proceedings

Extended Abstracts

Sparse Semi-Functional Partial Linear Single-Index Regression [†]

Silvia Novo [1,*], Germán Aneiros [1] and Philippe Vieu [2]

[1] MODES Research Group, CITIC, Universidade da Coruña, 15071 A Coruña, Spain; german.aneiros@udc.es
[2] Institut de Mathématiques, Université Paul Sabatier, 31062 Toulouse, France; philippe.vieu@math.univ-toulouse.fr
[*] Correspondence: s.novo@udc.es; Tel.: +34-981-167-000

[†] Presented at the XoveTIC Congress. A Coruña, Spain, 27–28 September 2018.

Published: 17 September 2018

Abstract: The variable selection problem is studied in the sparse semi-functional partial linear model, with single-index type influence of the functional covariate in the response. The penalized least squares procedure is employed for this task. Some properties of the resultant estimators are derived: the existence (and rate of convergence) of a consistent estimator for the parameters in the linear part and an oracle property for the variable selection method. Finally, a real data application illustrates the good performance of our procedure.

Keywords: functional data analysis; variable selection; sparse model; dimension reduction; functional single-index model; semiparametric model

1. Introduction

In many real problems, to predict the value of a random variable, observations of many other variables are available. However, in many cases, it is unknown which of them (very few) have a real influence in the response. In this practical framework, we need procedures able to select the relevant variables to avoid high-dimensionality problems. Reducing the complexity of the model becomes even more crucial when regression involves a functional variable too (data are functions, images…). Therefore, the main goal is the simplification of the model, which makes easier both its estimation and interpretation, without losing its predictive efficiency.

These practical problems have motived the peak of semiparametric models in the functional regression, together with the variable selection procedures. In [1] the penalized least squares method for estimation and variable selection is studied for the partial linear model with functional covariate. In this model, the real variables have a linear effect (involving interpretable coefficients that are the parameters) in the response, while the infinite-dimensional covariate has a nonlinear (nonparametric) influence. However, in real data applications, it would be interesting having parameters related to the functional variable to derive practical interpretations. This is one of the advantages of the semi-functional partial linear single-index model (SFPLSIM): the real covariates also affect in a linear way to the response, but the infinite-dimensional covariate influences it trough a projection in an unknown direction, after applying a nonlinear link function. This direction of projection behaves like a function-parameter that could have interesting interpretations. Some theoretical properties related to the nonparametric estimation of the functional single-index model are given in [2]. In this paper, we will study the sparse SFPLSIM, focusing in the variable selection problem. For this purpose, we will use the penalized least squares procedure for estimating the parameters of the lineal components and, simultaneously, selecting the relevant covariates. The properties of the estimators will be analysed from a theoretical point of view: we will set its convergence rates and the consistency for selecting the model. These results will be illustrated through a real data application.

2. The Model

The SFPLSIM is defined by the relationship

$$Y_i = X_{i1}\beta_{01} + \cdots + X_{ip_n}\beta_{0p_n} + m\left(\langle\theta_0, \mathcal{X}_i\rangle\right) + \varepsilon_i,\ \forall i = 1,\ldots,n, \tag{1}$$

where Y_i denotes a scalar response, $X_{i1},\ldots,X_{ip_n}$ are random covariates taking values in $\mathbb{R}$ and $\mathcal{X}_i$ is a functional random covariate valued in a separable Hilbert space $\mathcal{H}$ with inner product $\langle\cdot,\cdot\rangle$. $\boldsymbol{\beta}_0 = (\beta_{01},\ldots,\beta_{0p_n})^\top \in \mathbb{R}^{p_n}$, $\theta_0 \in \mathcal{H}$ and $m(\cdot)$ are a vector of unknown real parameters, an unknown functional direction and an unknown smooth real-valued function, respectively. Finally, ε_i is the random error, which verifies $\mathbb{E}\left(\varepsilon_i | X_{i1},\ldots,X_{ip_n},\mathcal{X}_i\right) = 0$.

3. The Penalized Least-Squares Estimators

For the purpose of simultaneously estimating β-parameters and selecting relevant X-covariates in the SFPLSIM (1), we will apply the penalized least-squares approach. For that, in a first step we transform the SFPLSIM in a linear model by extracting from Y_i and X_{ij} $(j = 1,\ldots,p_n)$ the effect of the functional covariate $\mathcal{X}_i$ when is projected on the direction θ_0. Specifically, denoting by $\boldsymbol{X}_i = \left(X_{i1}, X_{i2},\ldots,X_{ip_n}\right)^\top$, $\boldsymbol{X} = (\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)^\top$ and $\boldsymbol{Y} = (Y_1,\ldots,Y_n)^\top$, the fact that

$$Y_i - \mathbb{E}\left(Y_i | \langle\theta_0,\mathcal{X}_i\rangle\right) = \left(\boldsymbol{X}_i - \mathbb{E}\left(\boldsymbol{X}_i | \langle\theta_0,\mathcal{X}_i\rangle\right)\right)^\top \boldsymbol{\beta}_0 + \varepsilon_i,\ \forall i = 1,\ldots,n, \tag{2}$$

allows to consider the following approximate linear model (see Appendix A for understanding the notation):

$$\widetilde{\boldsymbol{Y}}_{\theta_0} \approx \widetilde{\boldsymbol{X}}_{\theta_0}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}, \tag{3}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1,\ldots,\varepsilon_n)^\top$. Then, in a second step, the penalized least-squares approach is applied to model (3). Specifically, $\boldsymbol{\beta}_0$ and θ_0 are estimated by considering a minimizer, $(\widehat{\boldsymbol{\beta}}_0, \widehat{\theta}_0)$, of the penalized profile least-squares function

$$\mathcal{Q}\left(\boldsymbol{\beta},\theta\right) = \frac{1}{2}\left(\widetilde{\boldsymbol{Y}}_\theta - \widetilde{\boldsymbol{X}}_\theta\boldsymbol{\beta}\right)^\top\left(\widetilde{\boldsymbol{Y}}_\theta - \widetilde{\boldsymbol{X}}_\theta\boldsymbol{\beta}\right) + n\sum_{j=1}^{p_n}\mathcal{P}_{\lambda_{jn}}\left(|\beta_j|\right),$$

where $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_{p_n})^\top$, $\mathcal{P}_{\lambda_{jn}}(\cdot)$ is a penalty function and $\lambda_{j_n} > 0$ is a tuning parameter. Note that, simultaneously to the parameter estimation, the previous procedure can be considered as a variable selection method: if $\widehat{\boldsymbol{\beta}}_{0j}$ is a non-null component of $\widehat{\boldsymbol{\beta}}_0$, then X_j is selected as an influential variable.

From now on, we will denote $J_n = \{1,\ldots,p_n\}$ and $S_n \subset J_n$ such that $\beta_{0j} \neq 0$ for $j \in S_n$ and $\beta_{0j} = 0$ for $j \in S_n^c = J_n/S_n$. In addition s_n will mean card(S_n) and we will assume that $S_n = \{1,\ldots,s_n\}$.

4. Asymptotic Theory

In this paper, the existence of the penalized estimator is established as well as the corresponding rates of convergence. In particular, under some assumptions, we proved that there exists a local minimizer $\left(\widehat{\boldsymbol{\beta}}_0, \widehat{\theta}_0\right)$ of $\mathcal{Q}\left(\boldsymbol{\beta},\theta\right)$ such that

$$\left\|\widehat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0\right\| = O_p\left(\sqrt{s_n}\left(n^{-1/2} + \delta_n\right)\right) \text{ where } \delta_n = \max_{j\in S_n}\left\{\left|\mathcal{P}'_{\lambda_{jn}}\left(|\beta_{0j}|\right)\right|\right\}. \tag{4}$$

Furthermore, the selected set of variables, $\widehat{S}_n = \{j \in J_n; \widehat{\beta}_{0j} \neq 0\}$, works as well (at least asymptotically) as it would do if the true set of relevant variables S_n was known. Specifically, $\mathbb{P}(\widehat{S}_n = S_n) \to 1$ as $n \to \infty$.

An application to real data is included, which shows the good performance of the presented method in terms of error of prediction.

Funding: The authors acknowledge partial support by MINECO grants MTM2014-52876-R and MTM2017-82724-R (EU ERDF support included). Additionally, financial support from the Xunta de Galicia (Centro Singular de Investigación de Galicia accreditation ED431G/01 2016-2019 and Grupos de Referencia Competitiva ED431C2016-015) and the European Union (European Regional Development Fund - ERDF), is gratefully acknowledged. The first author also thanks the financial support from the Xunta de Galicia and the European Union (European Social Fund - ESF), the reference of which is ED481A-2018/191.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

SFPLSIM	Semi-functional partial linear single index model

Appendix A. Notation

For any $(n \times q)$-matrix $\boldsymbol{A}$ $(q \geq 1)$, if $\boldsymbol{I}$ is the $(n \times n)$-identity-matrix, we denote

$$\widetilde{\boldsymbol{A}}_{\theta} = (\boldsymbol{I} - \boldsymbol{W}_{h,\theta})\,\boldsymbol{A}, \text{ where } \boldsymbol{W}_{h,\theta} = \left(w_{n,h,\theta}(\mathcal{X}_i, \mathcal{X}_j)\right)_{i,j},$$

with $w_{n,h,\theta}(\cdot,\cdot)$ being the weight function

$$w_{n,h,\theta}(\chi, \mathcal{X}_i) = \frac{K\left(d_{\theta}\left(\chi, \mathcal{X}_i\right)/h\right)}{\sum_{j=1}^{n} K\left(d_{\theta}\left(\chi, \mathcal{X}_j\right)/h\right)},$$

where $K : \mathbb{R}^+ \to \mathbb{R}^+$ is a kernel function, $h > 0$ is a smoothing parameter and, for $\theta \in \mathcal{H}$, $d_{\theta}(\cdot,\cdot)$ is the semimetric defined as

$$d_{\theta}\left(\chi, \chi'\right) = \left|\langle \theta, \chi - \chi' \rangle\right|, \forall \chi, \chi' \in \mathcal{H}.$$

References

1. Aneiros, G.; Ferraty, F.; Vieu, P. Variable selection in partial linear regression with functional covariate. *Statistics* **2015**, *49*, 1322–1347, doi:10.1080/02331888.2014.998675.
2. Novo, S.; Aneiros, G.; Vieu, P. Automatic and location-adaptive estimation in functional single-index regression. **2018**, in press.