

Consistent variable selection for functional regression models



Julian A.A. Collazos^a, Ronaldo Dias^a, Adriano Z. Zambom^{b,*}

^a Department of Statistics - State University of Campinas (UNICAMP), Rua Sergio Buarque de Holanda, 651, Distr. de Barao Geraldo, Campinas, Sao Paulo, Brazil

^b Department of Mathematics and Statistics - Loyola University Chicago, 1032 W. Sheridan Road, Chicago, IL 60660, United States

ARTICLE INFO

Article history:

Received 6 January 2015

Available online 19 June 2015

AMS subject classifications:

62G08

62G10

62G20

Keywords:

B-splines

Hypotheses testing

False discovery rate

Functional data

Likelihood ratio test

ABSTRACT

The dual problem of testing the predictive significance of a particular covariate, and identification of the set of relevant covariates is common in applied research and methodological investigations. To study this problem in the context of functional linear regression models with predictor variables observed over a grid and a scalar response, we consider basis expansions of the functional covariates and apply the likelihood ratio test. Based on p -values from testing each predictor, we propose a new variable selection method, which is consistent in selecting the relevant predictors from set of available predictors that is allowed to grow with the sample size n . Numerical simulations suggest that the proposed variable selection procedure outperforms existing methods found in the literature. A real dataset from weather stations in Japan is analyzed.

© 2016 Published by Elsevier Inc.

1. Introduction

In regression analysis, selecting the relevant set of predictors is a fundamental step for building a good predictive model. Including insignificant predictors results in over-complicated models with less predictive power and reduced ability to discern and interpret the influence of each variable. However, classical selection methods have to be adapted to the high-dimensional data sets which are becoming increasingly common in several areas of research.

When the data is observed at several time (or space) points, simple linear regression models cannot be directly used. Functional regression models (FRM) express the discrete observations of the predictor as a smooth function, and inference can then be made about a response variable based on the functional data [29]. Such models have become increasingly useful due to their large number of applications, see [18] for some fundamental results and Ferraty and Vieu [13] for a nonparametric approach. This high demand has recently leveraged important theoretical advances, see for example [19,14,20,12,3,16], to cite a few.

However, only a few authors have considered variable selection in functional regression analysis. Aneiros and Vieu [4] show how to perform variable selection using the continuous structure of the functional predictors by studying which of the discrete observed points should be incorporated. Using a partial linear model for multi-functional data, Aneiros and Vieu [5] propose a variable selection method based on the continuous specificity of the functional data. Cuevas [10, Section 5] presents an interesting overview of recent methods for functional data analysis including functional regression. Most recent contributions in regression for these models can be found in [7]. Another class of such methods uses regularization techniques, where the penalty simultaneously shrinks parameters and selects variables. Matsui and Konishi [24] studied the

* Corresponding author.

E-mail address: adriano.zambom@gmail.com (A.Z. Zambom).

group SCAD regularization for estimating and selecting functional regressors while Mingotti, Lillo and Romo [27] and Hong and Lian [17] generalized the Lasso for the case of scalar regressors and a functional response. Other recent contributions to the variable selection problem in functional models are Fan and Li [11], Aneiros, Ferraty, and Vieu [2], Gertheiss, Maity, and Staicu [15] and Ma, Song and Wang [23].

In this paper, we propose a different approach, exploiting the conceptual connection between model testing and variable selection: dropping a covariate from the model is equivalent to not rejecting the null hypothesis that its corresponding parameter(s) is equal to zero. Abramovich, Benjamini, Donoho and Johnstone [1] showed that the application of a false discovery rate (FDR) controlling procedure, such as Benjamini and Yekutieli [6], on p -values resulting from testing each null hypothesis can be translated into minimizing a model selection criterion. The extension and adaptation of the theory of hypothesis testing to functional models have been studied by several authors in the literature [9,33,32,22,25,28]. An interesting application can be found in [26], with results on the connection between p -values and variable selection in regression analysis.

The main objective of this paper is twofold: to study the asymptotic properties of the hypothesis test based on residual sum of squares for the relevance of a predictor in a multivariate functional regression model; and to propose a competitive variable selection procedure based on FDR (or Bonferroni) corrections applied on the p -values from the tests of each available functional predictor. The proposed test statistic is a likelihood ratio type test, where restricted and full models are estimated through the B-Splines basis expansions of both coefficients and functional predictors. We examine the shift (non-centrality parameter) of the distribution of the test statistic under the alternative hypothesis, which provides insight into the power of the test and induce the demonstration of consistency of the variable selection procedure.

The remainder of this paper is as follows. In Section 2, we formally describe the regression model with functional covariates and scalar response via basis expansions. In Section 3, we present the testing procedure and the variable selection method. In Section 4 we evaluate the finite sample performance of the proposed variable selection through simulation examples and a real application with weather data is considered in Section 5.

2. The functional regression model: FRM

Suppose that we have n observations $\{(y_i, \mathbf{x}_i(\mathbf{t})) : \mathbf{t} \in \mathcal{T}, i = 1, \dots, n\}$, where y_i is a scalar response, $\mathbf{x}_i(\mathbf{t}) = (x_{i1}(t_1), \dots, x_{iM}(t_M))$ are functional predictors and $\mathcal{T} = \mathcal{T}_1 \times \dots \times \mathcal{T}_M$. Each \mathcal{T}_m , $m = 1, \dots, M$, is a compact set in \mathbb{R} where the m th predictor may be observed. The functional predictors x_m , $m = 1, \dots, M$ are assumed to be in a fixed design so that in practice $t_m \in \mathcal{T}_m$ is a grid representing time or space. Suppose that each of the M functional predictors can be expressed as:

$$x_{im}(t_m) = \sum_{j=1}^{p_m} \omega_{imj} \phi_{mj}(t_m) = \mathbf{w}_{im}^T \boldsymbol{\phi}_m(t_m), \quad m = 1, \dots, M, \quad t_m \in \mathcal{T}_m, \quad (1)$$

where $\mathbf{w}_{im} = (\omega_{im1}, \dots, \omega_{imp_m})^T$ are the vectors of coefficients and $\boldsymbol{\phi}_m(t_m) = (\phi_{m1}(t_m), \dots, \phi_{mp_m}(t_m))^T$ are vectors of B-Splines basis functions. The basis functions and the p_m coefficients in (1) are assumed to be determined prior to the regression modeling through smoothing methods. In general this finite B-splines representation of a functional predictor is a good approximation of smooth functions, such as functions in the Sobolev Space (see [30]).

We consider the functional regression model [29] given by

$$y_i = \beta_0 + \sum_{m=1}^M \int_{\mathcal{T}_m} x_{im}(t_m) \beta_m(t_m) dt_m + \varepsilon_i, \quad (2)$$

where β_0 is a constant, ε_i , $i = 1, \dots, n$ are i.i.d. Gaussian noises with mean 0 and constant variance σ^2 , and $\beta_m(t_m)$ are functional coefficients that we assume can be represented through the basis expansion

$$\beta_m(t_m) = \sum_{j=1}^{p_m} b_{mj} \phi_{mj}(t_m) = \mathbf{b}_m^T \boldsymbol{\phi}_m(t_m), \quad m = 1, \dots, M, \quad t_m \in \mathcal{T}_m, \quad (3)$$

for the parameter vectors $\mathbf{b}_m = (b_{m1}, \dots, b_{mp_m})^T$. Thus the FRM in (2) can be re-expressed as a linear model in the following way

$$\begin{aligned} y_i &= \beta_0 + \sum_{m=1}^M \int_{\mathcal{T}_m} \mathbf{w}_{im}^T \boldsymbol{\phi}_m(t_m) \boldsymbol{\phi}_m^T(t_m) \mathbf{b}_m dt_m + \varepsilon_i = \beta_0 + \sum_{m=1}^M \mathbf{w}_{im}^T \int_{\mathcal{T}_m} \boldsymbol{\phi}_m(t_m) \boldsymbol{\phi}_m^T(t_m) dt_m \mathbf{b}_m + \varepsilon_i \\ &= \beta_0 + \sum_{m=1}^M \mathbf{w}_{im}^T \mathbf{J}_{\boldsymbol{\phi}_m} \mathbf{b}_m + \varepsilon_i = \mathbf{Z}_i^T \mathbf{b} + \varepsilon_i, \end{aligned}$$

or in matrix form $\mathbf{Y} = \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}$, where $\mathbf{Z}_i = (1, \mathbf{w}_{i1}^T \mathbf{J}_{\boldsymbol{\phi}_1}, \dots, \mathbf{w}_{iM}^T \mathbf{J}_{\boldsymbol{\phi}_M})^T$, $\mathbf{b} = (\beta_0, \mathbf{b}_1^T, \dots, \mathbf{b}_M^T)^T$, $\mathbf{Z} = (\mathbf{Z}_1, \dots, \mathbf{Z}_n)^T$, $\mathbf{J}_{\boldsymbol{\phi}_m} = \int_{\mathcal{T}_m} \boldsymbol{\phi}_m(t_m) \boldsymbol{\phi}_m^T(t_m) dt_m$ are $p_m \times p_m$ cross product matrices and $\boldsymbol{\epsilon}$ is the vector of error terms. Since we adopt B-splines basis expansions, the cross product matrix $\mathbf{J}_{\boldsymbol{\phi}_m}$ can be easily computed using the procedure in [21].

3. Methodology

3.1. Testing procedure

In this section we address the problem of testing the relevance of one individual functional predictor in the multivariate FRM. We consider testing the r th ($r \in \{1, \dots, M\}$) predictor through the following null hypothesis

$$H_0^r : \mathbf{b}_r = \mathbf{0} \quad \text{vs.} \quad H_a^r : \mathbf{b}_r \neq \mathbf{0}. \quad (4)$$

In linear models with normal errors, least squares estimates, which minimize the residual sum of squares, are equivalent to maximum likelihood estimates. For ease of notation, in this section, we omit from all statistics the index r that identifies the predictor being tested. Let ζ and Ω denote the spaces generated by the predictors under H_0 and H_a respectively. Note that $\zeta \subset \Omega$ and hence $\text{rank}(\Omega) = 1 + \sum_{m=1}^M p_m := k$ and $\text{rank}(\zeta) = k - p_r = 1 + \sum_{m=1}^M p_m - p_r := k_0$. We assume throughout this paper that the matrix \mathbf{Z} has full rank, that is, \mathbf{Z} has $k < n$ linearly independent columns (see also condition (C1) in Section 3.2). This assumption guarantees the existence and uniqueness of the least squares estimators. Let RSS_0 and RSS denote the residual sum of squares under H_0 and H_a respectively, that is,

$$RSS_0 = \sum_{i=1}^n (y_i - \mathbf{z}_i^T \hat{\mathbf{b}}^0)^2 \quad \text{and} \quad RSS = \sum_{i=1}^n (y_i - \mathbf{z}_i^T \hat{\mathbf{b}})^2, \quad (5)$$

where $\hat{\mathbf{b}}^0 = \hat{\mathbf{b}} - (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{A}^T (\mathbf{A} (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{A}^T)^{-1} \mathbf{A} \hat{\mathbf{b}}$ for a $p_r \times k$ matrix \mathbf{A} defining the null hypothesis, i.e., $\mathbf{A} \mathbf{b} = \mathbf{0}$ implies $\mathbf{b}_r = \mathbf{0}$.

For insight into the distribution of the test statistic and the non-centrality parameter presented below, it is useful to express the sum of squares RSS_0 and RSS as a quadratic form. We write $\hat{\mathbf{Y}}_0 = \mathbf{Z} \hat{\mathbf{b}}^0 = \mathbf{P}_0 \mathbf{Y}$ and $\hat{\mathbf{Y}} = \mathbf{Z} \hat{\mathbf{b}} = \mathbf{P} \mathbf{Y}$, where \mathbf{P}_0 and \mathbf{P} are the orthogonal projection matrices which project \mathbf{Y} onto the spaces ζ and Ω , respectively. We can then rewrite the residual sum of squares as $RSS_0 = \mathbf{Y}^T (\mathbf{I}_n - \mathbf{P}_0) \mathbf{Y}$ and $RSS = \mathbf{Y}^T (\mathbf{I}_n - \mathbf{P}) \mathbf{Y}$, so that $RSS_0 - RSS = \mathbf{Y}^T (\mathbf{P} - \mathbf{P}_0) \mathbf{Y}$. Since

$$\frac{RSS_0}{\sigma^2} \stackrel{H_0}{\sim} \chi_{n-k_0}^2 \quad \text{and} \quad \frac{RSS}{\sigma^2} \stackrel{H_0}{\sim} \chi_{n-k}^2,$$

in order to test H_0 in (4) we use the likelihood ratio statistic

$$T_L = -2 \ln \left[\frac{\tilde{L}_0}{\tilde{L}} \right] = -2 \left[-\frac{1}{2\tilde{\sigma}^2} RSS_0 + \frac{1}{2\tilde{\sigma}^2} RSS \right] = \frac{RSS_0 - RSS}{\tilde{\sigma}^2} \stackrel{H_0}{\underset{n \rightarrow \infty}{\rightarrow}} \chi_{k-k_0}^2 \quad (6)$$

in distribution, with $\tilde{\sigma}^2 = RSS/n \xrightarrow{P} \sigma^2$ the maximum likelihood ratio statistic. From the Normality assumption of the residuals and the fact that

$$\frac{1}{\sigma^2} E[RSS_0 - RSS] = \frac{1}{\sigma^2} [\sigma^2 \text{Tr}(\mathbf{P} - \mathbf{P}_0) + (\mathbf{Z} \mathbf{b})^T (\mathbf{P} - \mathbf{P}_0) \mathbf{Z} \mathbf{b}] = (k - k_0) + \delta = p_r + \delta,$$

where

$$\delta = \mathbf{b}^T \mathbf{Z}^T (\mathbf{P} - \mathbf{P}_0) \mathbf{Z} \mathbf{b} / \sigma^2, \quad (7)$$

the following proposition can be established.

Proposition 3.1 (Theorem 5.3c in [31]). Let RSS and RSS_0 be defined as in (5). Then, under the alternative hypothesis in (4)

$$\frac{RSS_0}{\sigma^2} \stackrel{H_0}{\sim} \chi_{n-k_0}^2(\delta) \quad \text{and} \quad \frac{RSS}{\sigma^2} \stackrel{H_0}{\sim} \chi_{n-k}^2, \quad \text{so that} \quad \frac{RSS_0 - RSS}{\sigma^2} \stackrel{H_0}{\sim} \chi_{k-k_0}^2(\delta).$$

Lemma 3.2 specifies the order of the non-centrality parameter of the distribution of $(RSS_0 - RSS)/\sigma^2$. Growing at the order of the sample size, multiplied by the significance size of the parameter being tested, the shift produced by the non-centrality parameter under H_a provides evidence for rejecting the null hypothesis. Using this result, Theorem 3.5 shows the consistency of the proposed variable selection procedure, which is described in Section 3.2.

Lemma 3.2. Let T_L be the likelihood ratio test statistic defined in (6) for testing H_0 in (4). For the alternative hypothesis, the non-centrality parameter δ defined in (7) is of order $\delta \sim c(n - k_0)$, for a constant c .

3.2. Consistent test based variable selection

In this section we describe a test-based variable selection method which is shown to consistently identify the set of relevant predictors. A similar procedure was used by Bunea, Wegkamp and Auguste [8] in the linear model setting, and by Zambom and Akritas [34] for a nonparametric model.

Let $I_M = \{1, \dots, M\}$ denote the set of indices of the M available functional predictors. Assume that the true underlying model is sparse in the sense that only a few predictors significantly relate to the response variable, while M is allowed to grow with n at a rate such that the following condition holds

$$\text{Condition (C1)} : k = 1 + \sum_{m=1}^M p_m \leq \sqrt{n}/\log(n).$$

Let $I_0 = \{m_1, \dots, m_{M_0}\}$ denote the (unknown) subset of indices corresponding to the M_0 significant predictors. The objective of the proposed variable selection method is to identify the subset I_0 , that is, to determine the set of functional variables with predictive significance.

Let T_L^r , $r = 1, \dots, M$, denote the likelihood test statistic defined in (6) for testing H_0^r in (4) and

$$\pi_r = 1 - \Psi(T_L^r) \quad (8)$$

the corresponding p -value, where $\Psi(\cdot)$ is the cumulative function of the $\chi_{p_r}^2$ distribution. The Bonferroni method yields $\hat{I} = \{m : \pi_m \leq q/M\}$ as the estimate of I_0 . The false discovery rate (FDR) procedure [6] computes

$$s = \max \left\{ j : \pi_{(j)} \leq \frac{j}{M} \frac{q}{\sum_{l=1}^M l^{-1}} \right\}, \quad (9)$$

where $\pi_{(1)} \leq \dots \leq \pi_{(M)}$ denote the ordered p -values and q is the choice of level, and rejects $H_0^{(j)}$, $j = 1, \dots, s$. If no such s exists, no hypothesis is rejected. The proposed variable selection method selects the predictors with indices corresponding to the s rejected null hypotheses. Hence, I_0 is estimated by the set \hat{I} of indices corresponding to the first s ordered p -values.

Let us now prove the consistency of the proposed variable selection method. Let R denote the total number of rejected hypothesis, so we have that $R = s\mathbb{1}$ (s in (9) exists), where $\mathbb{1}(\cdot)$ is the indicator function. Now, let V be the number of falsely rejected hypotheses, and set $Q = (V/R)\mathbb{1}(R > 0)$ for the proportion of falsely rejected hypotheses. By definition, the FDR is $E(Q)$, and $E(Q) \leq q(M - M_0)/M \leq q$, [6]. We consider consistent a procedure, and the estimated set \hat{I} , if $P(\hat{I} = I_0) \rightarrow 1$ as $n \rightarrow \infty$. Theorem 3.5, in connection with Lemmas 3.2–3.4, shows the consistency of \hat{I} .

Lemma 3.3. Let T_L^r and $\pi_r = 1 - \Psi(T_L^r)$ be the test statistic and the p -value defined as in (6) and (8) for testing H_0^r . Assume condition (C1) holds and define $A_n = \{|\hat{\sigma} - \sigma| \leq \sqrt{\log(n)/n}\}$.

- (a) For $r \notin I_0$ and any $0 < \gamma < 1$, we have $P(\{\pi_r \leq \gamma\} \cap A_n) = \gamma + O(\sqrt{\log(n)/n})$.
- (b) For $r \in I_0$ and $0 < \gamma < 1$, as $n \rightarrow \infty$, if $\gamma \geq 1/n$, we have $P(\{\pi_r > \gamma\} \cap A_n) = o(\gamma) + O(\sqrt{\log(n)/n})$.

Lemma 3.4. Let Γ_n be the event where the smallest M_0 p -values defined in (8) are the p -values corresponding to the M_0 significant functional predictors, with $I_0 = \{m_1, \dots, m_{M_0}\}$, that is

$$\Gamma_n = \left[\{\pi_{(1)}, \dots, \pi_{(M_0)}\} = \{\pi_{m_1}, \dots, \pi_{m_{M_0}}\} \right].$$

Then, if condition (C1) holds, $\lim_{n \rightarrow \infty} P(\Gamma_n) = 1$.

Theorem 3.5. Let δ be the non-centrality parameter defined in (7), and q the chosen bound of FDR in (9) or in Bonferroni corrections. Assume that condition (C1) holds and $q \rightarrow 0$ as $n \rightarrow \infty$, in such a way that $q \geq M \left(\sum_{l=1}^M l^{-1} \right) / (M_0 n)$ and $Mq/\log(M) \rightarrow 0$. Then, $\lim_{n \rightarrow \infty} P(\hat{I} = I_0) = 1$.

Note that the choice of $q \rightarrow 0$ is important for the consistency of the proposed method. For real datasets, a rule of thumb is to choose $q = O(1/M)$ if M is large relatively to the sample size n , otherwise choose $q = O(1/\sqrt{n})$. These choices guarantee the consistency of the variable selection while satisfying all assumptions and conditions. In the simulation study we explore different choices of this parameter.

4. Numerical simulations

Simulation studies were conducted to evaluate the finite sample performance of the proposed variable selection procedure. The Monte Carlo simulations in this section are based on 100 and 300 generated observations of six functional covariates and a scalar response $\{(x_{im}(t), y_i); t \in \tau_m, i = 1, \dots, n, m = 1, \dots, 6\}$, extending the simulation set up in [24] by including three extra functional predictors. We compared the performance of the proposed variable selection procedure with that of group SCAD and group LASSO proposed by Matsui and Konishi [24], and the Generalized Functional Linear Model

Table 1

Number of correctly selected models and AMSE.

c	n		T_L^{BC}			T_L^{FDR}			SCAD		LASSO		GFLM
			0.01	0.05	0.1	0.01	0.05	0.1	GCV	BIC	GCV	BIC	
0	100	Correct	88	79	65	87	74	58	82	82	80	83	77
		AMSE	(2.07)	(2.04)	(2.01)	(2.06)	(2.05)	(1.97)	(1.45)	(1.45)	(1.19)	(1.30)	(8.94)
	300	Correct	96	92	88	95	89	83	85	85	84	86	83
		AMSE	(1.93)	(1.98)	(1.89)	(1.92)	(1.97)	(1.91)	(1.31)	(1.31)	(1.04)	(1.16)	(8.51)
0.4	100	Correct	79	79	78	82	80	73	79	79	65	65	76
		AMSE	(2.61)	(2.98)	(2.77)	(2.88)	(3.01)	(2.82)	(5.60)	(5.60)	(5.67)	(5.70)	(11.37)
	300	Correct	96	94	90	95	92	88	83	83	71	80	84
		AMSE	(2.57)	(2.90)	(2.74)	(2.87)	(2.91)	(2.79)	(5.58)	(5.58)	(5.64)	(5.59)	(10.78)
0.8	100	Correct	83	81	80	83	81	79	83	83	72	74	83
		AMSE	(7.15)	(7.96)	(7.92)	(7.42)	(7.87)	(7.78)	(7.41)	(7.41)	(7.14)	(7.87)	(13.49)
	300	Correct	98	96	93	99	95	92	93	93	80	82	94
		AMSE	(7.08)	(7.10)	(7.01)	(7.09)	(7.11)	(7.14)	(7.27)	(7.27)	(7.17)	(7.32)	(12.05)

(GFLM) method in [15] with adaptive penalization. For comparison purposes, we used 6 basis functions for the estimation of the predictors and the functional parameters $\beta(\cdot)$ in all methods. First, we generated z_{im} corresponding to the predictor X_m in an equally spaced grid of 50 points in \mathcal{T}_m in the following way:

$$z_{im} = u_{im}(t_m) + \epsilon_{im}, \quad \epsilon_{im} \sim N(0, (0.025r_{x_{im}})^2),$$

where $r_{x_{im}} = \max_i(u_{im}(t_m)) - \min(u_{im}(t_m))$ and

$$\begin{aligned} u_{i1}(t) &= \cos(2\pi(t - a_1)) + a_2, & \mathcal{T}_1 &= [0, 1], & a_1 &\sim N(-4, 3^2), & a_2 &\sim N(7, 1.5^2), \\ u_{i2}(t) &= b_1 \sin(\pi t) + b_2, & \mathcal{T}_2 &= [0, \pi/3], & b_1 &\sim U(3, 7), & b_2 &\sim N(0, 1), \\ u_{i3}(t) &= c_1 t^3 + c_2 t^2 + c_3 t, & \mathcal{T}_3 &= [-1, 1], & c_1 &\sim N(-3, 1.2^2), & c_2 &\sim N(2, 0.5^2), & c_3 &\sim N(-2, 1), \\ u_{i4}(t) &= \sin(2(t - d_1)) + d_2 t, & \mathcal{T}_4 &= [0, \pi/3], & d_1 &\sim N(-2, 1), & d_2 &\sim N(3, 1.5^2), \\ u_{i5}(t) &= e_1 \cos(2t) + e_2 t, & \mathcal{T}_5 &= [-2, 1], & e_1 &\sim U(2, 7), & e_2 &\sim N(2, 0.4^2), \\ u_{i6}(t) &= f_1 e^{-t/3} + f_2 t + f_3, & \mathcal{T}_6 &= [-1, 1], & f_1 &\sim N(4, 2^2), & f_2 &\sim N(-3, 0.5^2), & f_3 &\sim N(1, 1). \end{aligned}$$

The scalar response y_i was generated as $y_i = g(\mathbf{u}_i) + \varepsilon_i$, where $g(\mathbf{u}_i) = \sum_{m=1}^6 \int_{\mathcal{T}_m} u_{im}(t) \beta_m(t) dt$, $\varepsilon_i \sim N(0, (0.05R_{y_i})^2)$ and $R_{y_i} = \max(g(\mathbf{u}_i)) - \min(g(\mathbf{u}_i))$. For a constant c taking values 0, 0.4 and 0.8, the coefficient functions $\beta_m(t)$ are given by

$$\beta_1(t) = \sin(t), \quad \beta_2(t) = \sin(2t), \quad \beta_3(t) = -ct^2, \quad \beta_4(t) = \sin(2t), \quad \beta_5(t) = c \sin(\pi t), \quad \beta_6(t) = 0.$$

Note that if $c = 0$ the true model specifies that only u_1 , u_2 and u_4 significantly relate to the response, corresponding to the predictors X_1 , X_2 and X_4 .

As the first step of our analysis, the random data z_{im} was converted into the functional data x_{im} using B-splines basis smoothing. For these data, we assumed the functional regression model

$$y_i = \sum_{m=1}^6 \int_{\mathcal{T}_m} x_{im}(t) \beta_m(t) dt + \varepsilon_i,$$

and applied the proposed variable selection method described in Section 3. With 100 Monte Carlo simulations, we computed the number of correctly selected models and the averages of the mean square errors (AMSE) for the proposed method with FDR and Bonferroni corrections, as well as for group LASSO, group SCAD and GFLM. The results in Table 1 suggest that when the sample size is relatively small ($n = 100$), all four methods seem to select the correct model about the same number of times, however as the sample size increases, the proposed variable selection procedure outperforms group SCAD, group LASSO and the GFLM. We note that restrictive choices of level for the tests tend to yield better results of the proposed method. For example, we observe that the choice of $q = 0.01$ delivers the highest number of correctly model selections. For $c = 0$ or $c = 0.8$, group SCAD and group LASSO have AMSE similar to that of the proposed procedure. However, for predictors included in the model with low significance ($c = 0.4$), the AMSE of group SCAD and group LASSO are about double the AMSE achieved by our procedure, while the GFLM delivers the highest AMSE in all models.

5. Real data example: Weather data

In this application, we consider weather data observed monthly at 79 weather stations in Japan. The dataset was obtained from <http://www.data.jma.go.jp/obd/stats/data/en/>, and includes monthly and annual total observations averaged from 1971 to 2000: monthly observed average temperatures (TEMP), average atmospheric pressure (PRESS), time of daylight (LIGHT), average humidity (HUMID), maximum temperature (MAX.TEMP), minimum temperature (MIN.TEMP) and annual

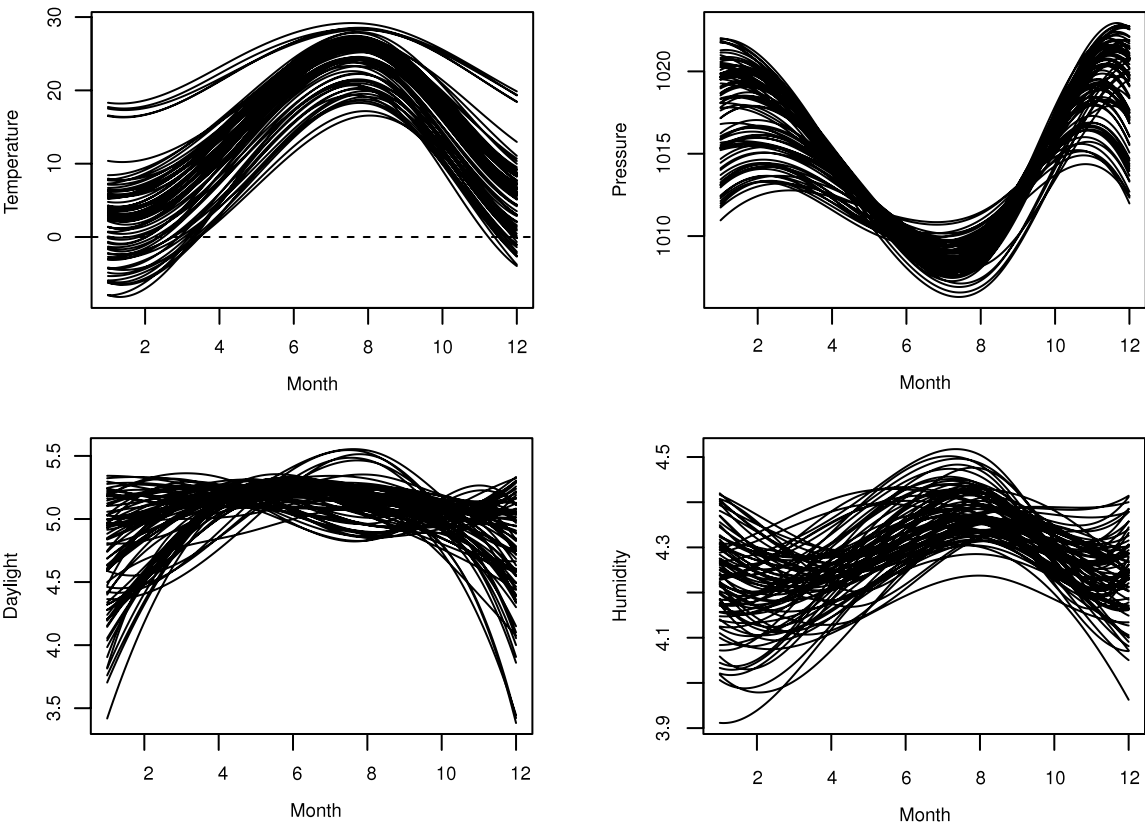


Fig. 1. Examples of smoothed functional covariates from weather data.

Table 2 Selected predictors for the weather dataset example.	
Method	Selected
T_L	PRESS, HUM, MAX.T
SCAD	LIGHT, HUM, MAX.T
LASSO	TEMP, LIGHT, HUM, MAX.T
GFLM	TEMP, PRESS, LIGHT

total precipitation. The dataset used in this analysis does not correspond to the one used in [24], rather we selected the 79 most reliable stations according to the aforementioned website.

The functional predictors, observed at a grid of 1–12 points, were fitted using 6 B-splines basis functions. Fig. 1 shows examples of the fitted functional predictors. The goal of this application is to select the functional covariates that significantly relate to annual total precipitation. We applied the proposed variable selection method and compared the results with those of the group SCAD, group LASSO and GFLM selection procedures, using the same number of basis functions.

The selected functional predictors for each method are shown in Table 2. Humidity and maximum temperature are selected by all methods except GFLM, however, differently from group SCAD and group LASSO, the proposed procedure and GFLM selected PRESS and did not select LIGHT. Atmospheric pressure is well known among meteorologists to be related to precipitation. Low and high air pressure systems are usually caused by unequal heating across the surface of the planet. A low pressure system is an area where the atmospheric pressure is lower than that of the area around it. The production of clouds and consequent precipitation are hence related to the wind, warm air and atmospheric lifting caused by low pressure systems.

In a simulation of 100 bootstrap samples from the weather data, we performed variable selection using the proposed method, group SCAD and group LASSO and GFLM. Table 3 shows the number of times each predictor was selected. While LIGHT was the third most selected predictor by group SCAD and group LASSO (about 70% of the time) and the most selected by GFLM, it was only the fourth most selected predictor when using the proposed procedure. On the other hand, pressure was selected most frequently by the proposed method, followed by humidity and maximum temperature. Our results meet the expectations of most specialized meteorology literature, which finds significant relation between pressure, humidity and maximum temperature with annual precipitation.

Table 3

Ratio of selection on 100 bootstrap samples of weather data.

Method	TEMP	PRESS	LIGHT	HUM	MAX.T	MIN.T
T_L (BC)	0.38	0.90	0.56	0.89	0.87	0.41
T_L (FDR)	0.40	0.90	0.58	0.87	0.86	0.45
SCAD (GCV)	0.37	0.23	0.65	0.81	0.81	0.24
SCAD (BIC)	0.37	0.21	0.75	0.81	0.83	0.23
LASSO (GCV)	0.45	0.35	0.62	0.78	0.80	0.25
LASSO (BIC)	0.45	0.34	0.75	0.81	0.81	0.23
GLM	0.73	0.67	0.79	0.47	0.47	0.21

Acknowledgments

This paper was partially supported by CNPq (grant 302956/2013-1), Fapesp (grant 2013/07375-0 and 2013/00506-1) and CAPES. We would also like to thank Michael G. Akritas and Nancy Lopes Garcia for their fruitful insights.

Appendix

Proof of Lemma 3.2. Since $(\mathbf{P} - \mathbf{P}_0)$ is idempotent, it is easy to show that the non-centrality parameter δ is equal to

$$\delta = \mathbf{b}^T \mathbf{Z}^T (\mathbf{P} - \mathbf{P}_0) \mathbf{Z} \mathbf{b} / \sigma^2 = \|\mathbf{Z} \mathbf{b} - \mathbf{P}_0 \mathbf{Z} \mathbf{b}\|^2 / \sigma^2.$$

Note that $E(\mathbf{Y}|\mathbf{Z}) = \mathbf{Z} \mathbf{b}$ is the vector of expected values conditional on \mathbf{Z} , which belongs to the subspace Ω , and $\mathbf{P}_0 \mathbf{Z} \mathbf{b}$ is its projection onto the restricted subspace ζ . Without loss of generality write $\mathbf{Z} \mathbf{b} = (\mathbf{Z}^0, \mathbf{Z}^1)(\mathbf{b}_{-r}, \mathbf{b}_r)$, where \mathbf{Z}^1 is the sub-matrix of \mathbf{Z} with columns corresponding to the parameters \mathbf{b}_r , and \mathbf{Z}^0 the remaining columns (similarly for \mathbf{b}_{-r}). Let $\tilde{\mathbf{Y}} = \mathbf{Z} \mathbf{b}$ so that $(\mathbf{P} - \mathbf{P}_0) \mathbf{Z} \mathbf{b} = \tilde{\mathbf{Y}} - \mathbf{P}_0 \tilde{\mathbf{Y}} = (\mathbf{I} - \mathbf{P}_0) \tilde{\mathbf{Y}}$. The quantity $(\mathbf{I} - \mathbf{P}_0) \tilde{\mathbf{Y}}$ is the residuals from the projection of $\tilde{\mathbf{Y}}$ onto the subspace ζ . This can be viewed as a linear model $\tilde{\mathbf{Y}} = E(\tilde{\mathbf{Y}}|\mathbf{Z}^0) + \tilde{\varepsilon}$, so that the mean squared error $\|(\mathbf{I} - \mathbf{P}_0) \tilde{\mathbf{Y}}\|^2 / (n - k_0) = \tilde{\mathbf{Y}}^T (\mathbf{I} - \mathbf{P}_0) \tilde{\mathbf{Y}} / (n - k_0) = \delta \sigma^2 / (n - k_0)$ will converge to the constant. This implies that $\delta \sim c(n - k_0)$.

Proof of Lemma 3.3. Part (a) Let $\Psi_{p_r}(\cdot)$ be the cumulative distribution function (c.d.f.) of the central $\chi_{p_r}^2$ distribution and $\Psi_{p_r}^{-1}(\cdot)$ its inverse. Also, denote the residual sum of squares under hypothesis H_0^r in (4) by RSS_0^r . Using the fact that $\lim_{n \rightarrow \infty} P(A_n) = 1$ (Lemma A.1 in [8]), we obtain $\lim_{n \rightarrow \infty} P(|\tilde{\sigma}^2 - \sigma^2| \geq \sigma \alpha) = 0$ for $\alpha = \sqrt{\log(n)/n}$. For all $r \notin I_0$, $\mathbf{b}_r = 0$, and for any $0 < \gamma < 1$ we find that

$$\begin{aligned} P(\{\pi_r \leq \gamma\} \cap A_n) &= P(\{1 - \Psi_{p_r}(T_L^r) \leq \gamma\} \cap A_n) = P(\{T_L^r \geq \Psi_{p_r}^{-1}(1 - \gamma)\} \cap A_n) \\ &= P\left(\left\{\frac{RSS_0^r - RSS}{\tilde{\sigma}^2} \geq \Psi_{p_r}^{-1}(1 - \gamma)\right\} \cap A_n\right) \leq P\left(\frac{RSS_0^r - RSS}{\sigma^2} \geq \left(1 - \frac{\alpha}{\sigma}\right) \Psi_{p_r}^{-1}(1 - \gamma)\right) \\ &= \gamma + O(\alpha). \end{aligned}$$

Part (b) Let $\alpha = \sqrt{\log(n)/n}$. For all $0 < \gamma < 1$,

$$\begin{aligned} P(\{\pi_r > \gamma\} \cap A_n) &= P(\{1 - \Psi_{p_r}(T_L^r) > \gamma\} \cap A_n) = P(\{T_L^r < \Psi_{p_r}^{-1}(1 - \gamma)\} \cap A_n) \\ &= P\left(\left\{\frac{RSS_0^r - RSS}{\tilde{\sigma}^2} < \Psi_{p_r}^{-1}(1 - \gamma)\right\} \cap A_n\right) \leq P\left(\frac{RSS_0^r - RSS}{\sigma^2} < \left(\frac{\alpha}{\sigma} + 1\right) \Psi_{p_r}^{-1}(1 - \gamma)\right). \end{aligned}$$

Under the alternative $(RSS_0^r - RSS)/\sigma^2$ has a non-central chi-square distribution with p_r degrees of freedom and non-centrality parameter δ , whose c.d.f. we denote by $\Psi_{p_r, \delta}(\cdot)$. Since $\delta \sim c(n - k_0)$ and $k \leq \sqrt{n}/\log(n)$, we conservatively have $\delta \sim c(n - \sqrt{n}/\log(n))$. For $\gamma \geq 1/n$, as $n \rightarrow \infty$ and hence $\delta \rightarrow \infty$, we have that

$$\begin{aligned} \Psi_{p_r, \delta}(\Psi_{p_r}^{-1}(1 - \gamma)) &= \sum_{j=0}^{\infty} \frac{\delta^j}{2^j j!} e^{-\frac{\delta}{2}} \Psi_{p_r+2j}(\Psi_{p_r}^{-1}(1 - \gamma)) \\ &= \sum_{j=0}^{\infty} \frac{\delta^j}{2^j j!} e^{-\frac{\delta}{2}} \left(1 - e^{-\Psi_{p_r}^{-1}(1 - \gamma)/2} \sum_{\ell=0}^{p_r/2+j-1} \frac{(\Psi_{p_r}^{-1}(1 - \gamma))^\ell}{2^\ell \ell!}\right) = o(\gamma), \end{aligned}$$

since the Poisson weights are dislocated to larger values of j at a rate of $\exp(n - \sqrt{n}/\log(n))$ while the values of $\Psi_{p_r+2j}(\Psi_{p_r}^{-1}(1 - \gamma))$ are dislocated at a rate slower than n , for the choice of γ (Note that even if γ was chosen to decrease at a slower rate than $\exp(-n)n^k$, the percentile $\Psi_{p_r}^{-1}(1 - \gamma)$ would increase slower than a linear rate in n , and $\Psi_{p_r, \delta}(\Psi_{p_r}^{-1}(1 - \gamma))$ would be $o(1)$.) Hence $P(\{\pi_r > \gamma\} \cap A_n) \leq \Psi_{p_r, \delta}((\frac{\alpha}{\sigma} + 1) \Psi_{p_r}^{-1}(1 - \gamma)) = o(\gamma) + O(\alpha)$. \square

Proof of Lemma 3.4. Since $\lim_{n \rightarrow \infty} P(A_n) = \lim_{n \rightarrow \infty} P(|\tilde{\sigma} - \sigma| \leq \alpha) = 1$, where $\alpha = \sqrt{\log(n)/n}$, it suffices to show that $\lim_{n \rightarrow \infty} P(\Gamma_n^c \cap A_n) = 0$. From Lemma 3.2, δ is of order $\sim cn$, so that for $\gamma = \alpha$

$$\begin{aligned} P(\Gamma_n^c \cap A_n) &\leq \sum_{m \in I_0} \sum_{k \notin I_0} P(\{\pi_k < \pi_m\} \cap A_n) \\ &\leq \sum_{m \in I_0} \sum_{k \notin I_0} [P(\{\pi_k \leq \gamma\} \cap A_n) + P(\{\pi_m > \gamma\} \cap A_n)] \\ &\leq \sum_{m \in I_0} \sum_{k \notin I_0} [\gamma + O(\alpha) + o(\gamma)] = M_0(M - M_0) [\gamma + O(\alpha) + o(\gamma)], \end{aligned}$$

where the last inequality follows from Lemma 3.3. Since $\gamma = \alpha$ we have $\lim_{n \rightarrow \infty} P(\Gamma_n^c \cap A_n) = 0$. \square

Proof of Theorem 3.5. We follow the proof in [8] to prove the theorem under FDR corrections. The case of Bonferroni corrections follows with similar steps. If \hat{I} is equal to I_0 , we have M_0 rejections ($R = M_0$) with none of them being erroneous ($V = 0$). Thus, the consistency of \hat{I} is verified by showing that

$$P(\hat{I} = I_0) = P(R = M_0, V = 0) \rightarrow 1, \quad \text{as } n \rightarrow \infty. \quad (10)$$

This follows by showing that both $P(R \neq M_0)$ and $P(V \geq 1)$ are asymptotically negligible. We have that [8, Lemma 2.1]

$$P(V \geq 1) \leq P(R \neq M_0) + \frac{M_0(M - M_0)}{M} q. \quad (11)$$

Hence, in order to show consistency of \hat{I} we need only show that $P(R \neq M_0) \rightarrow 0$. Let $q_M = q / \sum_{l=1}^M l^{-1}$ and note that $\{R \neq M_0\} = \cup_{m=M_0+1}^M \{\pi(m) \leq q_M m / M\} \cup \{\pi(M_0) > q_M M_0 / M\}$, so that

$$\begin{aligned} P(R \neq M_0) &\leq P(A_n^c) + P(\Gamma_n^c \cap A_n) + P\left(\left\{\pi(M_0) > q_M \frac{M_0}{M}\right\} \cap \Gamma_n \cap A_n\right) \\ &\quad + \sum_{m=M_0+1}^M P\left(\left\{\pi(m) \leq q_M \frac{m}{M}\right\} \cap \Gamma_n \cap A_n\right), \end{aligned} \quad (12)$$

where $A_n = \{|\tilde{\sigma} - \sigma| \leq \alpha\}$, with $\alpha = \sqrt{\log(n)/n}$, and Γ_n is the event defined in Lemma 3.4. The third term on the right hand side of (12) is equal to

$$\begin{aligned} P\left(\left\{\pi(M_0) > q_M \frac{M_0}{M}\right\} \cap \Gamma_n \cap A_n\right) &\leq M_0 \max_{m \in I_0} P\left(\left\{\pi_m > q_M \frac{M_0}{M}\right\} \cap A_n\right) \\ &= O\left(M_0 \left(o\left(\frac{q_M M_0}{M}\right) + \alpha\right)\right) = o(1), \quad \text{as } n \rightarrow \infty, \end{aligned}$$

by Lemma 3.3 and the assumptions of the theorem. For the last term in (12) we have

$$\begin{aligned} \sum_{m=M_0+1}^M P\left(\left\{\pi(m) \leq q_M \frac{m}{M}\right\} \cap \Gamma_n \cap A_n\right) &\leq \sum_{m=M_0+1}^M P(\{\pi(m) \leq q_M\} \cap \Gamma_n \cap A_n) \\ &\leq \sum_{m \notin I_0} P(\{\pi_m \leq q_M\} \cap A_n) = O\left((M - M_0) \left(\frac{q}{\log(M)} + \alpha\right)\right) \\ &= o(1), \quad \text{as } n \rightarrow \infty, \end{aligned}$$

by Lemma 3.3 and the assumptions of the theorem. This shows that $P(\{R \neq M_0\}) \rightarrow 0$. Following (11) with the choice of q , we can conclude that \hat{I} is consistent, i.e., $\lim_{n \rightarrow \infty} P(\hat{I} = I_0) = 1$. \square

References

- [1] F. Abramovich, Y. Benjamini, D.L. Donoho, I.M. Johnstone, Adapting to unknown sparsity by controlling the false discovery rate, *Ann. Statist.* 34 (2006) 584–653.
- [2] G. Aneiros, F. Ferraty, P. Vieu, Variable selection in semi-functional regression models, in: *Recent Advances in Functional Data Analysis and Related Topics*, in: Contributions to Statistics, vol. 57, 2011, pp. 17–22.
- [3] G. Aneiros, P. Vieu, Testing linearity in semi-parametric functional data analysis, *Comput. Statist.* 28 (2013) 413–434.
- [4] G. Aneiros, P. Vieu, Variable selection in infinite-dimensional problems, *Statist. Probab. Lett.* 94 (2014) 12–20.
- [5] G. Aneiros, P. Vieu, Partial linear modeling with multi-functional covariates, *Comput. Statist.* (ISSN: 1613-9658) (2015).
- [6] Y. Benjamini, D. Yekutieli, The control of the false discovery rate in multiple testing under dependency, *Ann. Statist.* 29 (2001) 1165–1188.
- [7] E.G. Bongiorno, E. Salinelli, A. Goia, P. Vieu, *Contributions in Infinite-Dimensional Statistics and Related Topics*, Società Editrice Esculapio, 2014.

- [8] F. Bunea, M. Wegkamp, A. Auguste, Consistent variable selection in high dimensional regression via multiple testing, *J. Statist. Plann. Inference* 136 (2006) 4349–4364.
- [9] H. Cardot, A. Goia, P. Sarda, Testing for no effect in functional linear regression models, some computational approaches, *Comm. Statist. Simulation Comput.* 33 (2004) 179–199.
- [10] A. Cuevas, A partial overview of the theory of statistics with functional data, *J. Statist. Plann. Inference* 147 (2014) 1–23.
- [11] J. Fan, R. Li, New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis, *J. Amer. Statist. Assoc.* 99 (2004) 710–723.
- [12] F. Ferraty, A. Laksaci, A. Tadj, P. Vieu, Rate of uniform consistency for nonparametric estimates with functional variables, *J. Statist. Plann. Inference* 140 (2010) 335–352.
- [13] F. Ferraty, P. Vieu, *Nonparametric Functional Data Analysis, Theory and Practice*, in: Springer Series in Statistics, 2006.
- [14] F. Ferraty, P. Vieu, Additive prediction and boosting for functional data, *Comput. Statist. Data Anal.* 53 (2009) 1400–1413.
- [15] J. Gertheiss, A. Maity, A.M. Staicu, Variable selection in generalized functional linear models, *Stat* 2 (2013) 86–101.
- [16] A. Goia, P. Vieu, A partitioned single functional index model, *Comput. Statist.* (ISSN: 1613-9658) (2014).
- [17] Z. Hong, H. Lian, Inference of genetic networks from time course expression data using functional regression with lasso penalty, *Comm. Statist. Theory Methods* 40 (2011) 1768–1779.
- [18] L. Horváth, P. Kokoszka, *Inference for Functional Data with Applications*, in: Springer Series in Statistics, 2012.
- [19] G.M. James, Generalized linear models with functional predictors, *J. R. Stat. Soc. Ser. B Stat. Methodol.* 64 (2002) 411–432.
- [20] G. James, J. Wang, J. Zhu, Functional linear regression that's interpretable, *Ann. Statist.* 37 (2009) 2083–2108.
- [21] M. Kayano, S. Konishi, Functional principal component analysis via regularized Gaussian basis expansions and its application to unbalanced data, *J. Statist. Plann. Inference* 139 (2009) 2388–2398.
- [22] D. Kong, A.M. Staicu, A. Maity, Classical Testing in Functional Linear Models, Technical Reports 2647, North Carolina State University, Dept. of Statistics, 2013, pp. 1–23.
- [23] S. Ma, Q. Song, L. Wang, Simultaneous variable selection and estimation in semiparametric modeling of longitudinal/clustering data, *Bernoulli* 19 (2013) 252–274.
- [24] H. Matsui, K. Konishi, Variable selection for functional regression models via the L_1 regularization, *Comput. Statist. Data Anal.* 55 (2011) 3304–3310.
- [25] M.W. McLean, G. Hooker, D. Ruppert, Restricted likelihood ratio tests for linearity in scalar-on-function regression, *Stat. Comput.* (2014). <http://dx.doi.org/10.1007/s11222-014-9473-1>.
- [26] N. Meinshausen, L. Meier, P. Bühlmann, p-Values for high-dimensional regression, *J. Amer. Statist. Assoc.* 104 (2009) 1671–1681.
- [27] N. Mingotti, R.E. Lillo, J. Romo, Lasso variable selection in functional regression, *Statistics and Econometrics Series 13*, Working paper 13–14, 2013.
- [28] G.M. Pomann, A.M. Staicu, S. Ghosh, Two sample hypothesis testing for functional data, North Carolina State University, Dept. of Statistics, Preprint, 2014, submitted for publication.
- [29] J.O. Ramsay, B.W. Silverman, *Functional Data Analysis*, second ed., Springer, New York, 2005.
- [30] U. Reif, Orthogonality of cardinal B-splines in weighted Sobolev spaces, *SIAM J. Math. Anal.* 28 (1997) 1258–1263.
- [31] A.C. Rencher, G.B. Schaalje, *Linear Models in Statistics*, second ed., Wiley, New Jersey, 2008.
- [32] B.J. Swihart, J. Goldsmith, C.M. Crainiceanu, Restricted likelihood ratio tests for functional effects in the functional linear model, *Technometrics* (2013). <http://dx.doi.org/10.1080/00401706.2013.863163>.
- [33] X. Yang, K. Nie, Hypothesis testing in functional linear regression models with Neyman's truncation and wavelet thresholding for longitudinal data, *Stat. Med.* 27 (2008) 845–863.
- [34] A.Z. Zambom, M.G. Akritas, Nonparametric lack-of-fit testing and consistent variable selection, *Statist. Sinica* 24 (2014) 1837–1858.