Statistica Sinica **30** (2020), 439-465
doi:https://doi.org/10.5705/ss.202017.0246

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 and* [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 a functional regression mostly appears within a functional linear regression with a scalar response; however, there is no direct correspondence to functional response models with many covariates. We reformulate the problem as a nonparametric function estimation problem under the notion of *functional sparsity*. Sparsity encapsulates interpretability in a regression with multiple inputs, and the problem of sparse estimation is well understood in the context of variable selection in a parametric setting. For nonparametric models, interpretability not only concerns the number of covariates involved, but also the zero regions in the *functional form*. Thus, the sparsity consideration is much more complex. To distinguish the types of sparsity in nonparametric models, we refer to the former as *global sparsity* and to the latter as *local sparsity*, both of which constitute functional sparsity. Most existing methods focus on directly extending the framework of parametric sparsity for linear models to nonparametric models to address one type of sparsity, but not both. We develop a penalized estimation procedure that simultaneously addresses both types of sparsity in a unified framework. We establish the asymptotic properties of estimation consistency and sparsistency of the proposed method. Our method is illustrated by means of a simulation study and real-data analysis, and is shown to outperform existing methods in terms of identifying both local and global sparsity.

Key words and phrases: 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 over time, recorded for multiple subjects with multiple predictors. The varying coefficient models (Hastie and Tibshirani (1993); Hoover et al.

(1998)) 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 a 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 enhancing the utility of the standard linear model formulation. For generality, we consider the predictors to be functions. However, note that the varying coefficient models are equally applicable when the predictors take scalar values.

The domain selection problem in functional regression is known to be intrinsically difficult (Müller (2016)). Most prior studies examine the problem as a functional linear regression with a scalar response and a single functional covariate. Hall and Hooker (2016) formulated the problem as a truncated regression model with a single unknown domain, in order to study the identifiability issues in nonparametric function estimation. James, Wang and Zhu (2009) approached the problem from the viewpoint of sparsity estimation as *interpretable* solutions. Using a 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 between zero and non-zero regions. However, as Zhou, Wang and Wang (2013) have noted, the overlapping contribution of each coefficient to the neighboring regions means the independent shrinkage of the coefficients does not necessarily induce zero values in the coefficient function, in general. Thus, the procedure tends to over-penalize. As a remedy, Zhou, Wang and Wang (2013) suggested a two-step estimation procedure. Wang and Kai (2015) studied a similar problem under a standard nonparametric regression, suggesting the need to distinguish the *functional* features from the parametric variable selection.

We consider the regression problem under a *functional response* variable with varying coefficient models, involving multiple domain selection under the general setting, where the true number of covariates is unknown. Although the views and approaches adopted in prior studies are quite different, the domain selection problem can be motivated as a means to enhance the interpretability of the model selection in nonparametric models. In this regard, we share the view that considering some form of sparsity could be useful. For nonparametric models,

however, interpretability not only concerns the number of covariates involved (Wang, Li and Huang (2008); Noh and Park (2010); Wei, Huang and Li (2011); Xue and Qu (2012)), but also the zero regions in the *functional form* (James, Wang and Zhu (2009); Zhou, Wang and Wang (2013)). To distinguish between the types of sparsity in nonparametric models, we refer to the former as *global sparsity* and to the latter as *local sparsity*, both of which constitute *functional sparsity* (Tu et al. (2012); Wang and Kai (2015)). More formally, a function has *global sparsity* if it is zero over the entire domain. This 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. Thus, this identifies an inactive period for the corresponding covariate. These notions of interpretability are used informally in a separate context of the analysis. Thus, the significance of local sparsity estimation has not been well recognized.

We reformulate the domain selection problem as a nonparametric function estimation problem under the unified theme of *functional sparsity*. Then, we propose a one-step penalized estimation procedure that automatically determines the type of functional sparsity (i.e., local or global). Although we distinguish between the two types of sparsity on a conceptual level, our unified formulation does not require this distinction for the implementation. We directly exploit the fact that global sparsity is a special case of local sparsity from the viewpoint of domain selection, but not the other way around. Furthermore, the consistency of the coefficient function estimation does not necessarily provide us with information on local sparsity. This feature distinguishes our approach from the majority of existing methods that target global sparsity, such as (Wei, Huang and Li (2011)). Note that there is a fundamental difference in the underlying assumption on sparsity between parametric and nonparametric models, because we focus on function estimation with *dependent* variables over an unknown domain. This difference was also recognized by Kneip, Poß and Sarda (2016). Moreover, for parametric sparsity, an underlying sparse vector is specified. In contrast, the *true* sparse representation for functional sparsity may not be well defined in the function approximation. These differences pose different conceptual challenges in the development. Our proposed penalized procedure resembles a type of parametric sparsity estimation. However, our analysis is not comparable with those that focus on high-dimensional parametric sparsity estimation (e.g., James, Wang and Zhu (2009)).

We provide a theoretical analysis of the proposed method. In particular, we show that the local sparsity can be recovered consistently, and even diluted with

the problem of global sparsity estimation. We study the properties of our proposed method under the standard assumptions on nonparametric smooth function estimation, and exploit the functional property in a more natural manner. In this way, we 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 that allows us to directly control the functional sparsity through the coefficient functions themselves, rather than using a pointwise evaluation. In Section 3, we study the large-sample properties of the proposed method and establish the consistency and sparsistency of the function estimates. Section 4 describes our simulation studies under different scenarios, and a real-data analysis is provided in Section 5, demonstrating the utility of functional sparsity in relation to the interpretability of the results. All technical assumptions and proofs are provided in the online Supplementary Material.

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 $\beta_i(t)$, for $i = 1, \ldots, p$, 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 the parameter estimation.

2.1. Least squares estimation under a b-spline approximation

B-spline approximations are widely used to estimate smooth nonparametric functions. For a detailed discussion on B-splines, see de Boor (2001) and Schumaker (1981). 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 $0 = \eta_0 < \eta_1 < \cdots < \eta_K < \eta_{K+1} = 1$. Note that K is the number of

interior knots and $J = K + d + 1$. Here, we adopt the definition of a B-spline as stated in Definition 4.12 of Schumaker (1981). In general, the performance of B-spline approximations has been well studied. For instance, under 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 (1981) for further detail.

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 $\boldsymbol{\alpha}$ (Huang, Wu and Zhou (2002)) 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 $\boldsymbol{U}_k = (\mathcal{U}_k^T(t_{k1}), \ldots, \mathcal{U}_k^T(t_{kn_k}))^T$. The weights ω_k, for $k = 1, \ldots, n$, are usually chosen as $\omega_k \equiv 1$ or $\omega_k \equiv 1/n_k$ (Huang, Wu and Zhou (2004)). In this study, for simplicity, we set equal weights to every subject (i.e., $\omega_k \equiv 1$). Setting $\boldsymbol{U} = (\boldsymbol{U}_1^T, \ldots, \boldsymbol{U}_n^T)^T$ 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, Wu and Zhou (2004) 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 (LSE) of $\boldsymbol{\alpha}$. Thus, the LSEs of the 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 $\widehat{\alpha}_{ij}^{LSE}$ denotes an entry of $\widehat{\boldsymbol{\alpha}}_{LSE}$. Here, we take a marginal approach

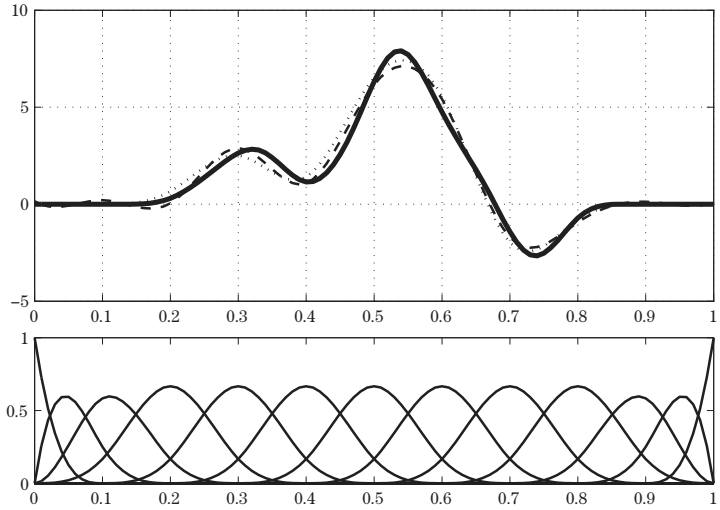

Figure 1. Top: A graphical display of a smooth function (solid thick line) and two approximating functions from a family of cubic B-spline basis functions with nine equally spaced interior knots. Bottom: A graphical display of the set of B-spline functions used in the approximation.

(Wu and Zhang (2006)) to construct the LSE criterion, without accounting for within-subject correlation. Proper modeling of the covariance structure would require further parametric assumptions (Diggle et al. (1994)) or nonparametric smoothing techniques (Wu and Zhang (2006)), which are not the focus of this study.

2.2. B-spline approximation and sparsity

From B-spline approximation theory, there exists a function of the form given in (2.2) that is very close to the true underlying function. However, this function is not capable of characterizing the 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 (2015) for further detail.

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, dotted) is depicted. In the bottom panel, a family of cubic B-spline basis functions with nine 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 good performance of the approximation. Furthermore, $\beta(t)$ is zero on [0, 0.1] and [0.9, 1], but its approximation is not zero, except for some singletons. From this

aspect, the 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 reflect local sparsity, with almost indistinguishable performance.

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

Inspired by the 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 a 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 a composite penalty

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

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

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

Equipping the least squares criterion with the penalty defined in (2.4), we

obtain the following 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 LSE (PLSE) $\widehat{\boldsymbol{\alpha}} = \widehat{\boldsymbol{\alpha}}(\lambda, \gamma)$ is defined as 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 a 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 $\boldsymbol{\alpha}$. Thus, we implement the following iterative algorithm, proposed by Huang et al. (2009), to minimize (2.5):

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

The motivation for this algorithm is a reparametrization of the nonconvex optimization problem into a complex optimization problem in terms of (θ, τ), which reaches an equivalent solution. In essence, the suggested algorithm performs an iteratively reweighted LASSO until convergence. 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 nonzero estimators of the coefficients α_{ij}. Then from 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 a sub-matrix of $\boldsymbol{U}$, with each column corresponding to the selected α_{ij}, and $\boldsymbol{\Theta}_S$ is a diagonal matrix,

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

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, Li and Huang (2008), the asymptotic variance of $\widehat{\boldsymbol{\alpha}}_S$ may be expressed as

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

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

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

where $\mathcal{B}_{iS}(t)$ is a 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; thus, asymptotic variances of $\widehat{\boldsymbol{\alpha}}_S$ and $\widehat{\beta}_i(t)$ are also tuning-parameter dependent. Although this is a naive estimator, as shown in our numerical studies, its approximation is nevertheless found to be effective in capturing the level of variability. Alternatively, we can estimate the full covariance function nonparametrically. However, owing to the additional complexity in the implementation with irregular design points, this is not very practical. The literature takes a more pragmatic approach of using a random-effects formulation (e.g., Wu and Zhang (2006)). However, the

difficulty of selecting the covariates in the random-effects terms under the current context of sparse function estimation outweighs the potential benefits; thus, we do not pursue this. Instead, we investigated using a fully nonparametric approach to estimate the covariance surface by means of a functional principal component analysis (Yao, Müller and Wang (2005)). However, our numerical study did not identify a clear advantage to this approach. Further investigation is left for future work.

2.5. Choice of tuning parameters

To fit the model using a finite sample, we consider how to calibrate the tuning parameters. The tuning parameter $\lambda > 0$ balances the trade-off between the goodness-of-fit and the 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 a small λ, we would select more variables and obtain better estimation results, but lose control of the functional sparsity. In the classical nonparametric approaches, criteria such as the AIC, BIC, and GCV (Wahba (1990)) are commonly used for model selection. Previous analyses noted that the AIC and GCV tend to select more variables, and thus are better suited to predictions. We use a BIC-type criterion in our analysis, reported in Section 4. To account for the increasing number of parameters when comparing models with varying dimensions, we use the extended BIC (EBIC) (Huang, Horowitz and Wei (2010)), which also penalizes the size of the full model. The EBIC is given by

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

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 nonzero estimates in $\widehat{\boldsymbol{\alpha}}(\lambda)$. Then, $\sum_{i=1}^{p} J_i = \sum_{i=1}^{p}(K_i+d+1)$ is the total number of parameters in the full model. Note that when $\nu = 0$, the EBIC is the same as the BIC, but when $\nu > 0$, the EBIC imposes greater penalty on overfitting. We use $\nu = 0.5$, as suggested in (Huang, Horowitz and Wei (2010)).

Note that the tuning parameter γ influences the performance of the group selection. A value of γ that is 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 functional sparsity in its solution. Unlike λ, however, $0 < \gamma < 1$ is more often viewed as a higher-level model parameter (often set equal to 0.5, (Huang et al. (2009))), similarly to how

the LASSO ($\gamma = 1$) estimator may be chosen over the Ridge ($\gamma = 2$) estimator in advance. Our theoretical results suggest that γ is closely related to λ, in the asymptotic sense, similarly to (Huang et al. (2009); Knight and Fu (2000)). Thus an adaptive selection of λ in a finite sample is expected to reflect this relation *automatically*. This is confirmed numerically (see Section 4), suggesting a value of $\gamma = 0.5$ as a rule of thumb.

In addition, because the parametric model formulation arises as an approximation to the nonparametric model, the parameter space to explore is not fixed, and is potentially very large. Even with known covariates, the spline approximation with the fully adaptive choice of the degree, knot locations and number of knots is impractical. Following a similar strategy to that in the literature (e.g., Huang, Wu and Zhou (2004); Wang, Li and Huang (2008)), 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. However, we found that the penalty was not effective in controlling the systematic increase in the parameter space and that the criterion favored the smallest possible K in the majority of cases. Instead, we select the number of knots K adaptively to the sample using 10-fold cross-validation without a penalty, leaving the potentially adaptive choice of sparsity to be controlled solely by the other tuning parameters.

3. Large-Sample Properties

We study the large-sample properties of our proposed PLSE $\widehat{\beta}_i(t)$, for $i = 1, \ldots, p$, when the number of sampled subjects n goes to infinity. In the proofs, we assume that the number of observations for each subject n_k is bounded. However, a similar argument can be applied to the case when n_k increases to infinity with n (Huang, Wu and Zhou (2004)). The number of interior knots increases with n; therefore, 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 (Huang, Wu and Zhou (2004); Wang, Li and Huang (2008)) are provided in the online Supplementary Material.

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

$$\mathcal{A}_{i1} = \left\{ g : \max_{t \in [\eta_{g-1}, \eta_g)} \mid \beta_i(t) \mid > C_i K_n^{-2} \right\},$$

$$\mathcal{A}_{i2} = \left\{ g : 0 \leq \max_{t \in [\eta_{g-1}, \eta_g)} \mid \beta_i(t) \mid \leq C_i K_n^{-2} \right\},$$

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

Theorem 6.27 of Schumaker (1981) shows that any smooth coefficient function $\beta_i(t)$ with a bounded second derivative has a B-spline approximant $\widetilde{\beta}_i(t)$ of the form given in (2.3) and that the approximation error is of order $O(K_{in}^{-2})$. We denote the sparse modification introduced in Wang and Kai (2015) by $\widetilde{\beta}_i^0(t)$, with coefficients $\widetilde{\boldsymbol{\alpha}}^0$.

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 the function space.

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

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

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

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

If (S1) *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 optimal (Stone (1982)). In fact, the result remains valid for more general classes of functions, for example, the collection of functions the derivatives of which satisfy 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 produces $\widehat{\boldsymbol{\alpha}}_{A_{il}} = \mathbf{0}$ to identify local sparsity with probability converging to one. In addition, $\beta_i(t) = 0$ for all t, then the proposed method yields $\widehat{\boldsymbol{\alpha}}_{A_{il}} = \mathbf{0}$, for all $l = 1, \ldots, K_i + 1$, with probability converging to one.

Theorem 2 (Sparsistency)**.** *Consider the following assumption:*

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

If (S2) *and the assumptions in Theorem* 1 *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 one as* n *goes to* ∞*.*

It is not surprising that our proposed method may yield a slightly more sparse functional estimate. This is because, 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. However, 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 (Zhao and Yu (2006)). However, for nonparametric models, particularly when local sparsity exists, selection consistency has not been widely studied. For the convenience of our discussion, we begin with some notation. For a coefficient function $\beta(t)$, let $N(\beta)$ and $S(\beta)$ denote the zero region and nonzero region, respectively. The (closed) support of β, denoted by $C(\beta)$, is defined as the closure of the nonzero region $S(\beta)$. Assume that $N(\beta)$ has finitely many singletons (as zero crossing), and that $C(\beta)$ 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 a result may not be of great interest, given that $\beta(t)$ lies in an infinite, not necessarily countable dimensional space. Next, consider a simple case of 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 $\mathcal{A}_{i2}$. The total length of the subintervals in $\mathcal{A}_{i2}$ converges to zero as n increases. For those intervals in $\mathcal{A}_{i1}$, a suitable choice of the constant C_i suggests 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 effects of the tuning parameters and the increasing dimension p on the 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 the increasing dimension p, given γ, assuming the number of relevant covariates, p_0, is set to four. In Scenario 3, we assess the performance with respect to K to study the effect of the adaptive choice of knots on the sparsity estimation. In addition, the relative performance is measured against that of the LSE and LASSO methods. The simulation results are summarized based on 400 replications. In each iteration, subjects are generated randomly 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 a constant–equal to one, $x_i(t)$, for $i = 2, 3, 4$, are similar to those considered in Huang, Wu and Zhou (2002): $x_2(t)$ 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 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 and a measurement error of $\epsilon_k(t) = \epsilon_k^{(1)}(t) + \epsilon_k^{(2)}(t)$. 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 is uncorrelated for different subjects. Then, $\epsilon_k^{(2)}(t)$ follows a normal distribution with mean zero and variance one, and is independent and identically distributed (i.i.d.).

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 a normal distribution with mean zero and variance $0.1\exp(t)$ to illustrate global sparsity. In Scenario 2, with increasing p, the extra predictors with zero coefficient functions are defined as $x_i(t) = Z_i(t) + 3/20\sum_{l=1}^{5} x_l(t)$, for $i = 6, \ldots, p$, where $Z_i(t)$ is i.i.d. from a standard normal distribution.

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

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

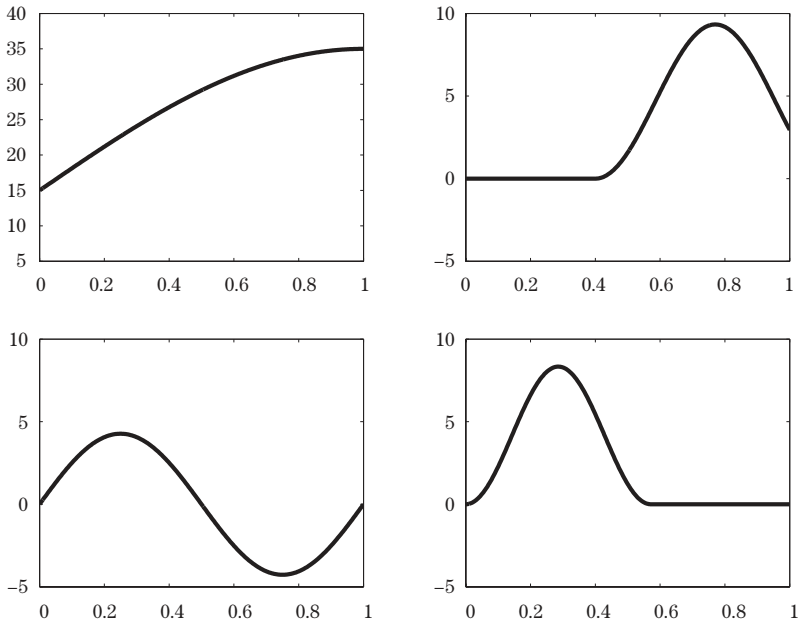

Figure 2. A graphical illustration of the coefficient functions β_i, for $i = 1, \ldots, 4$ (from left to right, top to bottom).

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

where $\widehat{\beta}_i^{(r)}$ is the estimated coefficient function from the rth repeated study. In addition, we use the following summary measures to compare the 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, whereas (c) and (d) summarize local sparsity.

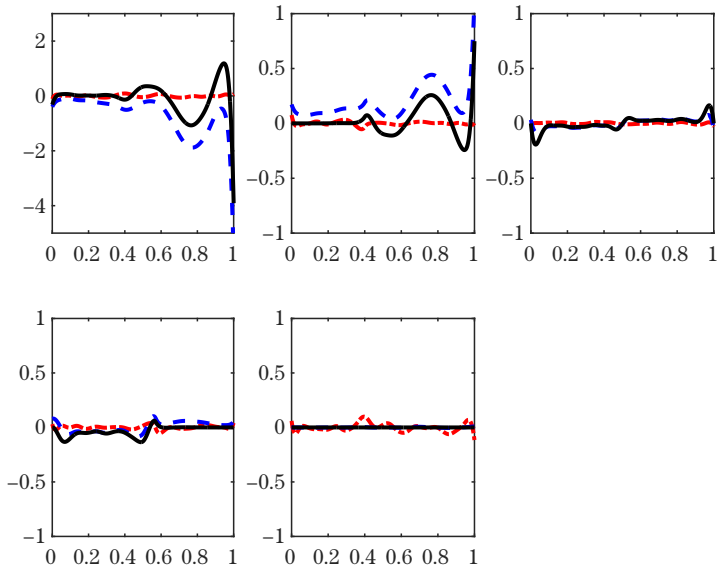

Figure 3. Comparison of the bias of the coefficient functions based on the LSE (dot-dashed), the LASSO (dashed), and $\text{PLSE}_{0.5}$ (solid) in Scenario 1, with $n = 200$. Note that $\text{PLSE}_{0.5}$ has zero bias in estimating the zero coefficient function $\beta_5(\cdot)$.

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 each coefficient function are summarized in Table 1. In general, as n increases, all methods yield decreasing MISE values. Notably, the results for the PLSE indicate comparable performance across different γ; in fact, the PLSE and LASSO methods show similar performance in terms of the function estimation. In addition, the PLSE with $\gamma = 0.35, 0.50$ successfully identifies the global sparsity of $\beta_5(\cdot)$ with zero MISE values for both choices of n, as does the PLSE with $\gamma = 0.75$ for $n = 200$. The bias of $\text{PLSE}_{0.5}$ (PLSE with $\gamma = 0.5$) and the LASSO and LSE methods with $n = 200$ is compared in Figure 3, showing that $\text{PLSE}_{0.5}$ has zero bias in estimating $\beta_5(\cdot)$.

Table 2 describes the performance when identifying local sparsity. 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 of the true model, the better. In contrast, the value of $I_{i,0}$ in the true model is the maximum error each method can make; thus, the smaller $I_{i,0}$, the better. In general, the LASSO

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

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

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

and PLSE methods show better performance in terms of functional sparsity. In addition, it can be seen that the PLSE with $\gamma = 0.35, 0.5, 0.75$ has an advantage in achieving both global and local sparsity, as compared with the LSE and LASSO methods. The case of $\gamma = 0.5$ performs slightly better than the others. For the

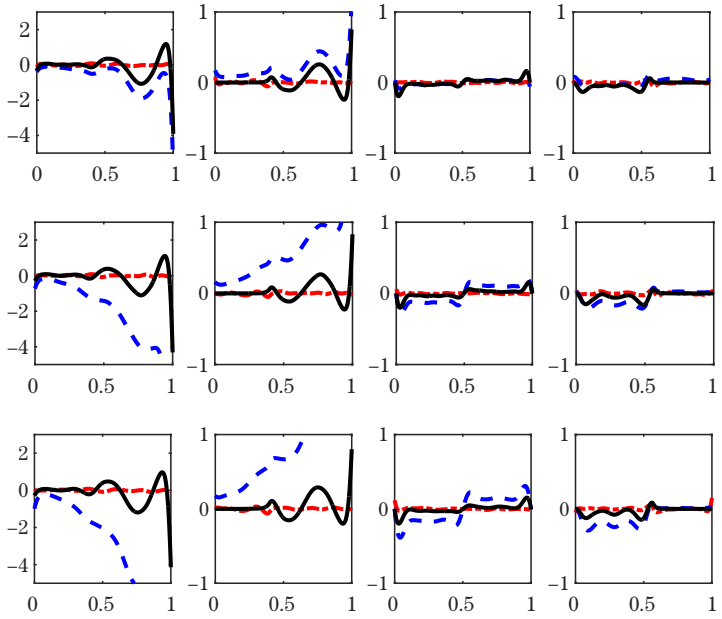

Figure 4. Comparison of the bias of the nonzero coefficient functions $\beta_1, \beta_2, \beta_3$, and β_4 (from left to right) based on the LSE (dot-dashed), LASSO (dashed), and $\text{PLSE}_{0.5}$ (solid), for $p = 5$ (top row), $p = 20$ (middle), and $p = 50$ (bottom) in Scenario 2.

remaining part, we use $\gamma = 0.5$ for the comparisons.

Scenario 2: Effect of dimension p

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

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

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

Method	MISE					$\max_{i\geq 6}$ 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	—
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
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
PLSE$_{0.5}$	0.7360	0.0396	0.0149	0.0205	0	2.7551e-05

Scenario 3: Effect of knots selection

The variation in the selection of the knots is expected to mainly influence the estimation of local sparsity. Increasing the number of knots helps to identify the boundary of local sparsity, but runs the risk of over-fitting nonzero estimates. Fine-tuning this parameter is much more delicate, because all model selection criteria are developed to control the squared error loss (MISE) as a goodness-of-fit, and thus are insensitive to the loss of missing local sparsity. That is, the balance between global and local sparsity is beyond the usual control of the bias-variance trade-off, and developing a new measure is still an open problem. Our knot selection based on cross-validation is essentially tuned toward 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 the results for fixed knots ($K = 11$) across the sample.

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 that in Scenario 2, with no major concern over the sensitivity of the knots selection in the comparison of the results.

In addition, in order to assess the usefulness of the asymptotic formula for the standard errors in (2.6), we calculate both the asymptotic and the empirical standard errors based on 400 repetitions. Then, we compare these standard errors in Figure 6 for an adaptive number of knots, and in Figure 7 for a fixed

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

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
$p = 5$														
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
$p = 20$														
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
$p = 50$														
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

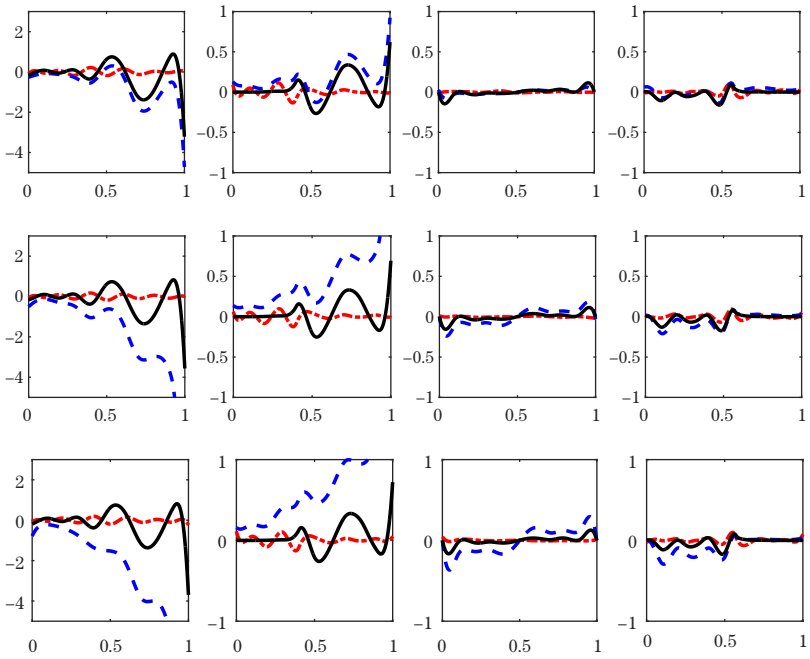

Figure 5. Comparison of the bias of the nonzero coefficient functions $\beta_1, \beta_2, \beta_3$, and β_4 (from left to right) based on the LSE (dot-dashed), LASSO (dashed), and $\text{PLSE}_{0.5}$ (solid), for $p = 5$ (top row), $p = 20$ (middle), and $p = 50$ (bottom) in Scenario 3.

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

Method	MISE					$\max_{i\geq 6} \text{MISE}_i$
	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	—
$\text{PLSE}_{0.5}$	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
$\text{PLSE}_{0.5}$	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
$\text{PLSE}_{0.5}$	0.7279	0.0417	0.0114	0.0150	0	1.9744e-05

number of knots, with the results showing good agreement. Thus, the variation in the number of knots greatly increases the variation in the estimation of the

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

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
$p = 5$														
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
$p = 20$														
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
$p = 50$														
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

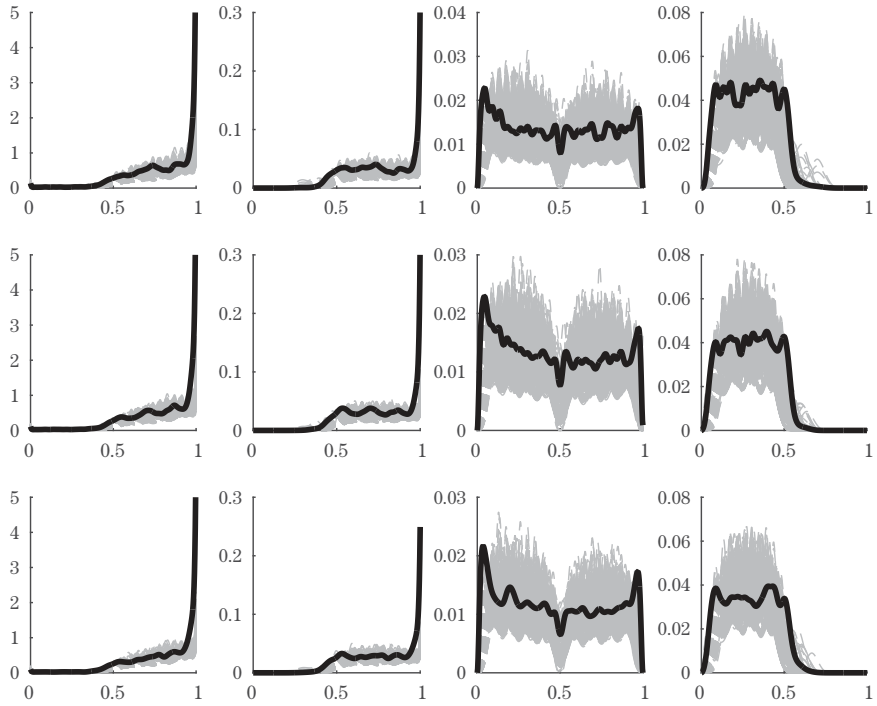

Figure 6. Asymptotic standard error (grey solid line) and empirical standard deviation (black solid line) of the coefficient functions with an adaptive number of knots in Scenario 2.

coefficient functions.

In summary, the simulation results demonstrate that our proposed method not only has an advantage in achieving local sparsity compared with the LASSO and LSE methods, but also ensures global sparsity for finite-dimensional models. Moreover, this advantage applies to models with an increasing dimension.

5. Real-Data Analysis

We demonstrate our method by analyzing yeast cell cycle gene expression data (Spellman et al. (1998); Lee et al. (2002)).

In the biological sciences, gene expression data are common. Scientists believe that transcription factors (TFs) might have an effect on a genome's cell cycle regulation. As a result, they have attempted to identify the 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 a gene expression and the physical binding of the TFs from chromatin immunoprecipitation (ChIP-chip)

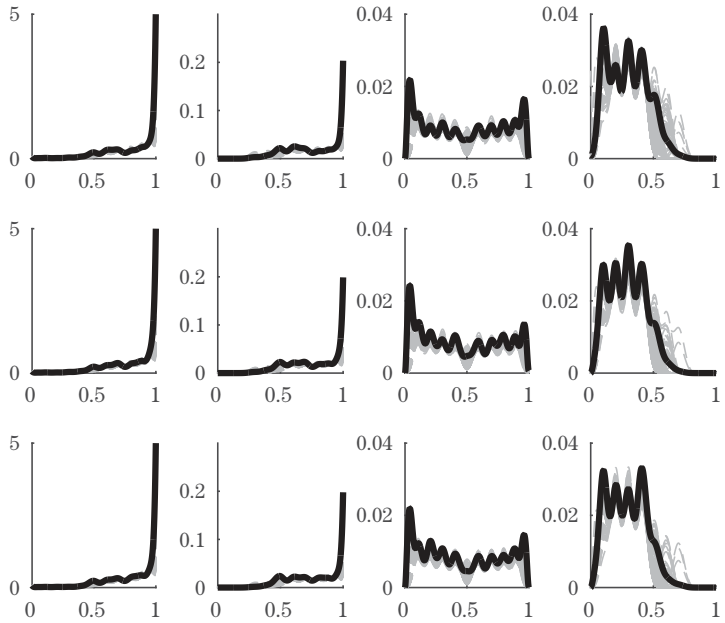

Figure 7. Asymptotic standard error (grey solid line) and empirical standard deviation (black solid line) of the coefficient functions with a fixed number of knots in Scenario 3.

data (Lee et al. (2002)). One set of gene expression data comes from an α-factor synchronization experiment of 542 genes, in which mRNA levels are measured every 7 minutes during a period of 119 minutes, resulting in 18 measurements in total (Spellman et al. (1998)). For our analysis, the time has been rescaled to [0, 1].

The ChIP-chip data contain the binding information of 106 TFs, of which 21 TFs are confirmed to be related to cell cycle regulation, by experiment. Wang, Chen and Li (2007) demonstrated that a variable selection procedure is able to identify some of the key TFs. It is believed that the effects of TFs vary during the cell cycle. In Chun and Keleş (2010), the authors considered a sparse partial least squares regression to study which TFs are important in a gene expression. However, they did not focus on the active periods of TFs. In this study, we apply our method to identify the key TFs, and to 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 can 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 TF i for gene k, for $i = 1, \ldots, 106$. Then, the varying coefficient model can be written as

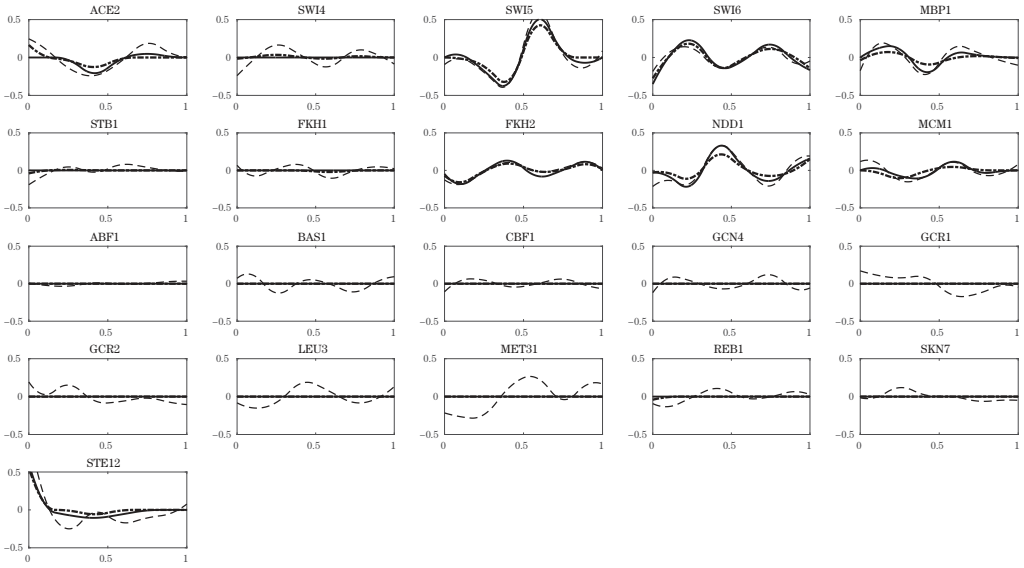

Figure 8. Subplots of estimated coefficient functions for the 21 confirmed TFs using the LSE (dashed), LASSO (dot-dashed), and $\text{PLSE}_{0.5}$ (solid).

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

where $\beta_i(t)$ models the effect of the ith TF on a gene expression at time t, and for the kth gene, ϵ_{kt} are independent over time.

Similarly to the simulation study, we apply our method together with LSE and LASSO methods, and compare the identification of the active period of each TF within the cell cycle process. Each coefficient function is approximated using quadratic B-splines defined on time interval [0, 1], with seven equally spaced knots. The number of knots is selected using cross-validation. It is not surprising that the LSE selects all TFs. The LASSO method identifies 32 TFs as important, while our proposed method identifies 16 TFs, which are a subset of those identified by the LASSO method. In Figure 8, the estimated coefficient functions for 21 experimentally confirmed TFs are shown. The figure shows that eight are selected by both methods. The LASSO method selects an additional four TFs, namely, SWI4, STB1, FKH1, and REB1, which show very low levels of activities. In Chun and Keleş (2010), 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 in the later period, and ACE2 is inactive in the early period.

Supplementary Material

The online Supplementary Material includes the technical assumptions and proofs of the theoretical properties of our proposed method.

Acknowledgments

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

References

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 (Statistical Methodology)* **72**, 3–25.

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

Diggle, P., Heagerty, P., Liang, K.-Y. and Zeger, S. (1994). *Analysis of Longitudinal Data*. OUP Oxford.

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

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

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.

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

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

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

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.

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

Kneip, A., Poß, D. and Sarda, P. (2016). Functional linear regression with points of impact. *The Annals of Statistics* **44**, 1–30.

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

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.

Müller, H.-G. (2016). Peter hall, functional data analysis and random objects. *The Annals of*

Statistics **44**, 1867–1887.

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

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

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.

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

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*, 477-497. Chapman and Hall/CRC.

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

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

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

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.

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

Wu, H. and Zhang, J.-T. (2006). *Nonparametric Regression Methods for Longitudinal Data Analysis: Mixed-effects Modelling Approaches*. Wiley.

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.

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.

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

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.

Department of Statistics, Colorado State University, Fort Collins, CO 80523, USA.

E-mail: catherine.tu@gmail.com

Department of Mathematics and Statistics, Fylde College, Lancaster University, Lancaster, LA1 4YF, U.K..

E-mail: juhyun.park@lancaster.ac.uk

Department of Statistics, Colorado State University, Fort Collins, CO 80523, USA.

E-mail: wanghn@stat.colostate.edu

(Received July 2015; accepted April 2018)