Journal of Multivariate Analysis 146 (2016) 63–71

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

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.

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).

http://dx.doi.org/10.1016/j.jmva.2015.06.007
0047-259X/© 2016 Published by Elsevier Inc.

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 $\{(y_i, \boldsymbol{x}_i(\mathbf{t})) : \mathbf{t} \in \boldsymbol{\mathcal{T}},\ i = 1, \ldots, n\}$, where y_i is a scalar response, $\boldsymbol{x}_i(\mathbf{t}) = (x_{i1}(t_1), \ldots, x_{iM}(t_M))$ are functional predictors and $\boldsymbol{\mathcal{T}} = \mathcal{T}_1 \times \cdots \times \mathcal{T}_M$. Each $\mathcal{T}_m,\ m = 1, \ldots, M$, is a compact set in $\mathbb{R}$ where the mth 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:

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

where $\boldsymbol{W}_{im} = (\omega_{im1}, \ldots, \omega_{imp_m})^T$ are the vectors of coefficients and $\boldsymbol{\phi}_m(t_m) = (\phi_{m1}(t_m), \ldots, \phi_{mp_m}(t_m))^T$ are vectors of B-Splines basis functions. The basis functions and the p_m coefficients in (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

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

where β_0 is a constant, $\varepsilon_i,\ i = 1, \ldots, n$ 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

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

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

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

or in matrix form $\mathbf{Y} = \mathbf{Zb} + \boldsymbol{\epsilon}$, where $\boldsymbol{Z}_i = (1, \boldsymbol{W}_{i1}^T \boldsymbol{J}_{\boldsymbol{\phi_1}}, \ldots, \boldsymbol{W}_{iM}^T \boldsymbol{J}_{\boldsymbol{\phi_M}})^T$, $\boldsymbol{b} = (\beta_0, \boldsymbol{b}_1^T, \ldots, \boldsymbol{b}_M^T)^T$, $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)^T$, $\boldsymbol{J}_{\boldsymbol{\phi_m}} = \int_{\mathcal{T}_m} \boldsymbol{\phi}_m(t_m)\boldsymbol{\phi}_m^T(t_m)dt_m$ 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 $\boldsymbol{J}_{\boldsymbol{\phi_m}}$ can be easily computed using the procedure in [21].