Component Selection and Smoothing in Smoothing Spline Analysis of Variance Models

By Yi Lin and Hao Helen Zhang

University of Wisconsin - Madison and North Carolina State University

Abstract

We propose a new method for model selection and model fitting in nonparametric regression models, in the framework of smoothing spline ANOVA. The "COSSO" is a method of regularization with the penalty functional being the sum of component norms, instead of the squared norm employed in the traditional smoothing spline method. The COSSO provides a unified framework for several recent proposals for model selection in linear models and smoothing spline ANOVA models. Theoretical properties, such as the existence and the rate of convergence of the COSSO estimator, are studied. In the special case of a tensor product design with periodic functions, a detailed analysis reveals that the COSSO applies a novel soft thresholding type operation to the function components and selects the correct model structure with probability tending to one. We give an equivalent formulation of the COSSO estimator which leads naturally to an iterative algorithm. We compare the COSSO with the MARS, a popular method that builds functional ANOVA models, in simulations and real examples. The COSSO gives very competitive performances in these studies.

Key words and phrases: LASSO, method of regularization, model selection, nonnegative garrote, nonparametric regression, variable selection.

1 Introduction

Consider the regression problem $y_i = f(x_i) + \epsilon_i, \quad i = 1, ..., n$, where f is the unknown regression function to be estimated, $x_i = (x_i^{(1)}, ..., x_i^{(d)})$'s are d dimensional vectors of covariates, and the ϵ's are independent noises with mean 0 and variance σ^2. Usually the estimator is judged in terms of prediction accuracy and interpretability.

In linear regression models it is assumed that $f(x) = \beta_0 + \sum_{j=1}^d \beta_j x^{(j)}$. Traditional approaches to variable selection include the best subset selection and the forward/backward stepwise selection. As pointed out by Breiman (1995), these methods suffer from instability and relative lack of accuracy. Several new and effective methods for variable selection in linear models have been proposed in recent years [Breiman (1995); Tibshirani (1996); Frank and Friedman (1993); Fan and Li (2001)]. Two methods, the nonnegative garrote by Breiman (1995) and the LASSO by Tibshirani (1996), are closely related to the method in our paper, and are reviewed in the following.

Assume that the $x_i^{(j)}$ are standardized so that $\sum_i x_i^{(j)}/n = 0$ and $\sum_i \{x_i^{(j)}\}^2/n = 1$. Let $\hat{\beta}^o = (\hat{\beta}_0^o, ..., \hat{\beta}_d^o)$ be the ordinary least square estimates. The nonnegative garrote solution is $(\hat\beta_0^o, r_1\hat\beta_1^o, ..., r_d\hat\beta_d^o)$, where $(r_1, ..., r_d)$ is the solution to

$$\min_{r_1,...,r_d} \sum_{i=1}^n \{y_i - \hat\beta_0^o - \sum_{j=1}^d r_j \hat\beta_j^o x_i^{(j)}\}^2, \quad \text{subject to} \quad r_j \ge 0, j = 1, ..., d, \text{ and } \sum_{j=1}^d r_j \le t.$$

Here $t \ge 0$ is a tuning parameter. The nonnegative garrote selects subset and shrinks the estimate at the same time. Breiman (1995) showed that the nonnegative garrote has consistently lower prediction error than subset selection with extensive simulation studies.

The Least Absolute Shrinkage and Selection Operator (LASSO) estimate $\hat\beta = (\hat\beta_0, ..., \hat\beta_d)$ is the minimizer of

$$\frac{1}{n}\sum_{i=1}^n \{y_i - \beta_0 - \sum_{j=1}^d \beta_j x_i^{(j)}\}^2 \quad \text{subject to} \quad \sum_{j=1}^d |\beta_j| \le t,$$

or equivalently, the minimizer of

$$\frac{1}{n}\sum_{i=1}^n \{y_i - \beta_0 - \sum_{j=1}^d \beta_j x_i^{(j)}\}^2 + \lambda \sum_{j=1}^d |\beta_j|,$$

where t or λ are tuning parameters. The LASSO is a penalized least squares method with the L_1 penalty on the coefficients. Tibshirani (1996) proposed and studied the LASSO. Bakin (1999) contained some useful developments. Knight and Fu (2000) studied the asymptotic properties of the LASSO.

We consider model selection in a more general nonparametric setting, within the smoothing spline analysis of variance (SS-ANOVA) framework [Wahba (1990); Wahba, Wang, Gu, Klein and Klein (1995); Gu (2002)]. In the SS-ANOVA we write

$$f(x) = b + \sum_{j=1}^d f_j(x^{(j)}) + \sum_{j<k} f_{jk}(x^{(j)}, x^{(k)}) + \cdots, \tag{1}$$

where b is a constant, f_j's are the main effects, f_{jk}'s are the two way interactions, and so on. The sequence is usually truncated somewhere to enhance interpretability. The identifiability of the terms in (1) is assured by side conditions through averaging operators. The SS-ANOVA generalizes the popular additive model by Hastie and Tibshirani (1990) and provides a general framework for nonparametric multivariate function estimation. One important question in the application of SS-ANOVA is to determine which variables or which ANOVA components should be included in the model. Gu (1992) proposed using cosine diagnostics as model checking tools after model fitting in Gaussian regression. Yau, Kohn and Wood (2002) presented a Bayesian method for variable selection in a nonparametric manner. Gunn and Kandola (2002) proposed a sparse kernel approach in a closely related framework. Zhang, Wahba, Lin, Voelker, Ferris, Klein and Klein (2002) gave a possible generalization of the LASSO to the SS-ANOVA for exponential families. They expanded each nonparametric component function as a linear combination of a large number of basis functions, and applied L_1 penalty to the coefficients of all the basis functions. The L_1 penalty gives a solution that is sparse in the coefficients. However, a separate model selection procedure has to be applied after model fitting, since sparsity in coefficients helps but does not guarantee the sparsity in SS-ANOVA components.

In this paper we consider a different approach for model selection and model fitting in the SS-ANOVA. This is a method of regularization with the penalty functional being the sum of component norms. We show that the approach provides a unified framework for several recent proposals for model selection in linear models and SS-ANOVA models. The new method reduces to the LASSO in linear models, and thus gives an alternative interpretation of the penalty term in the LASSO to being the L_1 norm of the coefficients: it is the sum of component norms. Our method will be referred to as the COmponent Selection and Smoothing Operator (COSSO). The general methodology is introduced in Section 2, where we also prove the existence of the COSSO estimate and give some rate of convergence results. In Section 3 we obtain an alternative formulation of the COSSO that is more suitable for computation. In Section 4, we consider the special case of a tensor product design with periodic functions. A detailed analysis in this special case sheds light on the mechanism of the COSSO in terms of component selection in the SS-ANOVA. In particular, we show in this case, that the COSSO applies a novel soft thresholding type operation to the function components and selects the correct model structure with probability tending to one. In Section 5, we present a COSSO algorithm that is based on iterating between the smoothing spline method and the nonnegative garrote. In Section 6, we consider the choice of the tuning parameter. Simulations are given in Section 7, where we compare the COSSO with the MARS developed by Friedman (1991), a popular algorithm that builds functional ANOVA models. Some real examples are given in Section 8, and Section 9 contains a discussion. The proofs are given in the Appendices.

2 The COSSO in smoothing spline ANOVA

2.1 The smoothing spline ANOVA

In the commonly used smoothing spline ANOVA model over $\mathcal{X} = [0,1]^d$, it is assumed that $f \in \mathcal{F}$, where $\mathcal{F}$ is a reproducing kernel Hilbert space (RKHS) corresponding to the decomposition (1). Let H^j be a function space of functions of $x^{(j)}$ over $[0,1]$ such that

$H^j = \{1\} \oplus \bar{H}^j$. Then the tensor product space of the H^j's is

$$\otimes_{j=1}^d H^j = \{1\} \oplus \sum_{j=1}^d \bar{H}^j \oplus \sum_{j<k} [\bar{H}^j \otimes \bar{H}^k] \oplus \cdots . \quad (2)$$

Each functional component in the SS-ANOVA decomposition (1) lies in a subspace in the orthogonal decomposition (2) of $\otimes_{j=1}^d H^j$. Typically only low order interactions are considered in the SS-ANOVA model for interpretability and visualization. The popular additive model is a special case in which $f(x^{(1)}, ..., x^{(d)}) = b + \sum_{j=1}^d f_j(x^{(j)})$, with $f_j \in \bar{H}^j$. In this case the selection of functional components is equivalent to variable selection. In more complex SS-ANOVA models model selection amounts to the selection of main effects and interaction terms in the SS-ANOVA decomposition. The interaction terms reside in the tensor product spaces of univariate function spaces. The reproducing kernel of a tensor product space is simply the product of the reproducing kernels of the individual spaces. This greatly facilitates the use of smoothing spline type method in such models.

A common example of the function space H^j of univariate functions, which is also used in this paper, is the second order Sobolev Hilbert space: $S = \{g : g, g'$ are absolutely continuous, $g'' \in \mathcal{L}_2[0, 1]\}$. When endowed with the norm

$$\|g\|^2 = \{\int_0^1 g(t)\ dt\}^2 + \{\int_0^1 g'(t)\ dt\}^2 + \int_0^1 \{g''(t)\}^2\ dt,$$

S can be decomposed as $S = \{1\} \oplus \bar{S}$, where $\bar{S}$ is a RKHS with the reproducing kernel $\bar{K}(s, t) = k_1(s)k_1(t) + k_2(s)k_2(t) - k_4(|s-t|)$, with $k_1(t) = t - 1/2$, $k_2(t) = \{k_1^2(t) - 1/12\}/2$, and $k_4(t) = \{k_1^4(t) - k_1^2(t)/2 + 7/240\}/24$. See Wahba (1990), Gu (2002).

2.2 The COSSO

In general, the function space in SS-ANOVA can be written as

$$\mathcal{F} = \{1\} \oplus \mathcal{F}_1, \quad \text{with} \quad \mathcal{F}_1 = \bigoplus_{\alpha=1}^p \mathcal{F}^\alpha, \quad (3)$$

where $\mathcal{F}^1$, ..., $\mathcal{F}^p$ are p orthogonal subspaces of $\mathcal{F}$. In the additive model $p = d$ and $\mathcal{F}^\alpha$'s are the main effect spaces. In the two way interaction model there are d main effect spaces and $d(d-1)/2$ two way interaction spaces, thus $p = d(d+1)/2$. We may further decompose the functional components into parametric and nonparametric parts, as is commonly done with the smoothing spline method. We do not pursue this in this paper as our emphasis is on the selection of functional components in the SS-ANOVA. However, the general idea of our procedure can still be applied with this further decomposition, and it may be helpful to select parametric and nonparametric components of the variables.

Denote the norm in the RKHS $\mathcal{F}$ by $\|\cdot\|$. A traditional smoothing spline type method finds $f \in \mathcal{F}$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \lambda \sum_{\alpha=1}^{p} \theta_\alpha^{-1} \|P^\alpha f\|^2, \tag{4}$$

where $P^\alpha f$ is the orthogonal projection of f onto $\mathcal{F}^\alpha$ and $\theta_\alpha \geq 0$. If $\theta_\alpha = 0$, then the minimizer is taken to satisfy $\|P^\alpha f\|^2 = 0$. We use the convention $0/0 = 0$ throughout this paper. The smoothing parameter λ is confounded with the θ's, but is usually included in the setup for computational purpose.

We propose the COSSO procedure that finds $f \in \mathcal{F}$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \tau_n^2 J(f), \quad \text{with} \quad J(f) = \sum_{\alpha=1}^{p} \|P^\alpha f\|, \tag{5}$$

where τ_n is a smoothing parameter. We sometimes suppress the dependence of τ on n in our notation. The penalty term $J(f)$ in the COSSO is a sum of RKHS norms, instead of the squared RKHS norm penalty employed in the smoothing spline. The penalty $J(f)$ is not a norm in $\mathcal{F}$. However, it is a pseudo-norm in the sense: for any f, g in $\mathcal{F}, J(f) \geq 0, J(cf) = |c|J(f), J(f+g) \leq J(f) + J(g)$; for any nonconstant f in $\mathcal{F}, J(f) > 0$. And we have that

$$\sum_{\alpha=1}^{p} \|P^\alpha f\|^2 \leq J^2(f) \leq p \sum_{\alpha=1}^{p} \|P^\alpha f\|^2. \tag{6}$$

Another difference between the COSSO and the common smoothing spline method is that there is only one smoothing parameter τ in the COSSO procedure (5), while in the common smoothing spline approach (4) there are multiple smoothing parameters.

The LASSO in linear models can be seen as a special case of the COSSO. For the input space $\mathcal{X} = [0,1]^d$, consider the linear function space $\mathcal{F} = \{1\} \oplus \{x^{(1)} - 1/2\} \oplus ... \oplus \{x^{(d)} - 1/2\}$, with the usual L_2 inner product on $\mathcal{F}$: $(f, g) = \int_{\mathcal{X}} fg$. The penalty term in the COSSO becomes $J(f) = (12)^{-1/2} \sum_{j=1}^{d} |\beta_j|$ for $f(x) = \beta_0 + \sum_{j=1}^{d} \beta_j x^{(j)}$. This is equivalent to the L_1 norm on the linear coefficients, leading to the LASSO estimator. Notice, however, we interpret the penalty as the sum of the norms of the function components, rather than the L_1 norm of the coefficients.

2.3 Existence of solution to the COSSO

The existence of the COSSO estimate is guaranteed by the following theorem.

THEOREM 1 *Let $\mathcal{F}$ be a RKHS of functions over an input space $\mathcal{X}$. Assume that $\mathcal{F}$ can be decomposed as in (3). Then there exists a minimizer of (5) in $\mathcal{F}$.*

It is of interest to characterize the conditions under which the solution to (5) is unique. It seems the uniqueness should follow under mild conditions on the design matrix of the input variables. We do not pursue this question here, and the development in this paper does not depend on the uniqueness of the COSSO estimate.

2.4 Asymptotic properties of the COSSO

In this section we assume a fixed design. Define $y = (y_1, ..., y_n)^{\mathrm{T}}$. With a little abuse of notations, let f stand for both the regression function and its functional values at data points, i.e., $f = (f(x_1), ..., f(x_n))^{\mathrm{T}}$. Define the norm $\|\cdot\|_n$ and inner product $\langle \cdot, \cdot \rangle_n$ in R^n as

$$\|f\|_n^2 = \frac{1}{n}\sum_{i=1}^n f^2(x_i), \quad \langle f, g\rangle_n = \frac{1}{n}\sum_{i=1}^n f(x_i)g(x_i);$$

then $\|y - f\|_n^2 = 1/n\sum_{i=1}^n \{y_i - f(x_i)\}^2$. The following theorem shows that the COSSO estimator in the additive model has a rate of convergence $n^{-2/5}$ if the tuning parameter is chosen appropriately.

THEOREM 2 *Consider the regression model $y_i = f_0(x_i) + \epsilon_i$, $i = 1, ..., n$, where x_i's are given covariates in $[0,1]^d$, and ϵ_i's are independent $N(0, \sigma^2)$ noises. Assume f_0 lies in $\mathcal{F} = \{1\} \oplus \mathcal{F}_1$, $\mathcal{F}_1 = \bigoplus_{j=1}^d \bar{S}^j$, with $S^j = \{1\} \oplus \bar{S}^j$ being the second order Sobolev space. Consider the COSSO estimate $\hat{f}$ as defined in (5). Then (i) if f_0 is not a constant, and $\tau_n^{-1} = O_p(n^{2/5})J^{3/10}(f_0)$, we have $\|\hat{f} - f_0\|_n = O_p(\tau_n)J^{1/2}(f_0)$; (ii) if f_0 is a constant, we have $\|\hat{f} - f_0\|_n = O_p(\max\{(n\tau_n)^{-2/3}, n^{-1/2}\})$.*

Results for more general ANOVA models can be obtained in a similar way. For example, since the tensor product space of two second order Sobolev spaces of univariate functions is a subspace of the second order Sobolev space of bivariate functions, we can obtain that the COSSO estimator in the two way interaction model has a rate of convergence that is at least as fast as $n^{-1/3}$. This rate is not the optimal rate for SS-ANOVA models. See Lin (2000). However, to prove better rates for the COSSO estimator, involved entropy calculations for tensor product spaces will be required.

3 An equivalent formulation

It can be shown that the solution to (5) is in a finite dimensional space, therefore the COSSO estimate can be computed directly from (5).

LEMMA 1 *Let $\hat{f} = \hat{b} + \sum_{\alpha=1}^{p} \hat{f}_\alpha$ be a minimizer of (5) in (3), with $\hat{f}_\alpha \in \mathcal{F}^\alpha$. Then $\hat{f}_\alpha \in$ span$\{R_\alpha(x_i, \cdot), i = 1, ..., n\}$, where $R_\alpha(\cdot, \cdot)$ is the reproducing kernel of the space $\mathcal{F}^\alpha$.*

However, it is possible to give an equivalent form of (5) that is easier to compute. Consider the problem of finding $\theta = (\theta_1, ..., \theta_p)^{\mathrm{T}}$ and $f \in \mathcal{F}$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n}\{y_i - f(x_i)\}^2 + \lambda_0 \sum_{\alpha=1}^{p} \theta_\alpha^{-1}\|P^\alpha f\|^2 + \lambda \sum_{\alpha=1}^{p} \theta_\alpha, \quad \text{subject to } \theta_\alpha \geq 0, \alpha = 1, ..., p, \quad (7)$$

where λ_0 is a constant that can be fixed at any positive value, and $\lambda_{(n)}$ is a smoothing parameter. We will fix λ_0 at some value for computational considerations.

LEMMA 2 *Set $\lambda = \tau^4/(4\lambda_0)$. (i) If $\hat{f}$ minimizes (5), set $\hat{\theta}_\alpha = \lambda_0^{1/2}\lambda^{-1/2}\|P^\alpha \hat{f}\|$, then the pair $(\hat{\theta}, \hat{f})$ minimizes (7). (ii) On the other hand, if a pair $(\hat{\theta}, \hat{f})$ minimizes (7), then $\hat{f}$ minimizes (5).*

The form of (7) is very similar to the common smoothing spline (4) with multiple smoothing parameters, except that there is an additional penalty on the θ's. Notice that there is only one smoothing parameter λ in (7). The θ's are part of the estimate, rather than free smoothing parameters. The additional penalty on θ's in (7) makes it possible to have some θ's be zeros, giving rise to zero function components in the COSSO estimate. In contrast, the common smoothing spline has two sets of smoothing parameters λ and θ's that are confounded. The common way to search for smoothing parameters iterates between λ and $\log\theta$'s, making it difficult to have zero components in the solution.

4 A special case with the tensor product design

In this section we consider the special case of a tensor product design with a SS-ANOVA model built from the second order Sobolev spaces of periodic functions (functions on the circle). The function space of the SS-ANOVA is then (2) with H's being the second order Sobolev spaces of periodic functions. This is a subspace of S consisting of periodic functions, and can be written as $T = \{1\} \oplus \bar{T}$, where

$$\bar{T} = \{f : f(t) = \sum_{\nu=1}^{\infty} a_\nu \sqrt{2}\cos 2\pi\nu t + \sum_{\nu=1}^{\infty} b_\nu \sqrt{2}\sin 2\pi\nu t, \quad \text{with} \quad \sum_{\nu=1}^{\infty}(a_\nu^2 + b_\nu^2)(2\pi\nu)^4 < \infty\}.$$

The reproducing kernel of $\bar{T}$ is $\bar{K}(s,t) = -k_4(|s-t|)$, and the norm in $\bar{T}$ is $\|g\|^2 = \int_0^1 \{g''(t)\}^2\, dt$ as used in Wahba (1990). For an even integer m, a useful approximate subspace of T is $T_m = \{1\} \oplus \bar{T}_m$, with

$$\bar{T}_m = \{f : f(t) = \sum_{\nu=1}^{m/2-1} a_\nu \sqrt{2} \cos 2\pi\nu t + \sum_{\nu=1}^{m/2-1} b_\nu \sqrt{2} \sin 2\pi\nu t + a_{m/2} \cos \pi m t\}.$$

Wahba (1990) used this subspace approximation to give a very instructive investigation of the filtering properties of the smoothing spline.

We assume that the data points follow a tensor product design. That is, the design points are

$$\{(x_{i_1,1}, x_{i_2,2}, ..., x_{i_d,d}) : i_k = 1, ..., n_k, k = 1, ..., d\},$$

where $x_{j,k} = j/n_k$, $j = 1, ..., n_k$, $k = 1, ..., d$. Therefore the total sample size is $n = n_1 \cdots n_d$. We assume the ϵ's in the regression model are independent with distribution $N(0, \sigma^2)$. We give a detailed treatment that sheds light on the mechanism of the COSSO for model selection, and show that under mild conditions, the COSSO selects the correct model structure with probability tending to one.

Without loss of generality, we fix $\lambda_0 = 1$ in the COSSO (7) and focus on the case of $d = 2$, $n_1 = n_2 = m$, and the SS-ANOVA model contains three components f_1, f_2, and f_{12}.

In this situation any $f \in \mathcal{F}$ can be written as $f(s,t) = b + f_1(s) + f_2(t) + f_{12}(s,t)$, with $f_1 \in \bar{T}^1$, $f_2 \in \bar{T}^2$, and $f_{12} \in \bar{T}^1 \otimes \bar{T}^2$, and (7) becomes

$$\begin{aligned} &\frac{1}{n}\sum_{k=1}^{m}\sum_{\ell=1}^{m}\{y_{kl} - f(x_{k,1}, x_{\ell,2})\}^2 + \theta_1^{-1}\int_0^1 \left\{\frac{\partial^2 f_1(s)}{\partial s^2}\right\}^2 ds + \theta_2^{-1}\int_0^1 \left\{\frac{\partial^2 f_2(t)}{\partial t^2}\right\}^2 dt \\ &+ \ \theta_{12}^{-1}\int_0^1\int_0^1 \left\{\frac{\partial^4 f_{12}(s,t)}{\partial s^2 t^2}\right\}^2 dsdt + \lambda(\theta_1 + \theta_2 + \theta_{12}), \quad \text{with} \ \ \theta_1 \geq 0, \theta_2 \geq 0, \theta_{12} \geq 0. \end{aligned}$$

We start with an argument for an approximate subspace of $\mathcal{F}$, and then move on to the space $\mathcal{F}$. Assume m is even, and define $\mathcal{F}_m = T_m^1 \otimes T_m^2 = \{1\} \oplus \bar{T}_m^1 \oplus \bar{T}_m^2 \oplus (\bar{T}_m^1 \otimes \bar{T}_m^2)$. We consider minimizing (7) in $\mathcal{F}_m$ first, as the argument is simple and instructive. Write $\gamma_1(t) = 1$, $\gamma_{2\nu}(t) = \sqrt{2}\cos(2\pi\nu t)$, $\gamma_{2\nu+1}(t) = \sqrt{2}\sin(2\pi\nu t)$, for $\nu = 1, ..., m/2 - 1$, and $\gamma_m(t) = \cos(\pi m t)$. Then any function in T_m can be written as $g(t) = \sum_{\nu=1}^m a_\nu \gamma_\nu(t)$. So any function in $\mathcal{F}_m$ can be written as

$$f(s,t) = \sum_{\mu=1}^{m}\sum_{\nu=1}^{m} a_{\mu\nu}\gamma_\mu(s)\gamma_\nu(t). \tag{8}$$

It is known that (see Wahba (1990), page 23)

$$\begin{aligned} m^{-1}\sum_{k=1}^{m}\gamma_\mu(k/m)\gamma_\nu(k/m) &= 1 \quad \text{if} \quad \mu=\nu=1,...,m; \\ &= 0 \quad \text{if} \quad \mu\neq\nu, \mu,\nu=1,...,m. \end{aligned}$$

Recall the definition of the inner product $\langle\cdot,\cdot\rangle_n$ of R^n defined in Section 2.4. Write $\gamma_{\mu\nu}(s,t)=\gamma_\mu(s)\gamma_\nu(t)$, and $\gamma_{\mu\nu}$ as the data vector corresponding to the function $\gamma_{\mu\nu}(s,t)$. From the above orthogonality relations and the tensor product design, we get

$$\begin{aligned} \langle\gamma_{\mu_1\nu_1},\gamma_{\mu_2\nu_2}\rangle_n &= 1 \quad \text{if} \quad \mu_1=\mu_2=1,...,m; \nu_1=\nu_2=1,...,m; \\ &= 0 \quad \text{if} \quad \mu_1\neq\mu_2 \text{ or } \nu_1\neq\nu_2, \mu_1,\nu_1,\mu_2,\nu_2=1,...,m. \end{aligned}$$

Therefore $\{\gamma_{\mu\nu},\mu=1,...,m;\nu=1,...,m\}$ form an orthonormal basis in R^n with respect to the norm $\|\cdot\|_n$. We then get from (8) that $a_{\mu\nu}=\langle f,\gamma_{\mu\nu}\rangle_n$. Write $z_{\mu\nu}=\langle y,\gamma_{\mu\nu}\rangle_n$. Then $z_{\mu\nu}=a_{\mu\nu}+\delta_{\mu\nu}$, where $\delta_{\mu\nu}\sim N(0,\sigma^2/n)$ are independent. The COSSO problem can be written as

$$\sum_{\mu=1}^{m}\sum_{\nu=1}^{m}(z_{\mu\nu}-a_{\mu\nu})^2+\theta_1^{-1}\sum_{\mu=2}^{m}q_{\mu1}a_{\mu1}^2+\theta_2^{-1}\sum_{\nu=2}^{m}q_{1\nu}a_{1\nu}^2+\theta_{12}^{-1}\sum_{\mu=2}^{m}\sum_{\nu=2}^{m}q_{\mu\nu}a_{\mu\nu}^2+\lambda(\theta_1+\theta_2+\theta_{12}), \tag{9}$$

with $q_{\mu\nu}\sim\mu^4\nu^4$ uniformly for $\mu\neq1$ or $\nu\neq1$, $\mu,\nu=1,...,m$. Here $\sim$ is read as "has the same order as." Therefore the minimizing $a_{\mu\nu}$ satisfies $\hat a_{11}=z_{11};\hat a_{\mu1}=z_{\mu1}\theta_1(\theta_1+q_{\mu1})^{-1}$, for $\mu\geq2$; $\hat a_{1\nu}=z_{1\nu}\theta_2(\theta_2+q_{1\nu})^{-1}$, for $\nu\geq2$; $\hat a_{\mu\nu}=z_{\mu\nu}\theta_{12}(\theta_{12}+q_{\mu\nu})^{-1}$, for $\mu\geq2,\nu\geq2$; and (9) becomes

$$\{\sum_{\mu=2}^{m}q_{\mu1}z_{\mu1}^2(q_{\mu1}+\theta_1)^{-1}+\lambda\theta_1\}+\{\sum_{\nu=2}^{m}q_{1\nu}z_{1\nu}^2(q_{1\nu}+\theta_2)^{-1}+\lambda\theta_2\}+\{\sum_{\mu=2}^{m}\sum_{\nu=2}^{m}q_{\mu\nu}z_{\mu\nu}^2(q_{\mu\nu}+\theta_{12})^{-1}+\lambda\theta_{12}\}.$$

We see that the three components can be minimized separately. Let us concentrate on θ_{12}, as θ_1 and θ_2 can be dealt with similarly. Let

$$A(\theta_{12})=\sum_{\mu=2}^{m}\sum_{\nu=2}^{m}q_{\mu\nu}z_{\mu\nu}^2(q_{\mu\nu}+\theta_{12})^{-1}+\lambda\theta_{12}.$$

Then $A'(\theta_{12})=\lambda-\sum_{\mu=2}^{m}\sum_{\nu=2}^{m}q_{\mu\nu}z_{\mu\nu}^2(q_{\mu\nu}+\theta_{12})^{-2}$, which increases as $\theta_{12}\geq0$ increases. Define $U=\sum_{\mu=2}^{m}\sum_{\nu=2}^{m}q_{\mu\nu}^{-1}z_{\mu\nu}^2$. If $U\leq\lambda$, then $A'(0)\geq0$, $A'(\theta_{12})>0$ for all $\theta_{12}>0$, and the minimizing $\hat\theta_{12}$ of A is 0; otherwise the minimizing $\hat\theta_{12}$ is larger than 0. We can see this is a soft thresholding type operation according to the magnitude of U. Notice $\hat\theta_{12}=0$

implies $\hat{f}_{12} = 0$.

Now we show that, if $f_{12} = 0$, and $n\lambda \to \infty$, then with probability tending to unity, $U \leq \lambda$, and therefore $\hat{\theta}_{12} = 0$. In this case $a_{\mu\nu} = 0$ for any pair (μ, ν) such that $\mu \geq 2$ and $\nu \geq 2$. So $E(U) \sim \sigma^2/n \sum_{\mu=2}^{m} \sum_{\nu=2}^{m} \mu^{-4}\nu^{-4} \sim n^{-1}\sigma^2, \mathrm{var}(U) = \sum_{\mu=2}^{m} \sum_{\nu=2}^{m} 2n^{-2}\sigma^4 q_{\mu\nu}^{-2} \sim n^{-2}\sigma^4$. Therefore when $n\lambda \to \infty$, by Chebyshev's inequality,

$$\mathrm{pr}(U > \lambda) \leq \mathrm{pr}(|U - E(U)| > \lambda - E(U)) \leq \mathrm{var}(U)/\{\lambda - E(U)\}^2 \to 0.$$

Next we show that, if $f_{12} \neq 0$, and $\lambda \to 0$, then with probability tending to unity, $U > \lambda$, and therefore $\hat{\theta}_{12} > 0$. In this case $a_{\mu_0,\nu_0} \neq 0$ for some $\mu_0 \geq 2$ and $\nu_0 \geq 2$, and then

$$\begin{aligned} E(U) &\geq E(q_{\mu_0,\nu_0}^{-1} z_{\mu_0,\nu_0}^2) \geq q_{\mu_0,\nu_0}^{-1} a_{\mu_0,\nu_0}^2; \\ \mathrm{var}(U) &= \sum_{\mu=2}^{m} \sum_{\nu=2}^{m} q_{\mu\nu}^{-2} \mathrm{var}(z_{\mu\nu}^2) = \sum_{\mu=2}^{m} \sum_{\nu=2}^{m} q_{\mu\nu}^{-2} (4n^{-1} a_{\mu\nu}^2 \sigma^2 + 2n^{-2}\sigma^4) \\ &\leq \{4n^{-1}\sigma^2 \sum_{\mu=2}^{m} \sum_{\nu=2}^{m} a_{\mu\nu}^2\} + 2n^{-2}\sigma^4 = 4n^{-1}\sigma^2 \|f_{12}\|_{L_2}^2 + 2n^{-2}\sigma^4 = O(n^{-1}). \end{aligned}$$

Therefore when $\lambda \to 0$, by Chebyshev's inequality, we get

$$\mathrm{pr}(U < \lambda) \leq \mathrm{pr}(|U - E(U)| > E(U) - \lambda) \leq \mathrm{var}(U)/\{E(U) - \lambda\}^2 \to 0.$$

We can see that the COSSO estimate smooths and at the same time thresholds components for model selection.

The argument for the function space $\mathcal{F}$ is similar but involves some technicality. We defer it to Appendix 2.

5 Algorithm

For any fixed θ, the COSSO (7) is equivalent to the smoothing spline (4). Therefore from the smoothing spline literature (for example,Wahba (1990)) it is well known the solution f has the form $f(x) = \sum_{i=1}^{n} c_i R_\theta(x_i, x) + b$, where $c = (c_1, ..., c_n)^{\mathrm{T}} \in R^n$, $b \in R$, and $R_\theta = \sum_{\alpha=1}^{p} \theta_\alpha R_\alpha$, with R_α being the reproducing kernel of $\mathcal{F}^\alpha$. With some abuse of notations, let R_α also stand for the $n \times n$ matrix $\{R_\alpha(x_i, x_j)\}$, $i = 1, ..., n$, $j = 1, ..., n$, let R_θ also stand for the matrix $\sum_{\alpha=1}^{p} \theta_\alpha R_\alpha$, and let 1_r be the column vector consisting of r ones. Then we can write $f = R_\theta c + b1_n$, and (7) can be expressed as

$$\frac{1}{n}(y - \sum_{\alpha=1}^{p} \theta_\alpha R_\alpha c - b1_n)^{\mathrm{T}}(y - \sum_{\alpha=1}^{p} \theta_\alpha R_\alpha c - b1_n) + \lambda_0 \sum_{\alpha=1}^{p} \theta_\alpha c^{\mathrm{T}} R_\alpha c + \lambda \sum_{\alpha=1}^{p} \theta_\alpha, \qquad (10)$$

where $\theta_\alpha \geq 0$, $\alpha = 1, ..., p$.

The form (10) turns out to be similar to the sparse kernel selection approach in Gunn and Kandola (2002). They used a different reproducing kernel and put penalty on all components including the constant b. They motivated their method by noting that the form of the penalty on the θ's in (10) tends to give sparse solutions for θ's, and gave empirical evidence to support the insight. Our method is motivated from a different formulation (5) which relates to the LASSO in linear models, and can be studied theoretically.

If θ's were fixed, then (10) can be written as

$$\min_{c,b}(y - R_\theta c - b1_n)^{\mathrm{T}}(y - R_\theta c - b1_n) + n\lambda_0 c^{\mathrm{T}} R_\theta c. \tag{11}$$

The solution to this smoothing spline problem is given in Wahba (1990).

On the other hand, if c and b were fixed, denote $g_\alpha = R_\alpha c$, and let G be the $n \times p$ matrix with the αth column being g_α. Simple calculation shows that the $\theta = (\theta_1, ..., \theta_p)^{\mathrm{T}}$ that minimizes (10) is the solution to

$$\min_\theta (z - G\theta)^{\mathrm{T}}(z - G\theta) + n\lambda \sum_{\alpha=1}^{p} \theta_\alpha \quad \text{subject to} \quad \theta_\alpha \geq 0, \alpha = 1, ..., p, \tag{12}$$

where $z = y - (1/2)n\lambda_0 c - b1_n$.

Therefore a reasonable scheme would be to iterate between (11) and (12). In each iteration (10) is decreased. Notice that (12) is equivalent to

$$\min_\theta (z - G\theta)^{\mathrm{T}}(z - G\theta) \quad \text{subject to} \quad \theta_\alpha \geq 0, \alpha = 1, ..., p; \sum_{\alpha=1}^{p} \theta_\alpha \leq M, \tag{13}$$

for some $M \geq 0$. If the algorithm that iterates between (11) and (12) converges, then the solution is also a fixed point for the algorithm that iterates between (11) and (13) for a fixed M. We prefer to iterate between (11) and (13) for computational considerations.

Notice that the formulation (13) is exactly the problem in calculating the nonnegative garrote estimate. Therefore our algorithm iterates between the smoothing spline and the nonnegative garrote. In our experience, it can take a large number of iterations for the algorithm to converge. In applications, though, we do not really need an exact solution of the COSSO. Our algorithm has a natural initial solution given by the smoothing spline, which is already a good estimate. By starting from this estimate and applying a limited number of iterations of our algorithm, we get what we view as an iterative improvement on the smoothing spline. This is in spirit similar to the basis pursuit in Chen, Donoho and Saunders (1998). We observe empirically that after the second iteration, the change between iterations is small but decreases slowly. This motivates us to consider the following one step update procedure:

1. Initialization: Fix $\theta_\alpha = 1$, $\alpha = 1, ..., p$.

2. Solve for c and b with (11).

3. For the c and b obtained in step 2, solve for θ with the nonnegative garrote (13).

4. With the new θ, solve for c and b with the smoothing spline (11).

This one step update procedure has the flavor of the one step maximum likelihood procedure, in which one step Newton-Raphson algorithm is applied to a good initial estimator and which is as efficient as the fully iterated maximum likelihood. A discussion of one step procedure and fully iterated procedure (in a different algorithm) can be found in Fan and Li (2001). In our experience, the one step update procedure and the fully iterated procedure have comparable estimation accuracy.

6 Choosing the tuning parameter

The generalized cross validation proposed by Craven and Wahba (1979) is one of the most popular methods for choosing smoothing parameters in the smoothing spline method. Let A be the smoothing matrix of the smoothing spline. That is, $\hat{y} = Ay$. The generalized cross validation estimate of the risk is

$$GCV = \frac{\|\hat{y} - y\|_n^2}{\{n^{-1}\text{tr}(I - A)\}^2}.$$

Tibshirani (1996) proposed a GCV-type criterion for choosing the tuning parameter for the LASSO through a ridge estimate approximation. This approximation is particularly easy to understand in light of the form (7) for the linear model $f(x) = \beta_0 + \sum_{j=1}^d \beta_j x^{(j)}$: fix the θ_j's at their estimated values $\hat{\theta}_j$'s, and calculate GCV for the corresponding ridge regression. This approximation ignores some variability in the estimation process. However, the simulation study in Tibshirani (1996) suggests that it is a useful approximation. This motivates our GCV-type criterion: We use the GCV score for the smoothing spline in (7) when θ's are fixed at the solution.

Another popular technique for choosing tuning parameters is the five or ten fold cross validation. The computation load of GCV is smaller. We compare the performances of these two criteria in the COSSO with simulations.

The following is the complete algorithm for the COSSO with adaptive tuning:

1. Fix $\theta_\alpha = 1$, $\alpha = 1, ..., p$. Solve the smoothing spline problem, and tune λ_0 according to CV or GCV. Fix λ_0 at the chosen value in all later steps.

2. For each fixed M in a reasonable range, apply the one step COSSO algorithm with M. Choose the best M according to CV or GCV. The solution corresponding to this chosen M is the final solution.

In our simulations, it is noticed that once λ_0 is fixed according to step 1, the optimal M seems to be close to the number of important components. This helps for determining the range of tuning for M. We tune M between 0 and 35.

7 Simulations

In this section we study the empirical performance of the COSSO estimate in terms of estimation accuracy and model selection. The COSSO estimate is compared with the MARS, which is a popular stepwise forward/backward procedure for building functional ANOVA models. The following four functions on $[0, 1]$ are used as building blocks of regression functions in some of the simulations:

$$g_1(t) = t; \quad g_2(t) = (2t-1)^2; \quad g_3(t) = \frac{\sin(2\pi t)}{2-\sin(2\pi t)};$$

$$g_4(t) = 0.1\sin(2\pi t) + 0.2\cos(2\pi t) + 0.3\sin^2(2\pi t) + 0.4\cos^3(2\pi t) + 0.5\sin^3(2\pi t).$$

We consider two covariance structures of the input vector X, with varying degrees of correlation:

Compound symmetry: Let $X^{(j)} = (W^{(j)} + tU)/(1+t)$, $j = 1, ..., d$, where $W^{(1)}$, ..., $W^{(d)}$ and U are i.i.d from Uniform(0,1). Therefore $\mathrm{corr}(X^{(j)}, X^{(k)}) = t^2/(1+t^2)$ for $j \neq k$. The uniform design corresponds to the case $t = 0$.

(trimmed) AR(1): Let $W^{(1)}$, ..., $W^{(d)}$ be i.i.d $N(0, 1)$, and let $X^{(1)} = W^{(1)}$, $X^{(j)} = \rho X^{(j-1)} + (1-\rho^2)^{1/2} W^{(j)}$, $j = 2, ..., d$. Trim $X^{(j)}$ in [-2.5, 2.5] and scale to [0,1].

Example 1. We consider a simple additive model in R^{10}. The underlying regression function is $f(x) = 5g_1(x^{(1)}) + 3g_2(x^{(2)}) + 4g_3(x^{(3)}) + 6g_4(x^{(4)})$. Therefore $X^{(5)}, ..., X^{(10)}$ are uninformative. We consider a sample size $n = 100$. To start with, we generate X uniformly from $[0, 1]^{10}$. We generate $y = f(x) + \epsilon$, where ϵ is a normal variate with mean 0 and variance 1.74. The standard deviation of the noise was chosen to give a signal to noise ratio 3 : 1 in the uniform case. For comparison, the variances of the component functions are $\mathrm{var}\{5g_1(X^{(1)})\} = 2.08$, $\mathrm{var}\{3g_2(X^{(2)})\} = 0.80$, $\mathrm{var}\{4g_3(X^{(3)})\} = 3.30$ and $\mathrm{var}\{6g_4(X^{(4)})\} = 9.45$.

We apply the COSSO with additive models (the additive COSSO) to the simulated data. Therefore there are 10 functional components in the model. Figure 1 shows how the magnitudes of the estimated components change with the tuning parameter M in one run. The magnitudes of the functional components are measured by their empirical L_1 norms, defined as $1/n \sum_{i=1}^{n} |\hat{f}_j(x_i^{(j)})|$ for $j = 1, ..., d$. The λ_0 in this run is fixed at 9.7656×10^{-6}. Both GCV and five-fold cross validation choose $M = 3.5$, giving a model of 5 terms in this run. The estimated function components are plotted along with the true function components in Figure 2. Notice the components are centered according to the ANOVA decomposition.

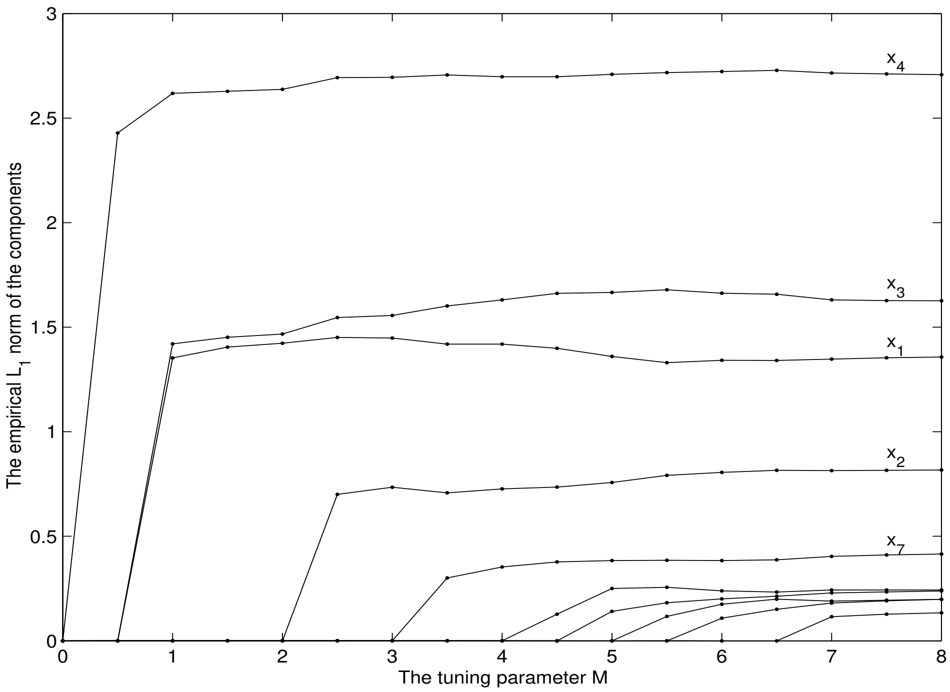

Figure 1: The empirical L_1 norm of the estimated components as plotted against the tuning parameter M in one run of Example 1.

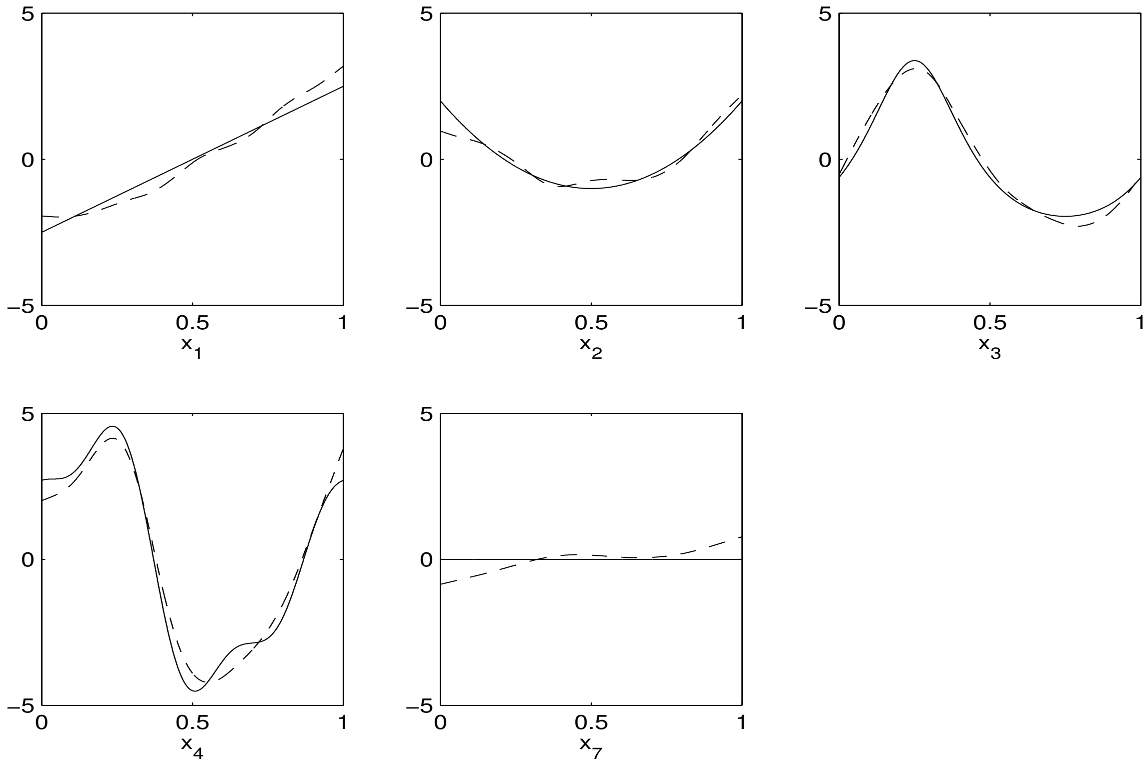

Figure 2: The estimated component functions (dashed line) and the true component functions (solid line) in one run of Example 1. Shown are the components for variables 1, 2, 3, 4, and 7. For the other variables, both the true and estimated component functions are zero.

We compare COSSO with GCV, COSSO with five-fold cross validation, and MARS. The measure of accuracy is the integrated squared error ISE $= E_X\{\hat{f}(X) - f(X)\}^2$. For each replication of the simulation study, the ISE is estimated by Monte Carlo integration using 10000 test points from the same distribution as the training points. We run our simulation 100 times and average. This is further done for several different settings with the compound symmetry covariance structure and the AR(1) covariance structure. The resulting average integrated squared error and its associated standard error (in parentheses) are given in Table 1. The matlab code for the COSSO is available from webpages of the authors (www.stat.wisc.edu/˜yilin or www.stat.ncsu.edu/˜hzhang2). Also included in the table is the average integrated squared error of MARS for additive models. The MARS simulations are done in R, with the function "mars" in the "mda" library contributed by Trevor Hastie and Robert Tibshirani. We can see the two COSSO procedures perform better than MARS in all the settings studied.

Table 1. The average integrated squared errors over 100 runs of Example 1 in different settings.

	Comp. symm.			AR(1)		
	$t = 0$	$t = 1$	$t = 3$	$\rho = -0.5$	$\rho = 0$	$\rho = 0.5$
COSSO(GCV)	0.93 (0.05)	0.92 (0.04)	0.97 (0.07)	0.94 (0.05)	1.04 (0.07)	0.98 (0.07)
COSSO(5CV)	0.80 (0.03)	0.97 (0.05)	1.07 (0.06)	1.03 (0.06)	1.03 (0.06)	0.98 (0.05)
MARS	1.57 (0.07)	1.24 (0.06)	1.30 (0.06)	1.32 (0.07)	1.34 (0.07)	1.36 (0.08)

To study the performance of the COSSO in terms of model selection, we determine in the uniform case the number of times each variable appears in the 100 chosen models (Table 2), and the number of terms in the 100 chosen models (Table 3). In our calculation we take θ to be zero if it is smaller than 10^{-6}. The COSSO with five-fold cross validation misses the second variable 6 times, but chooses the correct four variable model 84 times. The COSSO with GCV and MARS never miss any important variable, but tend to include uninformative variables in the chosen models. The COSSO with GCV chooses the correct four variable model 57 times, while MARS does that only 4 times.

Table 2. The frequency of appearance of the variables in the models chosen in 100 runs of Example 1 in the uniform setting.

	Variable									
	1	2	3	4	5	6	7	8	9	10
COSSO(GCV)	100	100	100	100	14	11	18	15	11	13
COSSO(5CV)	100	94	100	100	1	1	3	2	4	2
MARS	100	100	100	100	35	35	34	39	28	35

Table 3. The frequency of the size of the models chosen in 100 runs of Example 1 in the uniform setting.

	Model size								
	3	4	5	6	7	8	9	10	Mean
COSSO(GCV)	0	57	17	18	5	2	0	1	4.82
COSSO(5CV)	6	84	7	3	0	0	0	0	4.07
MARS	0	4	24	40	26	6	0	0	6.06

Table 4 gives the mean and standard deviation of the model sizes chosen by the methods in various settings. The settings considered are compound symmetry with $t = 1$ and 3 and trimmed AR(1) with $\rho = -0.5$, 0 and 0.5. The average model size chosen by the COSSO with five-fold cross validation is close to 4, the size of the true model. The COSSO with GCV selects slightly larger models. The models chosen by MARS are even larger.

Table 4. Mean and standard deviation of the model sizes chosen in 100 runs of Example 1.

	Comp. symm.		AR(1)		
	$t=1$	$t=3$	$\rho=-0.5$	$\rho=0$	$\rho=0.5$
COSSO(GCV)	4.8 (1.2)	4.8 (1.5)	4.7 (1.2)	4.8 (1.3)	4.6 (1.2)
COSSO(5CV)	4.1 (1.2)	4.4 (1.9)	4.1 (1.2)	4.0 (1.0)	3.8 (0.9)
MARS	6.3 (0.9)	6.2 (0.9)	6.1 (1.0)	6.1 (0.8)	5.9 (0.8)

Example 2. We consider a larger additive model with $d=60$. The regression function is

$$\begin{aligned} f(x) &= g_1(x^{(1)}) + g_2(x^{(2)}) + g_3(x^{(3)}) + g_4(x^{(4)}) \\ &+ 1.5g_1(x^{(5)}) + 1.5g_2(x^{(6)}) + 1.5g_3(x^{(7)}) + 1.5g_4(x^{(8)}) \\ &+ 2g_1(x^{(9)}) + 2g_2(x^{(10)}) + 2g_3(x^{(11)}) + 2g_4(x^{(12)}). \end{aligned}$$

Therefore there are 48 uninformative variables. The sample size is 500. The variance of the normal noise is set to 0.5184, to give a signal to noise ratio of 3 : 1 in the uniform case. For comparison, in the uniform setting $\mathrm{var}\{g_1(X^{(1)})\} = 0.08$, $\mathrm{var}\{g_2(X^{(2)})\} = 0.09$, $\mathrm{var}\{g_3(X^{(3)})\} = 0.21$ and $\mathrm{var}\{g_4(X^{(4)})\} = 0.26$. Both COSSO and MARS are run 100 times with additive models. The results are summarized in Tables 5 and 6. We see that the two COSSO procedures outperform MARS, with the COSSO with five-fold cross validation doing slightly better than the COSSO with GCV. The COSSO(5CV) is impressive in its ability to select the correct model size: the models chosen by it all have sizes close to 12, the size of the true model.

Table 5. The average integrated squared error over 100 runs of Example 2 and its standard error, in the unit of 10^{-3}.

	Comp. symm.		AR(1)	
	$t=0$	$t=1$	$\rho=0.5$	$\rho=-0.5$
COSSO(GCV)	201 (4)	178 (5)	199 (6)	183 (5)
COSSO(5CV)	144 (4)	162 (5)	153 (4)	149 (5)
MARS	353 (7)	302 (7)	286 (6)	280 (5)

Table 6. Mean and standard deviation of the model sizes chosen in 100 runs of Example 2.

	Comp. symm.		AR(1)	
	$t = 0$	$t = 1$	$\rho = 0.5$	$\rho = -0.5$
COSSO(GCV)	18.0 (4.1)	18.0 (4.1)	19.0 (5.1)	18.0 (4.3)
COSSO(5CV)	12.0 (0.2)	11.7 (1.4)	12.1 (1.4)	11.9 (1.0)
MARS	35.2 (2.3)	36.1 (2.1)	35.2 (2.5)	35.9 (2.4)

Example 3. We consider a 10 dimensional regression problem with several two way interactions:

$$f(x) = g_1(x^{(1)}) + g_2(x^{(2)}) + g_3(x^{(3)}) + g_4(x^{(4)}) + g_1(x^{(3)}x^{(4)}) + g_2(\frac{x^{(1)} + x^{(3)}}{2}) + g_3(x^{(1)}x^{(2)}).$$

We consider the uniform setting, and set the noise to be normal with standard deviation 0.2546, to give a signal to noise ratio of 3 : 1. The average integrated squared errors are given in Table 7 for sample sizes $n = 100, 200, 400$. Both the COSSO and MARS are run with the two way interaction model. We follow the advice in Friedman (1991) to set the cost for each basis function optimization to be 3 in the MARS for two way interaction models.

Table 7. The average integrated squared error over 100 runs of Example 3 and its standard error.

	$n = 100$	$n = 200$	$n = 400$
COSSO(GCV)	0.358 (0.009)	0.100 (0.003)	0.045 (0.001)
COSSO(5CV)	0.378 (0.005)	0.094 (0.004)	0.043 (0.001)
MARS	0.239 (0.008)	0.109 (0.003)	0.084 (0.001)

There are 55 function components in the COSSO. The COSSO does not do well when $n = 100$. It seems that there are too many function components for the COSSO to select from with 100 data points. MARS does not suffer from a small sample size so much as the COSSO. Part of the reason is that the MARS algorithm introduces a certain hierarchical order of the terms being searched from: only after a univariate basis function is included in the model, will the product of other terms with it become a candidate for inclusion in later steps. In contrast, the COSSO selects from all the function components and does not distinguish between main effects and interaction terms. Therefore the COSSO does not assume any hierarchical structure, and may not be efficient when the true model is hierarchical and the sample size is small. However, as the sample size increases, the COSSO procedures catch up quickly. Their performances are comparable to MARS when $n = 200$ and better than MARS when $n = 400$.

In the above examples we see that in general the COSSO with five-fold cross validation tends to do better than the COSSO with the GCV. We therefore recommend the use of five-fold cross validation with the COSSO unless the computation time is a crucial factor.

Example 4. The circuit example. This is an example from Friedman (1991). Of interest is the dependence of the impedance Z of a circuit and phase shift ϕ on components in the circuit. The true dependence is described by

$$\begin{aligned} Z &= [R^2 + \{\omega L - 1/(\omega C)\}^2]^{1/2}, \\ \phi &= \tan^{-1}\{\frac{\omega L - 1/(\omega C)}{R}\}. \end{aligned}$$

The input variables are uniform in the range

$$0 \le R \le 100, 40\pi \le \omega \le 560\pi, 0 \le L \le 1, 1 \le C \le 11,$$

and the noise is normal with the standard deviation set to give a signal to noise level $3:1$.

This is a relatively small problem with $d = 4$. All order of interactions are present. Friedman (1991) applied MARS with additive model, two way interaction model, and the saturated model to this example, and found that the performance of the two way interaction model was the best. We scale the input region to $[0,1]^4$ and apply the COSSO with five fold cross validation. With the small dimension, it is possible to apply the COSSO with the saturated model, which has $2^4 - 1 = 15$ function components. However, it turns out that the two way interaction COSSO does slightly better than the saturated model. We compare the integrated squared error of the two way interaction COSSO and that of the two way interaction MARS in Tables 8 and 9. It turns out that the COSSO performs much better than MARS.

Table 8. The average integrated squared error and its standard error for estimating the impedance Z, in the unit of 10^3.

	$n = 100$	$n = 200$	$n = 400$
COSSO	1.91 (0.12)	0.85 (0.05)	0.51 (0.03)
MARS	5.57 (0.41)	2.47 (0.16)	1.37 (0.08)

Table 9. The average integrated squared error and its standard error for estimating the phase shift ϕ, in the unit of 10^{-3}.

	$n = 100$	$n = 200$	$n = 400$
COSSO	12.98 (0.36)	7.96 (0.20)	5.36 (0.10)
MARS	20.59 (0.96)	12.60 (0.71)[1]	8.19 (0.14)[2]

1. Excluded one extreme outlier.

2. Excluded three extreme outliers.

8 Real examples

We apply the COSSO to three real datasets. They are the Ozone data, the Boston housing data, and the Tecator data. The first two datasets are available from the R library "mlbench". The Ozone data was used in Breiman and Friedman (1985), Buja, Hastie and Tibshirani (1989) and Breiman (1995). The daily maximum one-hour-average ozone reading and 8 meteorological variables were recorded in the Los Angeles basin for 330 days of 1976. The Boston housing data concerns housing values in suburbs of Boston. There are 12 input variables. The sample size is 506. The Tecator data is available from the datasets archive of StatLib at lib.stat.cmu.edu/datasets/. The data was recorded on a Tecator Infratec Food and Feed Analyzer working in the wavelength range 850 - 1050 nm by the Near Infrared Transmission (NIT) principle. 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. The absorbance is $-\log_{10}$ of the transmittance measured by the spectrometer. As suggested in the document, we use the first 13 principal components to predict the fat content. The total sample size is 215.

We apply the COSSO and the MARS on these datasets, and estimate the prediction squared errors $E[\{Y - \hat{f}(X)\}^2]$ by ten-fold cross validation. We select the tuning parameter by five-fold cross validation within the training set. The estimate obtained is then evaluated on the test set. We do this ten-fold cross validation five times and then average. For all three datasets, both the COSSO and the MARS find the two way interaction model has better prediction accuracy than the additive model. Therefore we choose to apply the two way interaction model and the results are given in Table 10. We can see the COSSO does considerably better than the MARS.

Table 10. The estimated prediction squared errors and their standard errors (in parentheses).

	Ozone	Boston	Tecator
COSSO	16.04 (0.06)	9.89 (0.08)	0.92 (0.02)
MARS	18.24 (0.45)	14.31 (0.34)	4.99 (1.07)

9 Discussion

The difference between the COSSO and the common smoothing spline mirrors that between the LASSO and the ridge regression. We have shown that the COSSO has attractive properties for model selection and estimation.

It might be possible to improve the performance of the GCV in the COSSO by taking into account the variability in the estimated θ's. However, how to account for the degrees of freedom for calculating the θ's is not clear, since the θ's are constrained instead of free parameters. This is a direction for further research.

The computation of LASSO is quite intensive for large data. The smoothing spline step in the COSSO alone requires $O(n^3)$ flops. A method that is commonly employed in the smoothing spline method to reduce computation load is the parsimonious bases approach used by Xiang and Wahba (1996), Ruppert and Carroll (2000), Lin, Wahba, Xiang, Gao, Klein and Klein (2000) and Yau, Kohn and Wood (2002). It has been shown that the number of basis terms can be much smaller than n without degrading the performance of the estimation. It should be possible to use this approach in the COSSO estimate for large problems.

It is possible to extend the COSSO idea to other settings, such as the generalized regression, hazard regression, and density estimation. The performances of the COSSO in these settings will be studied in the future.

It is possible to construct confidence intervals and error bars for the COSSO estimate by bootstrapping. However, that is computationally very intensive. It is of interest to see to what extent the Bayesian confidence interval of the smoothing spline in Wahba (1983) and Nychka (1988) can be adapted to the COSSO.

APPENDIX 1

Proof of Theorem 1. Denote the functional to be minimized in (5) by $A(f)$, then $A(f)$ is convex and continuous. Without loss of generality, we assume $\tau = 1$.

By (6) we have that $J(f) \geq \|f\|$, for any $f \in \mathcal{F}_1$. Let $R_{\mathcal{F}_1}$ be the reproducing kernel of $\mathcal{F}_1$ and $\langle \cdot, \cdot \rangle_{\mathcal{F}_1}$ be the inner product in $\mathcal{F}_1$. Denote $a = \max_{i=1}^n R_{\mathcal{F}_1}^{1/2}(x_i, x_i)$. By the definition of reproducing kernel, we have for any $f \in \mathcal{F}_1$ and $i = 1, ..., n$,

$$\begin{aligned} |f(x_i)| &= |\langle f(\cdot), R_{\mathcal{F}_1}(x_i, \cdot)\rangle_{\mathcal{F}_1}| \leq \|f\| \langle R_{\mathcal{F}_1}(x_i, \cdot), R_{\mathcal{F}_1}(x_i, \cdot)\rangle_{\mathcal{F}_1}^{1/2} \\ &= \|f\| R_{\mathcal{F}_1}^{1/2}(x_i, x_i) \leq a\|f\| \leq aJ(f). \end{aligned} \tag{A1}$$

Denote $\rho = \max_{i=1}^n (y_i^2 + |y_i| + 1)$. Consider the set

$$D = \{f \in \mathcal{F} : f = b + f_1, \text{with } \ b \in \{1\}, f_1 \in \mathcal{F}_1, J(f) \leq \rho, |b| \leq \rho^{1/2} + (a+1)\rho\}.$$

Then D is a closed, convex, and bounded set. Therefore by Theorem 4 of Tapia and Thompson (1978, page 162), there exists a minimizer of (5) in D. Denote the minimizer by $\bar{f}$. Then $A(\bar{f}) \leq A(0) < \rho$.

On the other hand, for any $f \in \mathcal{F}$ with $J(f) > \rho$, clearly $A(f) \geq J(f) > \rho$; for any $f \in \mathcal{F}$ with $J(f) \leq \rho$, $f = b + f_1$, $b \in \{1\}$, $f_1 \in \mathcal{F}$, and $|b| > \rho^{1/2} + (a+1)\rho$, we use (A1) to get that, for any $i = 1, ..., n$,

$$|b + f_1(x_i) - y_i| > [\rho^{1/2} + (a+1)\rho] - a\rho - \rho = \rho^{1/2}.$$

Therefore we have $A(f) > \rho$.

Hence for any $f \notin D$, we have $A(f) > A(\bar{f})$. Therefore $\bar{f}$ is a minimizer of (5) in $\mathcal{F}$.

PROOF OF THEOREM 2. The proof follows the approach in van de Geer (2000) and makes use of Theorem 10.2 of van de Geer (2000, page 169). For completeness we state the theorem first: (the original theorem is more general, and the following corresponds to the special case $\nu = 1$ and Gaussian noises of the original theorem):

LEMMA 3 (THEOREM 10.2 OF VAN DE GEER) *Consider a regression model $y_i = g_0(x_i) + \epsilon_i$, $i = 1, ..., n$, where g_0 is known to lie in a class $\mathcal{G}$ of functions, x_i's are given covariates in $[0,1]^d$, and ϵ_i's are independent $N(0, \sigma^2)$ noises. Let $I : \mathcal{G} \to [0, \infty)$ be a pseudo-norm on $\mathcal{G}$. Define $\hat{g} = \arg\min_{g \in \mathcal{G}} 1/n \sum_{i=1}^n \{y_i - g(x_i)\}^2 + \tau_n^2 I(g)$. Assume*

$$H_\infty(\delta, \{\frac{g - g_0}{I(g) + I(g_0)} : g \in \mathcal{G}, I(g) + I(g_0) > 0\}) \leq A\delta^{-\alpha} \tag{A2}$$

for all $\delta > 0$, $n \geq 1$ and some $A > 0$, $0 < \alpha < 2$. Here H_∞ stands for the entropy for the supreme norm. Then (i) if $I(g_0) > 0$, and $\tau_n^{-1} = O_p(n^{1/(2+\alpha)}) I^{(2-\alpha)/(4+2\alpha)}(g_0)$, we have $\|\hat{g} - g_0\|_n = O_p(\tau_n) I^{1/2}(g_0)$; (ii) if $I(g_0) = 0$, we have $\|\hat{g} - g_0\|_n = O_p(n^{-1/(2-\alpha)}) \tau_n^{-2\alpha/(2-\alpha)}$.

The above lemma can not be used directly since (A2) is not satisfied in our case. This problem can be dealt with with the following arguments.

For any $f \in \mathcal{F}$, we can write $f(x) = c + f_1(x^{(1)}) + ... + f_d(x^{(d)}) = c + g(x)$, such that $\sum_{i=1}^n f_j(x_i^{(j)}) = 0$, $j = 1, ..., d$. Similarly, write $f_0(x) = c_0 + f_{01}(x^{(1)}) + ... + f_{0d}(x^{(d)}) = c_0 + g_0(x)$, and $\hat{f}(x) = \hat{c} + \hat{f}_1(x^{(1)}) + ... + \hat{f}_d(x^{(d)}) = \hat{c} + \hat{g}(x)$. Then by construction $\sum_{i=1}^n \{g_0(x_i) - g(x_i)\} = 0$, we can write (5) as

$$(c_0 - c)^2 + \frac{2}{n}(c_0 - c)\sum_{i=1}^n \epsilon_i + \frac{1}{n}\sum_{i=1}^n \{g_0(x_i) + \epsilon_i - g(x_i)\}^2 + \tau_n^2 J(g).$$

Therefore, the minimizing $\hat{c}$ must minimize $(c_0 - c)^2 + 2/n(c_0 - c)\sum_{i=1}^n \epsilon_i$. That is, $\hat{c} = c_0 + 1/n \sum_{i=1}^n \epsilon_i$. Therefore $(\hat{c} - c_0)^2$ converges with rate n^{-1}. On the other hand, $\hat{g}$ must minimize

$$\frac{1}{n}\sum_{i=1}^n \{g_0(x_i) + \epsilon_i - g(x_i)\}^2 + \tau_n^2 J(g).$$

Let $\mathcal{G} = \{g \in \mathcal{F} : g(x) = f_1(x^{(1)}) + ... + f_d(x^{(d)}),\ \text{with}\ \sum_{i=1}^n f_j(x_i^{(j)}) = 0, j = 1, ..., d\}$. Then $g_0 \in \mathcal{G}$, $\hat{g} \in \mathcal{G}$. We can now apply Lemma 3 with $I = J$ and $\alpha = 1/2$. That (A2) is satisfied follows from Lemma 4 given below, since $J(g-g_0) \leq J(g)+J(g_0)$ for any $g \in \mathcal{G}$. The conclusion of Theorem 2 then follows from the conclusion of Lemma 3.

LEMMA 4

$$H_\infty(\delta, \{g \in \mathcal{G} : J(g) \leq 1\}) \leq Ad^{3/2}\delta^{-1/2},$$

for all $\delta > 0$, $n \geq 1$ and some $A > 0$ not depending on δ, n, or d.

PROOF OF LEMMA 4. Define $\mathcal{G}^j$ as the set of univariate functions of $x^{(j)}$:

$$\mathcal{G}^j = \{f_j \in S : J(f_j) \leq 1, \sum_{i=1}^n f_j(x_i^{(j)}) = 0\} = \{f_j : \{f_j(1)-f_j(0)\}^2 + \int_0^1 (f_j'')^2 \leq 1, \sum_{i=1}^n f_j(x_i^{(j)}) = 0\},$$

where S is the second order Sobolev space.

First, we show that for any $h \in \mathcal{G}^j$, we have $|h|_\infty \equiv \{\sup_{s\in[0,1]} |h(s)|\} \leq 1$. Since $\sum_{i=1}^n h(x_i^{(j)}) = 0$, there exists $a \in [0,1]$, such that $h(a) = 0$. Now if h is monotone, since $h \in \mathcal{G}^j$, we have $\max h - \min h = |h(1)-h(0)| \leq 1$, so $|h|_\infty \leq 1$; if h is not monotone, then $\max(h') > 0$ and $\min(h') < 0$. Now since $h \in \mathcal{G}^j$, we have

$$1 \geq \int_0^1 (h'')^2 \geq (\int_0^1 |h''|)^2 \geq \{\max(h') - \min(h')\}^2.$$

So $-1 \leq \min(h') < 0 < \max(h') \leq 1$. So $|h'|_\infty \leq 1$. Now since $h(a) = 0$, we have $|h|_\infty \leq 1$.

It then follows from Theorem 2.4 of van de Geer (2000, page 19)that

$$H_\infty(\delta, \mathcal{G}^j) \leq A\delta^{-1/2} \tag{A3}$$

for all $\delta > 0$, and $n \geq 1$, and some positive A not depending on δ and n.

By the definition of the $\mathcal{G}$ and $\mathcal{G}^j$, we see that in terms of the supreme norm, if each $\mathcal{G}^j$, $j = 1, ...d$, can be covered by N balls of radius δ, Then the set $\{g \in \mathcal{G} : J(g) \leq 1\}$ can be covered by N^d balls with radius $d\delta$. By (A3) we get

$$H_\infty(d\delta, \{g \in \mathcal{G} : J(g) \leq 1\}) \leq Ad\delta^{-1/2},$$

and the conclusion of the lemma follows.

PROOF OF LEMMA 1. For any $f \in \mathcal{F}$, we can write $f = b + \sum_{\alpha=1}^p f_\alpha$ with $f_\alpha \in \mathcal{F}^\alpha$. Let the projection of f_α onto $\text{span}\{R_\alpha(x_i, \cdot), i = 1, ..., n\} \subset \mathcal{F}^\alpha$ be denoted by g_α, and the orthogonal complement by h_α. Then $f_\alpha = g_\alpha + h_\alpha$, and $\|f_\alpha\|^2 = \|g_\alpha\|^2 + \|h_\alpha\|^2$, $\alpha = 1, ..., p$. Since $R = 1 + \sum_{\alpha=1}^p R_\alpha$ is the reproducing kernel of $\mathcal{F}$, we have, making use of the orthogonal structures,

$$f(x_i) = \langle 1 + \sum_{\alpha=1}^{p} R_\alpha(x_i, \cdot), b + \sum_{\alpha=1}^{p}(g_\alpha + h_\alpha)\rangle = b + \sum_{\alpha=1}^{p}\langle R_\alpha(x_i, \cdot), g_\alpha\rangle,$$

where $\langle \cdot, \cdot \rangle$ is the inner product in $\mathcal{F}$. Therefore (5) can be written as

$$\frac{1}{n}\sum_{i=1}^{n}\{y_i - b - \sum_{\alpha=1}^{p}\langle R_\alpha(x_i, \cdot), g_\alpha\rangle\}^2 + \tau^2 \sum_{\alpha=1}^{p}(\|g_\alpha\|^2 + \|h_\alpha\|^2)^{1/2}.$$

Therefore any minimizing f satisfies $h_\alpha = 0$, $\alpha = 1, ..., p$, and the conclusion of the lemma follows.

Proof of Lemma 2. Denote the functional in (5) by $A(f)$, and the functional in (7) by $B(\theta, f)$. For any $\alpha = 1, ..., p$, we have $\lambda_0 \theta_\alpha^{-1}\|P^\alpha f\|^2 + \lambda\theta_\alpha \geq 2\lambda_0^{1/2}\lambda^{1/2}\|P^\alpha f\| = \tau^2\|P^\alpha f\|$, for any $\theta_\alpha \geq 0$ and $f \in \mathcal{F}$, and the equality holds if and only if $\theta_\alpha = \lambda_0^{1/2}\lambda^{-1/2}\|P^\alpha f\|$. Therefore $B(\theta, f) \geq A(f)$ for any $\theta_\alpha \geq 0$, $\alpha = 1, ..., p$, and $f \in \mathcal{F}$, and the equality holds if and only if $\theta_\alpha = \lambda_0^{1/2}\lambda^{-1/2}\|P^\alpha f\|$, $\alpha = 1, ..., p$. The conclusion of the lemma follows.

APPENDIX 2

Further derivations in the tensor product design case. Now we consider the function space $\mathcal{F}$. Define $\Sigma = \{\bar{K}(x_{i,1}, x_{j,1})\}_{m\times m}$, the marginal kernel matrix corresponding to the reproducing kernel of $\bar{T}$. With a little abuse of notation, let R_j, $j = 1, 2$, also stand for the $n \times n$ matrix of the reproducing kernel R_j evaluated at the n data points, and same for R_{12}. Suppose the data points are permuted appropriately, we have $R_{12} = \Sigma \otimes \Sigma$, where $\otimes$ stands for the Kronecker product between matrices. See Chen (1991). Let 1_m be the column vector consisting of m ones. For the main effect spaces, we have, $R_1 = \Sigma \otimes (1_m 1_m^{\mathrm{T}})$ and $R_2 = (1_m 1_m^{\mathrm{T}}) \otimes \Sigma$.

Straightforward calculation gives $\Sigma 1_m = mt1_m$, where $t = 1/(720m^4)$. Let $\{\xi_1 = 1_m, \xi_2, ..., \xi_m\}$ be an orthonormal (with respect to the inner product $\langle\cdot,\cdot\rangle_m$ in R^m) eigensystem of Σ, with corresponding eigenvalues $m\eta_1, m\eta_2, ..., m\eta_m$, where $\eta_1 = t$, and $\eta_2 \geq \eta_3 \geq ... \geq \eta_m$. Then it is well known that $\eta_i \sim i^{-4}$, for $i \geq 2$. See Utreras (1983). Notice $\xi_1, \xi_2, ..., \xi_m$ are also the eigenvectors of $1_m 1_m^{\mathrm{T}}$, with corresponding eigenvalues being m, $0, ..., 0$. Write $\xi_{\mu\nu} = \xi_\mu \otimes \xi_\nu$. It is then easy to check that $\{\xi_{\mu\nu} : \mu, \nu = 1, ..., m\}$ form an eigensystem of R_1, R_2, and R_{12}. The eigenvalues of R_1, R_2, and R_{12} are, respectively,

$$\begin{aligned} r_{1,\mu 1} &= n\eta_\mu; \quad r_{1,\mu\nu} = 0 \quad \text{for } \mu \geq 1, \nu \geq 2; \\ r_{2,1\nu} &= n\eta_\nu; \quad r_{2,\mu\nu} = 0 \quad \text{for } \mu \geq 2, \nu \geq 1; \\ r_{12,\mu\nu} &= n\eta_\mu\eta_\nu \quad \text{for } \mu \geq 1, \nu \geq 1. \end{aligned}$$

It is clear that $\{\xi_{\mu\nu} : \mu, \nu = 1, ..., m\}$ is also an orthonormal basis in R^n with respect to the inner product $\langle\cdot,\cdot\rangle_n$. Consider the vector of length n of function values at the sample points: $f = (f(x_{k,1}, x_{\ell,2}) : k, \ell = 1, ..., m)^{\mathrm{T}}$. Let O be the $n \times n$ matrix with columns being the vectors $\xi_{\mu\nu}$, $\mu, \nu = 1, ..., m$. Then $O^{\mathrm{T}}O = nI$. Denote $a = (a_{\mu\nu} : \mu, \nu = 1, ..., m)^{\mathrm{T}} = (1/n)O^{\mathrm{T}}f$, and $z = (z_{\mu\nu} : \mu, \nu = 1, ..., m)^{\mathrm{T}} = (1/n)O^{\mathrm{T}}y$. That is,

$$a_{\mu\nu} = \langle f, \xi_{\mu\nu}\rangle_n, \quad z_{\mu\nu} = \langle y, \xi_{\mu\nu}\rangle_n;$$

then $f \in R^n$ can be expanded in terms of the orthonormal basis:

$$f = \sum_{\mu,\nu} a_{\mu\nu}\xi_{\mu\nu} = f_0 + f_1 + f_2 + f_{12},$$

where $f_0 = a_{11}\xi_{11}$, $f_1 = \sum_{\mu=2}^m a_{\mu 1}\xi_{\mu 1}$, $f_2 = \sum_{\nu=2}^m a_{1\nu}\xi_{1\nu}$, $f_{12} = \sum_{\mu=2}^m \sum_{\nu=2}^m a_{\mu\nu}\xi_{\mu\nu}$. Then all the components f_0, f_1, f_2, f_{12}, are orthogonal in R^n. Furthermore, we have $z_{\mu\nu} = a_{\mu\nu} + \delta_{\mu\nu}$, where $\delta_{\mu\nu} \sim N(0, \sigma^2/n)$, for $\mu \geq 1, \nu \geq 1$.

Now let us consider the COSSO estimate (10):

$$\frac{1}{n}(y - R_\theta c - b1_n)^{\mathrm{T}}(y - R_\theta c - b1_n) + c^{\mathrm{T}}R_\theta c + \lambda\sum_{\alpha=1}^p \theta_\alpha, \quad \text{subject to} \quad \theta_\alpha \geq 0, \alpha = 1, ..., p,$$

where $R_\theta = \sum_{\alpha=1}^p \theta_\alpha R_\alpha$. Let $s = O^{\mathrm{T}}c$, $D_\alpha = (1/n^2)O^{\mathrm{T}}R_\alpha O$. Then D_α is a diagonal matrix with diagonal elements being $r_{\alpha,\mu\nu}/n$, $\alpha = 1, 2$, or 12. The COSSO problem can be written as

$$(z - D_\theta s - (b, 0, ..., 0)^{\mathrm{T}})^{\mathrm{T}}(z - D_\theta s - (b, 0, ..., 0)^{\mathrm{T}}) + s^{\mathrm{T}}D_\theta s + \lambda\sum_{\alpha=1}^p \theta_\alpha,$$

where $D_\theta = \sum_{\alpha=1}^p \theta_\alpha D_\alpha$. It can then be shown by straightforward calculation that, for the minimizing s, b, and θ, $\hat{s}_{11} = 0$, $\hat{b} = z_{11}$, and s and θ minimize

$$\begin{aligned}
&\sum_{\mu\geq 2} [\{z_{\mu 1} - \eta_\mu(\theta_1 + \theta_{12}t)s_{\mu 1}\}^2 + \eta_\mu(\theta_1 + \theta_{12}t)s_{\mu 1}^2] \\
&+\sum_{\nu\geq 2} [\{z_{1\nu} - \eta_\nu(\theta_2 + \theta_{12}t)s_{1\nu}\}^2 + \eta_\nu(\theta_2 + \theta_{12}t)s_{1\nu}^2] \\
&+ \sum_{\mu\geq 2,\nu\geq 2} [(z_{\mu\nu} - \theta_{12}\eta_\mu\eta_\nu s_{\mu\nu})^2 + \theta_{12}\eta_\mu\eta_\nu s_{\mu\nu}^2] + \lambda(\theta_1 + \theta_2 + \theta_{12}).
\end{aligned}$$

Therefore at the minimum, we have

$$\begin{aligned}\hat{s}_{\mu 1} &= \{1+\eta_\mu(\theta_1+\theta_{12}t)\}^{-1}z_{\mu 1}, \quad \mu \geq 2;\\ \hat{s}_{1\nu} &= \{1+\eta_\nu(\theta_2+\theta_{12}t)\}^{-1}z_{1\nu}, \quad \nu \geq 2;\\ \hat{s}_{\mu\nu} &= (1+\eta_\mu\eta_\nu\theta_{12})^{-1}z_{\mu\nu}, \quad \mu \geq 2, \nu \geq 2;\end{aligned}$$

and θ's minimize

$$\begin{aligned}A(\theta_1,\theta_2,\theta_{12}) &= \sum_{\mu\geq 2} z_{\mu 1}^2(1+\eta_\mu\theta_1+\eta_\mu\theta_{12}t)^{-1} + \sum_{\nu\geq 2} z_{1\nu}^2(1+\eta_\nu\theta_2+\eta_\nu\theta_{12}t)^{-1}\\ &+ \sum_{\mu\geq 2,\nu\geq 2} z_{\mu\nu}^2(1+\theta_{12}\eta_\mu\eta_\nu)^{-1} + \lambda(\theta_1+\theta_2+\theta_{12}),\end{aligned}$$

subject to $\theta_1 \geq 0, \theta_2 \geq 0, \theta_{12} \geq 0$. We show that if $f_{12} = 0$ and $n\lambda \to \infty$, then $\hat{\theta}_{12} = 0$ with probability tending to one. The cases for θ_1 and θ_2 are similar. Direct calculation gives $\partial A/\partial\theta_{12} \geq \lambda - U$ for any $\theta_1 \geq 0$, $\theta_2 \geq 0$, and $\theta_{12} \geq 0$, where $U = t\sum_{\mu\geq 2}\eta_\mu z_{\mu 1}^2 + t\sum_{\nu\geq 2}\eta_\nu z_{1\nu}^2 + \sum_{\mu\geq 2,\nu\geq 2}\eta_\mu\eta_\nu z_{\mu\nu}^2$.

When $f_{12} = 0$, we have $a_{\mu\nu} = 0$, for $\mu \geq 2, \nu \geq 2$, and $z_{\mu 1} \sim N(a_{\mu 1}, \sigma^2/n), z_{1\nu} \sim N(a_{1\nu}, \sigma^2/n)$ and $z_{\mu\nu} \sim N(0, \sigma^2/n)$ for $\mu, \nu \geq 2$. Note $\sum_{\mu\geq 2} a_{\mu 1}^2 = ||f_1||_{L_2}^2$, $\sum_{\nu\geq 2} a_{1\nu}^2 = ||f_2||_{L_2}^2$, straightforward calculation gives $\mathrm{E}(U) = O(n^{-1})$ and $\mathrm{var}(U) = O(n^{-2})$. Therefore when $n\lambda \to \infty$, by Chebyshev's inequality,

$$\mathrm{pr}(U > \lambda) \leq \mathrm{pr}(|U - \mathrm{E}(U)| > \lambda - \mathrm{E}(U)) \leq \mathrm{var}(U)/\{\lambda - \mathrm{E}(U)\}^2 \to 0,$$

Hence with probability tending to unity, $U < \lambda$, and $\partial A/\partial\theta_{12} > 0$, for $\theta_1 \geq 0, \theta_2 \geq 0$, and $\theta_{12} \geq 0$. Therefore at the minimizer of A, we have $\hat{\theta}_{12} = 0$, which implies $\hat{f}_{12} = 0$.

That $\hat{f}_{12} \neq 0$ with probability tending to one when $f_{12} \neq 0$ and λ is appropriately chosen follows from the consistency results Theorem 2 and the comments thereafter in Section 2.4.

Acknowledgments. The authors wish to thank Grace Wahba for helpful comments.

References

Bakin, S. (1999). Adaptive regression and model selection in data mining problems. *Ph.D. dissertation, the Australian National University* .

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

Breiman, L. and Friedman, J. (1985). Estimating optimal transformations for multiple regression and correlation. *J. Am. 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.

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

Chen, Z. (1991). Interaction spline models and their convergence rates. *Ann. Statist.* **19** 1855–1868.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. *Numer. Math.* **31** 377–403.

Fan, J. and Li, R. Z. (2001). Variable selection via penalized likelihood. *J. Am. Statist. Assoc.* **96** 1348–1360.

Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35** 109–148.

Friedman, J. H. (1991). Multivariate adaptive regression splines (invited paper). *Ann. Statist.* **19** 1–141.

Gu, C. (1992). Diagnostics for nonparametric regression models with additive term. *J. Am. Statist. Assoc.* **87** 1051–1058.

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

Gu, C. and Xiang, D. (2001). Cross-validating non-gaussian data: Generalized approximate cross-validation revisited. *J. Comp. Graph. Statist.* **10** 581–591.

Gunn, S. R. and Kandola, J. S. (2002). Structural modeling with sparse kernels. *Mach. Learning* **48** 115–136.

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

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

Lin, X., Wahba, G., Xiang, D., Gao, F., Klein, R. and Klein, B. (2000). Smoothing spline ANOVA models for large data sets with Bernoulli observations and the randomized GACV. *Ann. Statist.* **28** 1570–1600.

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

Nychka, D. (1988). Bayesian confidence intervals for smoothing splines. *J. Am. Statist. Assoc.* **83** 1134–1143.

Ruppert, D. and Carroll, R. (2000). Spatially-adaptive penalties for spline fitting. *Australian and New Zealand Journal of Statistics* **45** 205–223.

Tapia, R. and Thompson, J. (1978). *Nonparametric Probability Density Estimation*. Baltimore, MD: Johns Hopkins University Press.

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

Utreras, F. (1983). Natural spline functions: their associated eigenvalue problem. *Numeri. Math.* **42** 107–117.

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

Wahba, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. R. Statist. Soc. B* **45** 133–150.

Wahba, G. (1990). *Spline Models for Observational Data*, vol. 59. SIAM. 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.

Wahba, G. and Wold, S. (1975). A completely automatic French curve. *Commun. Statist.* **4** 1–17.

Xiang, D. and Wahba, G. (1996). A generalized approximate cross validation for smoothing splines with non-Gaussian data. *Statist. Sinica* **6** 675–692.

Yau, P., Kohn, R. and Wood, S. (2002). Bayesian variable selection and model averaging in high dimensional multinomial nonparametric regression. *J. Comp. Graph. Statist.* To appear.

Zhang, H., Wahba, G., Lin, Y., Voelker, M., Ferris, M., Klein, R. and Klein, B. (2002). Variable selection and model building via likelihood basis pursuit. *Technical report, University of Wisconsin, Madison* .