In this section we show that the ACOSSO possesses a nonparametric analog of the oracle property.

Throughout this section we assume the regression model $y_i = f_0(\boldsymbol{x}_i) + \varepsilon_i$, $i = 1, \ldots, n$. The unknown regression function $f_0 \in \mathcal{F}$ is additive in the predictors so that $\mathcal{F} = \{1\} \oplus \mathcal{F}_1 \oplus \cdots \oplus \mathcal{F}_p$, where each $\mathcal{F}_j$ is a space of functions corresponding to x_j. We assume that the ε_i are independent with $\mathrm{E}\,\varepsilon_i = 0$, and are uniformly sub-Gaussian. Thus, following van de Geer (2000), there exists $K > 0$ and $C > 0$ such that

$$\sup_n \max_{i=1,\ldots,n} E\Big[\exp\Big(\frac{\varepsilon_i^2}{K}\Big)\Big] \le C. \tag{4.1}$$

Let $\mathcal{S}^2$ denote the RKHS of second order Sobolev space endowed with the norm in (2.5) with $\mathcal{S}^2 = \{1\} \oplus \bar{\mathcal{S}}^2$. Also, define the squared norm of a function at the design points as $\|f\|_n^2 = 1/n\sum_{i=1}^n f^2(\boldsymbol{x}_i)$. Let U be the set of indexes for all uninformative functional components in the model $f_0 = b + \sum_{j=1}^p P^j f_0$, $j = 1, \ldots, p$, i.e., $U = \{j : P^j f_0 \equiv 0\}$.

Theorem 1 below states the convergence rate of ACOSSO when used to estimate an additive model. We sometimes write w_j and λ as $w_{j,n}$ and λ_n, respectively, to explicitly denote the dependence on n. We also use the notation

$X_n \sim Y_n$ to mean $X_n/Y_n = O_p(1)$ and $Y_n/X_n = O_p(1)$ for some sequences X_n and Y_n. The proofs of Theorem 1 and the other results in this section are deferred to the appendix.

Theorem 1. (Convergence Rate) *Assume that* $f_0 \in \mathcal{F} = \{1\} \oplus \bar{\mathcal{S}}_1^2 \oplus \cdots \oplus \bar{\mathcal{S}}_p^2$, *where* $\bar{\mathcal{S}}_j^2$ *is the space* $\bar{\mathcal{S}}^2$ *corresponding to the* j^{th} *input variable,* x_j, *and that the* ε_i *are independent and satisfy* (4.1). *Consider the ACOSSO estimate,* $\hat{f}$, *defined in* (3.1). *Suppose that* $w_{j,n}^{-1} = O_p(1)$ *for* $j = 1, \ldots, p$, *that* $w_{j,n} = O_p(1)$ *for* $j \in U^c$, *and that* $\lambda_n^{-1} = O_p(n^{4/5})$. *For* $\hat{f}$, *defined at* (3.1)*:*

(i) *If* $P^j f_0 \neq 0$ *for some* j, *then* $\|\hat{f} - f_0\|_n = O_p(\lambda^{1/2} w_{*,n}^{1/2})$ *where* $w_{*,n} = \min\{w_{1,n}, \ldots, w_{p,n}\}$*;*

(ii) *If* $P^j f_0 = 0$ *for all* j, *then* $\|\hat{f} - f_0\|_n = O_p(\max\{n^{-1/2}, n^{-2/3}\lambda^{-1/3} w_{*,n}^{-1/3}\})$.

Corollary 1. (Optimal Convergence of ACOSSO) *Assume that* $f_0 \in \mathcal{F} = \{1\} \oplus \bar{\mathcal{S}}_1^2 \oplus \cdots \oplus \bar{\mathcal{S}}_p^2$, *and that the* ε_i *are independent and satisfy* (4.1). *Consider* $\hat{f}$ *with weights* $w_{j,n} = \|P^j \tilde{f}\|_{L_2}^{-\gamma}$, *for* $\tilde{f}$ *given by* (2.3) *with* $\boldsymbol{\theta} = \mathbf{1}_p$ *and* $\lambda_{0,n} \sim n^{-4/5}$. *If* $\gamma > 3/4$ *and* $\lambda_n \sim n^{-4/5}$, *then* $\|\hat{f} - f_0\|_n = O_p(n^{-2/5})$ *if* $P^j f_0 \neq 0$ *for some* j, *and* $\|\hat{f} - f_0\|_n = O_p(n^{-1/2})$ *otherwise.*

We now discuss the properties of the ACOSSO in terms of model selection. In Theorem 2 and Corollary 2 we consider functions in the second order Sobolev space of periodic functions, denoted $\mathcal{S}^2_{\text{per}}$, where $\mathcal{S}^2_{\text{per}} = \{1\} \oplus \bar{\mathcal{S}}^2_{\text{per}}$. We take the design points to be $\{x_{1,i_1}, x_{2,i_2}, \ldots, x_{p,i_p}\}_{i_j=1}^{m}{}_{j=1}^{p}$, where $x_{j,k} = k/m$, $k = 1, \ldots, m$, $j = 1, \ldots, p$, so the sample size is $n = m^p$. This set-up was also used by Lin and Zhang (2006) to examine the model selection properties of the COSSO.

Theorem 2. (Selection Consistency) *Assume a tensor product design and* $f_0 \in \mathcal{F} = \{1\} \oplus \bar{\mathcal{S}}^2_{per,1} \oplus \cdots \oplus \bar{\mathcal{S}}^2_{per,p}$, *where* $\bar{\mathcal{S}}^2_{per,j}$ *is the space* $\bar{\mathcal{S}}^2_{per}$ *corresponding to the* j^{th} *input variable,* x_j. *Also assume that the* ε_i *are independent and satisfy* (4.1). *Then* $\hat{f}$ *satisfies* $P^j \hat{f} \equiv 0$ *for all* $j \in U$ *with probability tending to one if and only if* $n w_{j,n}^2 \lambda_n^2 \xrightarrow{p} \infty$ *as* $n \to \infty$ *for all* $j \in U$.

We say a nonparametric regression estimator, $\hat{f}$, has the nonparametric *(np)-oracle* property if $\|\hat{f} - f_0\|_n \to 0$ at the optimal rate while $P^j \hat{f} \equiv 0$ for all $j \in U$ with probability tending to one. This means that the error associated with surface estimation has the same order as that for any other optimal estimator. A corollary to Theorem 2 states that the ACOSSO with weights given by (3.2) has the *np-oracle* property.

Corollary 2. (Nonparametric Oracle Property) *Assume a tensor product design and* $f_0 \in \mathcal{F}$ *where* $\mathcal{F} = \{1\} \oplus \bar{\mathcal{S}}^2_{per,1} \oplus \cdots \oplus \bar{\mathcal{S}}^2_{per,p}$*, and that the* ε_i *are independent and satisfy* (4.1)*. Define weights,* $w_{j,n} = \|P^j \tilde{f}\|^{-\gamma}_{L_2}$*, for* $\tilde{f}$ *given by the traditional smoothing spline with* $\lambda_0 \sim n^{-4/5}$*, and* $\gamma > 3/4$*. If also* $\lambda_n \sim n^{-4/5}$*, then the ACOSSO estimator has the np-oracle property.*

Remark 1. The derivation of the variable selection properties of adaptive COSSO requires detailed investigation on the eigen-properties of the reproducing kernel, which is generally intractable. In the case of Theorem 2 and Corollary 2 however, f periodic and $\boldsymbol{x}$ a tensor product design, the eigenfunctions and eigenvalues of the associated reproducing kernel have a particularly simple form. Results for this design are often instructive for general cases, as suggested in Wahba (1990). We conjecture that the selection consistency of the adaptive COSSO also holds more generally, and this is supported by numerical results in Section 6. The derivation of variable selection properties in the general case is a technically difficult problem which is worthy of future investigation. Neither the f periodic assumption nor $\boldsymbol{x}$ a tensor product design are required for establishing the MSE consistency of the adaptive COSSO estimator in Theorem 1 and Corollary 1.

Remark 2. The COSSO (the ACOSSO with $w_{j,n} = 1$ for all j and n) does not appear to enjoy the np-oracle property. Notice that by Theorem 2, λ_n must go to zero slower than $n^{-1/2}$ in order to achieve asymptotically correct variable selection. However, even if $\lambda_n = n^{-1/2}$, Theorem 1 implies that the convergence rate is $O_p(n^{-1/4})$ which is not optimal. These results are not surprising given that the linear model can be obtained as a special case of ACOSSO by using $\mathcal{F} = \{f : f = \beta_0 + \sum_{j=1}^{p} \beta_j(x_j - 1/2)\}$. For this $\mathcal{F}$ the COSSO reduces to the LASSO which is known not to have the oracle property (Knight and Fu (2000), Zou (2006)). In contrast, the ACOSSO reduces to the adaptive LASSO (Zou (2006)) that is known to achieve the oracle property.

Remark 3. The distribution of the error terms ε_i in Theorems 1 and 2 need only be independent with sub-Gaussian tails (4.1). The distributions need not be Gaussian, and need not even be the same for each ε_i. In particular, this allows for heteroskedastic errors.

Remark 4. Theorems 1 and 2 assume an additive model, in which case functional component selection is equivalent to variable selection. In higher order interaction models, the main effect for x_j and *all* of the interaction functional components involving x_j must be set to zero in order to eliminate x_j from the model and achieve true variable selection. Thus, when interactions are involved, we use the term *variable selection* to refer to functional component selection.

5. Computation

Since the ACOSSO in (3.1) may be viewed as the COSSO in (2.4) with an "adaptive" RKHS, the computation proceeds in a manner similar to that for the COSSO. We first present an equivalent formulation of the ACOSSO, then describe how to minimize this equivalent formulation for a fixed value of the tuning parameter. Discussion of tuning parameter selection is delayed until Section 5.3.

5.1. Equivalent formulation

Consider the problem of finding $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)'$ and $f \in \mathcal{F}$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n}[y_i - f(\boldsymbol{x}_i)]^2 + \lambda_0 \sum_{j=1}^{q} \theta_j^{-1} w_j^{2-\vartheta} \|P^j f\|_{\mathcal{F}}^2 + \lambda_1 \sum_{j=1}^{q} w_j^{\vartheta}\theta_j, \text{ subject to } \theta_j \geq 0\ \forall j, \tag{5.1}$$

where $0 \leq \vartheta \leq 2$, $\lambda_0 > 0$ is a fixed constant, and $\lambda_1 > 0$ is a smoothing parameter. That the above optimization problem is equivalent to (3.1) has important implications for computation, since (5.1) is easier to solve.

Lemma 1. *Set* $\lambda_1 = \lambda^2/(4\lambda_0)$. (*i*) *If* $\hat{f}$ *minimizes* (3.1) *and* $\hat{\theta}_j = \lambda_0^{1/2}\lambda_1^{-1/2}w_j^{1-\vartheta}$ $\|P^j\hat{f}\|_{\mathcal{F}}$, $j = 1, \ldots, q$, *then* $(\hat{\boldsymbol{\theta}}, \hat{f})$ *minimizes* (5.1). (*ii*) *If* $(\hat{\boldsymbol{\theta}}, \hat{f})$ *minimizes* (5.1), *then* $\hat{f}$ *minimizes* (3.1).

5.2. Computational algorithm

The form at (5.1) gives a class of equivalent problems for $\vartheta \in [0, 2]$. For simplicity we consider the case $\vartheta = 0$, since the ACOSSO can be then viewed as having the same form as the COSSO with an adaptive RKHS. For a given value of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)'$, the minimizer of (5.1) is the smoothing spline of (2.3) with θ_j replaced by $w_j^{-2}\theta_j$. Hence, it is known (Wahba (1990) for example) that the solution has the form $f(\boldsymbol{x}) = b + \sum_{i=1}^{n} c_i K_{\boldsymbol{w},\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{x}_i)$ where $\boldsymbol{c} \in \Re^n$, $b \in \Re$, and $K_{\boldsymbol{w},\boldsymbol{\theta}} = \sum_{j=1}^{q}(\theta_j/w_j^2)K_{\mathcal{F}_j}$, with $\mathcal{F}_j$ corresponding to the decomposition in (2.2).

Let $\boldsymbol{K}_j$ be the $n \times n$ matrix $\{K_{\mathcal{F}_j}(\boldsymbol{x}_i, \boldsymbol{x}_j)\}_{i,j=1}^{n}$ and let $\mathbf{1}_n$ be the column vector consisting of n ones. Write the vector $\boldsymbol{f} = (f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n))'$ as $\boldsymbol{f} = b\mathbf{1}_n + (\sum_{j=1}^{q}(\theta_j/w_j^2)\boldsymbol{K}_j)\boldsymbol{c}$, $\boldsymbol{y} = (y_1, \ldots, y_n)'$, and define $\|\boldsymbol{v}\|_n^2 = 1/n\sum_{i=1}^{n} v_i^2$ for a vector $\boldsymbol{v}$ of length n. Now, for fixed $\boldsymbol{\theta}$, minimizing (5.1) is equivalent to

$$\min_{b,\boldsymbol{c}} \left\{ \frac{1}{n} \left\| \boldsymbol{y} - b\mathbf{1}_n - \sum_{j=1}^{q} \theta_j w_j^{-2} \boldsymbol{K}_j \boldsymbol{c} \right\|_n^2 + \lambda_0 \sum_{j=1}^{q} \theta_j w_j^{-2} \boldsymbol{c}'\boldsymbol{K}_j\boldsymbol{c} \right\}, \tag{5.2}$$

which is just the traditional smoothing spline problem in Wahba (1990). On the other hand if b and $\boldsymbol{c}$ were fixed, the $\boldsymbol{\theta}$ that minimizes (5.1) is the same as the solution to

$$\min_{\boldsymbol{\theta}} \left\{ \|\boldsymbol{z} - \boldsymbol{G\theta}\|_n^2 + n\lambda_1 \sum_{j=1}^{q} \theta_j \right\}, \text{ subject to } \theta_j \geq 0 \; \forall j, \tag{5.3}$$

where $\boldsymbol{g}_j = w_j^{-2}\boldsymbol{K}_j\boldsymbol{c}$, $\boldsymbol{G}$ is the $n \times p$ matrix with the j^{th} column being $\boldsymbol{g}_j$, and $\boldsymbol{z} = \boldsymbol{y} - b\mathbf{1}_n - (n/2)\lambda_0\boldsymbol{c}$. Notice that (5.3) is equivalent to

$$\min_{\boldsymbol{\theta}} \|\boldsymbol{z} - \boldsymbol{G\theta}\|_n^2 \quad \text{subject to } \theta_j \geq 0 \; \forall j \text{ and } \sum_{j=1}^{q} \theta_j \leq M, \tag{5.4}$$

for some $M > 0$. The formulation in (5.4) is a quadratic programming problem with linear constraints for which there exist many algorithms (see Goldfarb and Idnani (1982) for example). A reasonable scheme is then to iterate between (5.2) and (5.4). In each iteration (5.1) is decreased. We have observed that after the second iteration the change between iterations is small and decreases slowly.

5.3. Selecting the tuning parameter

In (5.1) there is only one tuning parameter, λ_1, or equivalently M of (5.4); changing the value of λ_0 only scale shifts the value of M being used, so λ_0 can be fixed at any positive value. We choose to initially fix $\boldsymbol{\theta} = \mathbf{1}_q$ and to find λ_0 to minimize the GCV score of the smoothing spline problem in (5.2). This has the effect of placing the θ_j's on a scale so that M roughly translates into the number of non-zero components. Hence, it seems reasonable to tune M on $[0, 2q]$ for example.

We used 5-fold cross validation ($5CV$) in our examples to tune M, though we also found that the BIC criterion (Schwarz (1978)) was quite useful for selecting M. We approximated the effective degrees of freedom, ν, by $\nu = \text{tr}(\boldsymbol{S})$, where $\boldsymbol{S}$ is the weight matrix corresponding to the smoothing spline fit with $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. This type of approximation gives an under-estimate of the actual df, but has been demonstrated to be useful (Tibshirani (1996)). We have found that the ACOSSO with $5CV$ tends to over-select non-zero components, just as Zou, Hastie, and Tibshirani (2007) found that AIC-type criteria over-selected non-zero coefficients in the LASSO. They recommend using BIC with the LASSO when the goal is variable selection, as do we for the ACOSSO.

6. Simulated Data Results

In this section we study the empirical performance of the ACOSSO estimate and compare it to several other existing methods. We display the results of four different versions of the ACOSSO. All versions used weights w_j given by (3.2) with $\gamma = 2$, since we found that $\gamma = 2$ produced the best overall results among $\gamma \in \{1.0, 1.5, 2.0, 2.5, 3.0\}$. The initial estimate, $\tilde{f}$, was either the traditional smoothing spline or the COSSO with λ selected by GCV. We also used either $5CV$ or BIC to tune M. Hence, the four versions of ACOSSO are ACOSSO-5CV-T, ACOSSO-5CV-C, ACOSSO-BIC-T, and ACOSSO-BIC-C, where (-T) and (-C) stand for using the traditional smoothing spline and COSSO, respectively, as the initial estimate.

We include the methods COSSO, MARS, stepwise GAM, Random Forest (Breiman (2001)), and the Gradient Boosting Method (GBM) (Friedman (2001)). The tuning parameter for COSSO was chosen via $5CV$. To fit MARS models we used the 'polymars' procedure in the R-package 'polspline'. Stepwise GAM, Random Forest, and GBM fits were obtained using the R-packages 'gam', 'randomForest', and 'gbm', respectively. Input parameters for these methods, such as *gcv* for MARS, *n.trees* for GBM, etc., were appropriately set to give best results.

Note that Random Forest and GBM are both black box prediction machines that produce function estimates that are difficult to interpret, and they are not intended for variable selection. They are, however, well-known for making accurate predictions. Thus, we included them to demonstrate the utility of the ACOSSO even in situations where prediction is the goal.

We also included the results of the traditional smoothing spline (2.3) when fit with only the informative variables. That is, we set $\theta_j = 0$ if $P^j f = 0$, and $\theta_j = 1$ otherwise, then chose λ_0 by GCV. This is referred to as the ORACLE estimator. Notice that the ORACLE estimator is only available in simulations where we know which variables are informative. Though the ORACLE cannot be used in practice, it gives us a baseline for the best estimation risk we could hope to achieve with other methods.

Performance is measured in terms of estimation risk and model selection, specifically with the variables $\hat{R}$, $\bar{\alpha}$, $\bar{\beta}$. Here $\hat{R}$ is the average of ISE values over 100 realizations, where the ISE $= \mathrm{E}_{\boldsymbol{X}}\{f(\boldsymbol{X}) - \hat{f}(\boldsymbol{X})\}^2$ is calculated for each realization via a monte carlo integration with 1,000 points. Let $\hat{\alpha}_j$ be the proportion of realizations such that $P^j \hat{f} \neq 0$, $j = 1, \ldots, q$; then $\bar{\alpha} = 1/|U| \sum_{j \in U} \hat{\alpha}_j$ where $U = \{j : P^j f \equiv 0\}$ and $|U|$ is the number of elements in U. If $\hat{\beta}_j$ is the proportion of realizations such that $P^j \hat{f} = 0$, $j = 1, \ldots, q$, then $\bar{\beta} = 1/|U^c| \sum_{j \in U^c} \hat{\beta}_j$ where U^c is the complement of U. Model size is the number of functional components included in the model averaged over the 100 realizations.

We first present a very simple example to highlight the benefit of using the ACOSSO; then we repeat the same examples used in the COSSO paper to offer a direct comparison on examples where the COSSO is known to perform well.

The only difference here is that we have increased the noise level to make these problems a bit more challenging.

Example 1. Four functions on $[0,1]$ were used as building blocks of regression functions in the simulations:

$$g_1(t)=t; \quad g_2(t) = (2t-1)^2; \quad g_3(t) = \frac{\sin(2\pi t)}{2-\sin(2\pi t)};$$
$$g_4(t)=0.1\sin(2\pi t)+0.2\cos(2\pi t)+0.3\sin^2(2\pi t)+0.4\cos^3(2\pi t)+0.5\sin^3(2\pi t). \quad (6.1)$$

In this example $\boldsymbol{X} \in \Re^{10}$ and we took $n = 100$ observations from the model $y = f(\boldsymbol{X}) + \varepsilon$, where $f(\boldsymbol{x}) = 5g_1(x_1) + 3g_2(x_2) + 4g_3(x_3) + 6g_4(x_4)$, and $\varepsilon \overset{i.i.d.}{\sim} \mathcal{N}(0, 3.03)$. Therefore $X_5, \ldots, X_{10}$ were uninformative. We first considered the case where $\boldsymbol{X}$ was uniform in $[0,1]^{10}$, in which case the signal to noise ratio (SNR) was 3:1 (here we have adopted the variance definition for signal to noise ratio, $\text{SNR} = [\text{Var}\,(f(\boldsymbol{X}))]/\sigma^2$). For comparison, the variances of the functional components were $\text{Var}\,\{5g_1(X_1)\} = 2.08$, $\text{Var}\,\{3g_2(X_2)\} = 0.80$, $\text{Var}\,\{4g_3(X_3)\} = 3.30$ and $\text{Var}\,\{6g_4(X_4)\} = 9.45$.

For the estimations ACOSSO, COSSO, MARS, and GBM, we restricted $\hat{f}$ to be a strictly additive function; the Random Forest function does not have an option for this. There were 10 functional components considered for inclusion in the ACOSSO model. Figure 1 gives plots of y versus $x_1, \ldots, x_4$, along with the true $P^j f$ component curves for a realization from Example 1. The true component curves, $j = 1, \ldots, 4$, along with the estimates given by ACOSSO-5CV-T and COSSO are shown in Figure 2, here without the data for added clarity. Notice that the ACOSSO captured more of the features of the $P^3 f$ component and, particularly, the $P^4 f$ component, since the reduced penalty on these components allowed it more curvature. In addition, since the weights more easily allow for curvature on components that need it, M did not need to be large (relatively) to allow a good fit to components like $P^3 f$ and $P^4 f$. This had the effect that components with less curvature, like the linear $P^1 f$, could be estimated more accurately by ACOSSO than by COSSO, as seen in Figure 2.

Figure 3 shows how the magnitudes of the estimated components change with the tuning parameter M for both the COSSO and ACOSSO for this realization; magnitudes of the estimated components are measured by their L_2 norms $\|P^j \hat{f}\|_{L_2}$, dashed lines are drawn at the true values of $\|P^j f\|_{L_2}$ for reference. Notice that estimated functional component norms given by ACOSSO were closer to the true values than those given by the COSSO in general. Also, the uninformative components were more heavily penalized in the ACOSSO making it harder for them to enter the model.

Incidentally using *GCV* or 5*CV* for tuning parameter selection for the ACOSSO on the above realization gave $M = 3.81$ and $M = 4.54$, respectively,

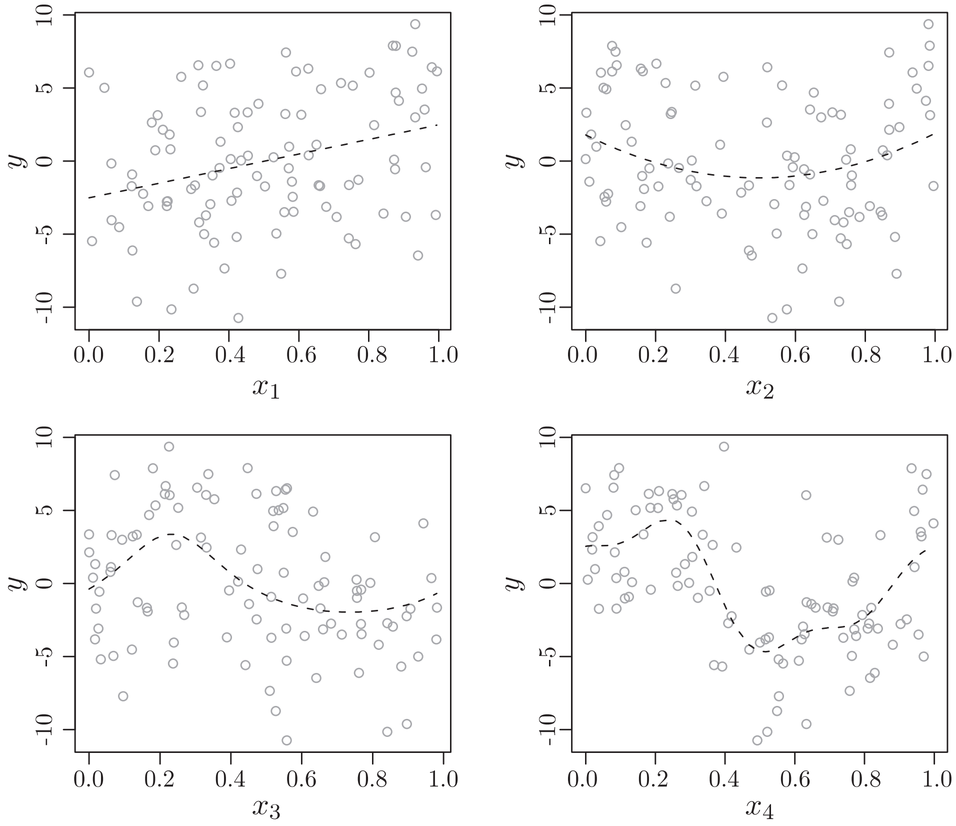

Figure 1. Plot of the true functional components, $P^j f$, $j = 1, \ldots, 4$, along with the data for a realization from Example 1.

both resulting in a model of five functional components for this run. The BIC method however gave $M = 2.97$, which resulted in the correct model of four functional components. This was a typical occurrence for realizations from this example, as can be seen in Table 1.

In Table 1 we can compare the risk and variable selection capability of the ACOSSO to the COSSO and other methods with $\boldsymbol{X}$ uniform on $[0, 1]^{10}$. All four of the ACOSSO methods did significantly better than COSSO and other methods in terms of risk. COSSO, MARS, GAM, Random Forest, and GBM had 131%, 180%, 150%, 349%, and 167% the risk of the ORACLE respectively, while the ACOSSO methods all had risk less than 108% that of the ORACLE. In terms of variable selection, the two ACOSSO-5CV methods had a much higher average type I error rate than the two ACOSSO-BIC methods and MARS. In fact ACOSSO-5CV-T had $\bar{\alpha} = 0.25$, which is quite high. Both ACOSSO-BIC methods, however, had $\bar{\alpha} \leq 0.03$, and had an average model size of close to 4.0, the correct number of components.

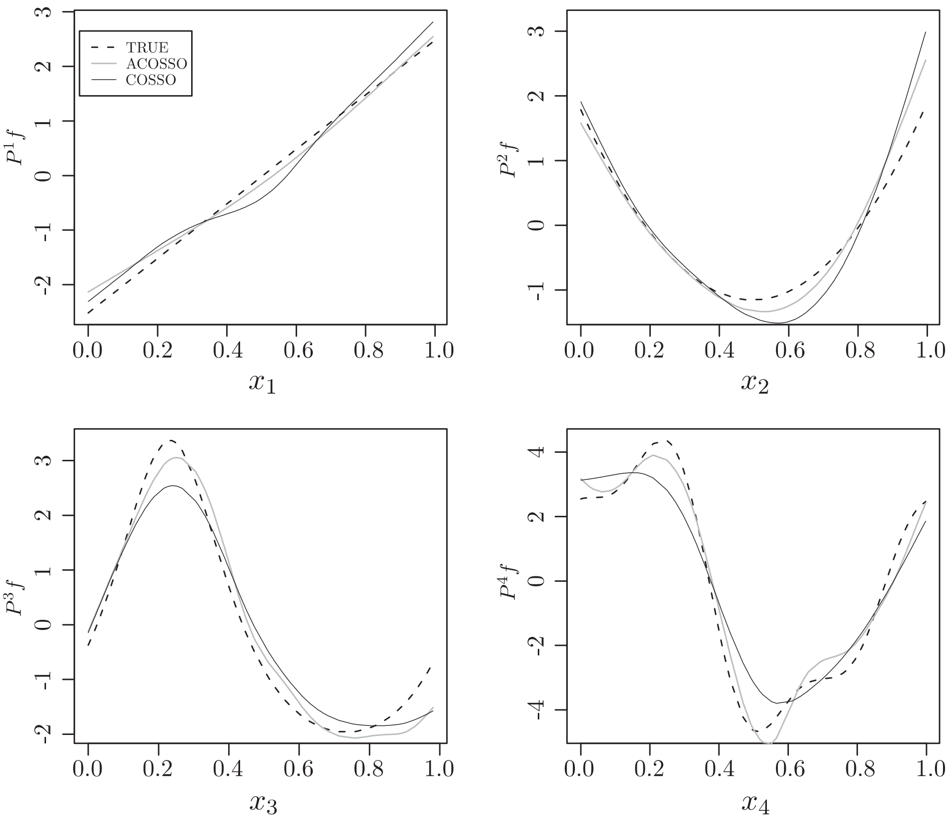

Figure 2. Plot of $P^j f$, $j = 1, \ldots, 4$, along with their estimates given by ACOSSO and COSSO for a realization from Example 1.

Table 1. Results of 100 realizations from Example 1 in the uniform case. Standard errors are given in parantheses.

	$\hat{R}$	$\bar{\alpha}$	$1 - \bar{\beta}$	model size
ACOSSO-5CV-T	1.204 (0.042)	0.252 (0.034)	0.972 (0.008)	5.4 (0.21)
ACOSSO-5CV-C	1.186 (0.048)	0.117 (0.017)	0.978 (0.007)	4.6 (0.11)
ACOSSO-BIC-T	1.257 (0.048)	0.032 (0.008)	0.912 (0.012)	3.8 (0.08)
ACOSSO-BIC-C	1.246 (0.064)	0.018 (0.006)	0.908 (0.014)	3.7 (0.07)
COSSO	1.523 (0.058)	0.095 (0.023)	0.935 (0.012)	4.3 (0.15)
MARS	2.057 (0.064)	0.050 (0.010)	0.848 (0.013)	3.7 (0.08)
GAM	1.743 (0.053)	0.197 (0.019)	0.805 (0.011)	4.4 (0.13)
Random Forest	4.050 (0.062)	NA	NA	10.0 (0.00)
GBM	1.935 (0.039)	NA	NA	10.0 (0.00)
ORACLE	1.160 (0.034)	0.000 (0.000)	1.000 (0.000)	4.0 (0.00)

Although the ACOSSO-5CV methods had higher $\bar{\alpha}$, they had better power than the other methods, as can be seen in the $1 - \bar{\beta}$ column of Table 1. Here

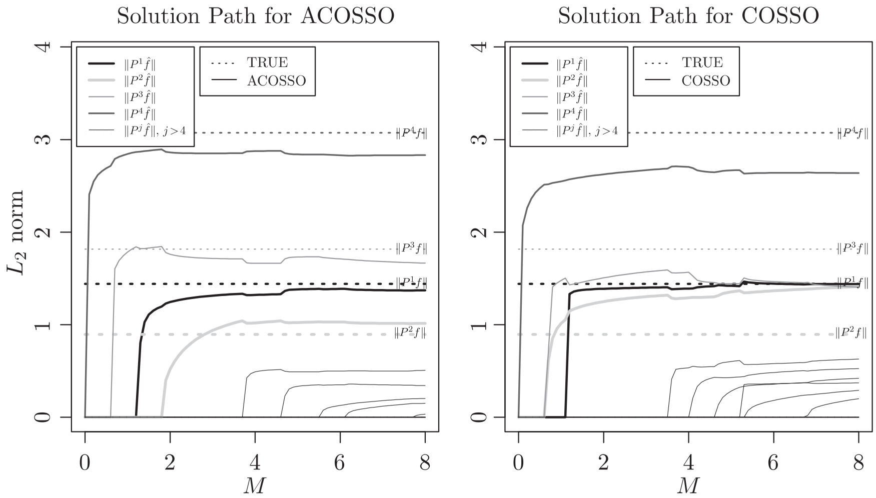

Figure 3. Plot of $\|P^j\hat{f}\|_{L_2}$ along with $\|P^jf\|_{L_2}$ by M for both ACOSSO and COSSO on a realization from Example 1.

$1-\bar{\beta}$ was almost completely determined by how well the methods do at including the second variable. The components P^1f, P^3f, and P^4f were included in the model nearly always for all of the methods. The percentage of the realizations that included P^2f was 65% for the ACOSSO-BIC methods, 75% for the COSSO, and only 40% and 23% for MARS and GAM, respectively. The percentage of the realizations that included P^2f was close to 90% for the ACOSSO-5CV methods but, as mentioned, the price paid was a higher type I error rate.

Example 2. Here $\boldsymbol{X} \in \Re^{60}$. We took 500 observations from $y = f(\boldsymbol{X}) + \varepsilon$. The regression function was additive in the predictors,

$$f(\boldsymbol{x}) = g_1(x_1) + g_2(x_2) + g_3(x_3) + g_4(x_4) + 1.5g_1(x_5) + 1.5g_2(x_6) + 1.5g_3(x_7)$$
$$+1.5g_4(x_8) + 2g_1(x_9) + 2g_2(x_{10}) + 2g_3(x_{11}) + 2g_4(x_{12}),$$

with $g_1, \ldots, g_4$ as given in (6.1). The noise variance was set to $\sigma^2 = 2.40$, yielding a SNR of 3:1 in the uniform case. Notice that $X_{13}, \ldots, X_{60}$ are uninformative. In this example, we took the two distributional families for the input vector $\boldsymbol{X}$ found in Lin and Zhang (2006).

Compound Symmetry: For an input $\boldsymbol{X} = (X_1, \ldots, X_p)$, let $X_j = (W_j + tU)/(1 + t), j = 1, \ldots, p$, where $W_1, \ldots, W_p$ and U are *i.i.d.* Unif$(0, 1)$. Thus, Corr $(X_j, X_k) = t^2/(1 + t^2)$ for $j \neq k$. The uniform distribution design corresponds to the case t = 0.

(*trimmed*) *AR*(1)*:* Let $W_1, \ldots, W_d$ be *i.i.d.* $\mathcal{N}(0, 1)$, and let $X_1 = W_1$, $X_j =$

Table 2. Estimation Risk based on 100 realizations from Example 2 under various covariance structures; standard errors are given in parantheses.

	Compound Symmetry			Trimmed AR(1)		
	$t = 0$	$t = 1$	$t = 3$	$\rho = -0.5$	$\rho = 0.0$	$\rho = 0.5$
ACOSSO-5CV-T	0.41 (0.01)	0.41 (0.01)	0.52 (0.01)	0.40 (0.01)	0.40 (0.01)	0.40 (0.01)
ACOSSO-5CV-C	0.42 (0.01)	0.40 (0.01)	0.43 (0.01)	0.40 (0.01)	0.40 (0.01)	0.40 (0.01)
ACOSSO-BIC-T	0.42 (0.01)	0.41 (0.01)	0.60 (0.02)	0.42 (0.01)	0.39 (0.01)	0.42 (0.01)
ACOSSO-BIC-C	0.42 (0.01)	0.42 (0.01)	0.47 (0.01)	0.45 (0.02)	0.39 (0.01)	0.43 (0.01)
COSSO	0.48 (0.01)	0.60 (0.01)	0.54 (0.01)	0.60 (0.02)	0.57 (0.01)	0.57 (0.01)
MARS	0.97 (0.02)	0.66 (0.01)	1.05 (0.02)	0.64 (0.01)	0.62 (0.01)	0.64 (0.01)
GAM	0.49 (0.01)	0.52 (0.01)	0.50 (0.01)	0.52 (0.01)	0.52 (0.01)	0.52 (0.01)
Random Forest	1.86 (0.01)	1.50 (0.01)	0.76 (0.00)	1.26 (0.01)	1.19 (0.01)	1.25 (0.01)
GBM	0.73 (0.01)	0.52 (0.01)	0.47 (0.00)	0.58 (0.01)	0.58 (0.01)	0.57 (0.01)
ORACLE	0.30 (0.00)	0.28 (0.00)	0.27 (0.00)	0.29 (0.00)	0.29 (0.01)	0.29 (0.01)

$\rho X_{j-1} + (1-\rho^2)^{1/2} W_j$, $j = 2, \ldots, p$. Then trim X_j to $[-2.5, 2.5]$ and scale to $[0, 1]$.

Table 2 shows the results of estimation risk for six distributions for the predictors. Again the ACOSSO methods had estimation risk much closer to ORACLE than the other methods. COSSO and GAM had very similar performance in this example and generally had the best risk among the other methods. One notable exception was the extremely high correlation case (Compound Symmetry, $t = 3$, where Corr $(\boldsymbol{X}_j, \boldsymbol{X}_k) = .9$ for $j \neq k$). Here ACOSSO-BIC-T and ACOSSO-5CV-T had risks near or above the risks of COSSO and GAM. GBM had the best risk in this particular case. However, the ACOSSO variants were substantially better overall than any of the other methods. A similar trend was also noticed (table not presented) for these six cases on the test function from Example 1.

Example 3. Here we considered a regression model with 10 predictors and several two way interactions, with

$$f(\boldsymbol{x}) = g_1(x_1) + g_2(x_2) + g_3(x_3) + g_4(x_4) + g_3(x_1x_2) + g_2(\frac{x_1+x_3}{2}) + g_1(x_3x_4)$$

so that $x_5, \ldots, x_{10}$ were uninformative. The noise variance was set at $\sigma^2 = 0.44098$ to give a SNR of 3:1. We considered only the uniform distribution on the predictors and evaluated performance at various sample sizes, $n = 100$, $n = 250$, and $n = 500$.

A summary of the estimation risk on 100, realizations can be found in Table 3. When $n = 100$, all of the other methods except Random Forest had substantially better risk than COSSO. However, the ACOSSO methods had risk comparable or better than the other methods, and less than half that of COSSO. The estimation risk for all methods improved as the sample size increased. However, stepwise

Table 3. Estimation Risk based on 100 realizations from Example 3 with $n = 100$, 250, and 500; standard errors are given in parantheses.

	$n = 100$	$n = 250$	$n = 500$
ACOSSO-5CV-T	0.139 (0.017)	0.055 (0.001)	0.034 (0.001)
ACOSSO-5CV-C	0.120 (0.011)	0.055 (0.001)	0.036 (0.001)
ACOSSO-BIC-T	0.200 (0.027)	0.054 (0.001)	0.034 (0.001)
ACOSSO-BIC-C	0.138 (0.016)	0.050 (0.001)	0.034 (0.001)
COSSO	0.290 (0.016)	0.093 (0.002)	0.057 (0.001)
MARS	0.245 (0.021)	0.149 (0.009)	0.110 (0.008)
GAM	0.149 (0.005)	0.137 (0.001)	0.136 (0.001)
Random Forest	0.297 (0.006)	0.190 (0.002)	0.148 (0.001)
GBM	0.126 (0.003)	0.084 (0.001)	0.065 (0.001)
ORACLE	0.071 (0.003)	0.042 (0.001)	0.029 (0.000)

Table 4. Average CPU time (in seconds) for each method to compute a model fit (including tuning parameter selection) for the simulations of Example 3.

	$n = 100$	$n = 250$	$n = 500$
ACOSSO-5CV-T	8.7	47.2	155.2
ACOSSO-5CV-C	10.5	49.7	163.6
ACOSSO-BIC-T	3.0	11.9	89.0
ACOSSO-BIC-C	3.9	26.6	118.4
COSSO	7.5	43.4	150.1
MARS	9.2	11.2	13.6
GAM	5.7	7.9	11.3
Random Forest	0.3	6.4	15.3
GBM	4.2	9.1	17.0

GAM did not improve from $n = 250$ to $n = 500$, probably because of its inability to model the interactions in this example. Also notice that the ACOSSO methods maintained close to half the risk of COSSO for all sample sizes. For $n = 500$ the ACOSSO methods had risks nearly the same as that of the ORACLE, and roughly half that of the next best methods (COSSO and GBM).

Computation Time. Table 4 gives the computation times (in seconds) for the various methods used in Example 3. In this example there were 10 predictors, but the total number of functional components, including interactions, was 55. The times given are the average over the 100 realizations, and include the time required for tuning parameter selection.

For larger sample sizes, ACOSSO and COSSO took significantly longer than the other methods. Considering the performance in the simulation examples and the computation times, the best overall ACOSSO variant seems to be ACOSSO-BIC-T. It is important to point out that other methods were computed via more polished R-packages that take advantage of the speed of compiled languages such as C or Fortran. The computing time for ACOSSO (and COSSO) could also be decreased substantially by introducing more efficient approximations, and by taking advantage of a compiled language.

7. Application to Data

We applied the ACOSSO to three datasets. We only report the results of using the two ACOSSO-BIC methods since they performed much better overall than the ACOSSO-5CV methods in our simulations. The Ozone data and Tecator data were also used by Lin and Zhang (2006). They are available from the datasets archive of StatLib at http://lib.stat.cmu.edu/datasets/. The Ozone data has been looked at by Breiman and Friedman (1995), Buja, Hastie, and Tibshirani (1989), and Breiman (1995). The data consists of the daily maximum one-hour-average ozone reading and eight meteorological variables recorded in the Los Angeles basin for 330 days of 1976. The Tecator data was recorded on a Tecator Infratec Food and Feed Analyzer. Each sample contains finely chopped pure meat with different moisture, fat, and protein contents. The input vector consists of a 100 channel spectrum of absorbances, $-\log_{10}$ of the transmittance measured by the spectrometer. As in Lin and Zhang (2006), we used the first 13 principal components to predict fat content. The total sample size is 215.

The third data set comes from a computer model for two phase fluid flow (Vaughn et al. (2000)). Uncertainty/sensitivity analysis of this model was carried out as part of the 1996 compliance certification application for the Waste Isolation Pilot Plant (WIPP) (Helton and Marietta, Editors (2000)). There were 31 uncertain variables that were inputs into the two-phase fluid flow analysis; see Storlie and Helton (2008) for a full description. We considered only a specific scenario which was part of the overall analysis. The variable BRNREPTC10K was used as the response. This variable corresponds to cumulative brine flow in m^3 into the waste repository at 10,000 years, assuming there was a drilling intrusion at 1000 years. The sample size is $n = 300$. This data set is available at http://www.stat.unm.edu/~storlie/acosso/.

We applied each of the methods on these data sets, and estimated the prediction risk, $\mathrm{E}\,[Y - f(\boldsymbol{X})]^2$, by ten-fold cross validation. We selected the tuning parameter using only data within the training set, i.e., a new value of the tuning parameter was selected for each of the ten training sets without using any data from the test sets. The estimate obtained was then evaluated on the test set. We repeated this ten-fold cross validation 50 times and averaged. The resulting prediction risk estimates along with standard errors are displayed in Table 5. The interaction model was used for all of the methods (except GAM) since it had better prediction accuracy on all three data sets than did the additive model.

Table 5. Estimated prediction risk for data examples; standard errors are given in parantheses. Risk for BRNREPTC10K for the WIPP data is in units of 100m^6.

	Ozone	Tecator	WIPP
ACOSSO-BIC-T	15.07 (0.07)	1.44 (0.02)	1.04 (0.00)
ACOSSO-BIC-C	14.81 (0.08)	1.38 (0.02)	1.05 (0.01)
COSSO	15.99 (0.06)	0.88 (0.02)	1.30 (0.01)
MARS	14.24 (0.12)	3.01 (0.17)	1.12 (0.01)
GAM	15.91 (0.12)	592.52 (4.26)	1.83 (0.01)
Random Forest	18.11 (0.07)	14.35 (0.10)	1.29 (0.01)
GBM	10.69 (0.00)	3.35 (0.00)	0.97 (0.00)

For the Ozone data set, the ACOSSO was comparable to MARS but better than COSSO, GAM and Random Forest, while GBM was the best method for prediction accuracy. For the Tecator data, both the COSSO and ACOSSO were much better than all of the other methods. There were several significant interactions so GAM performed poorly here. Interestingly, COSSO is better than ACOSSO here, the adaptive weights aren't always an advantage. In this case the advantage was likely due to the fact that 12 out of the 13 variables were selected for inclusion into the model, and around 62 out of the 91 total functional components were estimated to be non-zero, so this was not a very sparse model. Hence, using all weights equal to 1 (the COSSO) should work quite well here. In cases like this, it may be that using adaptive weights in the ACOSSO can detract from the COSSO fit by adding more noise to the estimation process. In contrast, the WIPP data set had only about 8 informative input variables of the 31 inputs and the ACOSSO significantly outperformed the COSSO and was comparable to GBM for prediction accuracy.

8. Conclusions and Further Work

We have developed the ACOSSO, a regularization method for simultaneous model fitting and variable selection in the context of nonparametric regression. The relationship between the ACOSSO and the COSSO is analogous to that between the adaptive LASSO and the LASSO, and we have explored a special case under which the ACOSSO has a nonparametric version of the oracle property, that which the COSSO does not appear to possess. In addition we have shown that the ACOSSO outperforms COSSO, MARS, and stepwise GAMs for variable selection and prediction in simulations and in some data examples. The ACOSSO also has competitive performance for prediction when compared with Random Forest and GBM. R code to fit ACOSSO models is available at `http://www.stat.unm.edu/~storlie/acosso/`.

It remains to show that ACOSSO has the np-oracle property under more general conditions such as random designs. It may also be possible to yet improve the performance of the ACOSSO by using a different weighting scheme. In addition, there are other ways to use the initial estimate, $\tilde{f}$, in the creation of the penalty term. These are topics for further research.

Acknowledgement

We would like to thank the editorial staff and the reviewers for their valuable comments and suggestions. We would also like to thank Yi Lin for his input. Curtis Storlie's research was partially supported by NSF grant DMS-0354189 and Sandia National Laboratories grant SURP-22858. Howard Bondell research was partially supported by NSF grant DMS-0705968. Brian Reich's research was partially supported by NSF grant DMS-0354189. Hao Helen Zhang's research is partially supported by NSF grant DMS-0645293, NIH/NCI R01 CA-085848, and NIH/NCI P01 CA-142538.

A. Appendix

A.1. Equivalent Form

Proof of Lemma 1. Denote the functional in (3.1) by $A(f)$ and the functional in (5.1) by $B(\boldsymbol{\theta}, f)$. Since $a + b \geq 2\sqrt{ab}$ for a, $b \geq 0$, with equality if and only if $a = b$, we have for each $j = 1, \ldots, q$,

$$\lambda_0 \theta_j^{-1} w_j^{2-\vartheta} \|P^j f\|_{\mathcal{F}}^2 + \lambda_1 w_j^{\vartheta} \theta_j \geq 2\lambda_0^{1/2} \lambda_1^{1/2} w_j \|P^j f\|_{\mathcal{F}} = \lambda w_j \|P^j f\|_{\mathcal{F}}$$

for any $\theta_j \geq 0$ and any $f \in \mathcal{F}$. Hence, $B(\boldsymbol{\theta}, f) \geq A(f)$ with equality only when $\theta_j = \lambda_0^{1/2} \lambda_1^{-1/2} w_j^{1-\vartheta} \|P^j \hat{f}\|_{\mathcal{F}}$, and the result follows.

A.2. Convergence rate

The proof of Theorem 1 uses the next Lemma; it is a generalization of Theorem 10.2 of van de Geer (2000). Consider the regression model $y_i = g_0(\boldsymbol{x}_i) + \varepsilon_i$, $i = 1, \ldots, n$, where g_0 is known to lie in a class of functions $\mathcal{G}$, the $\boldsymbol{x}_i$'s are given covariates in $[0, 1]^p$, and the ε_i's are independent and sub-Gaussian. Let $I_n : \mathcal{G} \to [0, \infty)$ be a pseudonorm on $\mathcal{G}$, and set $\hat{g}_n = \arg\min_{g \in \mathcal{G}} 1/n \sum_{i=1}^n (y_i - g(\boldsymbol{x}_i))^2 + \tau_n^2 I_n(g)$. Let $H_\infty(\delta, \mathcal{G})$ be the δ-entropy of the function class $\mathcal{G}$ under the supremum norm $\|g\|_\infty = \sup_{\boldsymbol{x}} |g(\boldsymbol{x})|$ (see page 17 of van de Geer (2000)).

Lemma 2. *Suppose there exists I_* such that $I_*(g) \leq I_n(g)$ for all $g \in \mathcal{G}$, $n \geq 1$, and suppose there exist constants $A > 0$ and $0 < \alpha < 2$ such that*

$$H_\infty\left(\delta, \left\{\frac{g - g_0}{I_*(g) + I_*(g_0)} : g \in \mathcal{G}, I_*(g) + I_*(g_0) > 0\right\}\right) \leq A\delta^{-\alpha} \qquad \text{(A.1)}$$

for all $\delta > 0$ *and* $n \geq 1$. *Then if* $I_*(g_0) > 0$ *and* $\tau_n^{-1} = O_p\left(n^{1/(2+\alpha)}\right) I_n^{(2-\alpha)/(4+2\alpha)}$ (g_0), *we have* $\|\hat{g}_n - g_0\| = O_p(\tau_n) I_n^{1/2}(g_0)$. *Moreover if* $I_n(g_0) = 0$ *for all* $n \geq 1$, *then* $\|\hat{g}_n - g_0\| = O_p(n^{-1/(2-\alpha)})\tau_n^{-2\alpha/(2-\alpha)}$.

Proof. This proof follows the same logic as the proof of Theorem 10.2 of van de Geer (2000). Notice that

$$\|\hat{g}_n - g_0\|_n^2 + \tau_n^2 I_n(\hat{g}_n) \leq 2(\varepsilon, \hat{g}_n - g_0)_n + \tau_n^2 I_n(g_0). \tag{A.2}$$

Condition (A.1), along with Lemma 8.4 in van de Geer (2000), guarantees that

$$\sup_{g \in \mathcal{G}} \frac{|(\varepsilon, \hat{g}_n - g_0)_n|}{\|\hat{g}_n - g_0\|_n^{1-\alpha/2}(I_*(g) + I_*(g_0))^{\alpha/2}} = O_p(n^{-1/2}). \tag{A.3}$$

Case (i) Suppose that $I_*(\hat{g}_n) > I_*(g_0)$. Then by (A.2) and (A.3) we have

$$\begin{aligned}\|\hat{g}_n - g_0\|_n^2 + \tau_n^2 I_n(\hat{g}_n) &\leq O_p(n^{-1/2})\|\hat{g}_n - g_0\|_n^{1-\alpha/2} I_*^{\alpha/2}(\hat{g}_n) + \tau_n^2 I_n(g_0)\\ &\leq O_p(n^{-1/2})\|\hat{g}_n - g_0\|_n^{1-\alpha/2} I_n^{\alpha/2}(\hat{g}_n) + \tau_n^2 I_n(g_0).\end{aligned}$$

The rest of the argument is identical to that on page 170 of van de Geer (2000). *Case* (ii) Suppose that $I_*(\hat{g}_n) \leq I_*(g_0)$ and $I_*(g_0) > 0$. By (A.2) and (A.3) we have

$$\begin{aligned}\|\hat{g}_n - g_0\|_n^2 &\leq O_p(n^{-1/2})\|\hat{g}_n - g_0\|_n^{1-\alpha/2} I_*^{\alpha/2}(g_0) + \tau_n^2 I_n(g_0)\\ &\leq O_p(n^{-1/2})\|\hat{g}_n - g_0\|_n^{1-\alpha/2} I_n^{\alpha/2}(g_0) + \tau_n^2 I_n(g_0).\end{aligned}$$

The remainder of this case is identical to that on page 170 of van de Geer (2000).

Proof of Theorem 1. The conditions of Lemma 2 do not hold directly for the $\mathcal{F}$ and $I_n(f) = \sum_{j=1}^p w_{j,n}\|P^j f\|_{\mathcal{F}}$ of Theorem 1. An orthogonality argument used in van de Geer (2000) and Lin and Zhang (2006) works to remedy this problem. For any $f \in \mathcal{F}$, we can write $f(\boldsymbol{x}) = b + g(\boldsymbol{x}) = b + f_1(x_1) + \cdots + f_p(x_p)$, such that $\sum_{i=1}^n f_j(x_{j,i}) = 0$, $j = 1, \ldots, p$. Similarly write $\hat{f}(\boldsymbol{x}) = \hat{b} + \hat{g}(\boldsymbol{x})$ and $f_0(\boldsymbol{x}) = b_0 + g_0(\boldsymbol{x})$. Then $\sum_{i=1}^n (g(\boldsymbol{x}_i) - g_0(\boldsymbol{x}_i)) = 0$, and we can write (3.1) as

$$(b_0 - b)^2 + \frac{2}{n}(b_0 - b)\sum_{i=1}^n \varepsilon_i + \frac{1}{n}\sum_{i=1}^n (g_0(\boldsymbol{x}_i) - g(\boldsymbol{x}_i))^2 + \lambda_n \sum_{j=1}^p w_{j,n}\|P^j g\|_{\mathcal{F}}.$$

Therefore $\hat{b}$ must minimize $(b_0 - b)^2 + 2/n(b_0 - b)\sum_{i=1}^n \varepsilon_i$, so that $\hat{b} = b_0 + 1/n\sum_i \varepsilon_i$. Hence, $(\hat{b} - b_0)^2 = O_p(n^{-1})$. On the other hand, $\hat{g}$ must minimize

$$\frac{1}{n}\sum_{i=1}^n (g_0(\boldsymbol{x}_i) - g(\boldsymbol{x}_i))^2 + \lambda_n \sum_{j=1}^p w_{j,n}\|P^j g\|_{\mathcal{F}} \tag{A.4}$$

over all $g \in \mathcal{G}$, where

$$\mathcal{G} = \{g \in \mathcal{F} : g(\boldsymbol{x}) = f_1(x_1) + \cdots + f_p(x_p),\ \sum_{i=1}^{n} f_j(x_{j,i}) = 0, j = 1, \ldots, p\}. \quad \text{(A.5)}$$

Now rewrite (A.4) as

$$\frac{1}{n}\sum_{i=1}^{n}(g_0(\boldsymbol{x}_i) - g(\boldsymbol{x}_i))^2 + \tilde{\lambda}_n \sum_{j=1}^{p} \tilde{w}_{j,n}\|P^j g\|_{\mathcal{F}}, \quad \text{(A.6)}$$

where $\tilde{\lambda}_n = \lambda_n w_{*,n}$, $w_{*,n} = \min\{w_{1,n}, \ldots, w_{p,n}\}$, and $\tilde{w}_{j,n} = w_{j,n}/w_{*,n}$.

The problem is reduced to showing that the conditions of Lemma 2 hold for $\tau_n^2 = \tilde{\lambda}_n$ and $I_n(g) = \sum_{j=1}^{p} \tilde{w}_{j,n}\|P^j g\|_{\mathcal{F}}$. However, notice that $\min\{\tilde{w}_{1,n}, \ldots, \tilde{w}_{p,n}\} = 1$ for all n, and this implies that $I_n(g) \geq I_*(g) = \sum_{j=1}^{p} \|P^j g\|_{\mathcal{F}}$ for all $g \in \mathcal{G}$ and $n \geq 1$. Also notice that the entropy bound in (A.1) holds whenever

$$H_\infty(\delta, \{g \in \mathcal{G} : I_*(g) \leq 1\}) \leq A\delta^{-\alpha}, \quad \text{(A.7)}$$

since $I_*(g - g_0) \leq I_*(g) + I_*(g_0)$ so that the set in brackets in (A.7) contains that in (A.1). Now (A.7) holds by Lemma 4 in the COSSO paper with $\alpha = 1/2$. We complete the proof by separately treating the cases U^c not empty and U^c empty.
Case (i) Suppose that $P^j f \neq 0$ for some j. Then $I_*(g_0) > 0$. Also, $w_{*,n}^{-1} = O_p(1)$ and $w_{j,n} = O_p(1)$ for $j \in U^c$ by assumption. This implies that $\tilde{w}_{j,n} = O_p(1)$ for $j \in U^c$, so that $I_n(g_0) = O_p(1)$. Also $\tilde{\lambda}_n^{-1} = O_p(1)\lambda_n^{-1} = O_p(n^{4/5})$. The result now follows from Lemma 2.
Case (ii) Suppose $P^j f = 0$ for all j. Then $I_n(g_0) = 0$ for all n and the result follows from Lemma 2.

Proof of Corollary 1. For the traditional smoothing spline with $\lambda_0 \sim n^{-4/5}$ it is known (Lin (2000)) that $\|P^j \tilde{f} - P^j f_0\|_{L_2} = O_p(n^{-2/5})$. This implies $|\|P^j \tilde{f}\|_{L_2} - \|P^j f_0\|_{L_2}| \leq O_p(n^{-2/5})$. Hence, $w_{j,n}^{-1} = O_p(1)$ for $j = 1, \ldots, p$ and $w_{j,n} = O_p(1)$ for $j \in U^c$, which also implies $w_{*,n} = O_p(1)$. The conditions of Theorem 1 are now satisfied and we have $\|f - f_n\| = O_p(n^{-2/5})$ if $P^j f \neq 0$ for some j. On the other hand, notice that $w_{j,n}^{-1} = O_p(n^{-2\gamma/5})$ for $j \in U$. Hence, $w_{*,n}^{-1} = O_p(n^{-2\gamma/5})$ whenever $P^j f = 0$ for all j, so that $n^{-2/3}\lambda_n^{-1/3} w_{*,n}^{-1/3} = O_p(n^{-1/2})$ for $\gamma > 3/4$, and the result follows.

A.3. Oracle property

Proof of Theorem 2. Let $\boldsymbol{\Sigma} = \{\bar{K}(x_{1,i}, x_{1,j})\}_{i,j=1}^{m}$ be the $m \times m$ Gram matrix corresponding to the reproducing kernel for $\bar{\mathcal{S}}^2_{\text{per}}$. Also let $\boldsymbol{K}_j$ be the $n \times n$ Gram matrix corresponding to the reproducing kernel for $\bar{\mathcal{S}}^2_{\text{per}}$ on variable x_j,

$j = 1, \ldots, p$. Let $\mathbf{1}_m$ be a vector of m ones. Assuming the observations are arranged with an appropriate permutation, we can write

$$\begin{aligned}
\boldsymbol{K}_1 &= \boldsymbol{\Sigma} \otimes (\mathbf{1}_m \mathbf{1}_m') \otimes \cdots \otimes (\mathbf{1}_m \mathbf{1}_m'), \\
\boldsymbol{K}_2 &= (\mathbf{1}_m \mathbf{1}_m') \otimes \boldsymbol{\Sigma} \otimes \cdots \otimes (\mathbf{1}_m \mathbf{1}_m'), \\
&\vdots \\
\boldsymbol{K}_p &= (\mathbf{1}_m \mathbf{1}_m') \otimes \cdots \otimes (\mathbf{1}_m \mathbf{1}_m') \otimes \boldsymbol{\Sigma},
\end{aligned}$$

where $\otimes$ here stands for the Kronecker product between two matrices.

Straightforward calculation shows that $\boldsymbol{\Sigma}\mathbf{1}_m = 1/(720m^3)\mathbf{1}_m$. Write the eigenvectors of $\boldsymbol{\Sigma}$ as $\{\boldsymbol{v}_1 = \mathbf{1}_m, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_m\}$, and let $\boldsymbol{\Upsilon}$ be the $m \times m$ matrix with these eigenvectors as its columns. The corresponding eigenvalues are $\{m\phi_1, m\phi_2, \ldots, m\phi_m\}$, where $\phi_1 = 1/(720m^4)$ and $\phi_2 \geq \phi_3 \geq \cdots \geq \phi_m$. It is known (Uteras (1983)) that $\phi_i \sim i^{-4}$ for $i \geq 2$. Notice that $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_m$ are also the eigenvectors of $(\mathbf{1}_m \mathbf{1}_m')$ with eigenvalues $m, 0, \ldots, 0$, respectively. Write $\boldsymbol{O} = \boldsymbol{\Upsilon} \otimes \boldsymbol{\Upsilon} \otimes \cdots \otimes \boldsymbol{\Upsilon}$ and let $\boldsymbol{\xi}_i$ be the i^{th} column of $\boldsymbol{O}$, $i = 1, \ldots, n$. It is easy to verify that $\{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n\}$ forms an eigensystem for each of $\boldsymbol{K}_1, \ldots, \boldsymbol{K}_p$.

Let $\{\boldsymbol{\zeta}_{1,j}, \ldots, \boldsymbol{\zeta}_{n,j}\}$ be the collection of vectors $\{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n\}$ sorted so that those corresponding to nonzero eigenvalues for $\boldsymbol{K}_j$ are listed first. Specifically, let

$$\begin{aligned}
\boldsymbol{\zeta}_{i,1} &= \boldsymbol{v}_i \otimes \mathbf{1}_m \otimes \cdots \otimes \mathbf{1}_m, \\
\boldsymbol{\zeta}_{i,2} &= \mathbf{1}_m \otimes \boldsymbol{v}_i \otimes \cdots \otimes \mathbf{1}_m, \\
&\vdots \\
\boldsymbol{\zeta}_{i,p} &= \mathbf{1}_m \otimes \mathbf{1}_m \otimes \cdots \otimes \boldsymbol{v}_i,
\end{aligned} \tag{A.8}$$

for $i = 1, \ldots, m$. Notice that each $\boldsymbol{\zeta}_{i,j}$, $i = 1, \ldots, m$, $j = 1, \ldots, p$, corresponds to a distinct $\boldsymbol{\xi}_k$, for some $k \in \{1, \ldots, n\}$. Let the first m elements of the collection $\{\boldsymbol{\zeta}_{1,j}, \ldots, \boldsymbol{\zeta}_{m,j}, \boldsymbol{\zeta}_{m+1,j}, \ldots, \boldsymbol{\zeta}_{n,j}\}$ be given by (A.8) and the remaining $n - m$ be given by the remaining $\boldsymbol{\xi}_i$ in any order. The corresponding eigenvalues are then

$$\eta_{i,j} = \begin{cases} n\phi_i & \text{for } i = 1, \ldots, m, \\ 0 & \text{for } i = m+1, \ldots, n. \end{cases}$$

It is clear that $\{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n\}$ is also an orthonormal basis in $\Re^n$ with respect to the inner product

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle_n = \frac{1}{n} \sum_i u_i v_i. \tag{A.9}$$

Let $\boldsymbol{f} = (f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n))'$, $\boldsymbol{a} = (1/n)\boldsymbol{O}'\boldsymbol{f}$, and $\boldsymbol{z} = (1/n)\boldsymbol{O}'\boldsymbol{y}$. That is, $z_i = \langle \boldsymbol{y}, \boldsymbol{\xi}_i \rangle_n$, $a_i = \langle \boldsymbol{f}, \boldsymbol{\xi}_i \rangle_n$, $\delta_i = \langle \boldsymbol{\varepsilon}, \boldsymbol{\xi}_i \rangle_n$, and we have that $z_i = a_i + \delta_i$. With some abuse of notation, also let $z_{i,j} = \left\langle \boldsymbol{y}, \boldsymbol{\zeta}_{i,j} \right\rangle_n$, $a_{i,j} = \left\langle \boldsymbol{f}, \boldsymbol{\zeta}_{i,j} \right\rangle_n$, and $\delta_{i,j} = \left\langle \boldsymbol{\varepsilon}, \boldsymbol{\zeta}_{i,j} \right\rangle_n$. Now, using $\vartheta = 2$ in (5.1), the ACOSSO estimate is the minimizer of

$$\frac{1}{n}(\boldsymbol{y} - \boldsymbol{K}_\theta \boldsymbol{c} - b\boldsymbol{1}_n)'(\boldsymbol{y} - \boldsymbol{K}_\theta \boldsymbol{c} - b\boldsymbol{1}_n) + \boldsymbol{c}'\boldsymbol{K}_\theta \boldsymbol{c} + \lambda_1 \sum_{j=1}^{p} w_j^2 \theta_j, \tag{A.10}$$

where $\boldsymbol{K}_\theta = \sum_{j=1}^{p} \theta_j \boldsymbol{K}_j$. Let $\boldsymbol{s} = \boldsymbol{O}'\boldsymbol{c}$ and $\boldsymbol{D}_j = (1/n^2)\boldsymbol{O}'\boldsymbol{K}_j\boldsymbol{O}$ be the diagonal matrix with diagonal elements ϕ_i. Then (A.10) is equivalent to

$$\left(\boldsymbol{z} - \boldsymbol{D}_\theta \boldsymbol{s} - (b, 0, \ldots, 0)'\right)' \left(\boldsymbol{z} - \boldsymbol{D}_\theta \boldsymbol{s} - (b, 0, \ldots, 0)'\right) + \boldsymbol{s}'\boldsymbol{D}_\theta \boldsymbol{s} + \lambda_1 \sum_{j=1}^{p} w_j^2 \theta_j, \tag{A.11}$$

where $\boldsymbol{D}_\theta = \sum_{j=1}^{p} \theta_j \boldsymbol{D}_j$. Straightforward calculation shows that this minimization problem is equivalent to

$$\ell(\boldsymbol{s}, \boldsymbol{\theta}) = \sum_{i=1}^{m} \sum_{j=1}^{p} (z_{ij} - \phi_i \theta_j s_{ij})^2 + \sum_{i=1}^{m} \sum_{j=1}^{p} \phi_i \theta_j s_{ij}^2 + \lambda_1 \sum_{j=1}^{p} w_j^2 \theta_j, \tag{A.12}$$

where $s_{ij} = \boldsymbol{\zeta}_{ij}'\boldsymbol{c}$, $i = 1, \ldots, m$, $j = 1, \ldots, p$, are distinct elements of $\boldsymbol{s}$.

We condition on $\boldsymbol{\theta}$ and then minimize over $\boldsymbol{s}$. Given $\boldsymbol{\theta}$, $\ell(\boldsymbol{s}, \boldsymbol{\theta})$ is a convex function of $\boldsymbol{s}$ and is minimized at $\hat{\boldsymbol{s}}(\boldsymbol{\theta}) = \{\hat{s}_{ij}(\theta_j)\}_{i=1}^{m}{}_{j=1}^{p}$, where $\hat{s}_{ij}(\theta_j) = z_{ij}(1 - \phi_i \theta_j)$. Inserting $\hat{\boldsymbol{s}}(\boldsymbol{\theta})$ into (A.12) gives

$$\begin{aligned} \ell(\hat{\boldsymbol{s}}(\boldsymbol{\theta}), \boldsymbol{\theta}) &= \sum_{i=1}^{m} \sum_{j=1}^{p} \frac{z_{ij}^2}{(1 + \phi_i \theta_j)^2} + \sum_{i=1}^{m} \sum_{j=1}^{p} \frac{\phi_i \theta_j z_{ij}^2}{(1 + \phi_i \theta_j)^2} + \lambda_1 \sum_{j=1}^{p} w_j^2 \theta_j \\ &= \sum_{i=1}^{m} \sum_{j=1}^{p} \frac{z_{ij}^2}{1 + \phi_i \theta_j} + \lambda_1 \sum_{j=1}^{p} w_j^2 \theta_j. \end{aligned} \tag{A.13}$$

Notice that $\ell(\hat{\boldsymbol{s}}(\boldsymbol{\theta}), \boldsymbol{\theta})$ is continuous in θ_j,

$$\frac{\partial^2 \ell(\hat{\boldsymbol{s}}(\boldsymbol{\theta}), \boldsymbol{\theta})}{\partial \theta_j^2} = 2 \sum_{i=1}^{m} \frac{z_{ij}^2 \phi_i^2}{(1 + \phi_i \theta_j)^3} > 0 \quad \text{for each } j, \tag{A.14}$$

and $\partial^2 \ell(\hat{\boldsymbol{s}}(\boldsymbol{\theta}), \boldsymbol{\theta}) / \partial \theta_j \partial \theta_k = 0$ for $j \neq k$. Therefore $\ell(\boldsymbol{s}, \boldsymbol{\theta})$ is convex and has a unique minimum, $\hat{\boldsymbol{\theta}}$.

Clearly, $P^j \hat{f} \equiv 0$ if and only if $\hat{\theta}_j = 0$. So it suffices to consider $\hat{\theta}_j$. As such, since we must have $\theta_j \geq 0$, the minimizer, $\hat{\theta}_j = 0$ if and only if

$$\left. \frac{\partial}{\partial \theta_j} \ell\left(\hat{\boldsymbol{s}}(\boldsymbol{\theta}), \boldsymbol{\theta}\right) \right|_{\theta_j = 0} \geq 0,$$

which is equivalent to

$$T = n\sum_{i=1}^{m}\phi_i z_{ij}^2 \le n w_{j,n}^2 \lambda_{1,n}. \tag{A.15}$$

If we assume that $P^j f = 0$, then we have $z_{ij} = \delta_{ij}$. We obtain bounds for $\mathrm{E}\,(T)$ and $\mathrm{Var}\,(T)$ to demonstrate that T is bounded in probability when $P^j f = 0$. To this end, we first obtain bounds for $\mathrm{E}\,(\delta_{ij}^2)$ and $\mathrm{Var}\,(\delta_{ij}^2)$. Recall that $\delta_{ij} = (1/n)\boldsymbol{\zeta}_{ij}'\boldsymbol{\varepsilon}$ and that the individual elements of $\boldsymbol{\varepsilon}$ are independent with $\mathrm{E}\,(\boldsymbol{\varepsilon}) = \mathbf{0}$. For notational convenience let $\boldsymbol{\xi} = \boldsymbol{\zeta}_{ij}$, some column of the $\boldsymbol{O}$ matrix and the $\boldsymbol{\xi}$'s are orthonormal with respect to the inner product in (A.9). Now,

$$\begin{aligned}\mathrm{E}\,(\delta_{ij}^2) &= \frac{1}{n^2}\mathrm{E}\,[(\boldsymbol{\xi}'\boldsymbol{\varepsilon})^2] = \frac{1}{n^2}\mathrm{E}\left(\sum_{a=1}^{n}\sum_{b=1}^{n}\xi_a\xi_b\varepsilon_a\varepsilon_b\right)\\ &= \frac{1}{n^2}\sum_{a=1}^{n}\xi_a^2\mathrm{E}\,(\varepsilon_a^2) \le \frac{1}{n^2}\sum_{a=1}^{n}\xi_a^2 M_1 = \frac{M_1}{n},\end{aligned} \tag{A.16}$$

where $M_1 = \max_a \mathrm{E}\,(\varepsilon_a^2)$, bounded because of the sub-Gaussian condition (4.1).

The variance of δ_{ij}^2 is

$$\begin{aligned}\mathrm{Var}\,(\delta_{ij}^2) &= \mathrm{Var}\left(\sum_{a=1}^{n}\sum_{b=1}^{n}\xi_a\xi_b\varepsilon_a\varepsilon_b\right)\\ &= \sum_{a=1}^{n}\sum_{b=1}^{n}\sum_{c=1}^{n}\sum_{d=1}^{n}\xi_a\xi_b\xi_c\xi_d\,\mathrm{Cov}\,(\varepsilon_a\varepsilon_b\,,\,\varepsilon_c\varepsilon_d).\end{aligned} \tag{A.17}$$

The ε_a's are independent, so $\mathrm{Cov}\,(\varepsilon_a\varepsilon_b\,,\,\varepsilon_c\varepsilon_d) \ne 0$ only in the three mutually exclusive cases (i) $a = b = c = d$, (ii) $a = c$ and $b = d$ with $a \ne b$, or (iii) $a = d$ and $b = c$ with $a \ne b$. Thus, (A.17) becomes,

$$\begin{aligned}\mathrm{Var}\,(\delta_{ij}^2) &= \frac{1}{n^4}\left[\sum_{a=1}^{n}\xi_a^4\,\mathrm{Cov}\,(\varepsilon_a^2\,,\,\varepsilon_a^2) + 2\sum_{a=1}^{n}\sum_{b\ne a}^{n}\xi_a^2\xi_b^2\,\mathrm{Cov}\,(\varepsilon_a\varepsilon_b\,,\,\varepsilon_a\varepsilon_b)\right]\\ &\le \frac{2}{n^4}\sum_{a=1}^{n}\sum_{b=1}^{n}\xi_a^2\xi_b^2\mathrm{Var}\,(\varepsilon_a\varepsilon_b) \le \frac{2M_2}{n^4}\left(\sum_{a=1}^{n}\xi_a^2\right)^2 = \frac{2M_2}{n^2},\end{aligned} \tag{A.18}$$

where $M_2 = \max_{a,b}\{\mathrm{Var}\,(\varepsilon_a\varepsilon_b)\}$, bounded because of the sub-Gaussian condition in (4.1). Notice that the derivations of the bounds in (A.16) and (A.18) do not depend on i or j, so the bounds in (A.16) and (A.18) are uniform for all i.

Using (A.16) we can write $\mathrm{E}\,(T)$ as

$$\mathrm{E}\,(T) = n\sum_{i=1}^{m}\phi_i\mathrm{E}\,[\delta_{ij}^2] \le M_1\sum_{i=1}^{m}\phi_i \sim M_1. \tag{A.19}$$

Further, we can use (A.18) to write Var (T) as

$$\text{Var}\,(T) = n^2\text{Var}\left(\sum_{i=1}^{m}\phi_i\delta_{ij}\right) = n^2\sum_{k=1}^{m}\sum_{l=1}^{m}\phi_k\phi_l\,\text{Cov}\,(\delta_{k,j}^2,\delta_{l,j}^2)$$
$$\leq 2M_2\sum_{k=1}^{m}\sum_{l=1}^{m}\phi_k\phi_l = 2M_2\left(\sum_{k=1}^{m}\phi_k\right)^2 \sim 2M_2. \tag{A.20}$$

Finally, as n increases, (A.19) and (A.20) guarantee that the left side of (A.15) is bounded in probability when $P^jf = 0$. Assuming that $nw_{j,n}^2\lambda_n^2 \xrightarrow{p} \infty$, or equivalently that $nw_{j,n}^2\lambda_{1,n} \xrightarrow{p} \infty$ by Lemma 1, the right side of (A.15) increases to ∞ in probability. Therefore, if $P^jf = 0$ then $\hat{\theta}_j = 0$ with probability tending to one. If, on the other hand, $nw_{j,n}^2\lambda_n^2 = O_p(1)$, then the probability that $T > nw_{j,n}^2\lambda_{1,n}$ converges to a positive constant. Hence the probability that $\hat{\theta}_j > 0$ converges to a positive constant.

Proof of Corollary 2. It is straightforward to to show that Theorem 1 still holds with $\mathcal{S}_{\text{per}}^2$ in place of $\mathcal{S}^2$. From the proof of Corollary 1, these weights also satisfy the conditions of Theorem 1. Since $w_{j,n}^{-1} = O_p(n^{-2\gamma/5})$ for $j \in U$, we have $nw_{j,n}^2\lambda_n^2 \xrightarrow{p} \infty$ for $j \in U$ whenever $\gamma > 3/4$. The conditions of Theorem 2 are now satisfied. In light of Theorem 1, we also know that if $P^jf \neq 0$, the probability that $P^j\hat{f} \neq 0$ also tends to one as the sample size increases due to consistency. Corollary 2 follows.

References

Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics.* Kluwer Academic Publishers, Norwell, MA.

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

Breiman, L. (2001). Random forests. *Machine Learning* **45**, 5–32.

Breiman, L. and Friedman, J. (1995). Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc.* **80**, 580–598.

Buja, A., Hastie, T. and Tibshirani, R. (1989). Linear smoothers and additive models (with discussion). *Ann. Statist.* **17**, 453-555.

Cleveland, W. (1979). Robust locally weighted fitting and smoothing scatterplots. *J. Amer. Statist. Assoc.* **74**, 829-836.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerical Math.* **31**, 377-403.

Efromovich, S. and Samarov, A. (2000). Adaptive estimation of the integral of squared regression derivatives. *Scand. J. Statist.* **27**, 335-351.

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

Friedman, J. (1991). Multivariate adaptive regression splines (with discussion). *Ann. Statist.* **19**, 1-141.

Friedman, J. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statist.* **29**, 1189-1232.

Goldfarb, D. and Idnani, A. (1982). Dual and primal-dual methods for solving strictly convex quadratic programs. In *Numerical Analysis* (Edited by J. P. Hennart). Springer-Verlag, Berlin.

Gu, C. (2002). *Smoothing Spline ANOVA Models.* Springer-Verlag, New York.

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

Helton, J. and Marietta, Editors, M. (2000). Special issue: The 1996 performance assessment for the Waste Isolation Pilot Plant. *Reliability Engineering and System Safety* **69**, 1-451.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.* **28**, 1356-1378.

Lin, Y. (2000). Tensor product space anova models. *Ann. Statist.* **28**, 734-755.

Lin, Y. and Zhang, H. (2006). Component selection and smoothing in smoothing spline analysis of variance models. *Ann. Statist.* **34**, 2272-2297.

Nadaraya, E. (1964). On estimating regression. *Theory Probab. Appl.* **9**, 141–142.

Schimek, M., ed. (2000). *Smoothing and Regression: Approaches, Computation, and Application.* John Wiley, New York.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.

Stone, C., Buja, A. and Hastie, T. (1994). The use of polynomial splines and their tensor-products in multivariate function estimation. *Ann. Statist.* **22**, 118-184.

Stone, C., Hansen, M., Kooperberg, C. and Truong, Y. (1997). Polynomial splines and their tensor products in extended linear modeling. *Ann. Statist.* **25**, 1371-1425.

Storlie, C. and Helton, J. (2008). Multiple predictor smoothing methods for sensitivity analysis: Example results. *Reliability Engineering and System Safety* **93**, 55-77.

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

Uteras, F. (1983). Natural spline functions: Their associated eigenvalue problem. *Numer. Math.* **42**, 107-117.

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

Vaughn, P., Bean, J., Helton, J., Lord, M., MacKinnon, R. and Schreiber, J. (2000). Representation of two-phase flow in the vicinity of the repository in the 1996 performance assessment for the Waste Isolation Pilot Plant. *Reliability Engineering and System Safety* **69**, 205-226.

Wahba, G. (1990). *Spline Models for Observational Data.* CBMS-NSF Regional Conference Series in Applied Mathematics.

Wahba, G., Wang, Y., Gu, C., Klein, R. and Klein, B. (1995). Smoothing spline anova for exponential families, with application to the WESDR. *Ann. Statist.* **23**, 1865-1895.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418-1429.

Zou, H., Hastie, T. and Tibshirani, R. (2007). On the "degrees of freedom" of the lasso. *Ann. Statist.* **35**, 2173-2192.

Department of Mathematics and Statistics, MSC03 2150, 1 University of New Mexico, Albuquerque, New Mexico, 87131-0001 (505) 331-9059, U.S.A.

E-mail: storlie@stat.unm.edu

Department of Statistics, North Carolina State University, 2311 Stinson Drive, Campus Box 8203, Raleigh, NC 27695-8203, U.S.A.

E-mail: bondell@stat.ncsu.edu

Department of Statistics, North Carolina State University, 2311 Stinson Drive, Campus Box 8203, Raleigh, NC 27695-8203, U.S.A.

E-mail: brian_reich@ncsu.edu

Department of Statistics, North Carolina State University, 2311 Stinson Drive, Campus Box 8203, Raleigh, NC 27695-8203, U.S.A.

E-mail: hzhang@stat.ncsu.edu

(Received September 2008; accepted October 2009)