1

Introduction to Functional Nonparametric Statistics

The main goal of this chapter is to familiarize the reader with both functional and nonparametric statistical notions. First and because of the novelty of this field of research, we propose some basic definitions in order to clarify the vocabulary on both functional data/variables and nonparametric modelling. Second, we fix some notations to unify the remaining of the book.

1.1 What is a Functional Variable?

There is actually an increasing number of situations coming from different fields of applied sciences (environmetrics, chemometrics, biometrics, medicine, econometrics, . . .) in which the collected data are curves. Indeed, the progress of the computing tools, both in terms of memory and computational capacities, allows us to deal with large sets of data. In particular, for a single phenomenon, we can observe a very large set of variables. For instance, look at the following usual situation where some random variable can be observed at several different times in the range (t_{min}, t_{max}). An observation can be expressed by the random family $\{X(t_j)\}_{j=1,\ldots,J}$. In modern statistics, the grid becomes finer and finer meaning that consecutive instants are closer and closer. One way to take this into account is to consider the data as an observation of the continuous family $\mathcal{X} = \{X(t);\ t \in (t_{min}, t_{max})\}$. This is exactly the case of the speech recognition dataset that we will treat deeply later in this monograph (see Section 2.2). Of course, other situations can be viewed similarly such as for instance the spectrometric curves presented in Section 2.1, for which the measurements concern different wavelengths instead of time points. Moreover, a new literature is emerging which deals with sparse functional data. In this situation, the number of measurements is small but the data are clearly of functional nature (see for instance the electrical consumption curves described in Section 2.3). To fix the ideas, we give the following general definition of a functional variable/data.

Definition 1.1. *A random variable $\mathcal{X}$ is called functional variable (f.v.) if it takes values in an infinite dimensional space (or functional space). An observation χ of $\mathcal{X}$ is called a functional data.*

Note that, when $\mathcal{X}$ (*resp.* χ) denotes a random curve (*resp.* its observation), we implicitly make the following identification $\mathcal{X} = \{\mathcal{X}(t);\ t \in T\}$ (*resp.* $\chi = \{\chi(t);\ t \in T\}$). In this situation, the functional feature comes directly from the observations. The situation when the variable is a curve is associated with an unidimensional set $T \subset \mathbb{R}$. Here, it is important to remark that the notion of functional variable covers a larger area than curves analysis. In particular, a functional variable can be a random surface, like for instance the grey levels of an image or a vector of curves (and in these cases T is a bidimensional set $T \subset \mathbb{R}^2$), or any other more complicated infinite dimensional mathematical object. Even if the real data used as supports throughout this book are all curves datasets (i.e., a set of curves data), all the methodology and theoretical advances to be presented later are potentially applicable to any other kind of functional data.

1.2 What are Functional Datasets?

Since the middle of the nineties, the increasing number of situations when functional variables can be observed has motivated different statistical developments, that we could quickly name as *Statistics for Functional Variables* (or *Data*). We are determinedly part of this statistical area since we will propose several methods involving statistical functional sample $\mathcal{X}_1, \ldots, \mathcal{X}_n$. Let us start with a precise definition of a functional dataset.

Definition 1.2. *A functional dataset $\chi_1, \ldots, \chi_n$ is the observation of n functional variables $\mathcal{X}_1, \ldots, \mathcal{X}_n$ identically distributed as $\mathcal{X}$.*

This definition covers many situations, the most popular being curves datasets. We will not investigate the question of how these functional data have been collected, which is linked with the discretization problems. According to the kind of the data, a preliminary stage consists in presenting them in a way which is well adapted to functional processing. As we will see, if the grid of the measurements is fine enough, this first important stage involves usual numerical approximation techniques (see for instance the case of spectrometric data presented in Chapter 3). In other standard cases, classical smoothing methods can be invoked (see for instance the phonemes data and the electrical consumption curves discussed in Chapter 3). There exist some other situations which need more sophisticated smoothing techniques, for instance when the repeated measures per subjects are very few (sparse data) and/or with irregular grid. This is obviously a parallel and complementary

field of research but this is far from our main purpose which is nonparametric statistical treatments of functional data. From now on, we will assume that we have at hand a sample of functional data.

1.3 What are Nonparametric Statistics for Functional Data?

Traditional statistical methods fail as soon as we deal with functional data. Indeed, if for instance we consider a sample of finely discretized curves, two crucial statistical problems appear. The first comes from the ratio between the size of the sample and the number of variables (each real variable corresponding to one discretized point). The second, is due to the existence of strong correlations between the variables and becomes an ill-conditioned problem in the context of multivariate linear model. So, there is a real necessity to develop statistical methods/models in order to take into account the functional structure of this kind of data. Most of existing statistical methods dealing with functional data use linear modelling for the object to be estimated. Key references on methodological aspects are those by [RS97] and [RS05], while applied issues are discussed by [RS02] and implementations are provided by [CFGR05]. Note also that, for some more specific problem, some theoretical support can be found in [B00].

On the other hand, nonparametric statistics have been developped intensively. Indeed, since the beginning of the sixties, a lot of attention has been paid to free-modelling (both in a free-distribution and in a free-parameter meaning) statistical models and/or methods. The functional feature of these methods comes from the nature of the object to be estimated (such as for instance a density function, a regression function, ...) which is not assumed to be parametrizable by a finite number of real quantities. In this setting, one is usually speaking of **Nonparametric Statistics** for which there is an abundant literature. For instance, the reader will find in [H90] a previous monograph for applied nonparametric regression, while [S00] and [AP03] present the state of the art in these fields. It appears clearly that these techniques concern only classical framework, namely real or multidimensional data.

However, recent advances are mixing nonparametric free-modelling ideas with functional data throughout a double infinite dimensional framework (see [FV03b] for bibliography). The main aim of this book is to describe both theoretical and practical issues of these recent methods through various statistical problems involving prediction, time series and classification. Before to go on, and in order to clarify the sense of our purpose, it is necessary to state precisely the meanings of the expressions parametric and nonparametric models.

There are many (different) ways for defining what is a nonparametric statistical model in finite dimensional context, and the border between nonparametric and parametric models may sometimes appear to be unclear (see the

introduction in [BL87] for more discussion). Here, we decided to start from the following definition of nonparametric model in finite dimensional context.

Definition 1.3. *Let $\boldsymbol{X}$ be a random vector valued in $\mathbb{R}^p$ and let ϕ be a function defined on $\mathbb{R}^p$ and depending on the distribution of $\boldsymbol{X}$. A model for the estimation of ϕ consists in introducing some constraint of the form*

$$\phi \in \mathcal{C}.$$

The model is called a parametric model for the estimation of ϕ if $\mathcal{C}$ is indexed by a finite number of elements of $\mathbb{R}$. Otherwise, the model is called a nonparametric model.

Our decision for choosing this definition was motivated by the fact that it makes definitively clear the border between parametric and nonparametric models, and also because this definition can be easily extended to the functional framework.

Definition 1.4. *Let $\mathcal{Z}$ be a random variable valued in some infinite dimensional space F and let ϕ be a mapping defined on F and depending on the distribution of $\mathcal{Z}$. A model for the estimation of ϕ consists in introducing some constraint of the form*

$$\phi \in \mathcal{C}.$$

The model is called a functional parametric model for the estimation of ϕ if $\mathcal{C}$ is indexed by a finite number of elements of F. Otherwise, the model is called a functional nonparametric model.

The appellation **Functional Nonparametric Statistics** covers all statistical backgrounds involving a nonparametric functional model. In the terminology **Functional Nonparametric Statistics**, the adjective **nonparametric** refers to the form of the set of constraints whereas the word **functional** is linked with the nature of the data. In other words, nonparametric aspects come from the infinite dimensional feature of the object to be estimated and functional designation is due to the infinite dimensional feature of the data. That is the reason why we may identify this framework to a double infinite dimensional context. Indeed, ϕ can be viewed as a nonlinear operator and one could use the terminology *model for operatorial estimation* by analogy with the multivariate terminology *model for functional estimation*.

To illustrate our purpose concerning these modelling aspects, we focus on the regression models

$$Y = r(X) + error, \tag{1.1}$$

where Y is a real random variable by considering various situations: linear (parametric) or nonparametric regression models with curves (i.e. $X = \mathcal{X} = \{X(t);\ t \in (0,1)\}$) or multivariate (i.e. $X = \boldsymbol{X} = (X^1, \ldots, X^p)$) data:

		MODELS	
		LINEAR	NONPARAMETRIC
DATA	MULTIVARIATE	*Example 1* $X \in \mathbb{R}^p$ $\mathcal{C} = \{r \text{ linear}\}$	*Example 2* $X \in \mathbb{R}^p$ $\mathcal{C} = \{r \text{ continuous}\}$
	FUNCTIONAL	*Example 3* $X \in F = L^2_{(0,1)}$ $\mathcal{C} = \{\chi \mapsto \int_0^1 \rho(t)\chi(t)dt \in \mathbb{R},\ \rho \in F\}$	*Example 4* $X \in F = L^2_{(0,1)}$ $\mathcal{C} = \{r \text{ continuous}\}$

Example 1 corresponds to the so-called multivariate linear regression model

$$Y = a_0 + \sum_{j=1}^{p} a_j\, X^j + error,$$

which is obviously a parametric model (with $p+1$ unknown real parameters $a_0, \ldots, a_p$). Example 2 refers to the classical multivariate nonparametric regression model

$$Y = r(X^1, \ldots, X^p) + error.$$

Now, Example 3 is exactly what [RS97] call *functional linear regression model for scalar response* namely

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

which can be reformulated as (1.1) with r being a continuous linear operator from F to $\mathbb{R}$ (by using the Riesz representation theorem). According to our definition, this is a functional parametric regression model, where $\rho(.)$ is the only one (functional) parameter. The last model (Example 4) can be written as (1.1) where r is just a continuous operator from F to $\mathbb{R}$. Example 4 is a functional nonparametric regression model according to Definition 1.4. This model will be treated with details in Chapter 5.

1.4 Some Notation

In the remaining of this book, χ will denote any non-random element of some infinite dimensional space E and $\boldsymbol{\mathcal{X}}$ a functional random variable valued in E. Similarly, $\{\chi_i\}_{i=1,\ldots,n}$ will denote the n observations of a sample $\{\boldsymbol{\mathcal{X}}_i\}_{i=1,\ldots,n}$ of f.r.v. Even if this monograph is devoted to the study of the nonparametric method for functional data, we will still need to introduce real or multivariate

variables. Instead of χ, $\boldsymbol{\mathcal{X}}$, χ_i and $\boldsymbol{\mathcal{X}}_i$, we will use boldfaced letters for vectors ($\boldsymbol{x}$, $\boldsymbol{X}$, $\boldsymbol{x_i}$ and $\boldsymbol{X_i}$) and standard letters for the real case (x, X, x_i, X_i).

In addition, any random variable considered in this book is defined on the same probability space $(\Omega, \mathcal{A}, P)$. Finally, except in Part IV, any statistical sample is implicitly assumed to be independent.

1.5 Scope of the Book

Once the general framework for nonparametric modelling of functional variable is given (see Part I), the book focuses on various statistical topics: predicting from a functional variable (see Part II), classifying a sample of functional data (see Part III) and statistics for dependent functional variables (see Part IV). All these statistical methods are developed in a free-parameter way (which includes free-distribution modelling). In this sense, this book is both completely different and complementary to the few other books existing on functional data ([RS97], [RS02] and [RS05]). Rather than set application against theory, this book is really an interface of these two features of statistics and each statistical topic is investigated from its deep theoretic foundations up to computational issues. For more details on practical aspects, the reader can refer to the companion website *http://www.lsp.ups-tlse.fr/staph/npfda*). This site contains functional datasets, case studies and routines written in two popular statistical languages: R (see [RDCT]) and S+ (see the comprehensive manual of [BCW88], as well as more recent literature in [C98], [KO05] or [VR00]).