Consistent Variable Selection for Functional Regression Models

Julian A. A. Collazos, Ronaldo Dias
Department of Statistics - State University of Campinas (UNICAMP)
Rua Sergio Buarque de Holanda, 651, Distr. de Barao Geraldo, Campinas, Sao Paulo, Brazil

Adriano Z. Zambom
Department of Statistics - Penn State University
323 Thomas Bldg., University Park, PA

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.

Keywords: B-splines, hypotheses testing, False Discovery Rate, Functional Data, likelihood ratio test

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 (Ramsay and Silverman, 2005). Such models have become increasingly useful due to their large number of applications, see Kokoszka and Horvath (2012) for some fundamental results and Ferraty and Vieu
(2006) for a nonparametric approach. This high demand has recently leveraged important theoretical advances, see for example James (2002), Ferraty and Vieu (2009), James, Wang and Zhu (2009), Ferraty, Laksaci, Tadj and Vieu (2010), and Aneiros and Vieu (2013), Goia and Vieu (2014), to cite a few.

However, only a few authors have considered variable selection in functional regression analysis. Aneiros and Vieu (2014) 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 (2015) propose a variable selection method based on the continuous specificity of the functional data. Cuevas (2014, 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 Bongiorno et al. (2014). Another class of such methods uses regularization techniques, where the penalty simultaneously shrinks parameters and selects variables. Matsui and Konishi (2011) studied the group SCAD regularization for estimating and selecting functional regressors while Mingotti, Lillo and Romo (2013) and Hong and Lian (2011) 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 (2004), Aneiros, Ferraty, and Vieu (2011), Gertheiss, Maity, and Staicu (2013) and Ma, Song and Wang (2013).

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 (2006) showed that the application of a false discovery rate (FDR) controlling procedure, such as Benjamini and Yekutieli (2001), 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 (Cardot, Goia, and Sarda, 2004, Yang and Nie, 2008, Swihart, Goldsmith and Crainiceanu, 2013, Kong, Staicu and Maity, 2013, McLean, Hooker and Ruppert, 2014, Pomann, Staicu and Ghosh, 2014). An interesting application can be found in Meinshausen, Meier and Buhlmann (2009), with results on the connection between p-values and variable selection in regression analysis.

The main objective of this paper is twofold: 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 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 \( \{(y_i, x_i(t)) : t \in \mathcal{T}, i = 1, \ldots, n\} \), where \( y_i \) is a scalar response, \( x_i(t) = (x_{i1}(t_1), \ldots, x_{iM}(t_M)) \) are functional predictors and \( \mathcal{T} = \mathcal{T}_1 \times \ldots \times \mathcal{T}_M \). Each \( \mathcal{T}_m, m = 1, \ldots, M, \) is a compact set in \( \mathbb{R} \) where the \( m \)-th predictor may be observed. The functional predictors \( x_m, m = 1, \ldots, M \) 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:

\[
x_{im}(t_m) = \sum_{j=1}^{\omega_{im}} \omega_{imj} \phi_{mj}(t_m), \quad m = 1, \ldots, M, t_m \in \mathcal{T}_m,
\]

where \( \phi_{mj}(t_m) = (\phi_{m1}(t_m), \ldots, \phi_{mp_m}(t_m))^T \) are vectors of B-Splines basis functions. The basis functions and the \( p_m \) coefficients in \( \phi_{mj}(t_m) \) 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 Reif, 1997).

We consider the functional regression model (Ramsay and Silverman, 2005) given by

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

where \( \beta_0 \) is a constant, \( \varepsilon_i, i = 1, \ldots, n \) are i.i.d. Gaussian noises with mean 0 and constant variance \( \sigma^2 \), and \( \beta_m(t_m) \) are functional coefficients that we assume can be represented through the basis expansion

\[
\beta_m(t_m) = \sum_{j=1}^{\omega_{im}} b_{mj} \phi_{mj}(t_m) = b_m^T \phi_{mj}(t_m), \quad m = 1, \ldots, M, t_m \in \mathcal{T}_m,
\]

for the parameter vectors \( b_m = (b_{m1}, \ldots, b_{mp_m})^T \). Thus the FRM in [2] can be re-expressed as a linear model in the following way

\[
y_i = \beta_0 + \sum_{m=1}^{M} \int_{\mathcal{T}_m} W_{im}^T \phi_{mj}(t_m) \phi_{mj}(t_m) b_m dt_m + \varepsilon_i = \beta_0 + \sum_{m=1}^{M} W_{im}^T \int_{\mathcal{T}_m} \phi_{mj}(t_m) \phi_{mj}(t_m) b_m dt_m + \varepsilon_i
\]

\[
= \beta_0 + \sum_{m=1}^{M} W_{im}^T \phi_{mj} b_m + \varepsilon_i = Z_i^T b + \varepsilon_i,
\]

or in matrix form \( Y = Zb + \varepsilon \), where \( Z_i = (1, W_{i1}^T \phi_{j1}, \ldots, W_{iM}^T \phi_{jM})^T \), \( b = (\beta_0, b_1^T, \ldots, b_M^T)^T \), \( Z = (Z_1, \ldots, Z_n)^T \), \( J_{\phi_m} = \int_{\mathcal{T}_m} \phi_{mj}(t_m) \phi_{mj}(t_m) dt_m \) are \( p_m \times p_m \) cross product matrices and \( \varepsilon \) is the vector of error terms. Since we adopt B-splines basis expansions, the cross product matrix \( J_{\phi_m} \) can be easily computed using the procedure in Kayano and Konishi (2009).
3. Methodology

3.1. Testing procedure

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

\[
H_0^r : b_r = 0 \quad vs \quad H_a^r : b_r \neq 0.
\]  

(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 \( \zeta \) and \( \Omega \) 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 \( Z \) has full rank, that is, \( Z \) 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,

\[
RSS_0 = \sum_{i=1}^{n} \left( y_i - Z_i^T \hat{b}^0 \right)^2 \quad \text{and} \quad RSS = \sum_{i=1}^{n} \left( y_i - Z_i^T \hat{b} \right)^2,
\]  

(5)

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

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 \( Y_0 = Z \hat{b}^0 = P_0 Y \) and \( \tilde{Y} = Z \hat{b} = PY \), where \( P_0 \) and \( P \) are the orthogonal projection matrices which project \( Y \) onto the spaces \( \zeta \) and \( \Omega \), respectively. We can then rewrite the residual sum of squares as \( RSS_0 = Y^T (I_n - P_0) Y \) and \( RSS = Y^T (I_n - P) Y \), so that \( RSS_0 - RSS = Y^T (P - P_0) Y \). Since

\[
\frac{RSS_0}{\bar{\sigma}^2} \overset{H_0}{\sim} \chi^2_{n-k_0} \quad \text{and} \quad \frac{RSS}{\bar{\sigma}^2} \overset{H_0}{\sim} \chi^2_{n-k},
\]

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

\[
T_L = -2 \ln \left( \frac{L_0}{L} \right) = -2 \left[ -\frac{1}{2\bar{\sigma}^2} RSS_0 + \frac{1}{2\bar{\sigma}^2} RSS \right] = \frac{RSS_0 - RSS}{\bar{\sigma}^2} \overset{H_0}{\sim} \chi^2_{n-k_0}
\]  

(6)

in distribution, with \( \bar{\sigma}^2 = RSS/n \overset{P}{\rightarrow} \sigma^2 \) the maximum likelihood ratio statistic. From the Normality assumption of the residuals and the fact that

\[
\frac{1}{\sigma^2} E[RSS_0 - RSS] = \frac{1}{\sigma^2} \left( \sigma^2 Tr(P - P_0) + (Z \hat{b})^T (P - P_0) Z \hat{b} \right) = (k - k_0) + \delta = p_r + \delta,
\]

where

\[
\delta = \hat{b}^T Z^T (P - P_0) Z \hat{b} / \sigma^2,
\]  

(7)

the following proposition can be established.
Proposition 3.1. (Theorem 5.3c in Rencher and Schaalje, 2008) Let \( \text{RSS} \) and \( \text{RSS}_0 \) be defined as in (5). Then, under the alternative hypothesis in (4)

\[
\frac{\text{RSS}_0}{\sigma^2} \sim \chi^2_{n-k_0}(\delta) \quad \text{and} \quad \frac{\text{RSS}}{\sigma^2} \sim \chi^2_{n-k},
\]

so that

\[
\frac{\text{RSS}_0 - \text{RSS}}{\sigma^2} \sim \chi^2_{k-k_0}(\delta).
\]

Lemma 3.2 specifies the order of the non-centrality parameter of the distribution of \((\text{RSS}_0 - \text{RSS})/\sigma^2\). 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 \( \delta \) 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 (2006) in the linear model setting, and by Zambom and Akritas, (2014) for a nonparametric model.

Let \( I_M = \{1, \ldots, M\} \) 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

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

Let \( I_0 = \{m_1, \ldots, m_{M_0}\} \) 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 \( T_{L,r} \), \( r = 1, \ldots, M \), 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 \( \Psi(.) \) is the cumulative function of the \( \chi^2_{p_r} \) 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 (Benjamini and Yekutieli, 2001) computes

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

where \( \pi_1 \leq \cdots \leq \pi(M) \) denote the ordered p-values and \( q \) is the choice of level, and rejects \( H_0^{(j)} \), \( j = 1, \ldots, s \). 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 $\tilde{I}$ 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 $R = s \mathbb{1}(s \in [q])$ exists, where $\mathbb{1}(.)$ 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 $E(Q) \leq q(M - M_0)/M \leq q$, (Benjamini and Yekutieli, 2001). We consider consistent a procedure, and the estimated set $\hat{I}$, if $P(\tilde{I} = I_0) \to 1$ as $n \to \infty$. Theorem 3.5 in connection with Lemmas 3.2 - 3.4 show the consistency of $\hat{I}$.

**Lemma 3.3.** Let $T^*_L$ and $\pi_r = 1 - \Psi(T^*_L)$ be the test statistic and the p-value defined as in (8) and (9) for testing $H^*_0$. 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 $n \to \infty$, 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 $\Gamma_n$ 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 $I_0 = \{m_1, ..., m_{M_0}\}$, that is

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

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

**Theorem 3.5.** Let $\delta$ be the non-centrality parameter defined in (7), and $q$ the chosen bound of FDR in (8) or in Bonferroni corrections. Assume that condition (C1) holds and $q \to 0$ as $n \to \infty$, in such a way that $q \geq M \left(\sum_{i=1}^M l_i^{-1}\right)/(M \log(M))$ and $M q / \log(M) \to 0$. Then, $\lim_{n \to \infty} P\left(\tilde{I} = I_0\right) = 1$.

Note that the choice of $q \to 0$ 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 $\{x_{im}(t), y_i\}$, extending the simulation set up in Matsui and Konishi (2011) 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 (2011), and the Generalized Functional Linear Model (GFLM) method in Gertheiss, et al. (2013) with adaptive penalization. For comparison purposes,
we used 6 basis functions for the estimation of the predictors and the functional parameters $\beta(.)$ in all methods. First, we generated $z_{im}$ corresponding to the predictor $X_m$ in an equally spaced grid of 50 points in $T_m$ 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

$$u_{i1}(t) = \cos(2\pi(t - a_1)) + a_2, \quad \mathcal{T}_1 = [0, 1], \quad a_1 \sim N(-4, 3^2), \quad a_2 \sim N(7, 1.5^2),$$

$$u_{i2}(t) = b_1 \sin(\pi t) + b_2, \quad \mathcal{T}_2 = [0, \pi/3], \quad b_1 \sim U(3, 7), \quad b_2 \sim N(0, 1),$$

$$u_{i3}(t) = c_1 t^3 + c_2 t^2 + c_3 t, \quad \mathcal{T}_3 = [-1, 1], \quad c_1 \sim N(-3, 1.2^2), \quad c_2 \sim N(2, 0.5^2), c_3 \sim N(-2, 1),$$

$$u_{i4}(t) = \sin(2(t - d_1)) + d_2 t, \quad \mathcal{T}_4 = [0, \pi/3], \quad d_1 \sim N(-2, 1), \quad d_2 \sim N(3, 1.5^2),$$

$$u_{i5}(t) = e_1 \cos(2t) + e_2 t, \quad \mathcal{T}_5 = [-2, 1], \quad e_1 \sim U(2, 7), \quad e_2 \sim N(2, 0.4^2),$$

$$u_{i6}(t) = f_1 e^{-t/3} + f_2 t + f_3, \quad \mathcal{T}_6 = [-1, 1], \quad f_1 \sim N(4, 2^2), \quad f_2 \sim N(-3, 0.5^2), f_3 \sim N(1, 1).$$

The scalar response $y_i$ was generated as $y_i = g(u_i) + \epsilon_i$, where $g(u_i) = \sum_{m=1}^{6} \int_{\mathcal{T}_m} u_{im}(t) \beta_m(t) dt$, $\epsilon_i \sim N(0, (0.05 R_y)^2)$ and $R_y = max(g(u_i)) - min(g(u_i))$. For a constant $c = 0.0, 0.4, 0.8$, 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

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

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 II 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, where 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.
Table 1: Number of correctly selected models and AMSE

| c | n | \( T^R_{BC} \) | \( T^F_{DR} \) | SCAD | LASSO | GFLM |
|---|---|---|---|---|---|---|
|   |   | \(.01\) | \(.05\) | \(.1\) | \(.01\) | \(.05\) | \(.1\) | 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) |
| .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) |
| .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) |

5. Real Data Example: Weather Data

In this application, we consider weather data observed monthly at 79 weather stations in Japan. The data set was obtained from [http://www.data.jma.go.jp/obd/stats/data/en/](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 total precipitation. The dataset used in this analysis does not correspond to the one used in Matsui and Konishi (2011), rather we selected the 79 most reliable stations according to the aforementioned website.

The functional predictors, observed at a grid of 1 to 12 points, were fitted using 6 B-splines basis functions. Figure 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.

![Figure 1 about here](image)

Figure 1: Examples of smoothed functional covariates from weather data

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.

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            |

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  |

**Acknowledgements** 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 $(P - P_0)$ is idempotent, it is easy to show that the non-centrality parameter $\delta$ is equal to

$$\delta = b^T Z^T (P - P_0) Z b / \sigma^2 = \|Z b - P_0 Z b\|^2 / \sigma^2.$$
Note that $E(Y|Z) = Zb$ is the vector of expected values conditional on $Z$, which belongs to the subspace $\Omega$, and $P_0Zb$ is its projection onto the restricted subspace $\zeta$. Without loss of generality write $Zb = (Z^0, Z^1)(b_{-r}, b_r)$, where $Z^1$ is the sub-matrix of $Z$ with columns corresponding to the parameters $b_r$, and $Z^0$ the remaining columns (similarly for $b_{-r}$). Let $\hat{Y} = Zb$ so that $(P - P_0)Zb = \hat{Y} - P_0 \hat{Y} = (I - P_0)\hat{Y}$. The quantity $(I - P_0)\hat{Y}$ is the residuals from the projection of $\hat{Y}$ onto the subspace $\zeta$. This can be viewed as a linear model $\tilde{Y} = E(\tilde{Y}|Z^0) + \varepsilon$, so that the mean squared error $|| (I - P_0)\hat{Y} ||^2 / (n-k_0) = \hat{Y}^T(I - P_0)\hat{Y} / (n-k_0) = \delta \sigma^2 / (n-k_0)$ will converge to the constant. This implies that $\delta \sim c(n-k_0)$.

Proof of Lemma 3.3

Part (a) Let $\Psi_{\alpha}(\cdot)$ be the cumulative distribution function (c.d.f.) of the central $\chi^2_{p_r}$ distribution and $\Psi_{\alpha}^{-1}(\cdot)$ its inverse. Also, denote the residual sum of squares under hypothesis $H_0$ in (1) by $RSS_0^*$. Using the fact that $\lim_{n \to \infty} P(A_n) = 1$ (Lemma A.1 in Bunea et al., 2006), we obtain $\lim_{n \to \infty} P(\delta^2 - \sigma^2 \geq \sigma \alpha) = 0$ for $\alpha = \sqrt{\log(n)/n}$. For any $r \notin I_0$, $b_r = 0$, and for any $0 < \gamma < 1$ we find that

$$P\left(\left\{ \pi_r \leq \gamma \right\} \cap A_n \right) = P\left(\left\{ 1 - \Psi_{\alpha}(T_L^r) \leq \gamma \right\} \cap A_n \right) = P\left(\left\{ T_L^r \geq \Psi_{\alpha}^{-1}(1 - \gamma) \right\} \cap A_n \right)$$

$$= P\left(\left\{ \frac{RSS_0^* - RSS}{\sigma^2} \geq \Psi_{\alpha}^{-1}(1 - \gamma) \right\} \cap A_n \right) \leq P\left(\frac{RSS_0^* - RSS}{\sigma^2} \geq \left(1 - \frac{\alpha}{\sigma}\right) \Psi_{\alpha}^{-1}(1 - \gamma) \right) = \gamma + O(\alpha).$$

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

$$P\left(\left\{ \pi_r > \gamma \right\} \cap A_n \right) = P\left(\left\{ 1 - \Psi_{\alpha}(T_L^r) > \gamma \right\} \cap A_n \right) = P\left(\left\{ T_L^r < \Psi_{\alpha}^{-1}(1 - \gamma) \right\} \cap A_n \right)$$

$$= P\left(\left\{ \frac{RSS_0^* - RSS}{\sigma^2} < \Psi_{\alpha}^{-1}(1 - \gamma) \right\} \cap A_n \right) \leq P\left(\frac{RSS_0^* - RSS}{\sigma^2} < \left(\frac{\alpha}{\sigma} + 1\right) \Psi_{\alpha}^{-1}(1 - \gamma) \right).$$

Under the alternative $(RSS_0^* - RSS)/\sigma^2$ has a non-central chi-square distribution with $p_r$ degrees of freedom and non-centrality parameter $\delta$, whose c.d.f. we denote by $\Psi_{\alpha,\delta}(\cdot)$. 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 $n \to \infty$ and hence $\delta \to \infty$, we have that

$$\Psi_{\alpha,\delta}(\Psi_{\alpha}^{-1}(1 - \gamma)) = \sum_{j=0}^{\infty} \frac{\delta^j}{2^j j!} e^{-\frac{1}{2}} \Psi_{\alpha+2j}(\Psi_{\alpha}^{-1}(1 - \gamma))$$

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

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 $\Psi_{\alpha+2j}(\Psi_{\alpha}^{-1}(1 - \gamma))$ are dislocated at a rate slower than $n$, for the choice of $\gamma$ (Note that even if $\gamma$ was chosen to decrease at a slower rate than $\exp(-n)n^k$, the percentile $\Psi_{\alpha}^{-1}(1 - \gamma)$ would increase slower than a linear rate in $n$, and $\Psi_{\alpha,\delta}(\Psi_{\alpha}^{-1}(1 - \gamma))$ would be $o(1)$). Hence $P\left(\left\{ \pi_r > \gamma \right\} \cap A_n \right) \leq \Psi_{\alpha,\delta}(\left(\frac{\alpha}{\sigma} + 1\right) \Psi_{\alpha}^{-1}(1 - \gamma)) = o(\gamma) + O(\alpha).$ \hfill \box

Proof of Lemma 3.4

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

\[
P(\Gamma_n^c \cap A_n) \leq \sum_{m \in I_0} \sum_{k \notin I_0} P(\{\pi_k < \pi_m \} \cap A_n)
\]

\[
\leq \sum_{m \in I_0} \sum_{k \notin I_0} [P(\{\pi_k \leq \gamma \} \cap A_n) + P(\{\pi_m > \gamma \} \cap A_n)]
\]

\[
\leq \sum_{m \in I_0} \sum_{k \notin I_0} [\gamma + O(\alpha) + o(\gamma)] = M_0(M - M_0) [\gamma + O(\alpha) + o(\gamma)],
\]

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

**Proof of Theorem 3.5**

We follow the proof in Bunea et. al. (2006) 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

\[ P(\hat{I} = I_0) = P(R = M_0, V = 0) \to 1, \text{ as } n \to \infty. \] (10)

This follows by showing that both \( P(R \neq M_0) \) and \( P(V \geq 1) \) are asymptotically negligible. We have that (Bunea et al. 2006, Lemma 2.1)

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

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

\[
P(R \neq M_0) \leq P(A_n^c) + P(\Gamma_n^c \cap A_n) + P(\left\{ \pi(M_0) > q_M \frac{M_0}{M} \right\} \cap \Gamma_n \cap A_n)
\]

\[
+ \sum_{m=M_0+1}^{M} P(\left\{ \pi(m) \leq q_M \frac{m}{M} \right\} \cap \Gamma_n \cap A_n), \tag{12}
\]

where \( A_n = \{|\tilde{\sigma} - \sigma| \leq \alpha\} \), with \( \alpha = \sqrt{\log(n)/n} \), and \( \Gamma_n \) is the event defined in Lemma 3.3. The third term on the right hand side of (12) is equal to

\[
P(\left\{ \pi(M_0) > q_M \frac{M_0}{M} \right\} \cap \Gamma_n \cap A_n) \leq M_0 \max_{m \in I_0} P(\left\{ \pi_m > q_M \frac{M_0}{M} \right\} \cap A_n)
\]

\[
= O \left( M_0 \left( o \left( \frac{q_M M_0}{M} \right) + \alpha \right) \right) = o(1), \text{ as } n \to \infty,
\]

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

\[
\sum_{m=M_0+1}^{M} P(\left\{ \pi(m) \leq q_M \frac{m}{M} \right\} \cap \Gamma_n \cap A_n) \leq \sum_{m=M_0+1}^{M} P(\left\{ \pi(m) \leq q_M \right\} \cap \Gamma_n \cap A_n)
\]

\[
\leq \sum_{m \in I_0} P(\{\pi_m \leq q_M \} \cap A_n) = O \left( (M - M_0) \left( \frac{q}{\log(M)} + \alpha \right) \right) = o(1), \text{ as } n \to \infty,
\]

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

**References**
References

[1] Abramovich, F., Benjamini, Y., Donoho, D.L., and Johnstone, I.M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *The Annals of Statistics*, 34, 584-653.

[3] Aneiros, G., Ferraty F. and Vieu, P. (2011) Variable Selection in Semi-Functional Regression Models. *Recent Advances in Functional Data Analysis and Related Topics-Contributions to Statistics* 57, 17-22.

[3] Aneiros, G. and Vieu, P. (2013). Testing linearity in semi-parametric functional data analysis. *Computational Statistics*, 28, 413-434.

[4] Aneiros, G., Vieu, P. (2014). Variable selection in infinite-dimensional problems. *Statistics and Probability Letters*, 94, 12-20.

[5] Aneiros, G., Vieu, P. (2015). Partial linear modeling with multi-functional covariates. *Computational Statistics*, Online ISSN: 1613-9658.

[6] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29, 1165-1188.

[7] Bongiorno, E. G., Salinelli, E., Goia, A. and Vieu, P. (2014). Contributions in infinite-dimensional statistics and related topics. *Societá Editrice Esculapio*.

[8] Bunea, F., Wegkamp, M., and Auguste, A. (2006). Consistent variable selection in high dimensional regression via multiple testing. *Journal of Statistical Planning and Inference* 136, 4349-4364.

[9] Cardot, H., Goia, A., and Sarda, P. (2004). Testing for No Effect in Functional Linear Regression Models, Some Computational Approaches. *Com. in Stat. - Simul. and Computation*, 33, 179-199.

[10] Cuevas, A. (2014). A partial overview of the theory of statistics with functional data. *J. of Statistical Planning and Inference*, 147, 1-23.

[11] Fan, J., and Li, R. (2004). New Estimation and Model Selection Procedures for Semiparametric Modeling in Longitudinal Data Analysis. *JASA*, 99, 710-723.

[12] Ferraty, F. and Vieu, P. (2006). Nonparametric Functional Data Analysis, Theory and Practice. *Springer Series in Statistics*.

[13] F. Ferraty and P. Vieu (2009). Additive prediction and boosting for functional data. *Computational Statistics & Data Analysis*, 53, 1400-1413.

[14] Ferraty, F., Laksaci, A., Tadj, A. and Vieu, P. (2010). Rate of uniform consistency for nonparametric estimates with functional variables. *J. Statistical Planning Inference*, 140, 335-352.
[15] Gertheiss, J., Maity, A., and Staicu, A.M. (2013). Variable Selection in Generalized Functional Linear Models. *Stat*, 2, 86-101.

[16] Goia, A. and Vieu, P. (2014). A partitioned Single Functional Index Model. *Computational Statistics*, Online ISSN 1613-9658.

[17] Horváth, L. and Kokoszka, P. (2012) Inference for Functional Data with Applications. *Springer Series in Statistics*.

[18] Hong, Z. and Lian, H. (2011). Inference of genetic networks from time course expression data using functional regression with lasso penalty. *Commun. in Statistics - Theory and Methods*, 40, 1768-1779.

[19] James, G. M. (2002). Generalized linear models with functional predictors. *JRSS-B*, 64, 411-432.

[20] James, G., Wang, J. and Zhu, J. (2009). Functional linear regression that’s interpretable. *Ann. Statist.*, 37, 2083-2108.

[21] Kayano, M., and Konishi, S. (2009). Functional principal component analysis via regularized Gaussian basis expansions and its application to unbalanced data. *JSPI*, 139, 2388-2398.

[22] Kong, D., Staicu, A.M., and Maity, A. (2013). Classical testing in functional linear models. *North Carolina State University, Dept. of Statistics*, Technical Reports 2647, 1-23.

[23] Ma, S., Song, Q. and Wang, L. (2013). Simultaneous variable selection and estimation in semiparametric modeling of longitudinal/clustered data. *Bernoulli*, 19, 252-274.

[24] Matsui, H., and Konishi, K. (2011). Variable selection for functional regression models via the $L_1$ regularization. *Computational Statistics and Data Analysis* 55, 3304-3310.

[25] McLean, M.W., Hooker, G., and Ruppert, D. (2014). Restricted likelihood ratio tests for linearity in scalar-on-function regression. *Statistics and Computing*, DOI: 10.1007/s11222-014-9473-1.

[26] Meinshausena, N., Meiera, L. and Buhlmann, P. (2009). p-Values for High-Dimensional Regression. *JASA*, 104, 1671-1681.

[27] Mingotti, N., Lillo, R. E., and Romo, J. (2013). Lasso variable selection in functional regression. *Statistics and Econometrics Series 13*, Working paper 13-14.

[28] Pomann, G.M., Staicu, A.M., and Ghosh, S. (2014). Two Sample Hypothesis Testing for Functional Data. *North Carolina State University, Dept. of Statistics*, Preprint Submitted.

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

[30] Reif, U. (1997). Orthogonality of cardinal B-splines in weighted Sobolev spaces. *SIAM J. Math. Anal.*, 28, 1258-1263.
[31] Rencher, A. C. and Schaalje, G. B. (2008). Linear Models in Statistics, 2nd ed. Wiley, New Jersey.

[32] Swihart, B.J., Goldsmith, J., and Crainiceanu, C.M. (2013). Restricted likelihood ratio tests for functional effects in the functional linear model. *Technometrics*. DOI: 10.1080/00401706.2013.863163.

[33] Yang, X., and Nie, K. (2008). Hypothesis testing in functional linear regression models with Neyman’s truncation and wavelet thresholding for longitudinal data. *Statistics in Medicine*, 27, 845-863.

[34] Zambom, A.Z., and Akritas, M.G. (2014). Nonparametric lack-of-fit testing and consistent variable selection. *Statistica Sinica* 24, 1837-1858.