Nonparametric Statistics for Functional Variables: A double infinitely dimensional challenge

Philippe Vieu
(joint work with Frédéric Ferraty)

Institut de Mathématiques
STAPH working group
Equipe de Statistique et Probabilités
Université Paul Sabatier
118 route de Narbonne, 31062 Toulouse, Cedex, FRANCE

Abstract. The aim of this talk is to present recent advances in nonparametric statistics for functional variables. This is a recent challenging field of Statistics having many potential applications in various fields of applied sciences, but needing sophisticated mathematical developments. The main mathematical difficulties come from the fact that both the observed variables and the target to be estimated can belong to infinite-dimensional (not necessarily linear) spaces. This talk will emphasize on the interest of semi-metric modelisation and of small ball probability considerations, for insuring good mathematical properties of the functional nonparametric methods.
While the main purpose is to emphasize on theoretical issues, the talk will be illustrated by means of a real curves data analysis problem coming from chemometrics.

Key words: Functional variable, Non-linear operator, Non-parametric Statistics, Semi-metric spaces, Small ball probability.

1 Introduction

Because of recent technological progess, applied scientists has now more and more often at hand sophisticated data such as images, curves, ..., and since a few years there is a real challenge for the statistical community for being able to propose models and methods for dealing with such kind of data. The aim of this talk will be to present some theoretical mathematical issues arising in these new kind of problems.

2 The functional feature of the variables

From one side, the stochastic modelling of such data (curves, images, ...) consists in assuming that have at hand a finite sample $X_1, \ldots, X_n$ of (most often but not always, independent) random variables taking values in some abstract measurable space $\mathcal{H}$ supposed to be of (possibly) infinite dimension. To fix the ideas, recall that the standard multivariate statistics case corresponds to the situation when $\mathcal{H} = \mathbb{R}^d$. In this sense, functional data modelling can be seen like a transition from finite to infinite dimensional problem. The non-existency of any measure being invariant by translations in infinite dimensional space (as Lebesgues measure could be in $\mathbb{R}^d$) makes unrealistic (and un-checkable in practice) any kind of model that would assume that the probability distribution of each X_i is absolutely continuous with respect to some given measure on $\mathcal{H}$. This is the main reason why theoretical advances in functional statistics are often made through small ball probability considerations.

3 The operatorial feature of the target object

From the other side, the complexity of functional data structures makes difficult the construction of parametric modelling. In order to insist on the main points linked with nonparametric modelling of functional data, we will stay in the simple regression problem when one wishes to use a functional variable X for predicting an other variable Y (supposed to be real valued for sake of simplicity of presentation). A regression model can be written as

$$Y = r(X) + \epsilon = E(Y|X) + \epsilon,$$

and the question is to estimate r from a sample $X_1, Y_1, \ldots, X_n, Y_n$. It is worth being noted that the object to be estimated (namely r) is a non-linear operator defined on the functional space $\mathcal{H}$. The adaptation of standard nonparametric ideas to this functional setting leads to construct general models of the form

$$r \in \mathcal{R},$$

where $\mathcal{R}$ is some class of operators which cannot be indexed by a finite number of elements of $\mathcal{H}$. Note that, this is much more general than a standard linearity and continuity assumption on r which would allow to summarize r by a single element $r* \in \mathcal{H}$. In this sense, nonparametric modelling is also an infintely dimensional problem. It will be seen that the good theoretical behaviour of nonparametric estimates depends on its local properties, and is therefore strongly linked to the kind of topology introduced in the space $\mathcal{H}$. We will see that, in many cases, the using of topologies induced by semi-metrics is uch more efficient than standard Hilbert or Banach topologies. Of course, this is also strongly linked with the samll ball considerations discussed before since this notion is directly depending on the topology.

4 Application in chemometrics

Spectrometric analysis is a typical application of functional data analysis. Even if the main purpose of the talk is to emphasize on theoretical issues, it will be illustrated by means of a real curves data analysis problem coming from chemometrics. This example will be used all along the talk to illustrate the interest of the functional methodology.

5 Conclusions

Finally, one could say that functional nonparametric statistics are doubly infinite dimensional problem, in the sense that both the observed variables X and the object to be estimated (for instance here r) are allowed to live in infinite dimensional spaces. Most of the results presented in this talk can be found in [1]. The chemometrical example (as well as much other ones) can be reproduced throught the free on line S+/R package (see [2]) which is the computational companion of [1]. It should finally be emphasized that the results presented here are part of the activities of the working group STAPH, created in topulouse some decade ago and interested in various “infinite dimensional” aspects of Statistics (see [3]).

References

[1] Ferraty and Vieu, 2006, *Nonparametric functional data analysis*, Springer Series in Statistics, New York.

[2] Ferraty and Vieu, 2006, *NPFDA: computational issues, S+/R programs and case studies*, feee package on line at http://www.lsp.ups-tlse.fr/staph/npfda/

[3] Staph, *Working Group on Functional and Operatorial Statistics*, http://www.lsp.ups-tlse.fr/staph/