Statistica Sinica **25** (2015), 1337-1354
doi:http://dx.doi.org/10.5705/ss.2014.035

FUNCTIONAL SPARSITY: GLOBAL VERSUS LOCAL

Haonan Wang and Bo Kai

Colorado State University and College of Charleston

Abstract: We consider the model selection problem in nonparametric regression. The notion of functional sparsity is a generalization of parameter sparsity in parametric models. In particular, two types of sparsity are studied, global and local sparsity. The goal is to produce a sparse estimate, that assigns zero values over regions where the true underlying function is zero. Most classical smoothing techniques yield consistent estimates with no sparsity. Here, a penalized least squares procedure, based on a basis function approximation and the group bridge penalty function, is proposed for simultaneous function estimation and zero subregion detection. Asymptotic properties, including both consistency in estimation and sparsistency in model selection, of the procedure are established. The methodology is illustrated through simulation studies and a case study.

Key words and phrases: Functional data analysis, group bridge, model selection, smoothing.

1. Introduction

In many scientific problems, classical regression and nonparametric smoothing techniques are implemented to model the relationship between the response variable and predictor variables. For better interpretability, researchers and practitioners are interested in identifying the predictor variables that are important and necessary for such model. It is also of great interest, and is our major goal, to identify information within each important predictor variable that actually relates to the response variable.

For instance, in neuroscience, information communication between different neurons, and hence between brain regions, is in the form of neuron spikes. Modeling the causal relationship between input and output neurons will enhance our understanding of brain cognitive functions. A key problem is to select the periods, within an observed time series from each input neuron, that are transformed into the output signal. Another example arises in human physiology, where interest is in modeling the functional relationship between the force exerted on an object and time. For a specific task performed, it is critical to identify the time point at which the force exerted deviates from the background force as such a point is an indicator of human response time to this task. See Ramsay, Wang, and Flanagan (1995) and Song et al. (2007b,a) for more details regarding these applications.

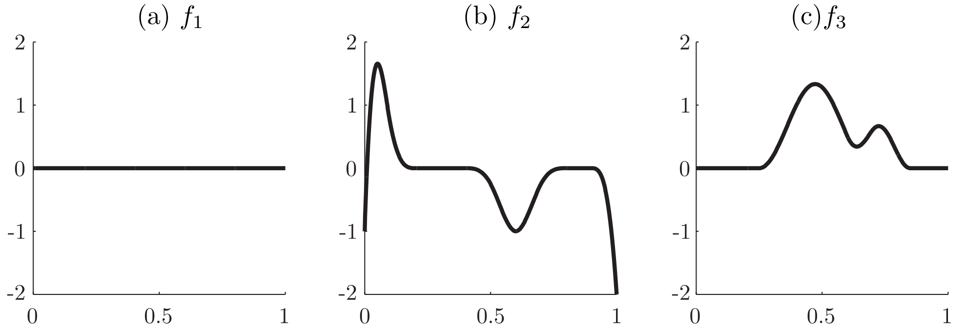

Figure 1. Here, f_1 is a function with global sparsity, while f_2 and f_3 possess local sparsity.

Consider the standard univariate nonparametric regression model

$$Y = f(X) + \varepsilon, \tag{1.1}$$

where Y is the response variable, X is the predictor variable and ε is the random error term with mean 0 and finite variance. In applications, a random sample of observations is obtained and denoted by $(X_1, Y_1), \ldots, (X_n, Y_n)$. The conditional mean, $f(x) = E(Y|X = x)$, is usually assumed to be an unknown smooth function. Here, we further assume that f is *sparse* as defined in Tu et al. (2012), and its zero region can be expressed as a finite union of subintervals. Those authors described two types of sparsity: global and local sparsity. In particular, if f is zero over its entire domain, it is called a function with global sparsity; if f is zero only over part of its domain (union of sub-intervals), it is called a function with local sparsity. When characterizing functional sparsity, singletons at which f is zero are not considered. For illustration, three functions with sparsity are shown in Figure 1. Here, f_1 is a function with global sparsity, while f_2 and f_3 are functions possessing local sparsity. Global sparsity suggests that there is no relationship between the predictor variable x and the response variable y. Local sparsity provides a new, flexible choice of the interpretability regarding such relationships: the response variable is pure noise when the predictor variable falls into the zero region. We propose and study a new estimation procedure that produces sparse and consistent estimates for functional relationships.

Nonparametric regression provides a class of powerful data-driven tools to explore the unknown relationship between response and predictor variables. There are many smoothing techniques developed for estimating a nonparametric function f, such as kernel smoothing (e.g., Härdle (1990); Wand and Jones (1995)), local polynomial regression (e.g., Fan and Gijbels (1996)), and spline smoothing (e.g., Wahba (1990); Eubank (1999)). Most existing smoothing methods are able to provide consistent estimates of f. But these known procedures cannot produce sparse solutions. We aim to develop a new approach to detect both types of sparsity. In particular, we would like to consistently produce zero estimates for f when no relationship between the response variable and the predictor variable is indicated.

Over the past several decades, sparsity for parametric regression models has been well defined (e.g., Fan and Li (2001)) and broadly studied in the context of variable selection. Among various variable selection procedures, regularization-based methods owe their popularity to improved estimation performance. In particular, they can achieve simultaneous model selection and parameter estimation. Popular penalization methods include least absolute shrinkage and selection operator (Lasso; Tibshirani (1996)), smoothly clipped absolute deviation (SCAD; Fan and Li (2001)), and the adaptive Lasso (Zou (2006)). Various techniques for group variable selection have been proposed. Yuan and Lin (2006) extended the idea of Lasso, and proposed group Lasso for selecting groups of variables. Huang et al. (2009) proposed a group bridge approach, which allows simultaneous variable selection at both group and individual (within-group) levels.

Recently, nonparametric estimation of functions with global sparsity has attracted attention. The usual first step is to approximate the unknown function by a set of basis functions, e.g., polynomial splines. See Huang (2003) for more detailed discussion on asymptotic theory for polynomial spline regression. Then, the nonparametric estimation problem for functions with global sparsity can be solved through various regularized methods developed for linear regression models. For instance, Wang, Li and Huang (2008) proposed a regularized estimation procedure for variable selection that combines basis function approximations and the group SCAD penalty. The proposed procedure can simultaneously select significant variables with time-varying effects and estimate the nonzero smooth coefficient functions. Huang, Horowitz and Wei (2010) proposed an adaptive group Lasso method for nonparametric additive models. Tu et al. (2012) further generalized the group bridge approach (Huang et al. (2009)), and proposed a sparse functional dynamic Multiple-Input-Single-Output model for analyzing neuron spike data. These methods perform quite well in identifying functions with global sparsity, and hence assign zero to those functions. The key to such satisfactory performance is the fact that the zero function lies in the linear space spanned by the basis functions, and corresponds to the zero vector of coefficients.

For functions with local sparsity, the situation is rather complicated. James, Wang, and Zhu (2009) have done pioneering work here in a functional linear model. They used a simple grid basis to approximate the nonparametric function, and implemented the Dantzig selector (Candes and Tao (2007); Bickel, Ritov, and Tsybakov (2009); James, Radchenko, and Lv (2009)) to determine whether or not the nonparametric function and its dth derivative are zero at each of the grid points. Zhou, Wang, and Wang (2013) have pointed out that this approach tends to yield estimates with large variation over nonzero region when the number of knots increases. Moreover, Zhou, Wang, and Wang (2013) improved this approach, and proposed a two-stage procedure. An initial estimator was obtained by the Dantzig selector, and a refinement procedure using a group SCAD approach was proposed.

In this paper, we propose a new one-step penalized procedure which is capable of simultaneous function estimation and sparse subregion detection. In particular, we take advantage of the local support property of B-spline basis functions, and propose an innovative overlapping group assignment of the vector coefficients. Our estimation procedure can be carried out by a well-developed algorithm proposed by Huang et al. (2009), and it is computationally tractable. We establish the asymptotic properties of our proposed method under standard smoothness assumptions. We prove that, under mild regularity conditions, our resulting penalized function estimate converges to the true underlying function at the optimal rate of convergence (Stone (1982)).

The rest of this paper is organized as follows. In Section 2, we introduce a regularized estimation procedure using both the basis expansion and the group bridge penalty, and discuss some practical issues, such as computation, tuning parameter selection, variance estimation. Asymptotic properties, including the consistency in estimation and sparsistency in model selection, are given in Section 3. In Section 4, we present several Monte Carlo simulation studies to evaluate the performance of the proposed procedure. Section 5 provides a data example. Conclusion and discussion are given in Section 6. Proofs and additional discussion are in the online supplement.

2. Methodology

2.1. Polynomial spline approximation and functional sparsity

Without loss of generality, we assume that the domain of $f(x)$ is $[0, 1]$. Polynomial splines are piecewise polynomials with the pieces joined smoothly at a set of interior knot points. In this paper, we adopt the B-spline basis functions due to their stable numerical properties. Details about spline functions can be found in de Boor (1978) and Schumaker (1981).

For B-splines, we partition the interval $[0, 1]$ into $M_n + 1$ subintervals by M_n interior knot points $0 < \kappa_1 < \cdots < \kappa_{M_n} < 1$, where M_n is allowed to increase with the sample size n. In addition, let $\kappa_0 = 0$ and $\kappa_{M_n+1} = 1$. The corresponding B-spline basis functions are denoted by $B_1(x), \ldots, B_{M_n+d+1}(x)$, where d is the degree of polynomial pieces. Each B-spline basis function has a local support and, for any adjacent knots κ_{j-1} and κ_j $(1 \leq j \leq M_n + 1)$, except for $d+1$ basis functions $B_j(x), \ldots, B_{j+d}(x)$, the basis functions vanish on the interval $[\kappa_{j-1}, \kappa_j]$.

Let $\mathbb{G}$ be the linear space of spline functions on $[0,1]$ spanned by the B-spline basis functions $\{B_k(x) : k = 1, \ldots, M_n + d + 1\}$. Suppose that $f(x)$ can be approximated by an element in $\mathbb{G}$ as

$$f(x) \approx \sum_{k=1}^{L_n} \gamma_k B_k(x), \tag{2.1}$$

where $L_n = M_n + d + 1$. In the special case that $f \in \mathbb{G}$, the approximation can be replaced with equality. The functional sparsity of f, including both global and local sparsities, can be fully characterized through the sparsity of its coefficient vector, the parametric representation of f in the linear space $\mathbb{G}$. Thus the zero function corresponds to the zero vector of coefficients of dimension L_n. Moreover, if $f(x) = 0$ for $x \in [\kappa_{j-1}, \kappa_j]$, then $\gamma_j = \cdots = \gamma_{j+d} = 0$.

For f not in the linear space $\mathbb{G}$, an accurate approximation is attainable for a sufficiently large sample (Schumaker (1981)). A natural question is whether such an approximant carries all necessary information regarding the sparsity of the true function $f(x)$. As will be shown in Lemma 1, there exists a function f_0 in $\mathbb{G}$ which provides an accurate approximation and also satisfies that, if $f(x) = 0$ for $x \in [\kappa_{j-1}, \kappa_j]$, so is $f_0(x)$. Consequently, the sparsity of f can be partially inferred through the sparsity of f_0, which can be represented by the parametric sparsity of the coefficient vector of the B-spline representation of f_0. This observation motivates our proposed penalized estimation procedure.

2.2. Penalized estimation procedure

The parameters in (2.1) can be estimated by minimizing the (un-penalized) least squares criterion

$$\frac{1}{n} \sum_{i=1}^{n} \Big(Y_i - \sum_{k=1}^{L_n} \gamma_k B_k(X_i) \Big)^2 = \frac{1}{n} \| \boldsymbol{Y} - \boldsymbol{B}\boldsymbol{\gamma} \|_2^2, \tag{2.2}$$

where $\boldsymbol{B}(x) = \big(B_1(x), \cdots, B_{L_n}(x)\big)^T$ is a vector of B-spline basis functions, $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_{L_n})^T$ is a vector of coefficients, $\boldsymbol{B}$ is the matrix $(\boldsymbol{B}(X_1), \cdots, \boldsymbol{B}(X_n))^T$, and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ is the response vector. The resulting estimated function is consistent. For more details about the rate of convergence of the least squares estimate, see Stone (1985), Huang (2001), and references therein.

Now suppose that $f(x)$ is sparse either locally or globally. We introduce a regularization term in addition to the least squares criterion (2.2), so that the resulting estimated function is correspondingly sparse. We adopt the group bridge penalty (Huang et al. (2009)) and the penalized least squares criterion

$$Q_n(\boldsymbol{\gamma}) = \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{B}\boldsymbol{\gamma}\|_2^2 + \lambda_n \sum_{j=1}^{M_n+1} \|\boldsymbol{\gamma}_{A_j}\|_1^{\alpha}, \tag{2.3}$$

where $0 < \alpha < 1$, λ_n is a regularization parameter, $A_j = \{j, j+1, \dots j+d\}$, $\boldsymbol{\gamma}_{A_j} = (\gamma_k, k \in A_j)^T$ and $\|\boldsymbol{\gamma}_{A_j}\|_1 = |\gamma_j| + \cdots + |\gamma_{j+d}|$. Let $\widehat{\boldsymbol{\gamma}}$ be a minimizer of $Q_n(\boldsymbol{\gamma})$ in (2.3). The corresponding penalized estimator of $f(x)$ is $\widehat{f}(x) = \boldsymbol{B}(x)^T\widehat{\boldsymbol{\gamma}}$. In classical linear regression model, Huang et al. (2009) pointed out that, if $\alpha = 1$, the group bridge penalty is the L_1 penalty and can only be used for individual variable selection. They also pointed out that, if $0 < \alpha < 1$, the group bridge penalty can be used for variable selection at both between-group and within-group levels simultaneously.

An alternative regularization term is the group Lasso penalty. However, the groups A_j in (2.3) overlap under our setting. Neither the original group Lasso (Yuan and Lin (2006)) nor the sparse group Lasso (Friedman, Hastie, and Tibshirani (2010); Simon et al. (2013)) can be applied to this problem because they were designed for non-overlapping groups. The group Lasso with overlapping groups has also been studied; see Jacob, Obozinski, and Vert (2009) and Liu and Ye (2010) for more details. Recently, Percival (2012) pointed out that the group Lasso method with overlapping groups may not be able to recover the true sparsity structure. In this paper, we include the group Lasso with overlapping groups as a competing method. Our simulation results support the theoretical findings of Percival (2012).

2.3. Computational aspects

Minimization of (2.3) is rather difficult since the group bridge penalty is not a convex function for $0 < \alpha < 1$. Here, we follow the iterative algorithm proposed by Huang et al. (2009) to find the minimizer for (2.3). It is outlined below.

1. Obtain an initial value of $\boldsymbol{\gamma}$, denoted by $\boldsymbol{\gamma}^{(0)}$.
2. For a given choice of λ_n, compute $\tau_n = \lambda_n^{1/(1-\alpha)}\alpha^{\alpha/(1-\alpha)}(1-\alpha)$.
3. For $s = 1, 2, \dots$, compute

$$\boldsymbol{\gamma}^{(s)} = \arg\min_{\boldsymbol{\gamma}} \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{B}\boldsymbol{\gamma}\|_2^2 + \sum_{j=1}^{M_n+1} \left(\theta_j^{(s)}\right)^{1-1/\alpha} \|\boldsymbol{\gamma}_{A_j}\|_1, \tag{2.4}$$

where

$$\theta_j^{(s)} = \left(\frac{1-\alpha}{\alpha\tau_n}\right)^{\alpha} \|\boldsymbol{\gamma}_{A_j}^{(s-1)}\|_1^{\alpha}, j = 1, \dots, M_n + 1.$$

4. Stop with convergence.

Huang et al. (2009) further pointed out that this algorithm always converges to a local minimizer depending on the initial value $\boldsymbol{\gamma}^{(0)}$. A natural first choice for the initial $\boldsymbol{\gamma}^{(0)}$ is the ordinary least squares estimator of (2.2). Step 3 can be carried out by the LARS algorithm (Efron et al. (2004)).

The choice of tuning parameters is crucial as it determines the performance of the proposed method. First, a sequence of knots needs to be selected for the B-spline basis. For simplicity, we use equally spaced knots and only select M_n, the number of interior knots. Similar to Huang, Wu, and Zhou (2004), we use K-fold cross-validation to select M_n. Then, an optimal λ_n, the tuning parameter that determines the sparsity of the resulting estimated function, needs to be selected. Since the number of coefficients L_n increases with n, we adopt a Bayesian information criterion (BIC) type procedure used in Huang, Horowitz and Wei (2010) to select λ_n:

$$BIC = \log\left\{\frac{\|\boldsymbol{y} - \boldsymbol{B}\widehat{\boldsymbol{\gamma}}\|_2^2}{n}\right\} + df \cdot \frac{\log(n)}{n} + \nu \cdot df \cdot \frac{\log(L_n)}{n}, \tag{2.5}$$

where $0 \le \nu \le 1$ is a constant and df is the total number of coefficients estimated as nonzero. Here we use $\nu = 0.5$ as suggested in Huang, Horowitz and Wei (2010).

2.4. Variance estimation

Consider the asymptotic variance of the proposed estimator. Let $\mathcal{C}$ be the set of indexes of selected coefficients and $\widehat{\gamma}_{\mathcal{C}}(\lambda_n)$ be the nonzero components of $\widehat{\boldsymbol{\gamma}}$ given λ_n. By (2.4) and the Karush-Kuhn-Tucker condition, we have

$$\widehat{\gamma}_{\mathcal{C}}(\lambda_n) = \{\boldsymbol{B}_{\lambda_n}^T \boldsymbol{B}_{\lambda_n} + \frac{n}{2}\boldsymbol{W}_{\lambda_n}\}^{-1}\boldsymbol{B}_{\lambda_n}^T \boldsymbol{y},$$

where $\boldsymbol{B}_{\lambda_n}$ is the sub-matrix of $\boldsymbol{B}$ by selecting the corresponding columns indexed by $\mathcal{C}$ for the given λ_n, and $\boldsymbol{W}_{\lambda_n}$ is the diagonal matrix with diagonal elements

$$\sum_{j:A_j \ni k} \frac{\widehat{\theta}_j^{(1-1/\alpha)}}{|\widehat{\gamma}_k|}, \qquad \widehat{\gamma}_k \neq 0.$$

Therefore, the asymptotic variance of $\widehat{\gamma}_{\mathcal{C}}(\lambda_n)$, defined as $\boldsymbol{avar}\{\widehat{\gamma}_{\mathcal{C}}(\lambda_n)\}$, can be approximated by

$$\{\boldsymbol{B}_{\lambda_n}^T \boldsymbol{B}_{\lambda_n} + \frac{n}{2}\boldsymbol{W}_{\lambda_n}\}^{-1}\boldsymbol{B}_{\lambda_n}^T \boldsymbol{B}_{\lambda_n}\{\boldsymbol{B}_{\lambda_n}^T \boldsymbol{B}_{\lambda_n} + \frac{n}{2}\boldsymbol{W}_{\lambda_n}\}^{-1}\widehat{\sigma}^2,$$

where $\widehat{\sigma}^2 = \|\boldsymbol{y} - \boldsymbol{B}\widehat{\boldsymbol{\gamma}}\|_2^2/n$. As $\widehat{f}(x) = \boldsymbol{B}(x)^T\widehat{\boldsymbol{\gamma}}$, the asymptotic variance of $\widehat{f}(x)$ is

$$\boldsymbol{B}_{\lambda_n}^T(x)\boldsymbol{avar}\{\widehat{\gamma}_{\mathcal{C}}(\lambda_n)\}\boldsymbol{B}_{\lambda_n}(x), \tag{2.6}$$

where $\boldsymbol{B}_{\lambda_n}(x)$ is the sub-vector of $\boldsymbol{B}(x)$ on selecting the corresponding entries indexed by $\mathcal{C}$ for the given λ_n.

3. Large Sample Properties

We show that the proposed estimator still converges to the true function at the optimal rate. Moreover, the proposed estimator correctly identifies the sparse pieces of the true function with probability converging to one under certain regularity conditions.

Let $\mathcal{H}_r$ be the collection of functions defined on $[0,1]$ whose δth derivative satisfies the Hölder condition of order ζ with $r \equiv \delta + \zeta$: there exists a constant C such that $|f^{(\delta)}(x_1) - f^{(\delta)}(x_2)| \leq C|x_1 - x_2|^{\zeta}$ for any $0 \leq x_1, x_2 \leq 1$. We need an assumption

(A.1) $f \in \mathcal{H}_r$ for some $r \geq 2$.

In the same notation, if (A.1) holds, there is a vector of dimension L_n, $\boldsymbol{\gamma}^* = (\gamma_1^*, \ldots, \gamma_{L_n}^*)^T$, so that the approximation error of $f^*(x)$ to $f(x)$ is, for a constant C^*,

$$\|f(x) - f^*(x)\|_\infty \leq C^* M_n^{-r}, \tag{3.1}$$

(Schumaker (1981)). This spline approximant, as well as the vector of coefficients $\boldsymbol{\gamma}^*$, may not possess any sparsity.

We introduce a *sparse modification* of $f^*(x)$ in $\mathbb{G}$, denoted by $f_0(x) = \boldsymbol{B}(x)^T\boldsymbol{\gamma}_0$. For convenience, we partition $\{1, \ldots, M_n+1\}$ into the sets

$$\begin{aligned}
&\mathcal{A}_1 = \{j : f(x) = 0, x \in [\kappa_{j-1}, \kappa_j]\}, \\
&\mathcal{A}_2 = \left\{j : 0 < \max_{x\in[\kappa_{j-1},\kappa_j]} |f(x)| \leq D M_n^{-r} \text{for some } D \text{ and } D > C\right\}, \\
&\mathcal{A}_3 = \{1, \ldots, M_n+1\} - \mathcal{A}_1 \cup \mathcal{A}_2.
\end{aligned}$$

Except for some singleton zeros of $f(x)$, we have

$$\bigcup_{j\in\mathcal{A}_1} [\kappa_{j-1}, \kappa_j] \subset \{x : f(x) = 0\} \subset \bigcup_{j\in\mathcal{A}_1\cup\mathcal{A}_2} [\kappa_{j-1}, \kappa_j].$$

It can be seen that $\bigcup_{j\in\mathcal{A}_1}[\kappa_{j-1}, \kappa_j]$ and $\bigcup_{j\in\mathcal{A}_1\cup\mathcal{A}_2}[\kappa_{j-1}, \kappa_j]$ provide a lower and an upper bound of the zero region of f. Here, $\bigcup_{j\in\mathcal{A}_2}[\kappa_{j-1}, \kappa_j]$ is a transition area which connects the zero region and nonzero region; if f is globally sparse, $\mathcal{A}_2$ and $\mathcal{A}_3$ are empty sets. If f has local sparsity, it can be shown that the total number of elements in $\mathcal{A}_2$ is of order $o(M_n)$. Consequently, the length of $\bigcup_{j\in\mathcal{A}_2}[\kappa_{j-1}, \kappa_j]$ converges to zero as n goes to infinity.

A sparse modification of $\boldsymbol{\gamma}^*$ can be defined as $\boldsymbol{\gamma}_0 = (\gamma_{0,1}, \ldots, \gamma_{0,L_n})^T$ with $\gamma_{0,k} = \gamma_k^* I\{k \notin \mathcal{B}_1\}$, where $\mathcal{B}_1 = \bigcup_{j\in\mathcal{A}_1\cup\mathcal{A}_2} A_j$. The resulting spline approximant,

$f_0(x) = \boldsymbol{B}(x)^T \boldsymbol{\gamma}_0$, reasonably preserves the sparsity of the function $f(x)$. In fact, $f_0(x)$ can be treated as a direct target of the estimation function $\widehat{f}(x)$. This suggests the decomposition

$$\widehat{f}(x) - f(x) = (\widehat{f}(x) - f_0(x)) + (f_0(x) - f(x)),$$

where $\widehat{f}(x) - f_0(x)$ and $f_0(x) - f(x)$ are the *estimation error* and the *approximation error* respectively.

Lemma 1. *Under* (A.1), *there exists a constant* C_0 *such that*

$$\|f(x) - f_0(x)\|_\infty \leq C_0 M_n^{-r}.$$

Theorem 1 (Convergence)**.** *Suppose that* (A.1) *holds, as does*

(A.2) $\{X_1, \ldots, X_n\}$ *is a random sample from a continuous density,* $f_X(x)$ *on [0,1], where* $f_X(x)$ *are uniformly bounded away from 0 and infinity.*

If $M_n \sim n^{1/(2r+1)}$ *and*

(A.3) $\lambda_n \alpha_n = O(n^{-1/2})$, *with* $\alpha_n = \left(\sum_{j \in \mathcal{A}_3} \|\boldsymbol{\gamma}_{0,A_j}\|_1^{2\alpha-2}\right)^{1/2}$,

then $\|\widehat{f} - f\|_2 = O_p\big(n^{-r/(2r+1)}\big)$.

The rate of convergence of our proposed function estimate is the same as the optimal rate of convergence for nonparametric regression (Stone (1982)). If $f(x) = 0$ for all $x \in [0,1]$, then $\mathcal{A}_2$ and $\mathcal{A}_3$ are empty sets, and (A.3) holds for all λ_n. On the other hand, if $f(x) \neq 0$ for all x, then $\mathcal{A}_1$ and $\mathcal{A}_2$ are empty and $\alpha_n = O(M_n^{1/2})$. Thus, we have $\lambda_n = O(n^{-(r+1)/(2r+1)})$. In general, if $f(x)$ possesses local sparsity and deviates from zero slowly, it requires a small λ_n to ensure the consistency of the resulting estimates.

Theorem 2 states that our proposed penalized least squares method will yield a sparse solution which is consistent with the sparsity of the true unknown function. That is, if $f(x) = 0, x \in [\kappa_i, \kappa_j]$, then our proposed method can recover such local sparsity with probability approaching 1.

Theorem 2 (Sparsistency)**.** *Suppose the assumptions of Theorem* 1 *and* (A.4) *hold, where*

(A.4) $\lambda_n n^{(2-\alpha)/2} M_n^{\alpha-1} \to \infty$.

With probability approaching 1, *we have* $(\widehat{\gamma}_{A_j} : j \in \mathcal{A}_1 \cup \mathcal{A}_2) = 0$.

As a direct consequence of Theorem 2, if $x \in \bigcup_{j \in \mathcal{A}_1} [\kappa_{j-1}, \kappa_j]$, our proposed procedure yields an estimate $\widehat{f}(x) = 0$ with large probability. Moreover, if $f(x) = 0$ for all x, it can be seen that the resulting estimator is also globally sparse. If a function $f(x)$ has local sparsity, $\mathcal{A}_2$ is nonempty, and $\bigcup_{j \in \mathcal{A}_2} [\kappa_{j-1}, \kappa_j]$ is identified as zero with large probability. Therefore, the penalized least squares estimate tends to yield a slightly more sparse function than the true function f. This occurs since, over the region $\bigcup_{j\in\mathcal{A}_2}[\kappa_{j-1},\kappa_j]$, the magnitude of $f(x)$ is comparable with the approximation error and is indistinguishable from zero. This is not crucial since the total length of the transition region $\bigcup_{j\in\mathcal{A}_2}[\kappa_{j-1},\kappa_j]$ tends to zero as n goes to infinity.

4. Simulation Studies

To evaluate the finite sample performance of the proposed procedure, we conducted three Monte Carlo simulation studies. In all numerical examples, we compared the proposed procedure (denoted as `Proposed`) with the un-penalized B-spline smoothing (denoted as `Un-Pen`) and the penalized B-splines with different penalties: the ordinary L_1 penalty (denoted as `Lasso`), the adaptive Lasso penalty (denoted as `aLasso`), the SCAD penalty (denoted as `SCAD`), and the group Lasso penalty (denoted as `gLasso`) with the grouping structure defined at (2.3). The group Lasso with overlapping groups is implemented by using the SLEP package (Liu, Ji, and Ye (2009)). Methods proposed by James, Wang, and Zhu (2009) and Zhou, Wang, and Wang (2013), originally developed for functional linear models, might be modified and extended to our problem, but this is not our main focus, and their methods are excluded from our comparisons.

In each example, a Monte Carlo experiment was conducted. For each of the 500 iterations, a data set with n observations was generated from the model

$$y = f(x) + \sigma\epsilon,$$

with ϵ standard normal. We considered sample sizes $n = 200$ and $n = 400$ with noise levels $\sigma = 0.2$ and $\sigma = 0.5$. As commonly adopted, we used cubic splines and set the bridge parameter as $\alpha = 0.5$ in all numerical examples. In each iteration, a 10-fold cross validation was used to select the number of knots, which were used in all competing methods. The regularization parameters were selected by the BIC procedure (2.5).

The performance of estimator $\widehat{f}(\cdot)$ was assessed via several different measures. The overall fitting performance was measured by the square root of average squared errors (RASE)

$$\mathrm{RASE} = \left\{ \frac{1}{n_{\mathrm{grid}}} \sum_{k=1}^{n_{\mathrm{grid}}} \left[\widehat{f}(x_k) - f(x_k) \right]^2 \right\}^{1/2},$$

where $\{x_k,\, k = 1, \ldots, n_{\mathrm{grid}}\}$ are the grid points at which the estimated function $\widehat{f}$ was evaluated. In our simulation, we set $n_{\mathrm{grid}} = 200$ with grid points evenly distributed over the interval. The mean and the standard deviation of RASE

Table 1. Simulation results for Example 1. The mean and standard deviation (in parentheses) of RASE, the average percentage of true zero intervals that were correctly identified (CZ) and the percentage that the estimated functions were zero over the entire domain (GZ) are reported.

	$n = 200$			$n = 400$		
	RASE	**CZ(%)**	**GZ(%)**	**RASE**	**CZ(%)**	**GZ(%)**
			$\sigma = 0.2$			
Un-Pen	0.037 (0.014)	0.00	0.00	0.026 (0.010)	0.00	0.00
Lasso	0.001 (0.005)	97.49	97.00	0.000 (0.003)	98.52	98.00
aLasso	0.002 (0.008)	95.86	94.40	0.001 (0.004)	97.88	96.40
SCAD	0.003 (0.011)	95.13	93.20	0.002 (0.008)	96.23	94.00
gLasso	0.000 (0.004)	98.90	98.60	0.000 (0.003)	99.15	98.80
Proposed	0.002 (0.007)	97.48	95.80	0.001 (0.005)	98.29	96.20
			$\sigma = 0.5$			
Un-Pen	0.092 (0.034)	0.00	0.00	0.065 (0.025)	0.00	0.00
Lasso	0.003 (0.015)	96.72	95.80	0.001 (0.008)	98.49	97.40
aLasso	0.005 (0.020)	95.89	94.60	0.002 (0.010)	98.51	97.40
SCAD	0.007 (0.028)	94.92	93.20	0.003 (0.015)	97.54	95.80
gLasso	0.002 (0.013)	98.00	97.20	0.001 (0.007)	98.95	98.00
Proposed	0.004 (0.018)	97.57	96.00	0.002 (0.011)	98.12	96.60

over the 500 replications are reported. To assess the performance of local sparsity detection, we used two numerical measures: the average percentage of true zero intervals correctly identified (CZ) and the average percentage of nonzero intervals falsely identified as zeros (FZ).

Example 1. To examine the performance of global sparsity detection, we took f to be the zero function. In this example, FZ is not reported because there are no false zeros. Instead, the percentage that the estimated functions were zero over the entire domain (GZ) is reported as an assessment of the performance of global sparsity detection.

The results of 500 Monte Carlo iterations are summarized in Table 1. It can be seen that all penalized estimators have comparable performance of fitting and detecting both global sparsity and local sparsity in this example. Here SCAD has larger average RASE and lower CZ and GZ compared to the other penalized estimators.

Example 2. Here the true f was the smooth function depicted in the center panel of Figure 1. This function lives in the linear space spanned by cubic B-splines with 9 interior knots. Let $\{B_k(x)\}_{k=1}^{13}$ be a set of cubic B-spline basis functions with 9 equally spaced interior knots, and let

$$\boldsymbol{b} = [-1, 3, 0, 0, 0, 0, 0, -1.5, 0, 0, 0, 0, -2]^T$$

Table 2. Simulation results for Example 2. The mean and standard deviation (in parentheses) of RASE, the average percentage of true zero intervals that were correctly identified (CZ) and the average percentage of nonzero intervals that were falsely identified as zeros (FZ) are reported.

	$n = 200$			$n = 400$		
	RASE	**CZ(%)**	**FZ(%)**	**RASE**	**CZ(%)**	**FZ(%)**
			$\boldsymbol{\sigma = 0.2}$			
Un-Pen	0.062 (0.019)	0.00	0.00	0.041 (0.010)	0.00	0.00
Lasso	0.070 (0.029)	25.00	0.00	0.046 (0.016)	20.65	0.00
aLasso	0.048 (0.028)	89.20	0.13	0.029 (0.016)	91.63	0.08
SCAD	0.046 (0.028)	93.02	0.12	0.028 (0.016)	94.02	0.09
gLasso	0.070 (0.029)	23.67	0.00	0.046 (0.016)	20.12	0.00
Proposed	0.045 (0.025)	94.02	0.12	0.028 (0.015)	95.60	0.08
			$\boldsymbol{\sigma = 0.5}$			
Un-Pen	0.153 (0.047)	0.00	0.00	0.103 (0.024)	0.00	0.00
Lasso	0.175 (0.059)	46.40	1.12	0.121 (0.039)	37.62	0.16
aLasso	0.138 (0.056)	85.97	1.21	0.090 (0.039)	89.83	0.59
SCAD	0.127 (0.059)	95.20	1.70	0.078 (0.037)	96.77	0.72
gLasso	0.174 (0.059)	44.72	1.15	0.122 (0.039)	37.23	0.13
Proposed	0.114 (0.052)	96.13	1.35	0.075 (0.030)	96.75	0.45

be the vector of coefficients. With $f(x) = \sum_{k=1}^{13} b_k B_k(x)$, f is zero on both $[0.2, 0.4]$ and $[0.8, 0.9]$.

The results of 500 Monte Carlo iterations are summarized in Table 2. The Lasso and the group Lasso estimators performed poorly in terms of the average percentage of correctly estimated zero intervals (CZ), which suggests that they are too conservative for local sparsity detection. The adaptive Lasso, SCAD, and the proposed estimator performed similarly. The proposed estimator outperformed the adaptive Lasso in terms of CZ and outperformed SCAD in terms of FZ, especially at the high noise level, and had the best overall performance.

To examine the performance of the asymptotic variance formula (2.6), the asymptotic and empirical standard deviations based on 500 replications are shown in Figure 2. In particular, the estimated function was evaluated at 200 grid points during each replication. The (pointwise) standard deviations of the estimated functions, shown in thick black line type, can be regarded as the true standard deviations. The estimated standard deviations in 500 replications are shown in gray color. The asymptotic variance formula performs quite well under all four settings.

Example 3. Here the true f was not from any linear space spanned by B-spline basis functions. We took

$$f(x) = g\Big(x; \frac{9}{2}, \frac{1}{4}, \frac{2}{3}\Big) + g\Big(x; 8, \frac{3}{5}, \frac{1}{3}\Big),$$

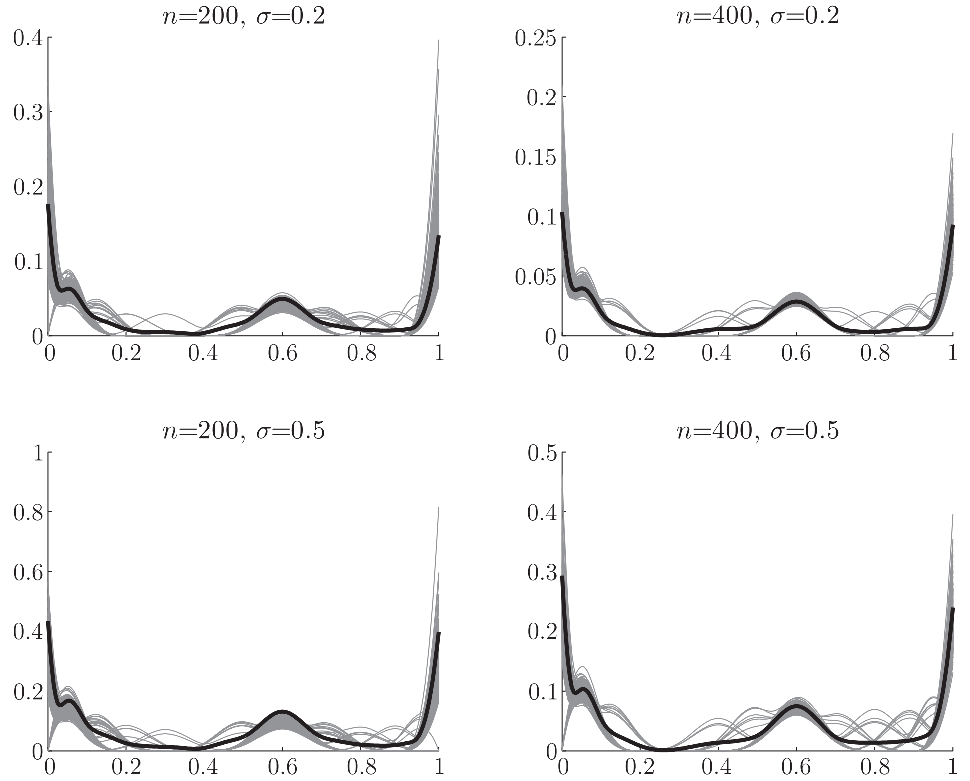

Figure 2. Empirical (black) and asymptotic (gray) pointwise standard deviations of the estimated functions.

where $g(x; a, b, c) = c \cdot (\sin(a\pi(x - b) - \pi/2) + 1) \cdot I(b < x < b + 2/a)$, $I(\cdot)$ the indicator function. This is f_3 in the right panel of Figure 1. It is zero over $[0, 0.25]$ and $[0.85, 1]$.

A summary of the simulation results from 500 iterations is given in Table 3. We have similar observations from Table 3: for local sparsity detection, the Lasso and the group Lasso estimators perform the worst; the adaptive Lasso estimator is better than Lasso and gLasso, but is still not as good as the proposed one; SCAD has competitive local sparsity detection capacity with the proposed one, but its false zeros rate is higher, especially under the high noise settings. Overall, the proposed estimator has the best performance.

5. Data Example

We applied our proposed procedure to the pinch force data set studied in Ramsay, Wang, and Flanagan (1995) and Ramsay and Silverman (2005). The data were collected by R. Flanagan at the MRC Applied Psychology Unit, Cam-

Table 3. Simulation results for Example 3. The mean and standard deviation (in parentheses) of RASE, the average percentage of true zero intervals that were correctly identified (CZ) and the average percentage of nonzero intervals that were falsely identified as zeros (FZ) are reported.

	$n = 200$			$n = 400$		
	RASE	**CZ(%)**	**FZ(%)**	**RASE**	**CZ(%)**	**FZ(%)**
			$\sigma = 0.2$			
Un-Pen	0.065 (0.021)	0.00	0.00	0.043 (0.009)	0.00	0.00
Lasso	0.054 (0.014)	75.99	0.00	0.039 (0.008)	78.95	0.00
aLasso	0.049 (0.015)	88.59	0.00	0.035 (0.008)	89.32	0.00
SCAD	0.049 (0.017)	91.58	0.00	0.035 (0.008)	89.84	0.00
gLasso	0.055 (0.014)	74.94	0.00	0.040 (0.009)	78.66	0.00
Proposed	0.048 (0.013)	93.15	0.00	0.035 (0.008)	91.17	0.00
			$\sigma = 0.5$			
Un-Pen	0.158 (0.046)	0.00	0.00	0.107 (0.022)	0.00	0.00
Lasso	0.119 (0.037)	79.90	0.07	0.084 (0.023)	81.41	0.00
aLasso	0.118 (0.042)	84.67	0.38	0.080 (0.028)	90.99	0.06
SCAD	0.122 (0.056)	90.16	0.58	0.077 (0.029)	95.57	0.08
gLasso	0.118 (0.035)	76.90	0.15	0.084 (0.023)	79.28	0.00
Proposed	0.107 (0.035)	94.65	0.35	0.076 (0.023)	95.50	0.04

bridge: twenty records of the force exerted on an object over time. A scatterplot of the data is displayed in Figure 3(a). Our interest here is to estimate the overall mean function. It can be seen that the curve is possibly sparse near both tails, expected given the nature of this experiment.

We fit the overall mean curve using `Un-Pen`, `Lasso`, `aLasso`, `SCAD`, `gLasso`, and `Proposed`. The number of interior knot points ($M = 20$) was selected by 10-fold cross-validation. The regularization parameters were selected by the BIC procedure (2.5). The fitted curves are depicted in Figure 3(b). Overall, the fitted mean curves from all methods were close to each other. A zoomed-in view in Figure 3(c) highlights the difference (around 0.145 seconds) in local sparsity among the fitted curves. The proposed procedure yields a more sparse fit, as desired over both tails, while the other five methods fail to produce sparse estimated curves. Figure 3(d) shows the fitted curve from the proposed method (solid line) with one pointwise standard deviation (dashed line) above and below.

Table 4 provides the estimated coefficients from the methods. It can be seen that `Lasso` and `gLasso` yield almost the same estimated coefficients (after rounding), which have least sparsity among the penalized methods. `aLasso` and `SCAD` produce more zeros in their estimated coefficients, but they only detect sparsity on the left tail, not on the right one, while the proposed method is the one that detects sparsity on both tails.

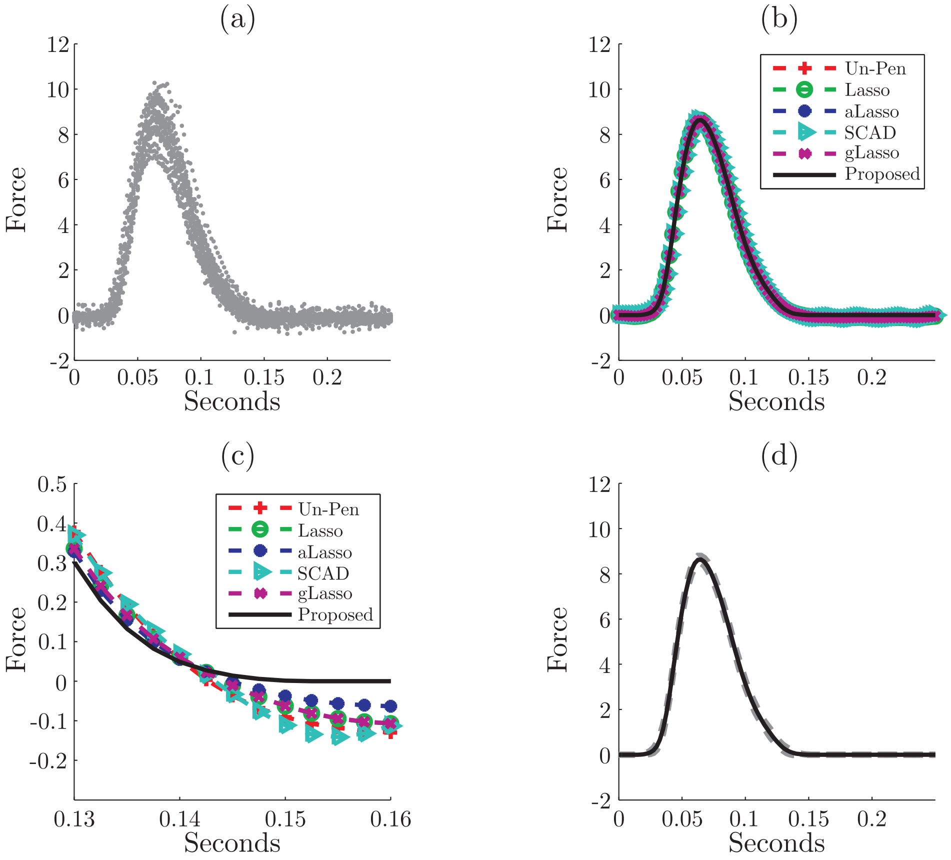

Figure 3. Data example. (a) A scatterplot of the original pinch force data. (b) Fitted mean curves from six competing methods: un-penalized method, Lasso, adaptive Lasso, SCAD, group Lasso and the proposed method. (c) A zoomed-in view of the fitted curves, as shown in (b), around 0.145 seconds. (d) The fitted mean curve from the proposed method (solid line) with one pointwise standard deviation above and below (dashed lines).

6. Conclusion and Discussion

We study the problem of sparse estimation in nonparametric regression, both global and local. We propose a one-step penalized least squares procedure, based on the basis function approximation and the group bridge penalty, for simultaneous functional estimation and model selection. We establish consistency in estimation and sparsistency in model selection of our proposed estimator. In simulation studies and a data example, we compare the proposed estimator with the un-penalized one and other alternatives. The results show that the proposed estimator performs the best among all, in terms of sparsity detection as well as the overall fitting to the true curve.

Table 4. Estimated coefficients for the pinch force data. Our proposed method can identify local sparsity near both tails.

$\widehat{\gamma}$	Un-Pen	Lasso	aLasso	SCAD	gLasso	Proposed
$\widehat{\gamma}_1$	-0.09	0	0	0	0	0
$\widehat{\gamma}_2$	-0.06	0	0	0	0	0
$\widehat{\gamma}_3$	-0.05	-0.09	0	0	-0.09	0
$\widehat{\gamma}_4$	-0.12	0	0	0	0	0
$\widehat{\gamma}_5$	0.60	0.48	0.48	0.52	0.48	0.46
$\widehat{\gamma}_6$	5.89	5.92	5.96	5.94	5.92	5.96
$\widehat{\gamma}_7$	9.08	9.02	9.04	9.05	9.02	9.03
$\widehat{\gamma}_8$	8.41	8.40	8.43	8.43	8.40	8.43
$\widehat{\gamma}_9$	6.51	6.47	6.49	6.50	6.47	6.48
$\widehat{\gamma}_{10}$	3.81	3.80	3.83	3.82	3.80	3.83
$\widehat{\gamma}_{11}$	2.05	2.01	2.02	2.05	2.01	2.00
$\widehat{\gamma}_{12}$	0.91	0.92	0.96	0.93	0.92	0.96
$\widehat{\gamma}_{13}$	0.29	0.22	0.19	0.27	0.22	0.15
$\widehat{\gamma}_{14}$	-0.04	0	0	0	0	0
$\widehat{\gamma}_{15}$	-0.13	-0.11	-0.07	-0.21	-0.11	0
$\widehat{\gamma}_{16}$	-0.13	-0.11	-0.06	0	-0.11	0
$\widehat{\gamma}_{17}$	-0.16	-0.14	-0.15	-0.27	-0.14	0
$\widehat{\gamma}_{18}$	-0.10	-0.08	0	0	-0.08	0
$\widehat{\gamma}_{19}$	-0.16	-0.14	-0.14	-0.19	-0.14	0
$\widehat{\gamma}_{20}$	-0.13	-0.11	-0.09	-0.17	-0.10	0
$\widehat{\gamma}_{21}$	-0.11	-0.09	0	0	-0.09	0
$\widehat{\gamma}_{22}$	-0.17	-0.14	-0.17	-0.30	-0.14	0
$\widehat{\gamma}_{23}$	-0.12	-0.06	0	0	-0.06	0
$\widehat{\gamma}_{24}$	-0.13	-0.04	0	0	-0.04	0

The proposed method can be extended to other regression models that contain nonparametric components, such as nonparametric additive models, varying coefficient models, semiparametric models, and even functional linear models. A reviewer noted that our proposed method can also be extended to estimate globally or locally constant functions by modifying the choice of penalty function. If a function $f(x) \in \mathbb{G}$ (the linear space of spline functions) is a constant between two adjacent knots, say $[\kappa_{j-1}, \kappa_j]$, we have $\gamma_j = \cdots = \gamma_{j+d}$. This suggests that we can group the coefficients $\{\gamma_j, \ldots, \gamma_{j+d}\}$ and detect their equality. Correspondingly, we can modify the group penalty by replacing the L_1-norm $\|\boldsymbol{\gamma}_{A_j}\|_1$ by $|\gamma_{j+1} - \gamma_j| + \cdots + |\gamma_{j+d} - \gamma_{j+d-1}|$. From a computational aspect, after a transformation of the B-spline basis functions, the optimization can be solved using a similar algorithm as we proposed in Section 2.3. Detailed discussion is provided in the online supplement.

Acknowledgement

The authors are very grateful to the Editor, an Associate Editor, and two referees for their constructive comments and suggestions. The research of Haonan Wang was partially supported by NSF grant DMS-1106975.

Supplementary Material

The online supplementary material includes proofs of the theoretical results and additional discussion of our proposed method.

References

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* **37**, 1705-1732.

Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35**, 2313-2351.

de Boor, C. (1978). *A Practical Guide to Splines.* Springer, New York.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.* **32**, 407-499.

Eubank, R. L. (1999). *Nonparametric Regression and Spline Smoothing.* Second Edition, Marcel Dekker, New York.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and its Application.* Chapman & Hall, London.

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

Friedman, J., Hastie, T. and Tibshirani, R. (2010). A note on the group lasso and a sparse group lasso. arXiv preprint arXiv:1001.0736.

Härdle, W. (1990). *Applied Nonparametric Regression.* Cambridge University Press, United Kingdom.

Huang, J., Horowitz, J. L. and Wei, F. (2010). Variable selection in nonparametric additive models. *Ann. Statist.* **38**, 2282-2313.

Huang, J., Ma, S., Xie, H. and Zhang, C.-H. (2009). A group bridge approach for variable selection. *Biometrika* **96**, 339-355.

Huang, J. Z. (2001). Concave extended linear modeling: A theoretical synthesis. *Statist. Sinica* **11**, 173-197.

Huang, J. Z. (2003). Local asymptotics for polynomial spline regression. *Ann. Statist.* **31**, 1600-1635.

Huang, J. Z., Wu, C. O. and Zhou, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statist. Sinica* **14**, 763-788.

Jacob, L., Obozinski, G. and Vert, J.-P. (2009). Group lasso with overlap and graph lasso. In *Proceedings of the* 26*th Annual International Conference on Machine Learning.* ACM, 433-440.

James, G., Radchenko, P. and Lv, J. (2009). DASSO: connections between the Dantzig selector and lasso. *J. Roy. Statist. Soc. Ser. B* **71**, 127-142.

James, G., Wang, J. and Zhu, J. (2009). Functional linear regression thats interpretable. *Ann. Statist.* **37**, 2083-2108.

Liu, J., Ji, S. and Ye, J. (2009). *SLEP: Sparse Learning with Efficient Projections.* Arizona State University.

Liu, J. and Ye, J. (2010). Fast overlapping group lasso. arXiv preprint arXiv:1009.0306.

Percival, D. (2012). Theoretical properties of the overlapping groups lasso. *Electronic J. Statist.* **6**, 269-288.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis.* Springer, New York.

Ramsay, J. O., Wang, X. and Flanagan, R. (1995). A functional data analysis of the pinch force of human fingers. *Appl. Statist.* **44**, 17-30.

Schumaker, L. L. (1981). *Spline Functions: Basic Theory.* Wiley.

Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. (2013). A sparse-group lasso. *J. Comput. Graph. Statist.* **22**, 231-245.

Song, D., Chan, R. H., Marmarelis, V. Z., Hampson, R. E., Deadwyler, S. A. and Berger, T. W. (2007a), Nonlinear dynamic modeling of spike train transformations for hippocampal-cortical prostheses. *IEEE Trans. Biomed. Engin.* **54**, 1053-1066.

Song, D., Chan, R. H., Marmarelis, V. Z., Hampson, R. E., Deadwyler, S. A. and Berger, T. W. (2007b). Statistical selection of multiple-input multiple-output nonlinear dynamic models of spike train transformation. *Proceedings of the* 29*th Annual International Conference of the IEEE EMBS.*

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10**, 1040-1053.

Stone, C. J. (1985). Additive regression and other nonparametric models'. *Ann. Statist.* **13**, 689-705.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Amer. Statist. Assoc.* **58**, 267-288.

Tu, C. Y., Song, D., Breidt, J. F., Berger, T. W. and Wang, H. (2012). *Functional Model Selection for Sparse Binary Time Series with Multiple Inputs.* Chapman and Hall/CRC.

Wahba, G. (1990). *Spline Models for Observational Data.* Society for Industrial Mathematics.

Wand, M. P. and Jones, M. C. (1995). *Kernel Smoothing.* Chapman and Hall/CRC.

Wang, L., Li, H. and Huang, J. Z. (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *J. Amer. Statist. Assoc.* **103**, 1556-1569.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. Roy. Statist. Soc. Ser. B* **68**, 49-67.

Zhou, J., Wang, N.-Y. and Wang, N. (2013). Functional linear model with zerovalue coefficient function at sub-regions. *Statist. Sinica* **23**, 25-50.

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

Department of Statistics, Colorado State University, Fort Collins, Colorado 80523, U.S.A.

E-mail: wanghn@stat.colostate.edu

Department of Mathematics, College of Charleston, Charleston, SC 29424, U.S.A.

E-mail: kaib@cofc.edu

(Received January 2014; accepted August 2014)