Statistica Sinica **16**(2006), 1021-1041

COMPONENT SELECTION AND SMOOTHING FOR NONPARAMETRIC REGRESSION IN EXPONENTIAL FAMILIES

Hao Helen Zhang and Yi Lin

North Carolina State University and University of Wisconsin at Madison

Abstract: We propose a new penalized likelihood method for model selection and nonparametric regression in exponential families. In the framework of smoothing spline ANOVA, our method employs a regularization with the penalty functional being the sum of the reproducing kernel Hilbert space norms of functional components in the ANOVA decomposition. It generalizes the LASSO in the linear regression to the nonparametric context, and conducts component selection and smoothing simultaneously. Continuous and categorical variables are treated in a unified fashion. We discuss the connection of the method to the traditional smoothing spline penalized likelihood estimation. We show that an equivalent formulation of the method leads naturally to an iterative algorithm. Simulations and examples are used to demonstrate the performances of the method.

Key words and phrases: Exponential family, LASSO, nonparametric regression, penalized likelihood, smoothing spline ANOVA.

1. Introduction

We consider the nonparametric regression and model selection problem in the exponential family framework. Suppose we are interested in predicting a response variable Y given d-dimensional input $\mathbf{X} = (X^{(1)}, \ldots, X^{(d)})$. Conditional on $\mathbf{X} = \mathbf{x}$, assume Y follows an exponential family distribution with the canonical density form

$$\exp\{yf(\mathbf{x}) - B(f(\mathbf{x})) + C(y)\}, \tag{1.1}$$

where B and C are known functions. The goal of the regression problem is to estimate $f(\mathbf{x})$ based on an independently and identically distributed sample $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$. In many practical situations, the number of input variables d is large and some of the input variables are superfluous. In such situations, effective variable selection can improve both the accuracy and the interpretability of the estimated model. Many variable selection methods have been proposed for the commonly studied linear regression model. Traditional methods include forward selection, backward elimination, and best subset selection. Recent developments include the nonnegative garotte (Breiman (1995)), LASSO (Tibshirani (1996)), SCAD (Fan and Li (2001)) and LARS (Efron, Hastie, Johnstone and Tibshirani (2004)). These methods conduct variable selection and coefficient shrinkage at the same time, and improve on the traditional methods in terms of estimation accuracy and stability of the solution.

The nonparametric regression model allows more flexibility than the linear model and is the topic of this paper. Many popular proposals for variable selection and estimation in nonparametric regression, such as CART (Breiman, Friedman, Olshen and Stone (1984)), TURBO (Friedman and Silverman (1989)), BRUTO (Hastie (1989)) and MARS (Friedman (1991)), use greedy search type algorithm for variable selection. This is similar to forward selection and backward elimination in the linear regression. The greedy search type algorithm can suffer from being myopic since it looks only one step ahead, thus may not take the globally optimal step. Given the successes of global penalized likelihood methods such as the LASSO and the SCAD in the linear model, it is desirable to develop global algorithms based on penalized likelihood for nonparametric regression models.

The smoothing spline ANOVA model (SS-ANOVA) provides a general framework for high dimensional function estimation, and has been successfully applied to many practical problems. See Wahba (1990), Wahba, Wang, Gu, Klein and Klein (1995) and Gu (2002). In this paper we consider a method of regularization in the SS-ANOVA model with the penalty being the sum of functional component norms. The proposed penalized likelihood method conducts simultaneous model selection and estimation. In the Gaussian regression context, regularization with this type of penalty has been studied by Lin and Zhang (2002) and was referred to as the COSSO penalty. In this paper we consider the more general setting of exponential family regression. This general framework allows the treatment of non-normal responses, binary and polychotomous responses, and event counts data.

Several different methods have been proposed for variable selection in the SS-ANOVA models. In the Gaussian regression setting, Gu (1992) proposed using cosine diagnostics as model checking tools after model fitting. For regression in exponential families, Zhang, Wahba, Lin, Voelker, Ferris, Klein and Klein (2004) proposed the likelihood basis pursuit which conducts variable selection by imposing the L_1 penalty on coefficients of basis functions. There is also related work in the machine learning literature in the framework of combining kernels. Bach, Lanckriet and Jordan (2004) and Bach, Thibaux and Jordan (2004) consider block 1-norm regularization for learning a sparse conic combination of kernels. Their formulation can also be used for variable selection in a nonparametric setting, and gave rise to a variational problem that is similar to ours. They developed an interesting method for approximately computing the regularization path.

This paper is organized as follows. Section 2 introduces the formulation of the COSSO-type penalized likelihood method. The existence of the estimator is established, and we also show that the penalized likelihood estimate has a finite representation form. Section 3 develops the computational algorithm. Simulations and examples are presented in Section 4 and Section 5. A summary is given in Section 6. The proofs of the theorems are given in the appendix.

2. Penalized Likelihood Method

2.1. Smoothing spline ANOVA model

The functional ANOVA decomposition of a multivariate function f is

$$f(\mathbf{x}) = b + \sum_{j=1}^{d} f_j(x^{(j)}) + \sum_{j=1}^{d} \sum_{k=j+1}^{d} f_{jk}(x^{(j)}, x^{(k)}) + \cdots, \tag{2.1}$$

where b is a constant, f_j's are the main effects, f_{jk}'s are the two-way interactions, and so on. The identifiability of the terms in (2.1) is assured by side conditions through averaging operators. In the smoothing spline ANOVA model, we assume $f_j \in H^{(j)}$, where $H^{(j)}$ is a reproducing kernel Hilbert space (RKHS) of functions of $x^{(j)}$, admitting an orthogonal decomposition $H^{(j)} = \{1\} \oplus \bar{H}^{(j)}$. The full function space is the tensor product space

$$\otimes_{j=1}^{d} H^{(j)} = \{1\} \oplus \left[\bigoplus_{j=1}^{d} \bar{H}^{(j)} \right] \oplus \left[\bigoplus_{j<k} (\bar{H}^{(j)} \otimes \bar{H}^{(k)}) \right] \oplus \cdots . \tag{2.2}$$

Each functional component in the SS-ANOVA decomposition (2.1) lies in a subspace in the orthogonal decomposition (2.2) of $\otimes_{j=1}^{d} H^{(j)}$. In the application of the SS-ANOVA model, usually only lower order interactions are retained in the decomposition for easy computation and interpretability. Correspondingly, the function space assumed for the SS-ANOVA model is a subspace $\mathcal{F}$ of $\otimes_{j=1}^{d} H^{(j)}$. We write $\mathcal{F}$ as

$$\mathcal{F} = \{1\} \oplus_{\alpha=1}^{p} \mathcal{F}^{\alpha}, \tag{2.3}$$

where $\mathcal{F}^1, \ldots, \mathcal{F}^p$ are p orthogonal subspaces of $\mathcal{F}$. For the additive model, $p = d$ and the $\mathcal{F}^{\alpha}$'s are the main effect subspaces. For the two-way interaction model, $p = d(d+1)/2$ and the $\mathcal{F}^{\alpha}$'s represent the main effect and two-way interaction subspaces. $\mathcal{F}$ is an RKHS with the norm $\|\cdot\|$ induced by the norm in $\otimes_{j=1}^{d} H^{(j)}$.

When $x^{(j)}$ is a continuous covariate, a typical example of $H^{(j)}$ is the second-order Sobolev Hilbert space $W_2[0,1] = \{h : h, h'$ are absolutely continuous, $h'' \in \mathcal{L}_2[0,1]\}$. The norm in $W_2[0,1]$ is

$$\|h\|^2 = \left\{ \int_0^1 h(t)dt \right\}^2 + \left\{ \int_0^1 h'(t)dt \right\}^2 + \int_0^1 \{h''(t)\}^2 dt.$$

The reproducing kernel of $W_2[0,1]$ is $K(s,t) = 1 + \bar{K}(s,t)$, where $\bar{K}(s,t) = k_1(s)k_1(t) + k_2(s)k_2(t) - k_4(|s-t|)$, $k_1(t) = t - 1/2$, $k_2(t) = \{k_1^2(t) - 1/12\}/2$, and $k_4(t) = \{k_1^4(t) - k_1^2(t)/2 + 7/240\}/24$. See Wahba (1990) and Gu (2002) for more on reproducing kernels.

When $x^{(j)}$ is a categorical covariate taking values on the discrete domain $\mathcal{X} = \{1, \ldots, L\}$, a function on $\mathcal{X}$ is simply an L-vector and the evaluation functional is the coordinator extraction. We define the squared norm of such an L-vector to be $1/L$ of the common Euclidean space squared norm. This definition ensures that functions on categorical variables with different numbers of categories have comparable norms. Under this norm we have $H^{(j)} = \{1\} \oplus \bar{H}^{(j)}$, where $\bar{H}^{(j)}$ is an RKHS with reproducing kernel $\bar{K}(s,t) = L\delta(s,t) - 1$. Here $\delta(s,t) = 1$ if $s = t$; $= 0$ otherwise.

2.2. COSSO penalized likelihood method

Let $l\{y, \eta\} = y\eta - B(\eta)$, the log likelihood corresponding to the exponential family distribution (1.1). Define the functional

$$L(f) = \frac{1}{n} \sum_{i=1}^{n} \big[- l\{y_i, f(\mathbf{x}_i)\} \big]. \tag{2.4}$$

The proposed COSSO penalized likelihood method solves

$$\min_{f \in \mathcal{F}} \; L(f) + \tau^2 J(f), \quad \text{with} \quad J(f) = \sum_{\alpha=1}^{p} \|P^{\alpha} f\|, \tag{2.5}$$

where $P^{\alpha} f$ is the orthogonal projection of f onto $\mathcal{F}^{\alpha}$, and $\tau > 0$ is the smoothing parameter. In the important special case of additive models, (2.5) becomes

$$\min_{f \in \mathcal{F}} \; L(f) + \tau^2 \sum_{j=1}^{d} \|f_j\|, \quad \text{with} \quad f(\mathbf{x}) = b + \sum_{j=1}^{d} f_j(x^{(j)}).$$

The penalty $J(f)$ in (2.5) is a sum of RKHS norms, instead of the squared RKHS norm penalty employed in smoothing splines. The LASSO in linear models can be seen as a special case of $J(f)$. For the input space $\mathcal{X} = [0,1]^d$, consider the linear function space $\mathcal{F} = \{1\} \oplus \{x^{(1)} - 1/2\} \oplus \cdots \oplus \{x^{(d)} - 1/2\}$, with the usual L_2 inner product on $\mathcal{F}$: $(f, g) = \int_{\mathcal{X}} fg$. The penalty term in (2.5) 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.

The functional $J(f)$ is convex in f. The existence of the COSSO penalized likelihood estimate is established under the following assumptions:

(i) $\rho_i \equiv \sup_{\eta} l(y_i, \eta) < \infty$, for any $i \in \{1, \ldots, n\}$;

(ii) there is a unique minimizer of $L(\eta)$ over $\eta \in R$.

Theorem 2.1. *Let* $(\mathbf{x}_i, y_i)$*,* $i = 1, \ldots, n$*, be i.i.d. pairs, and suppose* $y_i|\mathbf{x}_i$ *has an exponential family distribution of the form* (1.1)*. Then under assumptions* (i) *and* (ii)*, for any reproducing kernel Hilbert space* $\mathcal{F}$ *of functions with the decomposition* (2.3)*, there exists a minimizer of* (2.5) *over* $\mathcal{F}$*.*

Assumption (i) states that the likelihood based on a single observation is bounded from above. This is a necessary condition for the existence of a maximum likelihood estimate based on a single observation, and is satisfied by most commonly encountered exponential family distributions, including the binomial, Poisson, negative binomial, Gaussian, and gamma distribution with fixed shape parameter. In the case of a categorical variable Y, this assumption is always satisfied since the likelihood is less or equal to one. Assumption (ii) is usually satisfied in practical situations by the commonly used exponential families. One situation in which (ii) is violated occurs when we have a Bernoulli family and all y_i's happen to be one. In this extreme case, the minimizer is $\eta = \infty$, corresponding to the probability parameter being one (η is the log odds, the natural exponential family parameter for the Bernoulli family). Gu (2002) presented a proof for the existence of smoothing spline estimate under assumption (ii) with a general continuous and convex functional L. Gu's proof was based on two lemmas. The proof of the first lemma can be modified to accommodate the COSSO penalized likelihood estimate and is incorporated into our proof of Theorem 2.1. However, the proof of the second lemma is not applicable to the COSSO estimate. From our proof of Theorem 2.1, we can see that assumption (ii) can be relaxed to the existence of a minimizer b_0 of $L(\eta)$ over R, and that there exist $b_1 > b_0$ and $b_2 < b_0$ such that $L(b_0) < L(b_1)$ and $L(b_0) < L(b_2)$.

The following theorem shows that the solution to (2.5) lies in a finite dimensional space. The proof is similar to that of the representer theorem for smoothing splines (Kimeldorf and Wahba (1971)).

Theorem 2.2. *Let the minimizer of* (2.5) *be* $\hat{f} = \hat{b} + \sum_{\alpha=1}^{p} \hat{f}_\alpha$*, with* $\hat{f}_\alpha \in \mathcal{F}^\alpha$*. Then* $\hat{f}_\alpha \in$ *span*$\{R_\alpha(\mathbf{x}_i, \cdot), i = 1, \ldots, n\}$*, where* $R_\alpha(\cdot, \cdot)$ *is the reproducing kernel of* $\mathcal{F}^\alpha$*.*

2.3. Equivalent formulation

It is possible to compute the solution to (2.5) by using Theorem 2.2. Here we give an equivalent formulation of (2.5) that leads naturally to an iterative algorithm. Define $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^{\mathrm{T}}$ and let $\mathbf{0}$ be the vector of zeros. Consider

$$\min_{f \in \mathcal{F}, \boldsymbol{\theta} \geq \mathbf{0}} \frac{1}{n} \sum_{i=1}^{n} \big[-l\{y_i, f(\mathbf{x}_i)\} \big] + \lambda_0 \sum_{\alpha=1}^{p} \theta_\alpha^{-1} \|P^\alpha f\|^2 + \lambda \sum_{\alpha=1}^{p} \theta_\alpha, \tag{2.6}$$

where $\lambda_0 > 0$ is a constant, and λ is the smoothing parameter. If $\theta_\alpha = 0$, then the minimizer is taken to satisfy $\|P^\alpha f\|^2 = 0$. We take the convention $0/0 = 0$ throughout this paper.

Lemma 2.1. *Set* $\lambda = \tau^4/(4\lambda_0)$. (i) *If* $\hat{f}$ *minimizes* (2.5), *setting* $\hat{\theta}_\alpha = \lambda_0^{1/2}\lambda^{-1/2}\|P^\alpha \hat{f}\|$ *for* $\alpha = 1, \ldots, p$, *then the pair* $(\hat{\boldsymbol{\theta}}, \hat{f})$ *minimizes* (2.6). (ii) *On the other hand, if a pair* $(\hat{\boldsymbol{\theta}}, \hat{f})$ *minimizes* (2.6), *then* $\hat{f}$ *minimizes* (2.5).

The objective function in (2.6) is similar to that for a smoothing spline with multiple smoothing parameters, except for the additional penalty on θ's. This shows the connection between our method and smoothing splines. Note that λ is the only smoothing parameter in (2.6). Generally λ_0 is fixed at some value for computational convenience.

3. Algorithms

For fixed λ_0 and λ, the representer theorem for the smoothing spline states that the minimizer of (2.6) has the form $f(\mathbf{x}) = b + \sum_{i=1}^n c_i R_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{x})$, where $R_{\boldsymbol{\theta}} = \sum_{\alpha=1}^p \theta_\alpha R_\alpha$ and R_α is the reproducing kernel of $\mathcal{F}^\alpha$. With some abuse of notation, we use R_α for the matrix $\{R_\alpha(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1}^n$ and $R_{\boldsymbol{\theta}}$ for the matrix $\sum_{\alpha=1}^p \theta_\alpha R_\alpha$. Let $\boldsymbol{c} = (c_1, \ldots, c_n)^{\mathrm{T}}$, $\boldsymbol{f} = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))^{\mathrm{T}}$, and $\mathbf{1}_n$ be the vector of ones of length n. Then $\sum_{\alpha=1}^p \theta_\alpha^{-1}\|P^\alpha f\|^2 = \sum_{\alpha=1}^p \theta_\alpha \boldsymbol{c}^{\mathrm{T}} R_\alpha \boldsymbol{c} = \boldsymbol{c}^{\mathrm{T}} R_{\boldsymbol{\theta}} \boldsymbol{c}$, and (2.6) becomes

$$\min_{\boldsymbol{\theta} \geq \mathbf{0}, b, \boldsymbol{c}} \frac{1}{n} \sum_{i=1}^n \big[-l\{y_i, f(\mathbf{x}_i)\}\big] + \lambda_0 \boldsymbol{c}^{\mathrm{T}} R_{\boldsymbol{\theta}} \boldsymbol{c} + \lambda \mathbf{1}_p^{\mathrm{T}} \boldsymbol{\theta}. \tag{3.1}$$

We solve (3.1) using Newton-Raphson iteration. Given a current solution f^0, the conditional mean of Y given $\mathbf{x}_i$ is $\mu_i^0 = \dot{B}(f^0(\mathbf{x}_i))$ and the conditional variance is $V_i^0 = \ddot{B}(f^0(\mathbf{x}_i))$. Define $\nu_i = -y_i + \mu_i^0$ and $w_i = V_i^0$. The second-order Taylor expansion of $-y_i f(\mathbf{x}_i) + B(f(\mathbf{x}_i))$ at $f^0(\mathbf{x}_i)$ is

$$\begin{aligned} & -y_i f^0(x_i) + B(f^0(\mathbf{x}_i)) + \nu_i \big[f(\mathbf{x}_i) - f^0(\mathbf{x}_i)\big] + \frac{1}{2} w_i \big[f(\mathbf{x}_i) - f^0(\mathbf{x}_i)\big]^2 \\ = & \frac{1}{2} w_i \big[f(\mathbf{x}_i) - f^0(\mathbf{x}_i) + \frac{\nu_i}{w_i}\big]^2 + \beta_i, \end{aligned}$$

where β_i is independent of $f(\mathbf{x}_i)$. Define the adjusted dependent variable $z_i = f^0(\mathbf{x}_i) + (y_i - \mu_i^0)/w_i$, $\boldsymbol{z} = (z_1, \ldots, z_n)^{\mathrm{T}}$, and the weight matrix $W = \mathrm{diag}[w_1, \ldots, w_n]$. The Newton iteration update of (3.1) is to solve

$$\min_{\boldsymbol{\theta} \geq \mathbf{0}, b, \boldsymbol{c}} (\boldsymbol{z} - b\mathbf{1}_n - R_{\boldsymbol{\theta}} \boldsymbol{c})^{\mathrm{T}} W (\boldsymbol{z} - b\mathbf{1}_n - R_{\boldsymbol{\theta}} \boldsymbol{c}) + n\lambda_0 \boldsymbol{c}^{\mathrm{T}} R_\theta \boldsymbol{c} + n\lambda \mathbf{1}_p^{\mathrm{T}} \boldsymbol{\theta}. \tag{3.2}$$

The first part of (3.2) is a weighted least squares with the weights changing in each iteration. This iteratively-reweighted least squares procedure is commonly used for computing maximal likelihood estimates in generalized linear and additive models (Hastie and Tibshirani (1990)). We propose to minimize (3.2) by alternatively solving $(b, \boldsymbol{c})$ with $\boldsymbol{\theta}$ fixed and solving $\boldsymbol{\theta}$ with $(b, \boldsymbol{c})$ fixed. When $\boldsymbol{\theta}$ is fixed, the minimization problem is equivalent to solving the standard smoothing spline. When $(b, \boldsymbol{c})$ is fixed, we need to solve a quadratic programming (QP) under linear constraints. A similar algorithm was used in Lin and Zhang (2002) for Gaussian regression. In practice, one-step update is often sufficient to get a good approximate solution.

3.1. Full basis algorithm

1. With $\boldsymbol{\theta}$ fixed, solving (3.2) is equivalent to solving the standard smoothing spline
$$\min_{b,\boldsymbol{c}_w} ||\boldsymbol{z}_w - b\boldsymbol{s}_w - R_{w\boldsymbol{\theta}}\boldsymbol{c}_w||^2 + n\lambda_0 \boldsymbol{c}_w^{\mathrm{T}} R_{w\boldsymbol{\theta}}\boldsymbol{c}_w, \tag{3.3}$$
where $\boldsymbol{z}_w = W^{1/2}\boldsymbol{z}$,$R_{w\boldsymbol{\theta}} = W^{1/2}R_{\boldsymbol{\theta}}W^{1/2}$,$\boldsymbol{c}_w = W^{-1/2}\boldsymbol{c}$, and $\boldsymbol{s}_w = W^{1/2}\mathbf{1}_n$.
2. With $(b, \boldsymbol{c})$ fixed, (3.2) becomes
$$\min_{\boldsymbol{\theta}\geq\mathbf{0}} (\boldsymbol{z} - b\mathbf{1}_n - G\boldsymbol{\theta})^{\mathrm{T}} W (\boldsymbol{z} - b\mathbf{1}_n - G\boldsymbol{\theta}) + n\lambda_0 \boldsymbol{c}^{\mathrm{T}} G\boldsymbol{\theta} + n\lambda \mathbf{1}_p^{\mathrm{T}}\boldsymbol{\theta}, \tag{3.4}$$
where $G = [\boldsymbol{g}_1, \ldots, \boldsymbol{g}_p]$ with $\boldsymbol{g}_\alpha = R_\alpha \boldsymbol{c}$. Define $G_w = W^{1/2}G$ and $\boldsymbol{u}_w = \boldsymbol{z}_w - b\boldsymbol{s}_w - (n/2)\lambda_0\boldsymbol{c}_w$. It is easy to show that (3.4) is equivalent to, for some $M \geq 0$,
$$\min_{\boldsymbol{\theta}} ||\boldsymbol{u}_w - G_w\boldsymbol{\theta}||^2, \quad \text{subject to} \quad \mathbf{1}^{\mathrm{T}}\boldsymbol{\theta} \leq M, \quad \boldsymbol{\theta} \geq \mathbf{0}. \tag{3.5}$$

The tuning parameter M in (3.5) is equivalent to λ in (3.4). Note (3.5) has the same formulation as the nonnegative garrote approach (Breiman (1995)). The following gives the complete algorithm for solving the COSSO with fixed λ_0 and M. The issue of tuning parameter is very important and will be discussed later.

Algorithm 1.
Step 1: Initialize $f_i = \bar{y}$, $\mu_i = \dot{B}(f_i)$, $w_i = \ddot{B}(f_i)$, $z_i = f_i + (y_i - \mu_i)/w_i$ for $i = 1, \ldots, n$.
Step 2: Set $\boldsymbol{\theta} = \mathbf{1}_p$. Calculate $\boldsymbol{z}_w$, $\boldsymbol{s}_w$ and $R_{w\boldsymbol{\theta}}$. Solve (3.3) for $(b, \boldsymbol{c}_w)$.
Step 3: Calculate $\boldsymbol{u}_w = \boldsymbol{z}_w - bs_w - \frac{n}{2}\lambda_0\boldsymbol{c}_w$ and G_w. Solve (3.5) for $\boldsymbol{\theta}$.
Step 4: Update $R_{w\boldsymbol{\theta}}$ and solve (3.3) for $(b, \boldsymbol{c}_w)$ using current $\boldsymbol{\theta}$.
Step 5: Calculate $\boldsymbol{c} = W^{1/2}\boldsymbol{c}_w$ and $\boldsymbol{f} = b\mathbf{1}_n + R_{\boldsymbol{\theta}}\boldsymbol{c}$. Update the w_i's and z_i's.
Step 6: Go to step 2, until the convergence criterion is satisfied.

3.2. Subset basis algorithm

The computational complexity of the full basis algorithm in each iteration is $O(n^3)$. For large data sets, the implementation of the full basis algorithm can be slow. We propose an alternative subset basis algorithm to speed-up the computation. The idea is to minimize the objective function of (2.6) in a subspace spanned by some pre-selected N basis functions. This parsimonious basis approach has been used by Xiang and Wahba (1998), Ruppert and Carroll (2000), and Yau, Kohn and Wood (2002). In each iteration, the computational complexity of the subset basis algorithm with N basis functions is $O(nN^2)$.

Randomly take N points $\{\mathbf{x}_{1*}, \ldots, \mathbf{x}_{N*}\}$ from the data and use them to generate N basis functions. We will search for the minimizer of (2.6) in the subspace spanned by these basis functions. Then the solution has the form

$$f(\mathbf{x}) = b + \sum_{i=1}^{N} c_i R_{\boldsymbol{\theta}}(\mathbf{x}_{i*}, \mathbf{x}) = b + \sum_{i=1}^{N} c_i \sum_{\alpha=1}^{p} \theta_\alpha R_\alpha(\mathbf{x}_{i*}, \mathbf{x}). \tag{3.6}$$

Let $\boldsymbol{c} = (c_1, \ldots, c_N)^{\mathrm{T}}, R^{**}_\alpha$ be the matrix $\{R_\alpha(\mathbf{x}_{i*}, \mathbf{x}_{k*})\}_{i,k=1}^{N}$, and R^*_α be the matrix $\{R_\alpha(\mathbf{x}_i, \mathbf{x}_{k*})\}$, $i = 1, \ldots, n$ and $k = 1, \ldots, N$. Define $R^{**}_{\boldsymbol{\theta}} = \sum_{\alpha=1}^{p} \theta_\alpha R^{**}_\alpha$ and $R^*_{\boldsymbol{\theta}} = \sum_{\alpha=1}^{p} \theta_\alpha R^*_\alpha$. Then $\boldsymbol{f} = b\mathbf{1}_n + R^*_{\boldsymbol{\theta}}\boldsymbol{c}$, and (3.2) becomes

$$\min_{\boldsymbol{\theta} \geq \mathbf{0}, b, \boldsymbol{c}} (\boldsymbol{z} - b\mathbf{1}_n - R_{\boldsymbol{\theta}}\boldsymbol{c})^{\mathrm{T}} W (\boldsymbol{z} - b\mathbf{1}_n - R_{\boldsymbol{\theta}}\boldsymbol{c}) + n\lambda_0 \boldsymbol{c}^{\mathrm{T}} R^{**}_{\boldsymbol{\theta}} \boldsymbol{c} + n\lambda \mathbf{1}_p^{\mathrm{T}} \boldsymbol{\theta}. \tag{3.7}$$

Define the vectors $\boldsymbol{z}_w = W^{1/2}\boldsymbol{z}, \boldsymbol{s}_w = W^{1/2}\mathbf{1}_n, \boldsymbol{u}_w = \boldsymbol{z}_w - b\boldsymbol{s}_w$, and $R^*_{w\boldsymbol{\theta}} = W^{1/2}R^*_{\boldsymbol{\theta}}$.

1. With $\boldsymbol{\theta}$ fixed, (3.7) is equivalent to

$$\min_{b, \boldsymbol{c}} ||\boldsymbol{z}_w - b\boldsymbol{s}_w - R^*_{w\boldsymbol{\theta}}\boldsymbol{c}||^2 + n\lambda_0 \boldsymbol{c}^{\mathrm{T}} R^{**}_{\boldsymbol{\theta}} \boldsymbol{c}. \tag{3.8}$$

2. With $(b, \boldsymbol{c})$ fixed, let $G_w = [\boldsymbol{g}_{w1}, \ldots, \boldsymbol{g}_{wp}]$ with $\boldsymbol{g}_{w\alpha} = R^*_{w\alpha}\boldsymbol{c}$, and $\boldsymbol{h}_{\lambda_0}$ be the vector with the αth element $n\lambda_0 \boldsymbol{c}^{\mathrm{T}} R^{**}_\alpha \boldsymbol{c}$, $\alpha = 1, \ldots, p$. We need to solve

$$\min_{\boldsymbol{\theta}} ||\boldsymbol{u}_w - G_w\boldsymbol{\theta}||^2 + \boldsymbol{h}_{\lambda_0}^{\mathrm{T}}\boldsymbol{\theta}, \quad \text{subject to} \quad \mathbf{1}_p^{\mathrm{T}}\boldsymbol{\theta} \leq M, \quad \boldsymbol{\theta} \geq \mathbf{0}. \tag{3.9}$$

Algorithm 2. In Algorithm 1, replace (3.3) by (3.8), and replace (3.5) by (3.9).

When n is large, the subset basis algorithm can be much more efficient than the full basis algorithm when N is properly chosen. In the standard smoothing spline setting, Gu and Kim (2001) showed that N can be much smaller than n without degrading the performance of the estimation. This is also true for our method and can be seen from the numerical results. In practice, we suggest using the subset basis algorithm when n is large. We use the simple random subsampling scheme to select N basis points. Alternatively, a cluster algorithm in Xiang and Wahba (1998) can be used.

3.3. Smoothing parameter selection

Smoothing parameters balance the tradeoff between the likelihood fit and the penalty on function components. For an exponential family, the Kullback-Leibler (KL) distance between the distributions parameterized by f and $\hat{f}$ is $\mathrm{KL}(f, \hat{f}) = (1/n)\sum_{i=1}^{n}\big[\mu(\mathbf{x}_i)\{f(\mathbf{x}_i) - \hat{f}(\mathbf{x}_i)\} - \{B(f(\mathbf{x}_i)) - B(\hat{f}(\mathbf{x}_i))\}\big]$. Comparative KL distance is obtained by dropping terms that do not involve $\hat{f}$,

$$\mathrm{CKL}(f, \hat{f}) = \frac{1}{n}\sum_{i=1}^{n}\big[-\mu(\mathbf{x}_i)\hat{f}(\mathbf{x}_i) + B\{\hat{f}(\mathbf{x}_i)\}\big].$$

Since μ is unknown, we use five-fold cross validation (CV) to tune the smoothing parameter M in (3.5) or (3.9) adaptively. The grid search is applied. Here is the complete algorithm for the COSSO penalized likelihood fitting and tuning.

Step 1: Set $\boldsymbol{\theta} = \mathbf{1}_p$. Initialize f_i, μ_i, w_i, z_i, W in the same way as in Step 1 of Algorithm 1. For each fixed λ_0, repeat the following two steps.

(i) Calculate $\boldsymbol{z}_w, \boldsymbol{s}_w, R_{w\boldsymbol{\theta}}$ and solve (3.3) or (3.8) for $(b, \boldsymbol{c}_w)$.

(ii) Compute $\boldsymbol{f} = b\mathbf{1}_n + R_{\boldsymbol{\theta}}\boldsymbol{c}$. Update the w_i's and z_i's. Go to (i) until convergence.

Step 2: Choose the best λ_0 using the CV score, and fix it in all the later steps.

Step 3: For each fixed M in a reasonable range, implement Algorithm 1 (or 2 for large datasets). Choose the best M according to the CV score.

4. Simulations

We illustrate the performances of the proposed method with Bernoulli examples. Given $\mathbf{x}$, suppose the binary response Y takes value 1 with probability $p(\mathbf{x})$. To measure the estimation accuracy, the CKL distance between $\hat{p}(\mathbf{x})$ and $p(\mathbf{x})$ is calculated. We compute the expected misclassification rate (EMR) of the model by evaluating it on $10,000$ testing points. For comparison, the Bayes error for each example is reported. It is the classification error of the optimal classification rule based on the true p. The subset basis algorithm is used when $N < n$, and the full basis algorithm is used when $N = n$. We simulate 100 data sets for each example and report the average CKL, EMR, and model size. To give a sense of computational cost, we also report the time for one typical run (the total time on tuning and model fitting) for each example.

Example 1. Consider an additive model with ten continuous covariates independently generated from Unif$[0,1]$. The true logit function is

$$f(\mathbf{x}) = 3x^{(1)} + \pi \sin(\pi x^{(2)}) + 8(x^{(3)})^5 + \frac{2}{e-1}e^{x^{(4)}} - 6.$$

Thus $X^{(5)}, \ldots, X^{(10)}$ are uninformative. The sample size is 250. The Bayes classification error for this example is 0.216. We fit the additive model with the full basis algorithm, and the subset basis algorithms with different numbers of bases: $N = 25, 50, 100$. Figure 4.1 plots the true important functional components and their estimates fitted using the subset algorithm with $N = 100$. The 5th, 50th, 95th best estimates over 100 runs are ranked according to their EMR values. Notice the components are centered in the functional ANOVA decomposition. We can see that our method provides very good estimates for the important functional components.

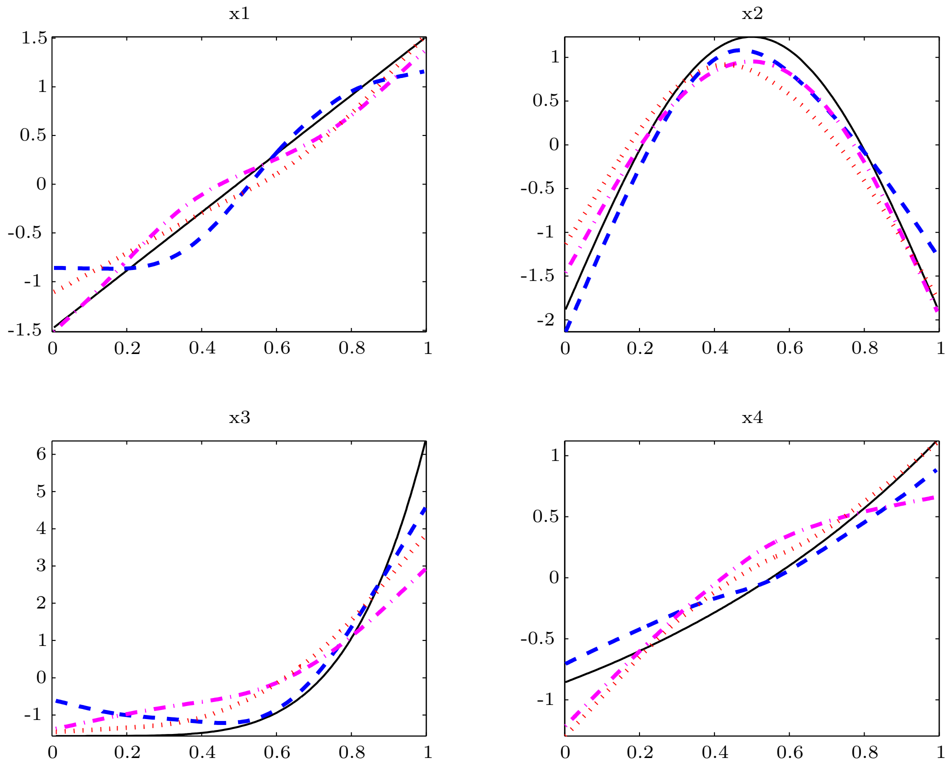

Figure 4.1. True (solid lines) and estimated component functions (dashed lines for the 5th best, dashed-dotted lines for the 50th best, dotted lines for the 95th best) over 100 runs. The additive model is fitted using the subset basis algorithm with $N = 100$.

Table 1 summarizes the average CKL and EMR of the fitted model over 100 runs. The values in the parentheses are the standard errors of the corresponding mean values. Table 2 shows the appearance frequency of each variable, the average model size, with the standard deviation of model size in parentheses. All the algorithms give similar results in terms of model estimation accuracy and variable selection. The method almost always selects $X^{(1)}$, $X^{(2)}$, $X^{(3)}$ (in more than 99% of the runs), and selects $X^{(4)}$ in at least 91% of the runs.

Table 1. Average CKL and EMR of the COSSO penalized likelihood estimates (Example 1).

N	CKL	EMR	Time
25	0.476 (0.018)	0.235 (0.018)	23.4
50	0.477 (0.017)	0.236 (0.016)	36.9
100	0.478 (0.017)	0.235 (0.016)	80.4
250 (full)	0.480 (0.016)	0.238 (0.011)	347.3

Table 2. Appearance frequency of the variables and the average model size (Example 1).

N	1	2	3	4	5	6	7	8	9	10	model size
25	99	99	100	91	10	8	11	9	13	14	4.54 (1.36)
50	100	100	100	92	14	10	16	12	13	16	4.73 (1.40)
100	100	100	100	93	12	10	17	12	14	17	4.75 (1.42)
250	100	100	100	93	12	10	17	12	14	17	4.75 (1.42)

Example 2. Consider situations where there exists some correlation among the covariates. Four functions on $[0,1]$ will be used to build the true underlying regression functions: $g_1(t) = t, g_2(t) = (2t-1)^2, g_3(t) = \sin(2\pi t)/(2-\sin(2\pi t))$, and $g_4(t) = 0.1\sin(2\pi t)+0.2\cos(2\pi t)+0.3\sin^2(2\pi t)+0.4\cos^3(2\pi t)+0.5\sin^3(2\pi t)$. Let $d = 10$ and the true logit function be $f(\mathbf{x}) = 5g_1(x^{(1)})+3g_2(x^{(2)})+4g_3(x^{(3)})+6g_4(x^{(4)})$. We consider two types of covariance structure.

(trimmed) AR(1): Generate $W_1,\ldots,W_{10}$ independently from $N(0,1)$. Define $X^{(1)} = W_1, X^{(j)} = \rho X^{(j-1)} + (1-\rho^2)^{1/2}W_j$ for $j = 2,\ldots,10$. Trim X's to $[-2.5, 2.5]$.

Compound Symmetry (CS): Generate $W_1,\ldots,W_{10}$, and U independently from Unif$[0,1]$. Define $X^{(j)} = (W_j + tU)/(1+t)$, for $j = 1,\ldots,10$. The constant $t \geq 0$ controls the degree of correlation, and $\mathrm{corr}(X^{(j)}, X^{(k)}) = t^2/(1+t^2)$ for any pair $j \neq k$.

1. Uncorrelated input variables: (trimmed) AR(1) with $\rho = 0$.

Tables 3 and 4 summarize the results from 100 simulations when there is no correlation among the covariates. The Bayes error is 0.134. We observe that, as the sample size increases, both CKL and EMR of the fitted model decrease substantially, and the model selects important variables more frequently. In all the settings, the subset basis algorithm performs as well as the full algorithm but can reduce the time substantially.

Table 3. Average CKL, EMR and time for AR(1) example with $\rho = 0$.

n	N	CKL	EMR	Time
100	50	0.462 (0.049)	0.221 (0.033)	24.8
	100	0.463 (0.049)	0.221 (0.032)	64.5
200	50	0.371 (0.029)	0.172 (0.020)	57.3
	100	0.371 (0.026)	0.172 (0.019)	185.9
	200	0.371 (0.027)	0.172 (0.020)	232.4
500	50	0.325 (0.017)	0.148 (0.014)	118.8
	100	0.326 (0.016)	0.148 (0.014)	221.7
	200	0.327 (0.016)	0.148 (0.014)	648.4
	500	0.328 (0.013)	0.151 (0.007)	1179.1

Table 4. Appearance frequency of variables and the average model size, AR(1) with $\rho = 0$.

n	N	1	2	3	4	5	6	7	8	9	10	model size
100	50	86	32	99	95	32	30	29	30	31	29	4.93 (2.07)
	100	87	33	99	95	35	33	30	31	32	30	5.05 (2.13)
200	50	93	40	100	100	10	14	13	7	12	12	4.01 (1.36)
	100	95	39	100	100	10	14	12	10	12	14	4.06 (1.43)
	200	95	39	100	100	10	14	12	10	12	16	4.08 (1.45)
500	50	100	85	100	100	6	8	6	8	7	3	4.23 (1.06)
	100	100	84	100	100	6	7	7	7	7	3	4.21 (1.05)
	200	100	83	100	100	6	7	7	7	7	3	4.20 (1.05)
	500	100	77	100	100	7	7	3	3	9	6	4.12 (0.99)

2. Correlated input variables.

Two types of covariance structures are considered: (trimmed) AR(1) with $\rho = 0.5$ and CS($t = 1$). The Bayes errors are respectively 0.149 and 0.142. Tables 5−8 show that the proposed method still performs very well in model building and estimation. When n is small, the method tends to choose some uninformative variables. As n grows, the method chooses the right model structure more frequently, and both CKL and EMR decrease quickly. The subset basis algorithm works more efficiently as well.

Table 5. Average CKL, EMR, and time for AR(1) example with $\rho = 0.5$.

n	N	CKL	EMR	Time
100	50	0.477 (0.054)	0.239 (0.035)	25.4
	100	0.466 (0.053)	0.231 (0.033)	95.9
200	50	0.477 (0.054)	0.239 (0.035)	79.1
	100	0.392 (0.027)	0.186 (0.018)	189.1
	200	0.392 (0.027)	0.186 (0.018)	224.7
500	50	0.351 (0.014)	0.165 (0.014)	118.8
	100	0.352 (0.013)	0.165 (0.013)	226.9
	200	0.352 (0.013)	0.165 (0.013)	668.7
	500	0.351 (0.011)	0.165 (0.006)	1171.0

Table 6. Variable appearance frequency and average model size for AR(1) with $\rho = 0.5$.

n	N	1	2	3	4	5	6	7	8	9	10	model size
100	50	82	45	100	90	38	27	31	33	22	27	4.95 (2.28)
	100	88	41	100	90	30	31	21	32	29	35	4.97 (2.10)
200	50	97	49	100	100	15	16	15	16	18	11	4.37 (1.69)
	100	99	48	100	100	14	18	15	18	17	12	4.40 (1.73)
	200	98	48	100	100	14	18	15	18	17	12	4.40 (1.73)
500	50	100	79	100	100	7	7	6	10	8	10	4.27 (1.25)
	100	100	81	100	100	4	10	7	11	6	10	4.29 (1.13)
	200	100	79	100	100	4	10	7	11	6	10	4.27 (1.14)
	500	100	78	100	100	12	4	11	5	9	7	4.26 (1.25)

Table 7. Average CKL, EMR, and time for CS($t = 1$) setting.

n	N	CKL	EMR	Time
100	50	0.468 (0.050)	0.222 (0.023)	18.5
	100	0.457 (0.049)	0.216 (0.026)	37.6
200	50	0.379 (0.022)	0.180 (0.017)	47.8
	100	0.380 (0.022)	0.180 (0.017)	98.7
	200	0.381 (0.023)	0.199 (0.016)	249.1
500	50	0.339 (0.012)	0.158 (0.007)	126.6
	100	0.337 (0.010)	0.158 (0.007)	252.7
	200	0.339 (0.012)	0.158 (0.007)	685.8
	500	0.339 (0.012)	0.158 (0.007)	1269.8

Table 8. Variable appearance frequency and average model size for CS($t = 1$) setting.

n	N	1	2	3	4	5	6	7	8	9	10	model size
100	50	74	38	100	81	38	33	30	33	34	35	4.96 (2.51)
	100	78	45	96	92	34	36	30	33	40	32	5.16 (2.49)
200	50	84	35	100	100	8	17	12	13	15	13	3.97 (1.49)
	100	83	37	100	100	11	18	14	12	17	14	4.06 (1.57)
	200	83	37	100	100	11	19	14	12	17	14	4.07 (1.57)
500	50	100	77	100	100	14	9	12	7	11	14	4.44 (1.43)
	100	100	71	100	100	6	7	6	7	5	6	4.08 (0.94)
	200	100	75	100	100	14	9	10	8	11	16	4.43 (1.34)
	500	100	75	100	100	14	9	10	8	11	15	4.42 (1.34)

Example 3. Categorical covariates included. Consider the case when both continuous and categorical predictors are present in the data. We first generate eleven variables from Unif$[0, 1]$, then make the last four variables categorical by setting

$$Z^{(1)} = I(X^{(8)} < 0.5) + 2I(X^{(8)} \geq 0.5),$$

$$Z^{(2)} = I(X^{(9)} < \frac{1}{3}) + 2I(\frac{1}{3} \leq X^{(9)} < \frac{2}{3}) + 3I(X^{(9)} > \frac{2}{3}),$$

$$Z^{(3)} = I(X^{(10)} < \frac{1}{4}) + 2I(\frac{1}{4} \leq X^{(9)} < \frac{2}{4}) + 3I(\frac{2}{4} \leq X^{(9)} < \frac{3}{4}) + 4I(X^{(9)} \geq \frac{3}{4}),$$

$$Z^{(4)} = I(X^{(11)} < \frac{1}{3}) + 2I(\frac{1}{3} \leq X^{(11)} < \frac{2}{3}) + 3I(X^{(11)} \geq \frac{2}{3}),$$

where I is the indicator function. The true logit function is constructed using the same building functions as in Example 2:

$$f(\mathbf{x}, \boldsymbol{z}) = 5g_1(x^{(1)}) + 3g_2(x^{(2)}) + 4g_3(x^{(3)}) + 6g_4(x^{(4)}) - 4.5z^{(1)} + 2.5\sqrt{z^{(3)}} - 2.4.$$

Table 9. Average CKL and EMR for CS(t=0) setting (Example 3).

n	N	CKL	EMR
100	100	0.536 (0.080)	0.265 (0.047)
200	100	0.351 (0.032)	0.160 (0.016)
	200	0.351 (0.033)	0.159 (0.017)
500	100	0.284 (0.009)	0.125 (0.006)
	200	0.284 (0.009)	0.125 (0.006)
	500	0.284 (0.009)	0.125 (0.006)

1. Uncorrelated input variables.

The correct model size is 6 and the Bayes error is 0.107. Table 10 shows that $Z^{(1)}$ is never missed; $Z^{(3)}$ is selected in 80% runs when $n = 200$, and in more than 98% runs when $n = 500$. Figure 4.2 plots the 5th,50th, 95th best estimates over 100 runs.

Table 10. Variable appearance frequency and average model size for CS($t = 0$) case.

n	N	x_1	x_2	x_3	x_4	x_5	x_6	x_7	z_1	z_2	z_3	z_4	model size
100	100	83	34	78	93	15	18	18	100	41	73	37	5.90 (2.37)
200	100	100	64	100	100	11	18	10	100	8	82	9	6.02 (1.41)
	200	100	73	99	100	7	16	16	100	8	79	5	6.03 (1.34)
500	100	100	100	100	100	19	10	13	100	2	98	1	6.43 (0.74)
	200	100	100	100	100	13	10	9	100	1	100	1	6.35 (0.64)
	500	100	100	100	100	19	11	13	100	2	98	1	6.44 (0.77)

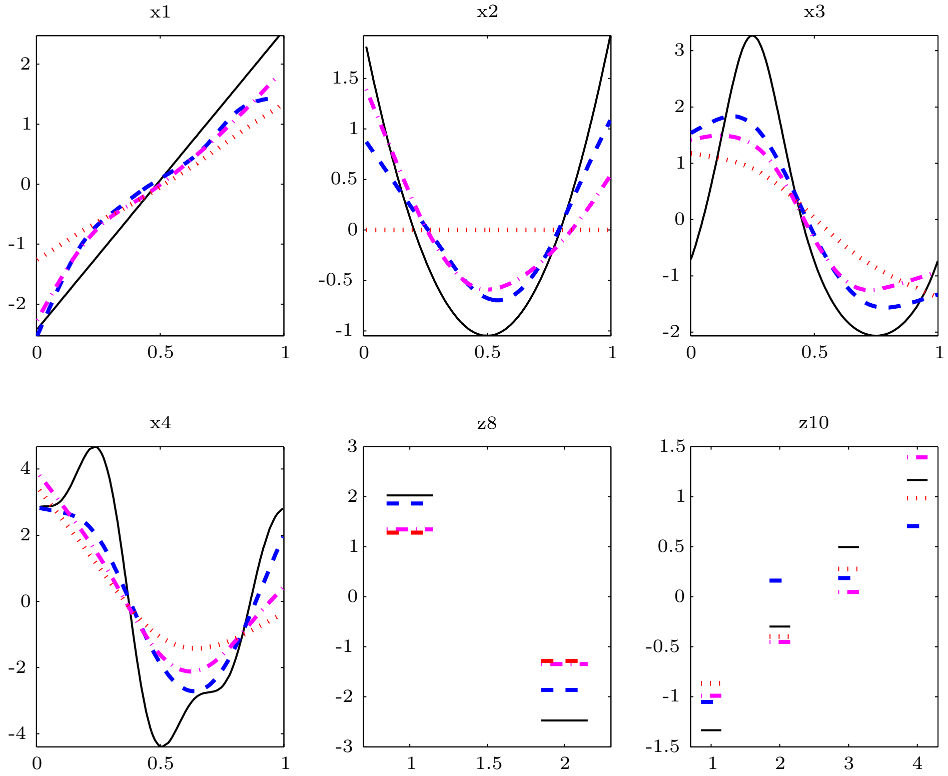

Figure 4.2. True (solid) and estimated functions (dashed, dashed-dotted, dotted respectively for 5th, 50th, 95th best) over 100 runs. Note that the 95th best run misses $X^{(2)}$.

2. Correlated input variables.

Consider the covariance structure CS($t = 1$). The Bayes error is 0.115. Tables 11 and 12 show that the method performs well.

Table 11. Average CKL and EMR for the CS($t = 1$) setting (Example 3).

n	N	CKL	EMR
100	100	0.486 (0.081)	0.231 (0.045)
200	100	0.367 (0.025)	0.165 (0.014)
	200	0.367 (0.026)	0.165 (0.014)
500	100	0.306 (0.013)	0.134 (0.007)
	200	0.306 (0.013)	0.134 (0.008)
	500	0.306 (0.014)	0.134 (0.008)

Table 12. Variable appearance frequency and the average model size for CS($t = 1$) setting.

n	N	x_1	x_2	x_3	x_4	x_5	x_6	x_7	z_1	z_2	z_3	z_4	model size
100	100	43	31	85	97	19	22	19	100	32	72	30	5.50 (2.40)
200	100	67	51	99	100	18	14	9	100	8	80	15	5.61 (1.74)
	200	67	51	99	100	18	13	10	100	8	80	14	5.60 (1.75)
500	100	97	83	100	100	15	12	11	100	4	98	5	6.25 (1.09)
	200	98	81	100	100	12	13	10	100	2	100	3	6.19 (1.00)
	500	98	82	100	100	12	13	10	100	2	100	3	6.20 (0.99)

Example 4. Two-way interaction model. Generate four continuous covariates independently from Unif$[0, 1]$. The true logit function contains an interaction term, with

$$f(\mathbf{x}) = 4x^{(1)} + \pi \sin\left(\pi x^{(1)}\right) + 6x^{(2)} - 8\left(x^{(2)}\right)^3 + 3\cos\left(2\pi(x^{(1)} - x^{(2)})\right) - 5.$$

The important components are $X^{(1)}, X^{(2)}$, and their interaction effect. The Bayes error is 0.155. The two-way interaction model is fitted for $n = 200$ and $n = 500$. Table 14 shows that the interaction term f_{12} is selected in all the runs.

Table 13. Average CKL, EMR, and times for Example 4.

n	N	CKL	EMR	Time
200	50	0.515 (0.005)	0.247 (0.004	30.59
	100	0.515 (0.005)	0.247 (0.004)	57.59
	200	0.514 (0.004)	0.248 (0.003)	99.71
500	50	0.448 (0.003)	0.192 (0.002)	187.30
	100	0.444 (0.008)	0.191 (0.007)	384.72

Table 14. Variable appearance frequency and average model size for Example 4.

n	N	x_1	x_2	x_3	x_4	x_1, x_2	x_1, x_3	x_1, x_4	x_2, x_3	x_2, x_4	x_3,x_4	model size
200	50	100	99	19	18	100	14	13	10	18	7	3.98 (1.56)
	100	100	99	19	18	100	14	13	10	18	7	3.98 (1.56)
	200	100	100	19	17	100	15	13	13	20	9	4.06 (1.64)
500	50	100	100	11	13	100	12	6	9	8	1	3.60 (0.74)
	100	100	100	13	15	100	15	11	10	12	5	3.81 (0.92)

5. Data Examples

We apply the COSSO penalized likelihood method to four benchmark datasets available at the UCI machine learning repository. Both the additive and two-way interaction COSSO are applied. When $n \leq 200$, we use the full basis algorithm; otherwise the subset basis algorithm is used with $N = 200$.

Cleveland Heart Disease. The problem concerns the prediction of the presence or absence of heart disease given various medical tests. There are 296 instances with seven categorical and six continuous covariates (contributed by R. Detrano).

BUPA Liver Disorder. The problem is to predict whether or not a male patient has a liver disorder based on blood tests and alcohol consumption. There are 345 observations with six continuous variables (donated by R. S. Forsyth).

PIMA Indian Diabetes. The problem is to predict the positive test for diabetes given a number of physiological measurements. The patients are females older than twenty-one years from Pima Indian heritage, Arizona. There are eight variables. Since the original data contains some impossible values (like zero bmi), we remove those instances and consider the remaining 532 observations (contributed by V. Sigillito).

Wisconsin Breast Cancer (WBC). The problem is to predict whether a tissue sample from a patient is malignant or benign. There are 699 observations and nine features; we use 683 complete records (donated by Dr. William H. Wolberg).

For each dataset, we estimate the mean error rate by ten-fold cross validation. In the ten-fold cross validation, the data is divided into ten parts, and each part in turn becomes the test set and the other nine parts form the training set. The tuning parameters are tuned with five-fold cross validation within the training set, then the performance of the algorithm is evaluated on the test set. Lim and Loh (2000) conducted a thorough comparison among thirty-three classification algorithms, including twenty-two decision tree, nine statistical, and two neural network algorithms, on these datasets. Their results are also based on ten-fold cross validation. Table 15 presents the EMRs of the additive COSSO, two-way interaction COSSO, and the best rule of the thirty-three algorithms reported in Lim and Loh (2000), which is referred to as "best(LL2000)". We can see from the table the performance of our method is comparable to the best performance of the thirty-three algorithms. We summarize the average model size of the additive and two-way interaction models in Table 16.

Table 15. EMRs of the additive, two-way interaction COSSO method, and the best(LL2000).

	additive COSSO	two-way COSSO	best(LL2000)
BUPA	**0.257** (0.022)	0.286 (0.023)	0.28
Cleveland	0.169 (0.022)	0.165 (0.018)	**0.14**
PIMA	0.220 (0.014)	**0.214** (0.014)	0.22
WBC	0.031 (0.008)	**0.026** (0.005)	0.03

Table 16. Average model size of the additive and two-way interaction COSSO models.

	BUPA	Cleveland	PIMA	WBC
additive COSSO	6.0 (0.03)	9.4 (2.95)	6.0 (0.94)	8.8 (0.63)
two-way COSSO	12.2 (4.9)	10.0 (3.43)	6.6 (2.16)	14.6 (1.27)

6. Summary

We propose a penalized likelihood method in a nonparametric generalized regression setting for simultaneous model estimation and variable selection. The special penalty form is able to shrink insignificant components to exact zeros. Our method handles continuous and categorical variables in a uniform manner, and demonstrates competitive performance in numerous examples. The subset basis algorithm is shown to be efficient for large datasets.

Acknowledgement

Hao Helen Zhang's research was partially supported by National Science Foundation grant DMS-0405913 and NCSU Faculty Research and Professional Development Fund. Yi Lin's work was partially supported by National Science Foundation grant DMS-0134987.

Appendix 1. Proof of Solution Existence

Proof of Theorem 2.1. Without loss of generality, we take $\tau = 1$. Denote the last part of (2.3) by $\mathcal{F}_1$. Because $\sum_{\alpha=1}^{p} \|P^\alpha f\|^2 \leq J^2(f)$, 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 $\mathcal{F}_1$. Let $a = \max_{i=1}^n R_{\mathcal{F}_1}^{1/2}(\mathbf{x}_i, \mathbf{x}_i)$. By the definition of a reproducing kernel we have, for any $f \in \mathcal{F}_1$ and $i = 1, \ldots, n$,

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

Denote the objective functional in (2.5) by $D(f)$. By the theory of exponential family distributions, B is convex and infinitely differentiable. Therefore from (A.1) we see that $D(f)$ is convex and continuous in $\mathcal{F}$. Let $\rho = (1/n) \sum_{i=1}^n \rho_i$, where the ρ_i's are defined in assumption (i) and are finite. It is easy to see that for any $f \in \mathcal{F}$, we have

$$D(f) \geq J(f) - \rho. \tag{A.2}$$

Denote the minimizer in assumption (ii) by b_0. For any $r > 0$, consider the set

$$E_r = \{f \in \mathcal{F} : f = b + f_1, \text{with } \ b \in \{1\}, f_1 \in \mathcal{F}_1, J(f) \leq \rho + B(0) + 1, |b - b_0| \leq r\}.$$

Then E_r is a closed, convex, and bounded set. By Theorem 4 of Tapia and Thompson (1978, p.162), there exists a minimizer of $D(f)$ in E_r. Denote this minimizer by $f_r = b_r + \bar{f}_r$, with $b_r \in \{1\}$ and $\bar{f}_r \in \mathcal{F}_1$. From (A.2) we have

$$J(f_r) \leq D(f_r) + \rho \leq D(b_0) + \rho \leq D(0) + \rho < B(0) + \rho + 1. \tag{A.3}$$

Now if $D(f)$ has no minimizer in $\mathcal{F}$, then f_r must be on the boundary of E_r. From (A.3) we must have $|b_r - b_0| = r$. Because D is convex and $D(f_r) \leq D(b_0)$, we have

$$D\{[\alpha b_r + (1-\alpha)b_0] + \alpha \bar{f}_r\} = D(\alpha f_r + (1-\alpha)b_0) \leq \alpha D(f_r) + (1-\alpha)D(b_0) \leq D(b_0), \tag{A.4}$$

for any $0 \leq \alpha \leq 1$. Now take a sequence $r_i \to \infty$ and set $\alpha_i = 1/r_i$. Then $\alpha_i \bar{f}_{r_i} \to 0$ in $\mathcal{F}$ since $J(\alpha_i \bar{f}_{r_i}) = \alpha_i J(\bar{f}_{r_i}) \leq \alpha_i(\rho + B(0) + 1) \to 0$. Since $|[\alpha_i b_{r_i} + (1 - \alpha_i)b_0] - b_0| = \alpha_i |b_{r_i} - b_0| = 1$, there exists a convergent subsequence of $[\alpha_i b_{r_i} + (1 - \alpha_i)b_0]$ converging to either $b_0 + 1$ or $b_0 - 1$. By looking at this subsequence, we get that $D(b_0 + 1) \leq D(b_0)$ or $D(b_0 - 1) \leq D(b_0)$ from (A.4), and that D is continuous in $\mathcal{F}$. That is, $L(b_0 + 1) \leq L(b_0)$ or $L(b_0 - 1) \leq L(b_0)$. This contradicts the uniqueness of the minimizer of $L(\eta)$ over $\{1\}$. Therefore $D(f)$ has a minimizer in $\mathcal{F}$.

Appendix 2. Proof of Representer Theorem

Proof of Theorem 2.2. 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 $\text{span}\{R_\alpha(\mathbf{x}_i, \cdot), i = 1, \ldots, n\} \subset \mathcal{F}^\alpha$

be denoted by g_α, and its orthogonal complement by h_α. Then $f_\alpha = g_\alpha + h_\alpha$, and $\|f_\alpha\|^2 = \|g_\alpha\|^2 + \|h_\alpha\|^2$, $\alpha = 1, \ldots, p$. Since the reproducing kernel of $\mathcal{F}$ is $R = 1 + \sum_{\alpha=1}^p R_\alpha$, we have

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

where $\langle \cdot, \cdot \rangle$ is the inner product in $\mathcal{F}$. Therefore (2.5) can be written as

$$\frac{1}{n}\sum_{i=1}^n \big[- l\{y_i, b + \sum_{\alpha=1}^p \langle R_\alpha(\mathbf{x}_i, \cdot), g_\alpha\rangle\}\big] + \tau^2 \sum_{\alpha=1}^p (\|g_\alpha\|^2 + \|h_\alpha\|^2)^{1/2}.$$

Then any minimizer f satisfies $h_\alpha = 0$, $\alpha = 1, \ldots, p$, and the theorem is proved.

Proof of Lemma 2.1. Denote the functional in (2.5) by $D(f)$, and the functional in (2.6) by $N(\theta, f)$. For any $\theta_\alpha \geq 0, f \in \mathcal{F}$, we have $\lambda_0 \theta_\alpha^{-1} \|P^\alpha f\|^2 + \lambda\theta_\alpha \geq 2\lambda_0^{1/2}\lambda^{1/2}\|P^\alpha f\| = \tau^2 \|P^\alpha f\|$, and equality holds if and only if $\theta_\alpha = \lambda_0^{1/2}\lambda^{-1/2}\|P^\alpha f\|$. Therefore $N(\boldsymbol{\theta}, f) \geq D(f)$ for any $\theta_\alpha \geq 0$, $\alpha = 1, \ldots, p$, and $f \in \mathcal{F}$, and the equality holds if and only if $\theta_\alpha = \lambda_0^{1/2}\lambda^{-1/2}\|P^\alpha f\|$, $\alpha = 1, \ldots, p$. The conclusion follows.

References

Bach, F., Lanckriet, G. R. and Jordan, M. I. (2004). Multiple kernel learning, conic duality, and the SMO algorithm. *Proceeding of the Twenty-First International Conference on Machine Learning.*

Bach, F., Thibaux, R. and Jordan, M. I. (2004). Computing regularization paths for learning multiple kernels. *Advances in Neural Information Processing Systems.* To appear.

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

Breiman, L., Friedman, J. H., Olshen, R. and Stone, C. J. (1984). *Classification and Regression Trees.* Wadsworth, Belmont, CA.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.* **32**, 407-451.

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

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

Friedman, J. H. and Silverman, B. W. (1989). Flexible parsimonious smoothing and additive modeling (with discussion). *Technometrics* **31**, 3-39.

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

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

Gu, C. and Kim, Y. J. (2001). Penalized likelihood regression: General formulation and efficient approximation. *Canadian Journal of Statistics* **30**, 619-628.

Hastie, T. (1989). Discussion of ”Flexible parsimonious smoothing and additive modeling” by J. Friedman and B. Silverman. *Technometrics* **31**, 3-39.

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

Kimeldorf, G. and Wahba, G. (1971). Some results on Tchebycheffian spline functions. *J. Math. Anal. Appl.* **33**, 82-85.

Lim, T. and Loh, W. Y. (2000). A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine Learning* **40**, 203-229.

Lin, Y. and Zhang, H. H. (2002). Component selection and smoothing in smoothing spline analysis of variance models. Tech. Rep. 1072, University of Wisconsin - Madison. Submitted.

Mangasarian, O. L. and Wolberg, W. H. (1990). Cancer diagnosis via linear programming. *SIAM News* **23-5**, 1-18.

Ruppert, D. and Carroll, R. (2000). Spatially-adaptive penalties for spline fitting. *Aust. N. Z. J. Stat.* **45**, 205-223.

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

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

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.

Xiang, D. and Wahba, G. (1998). Approximate smoothing spline methods for large data sets in the binary case. *Proceedings of ASA Joint Statistical Meetings, Biometrics Section*, 94-98.

Yau, P., Kohn, R. and Wood, S. (2002). Bayesian variable selection and model averaging in high dimensional multinomial nonparametric regression. *J. Comput. Graph. Statist.* **12**, 23-54.

Zhang, H. H., Wahba, G., Lin, Y., Voelker, M., Ferris, M., Klein, R. and Klein, B. (2004). Variable selection and model building via likelihood basis pursuit. *J. Amer. Statist. Assoc.* **99**, 659-672.

Department of Statistics, North Carolina State University, U.S.A.

E-mail: hzhang2@stat.ncsu.edu

Department of Statistics, University of Wisconsin at Madison, U.S.A.

E-mail: yilin@stat.wisc.edu

(Received August 2004; accepted April 2005)