Journal of Statistical Planning and Inference 147 (2014) 1–23

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

A partial overview of the theory of statistics with functional data

Antonio Cuevas*

Departamento de Matemáticas, Universidad Autónoma de Madrid, Spain

ARTICLE INFO

Article history:
Received 30 June 2012
Received in revised form
23 January 2013
Accepted 7 April 2013
Available online 16 April 2013

Keywords:
Functional data
Functional classification
Functional regression
Bootstrap in FDA

ABSTRACT

The theory and practice of statistical methods in situations where the available data are functions (instead of real numbers or vectors) is often referred to as *Functional Data Analysis* (FDA). This subject has become increasingly popular from the end of the 1990s and is now a major research field in statistics.

The aim of this expository paper is to offer a short tutorial as well as a partial survey of the state of the art in FDA theory. Both the selection of topics and the references list are far from exhaustive. Many interesting ideas and references have been left out for the sake of brevity and readability.

In summary, this paper provides:

(a) A discussion on the nature and treatment of the functional data.
(b) A review of some probabilistic tools especially suited for FDA.
(c) A discussion about how the usual centrality parameters, mean, median and mode, can be defined and estimated in the functional setting.
(d) Short accounts of the main ideas and current literature on regression, classification, dimension reduction and bootstrap methods in FDA.
(e) Some final comments regarding software for FDA.

1. Introduction

1.1. Some historical perspective

The problems of statistical inference can be primarily classified according to the nature of the sample space $\mathcal{X}$ (where the available data live) and that of the parameter space Θ, where the target "parameter" is supposed to belong. To a certain extent, the progress of the mathematical statistics can be described in terms of the conquest of new broader more sophisticated structures for $\mathcal{X}$ and Θ, in particular those corresponding to infinite-dimensional spaces. The denomination *Abstract Inference* was used by Grenander (1981) to provide a particularly insightful view of this progress towards generality in the statistical theory.

From this perspective, the theory of statistics with functional data, often denoted Functional Data Analysis (FDA), corresponds to a last-generation statistics where $\mathcal{X}$ (and, in many cases, also Θ) is an infinite-dimensional function space.

* Tel.: +34 914973810; fax: +34 914974889.
E-mail address: antonio.cuevas@uam.es

0378-3758/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jspi.2013.04.002

So, according to the above mentioned classification, FDA could be placed in the general development of statistical theory as indicated in the following informal sketch (where n denotes the sample size):

Statistical theory	$\mathcal{X}$	Θ	Dating back to
Classical parametric inf.	$\mathbb{R}$	$\Theta\subset\mathbb{R}$	1920s
Multivariate analysis	$\mathbb{R}^d$ $(n >> d)$	$\Theta\subset\mathbb{R}^k$ $(n >> k)$	1940s
Nonparametrics	$\mathbb{R}^d$ $(n >> d)$	A function space	1960s
High dimensional problems	$\mathbb{R}^d$ $(n < d)$	$\Theta\subset\mathbb{R}^k$	2000s
Functional Data Analysis	A function space	$\mathbb{R}^k$, or a function space	1990s

Hence, in simple words, we might say FDA refers usually to those statistical problems where the available data consist on a sample of n functions $x_1 = x_1(t), \ldots, x_n = x_n(t)$ defined on a compact interval of the real line, say $[0, 1]$. Additional (real or functional) variables are often incorporated, for example in the regression models.

Other more sophisticated models are possible, where [0, 1] is replaced by a d-dimensional interval or the sample functions are vector-valued. Also, FDA bears some affinity with those statistical problems, often referred to as "inference in stochastic processes" where the sample information is given by a partial trajectory $x(t)$, $t\in[0,T]$ of a stochastic process $\{X(t),\ t\geq 0\}$. In this case, the length T of the observation interval plays the role of the sample size n.

Of course the FDA theory has incorporated many standard tools of the classical parametric or multivariate statistics. For example, the dimension reduction tools. However, the infinite-dimensional nature of the sample space poses especial problems which allow us to classify FDA as a genuinely new branch of the statistical theory.

1.2. Some monographs and general references

The book by Ramsay and Silverman (2005), whose first edition was published in 1997, must be cited as a major landmark in the history of FDA. This book has a practical orientation, targeted to a wide scientific audience. It has developed a crucial role in the popularization of FDA. The associated software, freely provided by the authors, became soon an effective toolbox for an increasing number of researchers, flooded by a new abundance of experimental data coming from on-line monitoring of different experiments. Another book by the same authors, Ramsay and Silverman (2002), is focussed on the illustration of the main FDA techniques through the study of specific case studies with real data.

The book by Ferraty and Vieu (2006) represented a second-generation view of the subject. It incorporates further mathematical insights, including a more detailed treatment on the non-trivial asymptotic issues involved in FDA, together with some discussion of several relevant issues as the use of semi-metrics and the so-called "small ball probabilities" phenomenon, which is in the basis of many theoretical difficulties in FDA. However, again, the practical aspects played a major role among the aims of this book.

The FDA French school include many other references deserving mention: some researchers are grouped in the STAPH team (www.math.univtoulouse.fr/staph/) whose earliest contribution to the topic is perhaps the paper by Dauxois et al. (1982), a pioneering contribution to the study of principal components for functional data.

The paper by Bosq (1991) is another path-breaking reference in the topic (not considered here) of functional-valued time series. The corresponding general theory of auto-regressive functional processes is given in Bosq (2000). The monograph by Bosq and Blanke (2007) deals mainly with the use of nonparametric approaches in statistical functional problems. Whereas the orientation of these two books is mostly theoretical, they are both, in a way, extremely practical as they jointly provide a fascinating account of the main mathematical tools involved in FDA.

The book by Horváth and Kokoszka (2012) is a fresh addition to the current general literature on FDA. It offers a well-balanced mixture of theoretical aspects (e.g., the useful Chapter 2 on Hilbert space theory) and applications (in particular, detailed discussions of real data examples and up-to-date information on software). About 40% of the book length (from Chapters 13 to 18) is devoted to the analysis of dependent functional data, including functional time series, change point detection and spatial statistics with functional data.

Special issues devoted to FDA topics have been published by different journals, including *Statistica Sinica*, issue 14, 3 (2004). *Computational Statistics*, 22, 3 (2007), *Computational Statistics & Data Analysis*, 51, 10 (2007), *Journal of Multivariate Analysis*, 101, 2 (2010).

Among the survey and overview papers, let us mention, e.g., Rice (2004), Müller (2005), González-Manteiga and Vieu (2011) (which includes an extensive bibliography), Delsol et al. (2011) (especially oriented to practical issues and real-data applications), etc.

The recent collective book Ferraty and Romain (2011) consist of 16 chapters, from different authors, with up-to-date surveys of the main topics of FDA.

1.3. The purpose and contains of this paper

The aim is to provide a personal perspective of the current theory and practice of FDA. Such a view is necessarily limited by several obvious constrains, including the space limitations and the author's awareness of the different subjects.

This accounts for the use of the term "partial" in the title. In particular, no claim of bibliographical completeness is made (I apologize for any omissions). In fact, since FDA is a vast topic with many different facets, I realize that there would probably be room for another paper, almost disjoint with this one, written under a similar title by another author with different interests and experience.

This paper is a sort of mixture of a tutorial and an up-to-date survey of FDA: it is somewhat of a tutorial since it is not targeted to a readership of specialists in FDA but rather to a broader audience of statisticians, not necessarily familiar with the subject. This places readability and a certain degree of self-containedness as major priorities of this paper. In the survey aspect, an effort has been made to provide a reasonably wide range of topics and, as a rule, the most recent references have been preferred over the older ones.

The organization of the paper tries to evoke the familiar structure of so many classical books of statistics: thus, some "descriptive" aspects (not related to inference notions), concerning the structure and representation of the data are discussed in Section 2, the probabilistic basis of FDA theory is briefly outlined in Section 3; the location, centrality parameters (mean, median and mode) are discussed in Section 4; regression, classification and dimension reduction techniques are analyzed in Sections 5–7 respectively. The functional resampling methodologies are reviewed in Section 8. Some remarks on the available software are made in Section 9.

2. The data

Unlike the classical statistics (where there is usually little discussion on the nature of the data) in FDA one might reasonably ask: Do really exist such a thing as a functional data? The question has some relevance as, in practice, when a process $\{X(t), t\geq 0\}$ is monitored, one usually record the values in a discrete grid $t_1, \ldots, t_N$. So, at the end one always has a, possibly high-dimensional, vector observation $(x(t_1), \ldots, x(t_N))$. There are at least two reasons that could lead us to consider this as a functional data: first, the possibility (at least theoretical) of observing the phenomenon in a much finer grid and, in the limit, to observe $x(t)$ at any fixed instant t. Second, the choice of a functional model to approximately represent it. Hence, the FDA methodology can be seen as a new toolbox of functional models, appropriate to deal with those statistical experiments where the individual observations can be modeled as elements of a function space.

To fix ideas let us next consider the following example.

2.1. An example in experimental cardiology: the calcium curves

The mitochondrial calcium overload, a measure of the mitochondrial calcium ion Ca2+ levels, was measured in two groups (control and treatment) with 45 cells each. In the treatment group the cells received *Cariporide*, a drug which is expected increase the total Ca2+ load within the mitochondria (Fig. 1).

Higher levels indicate a better protection against the ischemia process.

This magnitude has been monitored in isolated mouse cardiac cells. In each cell, measurements were taken every 10 s (for 1 h) using a fluorescence imaging system. I am very grateful to David García-Dorado (Cardiology Service, Hospital Vall d'Hebron, Barcelona) for providing, and kindly explaining to me, these data.

The initial purpose of the experiment is to check the effect of *Cariporide* comparing the curves obtained in the treatment and the control group. The functional ANOVA procedure proposed in Cuevas et al. (2004) provided a p-value 0.04 for these data. So there is some statistical evidence of differences between both groups.

According to the way in which the data come to the researcher, it is usual to distinguish between two situations: the case of *densely observed curves* in which it is possible to observe the data $x(t)$ in an arbitrarily fine grid of points $(x(t_1), \ldots, x(t_N))$ and the case of *sparse data* where the $x_i(t)$ are observed in a given grid of points $x_i(t_{ij})$, although the motivation for the proposed methods still remains functional. Unless otherwise stated we will assume here the dense situation but we also comment several instances of the sparse case, which is particularly important in many practical applications.

2.2. Processing the data: basis representation and smoothing

Very often the "raw data" $(x(t_1), \ldots, x(t_N))$ require a preliminary treatment, before applying the FDA techniques. This is motivated either in terms of dimension reduction or in order to remove the noise present in the data measurements.

Basis representation is a very usual way to transform the data: assume that $x \in L^2[0,1]$. Then, if $\{e_k(t)\}$ is a orthonormal basis of that space, we may think of fitting a function $\tilde{x}$ from the raw data in the following way: we fix a number J of basis functions, typically smaller than n, and we define

$$\tilde{x}(t) = \sum_{j=1}^{J} c_j e_j(t),$$

where the "Fourier coefficients" c_j are chosen in order to minimize (see Ramsay and Silverman, 2005 for details)

$$\sum_{k=1}^{N} \left(x(t_k) - \sum_{j=1}^{J} c_j e_j(t_k) \right)^2$$

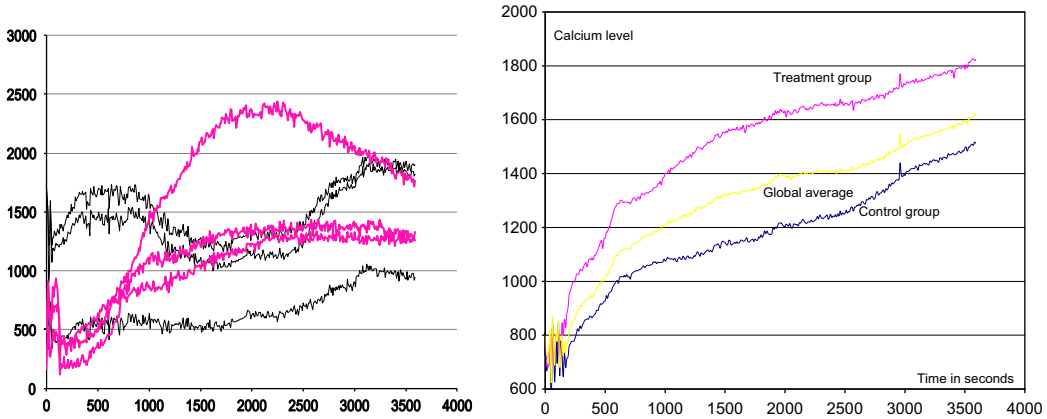

Fig. 1. On the left panel, six calcium overload curves: the three dark lines correspond to the control group, the three lighter curves are from the treatment group. The right panel shows the mean curves for both groups and the overall mean curve.

Then, this representation process can be summarized in terms of two transformations,

$$(x(t_1),\ldots,x(t_N))\mapsto(c_1,\ldots,c_J)\mapsto(\tilde{x}(t_1),\ldots,\tilde{x}(t_N)).$$

Hence, this procedure provides both a more compact representation of the data (as J is typically much smaller than N) and a "denoising" process, since $\tilde{x}$ can be considered as a smoothed version of the original data x.

Another usual smoothing procedure can be done via convolution with a kernel function. Again in numerical terms this amounts to a linear transformation

$$x=x(t)\mapsto\tilde{x}=\tilde{x}(t)=\sum_{j=1}^{N}S_j(t)x(t_j), \tag{1}$$

where $S_j(t)$ could be, for example, the Nadaraya–Watson weights

$$S_j(t)=\frac{K\left(\frac{t-t_j}{h}\right)}{\sum_{i=1}^{N}K\left(\frac{t-t_i}{h}\right)}:=W_{hN}(t-t_j),$$

associated with a kernel function K (a typical choice is the standard Gaussian density) and h is the bandwidth parameter which controls the desired smoothing degree. Note that (1) provides in fact an approximation to the functional convolution transformation $x(t)\mapsto N\int_0^1 W_{hN}(t-u)x(u)\,du$.

The choice of the orthogonal system $\{e_j\}$ in the basis representation procedure and the weight function W_{hN} in the convolution smoothing are important, non-trivial decisions whose discussion is beyond the scope of this survey. See, Ramsay and Silverman (2005) for more details.

Note that an important advantage of these regularization processes is the possibility of using the first or second order derivatives of the original data. This is extremely useful in practice, as sometimes the relevant information is included in the derivatives rather than in the data themselves. A well-known example is given by the spectrometric curves discussed, among others, by Ferraty and Vieu (2006).

2.3. Where do the data live?

Most theoretical developments require the assumption that the sample space $\mathcal{X}$ is a real separable Banach space whose norm is denoted by $\|\cdot\|$. In formal terms, our sample data will be observations drawn from an $\mathcal{X}-$valued random element X (that is, a measurable function) defined on some probability space $(\Omega,\mathcal{A},\mathbb{P})$. Separability ensures that a linear combination of $\mathcal{X}-$valued random elements is again a random element. Very often a structure of (separable) Hilbert space, with associated inner product $\langle\cdot,\cdot\rangle$, is needed for $\mathcal{X}$. This is the case, for example, if orthogonal expansions must be used.

Two standard choices for $\mathcal{X}$ are $\mathcal{C}[0,1]$, the Banach space of real continuous functions $x:[0,1]\to\mathbb{R}$ endowed with the supremum norm $\|x\|=\max_t|x(t)|$, and the Hilbert space $L^2[0,1]$ of square integrable real functions on $[0,1]$ endowed with the usual inner product $\langle x,y\rangle=\int_0^1 x(t)y(t)\,dt$.

Nevertheless, many other possibilities are available, depending on the particular aspect of the data we are interested in. These include the use of different seminorms, for example those based on the L^2-distance between the first or second derivatives of the observed functions. More details can be found for example in Ferraty and Vieu (2006).

3. Some probability background

Many classical books of mathematical statistics used to include some introductory chapters devoted to the probability foundations of the subject. Indeed there are some good reasons for this, since statistical inference can be seen as a second step after proposing a probability model for a random experiment. In the case of FDA, since the data are functions, the observed "random variables" are in fact stochastic processes, i.e., random elements taking values in a function space. Therefore some crucial notions such as expectation or stochastic convergences must be adapted (with highly non-trivial changes) to the infinite-dimensional framework. Other important notions, almost omnipresent in classical statistics (such as quantiles, density function, distribution function, maximum likelihood) simply do not work in the FDA setup or require a painstaking adaptation process. In this section we provide a short, non-systematic list of some key probability concepts which appear often in the FDA theory.

3.1. The concept of "functional expectation". Strong and weak integral

From the point of view of statistical inference, the functional data are (in most cases) observations drawn from a random element X defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and taking values in a separable Banach space $(\mathcal{X}, \|.\|)$. Of course, a suitable notion of expectation $\mathbb{E}(X)$ is also crucial here. The natural approach, based on a direct extension of the standard integral for real measurable functions with respect to a measure $\mathbb{P}$, can also be applied, with some obvious changes, here. Thus, the integral $\mathbb{E}(X) = \int_\Omega X\, d\mathbb{P}$ can be defined by the usual "ascending hierarchy" of complexity in the integrand: if X is an indicator function $X = \mathbb{I}_A$, then $\mathbb{E}(X) = \mathbb{P}(A)$. Now, when X is a simple function (a linear combination of indicators), $\mathbb{E}(X)$ is defined in the natural way, consistent with the desired property of linearity. When X is a general random element we express it as a limit of simple functions and define $\mathbb{E}(X)$ as the corresponding limit of expectations. This is the so-called *Bochner integral*, or *strong integral*. It can be shown that the expectation of X in Bochner sense does exist if and only if $\mathbb{E}\|X\| < \infty$. More details on the definition and properties of the strong integral can be found, e.g., in Bosq (2000).

As we are usually dealing with a random function $X = X(t, \omega)$, $t \in [0, 1]$, $\omega \in \Omega$, there is another natural way to define the function $\mathbb{E}(X) = \mathbb{E}(X)(t)$ which would consist just on computing, for each t, the ordinary expectation $\int_\Omega X(t, \omega)\, d\mathbb{P}(\omega)$. The general version of this idea leads to the so-called *Pettis integral* or *weak integral*. A random variable taking values a Banach space $\mathcal{X}$ is said to be weakly integrable if there exists an element of $\mathcal{X}$ denoted by EX such that $\mathbb{E}(x^*(X)) = x^*(EX)$ for all $x^* \in \mathcal{X}^*$ (the continuous dual of $\mathcal{X}$). It can be shown that if the Bochner integral does exist and is finite then it coincides with the weak integral.

3.2. Gaussian processes. Karhunen–Loève expansion

Typically our sample curves will be realizations of a random element (stochastic process) $X = X(t) = X(t, \omega)$ defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and taking values in a function (trajectory) space $\mathcal{X}$, say $\mathcal{X} = \mathcal{C}[0, 1]$ or $L^2[0, 1]$. Thus, different assumptions on the distribution of X lead to different probability models in FDA.

As in the case of classical inference models the Gaussian distribution plays an outstanding role in FDA. Let us recall that a random element X taking values in a Banach space $\mathcal{X}$ is said to be Gaussian if $x^*(X)$ is a real-valued Gaussian random variable for all $x^* \in \mathcal{X}^*$. In the case of a stochastic process $X = X(t)$ a equivalent definition can be established in terms of the finite-dimensional projections $(X(t_1), \ldots, X(t_k))$, for $t_1, \ldots, t_k \in [0, 1]$ and $k \in \mathbb{N}$: X is Gaussian if and only if all these marginals have Gaussian distributions. The distribution of a Gaussian process is uniquely determined by the mean function $m(t) = \mathbb{E}(X(t))$ and its covariance function $\gamma(s, t) = \mathrm{Cov}(X(s), X(t))$. Two examples of special interest are (standard) *Brownian motion* for which $m = 0$ and $\gamma(s, t) = \min(s, t)$ and the (standard) *Brownian bridge*, for which $m = 0$ and $\gamma(s, t) = \min(s, t)(1 - \max(s, t))$.

A useful tool in many FDA developments is the following result (see, e.g., Ash and Gardner, 1975, p. 38), which can be seen as a stochastic analog of the classical Fourier expansions for real functions.

Karhunen–Loève expansion. *Let $X = X(t)$, $t \in [0, 1]$, be a stochastic process with $\mathbb{E}(X(t)) = 0$ and $\mathbb{E}(X^2(t)) < \infty$ for all $t \in [0, 1]$. Suppose that the covariance function $\gamma(s, t)$ is continuous.*

Then $X(t)$ can be expressed in the form

$$X(t) = \sum_{k=1}^{\infty} Z_k e_k(t), \tag{2}$$

where the convergence is in L^2, uniform in t, $\{e_k\}_{k \in \mathbb{N}}$ is an orthonormal basis of $L^2[0, 1]$ given by the eigenfunctions of the covariance operator Γ, associated with $\gamma(s, t)$, whose corresponding eigenvalues are λ_k (that is $\lambda_k e_k(t) = \int_0^1 \gamma(s, t) e_k(s)\, ds$) and $Z_k = \int_0^1 X(t) e_k(t)\, dt$ is a sequence of orthogonal (uncorrelated) random variables with $\mathbb{E}(Z_k) = 0$, $\mathbb{E}(Z_i^2) = \lambda_i$.

As we will comment below, Karhunen–Loève expansion plays a fundamental role, e.g., in the FDA regression theory. Note that the Z_k's in (2) are always uncorrelated but the independence is only guaranteed in the Gaussian case (also, in this case, the convergence in (2) holds also a.s.). This a further reason why Gaussianity is a so common assumption in FDA. As argued by Delaigle and Hall (2010, p. 1173), "*Particularly in the infinite-dimensional setting of functional data analysis, it seems impossible to use effectively general models for random variables that are uncorrelated but not independent. Such an approach leads to cumbersome methods and does not seem to allow useful insight into theoretical properties.*" Let us note however that in

the setting of linear prediction for random variables taking values in a separable real Hilbert space $\mathbb{H}$, the assumption of *strong orthogonality* (more general than independence: see Bosq and Blanke, 2007, p. 231) can be used in a natural way to obtain the expression of the linear predictor of an $\mathbb{H}-$valued element Y from a sequence Z_n of strongly orthogonal $\mathbb{H}-$valued elements Z_n.

3.3. On (the lack of) densities in the functional setup

The density functions (with respect to the Lebesgue measure, μ_L) are an extremely useful tool in finite-dimensional statistics and probability. The point is that in many cases, it makes sense to assume that the distribution law P_X [given by $P_X(B)=\mathbb{P}(X\in B)$] of the random variable X of interest is absolutely continuous with respect to μ_L or, in other words it is a μ_L-continuous distribution (a standard notation for this is $P_X\ll\mu_L$). This means that $P_X(B)=0$ whenever $\mu_L(B)=0$, which from Radon–Nikodym theorem, is equivalent to ensure the existence of a *density function* f such that $P_X(B)=\int_B f\,d\mu_L$, for all Borel sets B. Among other advantages, the density functions allow us to easily characterize distributions, calculate probabilities and moments and define likelihood functions.

The lack of a simple natural way to define and calculate density functions in infinite-dimensional spaces is one of the sharpest differences between FDA and the standard statistical theory. The point is that in infinite-dimensional spaces there is no general simple way for defining (translation-invariant) "reference measures" playing a role similar to that of μ_L in $\mathbb{R}^d$. However, there are some problems (for example in binary supervised classification or logistic regression) where basically we have two probability measures μ_0 and μ_1 involved. In those situations it could make sense to take one of them as a reference to define the Radon–Nikodym derivatives. To be more specific, let Y be a dichotomous variable with $p=\mathbb{P}(Y=0)$ and $\mathbb{P}(Y=1)=1-p$, with $p\in(0,1)$. Let X be the (possibly functional) regressor variable used to predict Y. Let us denote by μ_i the conditional distribution of $X|Y=i$, for $i=0,1$. A major aim in several important statistical problems is to estimate the regression function $\eta(x)=\mathbb{E}(Y|X=x)=\mathbb{P}(Y=1|X=x)$. It can be seen (see Baíllo et al., 2011a, Theorem 1) that, if we assume that both measures μ_0 and μ_1 are equivalent (that is they both are mutually absolutely continuous with respect to each other) then

$$\eta(x)=\frac{1-p}{p\dfrac{d\mu_0}{d\mu_1}(x)+1-p}\quad\text{a.s.} \tag{3}$$

where $d\mu_0/d\mu_1(x)$ denotes the Radon–Nikodym derivative of μ_0 with respect to μ_1.

The hypothesis of equivalence of μ_0 and μ_1 is particularly meaningful in the case of Gaussian processes, in view of *Feldman–Hájek dichotomy* according to which two Gaussian measures μ_0 and μ_1 are either equivalent or mutually singular, that is, there exists a set A such that $\mu_1(A)=0$ and $\mu_0(A)=1$ (this is denoted $\mu_0\perp\mu_1$). It is worth mention that (unlike the case of finite-dimensional Gaussian distributions) the assumption of equivalence for two Gaussian processes is quite restrictive in the sense that it fails to hold (somewhat counterintuitively) in many simple standard examples: thus if μ_0 is the distribution of a standard Brownian motion $\{B(t),t\in[0,1]\}$ and μ_1 is the distribution of $\sigma^2B(t)$ with $\sigma^2\neq1$ then $\mu_0\perp\mu_1$; see Varberg (1961).

However, the most striking fact concerning general Radon–Nikodym derivatives is maybe that they have explicit, not too complicated expressions, in several important (Gaussian) cases. A relevant example is given by the following result; see, e.g., Mörters and Peres (2010, p. 24).

Cameron–Martin Theorem. *Let $F\in\mathcal{C}[0,1]$ such that $F(0)=0$. Let μ_0 and μ_F be the distribution of the standard Brownian motion B in $\mathcal{C}[0,1]$ and the distribution of the "Brownian with trend" F, $B_F(t)=F(t)+B(t)$, respectively. Denote by $\mathcal{D}[0,1]$ the Dirichlet space $\mathcal{D}[0,1]=\{F:[0,1]\to R:F(t)=\int_0^t f(s)\,ds$, for some $f\in L^2[0,1]\}$. Then,*

(a)*If $F\notin\mathcal{D}[0,1]$, then $\mu_F\perp\mu_0$.* (b) *If $F\in\mathcal{D}[0,1]$, then μ_F and μ_0 are equivalent. Moreover, in this case,*

$$\frac{d\mu_F}{d\mu_0}(B)=\exp\left(-\frac{1}{2}\int_0^1 F'(s)^2\,ds+\int_0^1 F'\,dB\right), \tag{4}$$

*for μ_0-*almost *all $B\in\mathcal{C}[0,1]$.*

If F' is not of bounded variation (so that standard integration-by-parts would not work), $\int_0^1 F'\,dB$ must be interpreted in the sense of *Paley–Wiener stochastic integral*.

As an example, let us consider the case $F(t)=ct$: expression (4) combined with Itô Formula leads to

$$\frac{d\mu_F}{d\mu_0}(B)=\exp\left(-\frac{c^2}{2}+cB(1)\right),\quad \mu_0-\text{a.s.}$$

Other more complicated (but still affordable) expressions of Radon–Nikodym densities, for more general cases of Gaussian processes with "triangular" covariance functions of type $\gamma(s,t)=u(\min(s,t))v(\max(s,t))$, can be found, e.g., in Varberg (1961) and Jørsboe (1968). The idea of using such explicit expressions of the densities $d\mu_0/d\mu_1$ in order to define plug-in estimators of the regression function (3) in binary classification problems has been considered in Baíllo et al. (2011a).

A novel approach to the use of densities in FDA is proposed by Delaigle and Hall (2010) who define a so-called *log-density* in terms of the densities of the first r score variables $\lambda_k^{-1/2}Z_k$ in the Karhunen–Loève expansion of X (2). See also Dabo-Niang et al. (2010) and references therein for further ideas on infinite-dimensional densities.

3.4. Some basic functional limit theorems

The earliest, best known functional versions of the Strong Law of Large Numbers (SLL) and the Central Limit Theorem (CLT) are due to Mourier (1953) and Varadhan (1962), respectively. The functional versions of SLL and CLT are essential tools in the FDA asymptotic theory. The basic statements, given below, are closely analog to those of their one-dimensional counterparts.

Functional strong law of large numbers. *Let $\{X_n\}$ be a sequence of iid random elements with values in a separable Banach space $\mathcal{X}$. If $\mathbb{E}\|X_1\| < \infty$ then*

$$\frac{\sum_{i=1}^n X_i}{n} \xrightarrow{a.s.} \mathbb{E}(X_1),$$

where $\mathbb{E}(X_1)$ denotes the strong expectation of X_1.

Functional central limit theorem. *Let $\{X_n\}$ be a sequence of iid random elements with values in a separable Hilbert space $\mathcal{X}$. If $\mathbb{E}\|X_1\|^2 < \infty$ then*

$$\sqrt{n}\left(\frac{\sum_{i=1}^n X_i}{n} - \mathbb{E}(X_1)\right) \xrightarrow{w} \mathcal{G}(0, \Gamma_{X_1}), \tag{5}$$

where $\mathcal{G}(0, \Gamma_{X_1})$ denotes the Gaussian probability measure on $\mathcal{X}$ with expectation 0 and covariance operator $\Gamma_{X_1}(x^*, y^*) = \mathrm{Cov}(x^*(X_1), y^*(X_1))$, for $x^*, y^* \in \mathcal{X}^*$ (the continuous dual of $\mathcal{X}$).

Much more general versions of these classical theorems are today available. See the paper by Hoffmann-Jorgensen and Pisier (1976) and the Chapters 7 and 10 in the monograph by Ledoux and Talagrand (2011) for a comprehensive treatment of these limit results.

3.5. Some useful inequalities

Some inequalities are behind many of the deepest results in statistics. The book by DasGupta (2008, Chapter 35) provides a summary of useful inequalities in the standard theory of statistics and probability. Let us mention here just a couple of examples of inequalities with applications in FDA. Our first example is the so-called *bounded differences* or *McDiarmid's* inequality. Its very general formulation makes it a useful tool, e.g., in the analysis of classification errors (see Boucheron et al., 2005).

McDiarmid's inequality (*McDiarmid, 1989*). *Suppose $g : \mathcal{X}^n \longrightarrow \mathbb{R}$ satisfies*

$$\sup_{x_1,\ldots,x_i,x'_i,\ldots,x_n} |g(x_1,\ldots,x_n) - g(x_1,\ldots,x_{i-1},x'_i,x_{i+1},\ldots,x_n)| \le c_i, \text{ for some given constants } c_i.$$

Then, if $X_1,\ldots,X_n$ are independent $\mathcal{X}-$valued random elements,

$$\mathbb{P}(|g(X_1,\ldots,X_n) - \mathbb{E}(g(X_1,\ldots,X_n))| \ge \epsilon) \le 2e^{-2\epsilon^2/\Sigma c_i^2}.$$

The following exponential inequality, due to Yurinskii (1976), has been used in the theory of principal functional components (Boente and Fraiman, 2000) and in the study of linear regression with functional response (Cuevas et al., 2002):

Yurinskii's inequality (*Yurinskii, 1976*). *Let $\{X_i\}$ be independent random elements taking values in a separable Hilbert space. Assume $\mathbb{E}(X_i) = 0$, $\mathbb{E}\|X_i\|^m \le (m!/2)b_i^2 A^{m-2}$ for some constants $b_i = b_i(n)$, $A = A(n)$ and all $m \ge 2$. Then, if $B_n = \sum_{i=1}^n b_i^2$,*

$$\mathbb{P}\left(\left\|\sum_{i=1}^n X_i\right\| > xB_n\right) \le 2\exp\left(-x^2\left[2 + 3.24\frac{xA}{B_n}\right]^{-1}\right)$$

Several other exponential inequalities are available for the functional case; see Bosq (2000, p. 49) for further Bernstein-type inequalities and Ferraty et al. (2011) for applications to the study of kernel regression estimates with functional response.

3.6. The random projections methodology

The following result, due to Cuesta-Albertos et al. (2007) is a useful generalization of the classical Cramér–Wold device.

Theorem (*Cuesta-Albertos et al., 2007*). *Let X, Y be random elements taking values in a separable Hilbert space $(\mathbb{H}, \langle\cdot\rangle)$ with distributions P and Q, respectively. Let μ be a non-degenerate Gaussian measure on $\mathbb{H}$. Denote by $\|\cdot\|$ the norm in $\mathbb{H}$ associated with the inner product $\langle\cdot\rangle$. Assume that*:

(a) *The absolute moments* $m_n = \mathbb{E}\|X\|^n = \int \|x\|^n\, dP(x)$ *satisfy the Carleman condition* $\sum_n m_n^{-1/n} = \infty$.
(b) $\mu\{v \in \mathbb{H} : \langle v, X\rangle \overset{d}{=} \langle v, Y\rangle\} > 0$, *where the notation* $\overset{d}{=}$ *stands for equality in distribution.*

Then, $P = Q$.

Roughly speaking, this result establishes that, under condition (a), a probability distribution in a Hilbert space is determined by its one-dimensional projections in any set of positive measure. In other words, if (a) holds, given a reference Gaussian measure μ, we have that, if two probability distributions are different, then the $\mu-$probability of finding two one-dimensional linear projections equally distributed is zero. The applications of this result in goodness of fit (in a FDA setup) are analyzed in Cuesta-Albertos et al. (2007). Another natural application, aimed to define new functional depth measures is considered in Cuesta-Albertos and Nieto-Reyes (2008) and Cuevas and Fraiman (2009).

3.7. Small balls probabilities

As mentioned above, the regression function $\eta(x) = \mathbb{E}(Y|X = x)$, where X is a functional explanatory variable and Y is a real-valued response is also a basic prediction tool in FDA. Then, its nonparametric estimation is a relevant issue. Interestingly, the classical nonparametric regression estimation methods can be adapted, in a quite simple and natural way (see Section 5 below), to the functional setting. However, some important theoretical issues arise due to the infinite-dimensional character of the data space. One of these is the slow convergence rates found in nonparametric FDA, at least for the standard choices ($L^2[0,1]$, $\mathcal{C}[0,1]$,…) of the sample space $\mathcal{X}$. The reason is intuitively simple: as the nonparametric methods are "of local nature" (that is, we essentially need to have sample points at a neighborhood of any point in order to make accurate estimations) it turns out that the convergence rates for these methods depend on the so-called *small balls probabilities*, that is, $\phi_x(h) = \mathbb{P}(X \in B(x,h))$, where $B(x,h)$ denotes the closed ball with center x and radius h; see Ferraty and Vieu (2011). When $\mathcal{X} = \mathbb{R}^d$, we typically have $\phi_x(h) \sim h^d$ which is really small for large values of d. This is basically the reason for the so-called *curse of dimensionality*: it is really difficult obtain good nonparametric estimations in high-dimensional spaces since, for d large, $\mathbb{R}^d$ is "almost empty" of sample points, unless unrealistically large sample sizes are available.

In the functional setting the situation is even worse. For example, for the standard Brownian motion in $\mathcal{C}[0,1]$ we have $\phi_x(h) \sim \exp(-\pi^2/8h^2)$ (see, e.g., Li and Shao, 2001). This suggests that sometimes the standard function spaces are "too large". A possibility is to use simpler models for the data structure by considering other distances or semi-distances; more details can be found in Ferraty and Vieu (2006).

4. Mean, median and mode

4.1. The definitions

Regarding the mean, similar to the real-valued case, the integral-based definition of the "functional" expectation (see Section 3) allows for the usual motivation in terms of projections: If X is a random element taking values in a Hilbert space $\mathcal{X}$ and $\mathbb{E}\|X\|^2 < \infty$, then the *mean* of X, $m = \mathbb{E}(X)$, fulfills $\mathbb{E}\|X - m\|^2 = \min_{a\in\mathcal{X}} \mathbb{E}\|X - a\|^2$; see, e.g., Bosq (2000, pp. 38–41), for a broader discussion of this, including the extension to conditional expectations. This property is behind the interpretation of m in terms of a "centrality value".

In a similar way, still keeping the analogy with the classical one-variate case, one could think of defining the *median* $M = M(X)$ as the minimizer of $\Psi(a) = \mathbb{E}(\|X - a\| - \|X\|)$. From the triangle inequality $\Psi(y)$ is always well-defined, even if $\mathbb{E}\|X\| = \infty$. Moreover, when $\mathbb{E}\|X\| < \infty$, we have $\mathbb{E}\|X - M\| = \min_{a\in\mathcal{X}} \mathbb{E}\|X - a\|$. It can also be seen that M does exist, provided that X takes values in a space $\mathcal{X}$ which is the continuous dual of a separable Banach space. In addition, uniqueness of M is guaranteed whenever $\mathcal{X}$ is a strictly convex space (i.e., $\|x + y\| < \|x\| + \|y\|$, when x and y are not proportional; this holds for Hilbert spaces or for $\mathcal{X} = L_p$, $1 < p < \infty$) provided that X is not concentrated on a "straight line" (of the form $\{\lambda a + b,\ \lambda \geq 0\}$ for some $a, b \in \mathcal{X}$) in $\mathcal{X}$. Also, under the symmetry assumption that the distributions of $X - m$ and $m - X$ coincide, one gets $m = M$; see Kemperman (1987), Vardi and Zhang (2000), Cadre (2001) and Gervini (2008) for closely related ideas and details.

The lack of a unique natural notion of density in the infinite-dimensional spaces is a major hurdle for the task of extending the notion of *mode* to the functional setting. However, the following simple notion of h-mode has been used in Dabo-Niang et al. (2007) and Cuevas et al. (2007): given a suitable kernel function K (e.g., the standard Gaussian density or the quadratic kernel $K(t) = \frac{3}{2}(1-t^2)\mathbb{I}_{[0,1]}(t)$) we could define the h-mode of X as

$$M_0 = \operatorname*{argmax}_a \mathbb{E}K\left(\frac{\|a - X\|}{h}\right).$$

The empirical version of this definition is useful for data analysis purposes. However, in principle, h must be fixed in advance as a constant. The point is that, unlike the finite-dimensional case, there is no natural density to be estimated making h tend to zero "slowly enough".

For other interesting, more sophisticated ideas on the notion of functional mode, see Gasser et al. (1998), Hall and Heckman (2002) and Dabo-Niang et al. (2010).

4.2. Sample versions

The above definitions for m, M and M_0, allow for direct empirical versions, based on a sample $X_1,\ldots,X_n$. Of course the idea is to replace everywhere the expectation with the corresponding empirical average. Thus we define

$$\hat{m}=\frac{1}{n}\sum_{i=1}^{n}X_i,\quad \hat{M}=\operatorname*{argmin}_{a}\sum_{i=1}^{n}\|X_i-a\|,\quad \hat{M}_0=\operatorname*{argmin}_{a}\sum_{i=1}^{n}K\left(\frac{\|a-X_i\|}{h}\right).$$

For practical and computational aspects regarding $\hat{M}$ and $\hat{M}_0$ we refer again to Gervini (2008), Cuevas et al. (2007) and Dabo-Niang et al. (2007). Cuesta-Albertos and Matrán (1989) and Cadre (2001) have analyzed the consistent estimation of M in separable Banach spaces; see also Cuesta-Albertos and Matrán (1988) for closely related results.

As for the estimation of m, the particular importance of this "function parameter" has led to consider different sample models in practice. For example, a quite natural assumption is to assume that our data come in fact in a "sparse" fashion which, moreover, is affected by noise. Thus we observe

$$Y_{ij}=X(t_{ij})+e_{ij},\quad j=1,\ldots,m_i,\ \ i=1,\ldots,n,$$

where the $t_{ij}\in[0,1]$ are some "design points" and the noise variables e_{ij} are supposed to be independent homoscedastic with mean zero. The optimal estimation of the mean function under different assumptions for the design t_{ij} has been analyzed by Cai and Yuan (2011). A similar model has been previously considered by other authors, e.g., Rice and Silverman (1991). The paper by Bunea et al. (2011) is also devoted to the estimation of the mean function of a Gaussian process under a sparse sampling model.

4.3. Depth-based notions of medians and quantiles

The study of the so-called *depth functions* has become a hot topic for research in FDA. Given a probability measure P on the sample space $\mathcal{X}$, a depth function, relative to P, is just a non-negative function $D(P,x)$ defined on $\mathcal{X}$ which indicates "how deep" is the observation x "inside" the distribution P. If $\mathbb{P}_n$ stands for the empirical distribution corresponding to a sample $X_1,\ldots,X_n$, then $D(\mathbb{P}_n,x)$ indicates how deep is the observation x inside the sample. One reason for the increasing popularity of depth notions in multivariate analysis and FDA is the fact that every empirical depth function provides a sort of ordering of the data, according to their relative depths, the deepest datum being the depth-median. In this way we have a standard procedure to define both empirical and population quantiles.

Let us briefly outline some basic ideas on this topic and how they can be adapted to the FDA framework. In the simplest case $\mathcal{X}=\mathbb{R}$, the usual (in various senses equivalent) depth functions are

$$D_0(P,x)=P(-\infty,x]P[x,\infty)\quad\text{and}\quad D_1(P,x)=\min(P[x,\infty),P(-\infty,x]).$$

In both cases, the corresponding median and quantiles lead to the usual definitions of these concepts. However, the depth notions are much more useful in the multivariate case $\mathcal{X}=\mathbb{R}^d$. The generalization of D_1 to this situation leads to the *Tukey's (or halfspace) depth* which is defined as the infimum of the probabilities of all closed half-spaces containing x. In other words, $D_T(P,x)$ is the infimum of all the D_1-depths of the one-dimensional projections of x, computed with respect to the corresponding one-dimensional marginals of P; see Zuo and Serfling (2000) for comparisons of D_T with other depth notions. A randomized version of D_T, aimed to deal with the high computational cost of D_T, has been proposed by Cuesta-Albertos and Nieto-Reyes (2008): let us take at random k iid projection directions $v_1,\ldots,v_k$ and define the *random Tukey depth* of x as

$$D_{RT}(P,x)=\inf_{1\le i\le k}D_1(P_{v_i},\langle v_i,x\rangle),$$

where P_{v_i} denotes the distribution of $\langle v_i,X\rangle$, the one-dimensional projection of X on the direction v_i. While this depth function turns out to be random (as it depends on the randomly chosen projection directions), the above mentioned generalization of Cramér–Wold Theorem (Cuesta-Albertos et al., 2007) provides a sound theoretical support for such methodology.

However, the interesting point here is the fact that the random Tukey depth can be extended (see Cuesta-Albertos and Nieto-Reyes, 2008), keeping an almost identical definition, to the cases where $\mathcal{X}$ is a separable Hilbert space (for example $\mathcal{X}=L^2[0,1]$). Obviously, the random vectors v_i should be replaced with elements in the continuous dual space $\mathcal{X}^*$. From Riesz Representation Theorem, these dual elements v are associated with kernel functions a defining the linear continuous operator $x\mapsto\int_0^1 a(s)x(s)\,ds$ (which replaces the finite-dimensional inner product). Hence, choosing the dual elements v's amounts to randomly pick up the corresponding kernel functions a's with an appropriate L^2 process.

A different approach to get an infinite-dimensional depth function from the one-dimensional depths of the projections is proposed by Cuevas and Fraiman (2009): the basic idea is just to define a new depth function by integrating out the one-

dimensional depths, rather than by calculating the maximum of them on a given number of projections. This leads to a definition of type

$$D(P,x) = \int D^1(P_f, f(x))\, dQ(f),$$

where D^1 is a one-dimensional depth function, f denotes an element in the continuous dual space, P_f is the (one-dimensional) distribution of $f(X)$ (the projection of X via the real linear continuous f) and Q is a probability measure on the continuous dual space of f's.

The practical aspects of this idea, as well as some comparisons with other methods, have been considered in Cuevas et al. (2007).

We shall not discuss here depth functions in further detail. Let us just remark the fact that every depth function has an associated notion of median (the deepest point), quantiles (defined in the obvious way after sorting the points according to their depths) and $\alpha-$trimmed mean (the average of the $100(1-\alpha)\%$ deepest points); see Fraiman and Pateiro-López (2012) for a recent new proposal relying on the use of projections.

Nowadays, the study of depth notions for high-dimensional and functional data is far from being an exhausted topic. Apart from the mentioned proposals, different ideas have been proposed: e.g., Chaudhuri (1996), Vardi and Zhang (2000), López-Pintado and Romo (2006, 2009), Li et al. (2012). Still, probably some more research is needed.

4.4. Impartial trimmed means

A different way (not involving depths or ranking in the observations) for defining trimmed means in general spaces is the impartial trimmed procedure introduced by Gordaliza (1991). This procedure has been adapted to the functional case by Cuesta-Albertos and Fraiman (2006); see the references of that paper for additional bibliography on the subject.

Given $\alpha\in(0,1)$ and a probability distribution P on the sample space $\mathcal{X}$, *the impartial* $\alpha-$trimmed *mean* m_P of P is defined by

$$\int \|x-m_P\|^2 \tau_P(x)\, dP(x) = \inf_{a\in\mathcal{X},\tau\in\mathcal{P}_\alpha} \int \|x-a\|^2 \tau(x)\, dP(x), \tag{6}$$

where $\mathcal{P}_\alpha$ is the set of $\alpha-$trim functions,

$$\left\{\tau : \mathcal{X}\to[0,1], \int \tau(y)\, dP(y) \geq 1-\alpha\right\}.$$

The empirical versions are obtained by replacing P with the corresponding empirical distributions. The idea behind (6) is to allow the distribution P (or the empirical $\mathbb{P}_n$) to decide what part of the sample space (or what part of the sample) must be trimmed. This accounts for the term "impartial". Sufficient conditions for the existence and qualitative robustness of m_P, as well as some computational issues are discussed in Cuesta-Albertos and Fraiman (2007).

5. Functional regression

According to the standard setup, a typical (random design) functional regression model has the form

$$Y = g(X) + \epsilon,$$

where X is the explanatory (functional) variable, Y is the output (response) variable, g is a (usually unknown) function and ϵ is the error which is often assumed to fulfill $\mathbb{E}(\epsilon|X) = 0$.

The aim is to estimate g from a random sample (X_i, Y_i), $i=1,\ldots,n$.

5.1. Functional linear regression with scalar response

We will consider the following model:

$$Y_i = \alpha + \int_0^1 \beta(t) X_i(t)\, dt + \epsilon_i, \quad i = 1,\ldots,n, \tag{7}$$

where $\beta\in L^2[0,1]$, $X_i = X_i(\omega,.)\in L^2[0,1]$ are iid with $\mathbb{E}\|X_i\|^2 < \infty$, ϵ_i are iid with $\mathbb{E}(\epsilon_i) = 0$ and $\mathbb{E}(X_i(t)\epsilon_i) = 0$, almost everywhere.

The main problem here is the estimation of β. A particularly insightful account of this model is given in the survey paper by Cardot and Sarda (2011). Among several other topics, these authors consider the estimation of β based on the functional principal components method. We summarize here the main ideas of this procedure. We need to introduce some concepts and notation.

Assume that $\gamma(s,t) = Cov(X_i(s), X_i(t))$ is continuous. The covariance operator of the underlying L^2-process $X(t)$ is given by

$$\Gamma u(t) = \int_0^1 \gamma(s,t) u(s)\, ds, \quad \forall u\in L^2[0,1].$$

In a equivalent way it can be expressed by

$$\Gamma u = \mathbb{E}(\langle X_i - \mathbb{E}X_i, u\rangle (X_i - \mathbb{E}(X_i)))$$

The operator Γ (see, e.g., Laha and Rohatgi, 1979) is non-negative, symmetric, Hilbert–Schmidt ($\sum_i \|\Gamma e_i\|^2 < \infty$) and hence compact. The eigenvalues λ_j of Γ are all positive and zero is their only accumulation point. If we denote by e_j the sequence of orthonormal eigenfunctions associated with the eigenvalues λ_j (we assume they are sorted in decreasing order), we may perform the above mentioned Karhunen–Loève decomposition,

$$X_i(t) - \mathbb{E}(X_i(t)) = \sum_{j=1}^{\infty} Z_j e_j(t),$$

where the Z_j are centered uncorrelated random variables with $V(Z_j) = \lambda_j$.

If we center the variables in (7) it can be seen (see Cardot and Sarda, 2011, p. 24 for details) that our function β must fulfill

$$\Delta = \Gamma\beta \tag{8}$$

in the sense that $\Gamma\beta$ coincides with $\mathbb{E}((X_1 - \mathbb{E}(X_1))(s)(Y_1 - \mathbb{E}(Y_1)))$, which is the kernel function which defines the operator

$$\Psi \mapsto \Delta\Psi = \mathbb{E}(\langle \Psi, (X_1 - \mathbb{E}(X_1))\rangle (Y_1 - \mathbb{E}(Y_1))).$$

Let us note the analogy with the ordinary multivariate regression model $Y = X\beta + \epsilon$, with $\beta \in \mathbb{R}^p$, for which β appears as the solution of the *normal equations* $X'X\beta = X'Y$. Expression (8) would be the functional analog of such equations. The trouble here is that the standard solution for the finite dimensional case ($\hat{\beta} = (X'X)^{-1}X'Y$) has no direct analog here since Γ is not an invertible operator as long as its range is an infinite-dimensional space. In particular, Eq. (8) does not identify β in a unique way, since for any solution β, we have that $\beta + \beta_0$ is also a solution, for any β_0 in the kernel of Γ (i.e., $\Gamma\beta_0 = 0$).

Then, we look for solutions in the closure of $\mathrm{Im}(\Gamma) = \{\Gamma x : \ x \in L^2[0,1]\}$; alternatively, we assume (w.l.o.g.) that $\ker(\Gamma) = \{0\}$. Expanding β in the orthonormal basis $\{e_i\}$ of eigenfunctions of Γ , we have $\beta = \sum_{j=1}^{\infty} \langle \beta, e_j\rangle e_j$. Using the normal equation (8),

$$\Delta e_j = \lambda_j \langle \beta, e_j\rangle, \quad j = 1, 2, \ldots \tag{9}$$

Eq. (9) implies

$$\beta = \sum_{j=1}^{\infty} \frac{\Delta e_j}{\lambda_j} e_j = \sum_{j=1}^{\infty} \frac{\mathbb{E}(Z_j(Y_1 - \mathbb{E}(Y_1))}{\lambda_j} e_j, \tag{10}$$

where Z_j is the j-coordinate of $X_1 - \mathbb{E}(X_1)$.

It turns out that the condition

$$\sum_{j=1}^{\infty} \frac{\mathbb{E}^2(Z_j(Y_1 - \mathbb{E}(Y_1))}{\lambda_j^2} < \infty$$

ensures the existence and uniqueness of a solution β for the normal equations in the closure of $\mathrm{Im}(\Gamma)$.

Thus, a *Functional Principal Component Regression Estimator* for β could be defined by

$$\hat{\beta} = \sum_{j=1}^{K} \frac{\Delta_n \hat{e}_j}{\hat{\lambda}_j} \hat{e}_j$$

where Δ_n is the empirical version of the cross-covariance operator Δ defined above, $\hat{\lambda}_j$ and $\hat{e}_j$ are the (K-largest) eigenvalues and eigenvectors of the empirical covariance operator Γ_n which estimates Γ. Of course, the infinite-dimensional Law of Large Numbers, established in Section 3.4, plays an important role in order to ensure the consistency of these estimators.

The above discussion just attempts to provide a sketch of some basic ideas involved in functional linear regression. Many more theoretical and practical details have been considered in the literature. For example, some authors have considered the above estimation procedure combined with a smoothing step (either in the curves X_i or in the estimator $\hat{\beta}$). We refer again to Cardot and Sarda (2011) for further information. A few additional references on this subject are Cardot et al. (2003), Cai and Hall (2006) and Crambes et al. (2009).

5.2. Functional linear regression with functional response

Here the basic model is

$$Y_i(t) = \int_0^1 \beta(s,t) X_i(s)\, ds + \epsilon_i(t), \quad i = 1, \ldots, n, \tag{11}$$

where $\beta(s,t)$ defines a Hilbert–Schmidt (hence, compact) integral operator, $x \mapsto \int_0^1 \beta(s,t)x(s)\, ds$ on $L^2[0,1]$, that is $\int_0^1 \int_0^1 \beta(s,t)^2\, ds\, dt < \infty$, $X_i = X_i(t) \in L^2[0,1]$ and typically $\epsilon_i(t)$ are iid L^2-processes with $\mathbb{E}(\epsilon_i) = 0$, and ϵ_i independent from (X_i, Y_i).

This model has been considered by Chiou et al. (2004) and Kokoszka et al. (2008), among others, under the usual assumption of *random design* for the explanatory variables that is, the X_i's are random variables not controlled by the user. In Cuevas et al. (2002) the model (11) has been analyzed assuming a fixed design. This assumption is very common in the standard multiple (finite-dimensional) regression theory. In the functional case a fixed design model could be justified, for example, in those cases where input functions $x(t)$ are given, non-random, signals entering in a communication channel and the outputs $Y(t)$ are the responses that the receiver gets, possibly distorted by the communication channel (through the linear operator β) and by the random noise $\epsilon(t)$. In this setup it makes sense to consider the so-called *calibration* (*or inverse regression*) *problem*, where a given unknown input $x_0(t)$ must be reconstructed from the responses $Y_1(t),\ldots,Y_n(t)$ obtained when it is repeatedly transmitted through the communication channel.

5.3. Functional nonparametric regression

Given a scalar random output Y and a (possibly functional) explanatory X it is natural to look for the "best approximation" of Y in terms of X. It is well-known that the minimizer (on the space of real measurable functions g such that $\mathbb{E}(g^2(X))<\infty$) of the L^2-mean error $\mathbb{E}(Y-g(X))^2$ is given by the *regression function*

$$\eta(x)=\mathbb{E}(Y|X=x).$$

The nonparametric estimation of η from a sample (X_i,Y_i) $i=1,\ldots,n$ is also possible (as in the case of an $\mathbb{R}^d-$valued X) by using suitable versions of the classical nonparametric regression estimators (kernel, k-NN, etc.). As pointed out in Section 1.1, in the statistical problems with finite-dimensional data, the term "nonparametric" roughly implies that the target of the inference is not restricted to live in a finite-dimensional space. In the FDA setup, the situation is not so clear since, as we have seen in the previous subsection, even the strong assumption that $\eta(X)$ is a linear continuous functional does not restrict the estimation to a finite dimensional space: a function β in $L^2[0,1]$ must be estimated. Still, we will use here the word "nonparametric" in a broader sense to imply that no strong assumption on the "shape" of the target function η is made, apart from standard analytic assumptions of continuity, differentiability, the conditions of finiteness of moments, etc.

Most nonparametric estimators have the form of a weighted average of the responses

$$\hat{\eta}(x)=\sum_{i=1}^{n} W_{ni}(x)Y_i, \tag{12}$$

where, in general, the weights $W_{ni}(x)$ will depend on the $X_1,\ldots,X_n$, especially from those closer to x.

A typical choice would be the Nadaraya–Watson weights (considered above, from a different point of view, as a possible tool for data smoothing),

$$W_{ni}(x)=\frac{K\left(\frac{D(x,X_i)}{h}\right)}{\sum_{j=1}^{n}K\left(\frac{D(x,X_j)}{h}\right)}, \tag{13}$$

where D denotes some distance or semi-distance, h is a smoothing parameter and K is a kernel function such as $K(t)=\frac{3}{2}(1-t^2)\mathbb{I}_{[0,1]}(t)$ or $K(t)=2\exp(-t^2/2)$. The choice (13) leads to the so-called *Nadaraya–Watson kernel estimator*.

Under some conditions, which include $|\eta(x)-\eta(x')|\leq CD(x,x')^s$, for some $C,s>0$, and

$$h:=h_n\to 0 \quad \text{and} \quad \frac{\log n}{n\phi_x(h)}\to 0, \tag{14}$$

a convergence rate of type

$$\hat{\eta}(x)-\eta(x)=O(h^s)+O\left(\sqrt{\frac{\log n}{n\phi_x(h)}}\right), \quad \text{a.s.} \tag{15}$$

holds (here $\phi_x(h)$ denotes the small ball probability defined in Section 3.7). See Ferraty and Vieu (2006, Chapter 6) and Ferraty and Vieu (2011) for details. Since, in most usual spaces, (14) requires to take slowly decreasing sequences h_n, we conclude that result (15) entails that the convergence of the estimator $\hat{\eta}(x)$ to the true value $\eta(x)$ will typically be very slow.

The k-nearest neighbors (k-NN) regression estimators are obtained from (12) with the weight choice

$$W_{ni}(x)=\frac{\mathbb{I}_{B(x,C_k(x))}(X_i)}{k}, \tag{16}$$

where $C_k(x)$ is the distance from x to the k-th nearest sample observation among $X_1,\ldots,X_n$. The intuitive idea is very simple: we estimate $\eta(x)$ as a local average of the responses Y_i whose corresponding X_i are "close" (i.e., among the k closest) to x. Here $k=k_n$ is an integer-valued smoothing parameter. In order to get consistency one typically needs $k_n\to\infty$ and $k_n/n\to 0$, as $n\to\infty$.

Convergence rates for the L^2-error $\mathbb{E}[(\hat{\eta}(X)-\eta(X))^2]$ of the k-NN estimator are given in Biau et al. (2010) for the case that X takes values in a separable Banach space. Of course, additional regularity conditions on the regression function and the underlying distribution are needed: different results are given in Biau et al. (2010) for Sobolev spaces, Besov spaces and reproducing kernels Hilbert spaces.

It is well-known that the finite-dimensional regression estimates (of kernel or k-NN type) enjoy the property or *weak universal consistency*. This means that, under some mild conditions on the kernel K and the smoothing parameter ($h\to 0$ and $nh^d\to\infty$), the kernel regression estimator fulfills $\mathbb{E}\|\hat{\eta}(X)-\eta(X)\|_2^2\to 0$ under the sole assumption that $\mathbb{E}(Y^2)<\infty$; see Györfi et al. (2002, p. 71). Similarly for the k-NN regression estimators the weak universal consistency holds whenever $k_n\to\infty$ $k_n/n\to 0$, and $\mathbb{E}(Y^2)<\infty$, provided that the ties in the data occur with probability 0. An important conceptual difference between the finite-dimensional regression estimates (of kernel or k-NN type) and their functional counterparts is the lack of this property of universal consistency in the infinite-dimensional framework. A recent study of this topic can be found in the paper by Forzani et al. (2012). We will comment more on this in the following section, in the supervised classification setting.

5.4. Optimal linear approximations and projections

The finite-dimensional theory of linear regression is known to have an elegant interpretation in terms of projections. These geometric ideas can be partially translated to the functional setting but, again, some difficulties arise, associated with the passage to infinite dimensions.

We next summarize (taking Chapter 10 in Bosq and Blanke, 2007 as a source) some relevant facts in this connection:

(a) Let $\mathbb{H}$ be a real separable Hilbert space. Let us denote by $L^2_{\mathbb{H}}=L^2_{\mathbb{H}}(\Omega,\mathcal{A},\mathbb{P})$ the space of $\mathbb{H}-$valued random elements X defined on the probability space $(\Omega,\mathcal{A},\mathbb{P})$ with $\mathbb{E}\|X\|^2<\infty$. Let us denote by $\mathcal{L}$ the space of linear continuous operators $l:\mathbb{H}\to\mathbb{H}$. Given X, $Y\in L^2_{\mathbb{H}}$, if we are looking for "the best linear approximation" of Y in terms of X, a natural choice would be $\Pi^X(Y)=\lambda(X)$, the orthogonal projection of Y on $\overline{\mathrm{sp}(X)}$, the linear closed subspace of $L^2_{\mathbb{H}}$ defined as the closure of $\{l(X):l\in\mathcal{L}\}$.

(b) By projecting on a closed subspace we guarantee the existence of $\lambda(X)$, although λ might fail to be itself linear and continuous. A necessary and sufficient condition (see Bosq and Blanke, 2007, p. 233) to ensure that there exists $l\in\mathcal{L}$ such that $l(X)=\lambda(X)$ a.s. is $\|\Gamma_{X,Y}(x)\|\le\alpha\|\Gamma_X(x)\|$ for some $\alpha\ge 0$ and every $x\in\mathbb{H}$, where $\Gamma_{X,Y}$ stands for the cross-covariance operator,

$$\Gamma_{X,Y}(x)=\mathbb{E}(\langle X-\mathbb{E}X,x\rangle(Y-\mathbb{E}Y)),$$

and $\Gamma_X=\Gamma_{X,X}$.

(c) In any case, one may always find a sequence of linear continuous operators $l_n\in\mathcal{L}$ and a linear subspace $\mathcal{V}\subset\mathbb{H}$ with $P_X(\mathcal{V})=1$ such that $l_n(x)\to\lambda(x)$ for all $x\in\mathcal{V}$. See Bosq and Blanke (2007, p. 231) for details.

(d) It is a well-known elementary result that, if (X,Y) is a Gaussian random variable taking values in $\mathbb{R}^d\times\mathbb{R}$, then the regression function $\eta(x)=\mathbb{E}(Y|X=x)$ is linear. An analogous result for the case where (X,Y) takes values in $\mathbb{H}\times\mathbb{H}$ (where $\mathbb{H}$ is a separable Hilbert space) is given in Mandelbaum (1984, Theorem 2).

5.5. Functional analysis of variance

The study of the classical *analysis of variance models* is closely associated with the theory of linear models. The functional counterpart of this theory has also received a considerable attention in the literature, starting from the simplest one-way model and including other more sophisticated two-ways models with interaction as well as the analysis of some real-data examples. Some references are Abramovich et al. (2004), Abramovich and Angelini (2005), Antoniadis and Sapatinas (2007), Cuevas et al. (2004), Cuesta-Albertos and Febrero-Bande (2010), Schott (2007) and Spitzner et al. (2003).

5.6. Other functional regression topics

Logistic regression models with a functional explanatory variable have been considered, among others, by Aguilera et al. (2006) and Lindquist and McKeague (2009).

Autoregressive linear models for functional time series are studied in Bosq (2000). Two recent references are Horváth et al. (2010) and Ruiz-Medina (2012).

Some measures to detect influential observations in the functional linear model are analyzed in Febrero-Bande et al. (2010).

The so-called Receiver Operating Characteristic (ROC) curve is an increasingly popular tool aimed to evaluate the performance the statistical processes of binary decision (in particular, in binary classification: see the next section). Inácio et al. (2012) have analyzed the extension of the ROC methodology for the case where there is a functional covariate information available.

The problem of detecting a change, during the observation period, in the operator which defines a functional linear model has been addressed by Horváth and Reeder (2012); see also Berkes et al. (2009).

6. Supervised and unsupervised functional classification

In statistics, the word *classification* has also the same usual double meaning as in the ordinary language, where this term stands for both "to assign (an element) to a particular class or category" and for "arrange (a group of elements) in classes according to shared characteristics". The first meaning would correspond to the statistical methodology called *supervised classification* or *discriminant analysis*. The second one would better suit to the *clustering methodology* which roughly corresponds to the unsupervised classification theory. Here the term "supervised" in the first problem refers to the fact that there is a "training" (sample) data set of elements which are assumed to be well-classified. Then the problem is to classify the new incoming elements. In the unsupervised class no such help is available: the problem is just to group the data into "clusters" of mutually alike elements.

A recent survey of classification methods in FDA is given in Baíllo et al. (2011b). We provide in this section an updated summary of that survey.

6.1. Functional discrimination

6.1.1. Statement of the problem; similarities and differences with the multivariate case

Suppose that a random element X, taking values on the sample space $\mathcal{X}$, can be observed in two populations P_0 and P_1. Denote by μ_i the distributions $X|Y=j$, $j=0,1$, where Y is a dichotomous variable indicating the membership to the population 1 or 0. The available data consist of a sample of independent observations $\{(X_i, Y_i), 1 \le i \le n\}$.

The problem is to decide, from the information provided by the training sample, if a new observation $X=x$ (for which the value of Y is not known) has been taken in P_0 or in P_1.

In the classical (multivariate) case, the sample space $\mathcal{X}$ where the random observations X_i take values is just the d-dimensional Euclidean space $\mathcal{X}=\mathbb{R}^d$. We are however concerned here with the (infinite-dimensional) functional situation in which $\mathcal{X}$ is a function space.

An example of functional classification is given by the ECG data set. It is available in http://alumni.cs.ucr.edu/~wli/selfTraining/. This data set consists of 2026 electrocardiogram curves recorded during one heartbeat. 1506 were labeled as "normal" and 520 are "pathological". Fig. 2 shows a few of these curves. The purpose of functional discrimination in this context would be to implement an "automatic" method aimed to classify a new incoming electrocardiogram curve as "normal" or not, on the basis of the data set of well-classified curves.

The mathematical aim is to find a "classifier" (or "classification rule") $g : \mathcal{X} \to \{0, 1\}$ that minimizes, at least asymptotically, the classification error, or "risk", $\mathbb{P}(g(X) \neq Y)$. It is well-known (e.g., Devroye et al., 1996, p. 11) that the optimal classification rule (sometimes called "Bayes rule") is

$$g^*(x) = \mathbb{I}_{\{\eta(x) > 1/2\}}, \tag{17}$$

where $\eta(x) = \mathbb{E}(Y|X=x)$. The minimal classification error $L^* = \mathbb{P}(g^*(X) \neq Y)$ is called *the Bayes error*. The Bayes rule is usually unknown but it can be approximated, in different ways, from the training data.

Expression (17) shows that the (supervised) classification and regression problems are quite close to each other. In fact, any estimator $\hat{\eta}(x)$ of the regression function provides a data-driven classifier by just replacing $\eta(x)$ with the corresponding estimator $\hat{\eta}(x)$. This is the so-called *plug-in methodology*. Thus, if we replace in (17) the regression function $\eta(x)$ with a nonparametric estimator of type (12) with weights (13) or (16), we get, respectively, a kernel and a k-NN classifier. The latter is simply defined as a "majority vote" rule: an observation x will be assigned to whenever the majority of the k sample observations closest to x belong to P_0 (ties are randomly broken).

6.1.2. On the linear classification procedures in the FDA setting

So far, the landscape we have outlined for functional discrimination is not very different from the analogous finite-dimensional problem: in particular, the kernel or k-NN classifiers are obtained in a similar way just replacing the Euclidean distance with an appropriate functional metric. However, the passage from the classical multivariate (typically low-dimensional) discrimination problem to the infinite-dimensional setting entails several important challenges. To begin with, the Fisher linear discriminant rule (which is by far the most popular in multivariate discrimination; see, e.g., Devroye et al., 1996, Chapter 4) cannot be straightforwardly extended to the functional case or to the high-dimensional case with highly correlated explanatory variables. The reason again (see Section 5.1) is the non-invertibility of the covariance operator. Hastie et al. (1995) tackle this problem using regularized versions of the covariance matrix aimed to achieve invertibility. Other adaptations of linear discrimination ideas to the functional setting can be found in James and Hastie (2001), Shin (2008) and Delaigle and Hall (2012a).

6.1.3. The lack of universal consistency

Another important difference between the finite-dimensional discrimination problem and its FDA counterpart concerns the important theoretical issue of consistency. Recall that a sequence

$$g_n(x) = g_n((X_1, Y_1), \ldots, (X_n, Y_n); x) \quad \text{with } g_n : \mathcal{X} \longrightarrow \{0, 1\}$$

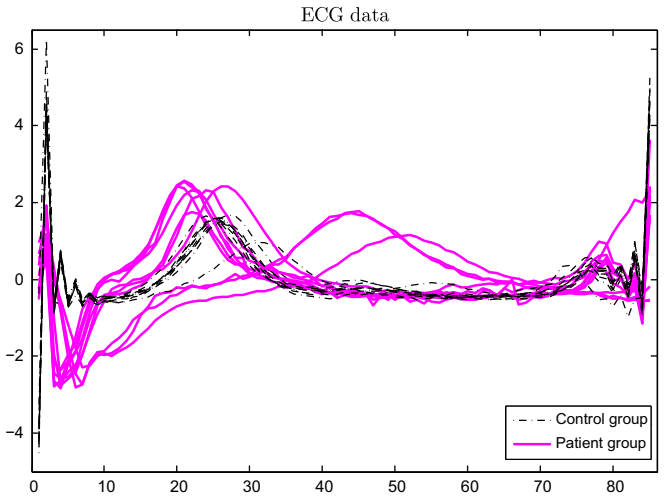

Fig. 2. Some ECG curves from the patients and the control group.

of classifiers is said to be *weakly consistent* if the *conditional classification error*

$$L_n = \mathbb{P}(g_n(X) \neq Y | (X_1, Y_1), \ldots, (X_n, Y_n))$$

fulfills

$$L_n \xrightarrow{P} L^* \quad \text{or, equivalently,} \quad \mathbb{E}(L_n) \longrightarrow L^*.$$

In the case $\mathcal{X} = \mathbb{R}^d$, the k-NN classifiers are weakly universally consistent, provided that $k = k_n \to \infty$ and $k/n \to 0$. This is a consequence of a celebrated result by Stone (1977) (see also Devroye et al., 1996, Chapter 11, for details). As for the kernel classifiers, under some mild assumptions on the kernel K, the strong universal consistency (i.e., $L_n \xrightarrow{a.s.} L^*$) holds (see Devroye et al., 1996, Theorem 10.1).

Unfortunately, these remarkable results of universal consistency are no longer valid in the functional setting. There are, however, some sufficient conditions for consistency in the recent literature. Thus, Cérou and Guyader (2006) have proved the following result for the k-NN classifier:

Theorem (*Cérou and Guyader, 2006*). *If $\mathcal{X}$ is a separable metric space, then $k \to \infty$ and $k/n \to 0$ entails $\mathbb{E}(L_n) \longrightarrow \mathbb{E}(L^*)$, provided that the following Besicovich condition holds*:

$$\frac{1}{P_X(B(X,\delta))} \int_{B(X,\delta)} |\eta - \eta(X)| \, dP_X \xrightarrow{\mathbb{P}} 0 \quad \text{as } \delta \to 0, \tag{18}$$

where $\xrightarrow{\mathbb{P}}$ stands for convergence in probability and P_X denotes the distribution of X.

Condition (18) is clearly reminiscent of the classical Lebesgue Differentiation Theorem (which always holds in $\mathbb{R}^d$). It can be seen that the continuity of η is a sufficient condition for (18). An analogous result has been proved by Abraham et al. (2006) for kernel classifiers (under some mild assumptions for the kernel K), assuming a condition similar to (18).

6.1.4. About depth-based classification

Given a (sample) depth measure $D(\mathbb{P}_n, x)$, if we want to classify a new incoming x_0, we might just evaluate the depth of x_0 in both sub-samples (from P_0 and from P_1) and assign x_0 according to the data set where it is more deeply placed. Typically the depth-based methods will fail in those cases where the populations are "nested" or extremely heteroscedastic. The recent paper by Li et al. (2012) offers and interesting proposal (still relying on depth functions through the so-called *DD-plots*) to successfully tackle these problems. Some other recent references on the use of depth functions in supervised classification are Cuevas et al. (2007) and López-Pintado and Romo (2006). The manuscript by Sguera et al. (2012) includes an quite extensive simulation study of many different proposals for depth-based classification.

6.1.5. A benchmark method for functional classification?

The paper by Hand (2006) convincingly argues, based on purely practical grounds, that the time-honored Fisher's linear classifier (dating from 1936) is still the reference for most users of finite-dimensional discrimination methods. Under the provocative title "Classifier technology and the illusion of progress", this author points out how in real-life low-dimensional classification problems, Fisher's method is not easily beaten by other more sophisticated classifiers, even if these are much better in some specific examples. It is not yet clear what method (if any) might play this benchmark role in high-dimensional or functional problems. The limited experience available suggests that k-NN (see Cuevas et al., 2007; Sguera

et al., 2012) and Partial Least Squares (a dimension reduction method: see Section 7.2 below) combined with linear classification provide quite good results in a broad range of situations.

6.1.6. The Gaussian case

As commented in Section 3.3, the special difficulties for the use density functions in the functional settings have some practical consequences. One of them is the corresponding lack of explicit expressions for optimal classification rules in FDA.

However, the simple expression (3) for $\eta(x)=\mathbb{P}(Y=1|X=x)$ allows us to identify the optimal rule, as long as we are able to give an explicit expression for the Radon–Nikodym derivative $d\mu_0/d\mu_1$ of the process μ_0 with respect to μ_1. It turns out that this is possible in some special cases of Gaussian processes with "triangular" covariance functions of type $\gamma(s,t)=u(\min(s,t))v(\max(s,t))$ (Varberg, 1961; Jørsboe, 1968). Of course, these Radon–Nikodym derivatives (and hence the optimal rule) depend on some unknown features of the process distribution (as, for example, the mean function) but these can be estimated. As a consequence we have a sort of functional version of the optimal parametric plug-in rules of the finite-dimensional cases. The theoretical and practical aspects of these ideas have been explored in Baíllo et al. (2011a).

6.2. Unsupervised functional classification (clustering)

Some well-known algorithms (hierarchical algorithms, k-means, etc.) for grouping in data clusters a given data set $\{X_1,\ldots,X_n\}$ are based on the mutual distances between the data. So they can be adapted, in a more or less straightforward way, to the case of infinite-dimensional data, provided that a suitable distance is defined. The difficulties arise mainly regarding the interpretation in population terms. For example, Hartigan's (1975) definition of a (population) cluster as a connected component of the λ-density level set $\{f\geq\lambda\}$ requires to have a density function (with respect to some natural dominant measure) characterizing the distribution of the data X_i. The hurdles associated with handling and estimation of infinite-dimensional densities have been commented above. Thus this definition is not operative in the functional setting. Also, some natural consistency results turn out to be more involved when established in infinite-dimensional versions.

In the rest of this subsection we will restrict the discussion to the so-called k-means procedure, perhaps the most popular clustering method. As we have just commented, the formal definition and its sample version for the functional setting are much the same as the corresponding multivariate versions.

Given a probability measure P on the space $\mathcal{X}$, an $\mathcal{X}-$valued random element X with distribution P and $k\in\mathbb{N}$, let us define the k-mean parameter, associated with P as any set $\{h_1^P,\ldots,h_k^P\}$ of k *cluster centers*, $h_i^P\in\mathcal{X}$ minimizing (on all possible sets $\{h_1,\ldots,h_k\}$, with $h_i\in\mathcal{X}$) the following expression:

$$I_k(P;h_1,\ldots,h_k):=\mathbb{E}\left(\min_{i=1,\ldots,k}\|X-h_i\|^2\right) \tag{19}$$

The intuitive idea is very simple: we are just looking for the cluster centers h_i such that the expected distance from a random observation X to its nearest center in $\{h_1,\ldots,h_k\}$ is minimal.

The sample version of (19), based on data $X_1\ldots,X_n$, leads to minimize on $\{h_1,\ldots,h_k\}$ the natural empirical approximation of $I_k(P;h_1,\ldots,h_k)$, that is,

$$I_k(\mathbb{P}_n;h_1,\ldots,h_k):=\frac{1}{n}\sum_{i=1}^{n}\|X_i-h_{c(i)}\|^2, \tag{20}$$

where $h_{c(i)}$ denotes the cluster center, in $\{h_1,\ldots,h_k\}$, closest to X_i.

Now, given the set of cluster centers $\{h_1^n,\ldots,h_k^n\}$, those observations having the same closest center are in the same cluster.

Of course, the exact computation of the optimal cluster centers, even in the empirical version (20), is a formidable task. Different approximate (often randomized) algorithms have been proposed.

The consistency (in the Hausdorff metric) of the sequence $\{h_1^n,\ldots,h_k^n\}$ to the population optimum $\{h_1^P,\ldots,h_k^P\}$ has been proved by Cuesta-Albertos and Matrán (1988) assuming that $\mathcal{X}$ is a uniformly convex Banach space.

Some interesting results of convergence of the "empirical minimal error" $I_k(\mathbb{P}_n;h_1,\ldots,h_k)$ to its population counterpart $I_k(P;h_1,\ldots,h_k)$ have been obtained by Biau et al. (2008).

The lack of robustness is known to be a drawback of the k-means procedure. A remedy could be to replace the square norm in (19) with another function Φ giving a smaller weight to the extreme observations. Thus, the criterion would become

$$I_k(P;h_1,\ldots,h_k):=\mathbb{E}\left(\Phi\left(\min_{i=1,\ldots,k}\|X-h_i\|\right)\right). \tag{21}$$

Additionally, or alternatively, the impartial trimming procedure commented in Section 4.4 can be also used to define *impartial trimmed k-means*. The adaptation of this idea to functional data has been considered in Cuesta-Albertos and Fraiman (2006, 2007).

A few other references on clustering for functional data are Heckman and Zamar (2000), James and Sugar (2003), Serban and Wasserman (2005) and Tarpey (2007).

7. On dimension reduction techniques in FDA

A natural idea in high-dimensional or functional data is to transform the sample data into elements of small dimensional spaces, thus allowing for a simpler statistical treatment. Of course the use of projections via linear functionals, either systematic or random (see Sections 3.6 and 4.3) is a possible, sometimes very useful, option. Let us now consider here other more standard versions of this idea (still relying on linear projections) which are explicitly based on the covariance structure of the data.

A non-linear approach to dimension reduction in functional data problems can be found in Chen and Müller (2012).

7.1. Functional principal components

Let X be a random element taking values in the sample space $\mathcal{X}=L^2[0,1]$. By analogy with the finite-dimensional case, the aim of *Functional Principal Components* (FPC) is to define orthonormal projection directions $\alpha_1,\ldots,\alpha_k\in L^2[0,1]$ such that the projections of X along these directions take as much variability as possible.

Thus the first principal component is given by the projection direction α_1 achieving maximum variance,

$$V(\langle\alpha_1,X\rangle)=\max\{V(\langle a,X\rangle):\ \|a\|=1\}$$

and, for $k>1$, the k-th principal component is defined by

$$V(\langle\alpha_k,X\rangle)=\max\{V(\langle a,X\rangle):\ \|a\|=1,\ \langle a,\alpha_j\rangle=0,\ \text{for } j=1,\ldots,k{-}1\} \tag{22}$$

Then, the essential idea would be to replace in the statistical treatment the original data X_i with the corresponding k dimensional vector of projections $(\langle\alpha_1,X_i\rangle,\ldots,\langle\alpha_k,X_i\rangle)$. As in the finite-dimensional case, it can be shown that the FPC directions α_j turn out to be an orthonormal basis of the covariance operator associated with the kernel function $\gamma(s,t)=\mathrm{Cov}(X(s),X(t))$. Also, it directly follows that the corresponding eigenvalues λ_j fulfill $\lambda_j=V(\langle\alpha_j,X\rangle)$.

The estimation of the FPC directions α_j and variances λ_j can be done using an appropriate estimator of the covariance operator. The study of the corresponding estimators $\hat{\lambda}_j$ and $\hat{\alpha}_j$ and their asymptotic properties was done by Dauxois et al. (1982), a pioneering reference in FDA.

Thus we would have an empirical version of the optimization problem (22) or, equivalently, to look for the functions $\hat{\alpha}_j$ with $\|\hat{\alpha}_j\|=1$ fulfilling

$$\Gamma_n\alpha_j=\lambda_j\alpha_j, \tag{23}$$

for some $\hat{\lambda}_j$, where Γ_n is an empirical estimator of Γ: the simplest one would be the operator associated with the empirical covariance function

$$\gamma_n(s,t)=\frac{1}{n}\sum_{i=1}^{n}((X_i(s)-\overline{X}(s))(X_i(t)-\overline{X}(t))).$$

However, in practice, the search for a solution of this problem requires to exclude from consideration the possibly too unsmooth solutions. This can be done at least in two ways:

(a) To smooth the data $X_i(t)$ and to use the "smoothed empirical" $\tilde{\gamma}_n$ associated with the smoothed data. For example the smoothing process could be done by convolution: we could define $X_{ih}=\int_0^1 K_h(t-s)X_i(s)\,ds$, K_h being an appropriate kernel, e.g., the Gaussian density $N(0,h^2)$. Then, the covariance operator would be estimated with the smoothed empirical version

$$\tilde{\gamma}_n(s,t)\coloneqq\gamma_{nh}(s,t)=\frac{1}{n}\sum_{i=1}^{n}((X_{ih}(s)-\overline{X}_h(s))(X_{ih}(s)-\overline{X}_h(t))).$$

This method has been analyzed by Boente and Fraiman (2000) by assuming a "densely recorded" setting where the values of the $X_i(t)$ are available at a grid of arbitrarily close points $t_1,\ldots,t_N$. A more recent paper by Hall et al. (2006) considers also the sparse model $Y_{ij}=X(t_{ij})+e_{ij}$, $j=1,\ldots,m_i$, $i=1,\ldots,n$.

(b) To solve a modified version of the optimization problem (22) aimed to penalize the "rough" solutions. The proposal by Silverman (1996) is to maximize

$$\frac{V(\langle\alpha,X\rangle)}{\|\alpha\|^2+\delta\|\alpha''\|^2},$$

$\delta>0$ being a roughness penalty.

Some additional references on FPC, using both splines smoothing, in the evaluation of the sampling FPC are James et al. (2000) and Zhou et al. (2008). The recent paper by Li and Hsing (2010) uses local linear smoothing.

Some robustness issues, as well as an interesting practical example, are discussed in Locantore et al. (1999).

7.2. Partial least squares

Partial least squares (PLS) is an increasingly popular method of dimension reduction which can be used in those statistical problems (regression, supervised classification, etc.) where an explanatory variable X and a response Y are involved. The basic idea is clearly reminiscent of that of Principal Components Analysis: we want to select "special" directions suitable to project the data along them retaining as much information as possible. The main difference is that in the PLS scenario the output variable is taken into account. In the multivariate case (where X and Y take values in $\mathbb{R}^d$ and $\mathbb{R}^q$, respectively) the first pair of PLS directions would be given by the unit vectors $b_1 \in \mathbb{R}^d$ and $c_1 \in \mathbb{R}^q$ maximizing

$$\frac{\mathrm{cov}^2(\langle b, X\rangle, \langle c, Y\rangle)}{\|b\|^2\|c\|^2}$$

an the remaining directions b_k, c_k are chosen in a similar way by imposing the orthogonality condition $b'_k b_j = 0$, for $j \neq k$.

The original PLS idea goes back to Wold (1975). However, initially this methodology was mainly used by applied statisticians, especially in chemometrics. The papers by Frank and Friedman (1993) and Barker and Rayens (2003), among others, helped to popularize the PLS methodology among the whole statistical community. We refer to these papers for important details about how the PLS method can be interpreted in terms of a penalized canonical correlation procedure and can be equivalently defined by minimizing an appropriate sum of squared residuals.

The PLS ideas (implemented in terms of sequential algorithms: see below) have been recently adapted to different functional settings, including regression (Preda and Saporta, 2005; Escabias et al., 2007; Reiss and Ogden, 2007) and supervised classification (Preda et al., 2007; Delaigle and Hall, 2012a).

The recent paper by Delaigle and Hall (2012b) provides an interesting perspective, as well as some new proposals, in the application of PLS techniques in the FDA scenario. We briefly summarize here the presentation of the PLS method (in terms of a minimization problem) given in Section 2 of that paper: the explanatory variable X and the scalar output Y are supposed to follow the standard functional regression model with scalar response,

$$Y = \alpha + \int_0^1 \beta(t)X(t)\,dt + \epsilon,$$

where $\alpha \in \mathbb{R}$, $\beta \in L^2[0,1]$, $\mathbb{E}(\epsilon|X) = 0$ and $X = X(t)$ is a L^2 process on $[0,1]$ with covariance function $\gamma(s,t)$. As usual, the aim is to estimate the regression function

$$\eta(x) = \mathbb{E}(Y|X = x) = \alpha + \int_0^1 \beta(t)x(t)\,dt$$

We will tackle the problem looking for a PLS basis of functions $\Psi_1, \ldots, \Psi_p, \ldots$ whose elements are iteratively chosen: the p-th PLS function Ψ_p is defined by maximizing on Ψ the functional

$$f(\Psi) = \mathrm{cov}(Y - \eta_{p-1}(X), \int_0^1 X(t)\Psi(t)\,dt), \tag{24}$$

subject to $\int_0^1\int_0^1 \Psi_p(s)\gamma(s,t)\Psi_p(t)\,ds\,dt = 1$ and $\int_0^1\int_0^1 \Psi_p(s)\gamma(s,t)\Psi_j(t)\,ds\,dt = 0$, for $j = 1, \ldots, p-1$, where $\eta_0(x) = \mathbb{E}(Y)$ and η_{p-1} is an approximation to $\eta(x)$ based on the previously obtained PLS functions $\Psi_1, \ldots, \Psi_{p-1}$, defined by

$$\eta_{p-1}(x) = \mathbb{E}(Y) + \sum_{j=1}^{p-1} \beta_j \int_0^1 (x(t) - m(t))\Psi_j(t)\,dt,$$

where $m(t) = \mathbb{E}(X(t))$, and the β_j are chosen in order to minimize on $v_1, \ldots, v_{p-1}$ the expected error

$$\mathbb{E}\left\{\left(\int_0^1 \beta(t)(X(t) - m(t))\,dt - \sum_{j=1}^{p-1} v_j \int_0^1 (X(t) - m(t))\Psi_j(t)\,dt\right)^2\right\}. \tag{25}$$

However, in practice, one would rather minimize the natural empirical version of (25) given by

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \overline{Y} - \sum_{j=1}^{p-1} v_j \int_0^1 (X_i(t) - \overline{X}(t))\Psi_j(t)\,dt\right)^2,$$

and, of course, the covariance function $\gamma(s,t)$ and the "covariance operator" in (24) must also be estimated with their empirical counterparts. The, non-trivial, algorithmic aspects involved in the estimation of the PLS basis $\{\Psi_j\}$ are also discussed in Delaigle and Hall (2012b).

So, as a result of the PLS process we end up with p members of the PLS basis $\Psi_1, \ldots, \Psi_p$ which can be used either to give an estimator $\eta_p(x)$ of the regression function $\eta(x)$ or to project the sample data along these directions thus replacing every infinite-dimensional datum x with its projections $(\langle x, \Psi_1\rangle, \ldots, \langle x, \Psi_p\rangle)$.

As a conclusion, it could be said that PLS is a promising methodology in the FDA scenario. In particular, the results in functional classification (e.g., Preda et al., 2007) are quite positive.

7.3. Variable selection

Variable selection can be seen as a radical way of dimension reduction, aimed to identify a (usually small) subset of "really important variables". In the FDA setup the aim would be just to replace the original data, of type $x(t), t\in[0,1]$ with low-dimensional vectors $(x(t_1),\ldots,x(t_k))$, where the points $t_1,\ldots t_k$ are common for all the available data and are selected according to the statistical technique (regression, classification, etc.) we are dealing with. An obvious advantage with respect to other dimension reduction procedures (e.g., those based on the use of general linear projections) is the ease of interpretability, as the reduction is made keeping the original variables rather than transformations of them.

The interest of this topic has been boosted by the increasing demand of statistical treatment for high-dimensional genomic data. The case of functional data is a bit different since they allow for the use of natural "functional" assumptions such as smoothness. However some ideas of variable selection can be considered in both frameworks.

A natural idea, used in different versions, arises in the setup of multivariate linear models $Y=X_1\beta_1+\cdots+X_d\beta_p+\epsilon$ with a large number p of explanatory variables. The variables X_i for which $\beta_i\simeq 0$ are clearly non-relevant in this model. A popular approach to the selection of the relevant variables is the so-called LASSO method, originally proposed by Tibshirani (1996). It is based on the idea of modifying the ordinary least-squares estimator of β in order to automatically get a large number of zeros in the estimated β_i's. More specifically, the lasso estimator is defined as

$$\hat{\beta}=\underset{\beta}{\operatorname{argmin}}\left(\sum_i\left(Y_i-\sum_j\beta_jX_{ij}\right)^2\right)\quad\text{subject to }\sum_j|\beta_j|\le\lambda,$$

where λ is a shrinking parameter ("lasso parameter") whose data-driven selection is discussed in Tibshirani (1996). The geometric motivation behind this procedure is also nicely illustrated in that paper, Section 2.3 (by considering the two-dimensional case): whereas the objective function has ellipsoidal contours, the constraint region is a rotated square. If we move the contours curves, the lasso solution arises in the "first touch" with the constraint region. This touching point will typically correspond to a corner, corresponding to a zero coefficient. In general, it is argued in Tibshirani (1996, p. 60) that, if $\hat{\beta}_j^0$ denote the full (unconstrained) least squares estimators of β_j, and $\sum_j\hat{\beta}_j^0=\lambda_0$, then a choice $\lambda=\lambda_0/2$ amounts roughly to finding the best subset of size $d/2$.

Other related methods, based on different penalty functions are available: see Fan and Lv (2010) for an overview. In particular the so-called *Dantzig selector*, proposed by Candes and Tao (2007) has received a considerable attention. It has been proved by Bickel et al. (2009) that the LASSO and the Dantzig methods are asymptotically equivalent. The use of related ideas in the functional scenario has been recently analyzed by Kneip and Sarda (2011).

8. The bootstrap in the functional setup

Today the bootstrap is perhaps the most popular among the resampling methodologies. The basic ideas behind bootstrap are now a commonplace for the statistical community. However, in order to present here a survey of results on functional bootstrap we need to briefly recall some fundamentals.

It is well-known that, essentially, the bootstrap is a tool aimed to approximate the sampling distribution of a (usually re-scaled) statistic. The goal of such approximations is often to construct confidence intervals or critical regions for hypothesis testing. In other cases, the bootstrap is used to assess variabilities (in terms of mean square error, quantile ranges, etc.).

Let $X_1,\ldots,X_n\ldots$ be iid $\mathcal{X}-$valued random observations drawn from a distribution P on $\mathcal{X}$. As usual, denote by $\mathbb{P}_n$ the empirical probability measure associated with $X_1,\ldots,X_n$. Suppose we are interested in the probability law under P, $\mathcal{L}(R_n;P)$, of a statistic $R_n:=R(X_1,\ldots,X_n;P)$. For example, R_n could be the pivotal quantity used to construct a confidence region. The bootstrap method follows just as a systematic exploitation of the plug-in paradigm: if some element of interest depends on the unknown distribution P, just replace *everywhere* P with the empirical $\mathbb{P}_n$.

Thus, the unknown law $\mathcal{L}(R_n;P)$ can be estimated by the distribution, $\mathcal{L}(R_n;\mathbb{P}_n)$ of $R(X_1^*,\ldots,X_n^*;\mathbb{P}_n)$, where $X_1^*,\ldots,X_n^*$ denotes a iid (bootstrap) sample drawn from $\mathbb{P}_n$.

Of course, the point is that the bootstrap distribution $\mathcal{L}(R_n;\mathbb{P}_n)$ can be explicitly known (or, in practice, approximated with arbitrary precision) from the data, as the number of possible bootstrap samples is finite (though very large) and all of them are available. Then, it is natural to say that the bootstrap methodology (asymptotically) works when

$$D(\mathcal{L}(R_n;P),\mathcal{L}(R_n;\mathbb{P}_n))\longrightarrow 0\quad\text{in probability, or almost surely,}\tag{26}$$

where D denotes any distance between probability measures metrizing the weak convergence (for example, the Prohorov metric or the Bounded Lipschitz metric).

The first asymptotic validity results of type (26) for the bootstrap method in the standard case where the X_i and R_n are real-valued, date back to the early eighties: Bickel and Freedman (1981), Singh (1981), and Parr (1985).

In the FDA scenario the bootstrap poses a twofold challenge: first, to extend such asymptotic validity results to a range of functional situations as wide as possible, including those with dependent data. Second, to assess the practical performance of the bootstrap approximations via extensive simulations and real data applications. We next briefly review both aspects. As we will see, the progress has been much more remarkable in the first aspect than in the second one. Some technical

details (which can be consulted in the references) are omitted in the following outline, just in order to convey the whole picture.

Let us first consider, as a relevant example, the simplest case of the (centered) sample mean,

$$R(X_1,\ldots,X_n;P)=\sqrt{n}(\overline{X}-m),$$

where we use here the notation of the CLT (5). In this case, the "bootstrapped version" of R_n would be

$$R(X_1^*,\ldots,X_n^*;\mathbb{P}_n)=\sqrt{n}(\overline{X}^*-\overline{X}),$$

$\overline{X}^*$ denotes the sample mean of the X_i^* (and, in general, the asterisk refers to calculations in the "bootstrap world" of the artificial observations drawn from $\mathbb{P}_n$). Now, under the standard assumptions of the functional Central Limit Theorem (5), a validity result of type (26) is equivalent to

$$\sqrt{n}(\overline{X}^*-\overline{X})\xrightarrow{w}\mathcal{G}(0,\Gamma_{X_1}) \quad \text{in probability, or almost surely} \tag{27}$$

As mentioned above, validity results of this type, and some others for empirical and quantile processes, L-statistics, etc., were first obtained in the real case by Bickel and Freedman (1981) and Singh (1981). An extension to the case of separable Banach spaces (which in fact arises as a corollary of the corresponding extension of Donsker Theorem) is due to Giné and Zinn (1990). Further results in this line can be found in Sheehy and Wellner (1992). Again, a conclusion of type (27) for random elements taking values in a Hilbert space is obtained in Politis and Romano (1994). This result holds also under some dependence conditions for the X_i.

Apart from their intrinsic interest, the results of type (27) concerning bootstrap validity for the sample mean are also an important intermediate step to get much more general results. The point is that some statistics of interest have an expression of type $\sqrt{n}(T(\mathbb{P}_n)-T(P))$, where T is a functional (defined in some subspace of probability measures which includes the empirical distributions) fulfilling a suitable differentiability assumption. Essentially, the point is that differentiability entails the possibility of performing local linear approximations, so that one can get

$$\sqrt{n}(T(\mathbb{P}_n^*)-T(\mathbb{P}_n))=\sqrt{n}(\overline{Y}^*-\overline{Y})+\sqrt{n}\,\mathrm{Rem}_n, \tag{28}$$

where the Y_i are new random elements constructed from the X_i and T and Rem_n denotes the remainder term in the first-order Taylor expansion. Then, since the first term in the right-hand side of (28) goes to the corresponding Gaussian limit [from the benchmark result (27)], the bootstrap validity will follow if we are able to show the stochastic convergence to zero of $\sqrt{n}\,\mathrm{Rem}_n$. In the standard, finite-dimensional, setup these ideas are nicely illustrated by Parr (1985). For more general results of this type, applicable to FDA problems, see, e.g., Dudley (1990) and van der Vaart and Wellner (1996), Section 3.9.3. Ferraty et al. (2012) discuss the application of bootstrap methods in a problem of functional regression with functional response.

The paper by McMurry and Politis (2011) provides a survey on theoretical and practical aspects of resampling methodologies (including bootstrap) in nonparametric functional estimation and FDA.

Let us finally note that the papers by Cuevas et al. (2006) and González-Manteiga and Martínez-Calvo (2011a) include some practical insights (based on simulations and real-data examples) on the use of the bootstrap for problems of functional estimation and functional regression, respectively.

9. Software for FDA: the R package fda.usc

The practical use of FDA methodologies, with almost no exception, relies heavily on the availability of a friendly, reasonably comprehensive, software. The R package `fda.usc`, prepared by Manuel Febrero and Manuel Oviedo (University of Santiago de Compostela, Spain) is a recent valuable contribution in this regard. This software incorporates and extends the previous R-package `fda` (see Ramsay et al., 2009) and the R-functions provided by Ferraty and Vieu (2006) as a complement for their book. The package `fda.usc` can be freely downloaded from the R official web site (www.r-project.org. It provides R-functions for most FDA topics covered in this paper. The paper by Febrero-Bande and Oviedo de la Fuente (2012) gives a detailed tutorial, illustrated with many examples, for its practical use.

The Matlab users can find Matlab code for FDA "embedded" within the R-software `fda`; see Ramsay et al. (2009).

Acknowledgments

This work was partially supported by Spanish Grant MTM2010-17366.

The corrections, insights and additional references provided by an anonymous referee led to a much improved manuscript.

I am deeply indebted to my co-workers in FDA subjects: A. Baíllo, J.R. Berrendero, J. Cuesta-Albertos, M. Febrero-Bande and R. Fraiman. This paper is dedicated to them.

References

Abraham, C., Biau, G., Cadre, B., 2006. On the kernel rule for function classification. Annals of the Institute of Statistical Mathematics 58, 619–633.

Abramovich, F., Antoniadis, A., Sapatinas, T., Vidakovic, B., 2004. Optimal testing in a fixed-effects functional analysis of variance model. International Journal of Wavelets, Multiresolution and Information Processing 2, 323–349.

Abramovich, F., Angelini, C., 2005. Testing in mixed-effects FANOVA Models. Journal of Statistical Planning and Inference 136, 4326–4348.

Aguilera, A.M., Escabias, M., Valderrama, M.J., 2006. Using principal components for estimating logistic regression with high dimensional multicollinear data. Computational Statistics & Data Analysis 50, 1905–1924.

Antoniadis, A., Sapatinas, T., 2007. Estimation and inference in functional mixed-effects models. Computational Statistics & Data Analysis 51, 4793–4813.

Ash, R.B., Gardner, M.F., 1975. Topics in Stochastic Processes. Academic Press, New York.

Baíllo, A., Cuesta-Albertos, J.A., Cuevas, A., 2011a. Supervised classification for a family of Gaussian functional models. Scandinavian Journal of Statistics 38, 480–498.

Baíllo, A., Cuevas, A., Fraiman, R., 2011b. Classification methods for functional data. In: Ferraty, F., Romain, Y. (Eds.), The Oxford Handbook of Functional Data Analysis. Oxford University Press, Oxford, pp. 259–297.

Barker, M., Rayens, W., 2003. Partial least squares for discrimination. Journal of Chemometrics 17, 166–173.

Berkes, I., Gabrys, R., Horváth, L., Kokoszka, P., 2009. Detecting changes in the mean of functional observations. Journal of the Royal Statistical Society B 71, 927–946.

Biau, G., Cérou, F., Guyader, A., 2010. Rates of convergence of the functional k-nearest neighbor estimate. IEEE Transactions on Information Theory 56, 2034–2040.

Biau, G., Devroye, L., Lugosi, G., 2008. On the performance of clustering in Hilbert spaces. IEEE Transactions on Information Theory 54, 781–790.

Bickel, P.J., Freedman, D.A., 1981. Some asymptotic theory for the bootstrap. Annals of Statistics 9, 1196–1217.

Bickel, P.J., Ritov, Y., Tsybakov, A.B., 2009. Simultaneous analysis of Lasso and Dantzig selector. Annals of Statistics 37, 1705–1732.

Boente, G., Fraiman, R., 2000. Kernel-based functional principal components. Statistics & Probability Letters 48, 335–345.

Bosq, D., 1991. Modelization, nonparametric estimation and prediction for continuous time processes. In: Roussas, G. (Ed.), Nonparametric Functional Estimation and Related Topics. NATO ASI Series. Kluwer.

Bosq, D., 2000. Linear Processes in Function Spaces. Theory and Applications. Lecture Notes in Statistics, vol. 149. Springer, Berlin.

Bosq, D., Blanke, D., 2007. Inference and Prediction in Large Dimensions. Wiley, Chichester.

Boucheron, S., Bousquet, O., Lugosi, G., 2005. Theory of classification: a survey of some recent advances. ESAIM: Probability and Statistics 9, 323–375.

Bunea, F., Ivanescu, A.E., Wegkamp, M.H., 2011. Adaptive inference for the mean of a Gaussian process in functional data. Journal of the Royal Statistical Society B 73, 531–558.

Cadre, B., 2001. Convergent estimators for the L1 median of a Banach valued random variable. Statistics 35, 509–521.

Cai, T., Hall, P., 2006. Prediction in functional linear regression. Annals of Statistics 34, 2159–2179.

Cai, T., Yuan, M., 2011. Optimal estimation of the mean function based on discretely sampled functional data: phase transition. Annals of Statistics 39, 2330–2355.

Candes, E., Tao, T., 2007. The Dantzig selector: statistical estimation when p is much larger than n (with discussion). Annals of Statistics 35, 2313–2404.

Cardot, H., Ferraty, F., Mas, A., Sarda, P., 2003. Testing hypotheses in the functional linear model. Scandinavian Journal of Statistics 30, 241–255.

Cardot, H., Sarda, P., 2011. Functional linear regression. In: Ferraty, F., Romain, Y. (Eds.), The Oxford Handbook of Functional Data Analysis. Oxford University Press, Oxford, pp. 21–46.

Cérou, F., Guyader, A., 2006. Nearest neighbor classification in infinite dimension. ESAIM: Probability and Statistics 10, 340–355.

Chaudhuri, P., 1996. On a geometric notion of quantiles for multivariate data. Journal of the American Statistical Association 91, 862–872.

Chen, D., Müller, H.G., 2012. Nonlinear manifold representations for functional data. Annals of Statistics 40, 1–29.

Chiou, J., Müller, H.G., Wang, J.L., 2004. Functional response models. Statistica Sinica 14, 675–693.

Crambes, C., Kneip, A., Sarda, P., 2009. Smoothing splines estimators for functional linear regression. Annals of Statistics 37, 35–72.

Cuesta-Albertos, J.A., Del Barrio, E., Fraiman, R., Matrán, C., 2007. The random projection method in goodness of fit for functional data. Computational Statistics & Data Analysis 51, 4814–4831.

Cuesta-Albertos, J.A., Febrero-Bande, M., 2010. Multiway ANOVA for functional data. Test 19, 537–557.

Cuesta-Albertos, J.A., Fraiman, R., Ransford, T., 2007. A sharp form of the Cramér–Wold theorem. Journal of Theoretical Probability 20, 201–209.

Cuesta-Albertos, J.A., Fraiman, R., 2006. Impartial trimmed means for functional data. In: Liu, R., Serfling, R., Souvaine, D. (Eds.), Data Depth: Robust Multivariate Statistical Analysis, Computational Geometry and Applications. DIMACS Series, vol. 72. American Mathematical Society, pp. 121–146.

Cuesta-Albertos, J.A., Fraiman, R., 2007. Impartial trimmed k-means for functional data. Computational Statistics & Data Analysis 51, 4864–4877.

Cuesta-Albertos, J.A., Matrán, C., 1988. The strong law of large numbers for k-means and best possible nets of Banach valued random variables. Probability Theory and Related Fields 78, 523–534.

Cuesta-Albertos, J.A., Matrán, C., 1989. Uniform consistency of r-means. Statistics & Probability Letters 6, 65–71.

Cuesta-Albertos, J.A., Nieto-Reyes, A., 2008. The random Tukey depth. Computational Statistics & Data Analysis 52, 4979–4988.

Cuevas, A., Febrero, F., Fraiman, R., 2002. Linear functional regression: the case of fixed design and functional response. Canadian Journal of Statistics 30, 285–300.

Cuevas, A., Febrero, F., Fraiman, R., 2004. An ANOVA test for functional data. Computational Statistics & Data Analysis 47, 111–122.

Cuevas, A., Febrero, M., Fraiman, R., 2006. On the use of the bootstrap for estimating functions with functional data. Computational Statistics & Data Analysis 51, 1063–1074.

Cuevas, A., Febrero, M., Fraiman, R., 2007. Robust estimation and classification for functional data via projection-based depth notions. Computational Statistics 22, 481–496.

Cuevas, A., Fraiman, R., 2009. On depth measures and dual statistics. A methodology for dealing with general data. Journal of Multivariate Analysis 100, 753–766.

Dabo-Niang, S., Ferraty, F., Vieu, P., 2007. On the using of modal curves for radar waveforms classification. Computational Statistics & Data Analysis 51, 4878–4890.

Dabo-Niang, S., Yao, A.F., Pischedda, L., Cuny, P., Gilbert, F., 2010. Spatial mode estimation for functional random fields with application to bioturbation problem. Stochastic Environmental Research and Risk Assessment 24, 487–497.

DasGupta, A., 2008. Asymptotic Theory of Statistics and Probability. Springer, New York.

Dauxois, J., Pousse, A., Romain, Y., 1982. Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. Journal of Multivariate Analysis 12, 136–154.

Delaigle, A., Hall, P., 2010. Defining probability density for a distribution of random functions. Annals of Statistics 38, 1171–1193.

Delaigle, A., Hall, P., 2012a. Achieving near perfect classification for functional data. Journal of the Royal Statistical Society B 74, 267–286.

Delaigle, A., Hall, P., 2012b. Methodology and theory for partial least squares applied to functional data. Annals of Statistics 40, 322–352.

Delsol, L., Ferraty, F., Martínez-Calvo, A., 2011. Functional data analysis: an interdisciplinary statistical topic. In: Gettler-Summa, M., Bottou, L., Goldfarb, B., Murtagh, F., Pardoux, C., Touati, M. (Eds.), Statistical Learning and Data Science. Chapman and Hall, CRC, Boca Raton, pp. 189–195.

Devroye, L., Györfi, L., Lugosi, G., 1996. A Probabilistic Theory of Pattern Recognition. Springer-Verlag, New York.

Dudley, R.M., 1990. Nonlinear functionals of empirical measures and the bootstrap. In: Eberlein, E., Kuelbs, J., Marcus, M.B. (Eds.), Probability in Banach Spaces, vol. 7 (Oberwolfach, 1988). Birkhäuser, Boston, pp. 63–82.

Escabias, M., Aguilera, A.M., Valderrama, M.J., 2007. Functional PLS logit regression model. Computational Statistics & Data Analysis 51, 4891–4902.

Fan, J., Lv, J., 2010. A selective overview of variable selection in high dimensional feature space. Statistica Sinica 20, 101–148.
Febrero-Bande, M., Galeano, P., González-Manteiga, W., 2010. Measures of influence for the functional linear model with scalar response. Journal of Multivariate Analysis 101, 327–339.
Febrero-Bande, M., Oviedo de la Fuente, M., 2012. Statistical computing in functional data analysis: the R package fda.usc. Journal of Statistical Software 51, 1–28.
Ferraty, F., Laksaci, A., Tadj, A., Vieu, P., 2011. Kernel regression with functional response. Electronic Journal of Statistics 5, 159–171.
Ferraty, F., Romain, Y. (Eds.), 2011. The Oxford Handbook of Functional Data Analysis. Oxford University Press, Oxford.
Ferraty, F., van Keilegom, I., Vieu, P., 2012. Regression when both response and predictor are functions. Journal of Multivariate Analysis 109, 10–28.
Ferraty, F., Vieu, P., 2006. Nonparametric Functional Data Analysis: Theory and Practice. Springer.
Ferraty, F., Vieu, P., 2011. Kernel regression estimation for functional data. In: Ferraty, F., Romain, Y. (Eds.), The Oxford Handbook of Functional Data Analysis. Oxford University Press, Oxford, pp. 72–129.
Fraiman, R., Pateiro-López, B., 2012. Quantiles for finite and infinite dimensional data. Journal of Multivariate Analysis 108, 1–14.
Frank, I.E., Friedman, J.H., 1993. A statistical view of some chemometrics regression tools (with discussion). Technometrics 35, 109–148.
Forzani, L., Fraiman, R., Llop, P., 2012. Consistent nonparametric regression for functional data under the Stone–Besicovich conditions. IEEE Transactions on Information Theory 58, 6697–6708.
Gasser, T., Hall, P., Presnell, B., 1998. Nonparametric estimation of the mode of a distribution of random curves. Journal of the Royal Statistical Society B 60, 681–691.
Gervini, D., 2008. Robust functional estimation using the spatial median and spherical principal components. Biometrika 95, 587–600.
Giné, E., Zinn, J., 1990. Bootstrapping general empirical measures. Annals of Probability 18, 851–869.
González-Manteiga, W., Martínez-Calvo, A., 2011a. Bootstrap in functional linear regression. Journal of Statistical Planning and Inference 141, 453–461.
González-Manteiga, W., Vieu, P., 2011. Methodological richness of functional data analysis. In: Gettler-Summa, M., Bottou, L., Goldfarb, B., Murtagh, F., Pardoux, C., Touati, M. (Eds.), Statistical Learning and Data Science. Chapman and Hall, CRC, pp. 197–203.
Gordaliza, A., 1991. Best approximations to random variables based on trimming procedures. Journal of Approximation Theory 64, 162–180.
Grenander, U., 1981. Abstract Inference. Wiley, New York.
Györfi, L., Kohler, M., Krzyizak, A., Walk, H., 2002. A Distribution-Free Theory of Nonparametric Regression. Springer-Verlag, New York.
Hall, P., Müller, H.G., Wang, J.L., 2006. Properties of principal component methods for functional and longitudinal data analysis. Annals of Statistics 34, 1493–1517.
Hall, P., Heckman, N.E., 2002. Estimating and depicting the structure of a distribution of random functions. Biometrika 89, 145–158.
Hand, D., 2006. Classifier technology and the illusion of progress. Statistical Science 21, 1–34.
Hartigan, J.A., 1975. Clustering Algorithms. Wiley, New York.
Hastie, T., Buja, A., Tibshirani, R., 1995. Penalized discriminant analysis. Annals of Statistics 23, 73–102.
Heckman, N.E., Zamar, R.H., 2000. Comparing the shapes of regression functions. Biometrika 87, 135–144.
Hoffmann-Jorgensen, J., Pisier, G., 1976. The law of large numbers and the central limit theorem in Banach spaces. Annals of Probability 4, 587–599.
Horváth, L., Husková, M., Kokoszka, P., 2010. Testing the stability of the functional autoregressive process. Journal of Multivariate Analysis 101, 352–367.
Horváth, L., Kokoszka, P., 2012. Inference for Functional Data with Applications. Springer, New York.
Horváth, L., Reeder, R., 2012. Detecting changes in functional linear models. Journal of Multivariate Analysis 111, 310–334.
Inácio, V., González-Manteiga, W., Febrero-Bande, M., Gude, F., Alonzo, T.A., Cadarso-Suárez, C., 2012. Extending induced ROC methodology to the functional context. Biostatistics 13, 594–608.
James, G.M., Hastie, T.J., 2001. Functional linear discriminant analysis for irregularly sampled curves. Journal of the Royal Statistical Society B 63, 533–550.
James, G.M., Hastie, T.J., Sugar, C.A., 2000. Principal component models for sparse functional data. Biometrika 87, 587–602.
James, G.M., Sugar, C.A., 2003. Clustering for sparsely sampled functional data. Journal of the American Statistical Association 98, 397–409.
Jørsboe, O.G., 1968. Equivalence or Singularity of Gaussian Measures on Function Spaces. Various Publications Series, No. 4, Matematisk Institut, Aarhus Universitet, Aarhus.
Kemperman, J.H.B., 1987. The median of a finite measure on a Banach space. In: Dodge, Y. (Ed.), Statistical Analysis Based on the L1-norm and Related Methods. North Holland, Amsterdam, pp. 217–230.
Kneip, A., Sarda, P., 2011. Factor models and variable selection in high-dimensional regression analysis. Annals of Statistics 39, 2410–2447.
Kokoszka, P., Maslova, I., Sojka, J., Zhu, L., 2008. Testing for lack of dependence in the functional linear model. Canadian Journal of Statistics 36, 1–16.
Laha, R.G., Rohatgi, V.K., 1979. Probability Theory. Wiley, New York.
Ledoux, M., Talagrand, M., 2011. Probability in Banach Spaces, second ed. Springer, Berlin.
Li, J., Cuesta-Albertos, J.A., Liu, R., 2012. DD-classifier: nonparametric classification procedure based on DD-plot. Journal of the American Statistical Association 107, 737–753.
Li, W.V., Shao, Q.M., 2001. Gaussian processes: inequalities, small ball probabilities and applications. In: Rao, C.R., Shanbhag, D. (Eds.), Stochastic Processes: Theory and Methods, Handbook of Statistics, vol. 19. Elsevier, New York, pp. 533–598.
Li, Y., Hsing, T., 2010. Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. Annals of Statistics 38, 3321–3351.
Lindquist, M.A., McKeague, I.W., 2009. Logistic regression with Brownian-like predictors. Journal of the American Statistical Association 104, 1575–1585.
Locantore, N., Marron, J.S., Simpson, D.G., Tripoli, N., Zhang, J.T., Cohen, K.L., 1999. Robust principal components for functional data. Test 8, 1–73.
López-Pintado, S., Romo, J., 2006. Depth based classification for functional data (2006). In: DIMACS Series in Discrete Mathematics, vol. 72, pp. 103–120.
López-Pintado, S., Romo, J., 2009. On the concept of depth for functional data. Journal of the American Statistical Association 104, 486–503.
Mandelbaum, A., 1984. Linear estimators and measurable linear transformations on a Hilbert space. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 65, 385–397.
McDiarmid, C., 1989. On the method of bounded differences. In: Siemons, J. (Ed.), Surveys in Combinatorics. London Mathematical Society Lecture Note Series, vol. 141. Cambridge University Press, pp. 148–188.
McMurry, T., Politis, D., 2011. Resampling methods for functional data. In: Ferraty, F., Romain, Y. (Eds.), The Oxford Handbook of Functional Data Analysis. Oxford University Press, Oxford, pp. 189–209.
Mörters, P., Peres, Y., 2010. Brownian Motion. Cambridge University Press, Cambridge.
Mourier, E., 1953. Elements aléatoires dans un espace de Banach. Annales de l'Institut Henri Poincaré 13, 161–244.
Müller, H.-G., 2005. Functional modelling and classification of longitudinal data. Scandinavian Journal of Statistics 32, 223–240.
Parr, W.C., 1985. The bootstrap: some large sample theory and connections with robustness. Statistics & Probability Letters 3, 97–100.
Politis, D.N., Romano, J.P., 1994. Limit theorems for weakly dependent Hilbert space valued random variables with application to the stationary bootstrap. Statistica Sinica 4, 461–476.
Preda, C., Saporta, G., 2005. PLS regression on a stochastic process. Computational Statistics & Data Analysis 48, 149–158.
Preda, C., Saporta, G., Lévéder, C., 2007. PLS classification of functional data. Computational Statistics 22, 223–235.
Ramsay, J.O., Silverman, B.W., 2002. Applied Functional Data Analysis. Methods and Case Studies. Springer, New York.
Ramsay, J.O., Hooker, G., Graves, S., 2009. Functional Data Analysis with R and MATLAB. Springer, New York.
Ramsay, J.O., Silverman, B.W., 2005. Functional Data Analysis, second ed. Springer, New York.
Reiss, P.T., Ogden, R.T., 2007. Functional principal component regression and functional partial least squares. Journal of the American Statistical Association 102, 984–996.

Rice, J., 2004. Functional and longitudinal data analysis: perspectives on smoothing. Statistica Sinica 14, 631–647.
Rice, J.A., Silverman, B.W., 1991. Estimating the mean and covariance structure nonparametrically when the data are curves. Journal of the Royal Statistical Society B 53, 233–243.
Ruiz-Medina, M.D., 2012. Spatial functional prediction from spatial autoregressive Hilbertian processes. Environmetrics 23, 119–128.
Schott, J.R., 2007. Some high-dimensional tests for a one-way ANOVA. Journal of Multivariate Analysis 98, 1825–1839.
Serban, N., Wasserman, L., 2005. CATS: clustering after transformation and smoothing. Journal of the American Statistical Association 100, 990–999.
Sguera, C., Galeano, P., Lillo, R., 2012. Spatial depth-based classification for functional data. Working Paper 12-09. Univ. Carlos III, Madrid.
Sheehy, A., Wellner, J.A., 1992. Uniform Donsker classes of functions. Annals of Probability 20, 1983–2030.
Shin, J., 2008. An extension of Fisher's discriminant analysis for stochastic processes. Journal of Multivariate Analysis 99, 1191–1216.
Silverman, B.W., 1996. Smoothed functional principal components by choice of norm. Annals of Statistics 24, 1–24.
Singh, K., 1981. On the asymptotic accuracy of Efron's bootstrap. Annals of Statistics 9, 1187–1195.
Spitzner, D.J., Marron, J.S., Essick, G.K., 2003. Mixed-model functional ANOVA for studying human tactile perception. Journal of the American Statistical Association 98, 263–272.
Stone, C., 1977. Consistent nonparametric regression. Annals of Statistics 8, 1348–1360.
Tarpey, T., 2007. Linear transformations and the k-means clustering algorithm: applications to clustering curves. American Statistician 61, 34–40.
Tibshirani, R., 1996. Regression shrinkage and selection via the LASSO. Journal of the Royal Statistical Society B 58, 267–288.
van der Vaart, A., Wellner, J., 1996. Weak Convergence and Empirical Processes. Springer, New York.
Varadhan, S.R.S., 1962. Limit theorems for sums of independent random variables with values in a Hilbert space. Sankhya A 24, 213–238.
Varberg, D.E., 1961. On equivalence of Gaussian measures. Pacific Journal of Mathematics 11, 751–762.
Vardi, Y., Zhang, C.H., 2000. The multivariate L1-median and associated data depth. Proceedings of the National Academy of Sciences 97, 1423–1426.
Wold, H., 1975. Soft modelling by latent variables: the non-linear iterative partial least squares (NIPALS) approach. In: Gani, J. (Ed.), Perspectives in Probability and Statistics, Papers in Honour of M. S. Bartlett. Academic Press, London.
Yurinskii, V.V., 1976. Exponential inequalities for sums of random vectors. Journal of Multivariate Analysis 6, 473–499.
Zhou, L., Huang, J.Z., Carroll, R.J., 2008. Joint modelling of paired sparse functional data using principal components. Biometrika 95, 601–619.
Zuo, Y., Serfling, R., 2000. General notions of statistical depth function. Annals of Statistics 28, 461–482.