Journal of Multivariate Analysis 146 (2016) 1–6

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Editorial

An introduction to recent advances in high/infinite dimensional statistics

Aldo Goia [a], Philippe Vieu [b,*]

[a] Dipartimento di Studi per l'Economia e l'Impresa, Università del Piemonte Orientale, Novara, Italy

[b] Institut de Mathématiques, Université Paul Sabatier, Toulouse, France

ARTICLE INFO

Article history:
Received 25 November 2015
Available online 11 December 2015

AMS subject classifications:
62-02

Keywords:
High dimensional statistics
Infinite dimensional statistics
Functional data analysis
Selected review

ABSTRACT

The aim of this short contribution is to present the various papers composing this Special Issue on Statistics in HD spaces, by casting them into their bibliographical context through some necessarily short and selected discussion of the current literature.

1. Introduction

The idea of this Special Issue on Statistics in High Dimensional Spaces matured during the preparation of the third edition of the International Workshop on Functional and Operatorial Statistics (IWFOS) organized in Stresa (Italy) in June 2014 (see [21]). While the two first editions of this event were concentrated around infinite dimensional issues, one of the main aims of the third edition was to promote links with another currently active field of statistics, namely, problems involving high (but finite) dimensional data. Thanks to the team of the *Journal of Multivariate Analysis*, and specifically to its Chief-Editor, Professor Jan de Leeuw, we were offered the possibility to produce this Special Issue in order to present the most innovative recent contributions in these fields. In order to cover a scope as large as possible, our project has not been restricted only to the IWFOS's participants but it has been widely opened and, at the end, a large proportion of the papers in this Issue did not arise from IWFOS's meeting.

Functional Data Analysis (FDA) is a branch of statistics toward which the interest of the scientific community has grown consistently, also in connection with the increasing number of situations in which theoretical and applied scientists have to face data of a continuous nature (functions, curves, images, surfaces, etc.). The field has been popularized in particular by means of the Ramsay–Silverman's books (see [84] and [85]), and in the past twenty years a wide scope of statistical contributions have been published (see for instance [23–25,51,64,66] and [95] for general monographs, [40,60,65] and [79] for methodological surveys, and [22] for a selected set of recent developments). In parallel, extensive and varied methodologies to deal with problems with "very high" dimension have been developed: with the words "very high" dimension, in general one wants to convey that the number of observed variables is large and even sometimes much higher than the number of statistical units (see for instance the monograph [29], the paper [27] for a discussion on challenging open questions in the area,

* Corresponding editor.
E-mail address: philippe.vieu@math.univ-toulouse.fr (P. Vieu).

http://dx.doi.org/10.1016/j.jmva.2015.12.001
0047-259X/

and [1,28] and [43] for selected surveys on few specific issues). Also in this case, one of the major stimulus for the interest in High Dimensional Statistics (HDS) is the ever increasing number of situations involving large numbers of observed variables.

Rather surprisingly, and despite FDA and HDS having the common feature of dealing with dimensional problems, such disciplines have been until now (mostly) developed independently from each other, and very few cross-cutting models and/or methods have been proposed with the intent to take advantage of both approaches. As argued in Section 7.3 of [40], one notable exception to this is represented by the links developed between variable selection techniques (previously expounded in HDS) and the analysis of impact points in a set of functional variables, and each area profited from such cross-cutting methods: for instance the usual HDS variable selection techniques (LASSO for instance) may help to develop new advances in FDA (see [3,7,9,16,48,76,72], and [96]), while in counterpart FDA ideas may help for improving variable selection procedures on HDS (see [34,68] and [97]). This is why, in the spirit of IWFOS2014's meeting (see again [21]), the aim of this Special Issue is double: firstly, it intends to present innovative recent contributions on FDA and HDS and, secondly, to promote contacts between both areas.

At the end, this introductory paper presents the set of selected papers for this Special Issue, either on FDA (see discussion in Section 2) or on HDS (see discussion in Section 3) or on some closely related fields (see Section 4) by putting them back into their bibliographic context. Within each of these sections, the contributions are grouped according to the main topic studied in them, and this implies that some of them may appear more than once.

2. Statistics for functional variables

Classification. Curves classification is a field that drew much attention in the last few years, and this Special Issue welcomes two contributions on this topic: one is on discrimination (i.e., supervised classification) and the other one on clustering (i.e., unsupervised classification). As an extension of the usual multivariate methodologies (see for instance [17] for recent advances), the paper by [37] proposes a classification aggregation algorithm adapted to functional data. In some sense this paper can be seen as a functional adaptation of the paper [18] which will be presented below in Section 4. In the setting of unsupervised classification, the contribution by [15] proposes a methodology to identify shapes (functional data being two-dimensional sets) based on the interpoint distance distribution, that is, the distribution of the distances between two randomly selected points in the figure.

Functional linear regression. Functional linear models have been widely studied in the literature for regression problems in which the predictor is infinite dimensional (see for instance [46] for a comparative bibliographical discussion). In this issue, two contributions are dealing with this kind of model in the more general context where one has observed more than one functional variable. The paper by [36] extends functional linear ideas to the case when both predictors and responses are containing multivariate random functions and the inferential properties of some multifunctional principal component regression type estimate are studied. The paper by [39] is more concerned with linear regression models when a very large number of functional covariates are available, and a functional variable selection procedure is discussed and shown to have a consistency property under general conditions allowing in particular the number of functional predictors to increase up to infinity: this paper is really typical of what we wish to highlight in this Special Issue about the interest of crossing both FDA and HD Statistics.

Nonparametric techniques for functional data. Ever since it was popularized by [51], nonparametric FDA has received great attention in the literature and, naturally, this Special Issue includes various contributions in this area. The paper by [54] is revisiting conditional distribution estimation with functional covariate (see e.g. [49] or [50]) with a view to estimating extreme conditional quantiles and tail index of the Weibull distribution, providing a nice adaptation to the functional context of earlier multivariate nonparametric conditional extreme quantiles theory (see e.g. [42]). Also the paper by [89] is on quantile estimation with extreme-value application, but it studies smoothing parameter selection by means of Bayesian ideas. While dealing with non necessarily independent data is of high interest in FDA (see e.g. [13,51], or [75] for a sample of references using different kinds of dependence), the important question of bandwidth choice has been addressed, as far as we know, only in regression and for i.i.d. data (see [83]), while for dependent data it has been only addressed in a finite-dimensional setting [62,63]. The main innovations in the paper [89] concern dependent functional data and quantile estimation. The contribution by [33] focuses on bandwidth selection in the regression model with scalar response, in which a functional adaptation of the minimax techniques (see e.g. [58]) is used to construct fully data-driven selectors. One of the main interests of this bandwidth selection rule is the fact that it provides location adaptive bandwidths which is known to be, in a functional setting, an important point for good finite-sample behaviour (see, e.g., [6,12,30] or [69]). The last contribution in this field is on functional kernel smoothers which play a major role in nonparametric FDA, as having been used for constructing various functional operator estimates (see e.g. [11,51,69] for, respectively, Nadaraya–Watson, local linear and kNN functional estimates). The paper [45] is proposing a new version of the Nadaraya–Watson functional kernel estimate being specifically adapted to positive response.

Semiparametrics in FDA. Semiparametric modelling for functional data is a currently active field whose main objectives are to balance the trade-off between the high sensitivity to dimension of the nonparametric models and the relative lack of flexibility of the linear models (see [57] for a short survey). Basically, in the literature one can find various different additive type models (see for instance [4,5,52,74,80,87]) or projection based models (see e.g. [35,47,56]) and the contribution by [2] is completing this literature by proposing a combination of these ideas through a so-called Generalized Functional Linear