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