Estimation of Functional Sparsity in Nonparametric Varying Coefficient Models for Longitudinal Data Analysis

Catherine Y. Tu[1], Juhyun Park[2], and Haonan Wang[1]

[1]Colorado State University, [2]Lancaster University

Abstract

We study the simultaneous domain selection problem for varying coefficient models as a functional regression model for longitudinal data with many covariates. The domain selection problem in functional regression mostly appears under the functional linear regression with scalar response, but there is no direct correspondence to functional response models with many covariates. We reformulate the problem as nonparametric function estimation under the notion of *functional sparsity*. Sparsity is the recurrent theme that encapsulates interpretability in the face of regression with multiple inputs, and the problem of sparse estimation is well understood in the parametric setting as variable selection. For nonparametric models, interpretability not only concerns the number of covariates involved but also the *functional form* of the estimates, and so the sparsity consideration is much more complex. To distinguish the types of sparsity in nonparametric models, we call the former *global sparsity* and the latter *local sparsity*, which constitute functional sparsity. Most existing methods focus on directly extending the framework of parametric sparsity for linear models to nonparametric function estimation to address one or the other, but not both. We develop a penalized estimation procedure that simultaneously addresses both types of sparsity in a unified framework. We establish asymptotic properties of estimation consistency and sparsistency of the proposed method. Our method is illustrated in simulation study and real data analysis, and is shown to outperform the existing methods in identifying both local sparsity and global sparsity.

Keywords: Functional sparsity, Group bridge, Longitudinal data, Model selection, Nonparametric regression

1 Introduction

We study the simultaneous domain selection problem for varying coefficient models as a functional regression model for longitudinal data where the response variable changes with time, recorded for

multiple subjects with multiple predictors. The varying coefficient models [5, 6] are defined as

$$y(t) = \boldsymbol{x}^T(t)\boldsymbol{\beta}(t) + \epsilon(t), \tag{1.1}$$

where $y(t)$ is the response at time t, $\boldsymbol{x}(t) = (x_1(t), \ldots, x_p(t))^T$ is the vector of predictors at time t, $\epsilon(t)$ is an error process independent of $\boldsymbol{x}(t)$ and $\boldsymbol{\beta}(t) = (\beta_1(t), \ldots, \beta_p(t))^T$ is a vector of time varying regression coefficient functions. This model assumes a linear relationship between the response and predictors at each observation time point but allows the coefficients to vary over time, thus greatly enhances the utility of the standard linear model formulation. For generality, we write as if the predictors are also functional, but note that the varying coefficient models are equally applicable when the predictors are scalar valued, and our methodology developed here can be directly applied.

The domain selection problem in functional regression is known to be intrinsically difficult [15]. So far the problem is mostly studied in functional linear regression with a scalar response and single functional covariate. Hall and Hooker [4] formulated the problem as a truncated regression model with single unknown domain, studying the identifiability issues and nonparametric function estimation problem. James et al. [11] on the other hand approached the problem from the viewpoint of sparsity estimation as *interpretable* solutions. Using grid approximation, they imposed parametric sparsity constraints on the derivatives of the underlying function at a large number of grid points, which produces an estimate that distinguishes zero and non-zero regions. As Zhou et al. [30] noted, due to overlapping contribution of each coefficient to neighboring regions, independent shrinkage of the coefficients does not necessarily induce zero values in the coefficient function in general, and thus the procedure tends to over-penalize. As a remedy, Zhou et al. [30] further suggested a two-step estimation procedure. Wang and Kai [22] studied a similar problem under standard nonparametric regression, suggesting the need of distinguishing *functional* features from parametric variable selection.

We consider the regression problem under *functional response* variable with varying coefficient models involving multiple domain selection under the general setting where the true number of covariates is also unknown. Although the views and approaches taken in the earlier development are quite different, the problem of domain selection could be motivated as a means to enhance interpretability in the face of model selection in nonparametric models. In this regard, we share the view that some form of sparsity consideration could be useful. For nonparametric models, however, interpretability not only concerns the number of covariates involved [16, 24, 25, 27] but also the *functional form* of the estimates [11, 30]. To distinguish the types of sparsity in nonparametric models, we call the former *global sparsity* and the latter *local sparsity*, which constitute *functional sparsity* [20, 22]. More formally,

a function has *global sparsity* if it is zero over the entire domain, and it indicates that the corresponding covariate is irrelevant to the response variable. A function has *local sparsity* if it is nonzero but remains zero for a set of intervals, and it identifies an inactive period of the corresponding covariate. These notions as interpretability were informally used in a rather separate context of the analysis, and thus the significance of local sparsity estimation was not well recognized.

We reformulate the domain selection problem as a nonparametric function estimation under the unified theme of *functional sparsity* and propose a one-step penalized estimation procedure that automatically determines the types of functional sparsity, being local or global. Although we distinguish the two types of sparsity in the conceptual level, our unified formulation does not require distinction between them in implementation. We directly exploit the fact that global sparsity is a special case of local sparsity in view of domain selection, but not the other way around. Furthermore, consistency of coefficient function estimation does not necessarily give information on local sparsity. This feature distinguishes our approach from the majority methods targeting global sparsity such as [25]. It is worth noting that there is a fundamental difference in the underlying assumption on sparsity between parametric and nonparametric models, as our focus is on *dense* function estimation with *dependent* variables with unknown domain. This difference was also recognized by Kneip et al. [12]. Moreover, for parametric sparsity, an underlying sparse vector is specified whereas for functional sparsity its *true* sparse representation may not be well defined in their respective function approximation. These differences are not only philosophical, but pose different conceptual challenges in the development. Our proposed penalized procedure resembles a type of parametric sparsity estimation, however, our analysis is not comparable to those with high dimensional parametric sparsity estimation point of view (e.g., [11]).

We provide a theoretical analysis of the proposed method and, in particular, show that the local sparsity can be consistently recovered, even diluted with the problem of global sparsity estimation. We study the properties of our proposed method under standard assumptions on nonparametric smooth function estimation and exploit the functional property in more natural manner, thus contribute to bridging the gap between parametric variable selection and nonparametric functional sparsity in a coherent manner.

Our formulation is given in Section 2. Our approach is a one-step procedure, and allows us to directly control functional sparsity through the coefficient functions themselves, rather than pointwise evaluation. In Section 3, we study large sample properties of the proposed method and establish consistency and sparsistency of function estimates. Section 4 describes simulation studies under different

scenarios, and a real data analysis is given in Section 5, demonstrating the utility of functional sparsity in relation to interpretability of the results. Technical assumptions and proofs are provided in the online supplement.

2 Methodology

Suppose that, for n randomly selected subjects, observations of the kth subject are obtained at $\{t_{kl}, l = 1, \ldots, n_k\}$, and the measurements satisfy the varying coefficient linear model relationship in (1.1)

$$y_k(t_{kl}) = \boldsymbol{x}_k^T(t_{kl})\boldsymbol{\beta}(t_{kl}) + \epsilon_k(t_{kl}), \tag{2.1}$$

where $\boldsymbol{x}_k(t_{kl}) = (x_1(t_{kl}), \ldots, x_p(t_{kl}))^T$ and $y_k(t_{kl})$ is the response of the kth subject at t_{kl}. We assume that $\beta_i(t), i = 1, \ldots, p$ are smooth coefficient functions with bounded second derivatives for $t \in \mathcal{T}$. We use spline approximations to represent $\boldsymbol{\beta}(t)$ and formulate a constrained optimization problem for parameter estimation.

2.1 Least squares estimation under B-spline approximation

B-spline approximation has been widely used for estimating smooth nonparametric functions. For detailed discussion about B-splines, see de Boor [2] and Schumaker [17]. Specifically, for a smooth function $\beta(t)$, $t \in [0, 1]$, its approximant can be written as

$$\widetilde{\beta}(t) = \sum_{j=1}^{J} \alpha_j B_j(t), \tag{2.2}$$

where $\{B_j(\cdot), j = 1, \ldots, J\}$ is a group of B-spline basis functions of degree $d \geq 1$ and knots $0 = \eta_0 < \eta_1 < \ldots < \eta_K < \eta_{K+1} = 1$. Notice that K is the number of interior knots and $J = K + d + 1$. Here we adopt the definition of B-spline as stated in Definition 4.12 of Schumaker [17]. In general, performance of B-spline approximation has been well studied. For instance, under some mild conditions, there exists a function $\widetilde{\beta}(t)$ of the form (2.2) such that the approximation error goes to zero. See Theorem 6.27 of Schumaker [17] for more details.

We write the B-spline approximation for each smooth nonparametric coefficient function as

$$\widetilde{\beta}_i(t) = \sum_{j=1}^{J_i} \alpha_{ij} B_{ij}(t) = \boldsymbol{B}_i(t)^T \boldsymbol{\alpha}_i, \quad t \in [0, 1], \quad i = 1, \ldots, p, \tag{2.3}$$

where $\boldsymbol{B}_i(t) = (B_{i1}(t), \ldots, B_{iJ_i}(t))^T$, $\boldsymbol{\alpha}_i = (\alpha_{i1}, \ldots, \alpha_{ij_i})^T$ and $J_i = K_i + d + 1$. Here K_i is the number of interior knots for $\widetilde{\beta}_i(t)$ which may vary over i. For simplicity, we assume that the knots are evenly distributed over $[0, 1]$. Define a block diagonal matrix $\mathcal{B}(t)$ as

$$\mathcal{B}(t) = \text{diag}\{\boldsymbol{B}_1^T(t), \ldots, \boldsymbol{B}_p^T(t)\}.$$

Using (2.3) in the varying coefficient model (2.1) leads to

$$y_k(t_{kl}) \approx \boldsymbol{x}_k^T(t_{kl})\mathcal{B}(t_{kl})\boldsymbol{\alpha} + \epsilon_k(t_{kl}) = \mathcal{U}_k(t_{kl})\boldsymbol{\alpha} + \epsilon_k(t_{kl})$$

where $\mathcal{U}_k(t_{kl}) = \boldsymbol{x}_k^T(t_{kl})\mathcal{B}(t_{kl})$ and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_p^T)^T$. The least squares criterion of $\boldsymbol{\alpha}$ [9] is defined as

$$\ell(\boldsymbol{\alpha}) = \sum_{k=1}^{n} \omega_k \|\boldsymbol{y}_k - \boldsymbol{U}_k\boldsymbol{\alpha}\|_2^2$$

where $\boldsymbol{y}_k = (y_k(t_{k1}), \ldots, y_k(t_{kn_k}))^T$ and $\boldsymbol{U}_k = (\mathcal{U}_k^T(t_{k1}), \ldots, \mathcal{U}_k^T(t_{kn_k}))^T$. Weights ω_k, $k = 1, \ldots, n$, are usually chosen as $\omega_k \equiv 1$ or $\omega_k \equiv 1/n_k$ [10]. In this paper, for simplicity, we set equal weights to every subject, i.e., $\omega_k \equiv 1$. Putting $\boldsymbol{U} = (\boldsymbol{U}_1^T, \ldots, \boldsymbol{U}_n^T)^T$ and $\boldsymbol{y} = (\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_n^T)^T$, the least squares criterion $l(\boldsymbol{\alpha})$ can be written in matrix form; that is, $l(\boldsymbol{\alpha}) = \|\boldsymbol{y} - \boldsymbol{U}\boldsymbol{\alpha}\|_2^2$. Huang et al. [10] proved that, under certain assumptions, the matrix $\boldsymbol{U}^T\boldsymbol{U}$ is invertible for fixed p. Consequently, $l(\boldsymbol{\alpha})$ has a unique minimizer

$$\widehat{\boldsymbol{\alpha}}_{LSE} = (\boldsymbol{U}^T\boldsymbol{U})^{-1}\boldsymbol{U}^T\boldsymbol{y},$$

which is the least squares estimator of $\boldsymbol{\alpha}$, and thus, the least squares estimators of coefficient functions are

$$\widehat{\beta}_i^{LSE}(t) = \sum_{j=1}^{J_i} \widehat{\alpha}_{ij}^{LSE} B_{ij}(t), \quad i = 1, \ldots, p,$$

where $\widehat{\alpha}_{ij}^{LSE}$'s are entries of $\widehat{\boldsymbol{\alpha}}_{LSE}$. Here, we take marginal approach [26] to construct the LSE criterion without accounting for within subject correlation. Proper modeling of covariance structure would require further parametric assumptions [3] or nonparametric smoothing techniques [26], which is not the focus of this paper.

2.2 B-spline approximation and Sparsity

From the B-spline approximation theory, there exists a function of the form (2.2) which is very close to the true underlying function. While, this function is not capable of characterizing functional sparsity of the true function. Here the term "functional sparsity" is a generalization of the "parameter sparsity" in regression models; see Wang and Kai [22] for more details.

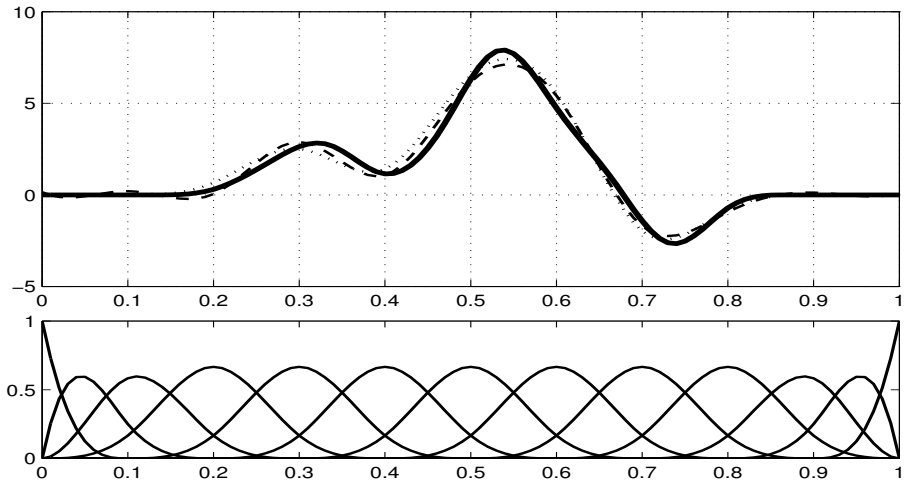

Figure 1: Top: a graphical display of a smooth function (solid thick line type) and two approximating functions from a family of cubic B-spline basis functions with 9 equally-spaced interior knots. Bottom: a graphical display of the set of B-spline functions used in the approximation.

For better illustration, we consider a toy example in Figure 1. Here, in the top panel, a smooth function $\beta(t)$ (thick line) with two spline approximants (dashed, dottted) are depicted. In the bottom, a family of cubic B-spline basis functions with 9 interior knots is shown. The "best" fitted function from the L_2 criterion is shown as the dashed line in the upper panel, which signifies a good performance of the approximation. We further note that $\beta(t)$ is zeros on $[0, 0.1]$ and $[0.9, 1]$; while, its approximation is not zero except for some singletons. From that aspect, this approximation does not capture the sparsity of the true underlying function. In contrast, the dotted curve depicted in the upper panel, also a linear combination of the B-spline basis functions, *automatically corrects* the function to preserve local sparsity with almost indistinguishable performance.

The other extreme case arises when the function is *close to* zero, for part or the whole of the interval. Our goal is to pursue a sparse solution, up to function approximation error, within the linear space spanned by B-spline basis functions. From nonparmetric estimation viewpoint, such solution preserves statistical accuracy and enhances interpretability; in fact, it is indistinguishable from the true underlying function.

Inspired by above observations on functional sparsity, we develop a new procedure that equips the least squares criterion with a regularization term. Usually, the regularization on parameters is expressed in terms of penalty function. Below we introduce a composite penalty based on the B-spline

approximation of the coefficient functions.

2.3 Penalized Least Squares Estimation with Composite Penalty

It is not too difficult to see that global sparsity corresponds to group variable selection of $\boldsymbol{\alpha}_i$ as a whole. To achieve local sparsity, these estimates need to be adjusted in such a way that some of the estimates could be exactly zero. As demonstrated in Section 2.2, we notice that for B-spline approximation, when $\alpha_j = 0$ for $j = l, \ldots, l+d$, the approximation $\widetilde{\beta}(t) = 0$ on the interval $[\eta_{l-1}, \eta_l)$, and especially, when $\alpha_j = 0$ for all j, $\widetilde{\beta}(t) = 0$ over the entire domain of $[0, M]$. This suggests local sparsity need to be imposed at the level of a group of neighboring coefficients. To incorporate global sparsity in varying coefficient model, there needs another layer of group structure. These considerations lead us to a composite penalty defined

$$L_1^{\gamma}(\boldsymbol{\alpha}) = \sum_{i=1}^{p} \sum_{m=1}^{K_i+1} \left(\sum_{j=m}^{m+d} |\alpha_{ij}| \right)^{\gamma},$$

which can be simply written as

$$L_1^{\gamma}(\boldsymbol{\alpha}) = \sum_{i=1}^{p} \sum_{g=1}^{G_i} \|\boldsymbol{\alpha}_{A_{ig}}\|_1^{\gamma}, \tag{2.4}$$

where $\boldsymbol{\alpha}_{A_{ig}} = (\alpha_{ig}, \ldots, \alpha_{i(g+d)})'$, $i = 1, \ldots, p$, $g = 1, \ldots, G_i$. The number of groups for the ith coefficient function is $G_i = K_i + 1$.

Equipping the least squares criterion with penalty (2.4), we obtain the penalized least squares (PLS) criterion

$$\mathrm{pl}(\boldsymbol{\alpha}) = \|\boldsymbol{y} - \boldsymbol{U}\boldsymbol{\alpha}\|_2^2 + \lambda \sum_{i=1}^{p} \sum_{g=1}^{G_i} \|\boldsymbol{\alpha}_{A_{ig}}\|_1^{\gamma}, \tag{2.5}$$

where $\lambda > 0$ and $0 < \gamma < 1$ are tuning parameters. The proposed penalized least squares estimator (PLSE) $\widehat{\boldsymbol{\alpha}} = \widehat{\boldsymbol{\alpha}}(\lambda, \gamma)$ is defined to be the minimizer of $\mathrm{pl}(\boldsymbol{\alpha})$. Consequently, the functional estimate of $\beta_i(t)$ is given by $\widehat{\beta}_i(t) = \boldsymbol{B}_i(t)^T \widehat{\boldsymbol{\alpha}}_i$, where $\widehat{\boldsymbol{\alpha}}_i$ is the subvector of $\widehat{\boldsymbol{\alpha}}$.

Note that, for $\gamma \in (0, 1)$, the penalized criterion $\mathrm{pl}(\boldsymbol{\alpha})$ is not a convex function of $\boldsymbol{\alpha}$. We implement the iterative algorithm proposed and studied by Huang et al. [8] to minimize (2.5). The algorithm is outlined as follows.

Step 1. Obtain an initial value $\boldsymbol{\alpha}^{(0)}$.

Step 2. For a given tuning parameter λ_n, and for $l = 1, 2, \ldots$, compute

$$\theta_{ig}^{(l)} = \left(\frac{1-\gamma}{\tau_n \gamma} \right)^{\gamma} \|\boldsymbol{\alpha}_{A_{ig}}^{(l-1)}\|_1^{\gamma}, \text{ for } i = 1, \ldots, p, \ g = 1, \ldots, G_i,$$

where $\tau_n = (\lambda_n)^{1/(1-\gamma)}\gamma^{\gamma/(1-\gamma)}(1-\gamma)$.

Step 3. Compute

$$\boldsymbol{\alpha}^{(l)} = \arg\min_{\boldsymbol{\alpha}} \|\boldsymbol{y} - \boldsymbol{U}\boldsymbol{\alpha}\|_2^2 + \sum_{i=1}^{p}\sum_{g=1}^{G_i}(\theta_{ig}^{(l)})^{1-1/\gamma}\|\boldsymbol{\alpha}_{A_{ig}}\|_1.$$

Step 4. Repeat steps 2 and 3 until convergence.

Note that unlike the standard LASSO, step 3 requires an overlapping LASSO. As the grouping does not change at each iteration, this can be easily solved with a simple linear transformation with grouping indicator matrix for $\boldsymbol{\alpha}$.

The motivation of this algorithm was given as a reparametrization of the non-convex optimization problem into a complex optimization problem in terms of (θ, τ) that reaches an equivalent solution. In essence, the suggested algorithm performs iteratively reweighted LASSO until convergence, and thus steps 2 and 3 can be expressed in more compact form. Given (λ_n, γ),

Step 1. Obtain an initial value $\boldsymbol{\alpha}^{(0)}$.

Step 2. For $l = 1, 2, \ldots$ define $\nu_{ig}^{(l)} = \gamma\|\boldsymbol{\alpha}_{A_{ig}}^{(l-1)}\|_1^{\gamma-1}$ for $i = 1, \ldots, p; g = 1, \ldots, p$.

Step 3. Solve

$$\boldsymbol{\alpha}^{(l)} = \arg\min_{\boldsymbol{\alpha}} \|\boldsymbol{y} - \boldsymbol{U}\boldsymbol{\alpha}\|_2^2 + \lambda_n\sum_{i=1}^{p}\sum_{g=1}^{G_i}\nu_{ig}^{(l)}\|\boldsymbol{\alpha}_{A_{ig}}\|_1.$$

Step 4 Repeat steps 2 and 3 until convergence.

2.4 Variance Estimation

In this section, we consider the problem of finding the asymptotic variance of our proposed estimator of the coefficient functions. Let $\widehat{\boldsymbol{\alpha}}_S$ denote the non-zero estimators of the coefficients α_{ij}'s, then by Step 3 in the aforementioned algorithm and the Karush-Kuhn-Tucker condition, we have

$$\widehat{\boldsymbol{\alpha}}_S = \left(\boldsymbol{U}_S^T\boldsymbol{U}_S + \frac{1}{2}\boldsymbol{\Theta}_S\right)^{-1}\boldsymbol{U}_S^T\boldsymbol{y},$$

where $\boldsymbol{U}_S$ is the sub-matrix of $\boldsymbol{U}$ with each column corresponding to the selected α_{ij}, and $\boldsymbol{\Theta}_S$ is a diagonal matrix

$$\text{diag}\{\sum_{g:A_{ig}\ni j}\widehat{\theta}_{ig}^{1-1/\gamma}/|\widehat{\alpha}_{ij}|, \text{ for } \widehat{\alpha}_{ij} \neq 0\}.$$

In the absence of covariance modeling of $\boldsymbol{y}$, we further approximate the variance of $\boldsymbol{y}$ by $\sigma^2\boldsymbol{I}$, where σ^2 can be estimated by $\widehat{\sigma}^2 = \|\boldsymbol{y} - \boldsymbol{U}\widehat{\boldsymbol{\alpha}}\|_2^2/n$. Thus, similar to Wang et al. [24], the asymptotic variance of $\widehat{\boldsymbol{\alpha}}_S$ may be expressed as

$$\text{avar}(\widehat{\boldsymbol{\alpha}}_S) = \left(\boldsymbol{U}_S^T\boldsymbol{U}_S + \frac{1}{2}\boldsymbol{\Theta}_S\right)^{-1}\boldsymbol{U}_S^T\boldsymbol{U}_S\left(\boldsymbol{U}_S^T\boldsymbol{U}_S + \frac{1}{2}\boldsymbol{\Theta}_S\right)^{-1}\widehat{\sigma}^2.$$

Let $\mathcal{B}_i(t)$ be the i-th row of the basis matrix $\mathcal{B}(t)$. Thus, the functional estimate of $\beta_i(t)$ can be written as $\widehat{\beta}_i(t) = \mathcal{B}_i(t)\widehat{\boldsymbol{\alpha}}$. Correspondingly, the asymptotic variance of $\widehat{\beta}_i(t)$ is

$$\text{avar}(\widehat{\beta}_i(t)) = \mathcal{B}_{iS}(t)\text{avar}(\widehat{\boldsymbol{\alpha}}_S)\mathcal{B}_{iS}^T(t), \tag{2.6}$$

where $\mathcal{B}_{iS}(t)$ is the sub-vector of $\mathcal{B}_i(t)$ with each element corresponding to the selected α_{ij}. Note that the estimator of $\boldsymbol{\alpha}$ depends on the choice of λ, so the asymptotic variances of $\widehat{\boldsymbol{\alpha}}_S$ and $\widehat{\beta}_i(t)$ are also tuning parameter dependent. Although this a naive estimator, as we shall see in numerical studies, its approximation nevertheless is found to be effective in capturing the level of variability. An alternative is to estimate the full covariance function nonparametrically but, due to its further complexity in implementation with irregular design points, it is not very practical. The literature takes a more pragmatic approach through random-effects formulation (e.g., Wu and Zhang [26]). However, the difficulty of selecting the covariates in the random effects terms under the current context of sparse function estimation outweighs the potential benefits, and we do not pursue this. Instead, we have investigated the usage of a fully nonparametric approach to estimating the covariance surface through a functional principal component analysis [28], however, no clear advantage was found through numerical study. Further investigation is left for future work.

2.5 Choice of Tuning Parameters

In order to fit the model with finite sample, we consider how to calibrate the tuning parameters. The tuning parameter $\lambda > 0$ balances the trade-off between goodness-of-fit and model complexity. When λ is large, we have strong penalization and thus are more likely to obtain a sparse solution with poor model fitting. With small λ, we would select more variables and get better estimation results but lose control of functional sparsity. In classical nonparametric approaches, the criteria such as AIC, BIC and GCV [21] are commonly used for model selection. It has been noted in previous analyses that the AIC and GCV criteria tend to select more variables, and are better suited for prediction purpose. We use a BIC-type criterion in our analysis reported in Section 4. To account for the increasing number of

parameters in comparing models with varying dimensions, we use the extended BIC (EBIC) [7], which also penalizes the size of the full model. The EBIC is given by

$$EBIC(\lambda) = \log\left(\|\boldsymbol{y} - \boldsymbol{U}\widehat{\boldsymbol{\alpha}}(\lambda)\|_2^2/N\right) + K(\lambda)\log(N)/N + \nu K(\lambda)\log(\sum_{i=1}^{p} J_i)/N,$$

where $N = \sum_{k=1}^{n} n_k$, $\widehat{\boldsymbol{\alpha}}(\lambda)$ is the penalized estimator of $\boldsymbol{\alpha}$ given λ, and $K(\lambda)$ is the total number of non-zero estimates in $\widehat{\boldsymbol{\alpha}}(\lambda)$. $\sum_{i=1}^{p} J_i = \sum_{i=1}^{p}(K_i + d + 1)$ is the total number of parameters in the full model. Note that when $\nu = 0$ the EBIC is the same as BIC, but when $\nu > 0$, EBIC puts more penalty on overfitting. We use $\nu = 0.5$ as suggested in [7].

We note that the tuning parameter γ influences the performance of group selection. A value of γ too small or too large could lead to inefficient group variable selection. When γ is close to 1, (2.4) is close to the L_1 penalty. Consequently, the minimizer of (2.5) may not achieve the functional sparsity in its solution. Unlike λ, however, $0 < \gamma < 1$ is more often viewed as a higher-level model parameter (often fixed as 0.5, [8]), in the same spirit as Lasso ($\gamma = 1$) estimator may be chosen over Ridge ($\gamma = 2$) estimator in advance. Our theoretical results suggest that γ is intricately linked to λ in asymptotic sense, similar to [8, 13], and thus the adaptive selection of λ in finite sample is expected to reflect this relation *automatically*. This is also confirmed numerically, shown in Section 4, and our experience also favors the choice of $\gamma = 0.5$ as a rule of thumb.

In addition, as the parametric model formulation arises as an approximation to the nonparametric model, the parameter space to explore is not fixed, and potentially very large. Even with known covariates, the fully adaptive spline approximation to choose the degree, knots location and the number of knots is impractical. Following a similar strategy in literature [e.g. 10, 24] we use equally spaced knots with cubic-splines and select the number of knots K adaptively. We attempted to simultaneously optimize the parameter K inside the model selection criterion but found out that, the penalty was not effective in controlling the systematic increase of parameter space and the criterion favored the smallest possible K in the majority cases. Instead, we select the number of knots K adaptively to the sample by 10-fold cross-validation without penalty, leaving the potentially adaptive choice of sparsity solely controlled by the other tuning parameters.

3 Large Sample Properties

We study large sample properties of our proposed penalized least squares estimator $\widehat{\beta}_i(t)$, $i = 1, \ldots, p$, when the number of sampled subjects n goes to infinity. We assume in the proofs that the number of

observations for each subject n_k is bounded but a similar argument can be applied to the case when n_k increases to infinity with n [10]. The number of interior knots increases with n, so we write $K_i = K_{in}$ for each $i = 1, \ldots, p$, and denote $K_n = \max_{0 \le i \le p} K_i$. The standard regularity conditions for varying coefficient linear models [10, 24] are given in the online supplement.

It is known that, by Theorem 6.27 of Schumaker [17], any smooth coefficient function $\beta_i(t)$ with bounded second derivative has a B-spline approximant $\widetilde{\beta}_i(t)$ of form (2.3) and the approximation error is of order $O(K_{in}^{-2})$. Denote its sparse modification introduced in Section 2.3 by $\widetilde{\beta}_i^0(t)$ with its coefficients $\widetilde{\boldsymbol{\alpha}}^0$.

For our mathematical convenience, we classify all group indices $\{1, \ldots, G_i\}$ for the coefficient function $\beta_i(t)$ into two groups defined as

$$\begin{aligned} \mathcal{A}_{i1} &= \{g : \max_{t \in [\eta_{g-1}, \eta_g)} \mid \beta_i(t) \mid > C_i K_n^{-2}\}, \\ \mathcal{A}_{i2} &= \{g : 0 \le \max_{t \in [\eta_{g-1}, \eta_g)} \mid \beta_i(t) \mid \le C_i K_n^{-2}\}, \end{aligned}$$

for some positive constant C_i. For sufficiently large C_i, the zero region $\{t : \beta_i(t) = 0\}$ is a subset of $\cup_{g \in \mathcal{A}_{i2}} [\eta_{g-1}, \eta_g)$.

Note that for a vector-valued square integrable function $A(t) = (a_1(t), \ldots, a_m(t))^T$ with $t \in [0, M]$, $\|A\|_2$ denotes the L_2 norm defined by $\|A\|_2 = (\sum_{l=1}^m \|a_l\|_2^2)^{1/2}$, where $\|a_l\|_2$ is the usual L_2 norm in function space.

Now, we establish the consistency of our proposed penalized estimator.

Theorem 1 (Consistency)**.** *Suppose that assumptions (A1)-(A6) in the online supplement are satisfied. For some* $0 < \gamma < 1$ *and* $K_n = O(n^{1/5})$*, if the following assumption*

(S1) *for* $\widetilde{\boldsymbol{\alpha}}^0$ *defined above,*

$$\lambda_n (d+1)^{1/2} \left(\sum_{i=1}^p \sum_{g \in \mathcal{A}_{i1}} \|\widetilde{\boldsymbol{\alpha}}_{A_{ig}}^0\|_1^{2(\gamma-1)} \right)^{1/2} = O(n^{1/2})$$

holds, then we have $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 = O_p(n^{-2/5})$*, where* $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$.

Assumption (S1) provides a bound on the rate of λ_n growing with n. The convergence rate established in Theorem 1 is essentially the optimal one [19]. In fact, the result remains valid for more general class of functions, e.g., the collection of functions whose derivatives satisfying the Hölder condition. Next, Theorem 2 states that our proposed penalized method is consistent in detecting functional

sparsity. That is if $\beta_i(t) = 0$ for $t \in [\eta_{l-1}, \eta_l)$, then the proposed estimator will produce $\widehat{\boldsymbol{\alpha}}_{A_{il}} = \mathbf{0}$ to identify local sparsity with probability converging to 1. And if $\beta_i(t) = 0$ for all t, then the proposed method will have $\widehat{\boldsymbol{\alpha}}_{A_{il}} = \mathbf{0}$ for all $l = 1, \ldots, K_i + 1$ with probability converging to 1.

Theorem 2 (Sparsistency)**.** *If assumptions in Theorem 1 and the following assumption*

(S2) $\lambda_n K_n^{\gamma-1} n^{-\gamma/2} \longrightarrow \infty$

are satisfied, then we have for every i, $i = 1, \ldots, p$, $(\widehat{\boldsymbol{\alpha}}_{A_{ig}} : g \in \mathcal{A}_{i2}) = \mathbf{0}$ with probability converging to 1 as n goes to ∞.

It is not surprising that our proposed method may yield a slightly more sparse functional estimate. This is due to the fact that, for all intervals with indices belonging to $\mathcal{A}_{i2}$, the value of $\beta_i(t)$ is quite small, the same order as the optimal rate, and is *indistinguishable* from zeros. Moreover, such intervals can be further partitioned into two groups, including the intervals on which the function is zero and the intervals on which the function is not always zero. While, the total length of the latter converges to zero as n increases.

The above discussion is related to the notion of *selection consistency*, an important and well studied problem of variable selection under parametric settings; for instance, see [29]. However, for nonparametric models, in particular, when local sparsity exists, selection consistency hasn't been widely studied. For the convenience of our discussion, we will begin with some notation. For a coefficient function $\beta(t)$, let $N(\beta)$ and $S(\beta)$ denote the zero region and non-zero region respectively. The (closed) support of β, denoted by $C(\beta)$, is defined as the closure of the non-zero region $S(\beta)$. Assume that $N(\beta)$ has finite many singletons (as zero crossing), and $C(\beta)$ can be expressed as a finite union of closed intervals.

If $\beta(t_0) \neq 0$ for some t_0, the consistency property in Theorem 1 and the smoothness constraint of the function and its estimate ensure that $\widehat{\beta}(t_0) \neq 0$ for sufficiently large n. However, such result may not be of great interest given the fact that $\beta(t)$ lies in an infinite, not necessarily countable, dimensional space. Next, consider a simple case that there is an interval $[a, b] \subset C(\beta)$ and $\beta(t) \neq 0$ for all $t \in [a, b]$. Thus, $\beta(t)$ is bounded away from zero over $[a, b]$. Similarly, as a consequence of Theorem 1, $\widehat{\beta}(t)$ is also bounded away from zero over $[a, b]$ for sufficiently large n. A more challenging case arises when $\beta(a) = 0$ and $\beta(t) \neq 0$ over $(a, b]$. We further assume that there is a sequence of knots such that $\eta_k \leq a < \eta_{k+1} \cdots < \eta_{k'} < b \leq \eta_{k'+1}$. The subinterval formed by two adjacent knots is either in $\mathcal{A}_{i1}$ or in $\mathcal{A}_{i2}$. It can be seen that the total length of the subintervals in $\mathcal{A}_{i2}$ converges to zero as n increases.

For those intervals in $\mathcal{A}_{i1}$, suitable choice of the constant C_i will suggest that the estimated function deviates from zero.

4 Simulation Study

We conducted simulation studies to assess the performance of our proposed method, with the main emphasis on understanding the impact of the tuning parameters and also the increasing dimension p on functional sparsity estimation. We consider three scenarios. In Scenario 1, we choose our tuning parameters (λ, K) as described in Section 2.5 and compare the results under various γ values. In Scenario 2, we assess the impact of increasing dimension p given γ, assuming the number of relevant covariates, p_0, is fixed to be 4. In Scenario 3, we assess the performance with respect to K, to study the effect of the adaptive choice of knots on sparsity estimation. In addition, the relative performance is measured against those from LSE and Lasso methods. The simulation results are summarized based on 400 replications. In each iteration, subjects are randomly generated according to the following varying coefficient model specification

$$y_k(t_{kl}) = \sum_{i=1}^{p} x_{ki}(t_{kl})\beta_i(t_{kl}) + \epsilon_k(t_{kl}), \quad l = 1, \ldots, n_k, \quad k = 1, 2, \ldots, n,$$

where $x_1(t)$ is constant 1, $x_i(t)$, $i = 2, 3, 4$ are similar to those considered in Huang et al. [9]: $x_2(t)$ is a uniform random variable over $[4t, 4t + 2]$; $x_3(t)$ conditioning on $x_2(t)$ is a normal random variable with mean zero and variance $(1 + x_2(t))/(2 + x_2(t))$; and $x_4(t)$, independent of $x_2(t)$ and $x_3(t)$, is Bernoulli(0.6). The number of measurements available varies across the subjects. For each subject, a sequence of 40 possible observation time points $\{(i - 0.5)/40 : i = 1, \ldots, 40\}$ is considered, but each time point has a chance of 0.4 being selected. We further added a random perturbation from $U(-0.5/40, 0.5/40)$ to each observation time. The random errors $\epsilon_k(t_{kl})$ are independent of the predictors but include serial correlation as well as measurement error as $\epsilon_k(t) = \epsilon_k^{(1)}(t) + \epsilon_k^{(2)}(t)$. The serial correlation component $\epsilon^{(1)}(t)$ is generated from a Gaussian process with mean zero and covariance function $cov(\epsilon_k^{(1)}(t), \epsilon_k^{(1)}(s)) = \exp(-10|t - s|)$ for the same subject k and uncorrelated for different subjects, and $\epsilon_k^{(2)}(t)$'s are iid from normal distribution with mean zero and variance 1.

The nonzero coefficient functions used in all scenarios are displayed in Figure 2. The coefficient functions do not belong to the B-spline function space.

In Scenario 1, we add a redundant variable $x_5(t)$ from normal distribution with mean zero and variance $0.1\exp(t)$ for illustration of global sparsity. In Scenario 2, with increasing p, the extra predictors

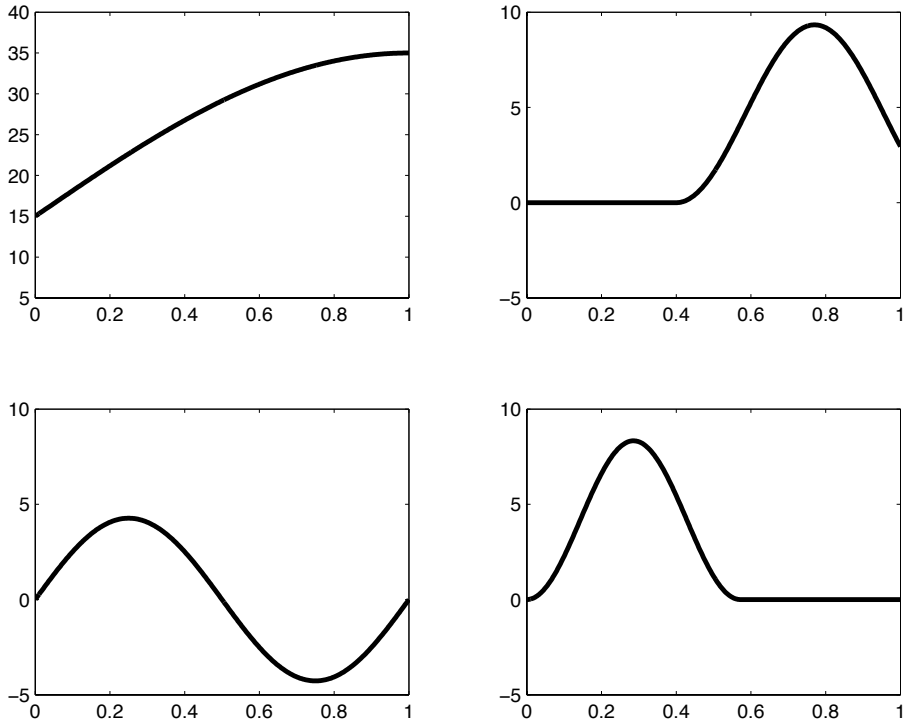

Figure 2: A graphical illustration of the coefficient functions β_i $i = 1, \ldots, 4$ (from left to right, top to bottom)

with zero coefficient functions are defined as $x_i(t) = Z_i(t) + 3/20 \sum_{l=1}^{5} x_l(t)$ for $i = 6, \ldots, p$ with $Z_i(t)$'s being iid from standard normal distribution.

The overall performance is measured in terms of bias and mean integrated squared error (MISE), based on $R = 400$ repetitions, computed as

$$\widehat{Bias}_i(u) = \frac{1}{R} \sum_{r=1}^{R} \widehat{\beta}_i^{(r)}(u) - \beta_i(u), \quad i = 1, \ldots, p, u \in [0, 1],$$

$$\widehat{MISE}_i = \frac{1}{R} \sum_{r=1}^{R} \int_0^1 (\widehat{\beta}_i^{(r)}(u) - \beta_i(u))^2 du, \quad i = 1, \ldots, p,$$

where $\widehat{\beta}_i^{(r)}$ is the estimated coefficient function from the rth repeated study. In addition, we introduce the following summary measures for comparison of functional sparsity:

(a) C_0: average number of correctly identified constant zero coefficient functions

(b) I_0: average number of incorrectly identified constant zero coefficient functions

(c) $C_{i,0}$: average length of correctly identified zero intervals for the ith coefficient function

(d) $I_{i,0}$: average length of incorrectly identified zero intervals for the ith coefficient function.

Note that (a) and (b) summarize global sparsity, while (c) and (d) summarize local sparsity.

Scenario 1: Effect of γ

Here, we consider the varying coefficient model with $p = 5$ and two different numbers of subjects $n = 100, 200$. In each iteration, our proposed PLSE method is implemented with $\gamma = 0.25, 0.35, 0.5$ and 0.75. The MISE values for every coefficient function are summarized in Table 1. In general, as n increases, all methods have smaller MISE values. Notably, the results for PLSE indicate comparable performances across different γ; in fact, PLSE and Lasso methods have similar performance in function estimation. In addition, PLSE with $\gamma = 0.35, 0.50$ can successfully identify the global sparsity of $\beta_5(\cdot)$ with zero MISE values for both choices of n, and so does PLSE with $\gamma = 0.75$ for $n = 200$. Bias of $\text{PLSE}_{0.5}$ (PLSE with $\gamma = 0.5$), Lasso and LSE with $n = 200$ is compared in Figure 3. It can be seen that $\text{PLSE}_{0.5}$ has zero bias in estimating $\beta_5(\cdot)$.

Method	MISE				
	β_1	β_2	β_3	β_4	β_5
			$\boldsymbol{n = 100}$		
LSE	0.9519	0.0825	0.0365	0.1145	1.3199
Lasso	2.9156	0.1636	0.0314	0.0591	0.0114
$\text{PLSE}_{0.25}$	1.1115	0.0686	0.0281	0.0613	0.1330
$\text{PLSE}_{0.35}$	1.2199	0.0633	0.0307	0.0440	0
$\text{PLSE}_{0.5}$	1.3156	0.0674	0.0319	0.0459	0
$\text{PLSE}_{0.75}$	1.8267	0.0948	0.0317	0.0471	0.0005
			$\boldsymbol{n = 200}$		
LSE	0.4232	0.0367	0.0165	0.0563	0.5745
Lasso	1.4561	0.0845	0.0153	0.0299	0.0041
$\text{PLSE}_{0.25}$	0.7259	0.0424	0.0138	0.0329	0.0731
$\text{PLSE}_{0.35}$	0.6421	0.0351	0.0152	0.0235	0
$\text{PLSE}_{0.5}$	0.7193	0.0382	0.0166	0.0250	0
$\text{PLSE}_{0.75}$	0.8615	0.0469	0.0157	0.0251	0

Table 1: Comparison of MISE for each coefficient function in Scenario 1.

In Table 2, performance in identifying local sparsity is demonstrated. The true values of sparsity in terms of $C_{i,0}$ and $I_{i,0}$ are given in the last row of *true model* as a reference. Hence, the closer the values of $C_{i,0}$ are to those in the true model, the better. On the contrary, the value of $I_{i,0}$ in true

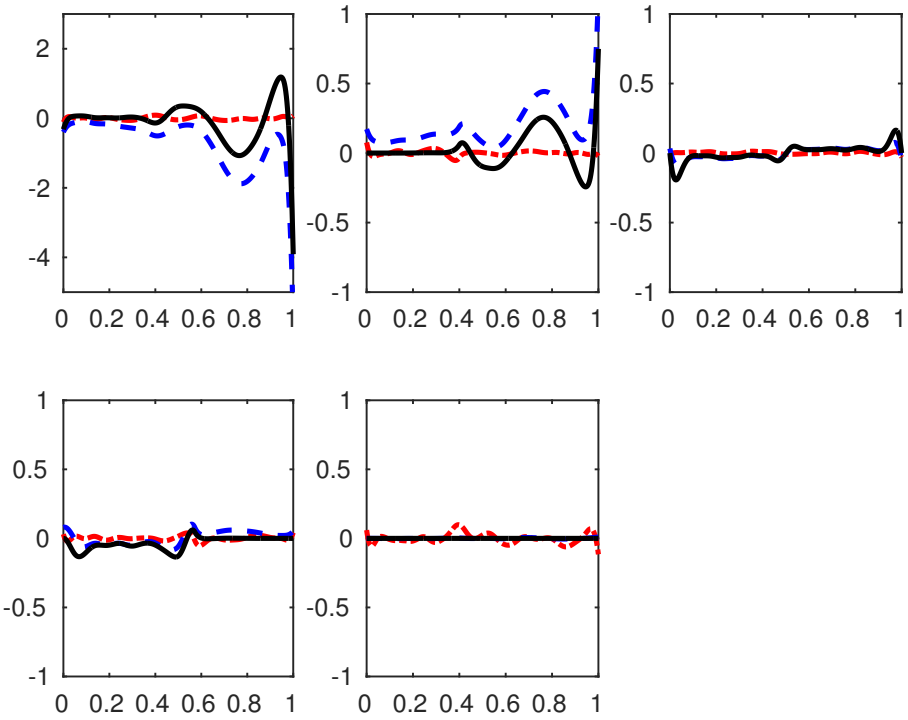

Figure 3: Comparison of bias of the coefficient functions based on LSE (dot-dashed), Lasso (dashed) and $\text{PLSE}_{0.5}$ (solid) in Scenario 1 with $n = 200$. Note that $\text{PLSE}_{0.5}$ has zero bias in estimating the zero coefficient function $\beta_5(\cdot)$.

model is the maximum error each method can make, so the smaller $I_{i,0}$, the better. In general, Lasso and PLSE have better performance in functional sparsity. In addition, it can be seen that PLSE with $\gamma = 0.35, 0.5, 0.75$ has an advantage in achieving both global and local sparsity compared with LSE and Lasso. The case that $\gamma = 0.5$ slightly outperforms the others. For the remaining part, we will use $\gamma = 0.5$ for comparison.

Scenario 2: Effect of dimension p

In this scenario, we study the effect of increasing p on the performance with a given sample size. In particular, we consider three different choices of $p = 5, 20$, and 50. Figure 4 and Table 3 show the results of bias and MISE respectively. The last column of MISE tables indicates the maximum MISE among the zero coefficient functions, as the selected variables vary between sample to sample. Compared to LSE and LASSO, the PLSE method shows remarkable stability in performance over

Method	$C_{1,0}$	$I_{1,0}$	$C_{2,0}$	$I_{2,0}$	$C_{3,0}$	$I_{3,0}$	$C_{4,0}$	$I_{4,0}$	$C_{5,0}$	$I_{5,0}$	C_0	I_0
						$\boldsymbol{n = 100}$						
LSE	0	0	0	0	0	0	0	0	0	0	0	0
Lasso	0	0	0.0219	0	0	0	0.0616	0.0009	0.8799	0	0.5675	0
$\text{PLSE}_{0.25}$	0	0	0.1468	0.0003	0	0	0.1626	0.0004	0.8005	0	0.4975	0
$\text{PLSE}_{0.35}$	0	0	0.3332	0.0048	0	0	0.3723	0.0062	1.0000	0	1	0
$\text{PLSE}_{0.5}$	0	0	0.3360	0.0040	0	0	0.3809	0.0072	1.0000	0	1	0
$\text{PLSE}_{0.75}$	0	0	0.2559	0.0005	0	0	0.3453	0.0042	0.9990	0	0.9975	0
						$\boldsymbol{n = 200}$						
LSE	0	0	0	0	0	0	0	0	0	0	0	0
Lasso	0	0	0.0166	0	0	0	0.0736	0.0004	0.9087	0	0.6850	0
$\text{PLSE}_{0.25}$	0	0	0.1299	0.0001	0	0	0.1433	0.0003	0.7510	0	0.4175	0
$\text{PLSE}_{0.35}$	0	0	0.3178	0.0022	0	0	0.3512	0.0030	1.0000	0	1	0
$\text{PLSE}_{0.5}$	0	0	0.3343	0.0025	0	0	0.3735	0.0047	1.0000	0	1	0
$\text{PLSE}_{0.75}$	0	0	0.2696	0.0005	0	0	0.3462	0.0026	1.0000	0	1	0
true model	0	1	0.4000	0.6000	0	1	0.4286	0.5714	1	0	1	4

Table 2: Sparsity summary measures (a)-(d) in Scenario 1. Here, for the true model, $C_{i,0}$, $i = 1, \ldots, 6$ are the lengths of zero intervals, $I_{i,0}$'s are the lengths of nonezero intervals, C_0 is the number of zero coefficient functions, and I_0 is the number of nonzero coefficient functions.

increasing dimension p.

Method	MISE					$\max_{i \geq 6} \text{MISE}_i$
	β_1	β_2	β_3	β_4	β_5	
			$\boldsymbol{p = 5}$			
LSE	0.4232	0.0367	0.0165	0.0563	0.5745	—
Lasso	1.4561	0.0845	0.0153	0.0299	0.0041	—
$\text{PLSE}_{0.5}$	0.7193	0.0382	0.0166	0.0250	0	—
			$\boldsymbol{p = 20}$			
LSE	0.5157	0.0434	0.0197	0.0612	0.6694	0.0151
Lasso	17.8758	0.8520	0.0331	0.0347	0	0.0016
$\text{PLSE}_{0.5}$	0.7422	0.0391	0.0166	0.0240	0	2.1886e-05
			$\boldsymbol{p = 50}$			
LSE	0.8269	0.0724	0.0292	0.0897	1.1484	0.0281
Lasso	35.1543	1.6475	0.0497	0.0415	0	8.4758e-04
$\text{PLSE}_{0.5}$	0.7360	0.0396	0.0149	0.0205	0	2.7551e-05

Table 3: Comparison of MISE for each coefficient function with $p = 5, 20$ and 50 in Scenario 2.

The performance of sparsity is summarized in Table 4. The additional two columns in $C_{i,0}$ and $I_{i,0}$ are added to summarize the performance on all other redundant variables, as an interval range of $[\min_{i \geq 6} C_{i,0}, \max_{i \geq 6} C_{i,0}]$ and $[\min_{i \geq 6} I_{i,0}, \max_{i \geq 6} I_{i,0}]$. Together with the global sparsity measure in C_0

Method	$C_{1,0}$	$I_{1,0}$	$C_{2,0}$	$I_{2,0}$	$C_{3,0}$	$I_{3,0}$	$C_{4,0}$	$I_{4,0}$	$C_{5,0}$	$I_{5,0}$	$[C_{i,0}^{(\min)}, C_{i,0}^{(\max)}]$	$[I_{i,0}^{(\min)}, I_{i,0}^{(\max)}]$	C_0	I_0
								$\boldsymbol{p = 5}$						
LSE	0	0	0	0	0	0	0	0	0	0	—	—	0	0
Lasso	0	0	0.0166	0	0	0	0.0736	0.0004	0.9087	0	—	—	0.6850	0
$\text{PLSE}_{0.5}$	0	0	0.3343	0.0025	0	0	0.3735	0.0047	1	0	—	—	1	0
true model	0	1	0.4000	0.6000	0	1	0.4286	0.5714	1	0	—	—	1	4
								$\boldsymbol{p = 20}$						
LSE	0	0	0	0	0	0	0	0	0	0	[0,0]	[0,0]	0	0
Lasso	0	0	0.0060	0	0	0	0.2398	0.0027	1	0	[0.4128, 0.5309]	[0,0]	2.1125	0
$\text{PLSE}_{0.5}$	0	0	0.3286	0.0015	0	0	0.3754	0.0056	1	0	[0.9985, 1.0000]	[0,0]	15.9675	0
true model	0	1	0.4000	0.6000	0	1	0.4286	0.5714	1	0	[1,1]	[0,0]	16	4
								$\boldsymbol{p = 50}$						
LSE	0	0	0	0	0	0	0	0	0	0	[0,0]	[0,0]	0	0
Lasso	0	0	0.0005	0	0	0	0.2949	0.0018	1	0	[0.6039, 0.8226]	[0,0]	20.4425	0
$\text{PLSE}_{0.5}$	0	0	0.3119	0.0005	0	0	0.3680	0.0015	1	0	[0.9971, 1.0000]	[0,0]	45.9000	0
true model	0	1	0.4000	0.6000	0	1	0.4286	0.5714	1	0	[1,1]	[0,0]	46	4

Table 4: Sparsity summary measures (a)-(d) for Scenario 2. Here, for the true model, $I_{i,0}$, $i = 1, \ldots, 4$ are the lengths of nonzero intervals, and $C_{i,0}$'s are the lengths of zero intervals.

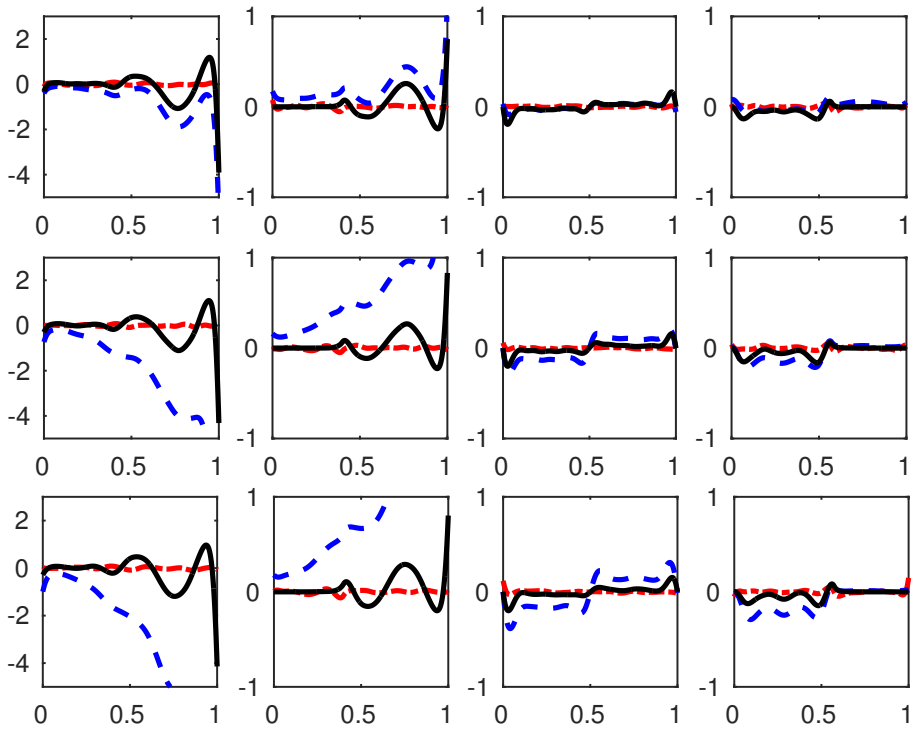

Figure 4: Comparison of bias of the nonzero coefficient functions $\beta_1, \beta_2, \beta_3$ and β_4 (from left to right) based on LSE (dot-dashed), Lasso (dashed) and $\text{PLSE}_{0.5}$ (solid) for $p = 5$ (top row), $p = 20$ (middle) and $p = 50$ (bottom) in Scenario 2.

and I_0, we can conclude that $\text{PLSE}_{0.5}$ systematically outperforms the other methods for all dimensions.

Scenario 3: Effect of knots selection

The variation in knots selection is expected to mainly influence the estimation of local sparsity. Increasing number of knots helps in identifying the boundary of local sparsity, with the risk of over-fitting non-zero estimates. Fine-tuning this parameter is much more delicate, as all model selection criteria are developed to control the squared error loss (MISE) as goodness of fit, and thus are insensitive to the loss of missing local sparsity. That is, balance between global and local sparsity is beyond the usual control of bias and variance trade-off, and developing a new measure is still an open problem. Our knot selection based on cross-validation is essentially tuned towards global sparsity. Here we assess the performance of our proposed estimator from the point of view of robustness to these variations. For comparison, we include results for fixed knots ($K = 11$) across the sample.

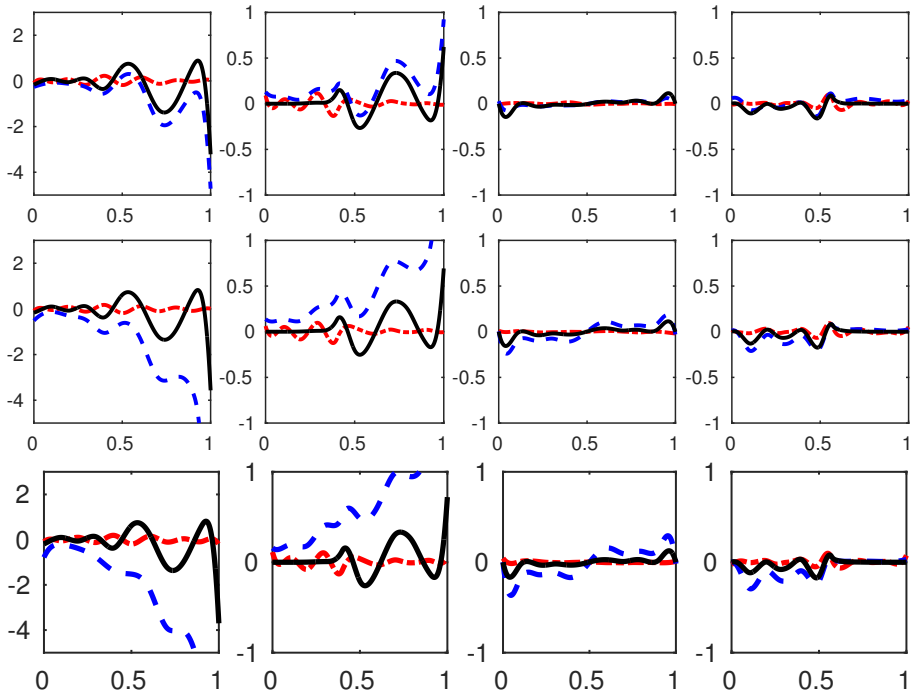

Figure 5: Comparison of bias of the nonzero coefficient functions $\beta_1, \beta_2, \beta_3$ and β_4 (from left to right) based on LSE (dot-dashed), Lasso (dashed) and $\text{PLSE}_{0.5}$ (solid) for $p = 5$ (top row), $p = 20$ (middle) and $p = 50$ (bottom) in Scenario 3.

Figure 5 and Table 5 summarize the bias and MISE. The sparsity summary is given in Table 6. We conclude that the overall performance is fairly comparable to those in Scenario 2 with no major concern over the sensitivity of the knots selection in comparison of the result.

Method	MISE					$\max_{i\geq 6}$ MISE$_i$
	β_1	β_2	β_3	β_4	β_5	
			$\boldsymbol{p=5}$			
LSE	0.2783	0.0253	0.0108	0.0379	0.3888	—
Lasso	1.2993	0.0753	0.0096	0.0195	0.0029	—
PLSE$_{0.5}$	0.6888	0.0405	0.0107	0.0154	0	—
			$\boldsymbol{p=20}$			
LSE	0.3376	0.0292	0.0127	0.0408	0.4429	0.0097
Lasso	11.6604	0.5420	0.0199	0.0211	0	0.0011
PLSE$_{0.5}$	0.7180	0.0412	0.0114	0.0156	0	2.1444e-05
			$\boldsymbol{p=50}$			
LSE	0.4521	0.0387	0.0167	0.0528	0.6608	0.0138
Lasso	30.2698	1.3683	0.0352	0.0290	0	7.0743e-04
PLSE$_{0.5}$	0.7279	0.0417	0.0114	0.0150	0	1.9744e-05

Table 5: Comparison of MISE for each coefficient function in Scenario 3 . Here, the number of knots is fixed to be 11.

In addition, in order to assess the usefulness of the asymptotic formula for the standard errors in (2.6), both asymptotic and empirical standard errors based on 400 repetitions are calculated, and compared in Figure 6 with adaptive number of knots and in Figure 7 with fixed number of knots, which show a good agreement between them. It can be seen that the variation in number of knots greatly increases the variation in estimation of coefficient functions.

In summary, the simulation results demonstrate that our proposed method not only has an advantage in achieving local sparsity compared with Lasso and LSE, but also can ensure global sparsity for finite dimensional models. Moreover, this advantage is carried onto models with increasing dimension.

5 Real Data Analysis

We demonstrate our method in an application to the analysis of yeast cell cycle gene expression data [14, 18].

In biological sciences, gene expression data are frequently collected. Scientists believe that transcription factors (TFs) might have effect on genome's cell cycle regulation. They have made great effort

Method	$C_{1,0}$	$I_{1,0}$	$C_{2,0}$	$I_{2,0}$	$C_{3,0}$	$I_{3,0}$	$C_{4,0}$	$I_{4,0}$	$C_{5,0}$	$I_{5,0}$	$[C_{i,0}^{(\min)}, C_{i,0}^{(\max)}]$	$[I_{i,0}^{(\min)}, I_{i,0}^{(\max)}]$	C_0	I_0
								$\boldsymbol{p = 5}$						
LSE	0	0	0	0	0	0	0	0	0	0	—	—	0	0
Lasso	0	0	0.0120	0	0	0	0.0660	0	0.9105	0	—	—	0.7800	0
$\text{PLSE}_{0.5}$	0	0	0.2647	0	0	0	0.3348	0	1.0000	0	—	—	1	0
true model	0	1	0.4000	0.6000	0	1	0.4286	0.5714	1	0	—	—	1	4
								$\boldsymbol{p = 20}$						
LSE	0	0	0	0	0	0	0	0	0	0	[0,0]	[0,0]	0	0
Lasso	0	0	0.0050	0	0	0	0.2087	0	1.0000	0	[0.3632,0.4610]	[0,0]	2.5075	0
$\text{PLSE}_{0.5}$	0	0	0.2682	0	0	0	0.3468	0	1.0000	0	[0.9980,1.0000]	[0,0]	15.9725	0
true model	0	1	0.4000	0.6000	0	1	0.4286	0.5714	1	0	[1,1]	[0,0]	16	4
								$\boldsymbol{p = 50}$						
LSE	0	0	0	0	0	0	0	0	0	0	[0,0]	[0,0]	0	0
Lasso	0	0	0.0010	0	0	0	0.2863	0	1.0000	0	[0.5870, 0.7960]	[0,0]	21.1400	0
$\text{PLSE}_{0.5}$	0	0	0.2717	0	0	0	0.3518	0	1.0000	0	[0.9970, 1.0000]	[0,0]	45.9300	0
true model	0	1	0.4000	0.6000	0	1	0.4286	0.5714	1	0	[1,1]	[0,0]	46	4

Table 6: Sparsity summary measures (a)-(d) for Scenario 3. Here, for the true model, $I_{i,0}$, $i = 1, \ldots, 4$ are the lengths of nonzero intervals, and $C_{i,0}$'s are the lengths of zero intervals.

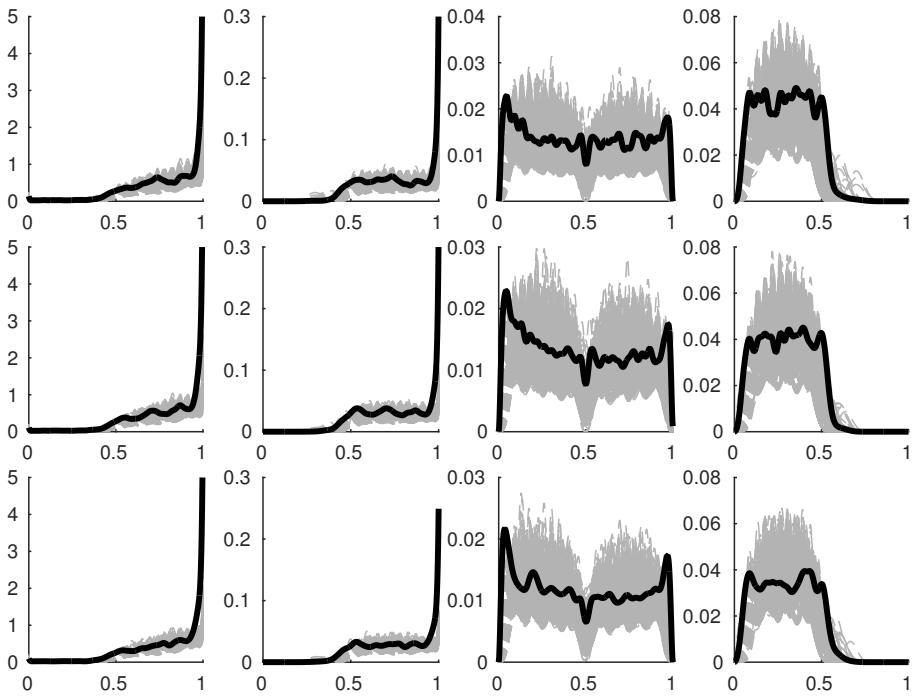

Figure 6: Asymptotic standard error (grey solid line), and empirical standard deviation (black solid line) of the coefficient functions with adaptive number of knots in Scenario 2.

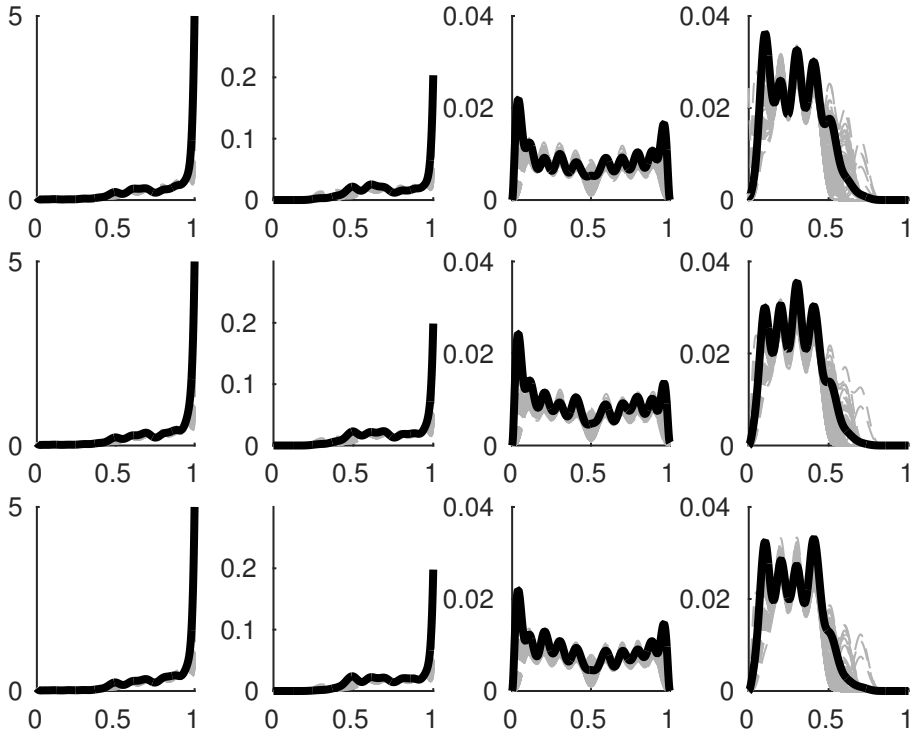

Figure 7: Asymptotic standard error (grey solid line), and empirical standard deviation (black solid line) of the coefficient functions with fixed number of knots in Scenario 3.

in identifying key TFs in the regulatory network based on a set of gene expression measurements. In this study, we analyze the relationship between the level of gene expression and the physical binding of TFs from chromatin immunoprecipitation (ChIP-chip) data [14]. One of the gene expression data comes from an α factor synchronization experiment of 542 genes, in which mRNA levels are measured every 7 minutes during 119 minutes, resulting in 18 measurements in total [18]. For our analysis, the time has been rescaled to [0,1].

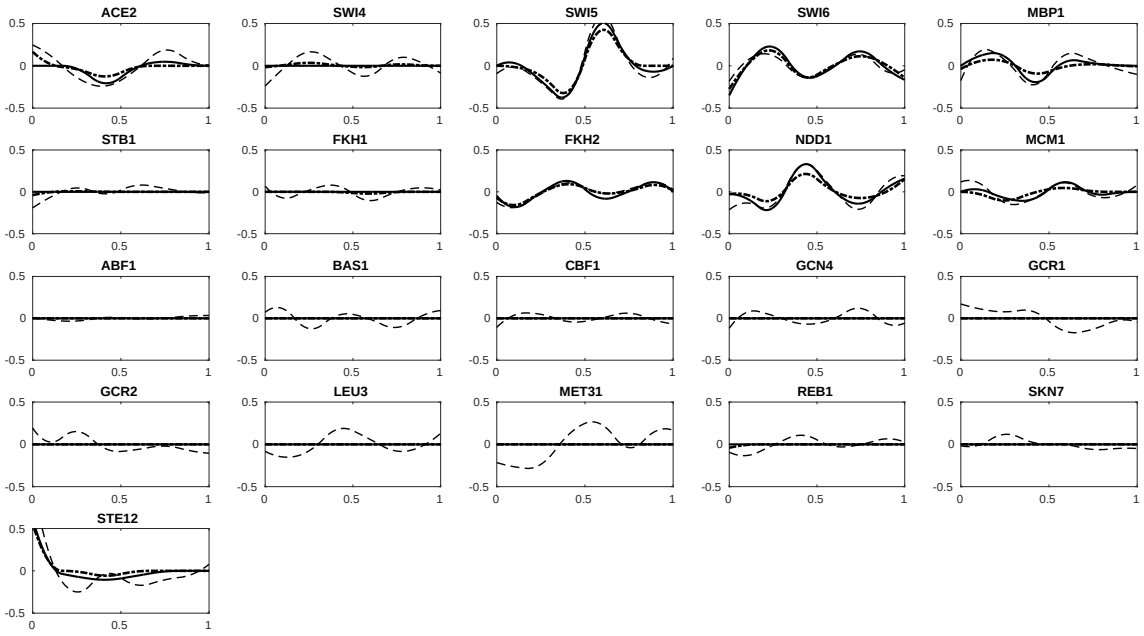

Figure 8: Subplots of estimated coefficient functions for the 21 confirmed TFs using LSE (dashed), Lasso (dot-dashed) and $\text{PLSE}_{0.5}$ (solid).

The ChIP-chip data contains the binding information of 106 transcription factors, among which 21 TFs are confirmed to be related to cell cycle regulation by experiment. Wang et al. [23] demonstrated that a variable selection procedure is able to identify some of those key TFs. It is believed that the effects of TFs vary during the cell cycle. In [1], the authors considered a sparse partial least squares regression to study which TFs are important in gene expression. But they did not focus on the active periods of TFs. In this paper we apply our method to identify the key TFs and estimate the effects of those selected TFs over time. In addition, our approach allows us to investigate whether active and inactive periods during the cycle could be identified for each TF. Let y_{kt} denote the gene expression level for gene k at time t for $k = 1, \ldots, 542$ and $t = 1, \ldots, 18$, and let x_{ki} denote the binding information

of transcription factor i for gene k, for $i = 1, \ldots, 106$. Then the varying coefficient model can be written as

$$y_{kt} = \beta_0(t) + \sum_{i=1}^{106} \beta_k(t) x_{ik} + \epsilon_{kt},$$

where $\beta_i(t)$ models the effect of the ith transcription factor on gene expression at time t, and for the k-th gene ϵ_{kt}'s are independent over time.

Similar to the simulation study, we apply our method together with LSE and Lasso methods and compared the identification of active period of each TF within the cell cycle process. Each coefficient function is approximated with quadratic B-splines defined on time interval $[0, 1]$ with seven equally spaced knots. The number of knots is selected by cross-validation. It is not surprising that LSE selects all TFs. The lasso method identifies 32 TFs as important, while our proposed method identifies 16 TFs, which are in fact the subset of those identified by the lasso method. In Figure 8, the estimated coefficient functions for 21 experimentally confirmed TFs are shown. From this figure, we could tell 8 of them are selected by both methods. The lasso method selects additional four TFs, namely, SWI4, STB1, FKH1, REB1, which show very low level of activities. In [1], the authors selected 32 TFs, 10 of which are verified TFs. In addition, our proposed method identifies some inactive periods for selected TFs. For example, STE12 tends to be inactive for the later period, and ACE2 is inactive at early period.

Acknowledgement

The authors are grateful to three referees and the associate editor for careful reading and helpful comments, which have lead to substantial improvements of the paper. The research of Haonan Wang was partially supported by NSF grants DMS-1106975, DMS-1521746 and DMS-1737795.

Supplementary Material

The online supplementary material includes technical assumptions and proofs of the theoretical properties of our proposed method.

References

[1] Chun, H. and Keleş, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B*, 72:3–25.

[2] de Boor, C. (2001). *A Practical Guide to Splines*. Springer.

[3] Diggle, P., Heagerty, P., Liang, K.-Y., and Zeger, S. (1994). *Analysis of longitudinal data*. OUP Oxford.

[4] Hall, P. and Hooker, G. (2016). Truncated linear models for functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(3):637–653.

[5] Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society: Series B*, 55:757–796.

[6] Hoover, D. R., Rice, J. A., Wu, C. O., and Yang, L. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, 85:809–822.

[7] Huang, J., Horowitz, J. L., and Wei, F. (2010). Variable selection in nonparametric additive models. *Annals of Statistics*, 38(4):2282–2313.

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

[9] Huang, J. Z., Wu, C. O., and Zhou, L. (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika*, 89:111–128.

[10] Huang, J. Z., Wu, C. O., and Zhou, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica*, 14:763–788.

[11] James, G. M., Wang, J., and Zhu, J. (2009). Functional linear regression that's interpretable. *Annals of Statistics*, 37:2083–2108.

[12] Kneip, A., Poß, D., and Sarda, P. (2016). Functional linear regression with points of impact. *Ann. Statist.*, 44(1):1–30.

[13] Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *Annals of Statistics*, 28:1356–1378.

[14] Lee, T., Rinaldi, N., Robert, F., Odom, D., Bar-Joseph, Z., Gerber, G., Hannett, N., Harbison, C., Thomson, C., Simon, I., Zeitlinger, J., Jennings, E., Murray, H., Gordon, D., Ren, B., Wyrick, J., Tagne, J., Volkert, T., Fraenkel, E., Gifford, D., and Young, R. (2002). Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, 298:799–804.

[15] Müller, H.-G. (2016). Peter hall, functional data analysis and random objects. *Ann. Statist.*, 44(5):1867–1887.

[16] Noh, H. S. and Park, B. U. (2010). Sparse varying coefficient models for longitudinal data. *Statistica Sinica*, 20:1183–1202.

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

[18] Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, 9:3273–3279.

[19] Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10:1040–1053.

[20] Tu, C. Y., Song, D., Breidt, F. J., Berger, T. W., and Wang, H. (2012). Functional model selection for sparse binary time series with multiple input. In *Economic Time Series: Modeling and Seasonality*, pages 477–497. Chapman and Hall/CRC.

[21] Wahba, G. (1990). *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics.

[22] Wang, H. and Kai, B. (2015). Functional sparsity: Global versus local. *Statistica Sinica*, 25:1337–1354.

[23] Wang, L., Chen, G., and Li, H. (2007). Group scad regression analysis for microarray time course gene expression data. *Bioinformatics*, 23:1486–1494.

[24] Wang, L., Li, H., and Huang, J. Z. (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *Journal of the American Statistical Association*, 103:1556–1569.

[25] Wei, F., Huang, J., and Li, H. (2011). Variable selection and estimation in high-dimensional varying-coefficient models. *Statistica Sinica*, 21:1515–1540.

[26] Wu, H. and Zhang, J.-T. (2006). *Nonparametric regression methods for longitudinal data analysis: mixed-effects modelling approaches*. Wiley.

[27] Xue, L. and Qu, A. (2012). Variable selection in high-dimensional varying-coefficient models with global optimality. *The Journal of Machine Learning Research*, 13:1973–1998.

[28] Yao, F., Müller, H.-G., and Wang, J.-L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100:577–590.

[29] Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563.

[30] Zhou, J., Wang, N.-Y., and Wang, N. (2013). Functional linear model with zero-value coefficient function at sub-regions. *Statistica Sinica*, 23:25–50.