3

What is a Well-Adapted Space for Functional Data?

Using functional data asks crucial statistical questions. Indeed, the larger is the space E in which the variable takes its values, the sparser are the data. In the case of functional data, we know (by the nature itself of the data) that E is an infinite dimensional space. So, this chapter focuses on this essential problem of high (i.e., infinite) dimensional data. Obviously, the sparseness notion is strongly linked with the way used to measure closeness between data, and we propose an original way to approach this question by mean of semi-metric considerations.

3.1 Closeness Notions

Proximities measures between mathematical objects play a major role in all statistical methods. In many situations, a classical norm can be used to measure the closeness between two objects. Because in a finite dimensional euclidean space (typically $\mathbb{R}^p$) there is an equivalence between all norms, the choice in the mathematical sense of this kind of measure is not crucial apart from some practical constraints (as, for instance, computational ease). Once a preliminary norm is fixed, it is clear that we can deduce a family of norms and from a statistical point of view, there remains one essential question: namely the choice among these different metrics. For instance, one of the most popular in $\mathbb{R}^p$ is the usual euclidean norm $\|.\|$ which is based on the sum of squares of the components of any vector. More precisely, let $\boldsymbol{x} = {}^t(x_1, \ldots, x_p)$ be a vector of $\mathbb{R}^p$; then, the classical euclidean norm is defined by

$$\|\boldsymbol{x}\|^2 = \sum_{j=1}^{p} (x_j)^2 = {}^t\boldsymbol{x}\,\boldsymbol{x}.$$

Of course, we can deduce a family of norms based on the euclidean norm by using different definite positive matrices $\boldsymbol{M}$, in the following way

$$\|\boldsymbol{x}\|_{\boldsymbol{M}}^2 = {}^t\boldsymbol{x}\,\boldsymbol{M}\,\boldsymbol{x}.$$

The choice of the norm comes to the same as the choice of $\boldsymbol{M}$.

Now, considering an infinite dimensional space, the equivalence between norms fails and the problem has to be attacked in a different way. In other words, in the functional context, the choice of the preliminary norm becomes crucial. Even more, considering normed or metric spaces can become too restrictive. In some situations and this is the case for our datasets, it appears that semi-metric spaces are better adapted than metric spaces. As we will see later, the shape of data and eventually exogene informations or the goal of the statistical study can help to drive the semi-metric selection. The aim of the next section is to show the benefit of considering semi-metrics as a closeness measure. Before going on, let us just recall some basic definitions.

Definition 3.1. $\|.\|$ *is a semi-norm on some space* F *as soon as:*
1) $\forall(\lambda, x) \in \mathbb{R} \times F$, $\|\lambda\, x\| = |\lambda|\, \|x\|$
2) $\forall(x, y) \in F \times F$, $\|x + y\| \leq \|x\| + \|y\|$.

Note that in fact, a semi-norm $\|.\|$ is a norm except that $\|x\| = 0 \not\Rightarrow x = 0$. Similarly, a semi-metric d can be defined to be a metric but such that $d(x, y) = 0 \not\Rightarrow x = y$.

Definition 3.2. d *is a semi-metric on some space* F *as soon as:*
1) $\forall x \in F$, $d(x, x) = 0$,
2) $\forall(x, y, z) \in F \times F \times F$, $d(x, y) \leq d(x, z) + d(z, y)$.

3.2 Semi-Metrics as Explanatory Tool

A large part of explanatory tools consists in displaying data in low-dimensional spaces. It is clear that the shape of such graphics depends strongly on the proximity measure. Look at the chemometric dataset. As was shown in Section 2.1, one can start the study by usual Principal Component Analysis (PCA), the proximity between subjects (spectrometric curves) being computed by means of the classical L_2-metric, which is defined for all observed curves χ_i and $\chi_{i'}$ by

$$\sqrt{\left(\int \left(\chi_i(t) - \chi_{i'}(t)\right)^2 \, dt\right)}.$$

Because the first axis has been interpreted as a factor scale (see Figure 2.3), it is pertinent to see the eigenspace spanned by axes 2 and 3 (see Figure 3.1).