Statistics and Probability Letters 136 (2018) 134–138

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 21 February 2018

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 $X = (X^1, \dots, X^p)$ 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 $t^1, \dots, t^p$ in such a way that it can also be seen as a specific high dimensional vector $X = (X^1, \dots, X^p) = (\chi(t^1), \dots, \chi(t^p))$. Despite of this apparently common structure, the underlying continuity feature of the curve makes the methodologies involving the discretized vector $(\chi(t^1), \dots, \chi(t^p))$ 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/© 2018 Elsevier B.V. All rights reserved.

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).

4.1. *Semi-parametric functional regression*

– What is a semi-parametric model? Semi-parametrics has been widely studied in the multivariate setting (see e.g. Härdle et al., 2004; Sperlich et al., 2006; Horowitz, 2009), but very few advances have been developed in functional one. Roughly speaking, a semi-parametric functional regression model consists in assuming that the unknown operator m in the model (3.1) can be expressed by means of one (or more) term(s) g_j each being of the form $g_j = (\phi_j(\chi, \theta_j))$, where the known operators ϕ_j take values in spaces $\mathcal{E}_j$ such that $dim(\mathcal{E}_j) \ll dim(\mathcal{E})$. The θ_j are unknown elements of $\mathcal{E}$ (this is the parametric feature of the model) and the g_j are smooth operators acting on the low dimension spaces $\mathcal{E}_j$ (this is the nonparametric feature of the model). The target is no more the high dimensional object m but only the lower dimensional ones g_j and θ_j. Hence one intents to reduce the low concentration problems discussed in Section 2, without suffering from the same lack of flexibility as linear modelling (see Section 3).

– A simple semi-parametric model: the single functional index model. This model is adapted from ideas in multivariate analysis (see Härdle et al., 1993) for earlier advances and Chapter 6 in Härdle et al. (2004) for a general discussion). If the functional variable χ is valued in some Hilbert space $(\mathcal{E}, \langle ; \rangle)$, it consists in assuming that χ acts on the response Y through some single functional direction $\theta \in \mathcal{E}$. This leads to the so-called Single Functional Index Model (SFIM), as introduced in Ferraty et al. (2003):

$$Y = m(\chi) + error = g(\langle \chi, \theta \rangle) + error. \tag{4.1}$$

In the usual curves setting where $\mathcal{E} = L^2([0, 1])$, the SFIM becomes

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

It is a reduction dimension statistical model since the question is no more to estimate the operator m but rather the function g (which acts on a 1-dimensional space) and the direction θ. When $\mathcal{E} = L^2([0, 1])$, both elements g and θ in the SFIM are standard real functions. From a predictive point of view, the model is insensitive to the dimensionality and the operator g can be estimated with 1-dimensional rates of convergence (the most recent results include Goia and Vieu, 2015; Ma, 2016), while from an exploratory point of view the model provides interpretable outputs (some real data analysis are presented in Chen et al., 2011 and Goia and Vieu, 2015).

– Complementary bibliography. Other semi-parametric multivariate ideas can be adapted to functional data (see Goia and Vieu, 2014 for a survey), including Projection Pursuit Regression (Ferraty et al., 2013; Chen et al., 2011), Slice Inverse Regression (see e.g. Ferré and Yao, 2005 and Zhang et al., 2017), Partial Linear Regression (see e.g. Aneiros and Vieu, 2006, 2015; Feng and Xue, 2016; Chiou et al., 2016 and Boente and Vahnovan, 2017) or Spatial regression models (see e.g. Sangalli et al., 2013 and Ettinger et al., 2016).

4.2. *Sparse functional regression*

– Sparse model in functional setting. A sparse regression model is based on the idea that only some (maybe just a small) area of the functional variable χ acts on the response Y. So, by looking at the discretized version $(\chi(t^1), \dots, \chi(t^p))$ of the functional element χ, the question turns to be a variable selection one. Variable selection in high dimensional regression has been widely studied, but despite of very similar formulations, the essence of the problem is quite different here : While in multivariate analysis an increasing number of variables means that new information have been collected, in functional data analysis it means that the same information has been collected in a more precise way. When $\mathcal{E} = L^2([0, 1])$, the dimension p is linked with the fineness of the grid on which the curves are observed.

– The sparse nonparametric functional regression model. It consists in writing

$$Y = \sum_{i=1}^{p} m^j(\chi(t^j)) + error, \tag{4.2}$$

and in introducing a sparsity set $\mathcal{S} = \{k, m^k \neq 0\}$ (keeping in mind that we should have $card(\mathcal{S}) \ll p$). The nonparametric feature of the model consists in simple regularity conditions on the unknown components m^j. It is clearly a dimensional reduction model since the targets m^j are elements acting on variables whose 1-dimensional type concentration (see Section 2). The asymptotics in Aneiros and Vieu (2017) shows that the m^j can be estimated at the usual 1-dimensional rate of convergence. In addition, the estimation of the set $\mathcal{S}$ gives direct exploratory informations on which part of the functional covariate has effect on the response (see Aneiros and Vieu, 2017 for a real data analysis).

– Additional literature. The recent literature on sparse functional modelling involves not only nonparametric but also linear regression (Aneiros and Vieu, 2014; Collazos et al., 2016), partial linear regression (Aneiros and Vieu, 2015), as well as discrimination and/or clustering (Fraiman et al., 2016; Berrendero et al., 2016b, a; Floriello and Vitelli, 2017). It is worth being noted that the word *sparse* maybe used in many other sense in the functional literature (see Aneiros and Vieu, 2016 for a discussion).

Acknowledgements

Developing links between both communities of Functional Data Analysis and Big Data Analysis has been of actuality in two recent meetings (see Bongiorno et al., 2014; Aneiros et al., 2017) and was pointed in a few recent works (see for instance Goia and Vieu, 2016 or Ahmed, 2017). This Special Issue takes it place in this process and the author wishes to thanks gratefully its Editors for offering the opportunity of presenting this discussion.

References

Ahmed, S.E., 2017. Big and Complex Data Analysis. Methodologies and Applications. In: Contributions to Statistics, Springer.

Aneiros, G., Bongiorno, E., Cao, R., Vieu, P., 2017. An introduction to the 4th edition of the international workshop on functional and operatorial statistics. In: Functional Statistics and Related Fields. In: Contributions to Statistics, Springer, pp. 1–6.

Aneiros, G., Vieu, P., 2006. Semi-functional partial linear regression. Statist. Probab. Lett. 76 (11), 1102–1110.

Aneiros, G., Vieu, P., 2014. Variable selection in infinite-dimensional problems. Statist. Probab. Lett. 94, 12–20.

Aneiros, G., Vieu, P., 2015. Partial linear modelling with multi-functional covariates. Comput. Statist. 30 (3), 647–671.

Aneiros, G., Vieu, P., 2016. Comments on: Probability enhanced effective dimension reduction for classifying sparse functional data. TEST 25 (1), 27–32.

Aneiros, G., Vieu, P., 2017. Nonparametric model for regression with functional covariate. J. Nonparametr. Stat. 28 (4), 839–859.

Berrendero, J., Cuevas, A., Pateiro, B., 2016a. Shape classification based on interpoint distance distributions. J. Multivariate Anal. 146, 237–247.

Berrendero, J., Cuevas, A., Torrecilla, J., 2016b. Variable selection in functional data classification: a maxima-hunting proposal. Statist. Sinica 26 (2), 619–638.

Boente, G., Vahnovan, A., 2017. Robust estimators in semi-functional partial linear regression models. J. Multivariate Anal. 154, 59–84.

Bongiorno, E., Goia, A., Salinelli, E., Vieu, P., 2014. An overview of IWFOS'2014. In: Contributions in Infinite-Dimensional Statistics and Related Topics. Esculapio, Bologna, pp. 1–5.

Bosq, D., 2000. Linear Processes in Function Spaces. Theory and Applications. In: Lecture Notes in Statistics, vol. 149, Springer-Verlag, New York.

Chen, D., Hall, P., Müller, H., 2011. Single and multiple index functional regression models with nonparametric link. Ann. Statist. 39, 1720–1747.

Chiou, J., Yang, Y., Chen, Y., 2016. Multivariate functional linear regression and prediction. J. Multivariate Anal. 146, 301–312.

Collazos, J., Dias, R., Zambom, A., 2016. Consistent variable selection for functional regression models. J. Multivariate Anal. 146, 63–71.

Cuevas, A., 2014. A partial overview of the theory of statistics with functional data. J. Statist. Plann. Inference 147, 1–23.

Demongeot, J., Naceri, A., Laksaci, A., Rachdi, M., 2017. Local linear regression modelization when all variables are curves. Statist. Probab. Lett. 121, 37–44.

Ettinger, B., Perotto, S., Sangalli, L., 2016. Spatial regression models over two-dimensional manifolds. Biometrika 103 (1), 71–88.

Febrero, M., Galeano, P., Gonzalez Manteiga, W., 2017. Functional principal component regression and functional partial least-squares regression: an overview and a comparative study. Int. Statist. Rev. 85, 61–83.

Feng, S., Xue, L., 2016. Partially functional linear varying coefficient model. Statistics 50 (4), 717–732.

Ferraty, F., Goia, A., Salinelli, E., Vieu, P., 2013. Functional projection pursuit regression. TEST 22 (2), 293–320.

Ferraty, F., Peuch, A., Vieu, P., 2003. Modéle á indice fonctionnel simple. C. R. Math. Acad. Sci. Paris 336 (12), 1025–1028 (in French).

Ferraty, F., Vieu, P., 2006. Nonparametric Functional Data Analysis. Theory and Practice. Springer-Verlag, New York.

Ferraty, F., Vieu, P., 2009. Additive prediction and boosting for functional data. Comput. Statist. Data Anal. 53 (4), 1400–1413.

Ferré, L., Yao, A.F., 2005. Smoothed functional inverse regression. Statist. Sinica 15 (3), 665–683.

Floriello, D., Vitelli, V., 2017. Sparse clustering of functional data. J. Multivariate Anal. 154, 1–18.

Fraiman, R., Gimenez, Y., Svarc, M., 2016. Feature selection for functional data. J. Multivariate Anal. 146, 191–208.

Geenens, G., 2011. Curse of dimensionality and related issues in nonparametric functional regression. Stat. Surv. 5, 30–43.

Goia, A., Vieu, P., 2014. Some advances on semi-parametric functional data modelling. In: Contributions in Infinite-Dimensional Statistics and Related Topics. Esculapio, Bologna, pp. 135–140.

Goia, A., Vieu, P., 2015. A partitioned single functional index model. Comput. Statist. 30 (3), 673–692.

Goia, A., Vieu, P., 2016. An introduction to recent advances in high/infinite dimensional statistics. J. Multivariate Anal. 146, 1–6.

González Manteiga, W., Vieu, P., 2007. Statistics for functional data. Comput. Statist. Data Anal. 51 (10), 4788–4792.

González Manteiga, W., Vieu, P., 2012. Methodological richness of functional data analysis. In: Statistical Learning and Data Science. In: Comput. Sci. Data Anal. Ser., CRC Press, Boca Raton, FL, pp. 197–203.

Greven, S., Scheipl, F., 2017. A general framework for functional regression modelling. Stat. Model. 17 (1–2), 1–35.

Härdle, W., Hall, P., Ichimura, H., 1993. Optimal smoothing in single-index models. Ann. Statist. 21 (1), 157–178.

Härdle, W., Müller, M., Sperlich, S., Werwatz, A., 2004. Nonparametric and Semi-Parametric Models. In: Springer Series in Statistics, Springer-Verlag, New York.

Horowitz, J., 2009. Semi-Parametric and Nonparametric Methods in Econometrics. In: Springer Series in Statistics, Springer, New York.

Horváth, L., Kokoszka, P., 2012. Inference for Functional Data with Applications. Springer, New York.

Hsing, T., Eubank, R., 2015. Theoretical Foundations to Functional Data Analysis with an Introduction to Linear Operators. In: Wiley Series in Probability and Statistics, Chichester/John Wiley & Sons.

Jacques, J., Preda, C., 2014. Functional data clustering: a survey. Adv. Data Anal. Classif. 8 (3), 231–255.

Kara-Zaitri, L., Laksaci, A., Rachdi, M., Vieu, P., 2017a. Data-driven kNN estimation in nonparametric functional data-analysis. J. Multivariate Anal. 153, 176–188.

Kara-Zaitri, L., Laksaci, A., Rachdi, M., Vieu, P., 2017b. Uniform in bandwidth consistency for various kernel estimators involving functional data. J. Nonparametr. Stat. 29 (1), 85–107.

Kirichenko, A., Nikitin, Y., 2014. Precise small deviations in L2 of some Gaussian processes appearing in the regression context. Cent. Eur. J. Math. 12 (11), 1674–1686.

Kokoszka, P., Oja, H., Park, B., Sangalli, L., 2017. Special issue on functional data analysis. Econom. Stat. 1, 99–100.

Li, W., Shao, Q., 2001. Gaussian processes:inequalities, small ball probabilities and applications. In: Stochastic Processes: Theory and Methods. In: Handbook of Statist., vol. 19, North-Holland, Amsterdam, pp. 533–597.

Ma, S., 2016. Estimation and inference in functional single-index models. Ann. Inst. Statist. Math. 68 (1), 181–208.

Marron, J.S., 2014. Object oriented data analysis: open problems regarding manifolds. In: Contributions in Infinite-Dimensional Statistics and Related Topics. Esculapio, Bologna, pp. 185–190.

Marron, J.S., Alonso, A., 2014. Overview of object oriented data analysis. Biom. J. 56 (5), 732–753.

Morris, J.S., 2015. Functional regression. Annu. Rev. Stat. Appl. 2, 321–359.

Müller, H.-G., 2016. Peter Hall, functional data analysis and random objects. Ann. Statist. 44 (5), 1867–1887.

Müller, H.-G., Wu, Y., Yao, F., 2013. Continuously additive models for nonlinear functional regression. Biometrika 100 (3), 607–622.

Müller, H.-G., Yao, F., 2008. Functional additive models. J. Amer. Statist. Assoc. 103 (484), 1534–1544.

Nagy, S., 2017. An overview of consistency results for depth functionals. In: Functional Statistics and Related Topics. In: Springer Contributions to Statistics, Springer.

Ramsay, J.O., Silverman, B.W., 2002. Applied Functional Data Analysis. Springer, New York.

Ramsay, J., Silverman, B., 2005. Functional Data Analysis, second ed.. In: Springer Series in Statistics, Springer, New York.

Reiss, P., Goldsmith, J., Shang, H.L., Ogden, R., 2017. Methods for scalar-on-function regression. Int. Statist. Rev. 85 (2), 228–249.

Sangalli, L., Ramsay, J., Ramsay, T., 2013. Spatial spline regression models. J. R. Stat. Soc. Ser. B Stat. Methodol. 75 (4), 681–703.

Shi, J.Q., Choi, T., 2011. Gaussian Process Regression Analysis for Functional Data. CRC Press, Boca Raton, FL, p. 2011.

Sperlich, S., Härdle, W., Aydinli, G., 2006. The Art of Semi-Parametrics. Selected Papers from the Conference Held in Berlin, 2003. In: Contributions to Statistics, Physica-Verlag/Springer, Heidelberg.

Stone, C., 1982. Optimal global rates of convergence for nonparametric regression. Ann. Statist. 10 (4), 1040–1053.

Valderrama, M., 2007. An overview to modelling functional data. Comput. Statist. 22 (3), 331–334.

Wang, J.L., Chiou, J.M., Müller, H.G., 2016. Review of functional data analysis. Annu. Rev. Stat. Appl. 3 (1), 257–295.

Zhang, J., 2013. Analysis of Variance for Functional Data. In: Chapman & Hall/CRC Monographs on Statistics & Applied Probability.

Zhang, X., Wang, C., Wu, Y., 2017. Functional envelope for model-free sufficient dimension reduction. Preprint.