Statistica Sinica **32** (2022), 653-671
doi:https://doi.org/10.5705/ss.202019.0409

A FUNCTIONAL INFORMATION CRITERION FOR REGION SELECTION IN FUNCTIONAL LINEAR MODELS

Yunxiang Huang[1,2] and Qihua Wang[1,2,3]

[1]*Chinese Academy of Sciences,* [2]*University of Chinese Academy of Sciences and* [3]*Zhejiang Gongshang University*

Abstract: To deal with the region selection problem in functional linear models, we propose a functional information criterion that can be used to identify the null region where the functional predictor has no contribution to the response. The null region identified by our proposal is shown to be asymptotically consistent under some mild conditions. In addition, we obtain the convergence rate of the length of the null region estimate, which has not been considered previously. The procedure is easily implementable in practice. The finite-sample performance is illustrated in applications to simulated and real data.

Key words and phrases: functional data, information criterion, model selection, spline, support estimate, variable selection.

1. Introduction

The functional linear model (FLM) has been widely adopted to investigate the relationship between a scalar response Y and a functional predictor $X(t)$ defined on a compact set $[0, T]$. Let $\{(Y_i, X_i(t)), i = 1, \ldots, n\}$ be independent observations of $(Y, X(t))$. The FLM is formulated as

$$Y_i = a + \int_0^T X_i(t)\beta(t)dt + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1.1}$$

where a is the intercept, β is an unknown slope function defined on the domain $[0, T]$, and the regression error ε_i is independent and identically distributed (i.i.d.) and independent of X_i, with mean zero and finite variance σ^2. The main concern usually focuses on estimating β and investigating the asymptotic properties of the estimators; see Ramsay and Silverman (2005), Hsing and Eubank (2015), Wang, Chiou and Müller (2016), Reiss et al. (2017) and the references therein for an overview of functional data analysis. However, few studies have considered the problem of identifying the null region on which $X(t)$ does not contribute to

Corresponding author: Qihua Wang, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, People's Republic of China, 100190. E-mail: qhwang@amss.ac.cn.

Y. In an FLM, excluding $X(t)$ on the null region from the model reduces the prediction variance. On the other hand, identifying the null region, and hence the active region in which $\beta(t) \neq 0$ almost everywhere, benefits interpretability.

In the pioneering work by James, Wang and Zhu (2009), the authors used a simple grid basis to approximate β, and then used the Dantzig selector proposed by Candes and Tao (2007) to determine whether β and its first several derivatives are zero at some discrete grid points. However, as discussed by Zhou, Wang and Wang (2013), the Dantzig selector requires a large number of knots to precisely identify the null region of β. Furthermore, when the grid size is large, the Dantzig selector tends to overparameterize the model. To overcome this difficulty, Zhou, Wang and Wang (2013) proposed a two-stage estimator by introducing a refinement step after obtaining an initial estimator of the null region by using the Dantzig selector. In the refinement stage, the authors use the group smoothly clipped absolute deviation (SCAD) penalty proposed by Wang, Chen and Li (2007) on β, and apply a boundary grid-search algorithm to refine the selected null region and to estimate β on the active region. Lin et al. (2017b) introduced a functional version of the SCAD penalty proposed by Fan and Li (2001), and proposed a one-stage procedure that simultaneously identifies the null region of β and estimates β on the active region. Lin et al. (2017a) also proposed a group variable selection method based on the grouped Lasso of Yuan and Lin (2006) after clustering. However, they do not provide theoretical results. Note that the above approaches are based on L^1-regularization methods and require a careful selection of tuning parameters, which both increase the computational complexity and make the methods difficult to use. On the other hand, because the regularization methods simultaneously identify the null region of β and estimate β on the active region, they require some smoothness assumptions on β to ensure the asymptotic properties of the estimator of β. However, these smoothness assumptions are not necessary if we are only interested in identifying the null region. Hall and Hooker (2016) considered a special case in which β is active on $[0, \theta]$, with $\theta \leq T$. To implement their methods, one needs to reconstruct a parametric model to approximate β. However, the authors fail to explain how to do so. In addition, as mentioned by the authors, the performance of their methods depends on the number of functional principal components chosen in the model. Grollemund et al. (2019) proposed a Bayesian step function estimation of the support of the slope function. In practice, this approach is computationally costly when the sample size is large, which may limit its implementation.

We propose a functional information criterion, called FICf, to identify the null region in functional linear models. The FICf can be viewed as a functional gen-

eralization of the general information criterion (GIC) developed by Shao (1997) for classical linear models. In particular, we use B-spline basis functions to approximate β and reformulate the null region identification problem as a variable selection problem, in contrast to existing methods in the literature. The tuning parameters in our procedure are easy to determine, which makes our method simple to implement and statistically stable. We prove that the null region identified by our proposal is asymptotically consistent, regardless of whether the true underlying β is continuous at the boundaries of the active region. To the best of our knowledge, this is the first work to extend the information criterion to the FLM to deal with the null region identification problem. We also obtain the convergence rate of the length of the null region estimate, which has not been considered previously, under some quite general additional assumptions.

The rest of the paper is organized as follows. We introduce the FICf approach and its practical issues in Section 2. The asymptotic properties are given in Section 3. Simulation studies are discussed in Section 4, followed by an application to real data in Section 5. Section 6 concludes the paper. Sketches of the proofs, details about the simulation studies, and some additional simulations and another application are provided in the supplementary material.

2. Methodology

2.1. Spline approximation

For convenience, we first review spline approximations. For more details, see, for example, de Boor (2001) and Schumaker (2007). Let $0 = t_0 < t_1 < \cdots < t_P = T$ be $(P + 1)$ evenly spaced knots on $[0, T]$ and $I_k = [t_{k-1}, t_k]$, for $k = 1, \ldots, P$. The B-spline basis functions associated with the knots of order $d + 1$ consist of $(d + P)$ piecewise polynomials of degree d, denoted by $\boldsymbol{B}_{dP}(t) = (B_1, \ldots, B_{d+P})^T(t)$. The number of the adjacent subintervals I_k that compose the support of each B-spline basis function in $\boldsymbol{B}_{dP}(t)$ is no more than $d+1$. This property is called the compact support property, and is crucial to our approach.

Given $\boldsymbol{B}_{dP}$, the true underlying slope function β can be approximated by a linear combination $\beta_S(t) = \boldsymbol{B}_{dP}^T(t)\boldsymbol{b}$, where $\boldsymbol{b} = (b_1, \ldots, b_{d+P}) \in \mathbb{R}^{d+P}$ are coefficients. Additionally, b_j is zero if the corresponding basis function $B_j(t)$ lies entirely inside the null region. See Lemma 1 for the accuracy of the spline approximation. By using the spline approximation, we can rewrite (1.1) as

$$\boldsymbol{Y} = a + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\varepsilon} + \boldsymbol{\varepsilon_e},$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$, $\boldsymbol{Z}$ is an $n \times (P + d)$ matrix with entries

$$z_{ij} = \int_0^T X_i(t) B_j(t) dt, \quad i = 1, \ldots, n, \ j = 1, \ldots, d + P,$$

$\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$, and $\boldsymbol{\varepsilon_e}$ is an $n \times 1$ vector with entries $\varepsilon_{e,i} = \int_0^T X_i(t)[\beta(t) - \beta_S(t)]dt$. With this expression, if the approximation error $\boldsymbol{\varepsilon_e}$ is small, we can take advantage of the compact support property and obtain the null region by identifying the zero coefficients of $\boldsymbol{b}$. The above consideration motivates the proposed model selection procedure.

2.2. The FICf method

Let $\mathcal{M}_n$ be a set of candidate models, where each $M \in \mathcal{M}_n$ is a subset of $\{1, \ldots, d+P\}$. Denote by $\boldsymbol{b}(M)$ the sub-vector of $\boldsymbol{b}$ with components M, and by $\boldsymbol{Z}(M)$ the sub-matrix that consists of columns M of $\boldsymbol{Z}$. The model corresponding to M is $\boldsymbol{\mu}(M) = \mathrm{E}(\boldsymbol{Y} \mid \boldsymbol{Z}(M)) = a + \boldsymbol{Z}(M)\boldsymbol{b}(M)$. The proposed FICf method selects the model that minimizes

$$\mathrm{FICf}_{n,\,P}(M) = \frac{1}{n}[S_n(M)]^2 + \frac{1}{n}\hat{\sigma}^2 p_{P,\,n}(\dim(M)) \tag{2.1}$$

over $\mathcal{M}_n$, where $n^{-1}[S_n(M)]^2 = n^{-1}\|\boldsymbol{Y} - \hat{\boldsymbol{\mu}}(M)\|^2$ is the within-sample mean squared error, $\hat{\boldsymbol{\mu}}(M)$ is an estimate of $\boldsymbol{\mu}$ under model M, $\hat{\sigma}^2$ is an estimate of σ^2, and $p_{P,\,n}(\dim(M))$ is a model complexity penalty that depends on both the dimension of M, denoted by $\dim(M)$, and $(P,\, n)$, and increases $\mathrm{FICf}_{n,\,P}(M)$ for overfitted models. Simply speaking, $p_{P,\,n}(\dim(M))$ affords a balance between good fit and model complexity. The result of the above method depends on the estimate $\hat{\boldsymbol{\mu}}(M)$. Nevertheless, the least squares estimator has a high variability when P is relatively large. For this reason, we use the smooth spline estimator of Cardot, Ferraty and Sarda (2003); see Section 3.

We say that $A(\beta)$ is the active region of β if $\lambda^*\{t \in A(\beta) : \beta(t) = 0\} = 0$, where λ^* denotes the Lebesgue measure. Similarly, we call $N(\beta)$ the null region if $\beta(t) \equiv 0$ on $N(\beta)$. Let $\hat{M}$ be the selected model that minimizes the FICf in (2.1). It remains to define the estimates of the active region and the null region. Let $\mathrm{supp}(B_j) = (a_j,\, b_j)$ be the support of B_j, for $j \in \{1, \ldots, d + P\}$. Natural estimates of $A(\beta)$ and $N(\beta)$ are $A_1(\hat{M}) = \cup_{j \in \hat{M}} \mathrm{supp}(B_j)$ and $N_1(\hat{M}) = \cup_{j \in \hat{M}^c} \mathrm{supp}(B_j)$, respectively, where $\hat{M}^c = \{1, \ldots, d + P\} \setminus \hat{M}$. However, unless $\hat{M}$ or $\hat{M}^c$ is empty, there is an overlap between $\hat{A}_1(\beta)$ and $\hat{N}_1(\beta)$. To remove such ambiguity, we define the active region estimate $A(\hat{M})$ and the null region

estimate $N(\hat{M})$ as follows:

$$A(\hat{M}) = \bigcup_{\alpha \in \hat{M}} \left[\frac{b_{\alpha-1} + a_\alpha}{2}, \frac{b_\alpha + a_{\alpha+1}}{2} \right], \text{ and } N(\hat{M}) = A(\hat{M}^c). \tag{2.2}$$

Here, we set $b_0 = 0$ and $a_{d+P+1} = T$. It is not hard to verify that $A(\hat{M})$ and $N(\hat{M})$ are disjoint, except for a set of measure zero.

To summarize, we first select the model $\hat{M}$ that minimizes the FICf in (2.1). The estimates of the null region and the active region are given in (2.2). It remains to choose the candidate model set $\mathcal{M}_n$, choose the penalty function $p_{P,n}(\dim(M))$, estimate $\hat{\sigma}^2$, and estimate $\hat{\boldsymbol{\mu}}(M)$ for each candidate model $M \in \mathcal{M}$, which we discuss in Section 2.3.

2.3. Computation issues

The computation of $\hat{M}$ requires to traversing all candidate models in $\mathcal{M}_n$ and calculating the corresponding FICf. However, the computation cost is high if $\mathcal{M}_n$ comprises all the subsets of $\{1, \ldots, d + P\}$ when P is large. To deal with this problem, note that a model with fewer active intervals has better interpretability. Thus, we impose the following restriction on $\mathcal{M}_n$, which can be regarded as a minimal length restriction on both the active and the null intervals.

Each $M \in \mathcal{M}_n$ consists of several sequences of adjacent integers. The length of each integer sequence is at least d_A, and there are at least d_N integers between two integer sequences.

Here, d_A, $d_N > 0$ are predefined tuning parameters. We discuss how choosing d_A and d_N later in this section. This minimal length restriction also helps us derive the asymptotic properties in Section 3.

In summary, we have the following algorithm for obtaining $\hat{M}$:

Step 1 Given the B-spline basis functions $\boldsymbol{B}_{dP}(t)$, compute the matrix $\boldsymbol{Z}$ using (2.1). Centralize $\boldsymbol{Y}$ and each column of $\boldsymbol{Z}$. Compute $\hat{\sigma}^2$, the estimate of the regression error variance.

Step 2 Compute FICf(M) in (2.1) for each candidate model $M \in \mathcal{M}_n$.

Step 3 Return the model with the least FICf as $\hat{M}$.

In Step 1, we can use any existing method to compute $\hat{\sigma}^2$ (e.g., Chapter 9 in Ramsay, Hooker and Graves (2009)). In practice, we regress Y on the functional principal component scores, and use the mean squared error as $\hat{\sigma}^2$. In Step 2, one needs a closed-form expression of the penalty $p_{P,n}(\dim(M))$ in (2.1). We

suggest

$$p_{P,n}(\dim(M)) = n^{7/9}\left(\frac{\dim(M)}{P}\right), \tag{2.3}$$

which results from the simulations in the Supplementary Material. See also the discussion after assumption (A11.2) in Section 3 for some theoretical interpretation.

We now turn to computing $[S_n(M)]^2$ in (2.1) in Step 2. Given a model M, the smooth spline estimator of Cardot, Ferraty and Sarda (2003) is the minimizer of

$$Q_{\lambda_n}(a(M), \boldsymbol{b}(M)) = \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - a(M) - \boldsymbol{Z}_i(M)\boldsymbol{b}(M)\right]^2 + \lambda_n \|D^m[\boldsymbol{B}_{dP}(M)^T\boldsymbol{b}(M)]\|_2^2, \tag{2.4}$$

where $\boldsymbol{Z}_i(M)$ is the ith row of $\boldsymbol{Z}(M)$, $\|\cdot\|_2$ denotes the L^2 norm on $[0, T]$, D^m is the mth-order differential operator, and $\boldsymbol{B}_{dP}(M)$ consists of components M of $\boldsymbol{B}_{dP}$. The roughness tuning parameter λ_n varies with the sample size n and balances the squared loss and the roughness of $\beta(M, t) = \boldsymbol{B}_{dP}(M, t)^T b(M)$ quantified by $\|D^m\beta\|_2^2$. Writing $\bar{Y} = n^{-1}\sum_{i=1}^n Y_i$ and $\bar{\boldsymbol{Z}}(M) = n^{-1}\sum_{i=1}^n \boldsymbol{Z}_i(M)$, it is not hard to see that $\hat{a}(M) = \bar{Y} - \bar{\boldsymbol{Z}}(M)\hat{\boldsymbol{b}}(M)$. Substituting this into (2.4) yields

$$Q_{\lambda_n}(\boldsymbol{b}(M)) = \frac{1}{n}\sum_{i=1}^{n}\left[(Y_i - \bar{Y}) - (\boldsymbol{Z}_i(M) - \bar{\boldsymbol{Z}}(M))\boldsymbol{b}\right]^2 + \lambda_n\|D^m\boldsymbol{B}_{dP}(M)^T\boldsymbol{b}(M)\|_2^2. \tag{2.5}$$

For simplicity of notation, we assume $\bar{Y} = 0$ and $\bar{\boldsymbol{Z}} = 0$ in rest of this section, which can be satisfied by subtracting the sample mean $(\bar{Y}, \bar{\boldsymbol{Z}})$ from $(Y_i, \boldsymbol{Z}_i)$, for $i = 1, \ldots, n$. Let J_m be an $(d + P) \times (d + P)$ matrix with entries $(J_m)_{ij} = \int_0^T D^m B_i(t) D^m B_j(t) dt$. The second term on the right side of (2.5) can be expressed as

$$\lambda_n\|D^m[\boldsymbol{B}_{dP}(M)^T\boldsymbol{b}(M)]\|_2^2 = \lambda_n\boldsymbol{b}(M)^T\boldsymbol{J}_m(M)\boldsymbol{b}(M),$$

where $\boldsymbol{J}_m(M)$ is a sub-matrix of $\boldsymbol{J}_m$ formed from rows M and columns M. From this, letting $\|\cdot\|$ be the Euclidean norm of a vector, the loss function in (2.5) becomes

$$Q_{\lambda_n}(\boldsymbol{b}(M)) = \frac{1}{n}\|(\boldsymbol{Y} - \boldsymbol{Z}(M)\boldsymbol{b}(M))\|^2 + \lambda_n\boldsymbol{b}(M)^T\boldsymbol{J}_m(M)\boldsymbol{b}(M), \tag{2.6}$$

which is a generalized ridge regression loss function. Minimizing $Q_{\lambda_n}(\boldsymbol{b}(M))$ in (2.6) gives

$$\hat{\boldsymbol{b}}(M) = (\boldsymbol{Z}(M)^T\boldsymbol{Z}(M) + \lambda_n \boldsymbol{J}_m(M))^{-1}\boldsymbol{Z}(M)^T\boldsymbol{Y}.$$

Writing $\boldsymbol{H}_{\lambda_n}(M) = \boldsymbol{Z}(M)(\boldsymbol{Z}(M)^T\boldsymbol{Z}(M)+\lambda_n\boldsymbol{J}_m(M))^{-1}\boldsymbol{Z}(M)^T$, we have $[S_n(M)]^2 = n^{-1}\|\boldsymbol{Y} - \boldsymbol{H}_{\lambda_n}(M)\boldsymbol{Y}\|^2$.

To implement the above algorithm, one needs to choose the following tuning parameters: the number of knots P and the degree d of the B-spline basis functions, the order of the derivation m on the right side of (2.5), the regularization parameter λ_n on the right side of (2.5), and the minimal lengths d_A and d_N in the restriction on the candidate model set. In fact, only λ_n is a key tuning parameter.

As commonly adopted, we use the cubic B-spline basis functions with $d = 3$, and set $m = 2$ in most cases; see Cardot, Ferraty and Sarda (2003) and Chapter 5 in Ramsay and Silverman (2005). As discussed in Cardot, Ferraty and Sarda (2003), the value of P is not crucial in an FLM, because overfitting can be controlled using the roughness penalty. On the other hand, when the sample size n is fixed, the penalty function $p_{P,n}(\dim(M))$ in (2.3) depends only on $\dim(M)/P$, which is close to $\lambda^*(A(M))/T$. For these reasons, the FICf is not sensitive to P. In practice, P needs to be large enough to capture the character of β. In addition, d_A and d_N should be small, but appropriately large to avoid over-fitting. Note too that the number of candidate models, which decides the computational complexity, depends only on P, d_A, and d_N. Therefore, we suggest a rule of thumb of simply fixing a large P (usually from 30 to 100), and setting $d_A = d_N$ to be about one-eighth of P. One can also use cross-validation to search for optimal P, d_A, and d_N, if required, which can be computed in parallel.

It remains to determine the tuning parameter λ_n for each candidate model $M \in \mathcal{M}$. Here, we use the generalized cross-validation (GCV, Craven and Wahba (1978)) method to select λ_n, which minimizes

$$\mathrm{GCV}(\lambda_n;\, M) = \frac{n^{-1}[S_n(M)]^2}{(1 - \mathrm{tr}(\boldsymbol{H}_{\lambda_n}(M))/n)^2}.$$

See Gu (2013) for more details about GCV in an FLM. The GCV method is fast in terms of computation and performs well in our simulations.

In practice, the proposed estimating procedure is computationally costly for very large P. In order to ease the computational cost, we can traverse the candidate model set $\mathcal{M}_n$ using the number of active intervals. Then we stop when the minimal FICf among the models with k active intervals is less than that with

$(k+1)$ active intervals. There is no theoretical guarantee that this approach will converge to the model minimizing (2.1), but we can still select a reasonable model with fewer active intervals.

3. Theoretical Properties

We first present the asymptotic properties of the smoothing spline estimator of β under relatively weak smoothness conditions. We assume that

(A1) The true underlying slope function β is active on finite open intervals. In addition, β is Lipschitz continuous on each active interval.

The first part of assumption (A1) is quite general. We introduce this assumption to exclude certain pathological cases, such as

$$\beta(t) = t^2 \sin\left(\frac{1}{t}\right) \mathrm{I}\left[\sin\left(\frac{1}{t}\right) > 0\right], \quad t \in (0,\, 1),$$

where $\mathrm{I}[\cdot]$ is the indicator function. We do not require the number of active intervals to be known. The second part of assumption (A1) is a relatively weak smoothness assumption compared with that in Zhou, Wang and Wang (2013) and Lin et al. (2017b). The following lemma ensures the existence and accuracy of the spline approximation.

Lemma 1. *Under assumption* (A1)*, there exists a* $\beta_S(t) = \boldsymbol{B}_{dP}^T(t)\boldsymbol{b}_S$*, for* $t \in [0,\, T]$*, such that* $\|\beta - \beta_S\|_2^2 < c_1 P^{-1}$ *for some* $c_1 > 0$*, and* $b_{s,\,j} = 0$ *if* $B_j(t)$ *lies entirely inside the null region* $N(\beta)$*.*

In addition, we make the following assumptions:

(A2) Let λ_{min} and λ_{max} be the minimum and maximum eigenvalues, respectively, of $n^{-1}\boldsymbol{Z}^T\boldsymbol{Z}$. There are constants $0 < c_2 \le c_3 < \infty$ such that $c_2 P^{-1} \le \lambda_{min} \le \lambda_{max} \le c_3 P^{-1}$ holds in probability as $n \to \infty$.

(A3) $\mathrm{E}\,\|X\|_2^2 < \infty$.

(A4) $P = o(n^{1/2})$, $\lambda_n = o(P^{-2m-1})$.

Assumption (A2) coincides with the assumption at the beginning of Section 2 in Shao (1997), which ensures the existence of $(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}$, and is analogous to condition (A_8) in Zhou, Wang and Wang (2013) and condition (C4) in Lin et al. (2017b); see also Zhu, Fung and He (2008) for some sufficient conditions such that assumption (A2) holds. Assumption (A3) is required to apply the central limit theorem on $n^{-1}\sum_{i=1}^n X_i(t)$ in the proof of Theorem 1; see, for instance,

Chapter 1 in van der Vaart and Wellner (1996) for details. With appropriately chosen tuning parameters P and λ_n in assumption (A4), the bias caused by the roughness penalty depending on λ_n is dominated by the approximation error. The following result gives the convergence rates for $\hat{\beta}$.

Theorem 1. *Under assumptions* (A1)–(A4)*, there is a unique minimizer* $(\hat{\boldsymbol{b}}(M), \hat{a}(M))$ *for* (2.4) *for each candidate model. If* $b(M)$ *includes all nonzero* b_j*, then we have* $\|\hat{\boldsymbol{b}}(M) - \boldsymbol{b}_S(M)\| = O_p(1)$*,* $|\hat{a}(M) - a| = O_p(P^{-1/2})$*, and* $\|\hat{\beta}(M) - \beta\|_2 = O_p(P^{-1/2})$*.*

The main difference between Theorem 1 and Theorem 3.1 in Cardot, Ferraty and Sarda (2003) is that the slope function β is allowed to be nondifferentiable or discontinuous at finite points in Theorem 1, whereas the slope function is supposed to be sufficiently smooth on $[0, T]$ in Cardot, Ferraty and Sarda (2003). In this sense, Theorem 1 is a generalization of Theorem 3.1 in Cardot, Ferraty and Sarda (2003).

Next, we give the asymptotic consistency property of our FICf approach. We say that a selection criterion or a selection method is region selection consistent if the null region of the selected model $N(\hat{M})$ satisfies

$$\Delta\lambda^*_\beta(\hat{M}) := \lambda^*\Big\{N(\beta)\triangle N(\hat{M})\Big\} \xrightarrow{p} 0, \text{ as } n \to \infty, \tag{3.1}$$

where the operation symbol $\triangle$ denotes the symmetric difference of two sets.

Similarly to the minimal length restriction in Section 2.3, we put the following restrictions on the candidate model set $\mathcal{M}_n$. Only the minimum length restriction on the null intervals is required here.

(A5) Each $M \in \mathcal{M}_n$ consists of several sequences of adjacent integers, with at least $PL(\mathcal{M}_n)$ integers between two integer sequences, where $L(\mathcal{M}_n) > 0$ is a parameter.

Let $l(\beta)$ be the length of the shortest null interval between two active intervals for the true underlying β. Clearly, $l(\beta)$ is bounded away from zero under assumption (A1). The following assumption on $L(\mathcal{M}_n)$ is required.

(A6) $L(\mathcal{M}_n) < l(\beta)$ is a predefined constant not depending on (P, n).

Assumption (A6) implies that some a priori information on $l(\beta)$ is required when developing the theoretical results. In the case that ε is Gaussian, such information is not needed, and assumption (A6) is satisfied automatically for large enough (P, n) by allowing $L(\mathcal{M}_n)$ to go to zero.

(A6′) $L(\mathcal{M}_n) = o(1)$. $[PL(\mathcal{M}_n)]^{-1} = o(1)$.

The reason for this is that the asymptotic properties of the FICf method depend on the tail behavior of ε.

Before proceeding further, we introduce some useful notation. We write $f = \Omega(g)$ if $g = O(f)$, $f = \omega(g)$ if $g = o(f)$, and $f = \Theta(g)$ if both $f = O(g)$ and $f = \Omega(g)$ hold. The corresponding order in the probability notation Ω_p, ω_p, and Θ_p, are defined in a similar way. We assume the following conditions:

(A7) $\mathrm{E}\,|\varepsilon|^l < \infty$ holds, for $l = 4([T/L(\mathcal{M}_n)]+1)$, where $[\cdot]$ is the floor function.

(A8) β has at most finite zeros on each active interval.

(A9.1) $p_{P,n}(\dim(M))$ is strictly monotonically increasing with respect to $\dim(M)$.

(A9.2) As $(P, n) \to \infty$, $n^{-1}p_{P,n}(P) = o(1)$.

(A9.3) As $(P, n) \to \infty$, $n^{-1}P[p_{P,n}(\dim(M_2)) - p_{P,n}(\dim(M_1))] \to \infty$, for M_1, $M_2 \in \mathcal{M}$, such that $\dim(M_2) - \dim(M_1) = \Theta(P)$.

Assumption (A7) coincides with equation (2.6) in Shao (1997), which can also be viewed as a tail condition for ε. Assumption (A8), like assumption (A1), excludes certain pathological cases. Under assumption (A8), it is not hard to show that for a given $l > 0$, there exists a $C(l) > 0$, such that

$$\inf_{E \subset A(\beta),\, \lambda^*(E) \geq l} \left[\int_E (\beta(t))^2 dt \right] \geq C(l).$$

Assumptions (A9.1)–(A9.3) provide some restrictions on $p_{P,n}(\dim(M))$. Assumption (A9.1) is trivial. To illustrate assumptions (A9.2) and (A9.3), the penalty of the model complexity $p_{(P,n)}$ is required to be dominated by the fitting error for underfitted models, but heavy enough to avoid overfitting. The following theorem gives the region selection consistency property of the FICf method.

Theorem 2. *Under assumptions* (A1)–(A8) *and* (A9.1)–(A9.3)*, the FICf in* (2.1) *is region selection consistent, with* $(a(M), \boldsymbol{b}(M))$ *estimated by minimizing* (2.4)*. When* ε *is Gaussian, assumption* (A6) *can be replaced by* (A6′)*.*

Clearly, the numerical performance depends on the estimate $\hat{\sigma}^2$, but $\hat{\sigma}^2$ is not required be to a consistent estimate of σ^2 in Theorem 2. Indeed, assumption (A6) or (A6′) is sufficient, but not necessary for Theorem 2. The key is to introduce an appropriate restriction on $\mathcal{M}_n$ to control its cardinality. The minimal null length assumptions are compatible with the algorithm in Section 2.3, and crucial to deriving the convergence rate of $\Delta\lambda^*_\beta(\hat{M})$ in Theorem 3. Therefore, we simply use assumption (A6) or (A6′) as a condition in Theorem 2.

We now introduce some additional regular conditions in order to develop the convergence rate of $\Delta\lambda^*_\beta(\hat{M})$. Letting $0 \le u_1 < v_1 < u_2 < v_2 < \cdots < u_q < v_q \le T$, suppose that β is active on $A(\beta) = \cup_{k=1}^q (u_k, v_k)$ and vanishes on $N(\beta) = [0, T] \setminus A(\beta)$. Here, we do not require q to be known.

(A10) For all u_i and v_i, for $i = 1, \ldots, q$, there are constants c_5, $c_6 > 0$ and $p \in \{0\} \cup [1, \infty)$ such that $|\beta(u_j + t)| > c_5 t^p$ and $|\beta(v_j - t)| > c_5 t^p$, for any $t \in (0, c_6)$.

(A11.1) As $(P, n) \to \infty$, $n^{-1}P[p_{P,n}(\dim(M_2)) - p_{P,n}(\dim(M_1))] \to \infty$, for M_1, $M_2 \in \mathcal{M}$, such that $\dim(M_2) - \dim(M_1) = \omega(P^{2p/(2p+1)})$.

(A11.2) As $(P, n) \to \infty$, $n^{-1}P^{1-\delta}[p_{P,n}(\dim(M_2)) - p_{P,n}(\dim(M_1))] \to 0$, for M_1, $M_2 \in \mathcal{M}$, such that $\dim(M_2) - \dim(M_1) = O(P^{(2p+\delta)/(2p+1)})$, for any $\delta \in (0, 1)$.

The parameter p in assumption (A10) represents the smoothness behavior of the true underlying β at the boundaries of the active intervals, and $p = 0$ if β is discontinuous. Note that β is assumed to be Lipschitz continuous on each active interval in assumption (A1), which rules out the case that β is continuous on $[0, T]$, with $p \in (0, 1)$. Under assumption (A10), we can choose an appropriate penalty $p_{P,n}$ in assumptions (A11.1) and (A11.2), which are sharper versions of assumptions (A9.2) and (A9.3), respectively, to obtain the optimal convergence rate of $\Delta\lambda^*_\beta(\hat{M})$, depending on p. For example, suppose $p = 1$, which seems to be the most general case, and $P = \Theta(n^{1/3})$. By assumptions (A11.1) and (A11.2), we can set $p_{P,n} = n^{7/9}(\dim(M)/P)$. This interprets the rationality of the penalty function in (2.3). Another noticeable case is $p = 0$. In this case, we can set $p_{P,n} = n \dim(M)/(P \log P)$. In general, a heavy penalty is required for a small p.

In order to ensure assumption (A6) holds without prior knowledge of $l(\beta)$ when ε is Gaussian, we need to regularize the behavior of zeros of β (if they exist) on $A(\beta)$. Denote the zeros by $t_1, \ldots, t_J$. Suppose there are constants c_7, $c_8 > 0$ and $p' \in [1, \infty)$ such that $|\beta(t'_j + t)| > c_7|t|^{p'}$, for all $1 \le j \le J$ and $|t| < c_8$. We assume that

(A12.1) As $(P, n) \to \infty$, $L(\mathcal{M}_n)P^{1/(2p'+1)} = \omega(1)$; and

(A12.2) As $(P, n) \to \infty$, $n^{-1}L(\mathcal{M}_n)^{(1-\delta)/(2p'+1)}[p_{P,n}(\dim(M_2)) - p_{P,n}(\dim(M_1))] \to 0$, for M_1, $M_2 \in \mathcal{M}$, such that $\dim(M_2) - \dim(M_1) = O(PL(\mathcal{M}_n)^{1-\delta})$, for any $\delta \in (0, 1)$.

Note that the forecast bias caused by excluding an interval of length $L(\mathcal{M}_n)$ inside $A(\beta)$ depends on both $L(\mathcal{M}_n)$ and the behavior of the zeros of β. If $p' \leq p$, it can be shown that assumption (A12.2) trivially holds under assumptions (A11.2) and (A12.1). Otherwise, $L(\mathcal{M}_n)$ should go to zero slowly enough such that the forecast bias caused by excluding an interval of length $L(\mathcal{M}_n)$ inside $A(\beta)$ still dominates $p_{P,n}$. Theorem 3 gives the convergence rate of $\Delta\lambda^*_\beta(\hat{M})$.

Theorem 3. *Under assumptions* (A1)–(A9.1), (A10), (A11.1), *and* (A11.2), *it follows that* $\Delta\lambda^*_\beta(\hat{M}) = o_p(P^{(-1+\delta_1)/(2p+1)})$ *and* $\int_{N(\hat{M})\setminus N(\beta)}[\beta(t)]^2dt = o(P^{-1+\delta_1})$, *for any* $\delta_1 > 0$. *When* ε *is Gaussian, assumption* (A6) *can be replaced by assumption* (A6′) *if* β *takes no zeros on the interior of* $A(\beta)$ *or* $p' \leq p$. *In the case of* $p' > p$, *assumption* (A6) *be replaced by assumptions* (A12.1) *and* (A12.2).

In Theorem 3, the convergence rate of $\Delta\lambda^*_\beta(\hat{M})$ depends on p in assumption (A10). In particular, a larger p causes a slower convergence rate. This result is not surprising, because a large p implies that β changes slowly at the boundaries of the active intervals, which blurs the boundaries between the active intervals and the null intervals.

4. Simulation Studies

To evaluate the finite-sample performance of our FICf procedure, we conducted simulation studies on the FLM in (1.1) with $T = 1$ and $\varepsilon_i \sim N(0, \sigma^2_\varepsilon)$. We consider five types of true underlying slope functions β. Owing to space constrains, we report only two cases in this section. The results of the other cases and the finite-sample performance of different penalties $p_{P,n}(\dim(M))$ in (2.1) are reported in the online Supplementary Material. The slope function β of the first case is the same as that used in Lin et al. (2017b).

Case I:

$$\beta(t) = \begin{cases} 2(1-t)\sin(2\pi(t+0.2)), & 0 \leq t \leq 0.3, \\ 0, & 0.3 < t \leq 0.7, \\ 2t\sin(2\pi(t-0.2)), & 0.7 < t \leq 1. \end{cases} \tag{4.1}$$

Note that β is smooth on $[0, 0.3) \cup (0.7, 1]$, vanishes on $[0.3, 0.7]$, and is nondifferentiable at $\{0.3, 0.7\}$.

Case II:

$$\beta(t) = \begin{cases} 2 - 8|t - 0.15|, & 0 \le t < 0.3, \\ 0, & 0.3 \le t \le 0.7, \\ -(2 - 8|t - 0.85|), & 0.7 < t \le 1. \end{cases} \tag{4.2}$$

The slope function β is not differentiable everywhere on the interior of the active region, which violates the smoothness assumptions in James, Wang and Zhu (2009), Zhou, Wang and Wang (2013), and Lin et al. (2017b). In addition, β is discontinuous at the boundaries of the null interval.

The predictor functions $\{X_i(t), i = 1, \ldots, n\}$ are generated from a linear combination of B-spline basis functions; that is, $X_i(t) = \sum_j x_{ij} B_j(t)$. The coefficients x_{ij} are generated from the standard normal distribution, and the B-spline basis functions are defined by 71 evenly spaced knots with order 5. The error term ε is Gaussian in the two cases, and its variance σ_ε^2 is fixed such that the signal-to-noise ratio $\mathrm{Var}[\int_0^T X(t)\beta(t)dt]/\sigma_\varepsilon^2 = 4$. We consider three sample sizes, $n = 150$, 450, 1000, and replicate 200 times for each case and sample size.

We compare our FICf method with competing methods, including the FLiRTI method of James, Wang and Zhu (2009), the two-stage method of Zhou, Wang and Wang (2013), the smooth and locally sparse method of Lin et al. (2017b), the Bayesian functional linear regression with sparse step functions (Bliss) method of Grollemund et al. (2019), and the FICf_0 method, which is similar to the FICf method, except $\boldsymbol{b}$ is estimated using the least squares estimator without the roughness penalty. The results for the ordered homogeneity pursuit Lasso method of Lin et al. (2017a) are not reported, because this approach performs badly in our simulations. For the FICf and FICf_0 methods, we use the smoothing spline estimator to estimate β and a for the selected model; these are estimated using the corresponding methods for the competing methods. Owing to the computational cost, we report the results for the Bliss method with a sample size of 150 only.

The performance of the region selection is measured by the length of the symmetric difference $\Delta\lambda_\beta^*(\hat{M})$, defined in (3.1). The summary of $\Delta\lambda_\beta^*(\hat{M})$ given in Table 1 suggests that the FICf method outperforms the other methods in terms of region selection. Note that the FICf method performs consistently better than the FICf_0, especially in the case of $n = 150$, which suggests that the roughness penalty plays as an important role in our FICf method. The proposed region selection procedure also improves both the estimation accuracy and the prediction accuracy in our simulations. See the Supplementary Material for further details.

Table 1. Simulation results of the length of the symmetric difference $\Delta\lambda^*_\beta(\hat{M})$ in (3.1) for Cases I and II. Each entry is the Monte Carlo average of 200 simulation replicates. The corresponding standard deviation is reported in parentheses. All values are multiplied by 100. FLiRTI: the method of James, Wang and Zhu (2009); Two-stage: the two-stage method of Zhou, Wang and Wang (2013); SLoS: the smooth and locally sparse method of Lin et al. (2017b); Bliss: the Bayesian functional linear regression with sparse step functions method of Grollemund et al. (2019); FICf: the proposed function information criterion method; FICf_0: similar to FICf, but not using the roughness penalty in the region selection.

	FLiRTI	Two-stage	SLoS	Bliss*	FICf_0	FICf
Case I						
$n = 150$	35.0(4.47)	11.5 (10.7)	12.2 (7.59)	6.22(2.85)	28.9 (9.23)	4.93(4.25)
$n = 450$	31.6(6.43)	9.54(9.26)	7.96(4.07)	–	13.6(7.55)	2.86(1.45)
$n = 1{,}000$	30.9(6.12)	6.79(7.96)	8.19(1.95)	–	9.10(2.05)	2.66(1.32)
Case II						
$n = 150$	31.2(6.73)	18.6 (12.5)	13.6 (3.02)	4.89(4.86)	27.5 (8.43)	7.07(4.23)
$n = 450$	28.9(6.64)	13.4 (9.73)	13.3 (2.93)	–	10.4 (6.56)	3.81(1.93)
$n = 1{,}000$	28.0(6.58)	12.1 (9.19)	12.2 (1.90)	–	4.71(1.44)	2.35(1.11)

*We report only the results of the Bliss method with a sample size of 150, owing to the computational cost.

5. Application to Beer Data

The beer data consist of 60 samples published by Nørgaard et al. (2000). A curve of near-infrared light absorbance from 1,100 to 2,250 nm, in steps of 2 nm, was measured for each sample. At the same time, the original extract concentration was recorded in degrees Plato. The main interest here is to predict the original extract concentration from the spectra curve. The original extract concentration is highly positively correlated with the alcohol percentage of beer, which serves as an important quality parameter in the brewery industry.

Figures 1a and 1b illustrate the spectra curves and the centralized spectra curves, respectively, for 10 randomly selected beer samples. By a priori visual inspection, it seems that the region larger than 1,400 nm is very noisy. As discussed in Nørgaard et al. (2000), this region is correlated with the O–H bond vibration of water, which almost inundates other signals.

Figure 1c (blue solid line) shows the smoothing spline estimate of β on $(1140, 1480) \cup (2150, 2235)$ identified by our FICf method. It suggests that the spectra curve from 1,480 nm to 2,150 nm makes no contribution to the original extract concentration, which coincides with the visual inspection. The active interval (1140, 1480) is consistent with the results in Nørgaard et al. (2000) and

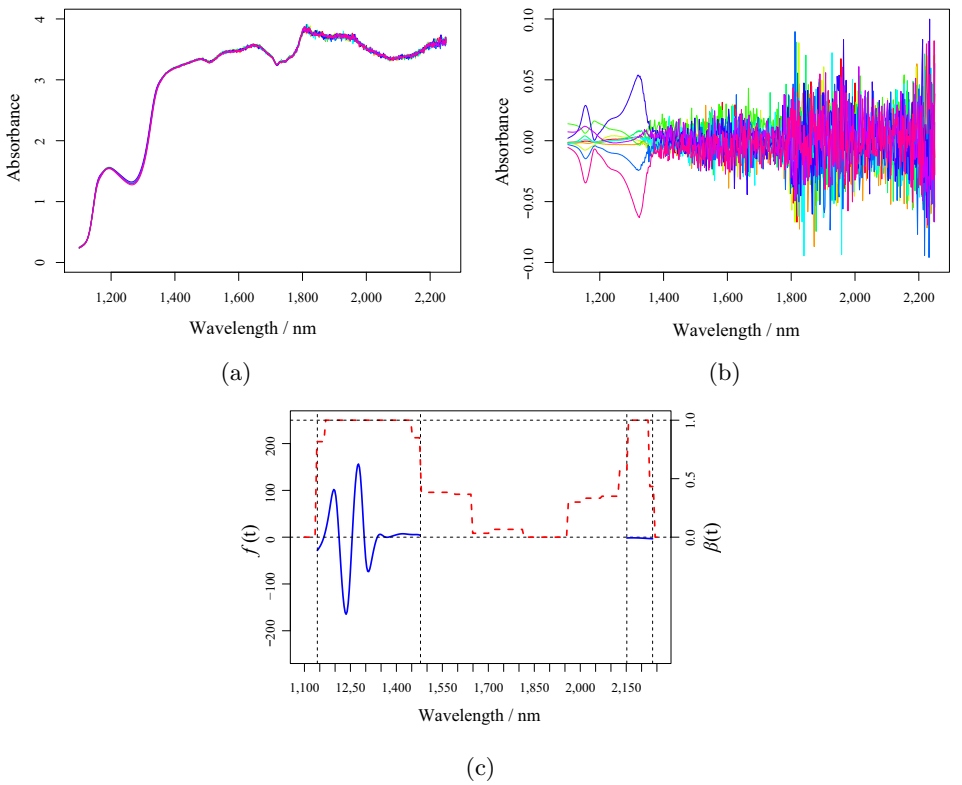

Figure 1. (a) (b) Spectra curves and centralized spectra curves, respectively of 10 random beer samples. (c) The smoothing spline estimate of the slope function β on the region selected using FICf (blue solid line) and the mean selecting frequency of the leave-one-out samples (red dashed line).

Lin et al. (2017a). As discussed in Nørgaard et al. (2000), this region is dominated by the C–H stretching overtone in organics. On the interval (2150, 2235), there is a very weak negative signal that results from the tones of the C–H bond on which the absorbance of water declines. See Smyth et al. (2008) for discussions in chemistry, and de Carvalho et al. (2016) for the near-infrared light spectra of water and alcohol.

The mean squared leave-one-out cross-validation errors of the FICf method and the competing methods are reported in Table 2. In this application, the FICf method has the lowest cross-validated error, followed by the two-stage method of Zhou, Wang and Wang (2013). The main difference between the latter two results is that the interval (2150, 2235) is excluded by the two-stage method. If we remove this interval from the model, the mean squared cross-validated error multiplied by 100 increases from 1.48 to 1.68, which implies that the negative

Table 2. The mean squared leave-one-out cross-validated errors of different methods for the beer data. The corresponding standard deviation is reported in parentheses. All values are multiplied by 100. Full: the smoothing estimate for the full model; OHPL: the ordered homogeneity pursuit LASSO method of Lin et al. (2017a); FLiRTI: the method of James, Wang and Zhu (2009); Two-stage: the two-stage method of Zhou, Wang and Wang (2013); SLoS: the smooth and locally sparse method of Lin et al. (2017b); Bliss-smooth: the smooth estimate of a Bayesian functional linear regression with sparse step functions of Grollemund et al. (2019); FICf: the smoothing spline estimate for the model selected using the proposed function information criterion method.

Full	OHPL	FLiRTI	Two-stage	SLoS	Bliss-smooth*	FICf
4.31(6.65)	3.79(5.37)	4.54(7.79)	1.55(1.95)	–†	3.98(7.27)	1.48(1.88)

*The mean and the standard deviation of the within-sample mean squared errors are reported instead for the Bliss-smooth method.
†The full model is selected.

relationship on that interval is informative in terms of predicting the original extract concentration.

To evaluate the reliability of the region selection procedure, we recommend a frequency-based measure. Let $\mathcal{X}_r$, for $r = 1, \ldots, R$, be R sets of the sample data, and $A(\hat{M}_r)$ be the active regions estimated from those sets. In this application, we use the leave-one-out samples for $\mathcal{X}_r$. We define the frequency of selection

$$f(t) = \frac{1}{R}\sum_{r=1}^{R} \mathrm{I}[t \in A(\hat{M}_r)],\ t \in [0,\, T].$$

Clearly, $f(t)$ has a value between zero and one. For a given t_0, a large $f(t_0)$ close to one indicates that t_0 is likely to belong to the active region, whereas a small $f(t_0)$ close to zero indicates that t_0 is likely to belong to the null region. Figure 1c (red dashed line) displays the frequency of selection, which confirms the correlation on the selected region. About 35% of the cross-validated selected regions contain the intervals $(1480, 1700)$ or $(1960, 2150)$. However, if we add either or both of these intervals in the model, the mean squared cross-validated error multiplied by 100 increases to 3.78, 3.41, and 4.25, respectively. Therefore, these intervals may be informative. However, this is beyond the scope of this study, and so is not explored further here.

6. Conclusion

Region selection in an FLM helps to reduce overfitting and improve interpretability. We have proposed an information criterion-based method called FICf

to identify the null region. To deal with the curse of dimensionality and the difficulty in calculation, we introduce a minimal length assumption in the algorithm that is also critical to developing the theoretical results. Note that we obtain a convergence rate for the symmetric difference between the null region of the true underlying slope function β and its estimate, which has not been investigated before.

Finally, although we have considered only the FLM in this study, the proposed information criterion-based method may be extended to the generalized FLM or nonlinear models. We assume that the functional data are fully observed in our approach. Further analysis is required for samples that are sparsely observed. In terms of model selection, one may also consider other problems of functional regression models, such as estimating the number of active intervals, selecting the shape of a slope functions, and selecting the points of impact of Kneip, Poss and Sarda (2016). These problems are more challenging for general functional data defined on higher-dimensional or even non-Euclidean domains, both theoretically and computationally.

Supplementary Material

The online Supplementary Material contains proofs of Lemma 1 and Theorems 1–3, as well as some details about the simulation studies, additional simulations, and another application.

Acknowledgments

The authors thank the co-editor, Dr. Hans-Georg Müller, and two anonymous referees for their constructive comments. Wang's research was supported by the National Natural Science Foundation of China (General program 11871460, Key program 11331011, and program for Innovative Research Group in China 61621003), and a grant from the Key Lab of Random Complex Structure and Data Science, CAS, China.

References

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

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

Craven, P. and Wahba, G. (1978). Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.* **31**, 377–403.

de Boor, C. (2001). *A Practical Guide to Splines*. Revised Edition. Springer-Verlag, New York.

de Carvalho, L. C., de Morais, C. D. L. M., de Lima, K. M. G., Júnior, L. C. C., Nascimento, P. A. M., de Faria, J. B. et al. (2016). Determination of the geographical origin and ethanol content of brazilian sugarcane spirit using near-infrared spectroscopy coupled with discriminant analysis. *Anal. Methods* **8**, 5658–5666.

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

Grollemund, P.-M., Abraham, C., Baragatti, M. and Pudlo, P. (2019). Bayesian functional linear regression with sparse step functions. *Bayesian Anal.* **14**, 111–135.

Gu, C. (2013). *Smoothing Spline ANOVA Models*. 2nd Edition. Springer, New York.

Hall, P. and Hooker, G. (2016). Truncated linear models for functional data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **78**, 637–653.

Hsing, T. and Eubank, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. John Wiley, Chichester.

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

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

Lin, Y.-W., Xiao, N., Wang, L.-L., Li, C.-Q. and Xu, Q.-S. (2017a). Ordered homogeneity pursuit lasso for group variable selection with applications to spectroscopic data. *Chemometr. Intell. Lab. Syst.* **168**, 62–71.

Lin, Z., Cao, J., Wang, L. and Wang, H. (2017b). Locally sparse estimator for functional linear regression models. *J. Comput. Graph. Statist.* **26**, 306–318.

Nørgaard, L., Saudland, A., Wagner, J., Nielsen, J. P., Munck, L. and Engelsen, S. B. (2000). Interval partial least-squares regression (*i*PLS): A comparative chemometric study with an example from near-infrared spectroscopy. *Appl. Spectrosc.* **54**, 413–419.

Ramsay, J. O., Hooker, G. and Graves, S. (2009). *Functional Data Analysis with R and MATLAB*. Use R! Springer, New York.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. 2nd Edition. Springer, New York.

Reiss, P. T., Goldsmith, J., Shang, H. L. and Ogden, R. T. (2017). Methods for scalar-on-function regression. *Int. Stat. Rev.* **85**, 228–249.

Schumaker, L. L. (2007). *Spline Functions: Basic Theory*. 3rd Edition. Cambridge University Press, Cambridge.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statist. Sinica* **7**, 221–264 with comments and a rejoinder by the author.

Smyth, H., Cozzolino, D., Cynkar, W., Dambergs, R., Sefton, M. and Gishen, M. (2008). Near infrared spectroscopy as a rapid tool to measure volatile aroma compounds in Riesling wine: Possibilities and limits. *Anal. Bioanal. Chem.* **390**, 1911–1916.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag, New York.

Wang, J.-L., Chiou, J.-M. and Müller, H.-G. (2016). Functional data analysis. *Annu. Rev. Stat. Appl.* **3**, 257–295.

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

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68**, 49–67.

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

Zhu, Z., Fung, W. K. and He, X. (2008). On the asymptotics of marginal regression splines with longitudinal data. *Biometrika* **95**, 907–917.

Yunxiang Huang

Academy of Mathematics and Systems Science, Chinese Academy of Sciences, No. 55, Zhongguancun E. Rd., Haidian District, Beijing, People's Republic of China, 100190.
University of Chinese Academy of Sciences, No.19(A) Yuquan Rd., Shijingshan District, Beijing, People's Republic of China, 100049.

E-mail: yxhuang@amss.ac.cn

Qihua Wang

Academy of Mathematics and Systems Science, Chinese Academy of Sciences, No. 55, Zhongguancun E. Rd., Haidian District, Beijing, People's Republic of China, 100190.
University of Chinese Academy of Sciences, No.19(A) Yuquan Rd., Shijingshan District, Beijing, People's Republic of China, 100049.
School of Statistics and Mathematics, Zhejiang Gongshang University Hangzhou, Zhejiang, People's Republic of China, 310018.

E-mail: qhwang@amss.ac.cn

(Received November 2019; accepted July 2020)