Consistent Variable Selection in Functional Linear Regression Model

Adriano Zanin Zambom
State University of Campinas, Campinas, Brazil - adriano.zambom@gmail.com

Julian A. A. Collazos*
State University of Campinas, Campinas, Brazil - jualacco@gmail.com

Ronaldo Dias
State University of Campinas, Campinas, Brazil - dias@ime.unicamp.br

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 in methodological investigations. For the functional linear regression model where the predictor variables are observed over a grid and the response is scalar, 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 proven to be consistent in selecting the set of relevant predictors. A real dataset from weather stations in Japan is analyzed.

Keywords: B-splines; hypotheses testing; false discovery rate; 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. 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, whose inner product with an unknown smooth coefficient function represents the effect of the predictor on the response variable. A class of such methods uses regularization techniques such as the Lasso (Tibshirani, 1996) and the group SCAD (Fan, & Li, 2001), where the penalty simultaneously shrinks parameters and selects variables. Matsui, & Konishi (2011) studied the group SCAD regularization for estimating and selecting functional regressors. Other recent contributions to the variable selection problem in the functional models are Aneiros et al. (2011), Gertheiss et al. (2013).
In this work we will be considered 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 et al. (2006) showed that application of the false discovery rate (FDR) controlling procedure of Benjamini, & Hochberg (1995) 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 (Swihart et al. (2013), Pomann et al. (2014)). This work aims at showing that a powerful hypothesis test statistic can be used to construct a competitive variable selection procedure by exploiting the aforementioned conceptual connection between model checking and variable selection. Thus, this work has two objectives: study the asymptotic properties of the hypothesis test based on error sum of squares for the relevance of a functional predictor in a multivariate functional regression model using B-spline basis expansions; and propose a competitive variable selection procedure based on the Benjamini, & Yekuteli (2001) FDR method (or Bonferroni) applied on the p-values from the tests of each available functional predictor.

2. The Functional Regression Model: FRM

Suppose that we have n observations $\{(y_i, \boldsymbol{X}_i(t)) : t \in \mathcal{T}, i = 1, ..., n\}$, where y_i is a scalar response and $\boldsymbol{X}_i(t) = (X_{i1}(t), ..., X_{iM}(t))$ are functional predictors. The set $\mathcal{T}$ is a compact set in $\mathbb{R}$ where the functional predictor may be observed, in general time or space. We assume that each of the M functional predictors can be expressed as:

$$X_{im}(t) = \sum_{j=1}^{p_m} \omega_{imj}\phi_{mj}(t) = \boldsymbol{W}_{im}^T \boldsymbol{\phi}_m(t), \quad m = 1, \ldots, M, \tag{1}$$

where $\boldsymbol{W}_{im} = (\omega_{im1}, ..., \omega_{imp_m})^T$ are the vectors of coefficients and
$\boldsymbol{\phi}_m(t) = (\phi_{m1}(t), ..., \phi_{mp_m}(t))^T$ are vectors of B-Splines basis functions. Estimating the smooth curves that represent the functional predictors can thus be easily accomplished by estimating the coefficients $\boldsymbol{W}_{im}$. We consider the functional regression model (Ramsay, & Silverman, (2005)) given by

$$y_i = \beta_0 + \sum_{m=1}^{M} \int_{\mathcal{T}} X_{im}(t)\beta_m(t)dt + \varepsilon_i, \tag{2}$$

where β_0 is a constant, ε_i is a Gaussian noise with mean 0 and constant variance σ^2, and $\beta_m(t)$ are functional coefficients that we assume can be represented through the basis expansion

$$\beta_m(t) = \sum_{j=1}^{p_m} b_{mj}\phi_{mj}(t) = \boldsymbol{b}_m^T \boldsymbol{\phi}_m(t), \quad m = 1, \ldots, M, \tag{3}$$

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

$$y_i = \beta_0 + \sum_{m=1}^{M} \int_{\mathcal{T}} \boldsymbol{W}_{im}^T \boldsymbol{\phi}_m(t)\boldsymbol{\phi}_m^T(t)\boldsymbol{b}_m dt + \varepsilon_i = \beta_0 + \sum_{m=1}^{M} \boldsymbol{W}_{im}^T \int_{\mathcal{T}} \boldsymbol{\phi}_m(t)\boldsymbol{\phi}_m^T(t)dt\boldsymbol{b}_m + \varepsilon_i \tag{4}$$

$$= \beta_0 + \sum_{m=1}^{M} \boldsymbol{W}_{im}^T \boldsymbol{J}_{\boldsymbol{\phi_m}} \boldsymbol{b}_m + \varepsilon_i = \boldsymbol{Z_i}^T \boldsymbol{b} + \varepsilon_i, \tag{5}$$

or in matrix form $\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}$, where $\boldsymbol{Z}_i = (1, \boldsymbol{W}_{i1}^T \boldsymbol{J}_{\boldsymbol{\phi_1}}, ..., \boldsymbol{W}_{iM}^T \boldsymbol{J}_{\boldsymbol{\phi_M}})^T$, $\boldsymbol{b} = (\beta_0, \boldsymbol{b}_1^T, ..., \boldsymbol{b}_M^T)^T$, $\boldsymbol{Z} = (\boldsymbol{Z}_1^T, \ldots, \boldsymbol{Z}_n^T)^T$, $\boldsymbol{J}_{\boldsymbol{\phi_m}} = \int_{\mathcal{T}} \boldsymbol{\phi}_m(t)\boldsymbol{\phi}_m^T(t)dt$ 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 $\boldsymbol{J}_{\boldsymbol{\phi_m}}$ can be easily computed. It follows from the assumption of Normality of the error terms that the FRM in (2), given a functional covariate X_i, has the following probability density function

$$f(y_i|X_i; \boldsymbol{b}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{(y_i - \boldsymbol{Z}_i^T\boldsymbol{b})^2}{2\sigma^2}\right\}. \tag{6}$$

The likelihood ratio test statistic, described below, is hence based on the normal likelihood function in terms of the residual sum of squares under the restricted and unrestricted models.

3. 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

$$H_0 : \boldsymbol{b}_r = \boldsymbol{0} \quad vs \quad H_a : \boldsymbol{b}_r \neq \boldsymbol{0}. \tag{7}$$

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 $\text{rank}(\Omega) = \sum_{m=1}^{M} p_m := k$ and $\text{rank}(\zeta) = k - p_r = \sum_{m=1}^{M} p_m - p_r := k_0$. Let RSS_0 and RSS denote the residual sum of squares under H_0 and H_a respectively, that is,

$$RSS_0 = \sum_{i=1}^{n} \left(y_i - \boldsymbol{Z}_i^T \hat{\boldsymbol{b}}^0\right)^2 \text{ and } RSS = \sum_{i=1}^{n} \left(y_i - \boldsymbol{Z}_i^T \hat{\boldsymbol{b}}\right)^2, \tag{8}$$

where $\hat{\boldsymbol{b}}^0 = \hat{\boldsymbol{b}} - (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{A}^T(\boldsymbol{A}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{A}^T)^{-1}\boldsymbol{A}\hat{\boldsymbol{b}}$ for a $p_r \times k_a$ matrix $\boldsymbol{A}$ defining the null hypothesis, i.e., $\boldsymbol{Ab} = \boldsymbol{0}$ implies $\boldsymbol{b}_r = \boldsymbol{0}$. For insight into the asymptotic distribution and the non-centrality parameter of the test statistic presented below, it is useful to express the sum of squares RSS_0 and RSS in quadratic form. We write $\hat{\boldsymbol{Y}}_0 = \boldsymbol{Z}\hat{\boldsymbol{b}}^0 = \boldsymbol{P}_0\boldsymbol{Y}$ and $\hat{\boldsymbol{Y}} = \boldsymbol{Z}\hat{\boldsymbol{b}} = \boldsymbol{PY}$, where $\boldsymbol{P}_0$ and $\boldsymbol{P}$ are the orthogonal projections of $\boldsymbol{Y}$ onto the spaces ζ and Ω, respectively. We can then rewrite the residual sum of squares as $RSS_0 = \boldsymbol{Y}^T(\boldsymbol{I}_n - \boldsymbol{P}_0)\boldsymbol{Y}$ and $RSS = \boldsymbol{Y}^T(\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{Y}$, so that $RSS_0 - RSS = \boldsymbol{Y}^T(\boldsymbol{P} - \boldsymbol{P}_0)\boldsymbol{Y}$. Since (Seber & Lee, Thm 2.7)

$$\frac{RSS_0}{\sigma^2} \overset{H_0}{\sim} \boldsymbol{\chi^2_{n-k_0}} \quad \text{and} \quad \frac{RSS}{\sigma^2} \overset{H_0}{\sim} \boldsymbol{\chi^2_{n-k}}, \tag{9}$$

in order to test H_0 in (7) we use the likelihood ratio statistic

$$T_L = -2Ln\left[\frac{\check{L}_0}{\tilde{L}}\right] = -2\left[-\frac{1}{2\tilde{\sigma}^2}RSS_0 + \frac{1}{2\tilde{\sigma}^2}RSS\right] = \frac{RSS_0 - RSS}{\tilde{\sigma}^2} \underset{n\to\infty}{\overset{H_0}{\to}} \boldsymbol{\chi^2_{k-k_0}}, \tag{10}$$

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

$$E\left[RSS_0 - RSS\right] = \frac{1}{\sigma^2}\left[\sigma^2 Tr(\boldsymbol{P} - \boldsymbol{P}_0) + (\boldsymbol{Zb})^T(\boldsymbol{P} - \boldsymbol{P}_0)\boldsymbol{Zb}\right] = (k - k_0) + \delta = p_r + \delta, \tag{11}$$

where

$$\delta = \boldsymbol{b}^T\boldsymbol{Z}^T(\boldsymbol{P} - \boldsymbol{P}_0)\boldsymbol{Zb}/\sigma^2, \tag{12}$$

it is straightforward to show that

$$\frac{RSS_0}{\sigma^2} \overset{H_a}{\sim} \boldsymbol{\chi^2_{n-k_0}}(\delta) \quad \text{and} \quad \frac{RSS}{\sigma^2} \overset{H_a}{\sim} \boldsymbol{\chi^2_{n-k}}, \tag{13}$$

and

$$T_L = \frac{RSS_0 - RSS}{\tilde{\sigma}^2} \underset{n\to\infty}{\overset{H_a}{\to}} \boldsymbol{\chi^2_{k-k_0}}(\delta), \tag{14}$$

that is, under the alternative hypothesis T_L has an asymptotic non-central chi-square distribution with $(k - k_0) = p_r$ degrees of freedom and non-centrality parameter δ. Note that, if adjusted by the degrees of freedom in the numerator and denominator, the proposed test statistic has a central F distribution under H_0 and a non-central F distribution under H_a, which will also converge to the designated asymptotic distribution.

Lemma 3.1 specifies the order of the non-centrality parameter of the asymptotic distribution of the test statistic. 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 information, we prove in Theorem 4.3 the consistency of the proposed variable selection procedure described below.

Lemma 3.1. Let T_L be the likelihood ratio test statistic defined in (10) for testing H_0 in (7). Under the alternative hypothesis the non-centrality parameter of the asymptotic distribution of T_L is of order

$$\delta = \frac{\boldsymbol{b}^T\boldsymbol{Z}^T(\boldsymbol{P} - \boldsymbol{P}_0)\boldsymbol{Zb}}{\sigma^2} \tag{15}$$

$$= O(n)\left(\Lambda^2 c_1 + \Lambda c_2 + c_3\right) = O(n). \tag{16}$$

where $\Lambda = \int_{\mathcal{T}} X_r(t)\beta_r(t)dt$ and $c_k, k = 1, 2, 3$, are constants with relation to $\beta_r(t)$.

4. 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 truly relevant predictors. A similar procedure was used by Bunea et al. (2006) in the linear model setting, and by Zambom, & Akritas, (2014) for a nonparametric regression model.

Let $I_M = \{1, ..., M\}$ denote the set of indices of the M available functional predictors. We assume that the true underlying model is sparse in the sense that only a few predictors significantly relate to the response variable. Let $I_0 = \{m_1, ..., m_{M_0}\}$ 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, $r = 1, ..., M$, denote the likelihood test statistic defined in (10) for testing H_0^r in (7) and

$$\pi_r = 1 - \Psi(T_L^r) \tag{17}$$

the corresponding p-value, where $\Psi(.)$ is the cumulative function of the $\boldsymbol{\chi^2_{p_r}}$ distribution. Let $\pi_{(1)} \leq ... \leq \pi_{(M)}$ denote the ordered p-values. The false discovery rate (FDR) procedure (Benjamini, & Hochberg (1995), Benjamini , & Yekutieli, (2001)) computes

$$k = \max \left\{ j : \pi_{(j)} \leq \frac{j}{M} \frac{q}{\sum_{l=1}^{d} l^{-1}} \right\} \tag{18}$$

for a choice of level q and rejects $H_0^{(j)}$, $j = 1, ..., k$. If no such k exists, no hypotheses are rejected. The proposed variable selection method selects the predictors with indices corresponding to the k rejected null hypotheses. Hence, I_0 is estimated by the set $\widehat{I}$ of indices corresponding to the first k 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

$$R = \begin{cases} k, & \text{if k in (18) exists,} \\ 0, & \text{otherwise.} \end{cases} \tag{19}$$

Now, let V be the number of falsely rejected hypotheses, and set

$$Q = \begin{cases} V/R & \text{if } R > 0, \\ 0 & \text{otherwise,} \end{cases} \tag{20}$$

for the proportion of falsely rejected hypotheses. By definition, the FDR is $E(Q)$, and $E(Q) \leq q(d-d_0)/d \leq q$. We consider consistent a procedure, and the estimated set $\hat{I}$, if $P(\hat{I} = I_0) \to 1$. The consistency result presented in Theorem 4.3 using the FDR or Bonferroni corrections for the p-values, allows the significance of the predictors to be diminishing with n, where the significance of a predictor is measured by $\Lambda_r = \int_{\mathcal{T}} X_r(t)\beta_r(t)dt$. To show consistency of the proposed method, we need first of two results given by the lemma 4.1 which indicate that the asymptotic distribution of π_r, $r \notin I_0$ is Uniform(0, 1), in part (a) and for $r \in I_0$, we can obtain a degenerate distribution $\pi_r \xrightarrow{p} 0$, in (b), and taking into account this, later we define the event Γ_n of the smallest p-values of the M_0 significant functional predictors to show that $\lim_{n\to\infty} P(\Gamma_n) = 1$ by lemma 4.2.

Lemma 4.1. Let T_L^r and $\pi_r = 1 - \Psi(T_L^r)$ be the test statistic and the p-value defined as in (10) and (17) for testing H_0^r, we have:

(a) For $r \notin I_0$ and any $\gamma > 0$, we have $P(\pi_r \leq \gamma) = \gamma + o(1)$.

(b) For $r \in I_0$, and $\gamma_n > 0, n \geq 1$. Then, as $n \to \infty$, if

$$\Lambda_r \to 0, \quad \delta \longrightarrow \infty, \quad \text{and} \quad \gamma_n > e^{-o(n\Lambda_r)},$$

we have $P(\pi_r > \gamma_n) = o(1)$.

Lemma 4.2. Let Γ_n be the event where the smallest M_0 p-values are the p-values corresponding to the M_0 significant functional predictors, with $I_0 = \{m_1, ..., m_{M_0}\}$, that is

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

Then $\lim_{n\to\infty} P(\Gamma_n) = 1$.

Theorem 4.3. Given δ in (12), and q the chosen bound of FDR in (18) or in Bonferroni corrections, assume that $\delta \to \infty$ and $q \to 0$, as $n \to \infty$, in such a way that

$$\Lambda_r \to 0, \quad q > \frac{M}{M_0}\left(\sum_{l=1}^{M} l^{-1}\right) e^{-o(n\Lambda_r)}.$$

Then, $\lim_{n\to\infty} P\left(\hat{I} = I_0\right) = 1$.

4. Real Data Application: Weather Data

The results are shown in the table 1 and 2 of the analysis from weather data in Japan. The estimated parameters of each functional predictor is shown in Table 1. Humidity and maximum temperature are selected by all methods, however, differently from group SCAD and group LASSO, the proposed procedure selected PRESS and did not select light. Atmospheric pressure is well known amount meteorologists to be related to precipitation. In a simulation of 100 bootstrap samples from the weather data, we performed variable selection using the proposed method, group SCAD and group LASSO. Table 2 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), it was the fourth most selected predictor when using the proposed procedure. However, pressure was selected most frequently by the proposed method, followed by humidity and maximum temperature.

5. Conclusions

We proposed a competitive variable selection procedure by connection between model checking and variable selection when the covariates of the model are functional and the response is scalar. We studied the consistency and asymptotic properties of the proposed method based on the FDR method (or Bonferroni). The proposed method is applied to the analysis of real data, selecting functional predictors effectively and closer to nature of the data.

References

Aneiros, G., Ferraty F., & 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.

Benjamini, Y., & and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B* **57**, 289-300.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statististics* **29**, 1165-1188.

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

Fan, J., & Li, R. (2001). Variable selection via Nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348-1360.

Table 1: Estimated parameters of weather data

Method	TEMP	PRESS	LIGHT	HUM	MAX.T	MIN.T
	0	-0.0908	0	-0.0018	-0.3305	0
	0	-0.0908	0	-0.0019	0.1762	0
T_L	0	-0.0347	0	0.0205	-0.1614	0
	0	0.0865	0	0.0278	0.2520	0
	0	-0.0630	0	-0.0591	-0.2483	0
	0	0.1764	0	0.0231	0.3556	0
	0	0	0.0009	-0.2416	-0.3631	0
	0	0	0.0004	0.0094	0.3594	0
SCAD	0	0	-0.0026	-0.0319	-0.2933	0
	0	0	0.0023	0.2064	0.1894	0
	0	0	-0.0025	-0.2743	-0.0415	0
	0	0	-0.0009	0.4421	0.2375	0
	-0.1097	0	-0.0018	0.1817	-0.1488	0
	0.1524	0	0.0019	-0.8557	0.1392	0
LASSO	-0.1521	0	-0.0034	0.9654	-0.1067	0
	0.1471	0	0.0030	-0.7670	0.0637	0
	-0.1313	0	-0.0030	0.4555	-0.0103	0
	0.1044	0	0.0020	0.0983	0.0964	0

Table 2: 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

Gertheiss, J., Maity, A., & Staicu, A.M. (2013). Variable Selection in Generalized Functional Linear Models. *Stat* **2**, 86-101.

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

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

Ramsay, J.O., & Silverman, B. W. (2005). *Functional Data Analysis*, 2nd ed. Springer-Verlag, New York.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58**, 267-288.

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