Statistica Sinica **23** (2013), 25-50
doi:http://dx.doi.org/10.5705/ss.2010.237

FUNCTIONAL LINEAR MODEL WITH ZERO-VALUE COEFFICIENT FUNCTION AT SUB-REGIONS

Jianhui Zhou, Nae-Yuh Wang and Naisyin Wang

University of Virginia, Johns Hopkins University and University of Michigan

Abstract: We propose a shrinkage method to estimate the coefficient function in a functional linear regression model when the value of the coefficient function is zero within certain sub-regions. Besides identifying the null region in which the coefficient function is zero, we also aim to perform estimation and inferences for the nonparametrically estimated coefficient function without over-shrinking the values. Our proposal consists of two stages. In stage one, the Dantzig selector is employed to provide initial location of the null region. In stage two, we propose a group SCAD approach to refine the estimated location of the null region and to provide the estimation and inference procedures for the coefficient function. Our considerations have certain advantages in this functional setup. One goal is to reduce the number of parameters employed in the model. With a one-stage procedure, it is needed to use a large number of knots in order to precisely identify the zero-coefficient region; however, the variation and estimation difficulties increase with the number of parameters. Owing to the additional refinement stage, we avoid this necessity and our estimator achieves superior numerical performance in practice. We show that our estimator enjoys the Oracle property; it identifies the null region with probability tending to 1, and it achieves the same asymptotic normality for the estimated coefficient function on the non-null region as the functional linear model estimator when the non-null region is known. Numerically, our refined estimator overcomes the shortcomings of the initial Dantzig estimator which tends to under-estimate the absolute scale of non-zero coefficients. The performance of the proposed method is illustrated in simulation studies. We apply the method in an analysis of data collected by the Johns Hopkins Precursors Study, where the primary interests are in estimating the strength of association between body mass index in midlife and the quality of life in physical functioning at old age, and in identifying the effective age ranges where such associations exist.

Key words and phrases: B-spline basis function, functional linear regression, group smoothly clipped absolute deviation approach, null region.

1. Introduction

We study the functional linear regression (FLR) model

$$Y_i = \mu + \sum_{d=1}^{D} \int_0^T X_{id}(t)\beta_d(t)dt + e_i, \tag{1.1}$$

where Y_i denotes the ith response, $X_{id}(t)$ are realizations of random processes $X_d(t)$, $\beta_d(t)$ are the corresponding smooth coefficient functions on $[0, T]$, and $e_i \sim N(0, \sigma^2)$ are random errors independent of $X_{id}(t)$, $i = 1, \ldots, n$. The response Y reflects the weighted cumulative effects of functional predictors $X_d(t)$, and the coefficient functions $\beta_d(t)$ represent the corresponding weights. In practice, it is often of interest to know which areas of $X_d(t)$ contribute to the value of Y and in what magnitude. That is, we are interested in learning the null region in which $\beta_d(t) = 0$, and in estimating the values of $\beta_d(t)$ when they are non-zero.

Regression models with functional predictors have application in functional data analysis (FDA), and lately in longitudinal data analysis (LDA) when the longitudinal covariate measurements are collected intensively. Ramsay and Silverman (2005) and Ferraty and Vieu (2006) reviewed theoretical and methodological developments and gave many examples. A non-exhaustive list of recent works includes the followings. Estimation of $\beta_d(t)$ with a spline approach was proposed by Cardot, Ferraty, and Sarda (2003). Crambes, Kneip, and Sarda (2009) proposed a smoothing spline estimator of $\beta_d(t)$ with a new penalty term to ensure existence of the estimator, and studied its asymptotic behavior. Fan and Zhang (2000) studied the FLR problem with a functional response. Cai and Hall (2006) investigated prediction issues in FLR. With an additional link function in model (1.1), Müller and Stadtmüller (2005) studied the generalized functional linear model. Yao, Müller, and Wang (2005) extended the scope of the problem to cover longitudinal data. James, Wang, and Zhu (2009) emphasized the importance of the interpretability of $\beta_d(t)$ and proposed to use a version of the Dantzig selector (Candes and Tao (2007)) for this purpose. They equated the problem of identifying zero-value regions of the corresponding order derivative of $\beta_d(t)$ to that of variable selection in a multiple linear regression setting; however, to precisely identify the null region of $\beta_d(t)$, the Dantzig selector needs to use a large number of knots, and the quality of the estimated $\beta_d(t)$ deteriorates on the non-null region with the increasing number of knots. It is known that with the number of parameters increasing and increasing with sample size, the variation of estimation increases. We illustrate this phenomenon in our specific setting in Section 4. Further, the asymptotic distribution, a property tends to be desired by a functional data analysis approach, is not reported for their estimator. We derived the asymptotic distribution for our proposed estimator in Section 3.

In this paper, we propose a two-stage estimator to simultaneously identify the null region of $\beta_d(t)$ and estimate $\beta_d(t)$ on the non-null region. The goal behind this approach is to avoid having a large number of parameters in either stage and still maintain a high quality of estimation performance. We roughly identify the null region at stage one, and adaptively regularize the estimate of $\beta_d(t)$ on the null and non-null regions at stage two, applying different group penalties. At

stage one, an initial estimator identifies and preferably over-estimates the null region; the desired precision is reached at the second stage. When the number of parameters is large, a direct implementation of the Dantzig selector gives poor estimation of the coefficient function in the non-null region. Our two-stage procedure simplies the problem, naturally reduces over-shrinking of the coefficient functions in the non-null region, and achieves superior numerical performance.

We structure the paper as follows. In Section 2, we present our proposed method as a B-spline approximation coupled with shrinkage, which takes the advantage of the local property of B-spline basis functions. In Section 3, we show that the proposed estimator enjoys the Oracle property and we give its asymptotic distribution. Simulation studies are reported in Section 4. In Section 5, we apply the proposed method to data from the Johns Hopkins Precursors Study to investigate the effect of body mass index at midlife on a quality of life index for physical functioning at old age. Concluding remarks are given in Section 6. Assumptions for the theoretical properties and sketch of the proofs are provided in the Appendix. The exact algorithms to implement the proposed method, detailed proofs, and additional numerical results are reported in a supplementary document.

2. Estimation of Coefficient Functions and Their Null Regions

To simplify the notation and presentation, we take $D = 1$, suppress the subscript d in model (1.1), and write

$$Y_i = \mu + \int_0^T X_i(t)\beta(t)dt + e_i. \tag{2.1}$$

We assume $e_i \sim N(0, \sigma^2)$ and $\mu = 0$. Without loss of generality, we let $\sigma = 1$. Generalization of the method to the cases $D > 1$ is discussed in Section 2.3. The asymptotic properties under (2.1) can be extended to model (1.1).

As in James, Wang, and Zhu (2009), we assume that the processes $X_i(t)$ are known while, in practice, $X_i(t)$ are usually not completely observable. Instead, the observations (Y_i, t_{ij}, X_{ij}) are available for $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$, where $X_{ij} = X_i(t_{ij})$. In practice, $X_i(t)$ are measured at discrete and perhaps irregular time points and it is common to include a pre-smoothing step. See Ramsay and Silverman (2005) for insight and illustrations, and Hall and Van Keilegom (2008) for theoretical considerations. We report the effects of pre-smoothing in the numerical studies.

For estimating the coefficient function, $\beta(t)$, as well as identifying its null region, denoted by $\mathcal{T}$, we use B-splines. Given $k_{0,n} + 1$ evenly-spaced knots, $0 = \tau_0 < \tau_1 < \tau_2 < \cdots < \tau_{k_{0,n}-1} < \tau_{k_{0,n}} = T$, let $I_j = [\tau_{j-1}, \tau_j]$ for $j =$

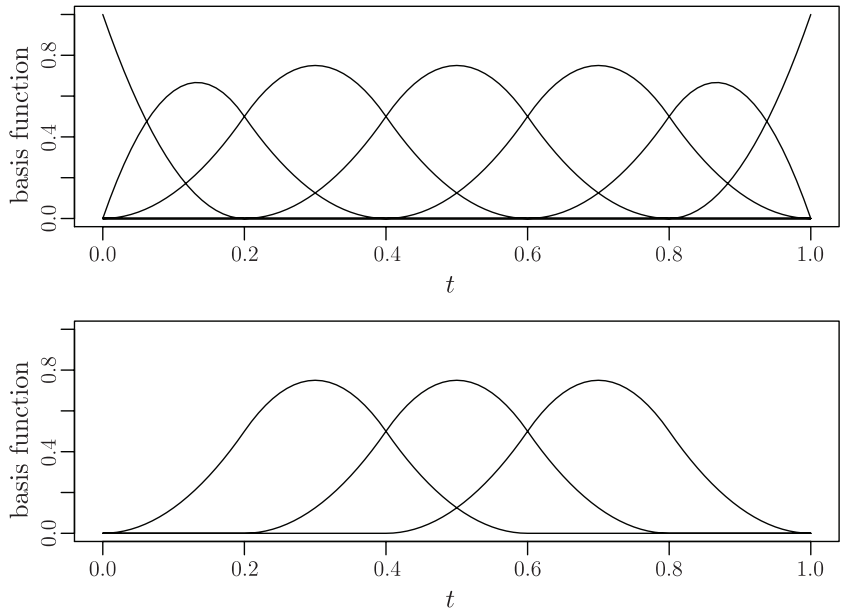

Figure 1. Local support of B-spline basis function.

$1, \ldots, k_{0,n}$. Associated with this set of knots, there are $(k_{0,n} + h)$ B-spline basis functions, $\boldsymbol{B}_0(t) = (B_{0,1}(t), \ldots, B_{0,k_{0,n}+h}(t))^T$, each of which is a piecewise polynomial of degree h with support on at most $h + 1$ subintervals I_j. The upper panel of Figure 1 shows the seven basis functions with $h = 2$ and knots $\{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$. Given the sample size n and the $k_{0,n} + 1$ knots, the coefficient function can be expressed as

$$\beta(t) = \boldsymbol{B}_0^T(t)\boldsymbol{b}_0 + e_0(t), \tag{2.2}$$

where $\boldsymbol{b}_0 = (b_{0,1}, \ldots, b_{0,k_{0,n}+h})^T$ is the B-spline approximation coefficient, $e_0(t)$ is an approximation error that is uniformly bounded on $[0, T]$ with the bound going to 0 as $k_{0,n}$ goes to infinity. For details on B-spline approximation, see Schumaker (1981). Using the B-spline approximation for $\beta(t)$ in (2.2), (2.1) can be re-written as

$$Y_i = \boldsymbol{z}_{0,i}^T\boldsymbol{b}_0 + \epsilon_{0,i}, \text{ or } \boldsymbol{Y} = \boldsymbol{Z}_0\boldsymbol{b}_0 + \boldsymbol{\epsilon}_0, \tag{2.3}$$

where $\boldsymbol{z}_{0,i}$ is a $(k_{0,n} + h) \times 1$ vector with the jth element $\int_0^T X_i(t)B_{0,j}(t)dt$, $\epsilon_{0,i} = e_i + \int_0^T X_i(t)e_0(t)dt$, and $\boldsymbol{Y}$ and $\boldsymbol{Z}_0$ are, respectively, the $n \times 1$ vector and $n \times (k_{0,n} + h)$ matrix that contain Y_i and $\boldsymbol{z}_{0,i}$ as entries.

Note that the choices of $\boldsymbol{B}_0(t)$, I_j, and consequently, the values of $\boldsymbol{b}_0$, $e_0(t)$, $\epsilon_{0,i}$ and $\boldsymbol{z}_{0,i}$, vary with the sample size n. To simplify notation, the subscript n

is suppressed; we use the subscript 0 to indicate the association with the initial stage.

2.1. Initial estimate of the null region

It is convenient to let the end points of $\mathcal{T}$ fall on the end points of certain subintervals I_j. With the initial knots $\{\tau_j\}_{j=0}^{k_{0,n}}$, assume each I_j is entirely contained either in $\mathcal{T}$ or in $\mathcal{T}^c$, the complement of $\mathcal{T}$. Having no prior information on the location of $\mathcal{T}$, we use a moderate number of evenly-spaced internal knots on $[0, T]$. The initial location of $\mathcal{T}$ can be roughly established through I_j and an estimate of $\boldsymbol{b}_0$ in (2.3).

We use the Dantzig selector to estimate $\boldsymbol{b}_0$ at this stage, $\text{argmin}_{\boldsymbol{b}}||\boldsymbol{b}||_{l_1}$ subject to $|\boldsymbol{Z}_{0,k}^T(\boldsymbol{Y} - \boldsymbol{Z}_0\boldsymbol{b})| \leq \lambda_D$, with $\boldsymbol{Z}_{0,k}^T$ being the kth column of $\boldsymbol{Z}_0$, $\lambda_D = \sqrt{2\log(k_{0,n}+h)}$, and $||\boldsymbol{b}||_{l_1}$ denoting the L_1 norm of $\boldsymbol{b}$. We denote this initial estimate of $\boldsymbol{b}_0$ by $\tilde{\boldsymbol{b}}_0$. The conditions of equivalence between the Dantzig selector and LASSO have been reported in James, Radchenko, and Lv (2009), Bickel, Ritov, and Tsybakov (2009), and references therein. It is not essential to have high precision during this stage, so other regularization estimators such as LASSO or SCAD can be used here, even though the simulation outcomes from James, Wang, and Zhu (2009) imply numerical advantages of the Dantzig selector over LASSO. For further details on the Dantzig selector, see Candes and Tao (2007).

With the B-spline basis supported on at most $h+1$ subintervals, the value of $\boldsymbol{B}_0^T(t)\boldsymbol{b}_0$ on a single subinterval I_j is determined by $h+1$ coefficients in $\boldsymbol{b}_0$. For example, in Figure 1, the estimated $\beta(t)$ for any $t \in [0.4, 0.6]$ depends only on the coefficients of the three basis functions in the lower panel. When the $h+1$ coefficients associated with I_j are all 0, the subinterval I_j is contained in the null region of $\boldsymbol{B}_0^T(t)\boldsymbol{b}_0$; otherwise, it is in the non-null region. In practice, even if $I_j \subset \mathcal{T}$, its associated $h+1$ coefficients in $\boldsymbol{b}_0$ might not all be 0, but we only need a rough estimate of null region at this stage. We simply use a small threshold value at the initial stage to identify the subintervals within the null region of $\beta(t)$. Thus, if the absolute values of all $h+1$ coefficients are smaller than d_n, we classify the corresponding I_j as part of $\mathcal{T}$. The union of all identified subintervals is taken as the initial estimate of $\mathcal{T}$, denoted by $\hat{\mathcal{T}}^{(0)}$. We let the threshold value d_n go to 0 as n goes to infinity. The rates of d_n and $k_{0,n}$ are given and discussed in Section 3, and their numerical determination is in Section 4.

2.2. Null region refinement and function estimation

The second stage of our estimator refines the estimated location of $\mathcal{T}$ and the estimate of $\beta(t)$ on $\mathcal{T}^c$. We develop a grouped smoothly-clipped absolute deviation (SCAD) method and a boundary grid-search algorithm at the second

stage to refine the null region and to achieve the estimate of $\beta(t)$ on $\mathcal{T}^c$. The asymptotic distribution of the estimated $\beta(t)$ can be naturally established. Our estimator overcomes the concerns of large numbers of parameters, maintains the sparse property, readily adopts the existing efficient computation algorithm, and achieves desired numerical and theoretical qualities.

In Stage 1, $k_{0,n}+1$ evenly-spaced internal knots are placed in $[0, T]$ to identify the initial estimate of $\mathcal{T}$; in Stage 2, we use a grid-search based algorithm to find the refined null region within $\hat{\mathcal{T}}^{(0)}$ by examining a sequence of working null regions $\mathcal{T}_w \subseteq \hat{\mathcal{T}}^{(0)}$. We let the search algorithm and the penalty determination in grouped SCAD share the same objective function so that we can conduct the evaluation jointly. A practical algorithm of specifying a sequence of $\mathcal{T}_w$'s is given in the supplementary document.

Having $\hat{\mathcal{T}}^{(0)}$, we remove all initial knots within $\hat{\mathcal{T}}^{(0)}$, and place $k_{1,n}+1$ evenly-spaced knots on $\hat{\mathcal{T}}^{(0),c}$, with $k_{1,n} < k_{0,n}$ in general. With the grid size in the boundary search procedure not related to $k_{1,n}$, the determination of grid size is a numerical decision. The smaller grid size gives more precise boundary but can lead to a computationally more demanding task. In our simulation studies, we used a grid size of $0.02T$, where T is the range of t.

Using the new set of knots corresponding to each $\mathcal{T}_w$, we generate a new set of B-spline basis functions $\boldsymbol{B}_1^w(t)$ and the corresponding new variables, $\boldsymbol{z}_{1,i}^w$, following the same procedure described below (2.3). Moreover, (2.1) can be re-written as

$$Y_i = \boldsymbol{z}_{1,i}^{wT}\boldsymbol{b}_1^w + \epsilon_{1,i}^w, \text{ or } \boldsymbol{Y} = \boldsymbol{Z}_1^w\boldsymbol{b}_1^w + \boldsymbol{\epsilon}_1^w, \tag{2.4}$$

where $\boldsymbol{b}_1^w$, $\boldsymbol{z}_{1,i}^w$ and $\boldsymbol{Z}_1^w$ are equivalently defined as $\boldsymbol{b}_0$, $\boldsymbol{z}_{0,i}$ and $\boldsymbol{Z}_0$, respectively, in (2.3), and the ith entry of $\boldsymbol{\epsilon}_1^w$ is $\epsilon_{1,i}^w = e_i + \int_0^T X_i(t)e_1^w(t)dt$, with the approximation error $e_1^w(t) = \beta(t) - \boldsymbol{B}_1^{wT}(t)\boldsymbol{b}_1^w$. As in (2.3), $\boldsymbol{Z}_1^w$, $\boldsymbol{b}_1^w$, $\boldsymbol{B}_1^w(t)$, and $\boldsymbol{\epsilon}_1^w$ vary with the sample size n. The subscript n is suppressed to simplify the notation, and the subscript 1 indicates the association with the refinement stage.

To estimate $\boldsymbol{b}_1^w$, we propose a group penalized least squares method. According to $\mathcal{T}_w$, the coefficients in $\boldsymbol{b}$ are divided into groups $\boldsymbol{b}_{N,w}$ (null) and $\boldsymbol{b}_{S,w}$ (signal). Specifically, if a subinterval is part of $\mathcal{T}_w$, then all coefficients b_i that are associated with this subinterval are put into $\boldsymbol{b}_{N,w}$; the group $\boldsymbol{b}_{S,w}$ contains the remainder.

Different penalty functions are available in the literature, including the L_2 penalty of ridge regression, the L_1 penalty of LASSO regression (Tibshirani (1996)), and the SCAD penalty function (Fan and Li (2001)). We use the SCAD penalty function with coefficients in the two groups $\boldsymbol{b}_{N,w}$ and $\boldsymbol{b}_{S,w}$ penalized separately. Using the L_2 penalty does not help us to identify $\mathcal{T}$, and Zou (2006) pointed out that the LASSO estimator is not consistent, which prevented us from

considering the group LASSO (Yuan and Lin (2006)). That SCAD penalizes less on coefficients with large absolute values allows the non-zero coefficient functions to be better estimated. With the division of $\boldsymbol{b}$ into $\boldsymbol{b}_{N,w}$ and $\boldsymbol{b}_{S,w}$, $\boldsymbol{b}_1^w$ is estimated as the minimizer of the objective function

$$\sum_{i=1}^{n}(Y_i - \boldsymbol{z}_{1,i}^w \boldsymbol{b})^2 + n\left\{p_\lambda(||\boldsymbol{b}_{N,w}||_{l_1}) + p_\lambda(||\boldsymbol{b}_{S,w}||_{l_1})\right\},$$

where $p_\lambda(\cdot)$ is the SCAD penalty of Fan and Li (2001), defined through its derivative $p'_\lambda(|\theta|) = \lambda\{I(|\theta| \leq \lambda) + [(a\lambda - |\theta|)_+/(a-1)\lambda]I(|\theta| > \lambda)\}$; a is usually taken as 3.7, and λ is a tuning parameter selected by the criterion $C(\mathcal{T}_w, \lambda)$, which can be the generalized cross validation criterion (GCV), Akaike's information criterion (AIC), the Bayesian information criterion (BIC; Schwarz), or the residual information criterion (RIC), as specified in the supplementary document. By calculating a criterion for each working $\mathcal{T}_w$, we simultaneously select the $\mathcal{T}_w$ and λ, that minimize it. Penalizing the coefficients, $\boldsymbol{b}$, in groups helps to shrink the coefficients in $\boldsymbol{b}_{N,w}$ to zero, simultaneously.

Given an initial value of $\boldsymbol{b}_1^w$, the local quadratic approximation (LQA) is used in the algorithm of Fan and Li (2001) to approximate the penalty function. However, as pointed out in Zou and Li (2008), the LQA estimator does not provide a sparse representation. Instead, Zou and Li (2008) proposed a one-step local linear approximation (LLA) and converted the SCAD problem to a LASSO regression that utilized the LARS algorithm of Efron et al. (2004) to get the sparse estimate. We use the LLA for the group penalty in our method for the same reason.

Let $\tilde{\boldsymbol{b}}_1$ be the initial estimate, which can be the ordinary least squares estimator, and $\tilde{\boldsymbol{b}}_{1N,w}$ and $\tilde{\boldsymbol{b}}_{1S,w}$ be the coefficients inside and outside the working null region $\mathcal{T}_w$, respectively. We approximate the group penalty $p_\lambda(||\boldsymbol{b}_{N,w}||_{l_1})$ by

$$p_\lambda(||\boldsymbol{b}_{N,w}||_{l_1}) \approx p_\lambda(||\tilde{\boldsymbol{b}}_{1N,w}||_{l_1}) + p'_\lambda(||\tilde{\boldsymbol{b}}_{1N,w}||_{l_1})\{||\boldsymbol{b}_{N,w}||_{l_1} - ||\tilde{\boldsymbol{b}}_{1N,w}||_{l_1}\}.$$

The penalty $p_\lambda(||\boldsymbol{b}_{S,w}||_{l_1})$ is approximated equivalently. Using these approximations, we estimate $\boldsymbol{b}_1^w$ by minimizing the objective function

$$Q_n(\mathcal{T}_w, \lambda, \boldsymbol{b}) = \sum_{i=1}^{n}(Y_i - \boldsymbol{z}_{1,i}^w \boldsymbol{b})^2 + n\left\{p'_\lambda(||\tilde{\boldsymbol{b}}_{1N,w}||_{l_1})||\boldsymbol{b}_{N,w}||_{l_1} + p'_\lambda(||\tilde{\boldsymbol{b}}_{1S,w}||_{l_1})||\boldsymbol{b}_{S,w}||_{l_1}\right\}. \tag{2.5}$$

With $||\boldsymbol{b}_{N,l}||_{l_w}$ and $||\boldsymbol{b}_{S,l}||_{l_w}$ as the L_1 norms of $\boldsymbol{b}_{N,w}$ and $\boldsymbol{b}_{S,w}$ and $p'_\lambda(||\tilde{\boldsymbol{b}}_{1N,w}||_{l_1})$ and $p'_\lambda(||\tilde{\boldsymbol{b}}_{1S,w}||_{l_1})$ as weights, the task of minimizing $Q_n(\boldsymbol{b})$ can be carried out using the adaptive LASSO of Zou (2006) and the efficient LARS algorithm of Efron et al. (2004). By the LLA, the coefficients within the same group are penalized

with the weights according to their group memberships. When $p'_\lambda(||\tilde{\boldsymbol{b}}_{1S,w}||_{l_1}) \to 0$ as $\lambda \to 0$, the coefficients in $\boldsymbol{b}_{S,w}$ are barely penalized. As a result, the estimation bias of $\beta(t)$ on $\mathcal{T}_w^c$, induced by the shrinkage penalty, can be greatly reduced.

For a fixed dimension of the regression parameters, we note that Wang, Chen, and Li (2007) developed an alternative group SCAD estimator using the L_2-norm and the local quadratic approximation. Their estimator does not have a sparse representation, so our procedure is more suitable here.

To summarize, we take the refined estimate of the null region, $\hat{\mathcal{T}}$, as

$$\hat{\mathcal{T}} = \arg \min_{\mathcal{T}_w \subseteq \hat{\mathcal{T}}^{(0)}} C(\mathcal{T}_w, \arg\min_{\lambda>0} C(\mathcal{T}_w, \lambda)). \tag{2.6}$$

With

$$\hat{\boldsymbol{b}}_1 = \arg\min_{\boldsymbol{b}} Q_n(\hat{\mathcal{T}}, \arg\min_{\lambda>0} C(\hat{\mathcal{T}}, \lambda), \boldsymbol{b}), \tag{2.7}$$

the refined estimate of $\beta(t)$ is

$$\hat{\beta}(t) = \boldsymbol{B}_1^T(t)\hat{\boldsymbol{b}}_1,$$

where $\boldsymbol{B}_1(t)$ are the B-spline basis functions associated with $\hat{\mathcal{T}}$.

2.3. Generalization to $D > 1$

We have taken $D = 1$ in (1.1). When $D > 1$, all steps can be carried out without much modification, as follows. After obtaining variables $\boldsymbol{z}_{0,i}$ for each $\beta_d(t)$, the Dantzig selector can be applied to obtain initial estimates simultaneously for each $\beta_d(t)$ by including the variables $\boldsymbol{z}_{0,i}$ of all $\beta_d(t)$ in (2.3). The adaptive knots and the variables $\boldsymbol{z}_{1,i}$ in (2.4) can then be obtained one by one for each d. Finally, after having the variables $\boldsymbol{z}_{1,i}^w$ of all $\beta_d(t)$ in (2.4), the refinement procedure with group SCAD is performed with the coefficient $\boldsymbol{b}$ in $Q_n(\mathcal{T}_w, \lambda, \boldsymbol{b})$ partitioned into $2D$ groups, two groups being associated with each $\beta_d(t)$.

3. Oracle Property

In this section, we show that, under certain conditions, our estimator enjoys the Oracle property for identifying $\mathcal{T}$ and for estimating $\beta(t)$ on $\mathcal{T}^c$. The conditions and the proofs of the theorems are deferred to the Appendix or to the supplementary document.

Recall that the parameters $\boldsymbol{b}_0$ and $\boldsymbol{b}_1$ in (2.3) and (2.4) vary with the sample size n. In the asymptotic studies, we denote them as $\boldsymbol{b}_0(n)$ and $\boldsymbol{b}_1(n)$, respectively. Similarly, $B_1(t)$ is denoted as $B_1(n, t)$, the initial estimator of $\boldsymbol{b}_0(n)$ as $\tilde{\boldsymbol{b}}_0(n)$, and the refined estimator of $\boldsymbol{b}_1(n)$ as $\hat{\boldsymbol{b}}_1(n)$. The tuning parameter λ is denoted as λ_n.

For the estimator in the initial stage, we have an asymptotic result.

Theorem 1. *Let* $\tilde{\boldsymbol{b}}_0(n) = (\tilde{b}_{0,1}(n), \ldots, \tilde{b}_{0,k_{0,n}+h}(n))^T$ *be the Dantzig estimate of* $\boldsymbol{b}_0(n)$*, with* $e_i \sim N(0,1)$ *and* $\mu = 0$*. For the tuning parameter* $\lambda_D(n) = \sqrt{2\log(k_{0,n}+h)}$ *in the Dantzig selector, and under the conditions* A_1*-*A_6 *in the Appendix, we have*

(i) $||\tilde{\boldsymbol{b}}_0(n) - \boldsymbol{b}_0(n)||_{l_2} = O_p\{n^{-1/2}k_{0,n}(\log k_{0,n})^{1/2}\}$;

(ii) $\sup|\tilde{b}_{0,j}(n)| = O_p\{n^{-1/2}k_{0,n}(\log k_{0,n})^{1/2}\}$ *for* $b_{0,j}(n)$ *associated with* $\mathcal{T}$*;*

(iii) *With probability tending to* 1*,*

$$\mathcal{T} \subseteq \hat{\mathcal{T}}^{(0)} \quad \text{and} \quad \hat{\mathcal{T}}^{(0)} \cap \mathcal{T}^c \subseteq \Omega(k_{0,n}),$$

where, with $r \geq 3$ *as in the condition* A_1*,* $\Omega(k_{0,n}) = \{t \in [0,T] : 0 < |\beta(t)| < k_{0,n}^{-r+2}\}$ *is a sub-region of* $[0,T]$*, converging to the empty set as* $n \to \infty$*.*

Theorem 1 shows that the Dantzig selector estimate of $\boldsymbol{b}_0(n)$ is consistent. The L_2 convergence rate follows from Bickel, Ritov, and Tsybakov (2009); see also Raskutti, Wainwright, and Yu (2010); The sup-norm convergence rate can be derived following Lounici (2008); proof of part (iii) is given in the supplementary document.

Let the $n \times (k_{1,n}+h)$ matrix $\boldsymbol{Z}_1(n) = (\boldsymbol{z}_{1,1}, \boldsymbol{z}_{1,2}, \ldots, \boldsymbol{z}_{1,n})^T$, where $\boldsymbol{z}_{1,i}$ is defined in (2.4). Recall that the estimate of the coefficient function is $\hat{\beta}(t) = \boldsymbol{B}_1^T(n,t)\hat{\boldsymbol{b}}_1(n)$, where $\boldsymbol{B}_1(n,t)$ contains the B-spline basis functions at the refinement stage. Further, divide the basis functions $\boldsymbol{B}_1(n,t)$ into $\boldsymbol{B}_{1N}(n,t)$ and $\boldsymbol{B}_{1S}(n,t)$ and the matrix $\boldsymbol{Z}_1(n)$ into $\boldsymbol{Z}_{1N}(n)$ and $\boldsymbol{Z}_{1S}(n)$, according to the membership of their corresponding coefficients $\boldsymbol{b}_{1N}(n)$, related to the null region $\hat{\mathcal{T}}^{(0)}$, and $\boldsymbol{b}_{1S}(n)$.

For the estimator in the refinement stage, we have an asymptotic result.

Theorem 2. *Assume the conditions* A_1*-*A_8 *in the Appendix and an initial estimator with the rate* $||\tilde{\boldsymbol{b}}_1(n) - \boldsymbol{b}_1(n)||_{l_2} = O_p(n^{-1/2}k_{1,n})$*, with* $e_i \sim N(0,1)$ *and* $\mu = 0$*.*

(i) *For* $t \in \mathcal{T}$*, we have* $\hat{\beta}(t) = 0$ *, with probability tending to* 1*.*

(ii) *For* $t \in \mathcal{T}^c$*, we have*

$$\Big(\frac{n}{k_{1,n}}\Big)^{1/2}[\hat{\beta}(t) - \beta(t) - \mathcal{B}_n(t) - \mathcal{W}_n(t)] \xrightarrow{\mathcal{D}} N[0, \sigma^2(t)],$$

where $\mathcal{B}_n(t)$ *denotes the estimation bias,* $\mathcal{W}_n(t) = \beta(t) - \boldsymbol{B}_1^T(n,t)\boldsymbol{b}_1(n)$ *is the B-spline approximation error, and*

$$\sigma^2(t) = \lim_{n\to\infty} \boldsymbol{B}_{1S}^T(n,t)[\Big(\frac{k_{1,n}}{n}\Big)\boldsymbol{Z}_{1S}^T(n)\boldsymbol{Z}_{1S}(n)]^{-1}\boldsymbol{B}_{1S}(n,t),$$

$$\Big(\frac{n}{k_{1,n}}\Big)^{1/2}|\mathcal{B}_n(t)| = O_p(n^{1/2}k_{1,n}^{-r}),$$
$$\Big(\frac{n}{k_{1,n}}\Big)^{1/2}|\mathcal{W}_n(t)| = O_p(n^{1/2}k_{1,n}^{-r-1/2}).$$

(iii) *With* $n^{-1}k_{1,n}^{2r} \to \infty$ *in* A_5*, we have*

$$\Big(\frac{n}{k_{1,n}}\Big)^{1/2}[\hat{\beta}(t) - \beta(t)] \xrightarrow{\mathcal{D}} N[0, \sigma^2(t)].$$

The assumed rate for the initial estimator $\tilde{\boldsymbol{b}}_1(n)$ is verified for the ordinal least squares estimator in the supplementary document.

Properties in Theorems 1 and 2 correspond to the Oracle properties reported in Zou (2006). Precisely, with a large n, we are able to identify the correct sub-regions in which the coefficient functions are non-zero; furthermore, we are able to estimate the values of the coefficient functions on the non-null regions as if we knew their exact locations.

4. Finite Sample Numerical Performances

We conducted simulation studies to evaluate the finite sample performance of our estimators.

Study 1. In this simulation study, we considered two covariate functions and the model

$$Y_i = 2 + \int_0^{10} X_{i1}(t)\beta_1(t)dt + \int_0^{10} X_{i2}(t)\beta_2(t)dt + e_i,$$

for $i = 1, \ldots, n$, with random errors $e_i \sim N(0, 1)$. The covariate functions $X_{i1}(t)$ and $X_{i2}(t)$ were quadratic spline functions on $[0, 10]$ with 50 equally-spaced knots and the corresponding coefficients were generated uniformly from $[-5, 5]$. We took $m = 50$ observations within $[0, 10]$ from the true functions as the observed data and re-constructed $X_{i1}(t)$ and $X_{i2}(t)$ by B-spline approximations. We used coefficient functions $\beta_1(t)$ and $\beta_2(t)$ as follows.

1. $\beta_1(t)$ was a piecewise quadratic function generated from quadratic B-spline functions with evenly-spaced knots at $\{0.0, 0.5, \ldots, 9.5, 10.0\}$, while its coefficients were from a 22×1 vector with the last eight entries $(2.0, 2.3, 2.5, 2.7, 2.7, 2.7, 2.6, 2.6)^T$, and the rest of the entries zero.
2. The non-zero part of $\beta_2(t)$ was a Trigonometric function:

$$\beta_2(t) \;=\; \begin{cases} 3.5\sin\left\{\frac{\pi(t+3)}{5}\right\} & \text{if } t > 7; \\ 0 & \text{if } t \le 7. \end{cases}$$

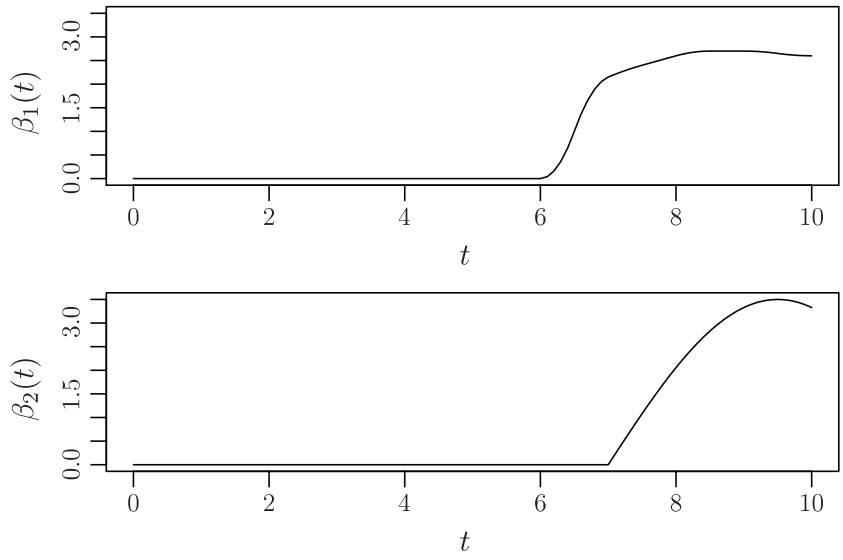

Figure 2. Plots of $\beta_1(t)$ and $\beta_2(t)$ for Study 1.

These coefficient functions are plotted in Figure 2. The shape of the function $\beta_1(t)$ is commonly observed in biological studies; it indicates the pattern of growing into a stable state. As shown in the figure, $\beta_1(t) = 0$ on $[0, 6]$ and $\beta_2(t) = 0$ on $[0, 7]$.

In each study, we generated 250 data sets with sample size $n = 150$. To evaluate the performance of the estimator, the estimated functions $\hat{\beta}_1(t)$ and $\hat{\beta}_2(t)$ were compared to the corresponding true functions. For the comparison, we report two quantities,

$$A_0 = \frac{1}{l_0}\int_{\mathcal{T}} |\hat{\beta}_d(t) - \beta_d(t)|dt \qquad A_1 = \frac{1}{l_1}\int_{\mathcal{T}^c} |\hat{\beta}_d(t) - \beta_d(t)|dt.$$

For $\beta_1(t)$, $\mathcal{T} = [0, 6]$, $\mathcal{T}^c = (6, 10]$, and $l_0 = 6$, $l_1 = 4$. For $\beta_2(t)$, $\mathcal{T} = [0, 7]$, $\mathcal{T}^c = (7, 10]$, and $l_0 = 7$, $l_1 = 3$. The quantity A_0 measures the integrated absolute differences between the estimated coefficient functions and the true functions on the null regions, while A_1 measures it on the non-null regions.

We first evaluate the performance of the Dantzig estimates of $\beta_d(t)$ $(d = 1, 2)$ with different numbers of knots. Ideally, in order to reach high precision in determining the null region with a direct application of the Dantzig selector, one would need a very large number of knots. However, performance also deteriorates when the number of parameters is very large. Here, we can decide the number of variables in our problem, and design the procedure to improve numerical performance. Below, we show the effect of the total number of knots, $k_{0,n} + 1$, on

Table 1. Integrated absolute biases of the least squares and the Dantzig estimates for Study 1. Each entry is the Monte Carlo average of A_j, $j = 0$ or 1; the corresponding standard deviation is reported in parentheses.

		$\beta_1(t)$		$\beta_2(t)$	
Estimator	$k_{0,n}$	A_0	A_1	A_0	A_1
Least Squares	50	2.220 (1.414)	3.272 (2.150)	1.905 (1.253)	3.982 (2.783)
Dantzig Selector	50	0.005 (0.010)	0.695 (0.090)	0.004 (0.007)	0.792 (0.114)
Dantzig Selector	100	0.010 (0.019)	1.416 (0.100)	0.009 (0.014)	1.687 (0.136)
Dantzig Selector	200	0.008 (0.021)	2.318 (0.102)	0.006 (0.013)	2.661 (0.141)
Dantzig Selector	300	0.007 (0.020)	2.724 (0.104)	0.004 (0.011)	3.120 (0.133)

the performance of the Dantzig estimator of $\beta_d(t)$, varying the number of knots among $k_{0,n} = 50$, 100, 200, and 300.

The performances of the estimated coefficient functions by least squares with $k_{0,n} = 50$, and the Dantzig selector with $k_{0,n} = 50$, 100, 200, and 300, are summarized in Table 1. The entries of the table give the Monte Carlo averages of A_0 and A_1 over the 250 generated data sets, while the corresponding standard deviation is reported in parentheses. Since the same $X_{i1}(t)$ and $X_{i2}(t)$ were generated for each value of $k_{0,n}$, the results of the Dantzig estimator in Table 1 differ only by $k_{0,n}$. Table 1 shows that the least square estimator performed poorly on both regions, even with 50 knots. The Dantzig estimator performed well on the null region throughout, which is indicated by the small values of A_0. However, its estimation of $\beta_d(t)$ on the non-null regions was increasingly worse with growth in the number of knots. The poor performance of the Dantzig estimator on the non-null regions in Table 1 indicates the necessity of a refining stage.

We next evaluate the performance of the proposed one-step group SCAD estimator. The method, as described in Sections 2.1 to 2.3 with $D = 2$, refined the initial null region estimates and estimated $\beta_d(t)$ on the non-null regions simultaneously, using the proposed algorithm. Based on the theoretical rates in Section 3, we let $k_{0,n} = c_{k0}n^{0.23}$, $k_{1,n} = c_{k1}n^{0.20}$, and $d_n = c_d n^{-0.25}$. We used a 10-fold cross validation on five data sets to choose the c's, and then fixed them throughout the simulation. With the grid-search algorithm to determine the boundary of null regions, we do not expect outcomes to be sensitive to the choices of c_{k0} and c_d, and a bad choice could simply lead to more computation in locating the boundaries at the second stage. This is indeed the case. The 10-fold CV results are indifferent over the range of c_d we tried, c_d varying from 0.05 to 0.5 (d_n varying between 0.014 and 0.143). The CV values varied near their minimums for $k_{0,n}$ between 35 and 55 for all five data sets. We used $k_{0,n} = 50 \approx 13n^{0.23}$, $k_{1,n} = 12 \approx 4.5n^{0.20}$, and $d_n = 0.2n^{-0.25}$. For comparison,

Table 2. Integrated absolute biases of the least squares, the Dantzig selector, the adaptive LASSO (adpLASSO), and the one-step group SCAD (gSCAD) estimates for Study 1. Each entry is the Monte Carlo average of A_j, $j = 0$ or 1; the corresponding standard deviation is reported in parentheses.

	$\beta_1(t)$		$\beta_2(t)$	
Estimator	A_0	A_1	A_0	A_1
Oracle Estimator	-	0.157 (0.041)	-	0.166 (0.046)
Least Squares	2.205 (1.432)	3.283 (2.549)	1.963 (1.256)	4.088 (2.716)
Dantizig Selector	0.006 (0.013)	0.692 (0.094)	0.006 (0.010)	0.821 (0.132)
adpLASSO AIC	0.041 (0.030)	0.193 (0.059)	0.036 (0.028)	0.214 (0.069)
adpLASSO BIC	0.031 (0.031)	0.212 (0.059)	0.025 (0.029)	0.240 (0.074)
gSCAD AIC	0.024 (0.033)	0.143 (0.038)	0.024 (0.030)	0.155 (0.048)
gSCAD BIC	0.004 (0.013)	0.140 (0.037)	0.003 (0.009)	0.154 (0.049)

we calculated the adaptive LASSO estimates of $\beta_d(t)$ by estimating the B-spline coefficients associated with the same adaptive knots, with the adaptive weights (Zou (2006)) being the inverse of absolute values of initial estimates by least squares. The criterion $C(\mathcal{T}_w, \lambda)$ was specified as GCV, AIC, BIC, or RIC, defined in the supplementary document. The same set of criteria were also used to select the tuning parameter in adaptive LASSO. The performances of estimated coefficient functions with AIC and BIC criteria, as well as the Oracle estimates that used the known null regions, are summarized in Table 2. The results with GCV and RIC, which are comparable to those with AIC and BIC, are reported in the supplementary document. The outcomes show that the criteria work about equally well. We tried different $k_{0,n}$ values but, as expected, they do not play much of a role in our procedure.

It may seem surprising that the Oracle estimator gives slightly larger A_1 values than the proposed estimators. This is due to the fact that the proposed estimators shrink the small non-zero values on the boundary toward zero, and consequently achieve less varied and better estimated outcomes near the null boundary.

The initial null regions of $\beta_1(t)$ and $\beta_2(t)$ identified by the Dantzig selector and the refined null regions of the one-step group SCAD method are shown in Table 3, where the averages of the lower and upper limits of the estimated null regions over the 250 generated data sets, as well as the corresponding standard deviations in parentheses, are summarized. The true null regions are $[0, 6]$ for $\beta_1(t)$ and $[0, 7]$ for $\beta_2(t)$, respectively.

As shown in Tables 2 and 3, the least squares estimates of $\beta_d(t)$ are poor on both null and non-null regions of $\beta_d(t)$. The Dantzig selector tends to have less favorable performance on the non-null regions of $\beta_d(t)$. Moreover, the initial

Table 3. Null region estimates for Study 1. Each entry is the Monte Carlo average of estimated boundary of the null region; the corresponding standard deviation is reported in parentheses.

	$\beta_1(t)$		$\beta_2(t)$	
Estimator	lower	upper	lower	upper
Dantzig Selector	0.008 (0.064)	6.230 (0.175)	0.002 (0.038)	7.123 (0.202)
gSCAD AIC	0.011 (0.091)	5.773 (0.479)	0.004 (0.063)	6.666 (0.528)
gSCAD BIC	0.010 (0.082)	6.058 (0.171)	0.003 (0.051)	6.951 (0.181)

null region identified by it tends to be larger than the true region. The one-step group SCAD method seems to satisfactorily estimate the coefficient functions on both null and non-null regions. Furthermore, by refining the null region initial estimates, our method, particularly using the BIC or RIC criterion, tends to identify the null regions of $\beta_d(t)$ with a greater accuracy than the Dantzig selector. The one-step group SCAD method also outperforms adaptive LASSO on both null and non-null regions.

The full simulation results with GCV, AIC, BIC, and RIC criteria, reported in the supplementary document, show that the performances of the criteria in the method are, in general, satisfactory; see Wang, Li, and Tsai (2007) for another report of using these four criteria in comparison of performances of SCAD methods. The results show that BIC and RIC are less conservative in identifying the null regions. This is expected because the BIC and RIC criteria put heavier penalty on the effective number of parameters than the AIC criterion (Shi and Tsai (2002)).

One advantage of the proposed estimators is that they readily provide inferences for non-zero coefficient functions. Here, we computed the variance of the point estimator given in Theorem 2 and constructed the pointwise 95% confidence interval according to the asymptotic normality results in Theorem 2 for t in the non-null regions. At each point, the true value of coefficient function $\beta_1(t)$ or $\beta_2(t)$ was compared with the computed pointwise 95% confidence interval, and the coverage probabilities (CP) of the confidence intervals were computed over the 250 generated data sets. Besides CP, we also calculated the Monte Carlo biases, standard deviations (SD), and mean square errors (MSE) for points in the non-null regions. For $\beta_1(t)$, we took $t = 6.1, 6.2, \ldots, 10.0$, and for $\beta_2(t)$, $t = 7.1, 7.2, \ldots, 10.0$. Using AIC and BIC, the entries of Table 4 give the averages of Monte Carlo biases, SDs, MSEs, and CP over these points, while the corresponding standard deviation is reported in parentheses. The coverage probabilities of the confidence intervals, computed over the 250 generated data sets at these points in the non-null regions of $\beta_1(t)$ and $\beta_2(t)$, are plotted in Figure 3, for AIC and BIC. Table 4 shows that the estimators based on AIC and BIC

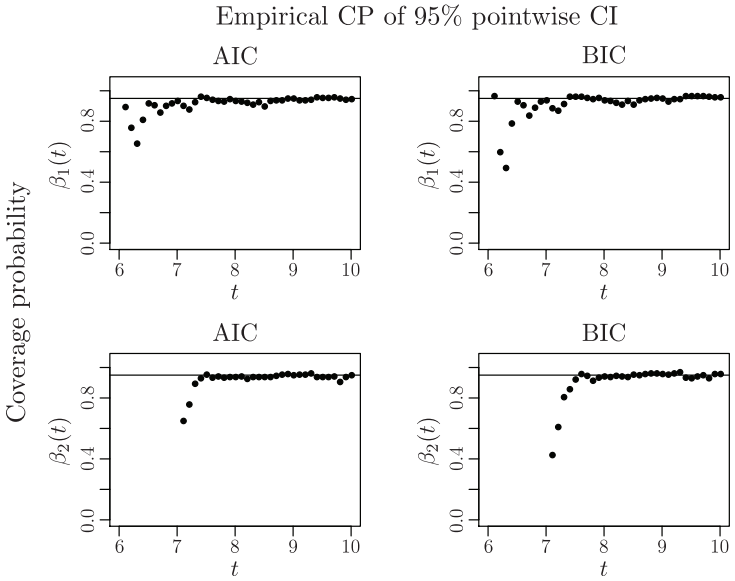

Figure 3. Empirical coverage probabilities (CP) of 95% pointwise confidence intervals for coefficient estimate over non-null region of $\beta_1(t)$ and $\beta_2(t)$ for Study 1, by AIC and BIC. For $\beta_1(t)$, the points are taken at $t = 6.1, 6.2, \ldots, 10.0$; for $\beta_2(t)$, the points are taken at $t = 7.1, 7.2, \ldots, 10.0$.

Table 4. Monte Carlo bias, standard deviation (SD), mean squared error (MSE), and empirical coverage probability (CP) of 95% pointwise confidence intervals of group SCAD (gSCAD) estimates for Study 1. Each entry is the average over the selected points in the non-null region of $\beta_1(t)$ or $\beta_2(t)$; the corresponding standard deviation is reported in parentheses.

	$\beta_1(t)$			
Estimator	Ave. MC Bias	Ave. MC SD	Ave. MC MSE	CP
gSCAD AIC	0.004 (0.013)	0.201 (0.213)	0.085 (0.331)	0.932 (0.047)
gSCAD BIC	-0.001 (0.019)	0.195 (0.218)	0.084 (0.339)	0.928 (0.094)
	$\beta_2(t)$			
Estimator	Ave. MC Bias	Ave. MC SD	Ave. MC MSE	CP
gSCAD AIC	-0.006 (0.031)	0.224 (0.247)	0.110 (0.394)	0.924 (0.053)
gSCAD BIC	-0.012 (0.043)	0.221 (0.242)	0.107 (0.378)	0.915 (0.098)

perform similarly on the non-null regions and all yield coverage probabilities of the pointwise confidence intervals close to the nominal level 0.95. The slightly lower coverage probabilities than the nominal level is due to the shrinkage of the estimates for t close to the boundary of the null regions, as shown in Figure 3. The GCV and RIC criteria have similar performances to AIC and BIC; their results are reported in the supplementary document.

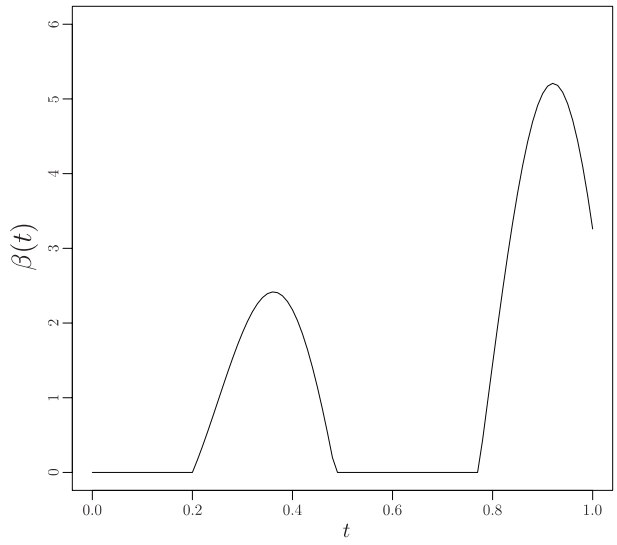

Figure 4. Plot of $\beta(t)$ for Study 2.

We also briefly investigated the effects of having irregularly spaced time points by dropping 10% and 20% of the observations on $X_{i1}(t)$ and $X_{i2}(t)$ at random from the simulated data, followed by a pre-smoothing step. Our procedure was then applied to the pre-smoothed data. The relative performance of least squares, the Dantzig selector, the adaptive LASSO, and the one-step group SCAD, were similar to what is seen in Tables 1-4. We noted a slight decrease in the average coverage probabilities. For example, for group SCAD with the AIC criterion, the coverage probabilities for $\beta_1(t)$ and $\beta_2(t)$ were 90.6% and 88.7% for 10% dropped, and 87.7% and 85.6% for 20% dropped. A further investigation indicated that the decrease in average coverage probabilities mainly occurred near the boundary of null regions. This observation implies that the performance of our estimators near the boundary of null regions tends to be subject to the influence of the locations and numbers of observations on the covariate functions $X_{i1}(t)$ and $X_{i2}(t)$.

Study 2. We conducted a second simulation study with

$$Y_i = 2 + \int_0^1 X_i(t)\beta(t)dt + e_i,$$

and $e_i \sim N(0, 0.25)$. The covariate functions $X_i(t)$ were generated and reconstructed on $[0, 1]$ by the same methods as in Study 1. The coefficient function was $\beta(t) = \max[0, 8\log(t+1)\sin\{3.5\pi(t-0.2)\}]$, shown in Figure 4. Its null region is $\mathcal{T} = [0.0, 0.2] \bigcup [0.486, 0.771]$.

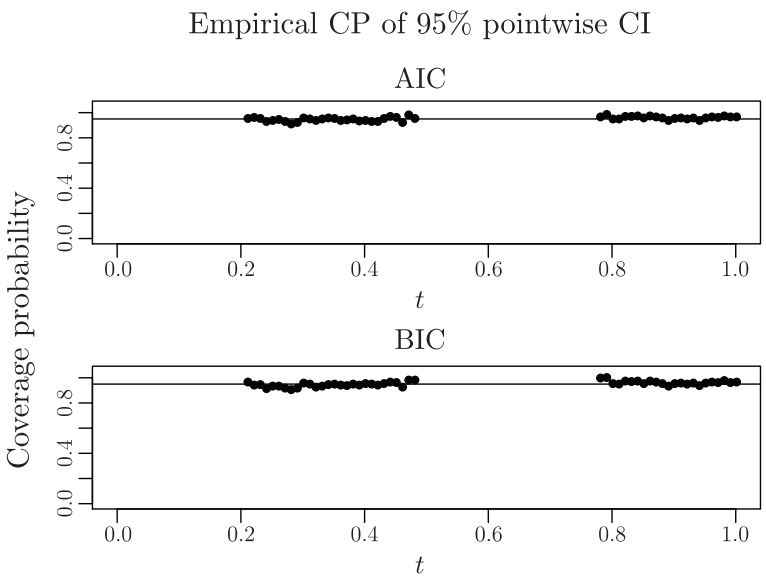

Figure 5. Empirical coverage probabilities (CP) of 95% pointwise confidence intervals for coefficient estimate over non-null region of $\beta(t)$ for Study 2, by AIC and BIC. The points are taken at $t = 0.21, 0.22, \ldots, 0.48, 0.78, 0.79, \ldots, 0.99, 1.00$.

We generated 250 data sets with sample size $n = 500$. The 10-fold CV results suggested the use of a much smaller $k_{0,n}$; we took $k_{0,n} = 20$, $k_{1,n} = 10$, and $d_n = n^{-0.25}$. The equivalent results to those in Study 1 are reported in Tables 5 to 7. The 95% CI overage probabilities (CP) are reported over $t = 0.21, 0.22, \ldots, 0.48, 0.78, 0.79, \ldots, 1.00$ within $\mathcal{T}^c$. Tables 5 and 6 show that the method has better performance overall in estimating $\beta(t)$ and identifying the null region of $\beta(t)$ than the other methods, and behaves similarly as the oracle estimator. The Dantzig estimator tends to yield larger A_1, as usual. We repeated the simulation with $k_{0,n} = 50$ while keeping the rest of the setup unchanged. The Monte Carlo averages of A_0 for least-squares, Dantzig, adaptive LASSO (AIC), and grouped SCAD (AIC) were 3.755, 0.001, 0.198, and 0.043, respectively, while the corresponding values for A_1 were 3.840, 0.849, 0.268, and 0.246. Similar performances were obtained using other criteria. We note that the outcomes for adaptive LASSO and our method change little, mainly due to the slight changes of boundaries and the knot locations at the second stage. The least-squares and Dantzig estimators again suffered from the large number of parameters.

5. The Johns Hopkins Precursors Study

The impact of cumulative lifelong risk exposure on quality of life (QoL) in old age is of great interest to researchers. We applied our method to data from

Table 5. Integrated absolute biases of the least squares, the Dantzig selector, the adaptive LASSO (adpLASSO), and the one-step group SCAD (gSCAD) estimates for Study 2. Each entry is the Monte Carlo average of A_j, $j = 0$ or 1; the corresponding standard deviation is reported in parentheses.

Estimator	A_0	A_1
Oracle Estimator	-	0.257 (0.054)
Least Squares	0.246 (0.060)	0.240 (0.054)
Dantzig Selector	0.006 (0.007)	0.485 (0.069)
adpLASSO AIC	0.066 (0.063)	0.246 (0.063)
adpLASSO BIC	0.023 (0.041)	0.278 (0.079)
gSCAD AIC	0.038 (0.076)	0.230 (0.054)
gSCAD BIC	0.009 (0.020)	0.226 (0.056)

Table 6. Null region estimates for Study 2. Each entry is the Monte Carlo average of estimated boundary of the null region; the corresponding standard deviation is reported in parentheses.

	$[0.000, 0.200]$		$[0.486, 0.771]$	
Estimator	lower	upper	lower	upper
Dantzig Selector	0.001 (0.009)	0.199 (0.016)	0.502 (0.014)	0.749 (0.008)
gSCAD AIC	0.001 (0.009)	0.194 (0.021)	0.507 (0.019)	0.744 (0.016)
gSCAD BIC	0.001 (0.009)	0.199 (0.016)	0.502 (0.014)	0.749 (0.008)

Table 7. Monte Carlo bias, standard deviation (SD), mean squared error (MSE), and empirical coverage probability (CP) of 95% pointwise confidence intervals of group SCAD (gSCAD) estimates for Study 2. Each entry is the average over the selected points in the non-null region of $\beta_1(t)$ or $\beta_2(t)$; the corresponding standard deviation is reported in parentheses.

	$\beta_1(t)$			
Estimator	Ave. MC Bias	Ave. MC SD	Ave. MC MSE	CP
gSCAD AIC	-0.012 (0.055)	0.296 (0.173)	0.120 (0.265)	0.950 (0.016)
gSCAD BIC	-0.020 (0.072)	0.286 (0.183)	0.120 (0.272)	0.951 (0.020)

the Johns Hopkins Precursors Study to investigate the effect of body mass index (BMI) at midlife on the quality of life at older ages. In this, we focused on the outcome of physical functioning (PF), an important QoL measure among the elderly, collected through the SF-36 health survey questionnaires (Ware and Sherbourne (1992)), with a score that ranged from 0 to 100. We restricted our analysis to data from 107 participants who had their PF scores assessed between 70 and 76 years of age. The age range of interest for the BMI was 40 to 70 years. The transformed PF score, $y = 10 * asin\sqrt{PF/100}$ was used as the response in model (2.1).

To obtain each participant's trajectory of BMI on $[40, 70]$, we first pre-

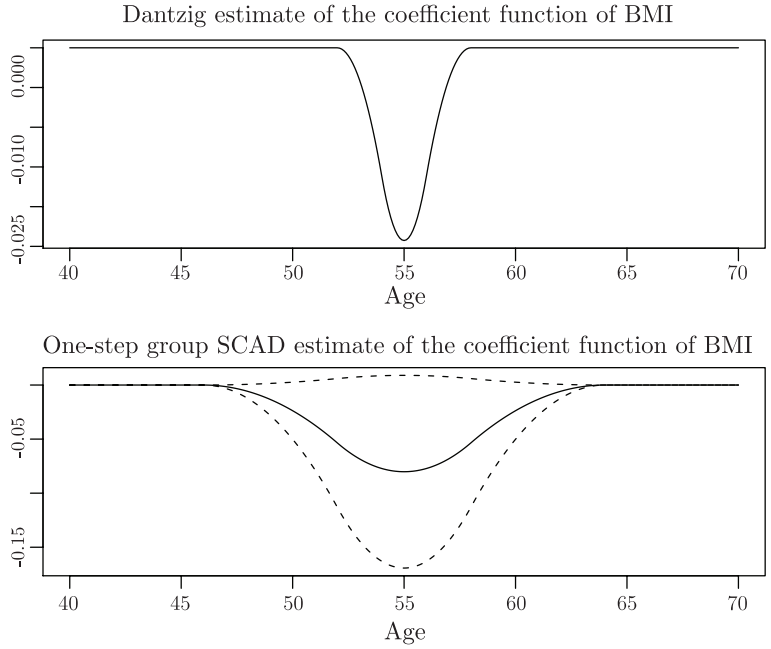

Figure 6. The estimated coefficient function for BMI in the Johns Hopkins Precursors Study. The upper panel shows the initial estimate by Dantzig selector. The lower panel shows the refined estimate by the proposed one-step group SCAD estimator with the dotted lines being the 95% pointwise CI for it in the refined non-null region $[46, 64]$. The AIC and RIC criteria yield the same refined estimates.

smoothed the available BMI records for each subject (Ramsay and Silverman (2005)). The pre-smoothed and centered BMI trajectories on $[38, 72]$ were then fitted by quadratic B-splines with evenly-spaced knots of $\{38, 40, \ldots, 70, 72\}$. In the initial stage of the method, the functional coefficient on $[40, 70]$ was approximated by quadratic splines with evenly-spaced knots of $\{40, 42, \ldots, 68, 70\}$. The Dantzig selector, with the tuning parameter selected by leave-one-out cross validation, yielded the initial null region $[40, 52] \cup [58, 70]$. The initial estimate of $\beta(t)$ by the Dantzig selector is plotted in Figure 6.

During the refining stage, we specified the working null regions as $[40, 52 - l] \cup [58 + l, 70]$ and let $l = 0, \ldots, 10$. Using the method proposed in Section 2.2, the refined null region selected by both AIC and RIC was $[40, 46] \cup [64, 70]$. The two criteria also led to identical refined estimates of $\beta(t)$ on [40, 70], as well as the same pointwise 95% confidence intervals; they are plotted in Figure 6.

Figure 6 shows that greater values of BMI between ages 46 and 64 seem to be associated with greater decrease in the PF scores in early to mid 70 years of age. The 95% pointwise confidence intervals of $\beta(t)$ on $[46, 64]$, albeit not

significant, are almost all in the negative range. In contrast to our approach, the Dantzig selector identified a larger null region and shrunk the estimated coefficient function on the non-null region toward zero. This is consistent with what we observed in the simulation study. The zero coefficient function in the mid-forties and younger implies that a greater BMI at this age range does not contribute much to additional risk of decreasing PF scores, after factoring in body weight patterns in the subsequent two decades. On the other hand, the zero coefficient function after the mid-sixties could be due to a mixture of two forces; high BMI is harmful in general, but being too thin may not be a good sign among the elderly either.

The non-significant finding could be due to the relatively small number of participants in our sample and the modest association between BMI and the PF scores, not uncommon in this kind of study. To better understand whether we had sufficient power to confirm a modest association, we conducted a small power analysis. In the analysis, we specified the true coefficient function as the curve estimated by the proposed method. Then, conditional on the available BMI records, we generated new PF scores according to the fitted model and applied the method to calculate 95% confidence intervals. The proportions of pointwise 95% confidence intervals that were completely negative at ages 50, 55, and 60, with the λ in SCAD penalty being fixed at the value selected in the data analysis, were 0.496, 0.256, and 0.522, respectively. The finding from the Precursors Study data nevertheless suggests the potential for added benefit of slower decline in physical functioning at old age in keeping a healthy body weight in midlife.

6. Concluding Remarks

With the development in variable selection when the number of predictors is large, we advance a method to estimate coefficient functions in functional linear model when the values of the functions are zero in certain sub-regions. Our aim is a functional data analysis (FDA) tool in which the final estimator on the non-null region behaves just like a regular FDA solution. Our estimator successfully attains the properties we desire: it maintains the sparsity and Oracle properties in the estimated coefficient functions, asymptotic normality applies, and it achieves superior numerical performances compared to existing alternatives in both identification of $\mathcal{T}$ and the estimation of $\beta(t)$. The proposed procedure borrows strength from existing efficient algorithms and can be easily carried out.

An additional point is that in functional data analysis, we select the number of basis functions and determine the number of parameters needed. When the number is unavoidably large, as in the variable selection problems in genomic studies, even the best estimator can be poor. When we reduce the number

of parameters, we simplify the nature of the problem and consequently obtain improved results. There are alternatives to what we have proposed. In our method, as indicated in the supplementary document, we shrank the limits of each null interval in a symmetric way with a grid size of $0.02T$. More effective search ideas, such as a combination of shrinking and expanding around the limits, could be conducted and should lead to some improvement. The performance of the estimates on the non-null region can be further improved by an adaptive selection of knots, as in the regular B-spline smoothing estimation. We have not pursued these directions here.

Acknowledgement

Zhou's research was supported by the National Science Foundation (DMS-0906665). N.-Y. Wang's research was supported by grants from the National Institutes of Health (UL1 RR025005 and P60 DK79637). N. Wang's research was supported by a grant from the National Cancer Institute (CA74552). The Johns Hopkins Precursors Study is supported by a grant from the National Institute on Aging (R01 AG01760).

Supplementary Document

SuppDoc.pdf describes the aglorithms, the technical proofs, and provides additional tables and figures for the outcomes of the numerical studies.

Appendix: Assumptions and Technical Details

We provide the assumptions behind the theoretical properties and sketch their proofs in the Appendix; refer to the supplementary document for further details.

A.1. Notation and assumptions

Recall that $k_{0,n}+1$, $k_{1,n}+1$ are the numbers of knots, and $\boldsymbol{Z}_0(n)$ and $\boldsymbol{Z}_1(n)$ are the design matrices in models (2.3) and (2.4) for the two stages, respectively.

Let $\boldsymbol{Z}_0^*(n)$ be a standardized version of $\boldsymbol{Z}_0(n)$ such that $\Psi_{z0*}=n^{-1}\boldsymbol{Z}_0^{*T}(n)\boldsymbol{Z}_0^*(n)$ has its diagonal elements, $\Psi_{z0*}(j,j)\equiv 1$, for all j. Take $\delta_{b_0}(n)=\tilde{\boldsymbol{b}}_0(n)-\boldsymbol{b}_0(n)$, and let $\delta_{b_0,\mathcal{T}}(n)$ and $\delta_{b_0,\mathcal{T}^c}(n)$ be $\delta_{b_0}(n)$, corresponding to null and signal regions, respectively, with $s(n)$ the number of non-zero coefficients in $\boldsymbol{b}_0(n)$. Recall the definitions of $\boldsymbol{Z}_{1N}(n)$ and $\boldsymbol{Z}_{1S}(n)$ in Section 3. To show the Oracle properties of the proposed estimator, consider the following conditions.

A_1: $\beta(t)$ has rth ($r\geq 3$) bounded derivative on $[0,T]$.

A_2: $\int_0^T |X_i(t)|dt \leq M' k_{1,n}^{r-1}$, for $i=1,\ldots,n$ and some constant $M'<\infty$.

A_3: For some integer s less than $k_{0,n}+h$ and non-zero $\delta_{b_o}(n)$,

$$\min_{|s(n)|\le s}\ \min_{|\delta_{b_0,\mathcal{T}}(n)|_{l_1}\le|\delta_{b_0,\mathcal{T}^c}(n)|_{l_1}} \frac{|\boldsymbol{Z}_0^{*T}(n)\delta_{b_0}(n)|_{l_2}}{|\sqrt{n}|\delta_{b_0,\mathcal{T}^c}(n)|_{l_2}} > 0.$$

A_4: For a constant $\gamma > 1$ and $s \ge s(n)$,

$$\max_{i\neq j}\Psi_{z0*}(i,j) \le \frac{1}{3\gamma s}.$$

A_5: For $k_{0,n}$, $n^{-1}k_{0,n}^{2r-2}\log k_{0,n}\to 0$ and $n^{-1}k_{0,n}^{2r}\log k_{0,n}\to\infty$.
For $k_{1,n}$, $n^{-1}k_{0,n}^{2r-4}k_{1,n}^{3}\to\infty$.

A_6: For the threshold value d_n, $d_n n^{1/2}k_{0,n}^{-1}(\log k_{0,n})^{-1/2}\to\infty$ and $d_n k_{0,n}^{r-2}\to 0$.

A_7: For the tuning parameter λ_n, $\lambda_n\to 0$ and $\lambda_n k_{0,n}^{r-2}\to\infty$.

A_8: There are constants $0<c_1'<c_2'$ such that

$$c_1' k_{1,n}^{-1} \le \lambda_{min}\{n^{-1}\boldsymbol{Z}_1^T(n)\boldsymbol{Z}_1(n)\} \le \lambda_{max}\{n^{-1}\boldsymbol{Z}_1^T(n)\boldsymbol{Z}_1(n)\} \le c_2' k_{1,n}^{-1},$$

where $\lambda_{min}(A)$ and $\lambda_{max}(A)$ denote the smallest and largest eigenvalues of the matrix A. The same eigenvalue condition as for $n^{-1}\boldsymbol{Z}_1^T(n)\boldsymbol{Z}_1(n)$ holds for the matrices $n^{-1}\boldsymbol{Z}_{1N}^T(n)\boldsymbol{Z}_{1N}(n)$ and $n^{-1}\boldsymbol{Z}_{1S}^T(n)\boldsymbol{Z}_{1S}(n)$. In addition,

$$\lambda_{max}\{n^{-1}\boldsymbol{Z}_{1N}^T(n)\boldsymbol{Z}_{1S}(n)\boldsymbol{Z}_{1S}^T(n)\boldsymbol{Z}_{1N}(n)\} < c_3 k_{1,n}^{-1},$$

for a constant $c_3 > 0$.

Condition A_2 is weaker than one assumed in James, Wang, and Zhu (2009). The restricted eigenvalue condition A_3 from Bickel, Ritov, and Tsybakov (2009) controls the singularity of the first stage design matrix to ensure the L_2 rate, while A_4 is required to warrant the sup-norm rate in Theorem 1. The rate of the threshold value d_n given in condition A_6 guarantees that, with probability tending to 1 as $n\to\infty$, $\mathcal{T}$ is correctly identified by $\hat{\mathcal{T}}^{(0)}$. Condition A_8 is analogous to one in Fan and Peng (2004) when the number of predictors increases with n. It appears as lemmas in Zhou, Shen, and Wolfe (1998) and Zhu, Fung, and He (2008); to avoid redundancy, we simply use it as a condition.

A.2. Sketch Proof of Theorem 2

We use $a_n > O_p(b_n)$ and $a_n \ge O_p(b_n)$ to denote that, as $n\to\infty$ with probability tending to 1, $b_n/a_n\to 0$ and b_n/a_n is bounded from above, respectively. Here we sketch the key steps in the proof of Theorem 2. Recall that $\boldsymbol{b}_{1N}(n)$ and

$\boldsymbol{b}_{1S}(n)$ are the division of $\boldsymbol{b}_1(n)$ according to $\hat{\mathcal{T}}^{(0)}$. Since $\boldsymbol{b}_{1N}(n)$ contains the coefficients associated with $\hat{\mathcal{T}}^{(0)}$, by Theorem 1 (iii), these coefficients are either associated with $\mathcal{T}$ or with $\Omega(k_{0,n})$. Consequently, by A_5, $||\boldsymbol{b}_{1N}(n)||_{l_1} = O_p(k_{0,n}^{-r+2})$. Following the proof of Part (iii) of Theorem 1 in the supplementary document, it is easy to see that $||\boldsymbol{b}_{1S}(n)||_{l_1} \geq O_p(1)$.

We assume the initial value $\tilde{\boldsymbol{b}}_1(n)$ satisfies $||\tilde{\boldsymbol{b}}_1(n) - \boldsymbol{b}_1(n)||_{l_2} = O_p(n^{-1/2}k_{1,n})$. Note that $\tilde{\boldsymbol{b}}_{1N}(n)$ and $\tilde{\boldsymbol{b}}_{1S}(n)$ are the division of $\tilde{\boldsymbol{b}}_1(n)$ according to $\hat{\mathcal{T}}^{(0)}$. Given $||\tilde{\boldsymbol{b}}_{1N}(n) - \boldsymbol{b}_{1N}(n)||_{l_1} \leq C||\tilde{\boldsymbol{b}}_{1N}(n) - \boldsymbol{b}_{1N}(n)||_{l_2} = O_p(n^{-1/2}k_{1,n})$, $||\boldsymbol{b}_{1N}(n)||_{l_1} = O_p(k_{0,n}^{-r+2})$, and A_5, we have that $||\tilde{\boldsymbol{b}}_{1N}(n)||_{l_1} = O_p(k_{0,n}^{-r+2})$ and $||\tilde{\boldsymbol{b}}_{1S}(n)||_{l_2} \geq O_p(1)$. Given A_7, with probability tending to 1, we have that $p'_{\lambda_n}(||\tilde{\boldsymbol{b}}_{1N}(n)||_{l_1}) = \lambda_n$ and $p'_{\lambda_n}(||\tilde{\boldsymbol{b}}_{1S}(n)||_{l_1}) = 0$.

Let Q_n be $Q_n(\hat{\mathcal{T}}, \lambda, \boldsymbol{b})$ in (2.5), and focus on the expansion of $Q_n\{\hat{\boldsymbol{b}}_1(n)\} - Q_n\{\boldsymbol{b}_1(n)\}$ as

$$
\begin{aligned}
&[\hat{\boldsymbol{b}}_1(n) - \boldsymbol{b}_1(n)]^T \boldsymbol{Z}_1^T(n)\boldsymbol{Z}_1(n)[\hat{\boldsymbol{b}}_1(n) - \boldsymbol{b}_1(n)] - 2(\boldsymbol{Z}_1^T(n)\boldsymbol{\epsilon}_1(n))^T[\hat{\boldsymbol{b}}_1(n) - \boldsymbol{b}_1(n)] \\
&\quad + n\lambda_n(||\hat{\boldsymbol{b}}_{1N}(n)||_{l_1} - ||\boldsymbol{b}_{1N}(n)||_{l_1}) \\
&\geq [\hat{\boldsymbol{b}}_1(n) - \boldsymbol{b}_1(n)]^T \boldsymbol{Z}_1^T(n)\boldsymbol{Z}_1(n)[\hat{\boldsymbol{b}}_1(n) - \boldsymbol{b}_1(n)] - 2(\boldsymbol{Z}_1^T(n)\boldsymbol{\epsilon}_1(n))^T[\hat{\boldsymbol{b}}_1(n) - \boldsymbol{b}_1(n)] \\
&\quad + n\lambda_n(||\hat{\boldsymbol{b}}_{1N}(n) - \boldsymbol{b}_{1N}(n)||_{l_1} - 2||\boldsymbol{b}_{1N}(n)||_{l_1}),
\end{aligned}
$$

where $\hat{\boldsymbol{b}}_{1N}(n), \hat{\boldsymbol{b}}_{1S}(n)$ and $\boldsymbol{b}_{1N}(n), \boldsymbol{b}_{1S}(n)$ are the divisions of $\hat{\boldsymbol{b}}_1(n)$ and $\boldsymbol{b}_1(n)$, respectively, according to their association with $\hat{\mathcal{T}}^{(0)}$. Handling the null and non-null coefficients separately, we show, with some detailed derivation, that a non-optimal bound for the convergence rate of $\hat{\boldsymbol{b}}_1(n)$ holds as $||\hat{\boldsymbol{b}}_1(n) - \boldsymbol{b}_1(n)||_{l_2} \leq O_p(n^{-1/2}k_{1,n}^{3/2})$; this rate is sufficient for us to use in the proof of Theorem 2. The proof of $\hat{b}_{1,j}(n) = 0$, with probability tending to 1, for any $\hat{b}_{1,j}(n)$ associated with $\hat{\mathcal{T}}^{(0)}$, is a direct consequence of this convergence. Part (i) is proved.

We have $p'_{\lambda_n}(||\tilde{\boldsymbol{b}}_{1N}(n)||_{l_1}) = \lambda_n$ and $p'_{\lambda_n}(||\tilde{\boldsymbol{b}}_{1S}(n)||_{l_1}) = 0$ with probability tending to 1. We now give the key steps that lead to the asymptotic distribution of $\hat{\beta}(t)$ for $t \in \mathcal{T}^c$. For large n, we have

$$
\begin{aligned}
&\Big(\frac{n}{k_{1,n}}\Big)^{1/2}(\hat{\beta}(t) - \beta(t)) \\
&\quad = \Big(\frac{n}{k_{1,n}}\Big)^{1/2}\boldsymbol{B}_{1S}^T(n,t)\{\hat{\boldsymbol{b}}_{1S}(n) - \boldsymbol{b}_{1S}(n)\} + \Big(\frac{n}{k_{1,n}}\Big)^{1/2}\{\boldsymbol{B}_1^T(n,t)\boldsymbol{b}_1(n) - \beta(t)\} \\
&\quad = \boldsymbol{B}_{1S}^T(n,t)\{\Big(\frac{k_{1,n}}{n}\Big)\boldsymbol{Z}_{1S}^T(n)\boldsymbol{Z}_{1S}(n)\}^{-1}\{(\frac{n}{k_{1,n}})^{-1/2}\boldsymbol{Z}_{1S}^T(n)\boldsymbol{e}(n)\} \\
&\qquad + \boldsymbol{B}_{1S}^T(n,t)\{\Big(\frac{k_{1,n}}{n}\Big)\boldsymbol{Z}_{1S}^T(n)\boldsymbol{Z}_{1S}(n)\}^{-1}[(\frac{n}{k_{1,n}})^{-1/2}\boldsymbol{Z}_{1S}^T(n)\{\boldsymbol{\epsilon}_1(n) - \boldsymbol{e}(n)\}] \\
&\qquad + \boldsymbol{B}_{1S}^T(n,t)\{\Big(\frac{k_{1,n}}{n}\Big)\boldsymbol{Z}_{1S}^T(n)\boldsymbol{Z}_{1S}(n)\}^{-1}\{(\frac{n}{k_{1,n}})^{-1/2}\boldsymbol{Z}_{1S}^T(n)\boldsymbol{Z}_{1N}(n)\boldsymbol{b}_{1N}(n)\}
\end{aligned}
$$

$$+\Big(\frac{n}{k_{1,n}}\Big)^{1/2}\{\boldsymbol{B}_1^T(n,t)\boldsymbol{b}_1(n)-\beta(t)\}$$
$$=U_n(t)+\Big(\frac{n}{k_{1,n}}\Big)^{1/2}\mathcal{B}'_n(t)+\Big(\frac{n}{k_{1,n}}\Big)^{1/2}\mathcal{B}''_n(t)+\Big(\frac{n}{k_{1,n}}\Big)^{1/2}\mathcal{W}_n(t).$$

By Huang (1998), $U_n(t)$ is the asymptotic normal component, $\mathcal{B}_n(t)=\mathcal{B}'_n(t)+\mathcal{B}''_n(t)$ is the estimation bias, and $W_n(t)$ contains the spline approximation error.

Given that $\boldsymbol{e}(n)\sim N(0,I_n)$, we have that, for $t\in\mathcal{T}^c$,

$$U_n(t)\xrightarrow{\mathcal{D}}N[0,\sigma^2(t)],$$

where $\sigma^2(t)=\lim_{n\to\infty}\boldsymbol{B}_{1S}^T(n,t)\{(k_{1,n}/n)\boldsymbol{Z}_{1S}^T(n)\boldsymbol{Z}_{1S}(n)\}^{-1}\boldsymbol{B}_{1S}(n,t)$.

By A_8, $\lambda_{max}((k_{1,n}/n)\boldsymbol{Z}_{1S}(n)\boldsymbol{Z}_{1S}^T(n))\le c_2'$. Thus, we have $\sup|\epsilon_{1,i}-e_i|\le M'Ck_{1,n}^{-r}$ for some constant C, and $||(n/k_{1,n})^{-1/2}\boldsymbol{Z}_{1S}^T(n)(\boldsymbol{\epsilon}_1(n)-\boldsymbol{e}(n))||_{l_2}\le C'n^{1/2}k_{1,n}^{-r}$ for some constant C'. Then

$$\Big(\frac{n}{k_{1,n}}\Big)^{1/2}|\mathcal{B}'_n(t)|=O_p(n^{1/2}k_{1,n}^{-r}).$$

Also by A_8, we have

$$\Big(\frac{n}{k_{1,n}}\Big)^{-1}\boldsymbol{b}_{1N}^T(n)\boldsymbol{Z}_{1N}^T(n)\boldsymbol{Z}_{1S}(n)\boldsymbol{Z}_{1S}^T(n)\boldsymbol{Z}_{1N}(n)\boldsymbol{b}_{1N}(n)\le c_2'^2||\boldsymbol{b}_{1N}(n)||_{l_2}^2.$$

By A_5, each coefficient in $\boldsymbol{b}_{1N}(n)$ is bounded by $C'k_{0,n}^{-r+2}$ for some constant C'. Combined with the fact that $k_{0,n}^{-r+2}=o_p(1)$ by A_7, we can show that

$$\Big(\frac{n}{k_{1,n}}\Big)^{1/2}|\mathcal{B}''_n(t)|=o_p(1).$$

Therefore, we have

$$\Big(\frac{n}{k_{1,n}}\Big)^{1/2}|\mathcal{B}_n(t)|=O_p(n^{1/2}k_{1,n}^{-r}).$$

The term $\mathcal{W}_n(t)$ is the B-spline approximation error at $\beta(t)$. Given A_1 and the B-spline approximation property, we have

$$\Big(\frac{n}{k_{1,n}}\Big)^{1/2}|\mathcal{W}_n(t)|=O_p(n^{1/2}k_{1,n}^{-r-1/2}).$$

Therefore we have, for $t\in\mathcal{T}^c$,

$$\Big(\frac{n}{k_{1,n}}\Big)^{1/2}[\hat{\beta}(t)-\beta(t)-\mathcal{B}_n(t)-\mathcal{W}_n(t)]\xrightarrow{\mathcal{D}}N[0,\sigma^2(t)].$$

Part (ii) is proved.

Assuming the additional stronger condition $n^{-1}k_{1,n}^{2r} \to \infty$ in A_5, it follows that $(n/k_{1,n})^{1/2}|\mathcal{B}_n(t)| = o_p(1)$ and $(n/k_{1,n})^{1/2}|\mathcal{W}_n(t)| = o_p(1)$. Therefore we have, for $t \in \mathcal{T}^c$,

$$\Big(\frac{n}{k_{1,n}}\Big)^{1/2}[\hat{\beta}(t) - \beta(t)] \xrightarrow{\mathcal{D}} N[0, \sigma^2(t)].$$

Part (iii) is proved.

References

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

Cai, T. and Hall, P. (2006). Prediction in functional linear regression. *Ann. Statist.* **34**, 2159-2179.

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

Cardot, H., Ferraty, F. and Sarda, P. (2003). Spline estimators for the functional linear model. *Statist. Sinica* **13**, 571-591.

Crambes, C., Kneip, A. and Sarda, P. (2009). Smoothing spline estimators for functional linear regression. *Ann. Statist.* **37**, 35-72.

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

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

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32**, 928-961.

Fan, J. and Zhang, J. (2000). Two-step estimation of functional linear models with application to longitudinal data. *J. Roy. Statist. Soc. Ser. B* **62**, 303-322.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice.* Springer-Verlag, New York.

Hall, P. and Van Keilegom I. (2008). Two-sample tests in functional data analysis starting from discrete data. *Statist. Sinica* **17**, 1511-1531.

Huang, J. Z. (1998). Projection estimation in multiple regression with application to functional anova models. *Ann. Statist.* **26**, 242-272.

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 that's interpretable. *Ann. Statist.* **37**, 2083-2108.

Lounici, K. (2008). Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic J. Statist.* **2**, 90-102.

Müller, H.-G. and Stadtmüller, U. (2005). Generalized functional linear models. *Ann. Statist.* **22**, 774-805.

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

Raskutti, G., Wainwright, M. J. and Yu, B. (2010). Restricted eigenvalue conditions for correlated Gaussian designs. *J. Machine Learning Research* **11**, 2241-2259.

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

Shi, P. and Tsai, C. L. (2002). Regression model selection–a residual likelihood approach. *J. Roy. Statist. Soc. Ser. B* **64**, 237-252.

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

Wang, H., Li, R. and Tsai, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94**, 553-568.

Wang, L., Chen, G. and Li, H. (2007). Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics* **23**, 1486-1494.

Ware, J. E. and Sherbourne, C. D. (1992). The MOS 36-item short-form health survey (SF-36): I. conceptual framework and item selection. *Med. Care* **30**, 473-483.

Yao, F., Müller, H. G. and Wang, J. L. (2005). Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc.* **100**, 577-590.

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, S., Shen, X. and Wolfe, D. A. (1998). Local asymptotics for regression splines and confidence regions. *Ann. Statist.* **26**, 1760-1782.

Zhu, Z., Fung, W. K. and He, X. (2008). On the asymptotics of marginal regression splines with longitudinal data. *Biometrika* **94**, 907-917.

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

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *Ann. Statist.* **36**, 1509-1533.

Department of Statistics, University of Virginia, Charlottesville, VA 22904, U.S.A.

E-mail: jz9p@virginia.edu

Department of Medicine, Johns Hopkins University, Baltimore, MD 21287, U.S.A.

E-mail: naeyuh@jhmi.edu

Department of Statistics, University of Michigan, Ann Arbor, MI 48109, U.S.A.

E-mail: nwangaa@umich.edu

(Received October 2010; accepted March 2012)