proceedings

MDPI

Proceedings

Fast Algorithm for Impact Point Selection in Semiparametric Functional Models †

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

[1] MODES Research Group, CITIC, Universidade da Coruña, 15071 A Coruña, Spain
[2] Institut de Mathématiques, Université Paul Sabatier, 31062 Toulouse, France
* Correspondence: s.novo@udc.es; Tel.: +34-981-167-000-1301
† Presented at the 2nd XoveTIC Conference, A Coruña, Spain, 5–6 September 2019.

Published: 31 July 2019

Abstract: A new sparse semiparametric functional model is proposed, which tries to incorporate the influence of two functional variables in a scalar response in a quite simple and interpretable way. One of the functional variables is included trough a single-index structure and the other one linearly, but trough the high-dimensional vector of its discretized observations. For this model, a new algorithm for impact point selection in the linear part and for the model estimation is proposed. This procedure is based on the functional origin of the linear covariates. Some asymptotic results will ensure the good performance of the method. The computational efficiency of the algorithm, without loss of predictive power, will be showed trough a simulation study and a real data application, by comparing its results with those obtained trough the standard PLS method.

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

1. Introduction

In the BIG data era, it is more and more frequent having observations of variables measured in a continuous support (data are curves, images). This informative richness provided by the functional variables makes very usual found them in regression problems. In many situations, we have a scalar variable of interest and we want to know which points of a functional variable are the most influential (points of impact) on this scalar variable (see [1]). The problem is that the functional variables usually are observed in many points and standard variable selection methods in the multidimensional context can provide inadequate results. On the one hand, these procedures are affected by the dependence between observations, which in this case is directly derived from its functional origin. On the other hand, the great quantity of observations makes difficult obtaining results in reasonable amount of time.

In this work, we are going to focus on a regression model with scalar response which incorporates the influence of two functional variables: one of them is included trough a single-index type structure (see for details [2,3]) and the other one, linearly, but trough a high-dimensional vector formed by its discretized observations (see [1,4] for details and motivation of this structure). In this way we obtain a very flexible model, which combines interpretable estimations with dimension reduction. For this model, the so-called Multi-functional Partial Linear Single-Index Model (MFPLSIM), we work in the framework where we have a very big number of linear covariates but only a few of them have a real influence in the response (sparse context). Accordingly, we are going to develop an efficient algorithm for impact point selection in the linear part and for the estimation of the model (the Fast Algorithm for Sparse Semiparametric Multifunctional Regression- FASSMR), which takes advantage of the functional origin of these scalar variables included in the linear part. The good practical behaviour of the proposed methodology will be showed trough a simulation study and a real data application. In both cases,

we will show its computational efficiency, without loss of predictive power, by comparing its results with the standard PLS procedure. Furthermore, some asymptotic results will support theoretically the FASSMR.

2. The Model

The MFPLSIM is defined by the relationship

$$Y = \sum_{j=1}^{p_n} \beta_{0j}\zeta(t_j) + m\left(\langle\theta_0, \mathcal{X}\rangle\right) + \varepsilon, \tag{1}$$

where Y is a real random response, $\mathcal{X}$ denote a random curve defined on some Hilbert space $\mathcal{H}$ with inner product $\langle\cdot,\cdot\rangle$ and ζ denote another random curve defined on some interval $[c,d]$. The curve ζ is observed in the points $c \leq t_1 < \cdots < t_{p_n} \leq d$ and denote by $\zeta(t_j)$, $j = 1,\ldots,p_n$, its discretized observations; $\left(\beta_{01},\ldots,\beta_{0p_n}\right)^\top$ is a vector of unknown coefficients, m is an unknown link function and θ_0 denotes an unknown curve in $\mathcal{H}$. Finally, ε is the random error, which verifies $\mathbb{E}\left(\varepsilon|\zeta(t_1),\ldots,\zeta(t_{p_n}),\mathcal{X}\right) = 0$. In model (1), we assume that only a few points of the curve ζ have an effect on the response Y. Then, we denote $S_n = \{j = 1,\ldots,p_n, \quad \text{such that } \beta_{0j} \neq 0\}$, and it is verified that $\sharp S_n = s_n = o(p_n)$.

3. The FASSMR

Our procedure is based on the fact that the variables $\zeta(t_j)$, $j = 1,\ldots,p_n$, come from the discretization of the functional variable ζ. Then, when t_j is close from t_k, the two corresponding variables $\zeta(t_j)$ and $\zeta(t_k)$ roughly contain the same information on the response. As consequence, some variables can be discarded before applying the variable selection procedure.

For presenting the FASSMR, let us assume that we have a statistical sample of size n, $\{(\zeta_i,\mathcal{X}_i,Y_i), \quad i = 1,\ldots,n\}$ i.i.d. as $(\zeta,\mathcal{X},Y)$. We will consider, without lost of generality, that p_n can be factorized in the following way: $p_n = q_n w_n$ with q_n and w_n integers. The previous considerations allow us present the following set of variables

$$\mathcal{R}_n^{\mathbf{1}} = \{\zeta(t_k^{\mathbf{1}}) = \zeta(t_{[(2k-1)q_n/2]}), \quad k = 1,\ldots,w_n\},$$

where $[z]$ denotes the smallest integer not less than $z \in \mathbb{R}$. Note that the correlation between consecutive variables inside of $\mathcal{R}_n^{\mathbf{1}}$ is much less important than in the whole set of p_n initial linear covariates. As consequence, the variable selection procedure will be carried out in variables belonging to $\mathcal{R}_n^{\mathbf{1}}$. In other words, we will considerer the following model with only w_n linear covariates

$$Y_i = \sum_{k=1}^{w_n} \beta_{0k}^{\mathbf{1}}\zeta_i(t_k^{\mathbf{1}}) + m^{\mathbf{1}}\left(\left\langle\theta_0^{\mathbf{1}},\mathcal{X}_i\right\rangle\right) + \varepsilon_i^{\mathbf{1}}. \tag{2}$$

Then, variable selection task can be developed following the standard procedure described in [5] and detailed in [6], which is based on transforming the model (2) into a linear one and applying the PLS procedure. We denote by $(\widehat{\boldsymbol{\beta}}_0^{\mathbf{1}},\widehat{\theta}_0^{\mathbf{1}})$, the estimation of the parameters of model (2) where $\widehat{\boldsymbol{\beta}}_0^{\mathbf{1}} = (\widehat{\beta}_{01}^{\mathbf{1}},\ldots,\widehat{\beta}_{0w_n}^{\mathbf{1}})^\top$. Then, $\zeta(t_k^{\mathbf{1}})$ is selected in $\mathcal{R}_n^{\mathbf{1}}$ if and only if $\widehat{\beta}_{0k}^{\mathbf{1}} \neq 0$.

Considering the whole set of initial of p_n linear covariates, that is, returning to model (1), a variable $\zeta(t_j) \in \{\zeta(t_1),\ldots,\zeta(t_{p_n})\}$ is selected if and only if it belongs to $\mathcal{R}_n^{\mathbf{1}}$ and its estimated coefficient, which can be denoted by $\widehat{\beta}_{0k_j}^{\mathbf{1}}$, is non null. Then, $\widehat{S}_n = \{j = 1,\ldots,p_n, \quad \text{such that } \zeta(t_j) = \zeta(t_{k_j}^{\mathbf{1}}) \in \mathcal{R}_n^{\mathbf{1}}$ and $\widehat{\beta}_{0k_j}^{\mathbf{1}} \neq 0\}$ and $\widehat{\beta}_{0j} = \widehat{\beta}_{0k_j}^{\mathbf{1}}$ if $j \in \widehat{\mathcal{S}}_n$ and $\widehat{\beta}_{0j} = 0$ otherwise. Finally, $\widehat{\theta}_0 = \widehat{\theta}_0^{\mathbf{1}}$ and an estimator of the function $m_{\theta_0}(\cdot) \equiv m(\langle\theta_0,\chi\rangle)$, denoted by $\widehat{m}_{\widehat{\theta}_0}(\chi)$, can be obtained by smoothing the residuals from the parametric fit (see Appendix A).

4. Theory, Simulation and Real Data Application Conclusions

The good behaviour of the proposed algorithm will be ensured theoretically. Furthermore, from the simulation study it can be seen that the FASSMR allows us to obtain the variable selection and estimation of model (1) in a reasonable amount of time, even for very big values of p_n. As will be derived from the simulation study, the developed algorithm clearly overpasses standard PLS procedure in terms of computational time without loss in prediction power. A real data application will also illustrate the flexibility and applicability of model (1) together with the FASSMR estimation.

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:

FASSMR	Fast Algorithm for Sparse Semiparametric Multi-functional Regression
i.i.d.	Independent and identically distributed
MFPLSIM	Multi-functional Partial Linear Single-Index Model
PLS	Penalized Least Squares

Appendix A

Denoting by $\widehat{\boldsymbol{\beta}}_0$ the vector of estimated parameters,

$$\widehat{m}_{\widehat{\theta}_0}(\chi) \equiv \widehat{m}\left(\left\langle \widehat{\theta}_0, \chi \right\rangle\right) = \frac{\sum_{i=1}^{n}\left(Y_i - \boldsymbol{\zeta}_i^{\top}\widehat{\boldsymbol{\beta}}_0\right) K\left(d_{\widehat{\theta}_0}\left(\chi, \mathcal{X}_i\right)/h\right)}{\sum_{i=1}^{n} K\left(d_{\widehat{\theta}_0}\left(\chi, \mathcal{X}_i\right)/h\right)},$$

where we have denoted $\boldsymbol{\zeta}_i = \left(\zeta_i(t_1), \ldots, \zeta_i(t_{p_n})\right)^{\top}$, $h > 0$ is a bandwidth, K is a kernel and, for any $\theta \in \mathcal{H}$, $d_\theta(\cdot,\cdot)$ is the semimetric defined as $d_\theta(\chi, \chi') = |\langle \theta, \chi - \chi' \rangle|$ for each $\chi, \chi' \in \mathcal{H}$.

References

1. Aneiros, G.; Vieu, P. Variable selection in infinite-dimensional problems. *Stat. Probab. Lett.* **2014**, *94*, 12–20.
2. Ait-Saïdi, A.; Ferraty, F.; Kassa, R.; Vieu, P. Cross-Validated Estimations in the Single-Functional Index Model. *Statistics* **2008**, *42*, 475–494.
3. Novo, S.; Aneiros, G.; Vieu, P. Automatic and location-adaptive estimation in functional single-index regression. *J. Nonparametric Stat.* **2019**, *31*, 364–392.
4. Aneiros, G.; Vieu, P. Partial linear modelling with multi-functional covariates. *Comput. Stat.* **2015**, *30*, 647–671.
5. Novo, S.; Aneiros, G.; Vieu, P. Sparse Semi-Functional Partial Linear Single-Index Regression. *Proceedings* **2018**, *2*, 1190.
6. Novo, S.; Aneiros, G.; Vieu, P. Sparse semiparametric regression when predictors are mixture of functional and high-dimensional variables. preprint.