Title	A Functional Information Criterion for Region Selection in Functional Linear Models
Manuscript ID	SS-2019-0409
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202019.0409
Complete List of Authors	Yunxiang Huang and Qihua Wang
Corresponding Author	Qihua Wang
E-mail	qhwang@amss.ac.cn

A Functional Information Criterion for Region Selection in Functional Linear Models

Yunxiang Huang and Qihua Wang

Academy of Mathematics and Systems Science, Chinese Academy of Sciences

University of Chinese Academy of Sciences

Zhejiang Gongshang University

Abstract:

To deal with the region selection problem in functional linear models, a functional information criterion is suggested to identify the null region where the functional predictor has no contribution to the response. It is shown that the null region identified by our proposal is 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 before. 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, spline, support estimate, variable selection, model selection, information criterion

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 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 estimators. We refer the readers to Ramsay and Silverman (2005); Hsing and Eubank (2015); Wang et al. (2016); Reiss et al. (2017) and the references therein for an overview of functional data analysis. Much less work considers the problem of identifying the null region on which $X(t)$ has no contribution to Y. In 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 pioneer work by James et al. (2009), the authors used a simple grid basis to approximate β, and utilized 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 et al. (2013), the Dantzig selector requires a large number of knots to precisely identify the null region of β. But when the grid size is large, the Dantzig selector tends to overparameterize the model. To overcome the difficulty, Zhou et al. (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 employed a group smoothly clipped absolute deviation penalty proposed by Wang et al. (2007) on β and applied a boundary grid-search algorithm to refine the selected the null region and to estimate β on the active region. Lin et al. (2017) introduced a functional version of the smoothly clipped absolute deviation penalty proposed by Fan and Li (2001), and proposed a one-stage procedure which simultaneously identifies null region of β and estimates β on active region. Lin et al. (2017) proposed a group variable selection method by using grouped LASSO proposed by Yuan and Lin (2006) after clustering. But they did not give theoretical results. Note that the above approaches are based on L^1-regularization methods and require careful selection of tuning parameters, which both increase the computational complexity and make the methods difficult to use. On the other hand, since the regularization methods simultaneously identify the null region of β and estimate β on the active region, it requires some smoothness assumptions on β to ensure the asymptotic properties of the estimator of β. However, such smoothness assumptions seem not logically necessary if we are only interested in identifying the null region. Hall and Hooker (2016) considered a special case that β is active on $[0, \theta]$ with $\theta \leq T$. To implement their methods, one needs to reconstruct a parametric model to approximate β. Yet the authors did not give the details of how to reconstruct the parametric model. Also as mentioned by the authors, the performances of their methods depend on the number of the 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. This may limit the implement of the method proposed in Grollemund et al. (2019).

In this paper, 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 generalization of the general information criterion (GIC) developed by Shao (1997) in 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, which is different from those methods in literature. The tuning parameters in our procedure are easy to determine, which makes our method implementationally simple and statistically stable. We prove that the null region identified by our proposal is asymptotically consistent no matter whether or not the true underlying β is continuous at the boundaries of the active region. To the best of our knowledge, the current paper is the first work that extends information criterion to 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 before, under some quite general additional assumptions.

The rest of the article 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. In Section 6 we give some concluding remarks. The sketch of the proofs, as well as some details about the simulation studies, and some additional simulations and another application are delayed to the supplementary material.

2. Methodology

2.1 Spline Approximation

For the convenience to introduce our work, we first make a brief review of spline approximation. 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 with 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 which compose the support of each B-spline basis function in $\boldsymbol{B}_{dP}(t)$ is no more than $d+1$. Such property is called the compact support property, which is crucial for 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 $B_j(t)$ entirely lies inside the null region. See Lemma 1 for the accuracy of the spline approximation. By using the spline approximation, we can re-write (1.1) as

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

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

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

$\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^T$, and $\boldsymbol{\varepsilon_e}$ is a $n \times 1$ 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 through 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 $\{1, \dots, d+P\}$. Denote by $\boldsymbol{b}(M)$ the sub-vector of $\boldsymbol{b}$ with components M, and $\boldsymbol{Z}(M)$ 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

$$\mathrm{FICf}_{n,\,P}(M) = \frac{1}{n}[S_n(M)]^2 + \frac{1}{n}\hat{\sigma}^2 p_{P,\,n}(\dim(M)), \tag{2.2}$$

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 which 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 in Cardot et al. (2003). We shall give the details of the estimating procedure later in 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 (2.2). 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 $j \in \{1, \ldots, d+P\}$. Nature 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 $\hat{M}^c = \{1, \ldots, d+P\} \setminus \hat{M}$. Yet unless $\hat{M}$ or $\hat{M}^c$ is empty, there is a 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:

$$A(\hat{M}) = \bigcup_{\alpha\in\hat{M}} \left[\frac{b_{\alpha-1} + a_\alpha}{2}, \frac{b_\alpha + a_{\alpha+1}}{2}\right], \text{ and } N(\hat{M}) = A(\hat{M}^c). \tag{2.3}$$

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

2.3 Computation Issues

For the computation of $\hat{M}$, it requires to traverse all the candidate models in $\mathcal{M}_n$ and to calculate the corresponding FICf. However, the computation cost is high if $\mathcal{M}_n$ is made up of all the subsets of $\{1, \ldots, d + P\}$ when P is large. To deal with the problem, we note that a model with less active intervals has a better interpretability. Based on this consideration, we impose the following restriction on $\mathcal{M}_n$, which can be regarded as a minimal length restriction on both the active and 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 shall discuss how to choose d_A and d_N later in this section. Such minimal length restriction also helps us to derive the asymptotic properties in Section 3.

In summary, we have the following algorithm to obtain $\hat{M}$.

Step 1 Given the B-spline basis functions $\boldsymbol{B}_{dP}(t)$, compute matrix $\boldsymbol{Z}$ by 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 (2.2) 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 $\hat{\sigma}^2$ (See, for example, Chapter 9 in Ramsay et al. (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 (2.2). We suggest

$$p_{P,n}(\dim(M)) = n^{7/9}\left(\frac{\dim(M)}{P}\right), \tag{2.4}$$

which results from the simulations in the supplementary material. Also see the discussion after assumption (A11.2) in Section 3 for some theoretical interpretation.

We now turn to computing $[S_n(M)]^2$ in (2.2) in Step 2. Given a model M, the smooth spline estimator in Cardot et al. (2003) is the minimizer of

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

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, varying with the sample size n, 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 $\hat{a}(M) = \bar{Y} - \bar{\boldsymbol{Z}}(M)\hat{\boldsymbol{b}}(M)$. Plugging this into (2.5) yields

$$\begin{aligned} Q_{\lambda_n}(\boldsymbol{b}(M)) =& \frac{1}{n}\sum_{i=1}^n \left[(Y_i - \bar{Y}) - (\boldsymbol{Z}_i(M) - \bar{\boldsymbol{Z}}(M))\boldsymbol{b}\right]^2 \\ &+ \lambda_n \|D^m \boldsymbol{B}_{dP}(M)^T \boldsymbol{b}(M)\|_2^2. \end{aligned} \quad (2.6)$$

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 $(Y_i, \boldsymbol{Z}_i), i = 1, \ldots, n$. 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 (2.6) can be expressed as

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

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 (2.6) becomes

$$Q_{\lambda_n}(\boldsymbol{b}(M)) = \frac{1}{n}\|(\boldsymbol{Y} - \boldsymbol{Z}(M)\boldsymbol{b}(M))\|^2 + \lambda_n \boldsymbol{b}(M)^T \boldsymbol{J}_m(M)\boldsymbol{b}(M), \quad (2.7)$$

which is a generalized ridge regression loss function. Minimizing $Q_{\lambda_n}(\boldsymbol{b}(M))$ in (2.7) 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 derivation m on the right side of (2.6), the regularization parameter λ_n on the right side of (2.6), 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 et al. (2003) and Chapter 5 in Ramsay and Silverman (2005) for discussions. As discussed in Cardot et al. (2003), the value of P is not crucial in FLM since overfitting can be controlled by the roughness penalty. On the other hand, when the sample size n is fixed, the penalty function $p_{P,n}(\dim(M))$ in (2.4) only depends on $\dim(M)/P$ which is close to $\lambda^*(A(M))/T$. For these reasons, FICf is not sensitive to P. In practice, P is required to be large enough to capture the character of β, while d_A and d_N should be small. Yet appropriately large d_A and d_N help to avoid over-fitting. Also note that the number of the candidate models, which decides the computational complexity, only depends on P, d_A and d_N. For these considerations, 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 in practice. More sophisticatedly, one can 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) 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 FLM. The GCV method is fast in computation and performs well in our simulations.

In practice, the proposed estimating procedure is computational costly for very large P. In order to ease the computational cost, we may traverse the candidate model set $\mathcal{M}_n$ by the number of the active intervals and stop if the minimal FICf among the models with k active intervals is less than that with $(k+1)$ active intervals. There is no theoretical guarantee for this approach converging to the model minimizing (2.2), but we can still select a reasonable model with less 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. β is Lipschitz continuous on each active interval.

The first part of assumption (A1) is quite general. We introduce this assumption to excludes some pathological cases like

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

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

Lemma 1. *Under assumption (A1), there exist a* $\beta_S(t) = \boldsymbol{B}_{dP}^T(t)\boldsymbol{b}_S$, $t \in [0,\, T]$ *such that* $\|\beta - \beta_S\|_2^2 < c_1 P^{-1}$ *for some* $c_1 > 0$ *and* $b_{s,j} = 0$ *if* $B_j(t)$ *entirely lies 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 of $n^{-1}\boldsymbol{Z}^T\boldsymbol{Z}$, respectively. There are constants $0 < c_2 \leq c_3 < \infty$ such that $c_2 P^{-1} \leq \lambda_{min} \leq \lambda_{max} \leq c_3 P^{-1}$ 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 et al. (2013) and condition (C4) in Lin et al. (2017). Also see Zhu et al. (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 convergence rates of $\hat{\beta}$.

Theorem 1. *Under assumptions (A1) – (A4), there is a unique minimizer* $(\hat{\boldsymbol{b}}(M), \hat{a}(M))$ *for* (2.5) *for each candidate model. If* $b(M)$ *includes all non-zero* 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 of Theorem 1 and Theorem 3.1 in Cardot et al. (2003) is that the slope function β is allowed to be non-differentiable or discontinuous at finite points in Theorem 1, while the slope function is supposed to be sufficiently smooth on $[0, T]$ in Cardot et al. (2003). In this sense, Theorem 1 is a generalization of Theorem 3.1 in Cardot et al. (2003).

Next, we shall 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

$$\Delta\lambda_{\beta}^{*}(\hat{M}) := \lambda^{*}\{N(\beta) \bigtriangleup N(\hat{M})\} \xrightarrow{p} 0, \text{ as } n \to \infty, \tag{3.8}$$

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

Similar 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 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 priori information of $l(\beta)$ is required when developing the theoretical results. In the case that ε is Gaussian, such priori information is not needed and assumption (A6) can be 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 notations. 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 probability notations 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) is coincided with equation (2.6) in Shao (1997), which can also be viewed as a tail condition for ε. Assumption (A8), like assumption (A1), excludes some 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 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* (2.2) *is region selection consistent with* $(a(M), \boldsymbol{b}(M))$ *estimated by minimizing* (2.5). *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 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 the cardinality of $\mathcal{M}_n$. However, the minimal null length assumptions are compatible with the algorithm in Section 2.3. In the meantime, those assumptions are also crucial to derive the convergence rate of $\Delta\lambda^*_\beta(\hat{M})$ in Theorem 3. For these reasons, we simply use assumption (A6) or (A6') as a condition in Theorem 2.

We now introduce some additional regular conditions 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 vanished on $N(\beta) = [0, T] \setminus A(\beta)$. Here we do not require q is known.

(A10) For all u_i and v_i, $i = 1, \ldots, q$, 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 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 the appropriate penalty $p_{P,n}$ in assumptions (A11.1) and (A11.2), which are shaper versions of assumptions (A9.2) and (A9.3), 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 (2.4). 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 (if exist) of β on $A(\beta)$. Denote the zeros by $t_1, \ldots, t_J$. Suppose that 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 $1 \leq j \leq J$ 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 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* $\int_{N(\hat{M})\setminus N(\beta)}[\beta(t)]^2\, dt = o(P^{-1+\delta_1})$ *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). Especially, a larger p causes a slower convergence rate.

This result is not surprising since 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 $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$. We consider five different types of true underlying slope functions β. Due to the length of the paper, we only report 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 (2.2) are reported in the online supplement. The slope function β of the first case is same as that used in Lin et al. (2017).

Case I:

$$\beta(t) = \begin{cases} 2(1-t)\sin(2\pi(t+0.2)), & 0 \leq t \leq 0.3, \\ 0, & 0.3 < t \leq 0.7, \\ 2t\sin(2\pi(t-0.2)), & 0.7 < t \leq 1. \end{cases} \tag{4.9}$$

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

Case II:

$$\beta(t) = \begin{cases} 2 - 8|t - 0.15|, & 0 \leq t < 0.3, \\ 0, & 0.3 \leq t \leq 0.7, \\ -(2 - 8|t - 0.85|), & 0.7 < t \leq 1. \end{cases} \tag{4.10}$$

The slope function β is not differentiable everywhere on the interior of the active region, which violates the smoothness assumptions in James et al. (2009); Zhou et al. (2013); Lin et al. (2017). 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 is 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 in James et al. (2009), the two-stage method in Zhou et al. (2013), the smooth and locally sparse method in Lin et al. (2017), the Bayesian functional linear regression with sparse step functions (Bliss) method in Grollemund et al. (2019), and the FICf$_0$ method which is similar to FICf but $\boldsymbol{b}$ is estimated by the least squares estimator without the roughness penalty. The results of the ordered homogeneity pursuit LASSO method in Lin et al. (2017) are not reported since 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, while those are estimated by the corresponding methods for the competing methods. Due to the computational cost, we only report the results of the Bliss method with sample size of 150.

The performance of region selection is measured by the length of the symmetric difference $\Delta\lambda_{\beta}^{*}(\hat{M})$ defined in (3.8). The summary of $\Delta\lambda_{\beta}^{*}(\hat{M})$ given in Table 1 suggests that the FICf method outperforms other methods in region selection in terms of $\Delta\lambda_{\beta}^{*}(\hat{M})$. Note that FICf performs consistently better than 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 details.

(*Insert Table 1 here.*)

5. Application for Beer Data

The beer data consisted of 60 samples published by Nørgaard et al. (2000). A curve of near-infrared light absorbance from 1100 to 2250 nm in steps of 2 nm was measured for each sample. At the same time, the original extract concentration in plato was also recorded. 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 brewery industry.

Figures 1(a) and 1(b) illustrate the spectra curves and the centralized spectra curves for 10 randomly selected beer samples, respectively. By a priori visual inspection, it seems that the region larger than 1400 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.

(*Insert Figure 1 here.*)

Figure 1(c) (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 1480 nm to 2150 nm has no contribution on 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); Lin et al. (2017). 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 which 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 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 least cross-validated error, followed by the two-stage method in Zhou et al. (2013). The main difference between the results of FICf and the two-stage method 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 relationship on that interval is informative in prediction of the original extract concentration.

(*Insert Table 2 here.*)

To evaluate the reliability of the region selection procedure, we recommend a frequency-based measure. Let $\mathcal{X}_r$, $r = 1, \ldots, R$ 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 0 and 1. For a given t_0, a large $f(t_0)$ close to 1 indicates that t_0 is likely to belong to the active region, while a small $f(t_0)$ close to 0 indicates that t_0 is likely to belong to the null region. Figure 1(c) (red dashed line) displays the frequency of selection, which confirms the correlation on the selected region. There are about 35% of the cross-validated selected regions containing the intervals $(1480, 1700)$ or $(1960, 2150)$. However, if we add either or both of those intervals in the model, the mean squared cross-validated error multiplied by 100 increases to 3.78, 3.41 and 4.25, respectively. For this reason, those intervals may be informative but we will not explore that here.

6. Concluding Remarks

Region selection in FLM is significant to reduce overfitting and to 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, which is also critical in developing the theoretical results. A point that should be stressed is that we obtain a convergence rate of the symmetric difference between the null region of the true underlying slope function β and its estimate, which has not been investigated before.

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

Supplementary Materials Supplementary materials available at Statistica Sinica online contain the proofs of Lemma 1 and Theorems 1–3, as well as some details about the simulation studies, and some additional simulations and another application.

Acknowledgements 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 Creative Research Group in China 61621003), a grant from the Key Lab of Random Complex Structure and Data Science, CAS, China.

References

Candes, E. and T. Tao (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist. 35*(6), 2313–2351.

Cardot, H., F. Ferraty, and P. Sarda (2003). Spline estimators for the functional linear model. *Statist. Sinica 13*(3), 571–591.

Craven, P. and G. Wahba (1978). Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math. 31*(4), 377–403.

de Boor, C. (2001). *A practical guide to splines* (Revised ed.), Volume 27 of *Applied Mathematical Sciences*. Springer-Verlag, New York.

De Carvalho, L. C., C. D. L. M. De Morais, K. M. G. De Lima, L. C. C. Júnior, P. A. M. Nascimento, J. B. De Faria, and G. H. de Almeida Teixeira (2016). Determination of the geographical origin and ethanol content of brazilian sugarcane spirit using near-infrared spectroscopy coupled with discriminant analysis. *Anal. Methods 8*(28), 5658–5666.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc. 96*(456), 1348–1360.

Grollemund, P.-M., C. Abraham, M. Baragatti, and P. Pudlo (2019). Bayesian functional linear regression with sparse step functions. *Bayesian Anal. 14*(1), 111–135.

Gu, C. (2013). *Smoothing spline ANOVA models* (Second ed.), Volume 297 of *Springer Series in Statistics*. Springer, New York.

Hall, P. and G. Hooker (2016). Truncated linear models for functional data. *J. R. Stat. Soc. Ser. B. Stat. Methodol. 78*(3), 637–653.

Hsing, T. and R. Eubank (2015). *Theoretical foundations of functional data analysis, with an introduction to linear operators*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester.

James, G. M., J. Wang, and J. Zhu (2009). Functional linear regression that's interpretable. *Ann. Statist. 37*(5A), 2083–2108.

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

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

Lin, Z., J. Cao, L. Wang, and H. Wang (2017). Locally sparse estimator for functional linear regression models. *J. Comput. Graph. Statist. 26*(2), 306–318.

Nørgaard, L., A. Saudland, J. Wagner, J. P. Nielsen, L. Munck, and S. B. Engelsen (2000). Interval partial least-squares regression (i pls): A comparative chemometric study with an example from near-infrared spectroscopy. *Appl. Spectrosc. 54*(3), 413–419.

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

Ramsay, J. O. and B. W. Silverman (2005). *Functional data analysis* (Second ed.). Springer Series in Statistics. Springer, New York.

Reiss, P. T., J. Goldsmith, H. L. Shang, and R. T. Ogden (2017). Methods for scalar-on-function regression. *Int. Stat. Rev. 85*(2), 228–249.

Schumaker, L. L. (2007). *Spline functions: basic theory* (Third ed.). Cambridge Mathematical Library. Cambridge University Press, Cambridge.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statist. Sinica 7*(2), 221–264. With comments and a rejoinder by the author.

Smyth, H., D. Cozzolino, W. Cynkar, R. Dambergs, M. Sefton, and M. Gishen (2008). Near infrared spectroscopy as a rapid tool to measure volatile aroma compounds in riesling wine: possibilities and limits. *Anal. Bioanal. Chem. 390*(7), 1911–1916.

van der Vaart, A. W. and J. A. Wellner (1996). *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York. With applications to statistics.

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

Wang, L., G. Chen, and H. Li (2007). Group scad regression analysis for microarray time course gene expression data. *Bioinformatics 23*(12), 1486–1494.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol. 68*(1), 49–67.

Zhou, J., N.-Y. Wang, and N. Wang (2013). Functional linear model with zero-value coefficient function at sub-regions. *Statist. Sinica 23*(1), 25–50.

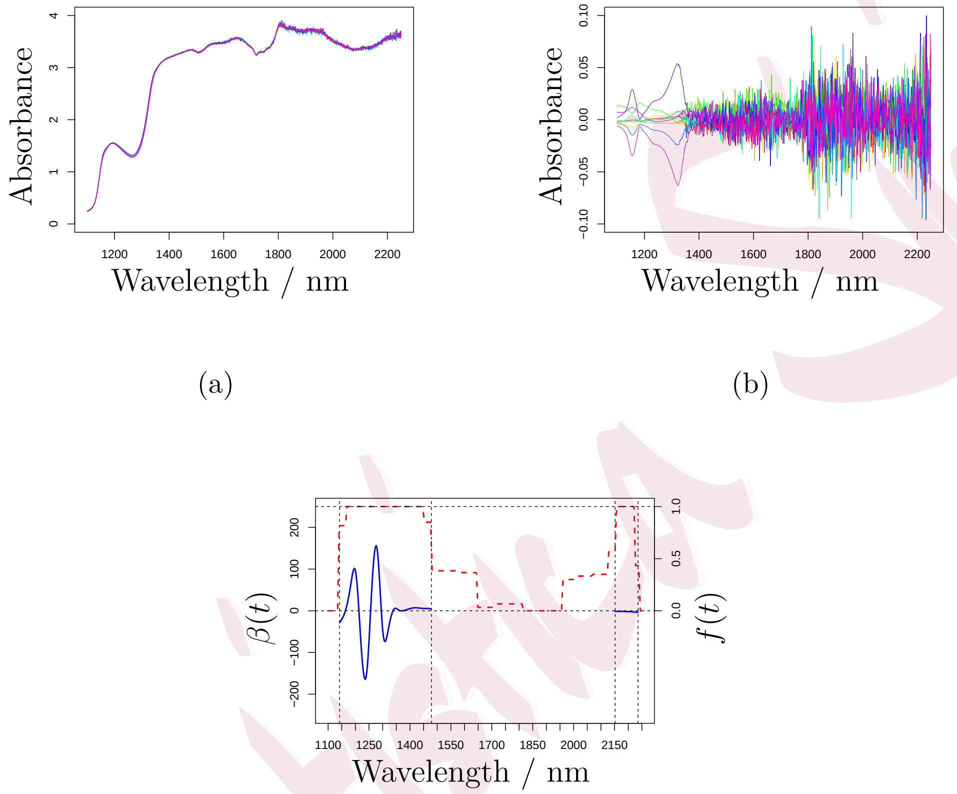
Zhu, Z., W. K. Fung, and X. He (2008). On the asymptotics of marginal regression splines with longitudinal data. *Biometrika 95*(4), 907–917.

Institute of Applied Mathematics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100080, People's Republic of China; University of Chinese Academy of Sciences, Beijing 100049, People's Republic of China;

E-mail: (yxhuang@amss.ac.cn)

Institute of Applied Mathematics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100080, People's Republic of China; University of Chinese Academy of Sciences, Beijing 100049, People's Republic of China; Zhejiang Gongshang University, Hangzhou, Zhejiang 310018, People's Republic of China;

E-mail: (qhwang@amss.ac.ac.cn)



(a)

(b)

(c)

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

Table 1: Simulation results of the length of the symmetric difference $\Delta\lambda^*_\beta(\hat{M})$ in (3.8) 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 the values are multiplied by 100. FLiRTI: the method in James et al. (2009); Two-stage: the two-stage method in Zhou et al. (2013); SLoS: the smooth and locally Sparse method in Lin et al. (2017); Bliss: the Bayesian functional Linear regression with Sparse Step functions method in Grollemund et al. (2019); FICf: the proposed function information criterion method; FICf$_0$: similar to FICf but not using the roughness penalty in 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)
$n = 1000$	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)
$n = 1000$	28.0(6.58)	12.1(9.19)	12.2(1.90)	–	4.71(1.44)	2.35(1.11)

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

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 the values are multiplied by 100. Full: the smoothing estimate for the full model; OHPL: the ordered homogeneity pursuit LASSO method in Lin et al. (2017); FLiRTI: the method in James et al. (2009); Two-stage: the two-stage method in Zhou et al. (2013); SLoS: the smooth and locally Sparse method in Lin et al. (2017); Bliss-smooth: the smooth estimate of Bayesian functional Linear regression with Sparse Step functions in Grollemund et al. (2019); FICf: the smoothing spline estimate for the model selected by 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 squared errors are reported instead for the Bliss-smooth method.

† The full model is selected.