ARTICLE IN PRESS

Statistics and Probability Letters ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

On dimension reduction models for functional data

Philippe Vieu

Institut de Mathématiques de Toulouse, UMR5219, Université Paul Sabatier, F-31062 Toulouse Cedex 9, France

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Functional data
Dimension reduction
Semi-parametrics
Sparse regression

ABSTRACT

This contribution is part of the recent links between Functional Data and Big Data communities. A selected survey highlights how earlier ideas in high dimensional problems can be adapted in functional setting.

1. Functional data and big data: a short introduction

In modern applied sciences one observes variables whose complexity is each day higher and higher. In multivariate analysis, the observed variable is a vector $X = (X^1, \ldots, X^p)$ and the dataset is usually called "big dataset" if the dimension p is "much higher" than the sample size itself n (this is denoted by $p \gg n$). In curves analysis the statistical variable is a curve $\{\chi = \chi(t), t \in I\}$, and more generally in Functional Data Analysis (FDA) the variable is an object χ taking values in some infinite dimensional space. In this sense, a functional dataset is also a "big dataset" since the dimension is infinite. In some part of the literature (see Marron, 2014; Marron and Alonso, 2014) the analysis of complex infinite dimensional objects is also called Data Oriented Object Analysis. In practice a functional element $\{\chi = \chi(t), t \in I\}$ is observed on a finite grid $t^1, \ldots, t^p$ in such a way that it can also be seen as a specific high dimensional vector $X = (X^1, \ldots, X^p) = (\chi(t^1), \ldots, \chi(t^p))$. Despite of this apparently common structure, the underlying continuity feature of the curve makes the methodologies involving the discretized vector $(\chi(t^1), \ldots, \chi(t^p))$ somewhat different from those for standard vectors X. This is probably the reason why during a long time both areas, namely FDA and High Dimensional Statistics (HDS), grew independently one from each other. This contribution aims to strengthen these links between FDA and HDS by discussing two kinds of methodologies for functional regression putting down roots in earlier literature in high multivariate data analysis.

FDA has been popularized twenty years ago by J. Ramsay and B. Silverman's book (see Ramsay and Silverman, 2002, 2005). Nowadays many statistical questions arising before for multivariate samples have been addressed in the functional framework, including time series analysis (see Bosq, 2000), non-parametric statistics (see Ferraty and Vieu, 2006), variance analysis (see Zhang, 2013), …. A wider scope of the literature can be found in recent monographies (see e.g. Shi and Choi, 2011; Horváth and Kokoszka, 2012; Hsing and Eubank, 2015) or survey papers (see e.g. Geenens, 2011; Cuevas, 2014; Jacques and Preda, 2014; Müller, 2016; Wang et al., 2016; Reiss et al., 2017; Kokoszka et al., 2017 or Nagy, 2017). Any methodology intending to deal with functional data has to front with the question of the dimensionality of the data (see discussion Section 2). For seakness of shortness we restrict our purpose to regression (see Section 3) and we discuss dimensional reduction regression models along Section 4 which is the main part of this paper. Again for size of size reasons, we pay greatest attention to two kinds of models combining exploratory and explanatory interests, namely semi-parametric models (see Section 4.1) and sparse models (see Section 4.2). While the main point is on links between FDA and HDS, this contribution is also the opportunity for a short and selected review on FDA but without so much attention to applications. A sample of discussions being oriented more towards applications includes (Ramsay and Silverman, 2002; González Manteiga and Vieu, 2007; Valderrama, 2007; González Manteiga and Vieu, 2012).

E-mail address: philippe.vieu@math.univ-toulouse.fr.

https://doi.org/10.1016/j.spl.2018.02.032
0167-7152/

2. The impact of the dimension on the concentration of variables

The question of the dimension is characterized by the fact that a sample of data is more and more sparse as its dimension increases, making the construction of statistical procedures harder and harder. This is confirmed by having a look on the probability distribution of the variable X. Let $\epsilon > 0$ fixed. If X is real valued, then its distribution is characterized by the function $F_X(\epsilon) = P(X \leq \epsilon)$, and as long as it is continuous with respect to Lebesgue measure one has:

$$P(X \in]x_0 - \epsilon x_0 + \epsilon[) = F_X(x_0 + \epsilon) - F_X(x_0 + \epsilon) \sim C\epsilon.$$

In multi-dimensional setting X takes values in $\mathbb{R}^p$, and this becomes:

$$P(X \in \mathcal{B}(x_0, \epsilon)) \sim C\epsilon^p.$$

So, the concentration is exponentially decreasing with the dimension p. This has been pointed out along the eighties as being of particularly bad effects on nonparametric smoothing techniques even for very small values of p (see Stone, 1982). It is admitted that nonparametrics is in most situations out of purpose as long as $p > 4$ or 5! In big data setting (when $p \gg n$) this is even more dramatical and does not have only impacts on nonparametrics but on any statistical procedure!

In the functional framework X takes values in an infinite dimensional space $\mathcal{E}$ and there is a wide literature (see e.g. Li and Shao, 2001 or Kirichenko and Nikitin, 2014) showing that for specific infinite dimensional processes (and some specific metric topologies) one has exponential-type small ball probability:

$$P(X \in \mathcal{B}(x_0, \epsilon)) \sim C_1 e^{-\frac{1}{\epsilon^{C_2}} \log(\frac{1}{\epsilon})^{C_3}},$$

supporting the idea that dimensional effects are even worst.

3. Functional regression

In functional regression, the infinite dimensional variable χ has to be used to explain and/or predict a response Y. The basic model can be written as

$$Y = m(\chi) + error, \tag{3.1}$$

and the flexibility of the model depends on the generality of the mathematical conditions assumed on m. Recent survey papers on functional regression include (Morris, 2015; Reiss et al., 2017; Greven and Scheipl, 2017). In nonparametric models, only smoothness conditions are made on m, and the problem is to estimate a non linear operator acting on the functional space $\mathcal{E}$. Earlier advances on nonparametric functional regression are provided in Ferraty and Vieu (2006) when kernel smoothing techniques are used (see Kara-Zaitri et al., 2017a for recent advances), while the literature covers now various alternative smoothers such as kNN (see Kara-Zaitri et al., 2017b) or local linear regressors (see Demongeot et al., 2017). In an other hand, a parametric model makes stronger assumptions on m changing the problem into the simpler one of estimating some element of $\mathcal{E}$. To fix the ideas, if $(\mathcal{E}, \langle ; \rangle)$ is an Hilbert space the linear model has the simple form

$$m(.) = \langle .; \theta \rangle, \text{ for some } \theta \in \mathcal{E}.$$

Earlier advances can be found in Ramsay and Silverman (2005) and a recent overview is provided in Febrero et al. (2017). In the curves setting where $\mathcal{E} = L^2([0, 1])$, this model becomes

$$Y = \int_0^1 X(t)\theta(t)dt + error.$$

From one side the nonparametric approach is much more flexible than the linear one, but in an other hand the parametric approach has the advantage of being less impacted by the dimensionality since the target (namely θ) is of low dimension than the operator m. For instance, when $\mathcal{E} = L^2([0, 1])$ the target θ is a 1-dimensional object. Moreover, the linear modelling provides an easily representable output θ. The aim of dimensionality reduction models is to balance flexibility and dimensionality sensitivity in order to capture all advantages of linear and nonparametric approaches.

4. Dimension reduction models for functional regression

A dimensionality reduction model imposes assumptions on the unknown regression operator allowing to characterize it by means of one (or more) new operator(s) acting on new space(s) being of low dimension. For reasons of shortness we discuss only to two specific dimension reduction ideas (semi-parametric and sparse ones). Other reduction dimension models based on additive have been developed in many directions for FDA (see e.g. Müller and Yao, 2008; Ferraty and Vieu, 2009; Müller et al., 2013).

Please cite this article in press as: Vieu P., On dimension reduction models for functional data. Statistics and Probability Letters (2018), https://doi.org/10.1016/j.spl.2018.02.032.