ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Master thesis

# A probabilistic approach to the classification of censored functional data

William Borgeaud dit Avocat

supervised by
Prof. Victor Panaretos

January 14, 2018

# Contents

# Chapter 1

# Introduction to Functional Data Analysis

## 1.1 What are functional data?

We will use the term *functional data* for any data living in some infinite dimensional space. This definition is intentionally vague and encompasses various types of data. By infinite dimensional space, we mean any model of space that does not satisfy some finiteness conditions. For example, infinite dimensional vector spaces or manifolds. Here are some concrete examples of such data and spaces:

1. Functions $f : \mathbb{R} \to \mathbb{R} \ \rightsquigarrow \ \mathbb{R}^{\mathbb{R}}$.

2. Square integrable measurable functions $f : [0,1] \to \mathbb{R} \ \rightsquigarrow \ \text{Ł}^2[0,1]$.

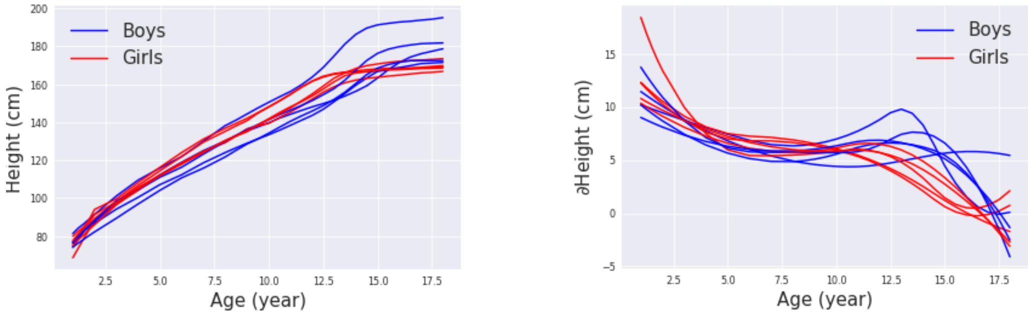
3. Continuous closed curves in the plane $f : [0,1] \to \mathbb{R}^2$, $f(0) = f(1) \ \rightsquigarrow \ \mathcal{C}(\mathbb{S}^1, \mathbb{R}^2)$.

4. Amplitudes of square integrable measurable functions $f : [0,1] \to \mathbb{R} \rightsquigarrow \text{Ł}^2[0,1]/\mathrm{Diffeo}([0,1])$, where $\text{Ł}^2[0,1] \curvearrowleft \mathrm{Diffeo}([0,1])$ by $f \cdot \gamma = (f \circ \gamma)\sqrt{\frac{d\gamma}{dt}}$.

These four examples show different type of spaces one can study. The first one is a vector space of uncountable dimension. With its standard topology it is not metrizable. The second is a separable Hilbert space with inner product $\langle f, g \rangle = \int_0^1 f(t)g(t)dt$. The third is a separable Banach space with norm $\|f\| = \max_{z \in \mathbb{S}^1} \|f(z)\|_{\mathbb{R}^2}$. Finally, the last is a Hilbert manifold, i.e., a separable topological space locally homeomorphic to a Hilbert space.

Evidently, some spaces are easier to work with than others. For instance, it is preferable to have a metric to measure the discrepancies between data points. Banach spaces and Hilbert manifold are metrizable, but functions from and to these spaces can be hard to analyze. The extensive theory of operator on Hilbert spaces make them good candidates for functional data analysis, and they are indeed the spaces of choice in most of the literature.

In this project, we focus on the separable Hilbert spaces. It is well-known that all such spaces are isomorphic to the space $\ell^2(\mathbb{R})$ of square summable sequences. However, we will mainly focus on the space $\text{Ł}^2[0,1]$ as the most common type of functional data are functions depending on time. A famous example of such data is the growth data set from the Berkeley Growth Study, see [Tuddenham and Snyder, 1954]. It contains the heights of girls and boys from birth to their 18th year. Graphs of these heigths and their derivative for 5 girls and boys are shown in Figure 1.1.



(a) Height against age

(b) Derivative of height against age

Figure 1.1

Questions one may ask on these data are: What are the mean and variance of these curves, and what do these notions mean? Can we classify a given curve as coming from a boy or a girl? How do we deal with the fact that individuals don't have the same "biological clock"?
Functional data analysis deals with this kind of questions. In the next section, we will expose the theory needed to answer the first one.

## 1.2 Stochastic processes on $[0, 1]$

We want to study random functions in $\text{Ł}^2[0, 1]$. There are two possible definitions:

- A random function in $L^2[0, 1]$ is a random element in $(L^2[0, 1], \mathcal{B}(\text{Ł}^2[0, 1]))$, i.e., a measurable map
$$X : (\Omega, \mathcal{F}, \mathbb{P}) \to (L^2[0, 1], \mathcal{B}(\text{Ł}^2[0, 1]))$$
for some probability space $(\Omega, \mathcal{F}, \mathbb{P})$.
- A random function in $L^2[0, 1]$ is a stochastic process on $[0, 1]$, i.e., a collection $\{X(t) \mid t \in [0, 1]\}$ of random variables defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$, such that
$$\int_0^1 X^2(t, \omega)dt < \infty \text{ for almost all } \omega.$$

The first definition is more abstract and somewhat more elegant. However, it is not so practical since, for example, the evaluation at a point is not well-defined in $L^2[0, 1]$. On the other hand, it can readily be generalized to other separable Hilbert spaces, or other function spaces. We will use this definition in Chapter 3 to develop the theory of infinite dimensional Gaussian random variables.
The second one is closer to what is observed in practice. Moreover, if $X(t, \omega)$ is measurable on the product $\sigma-$algebra $\mathcal{B}([0, 1]) \otimes \mathcal{F}$, one can show (see [Hsing and Eubank, 2015, Theorem 7.4.1]) that
$$X : \Omega \to L^2[0, 1] : \omega \mapsto X(\cdot, \omega)$$
is measurable, and thus is a random function with respect to the first definition.
We will use the second definition in our exposition, assuming when needed that $X$ is also a random element of $\mathbb{L}^2[0, 1]$ to avoid unnecessary technicalities. Most of the result given here can be found in [Hsing and Eubank, 2015].
One can readily define the mean of a process $X$ as the pointwise mean:
$$\mu(t) = \mathbb{E}[X(t)], \;\; t \in [0, 1].$$

The information on the (co)variance of $X$ is stored in the covariance kernel, a function $K : [0, 1]^2 \to \mathbb{R}$ defined by
$$K(s, t) = \text{Cov}[X(s), X(t)].$$

When $\mu$ and $K$ are well-defined, we call $X$ a second-order process. In multivariate analysis, the covariance matrix of a random vector is always positive semidefinite. For second-order processes, we have the following analogous result:

**Proposition 1.** The covariance kernel $K$ of a second-order process is symmetric and positive semidefinite, that is

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K(t_i, t_j) \geq 0$$

holds for any $n \in \mathbb{N}, t \in [0,1]^n, c \in \mathbb{R}^n$.

*Proof.* The symmetry of $K$ is a direct consequence of the symmetry of the covariance. For the positive semidefiniteness, we have

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j K(t_i, t_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j \mathrm{Cov}[X(t_i), X(t_j)] = \mathrm{Var}\left[\sum_{i=1}^{n} c_i X(t_i)\right] \geq 0.$$

□

We now assume that $X$ is mean-squared continuous, that is

$$\lim_{s \to t} \mathbb{E}[X(s) - X(t)]^2 = 0$$

for all $s, t \in [0,1]$. Intuitively, this is equivalent to saying that the mean of $X$ and its variation vary continuously. This is made formal by the following theorem, proved in [Hsing and Eubank, 2015, Theorem 7.2.4].

**Theorem 1.** The process $X$ is mean-squared continuous if and only if the mean $\mu$ and covariance kernel $K$ of $X$ are continuous.

This can be used to show that $\mathbb{E}\|X\|^2 < \infty$, or equivalently that $K \in \mathbb{L}^2([0,1]^2)$. Then, one can define the covariance operator

$$\mathcal{K} : \mathbb{L}^2[0,1] \to \mathbb{L}^2[0,1] : f \mapsto \mathcal{K}f(t) = \int_0^1 K(s,t) f(s) ds.$$

This can be seen as a generalization of the covariance matrix, seen as a linear operator, in the multivariate setting. The covariance operator has the following properties:

**Theorem 2.**

1. $\mathcal{K}$ is well-defined, i.e., $\mathcal{K}f \in \mathbb{L}^2[0,1]$ for all $f \in \mathbb{L}^2[0,1]$.
2. $\mathcal{K}$ is self-adjoint.
3. $\mathcal{K}$ is positive semidefinite.
4. $\mathcal{K}$ is trace class with $\mathrm{Tr}(\mathcal{K}) = \int_0^1 K(t,t) dt$.
5. $\mathcal{K}$ is Hilbert-Schmidt with norm $\|\mathcal{K}\|_{HS} = \|K\|_{\mathbb{L}^2}$.

*Proof.* Since the definition of $K$ and thus that of $\mathcal{K}$ do not depend on $\mu$, we assume $\mu \equiv 0$.

1. For all $f \in \mathbb{L}^2[0,1]$,

$$\int_0^1 (\mathcal{K}f)^2(t) dt = \int_0^1 \left(\int_0^1 K(s,t) f(s) ds\right)^2 dt \leq \iint K^2(s,t) ds dt \cdot \int_0^1 f^2(t) dt < \infty.$$

Moreover, this shows that the operator norm of $\mathcal{K}$ is at most that of $K$ in $\mathbb{L}^2([0,1]^2)$.

2. For any $f, g \in \mathbb{L}^2[0,1]$,

$$\langle \mathcal{K}f, g\rangle = \iint K(s,t)f(s)g(t)dsdt = \iint K(t,s)g(t)f(s)dtds = \langle f, \mathcal{K}g\rangle,$$

where we have used the symmetry of $K$.

3. For $n \in \mathbb{N}$, using the compactness of $[0,1]^2$ and the continuity of $K$, one can find a finite sequence

$$\{0 \leq x_1 < x_2 < ... < x_r \leq 1\} \subset [0,1]$$

such that

$$|K(s,t) - K(x_i, x_j)| \leq \frac{1}{n} \text{ if } x_i \leq s < x_{i+1},\ x_j \leq t < x_{j+1}.$$

Let $K_n$ denote the induced piecewise constant kernel. Then,

$$\|K - K_n\|_{\mathbb{L}^2} \leq \frac{1}{n}$$

, and so, using the bound for the operator norm found above, we get

$$|\langle \mathcal{K}f, f\rangle - \langle \mathcal{K}_n f, f\rangle| \leq \frac{1}{n}\|f\|^2.$$

Finally, we use the positive semidefiniteness of $K$ to get

$$\langle \mathcal{K}_n f, f\rangle = \sum_{i=1}^{r}\sum_{j=1}^{r} K(x_i, x_j)\int_{x_i}^{x_{i+1}} f \int_{x_j}^{x_{j+1}} f \geq 0,$$

and so $\langle \mathcal{K}f, f\rangle \geq 0$.

4. Let $\{\phi_i\}$ be a complete orthonormal system of $\mathbb{L}^2[0,1]$. Then

$$\langle \mathcal{K}\phi_i, \phi_i\rangle = \iint \mathbb{E}[X(s)X(t)]\phi_i(s)\phi_i(t)dsdt = \mathbb{E}\left[\iint X(s)\phi_i(s)X(t)\phi_i(t)dsdt\right] = \mathbb{E}[\langle X, \phi_i\rangle^2].$$

Thus,

$$\mathrm{Tr}(\mathcal{K}) = \sum_{i=1}^{\infty}\langle \mathcal{K}\phi_i, \phi_i\rangle = \sum_{i=1}^{\infty}\mathbb{E}[\langle X, \phi_i\rangle^2] = \mathbb{E}\left[\sum_{i=1}^{\infty}\langle X, \phi_i\rangle^2\right] = \mathbb{E}\|X\|^2 = \int_0^1 K(t,t)dt.$$

5. For any $s \in [0,1]$, the slice function $K_s(t) = K(s,t)$ is in $\mathbb{L}^2[0,1]$. Let $\{\phi_i\}$ be a complete orthonormal system of $\mathbb{L}^2[0,1]$. Then,

$$\mathcal{K}\phi_i(s) = \int_0^1 K(s,t)\phi_i(t)dxdy = \int_0^1 K_s(t)\phi_i(t)dxdy = \langle K_s, \phi_i\rangle.$$

Thus,

$$\begin{aligned}\|\mathcal{K}\|_{HS}^2 &= \sum_{i=1}^{\infty}\|\mathcal{K}\phi_i\|^2 = \sum_{i=1}^{\infty}\int_0^1 \langle K_s, \phi_i\rangle^2 ds \\ &= \int_0^1 \sum_{i=1}^{\infty}\langle K_s, \phi_i\rangle^2 ds = \int_0^1 \|K_s\|^2 ds \\ &= \int_0^1\int_0^1 K^2(s,t)dtds = \|K\|_{\mathbb{L}^2}.\end{aligned}$$

□

An essential result in the study of positive kernels is Mercer's theorem, which relates the kernel $K$ to the spectrum of $\mathcal{K}$.

**Theorem 3** (Mercer)**.** Let $K$ be a continuous symmetric positive semidefinite kernel and let $\mathcal{K}$ be its associated operator on $\mathbb{L}^2[0,1]$. By the previous theorem, $\mathcal{K}$ is a compact self-adjoint operator and thus admits a complete orthonormal system of eigenfunctions $\{\lambda_i, \phi_i\}$. Then,

$$K(s,t) = \sum_{i=1}^{\infty} \lambda_i \phi_i(s)\phi_i(t),$$

where the convergence is absolute and uniform.

We are now ready to prove the most important result in functional data analysis, the Karhunen–Loève theorem.

**Theorem 4** (Karhunen–Loève)**.** Let $X$ be a second-order mean-squared continuous process with mean $\mu$ and covariance kernel $K$. Let $\{\lambda_i, \phi_i\}$ be the complete orthonormal basis given by Mercer's theorem. Then, the random variables

$$\xi_i = \langle X - \mu, \phi_i \rangle$$

have mean 0 and are uncorrelated with

$$\mathbb{E}[\xi_i \xi_j] = \lambda_i \delta_{ij}.$$

Moreover, the sequence of processes

$$X_n(t) = \mu(t) + \sum_{i=1}^{n} \xi_i \phi_i(t)$$

converges uniformly to $X(t)$, that is

$$\sup_{t\in[0,1]} \mathbb{E}\left[X(t) - X_n(t)\right]^2 \overset{n\to\infty}{\longrightarrow} 0.$$

*Proof.* By replacing $X$ and $X_n$ with $X - \mu$ and $X_n - \mu$ in the statement of the theorem, we can assume without loss of generality that $\mu \equiv 0$.We have

$$\mathbb{E}[\xi_i \xi_j] = \mathbb{E}[\langle X - \mu, \phi_i\rangle \langle X - \mu, \phi_j \rangle] = \langle \mathcal{K}\phi_i, \phi_j\rangle = \lambda_i \delta_{ij},$$

so that the $\xi_i$'s are uncorrelated with variance $\lambda_i$. For any $t$ in $[0,1]$, we have

$$\begin{aligned}
\mathbb{E}\left[X(t) - X_n(t)\right]^2 &= \mathbb{E}\left[X(t)\right]^2 - 2\mathbb{E}\left[X(t)X_n(t)\right] + \mathbb{E}\left[X_n(t)\right]^2 \\
&= K(t,t) - 2\sum_{i=1}^{n} \phi_i(t)\mathbb{E}[\xi_i X(t)] + \sum_{i=1}^{n} \mathrm{Var}[\xi_i \phi_i(t)] \\
&= K(t,t) - 2\sum_{i=1}^{n} \phi_i(t)\mathbb{E}\left[X(t)\int X(s)\phi_i(s)ds\right] + \sum_{i=1}^{n} \lambda_i \phi_i^2(t) \\
&= K(t,t) - 2\sum_{i=1}^{n} \lambda_i \phi_i^2(t) + \sum_{i=1}^{n} \lambda_i \phi_i^2(t) \\
&= K(t,t) - \sum_{i=1}^{n} \lambda_i \phi_i^2(t).
\end{aligned}$$

The last term tends uniformly to 0 by Mercer's theorem. □

This theorem gives a stochastic analogue of the fact that $\mathbb{L}^2[0,1]$ is isomorphic to $\ell^2(\mathbb{R})$. More precisely, it states that the variation of $X$ around its mean can be expressed in terms of the countable sequence of random variables $\{\xi_i\}$ and the eigenfunctions $\{\phi_i\}$. We now give a concrete example of this theorem in the case of a standard Wiener process (following [Hsing and Eubank, 2015, Example 4.6.3]).

*Example* 1. Let $W$ be a standard Wiener process on $[0,1]$. It is well-known that its covariance kernel is given by

$$K(s,t) = \min(s,t).$$

Thus, its covariance operator is given by

$$\mathcal{K}f(t) = \int_0^1 \min(s,t) f(s) ds = \int_0^t s f(s) ds + t \int_t^1 f(s) ds.$$

To find the eigenpairs of $\mathcal{K}$, we solve $\mathcal{K}\phi = \lambda\phi$ for $\lambda$ and $\phi$:

$$\int_0^t s\phi(s)ds + t\int_t^1 \phi(s)ds = \lambda\phi(t), \ \forall t \in [0,1]. \tag{1.1}$$

Differentiating with respect to $t$ yields

$$t\phi(t) + \int_t^1 \phi(s)ds - t\phi(t) = \int_t^1 \phi(s)ds = \lambda\phi'(t). \tag{1.2}$$

Differentiating again yields $\phi(t) = -\lambda\phi''(t)$. The general solution of this differential equation is given by

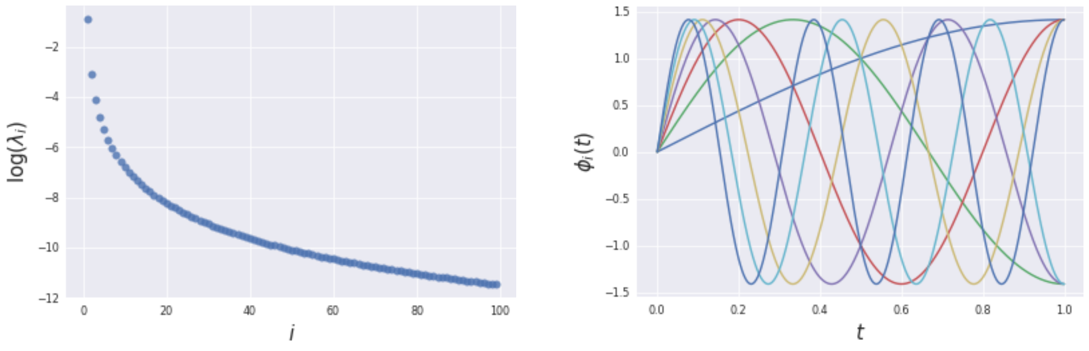
$$\phi(t) = a\sin\left(\frac{t}{\sqrt{\lambda}}\right) + b\cos\left(\frac{t}{\sqrt{\lambda}}\right).$$

Equation 1.1 implies that $\phi(0) = 0$, so that $b = 0$. Then, Equation 1.2 gives $\phi'(1) = 0$, so that $a\cos(1/\sqrt{\lambda}) = 0$. This implies

$$\frac{1}{\sqrt{\lambda}} \in \left\{\frac{(2i-1)\pi}{2} \ \mid \ j = 1, 2, ...\right\}.$$

Finally, combining all of this gives the following orthonormal system of eigenpairs, shown in Figure 1.2.

$$\lambda_i = \frac{4}{((2i-1)\pi)^2}, \ \phi_i(t) = \sqrt{2}\sin\left(\frac{(2i-1)\pi}{2}t\right).$$



(a) Logarithm of the first 100 eigenvalues of $\mathcal{K}$.

(b) First 7 eigenfunctions of $\mathcal{K}$.

Figure 1.2

Since $W$ is a Gaussian process, the random variables $xi_i = \langle W, \phi_i \rangle$ are Gaussian in the Karhunen–Loève theorem. Therefore, we can explicitly generate the finite sum approximations $W_n$. An example is shown in Figure 1.3.

Two of the most important theorems in classical statistics are the strong law of large numbers and the central limit theorem. Fortunately, these two results still hold in the functional setting. We state them here without proof, which can be found in [Hsing and Eubank, 2015, Theorem 7.7.2 and 7.7.6].

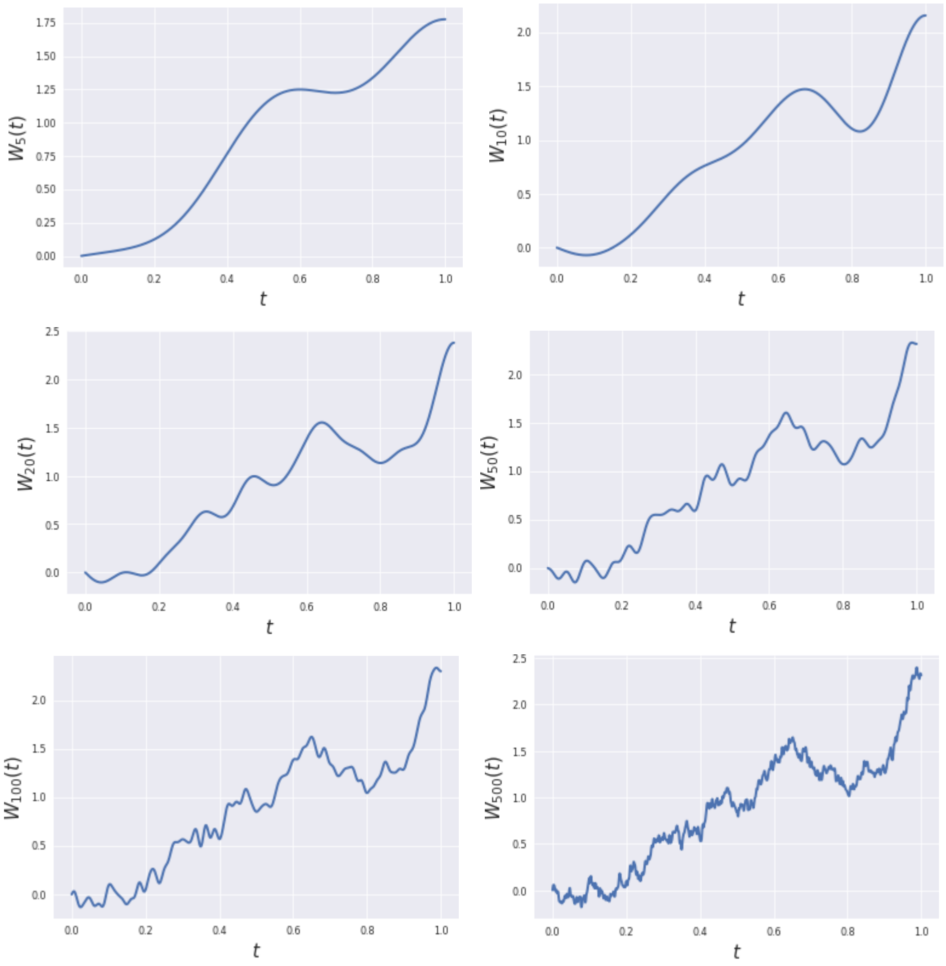


Figure 1.3: Finite sum approximations $W_n$ of $W$ for $n = 5, 10, 20, 50, 100, 500$.

**Theorem 5** (Strong law of large numbers)**.** Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of i.i.d stochastic processes with $\mathbb{E}\|X_1\|_{\mathbb{L}^2} < \infty$ and $\mathbb{E}[X_1] = \mu$. Let $S_n = X_1 + ... + X_n$ be the sequence of partial sums. Then,

$$\frac{S_n}{n} \xrightarrow{a.s.} \mu.$$

**Theorem 6** (Central limit theorem)**.** Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of i.i.d stochastic processes with $\mathbb{E}\|X_1\|_{\mathbb{L}^2}^2 < \infty$, $\mathbb{E}[X_1] = \mu$ and covariance operator $\mathcal{K}$. Then,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu) \xrightarrow{dist} N_{0,\mathcal{K}},$$

where $N_{0,\mathcal{K}}$ denotes a Gaussian measure om $\mathbb{L}^2[0,1]$, meaning that for any $f \in \mathbb{L}^2[0,1]$,

$$\left\langle f, \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu) \right\rangle \xrightarrow{dist} N_{0,\langle \mathcal{K}f, f\rangle},$$

see Chapter 3.

## 1.3 Statistical analysis of functional data

We now take on the task of estimating some parameters of a process from observing a finite sample. Our exposition follows that of [Horváth and Kokoszka, 2012]. Our setup will be as follows. We

observe $N$ i.i.d. samples $X_1, ..., X_N$ coming from a second-order mean-squared continuous process. In this section, we will see how to estimate the mean, covariance kernel and eigenpairs of the covariance operator of this process.
We start with the mean $\mu = \mathbb{E}[X_1]$. As in the finite dimensional setting, we estimate it using the sample mean

$$\hat{\mu}(t) = \frac{1}{N} \sum_{i=1}^{N} X_i(t).$$

The next theorem shows that $\hat{\mu}$ is a unbiased $\mathbb{L}^2$-consistent estimator.

**Theorem 7.** The process $\hat{\mu}$ satisfies $\mathbb{E}[\hat{\mu}] = \mu$ and

$$\mathbb{E}\|\hat{\mu} - \mu\|^2 = O\left(\frac{1}{N}\right).$$

Moreover,

$$\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{dist} N_{0,\mathcal{K}}.$$

*Proof.* For any $t \in [0, 1]$,

$$\mathbb{E}[\hat{mu}(t)] = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}[X_i(t)] = \mu(t),$$

whence $\mathbb{E}[\hat{\mu}] = \mu$. We note that $\hat{\mu} - \mu = N^{-1} \sum_{i=1}^{N} (X_i - \mu)$, so that

$$\mathbb{E}\|\hat{\mu}-\mu\|^2 = \frac{1}{N^2} \sum_{i,j=1}^{N} \mathbb{E}\left[\langle X_i - \mu, X_j - \mu\rangle\right] = \frac{1}{N^2} \sum_{i=1}^{N} \mathbb{E}\left[\langle X_i - \mu, X_i - \mu\rangle\right] = \frac{1}{N}\mathbb{E}\|X_1-\mu\|^2 = O\left(\frac{1}{N}\right),$$

where we used that $X_i$ and $X_j$ are independent for $i \neq j$, so that

$$\mathbb{E}\left[\langle X_i - \mu, X_j - \mu\rangle\right] = 0.$$

The last statement follows immediately from the central limit theorem. □

The estimation of the covariance kernel is also performed similarly to the finite dimensional setting. We estimate it with

$$\hat{K}(s, t) = \frac{1}{N} \sum_{i=1}^{N} (X_i(s) - \hat{\mu}(s))(X_i(t) - \hat{\mu}(t)).$$

As in the finite dimensional case, this estimator is biased with

$$\mathbb{E}[\hat{K}(s, t)] = \frac{N-1}{N} K(s, t),$$

which is negligible when $N$ is large. We estimate the covariance operator by the induced operator of $\hat{K}$, i.e., with

$$\hat{\mathcal{K}} f(t) = \int_0^1 \hat{K}(s, t) f(s) ds = \frac{1}{N} \sum_{i=1}^{N} (X_i(t) - \hat{\mu}(t)) \langle X_i - \hat{\mu}, f\rangle.$$

This last expression shows that the range of $\hat{\mathcal{K}}$ is contained in the span of $\{X_1, ..., X_N\}$ and is thus finite dimensional. This is the source of many problems in functional data analysis, since the true covariance operator is often infinite dimensional.
To analyze this estimator, we need to choose an operator space in which we can work. As was shown in Theorem 2, the true covariance operator $\mathcal{K}$ lives in the spaces of bounded, Hilbert-Schmidt and trace class operators. We note that $\hat{\mathcal{K}}$ also belongs to these spaces since it has finite dimensional range. Of those, the space of Hilbert-Schmidt operators $\mathcal{S}(\mathbb{L}^2[0, 1])$ is the most convenient since it is itself a separable Hilbert space which makes the computations easier. To make them even easier, we assume that $\mu \equiv \hat{\mu} \equiv 0$. This assumption is not too optimistic in practice since $\hat{\mu}$ converges quite rapidly to $\mu$. The next theorem shows that $\hat{\mathcal{K}}$ has good properties.

**Theorem 8.** If $\mathbb{E}\|X_1\|^4 < \infty$, the estimator $\hat{\mathcal{K}}$ satisfies:

1. $\hat{\mathcal{K}} \xrightarrow{a.s.} \mathcal{K}$,
2. $\mathbb{E}\|\hat{\mathcal{K}} - \mathcal{K}\|_{HS}^2 \leq \frac{1}{N}\mathbb{E}\|X_1\|^4$,

*Proof.* 1. As noted above, $\hat{K}$ can be written as $N^{-1}\sum_{i=1}^N X_i\langle X_i, \cdot\rangle$. Consider the operators

$$\mathcal{K}_i = X_i\langle X_i, \cdot\rangle - \mathcal{K},\ i = 1, ..., N.$$

Then, $\hat{\mathcal{K}} - \mathcal{K} = N^{-1}\sum_{i=1}^N \mathcal{K}_i$ and $\mathbb{E}[\mathcal{K}_i] = 0$ for $i = 1, ..., N$. Therefore, by the strong law of large numbers,

$$\hat{\mathcal{K}} - \mathcal{K} \xrightarrow{a.s} 0 \iff \hat{\mathcal{K}} \xrightarrow{a.s} \mathcal{K}.$$

2. Let $\{\phi_k\}$ is a complete orthonormal system of $\mathbb{L}^2[0, 1]$,

$$\|X_i\langle X_i, \cdot\rangle\|_{HS}^2 = \sum_{k=1}^{\infty} \|X_i\langle X_i, \phi_k\rangle\|^2 = \|X_i\|^2 \sum_{k=1}^{\infty}\langle X_i, \phi_k\rangle^2 = \|X_i\|^4.$$

Then, by similar computations as those carried out in the proof of Theorem 7, one can show that (see [Horváth and Kokoszka, 2012, Theorem 2.5])

$$\mathbb{E}\|\hat{\mathcal{K}} - \mathcal{K}\|_{HS}^2 = \mathbb{E}\left\langle \frac{1}{N}\sum_{i=1}^N \mathcal{K}_i, \frac{1}{N}\sum_{i=1}^N \mathcal{K}_i\right\rangle_{HS} \leq \mathbb{E}\|X_1\|^4.$$

□

One of the most popular tool in functional data analysis is the functional principal component analysis (FPCA) method. This method uses the Karhunen-Loève theorem to reduce the dimension of the data to infinite to finite. Similarly to the PCA method in multivariate analysis, we approximate the process $X$ using only the $r$ principal components (we still assume $\mathbb{E}[X] \equiv 0$):

$$X \simeq X^r = \sum_{i=1}^r \xi_i\phi_i.$$

The process $X^r$ has covariance operator $\mathcal{K}^r = \sum_{i=1}^r \lambda_i\phi_i \otimes \phi_i$ and therefore "explain" a ratio of

$$\frac{\mathrm{Tr}(\mathcal{K}^r)}{\mathrm{Tr}(\mathcal{K})} = \frac{\sum_{i=1}^r \lambda_i}{\sum_{i=1}^\infty \lambda_i}$$

of the variance of $X$.
In practice, one does not have access to the eigenpairs $\{\lambda_i, \phi_i\}_{i=1}^r$ of $\mathcal{K}$ and we thus need a way to approximate them. The obvious way to do so is by using the eigenpairs $\{\hat{\lambda}_i, \hat{\phi}_i\}_{i=1}^r$ of $\mathcal{K}$. We assume that $\lambda_1 > ... > \lambda_r > 0$ to make the problem identifiable. Moreover, we assume that we can the estimates $\hat{\phi}_i$ have good signs, i.e., that $\langle\phi_i, \hat{\phi}_i\rangle \geq 0$ for all $i$. In practice, one is only interested in the eigenspaces, so this last assumption is not to restrictive.
Then, the next theorem, proved in [Horváth and Kokoszka, 2012, Section 2.7], shows that $\{\hat{\lambda}_i, \hat{\phi}_i\}_{i=1}^r$ are consistent estimators of $\{\lambda_i, \phi_i\}_{i=1}^r$.

**Theorem 9.** Suppose $\mathbb{E}\|X_1\|^4 < \infty$, $\lambda_1 > ... > \lambda_r > 0$, and $\langle\phi_i, \hat{\phi}_i\rangle \geq 0$ for all $i$. Then, for all $i = 1, ..., r$,

$$\limsup_{N\to\infty} N\mathbb{E}\|\hat{\phi}_i - \phi_i\|^2 < \infty \quad\text{and}\quad \limsup_{N\to\infty} N\mathbb{E}|\hat{\lambda}_i - \lambda_i|^2 < \infty.$$

Consequently, given a new observation $\tilde{X}$ one can estimate the component scores $\langle\tilde{X}, \phi_i\rangle \sim \xi_i$ with $\langle\tilde{X}, \hat{\phi}_i\rangle$ and get

$$\sup_{i=1...r} \mathbb{E}\left(\langle\tilde{X}, \hat{\phi}_i\rangle - \langle\tilde{X}, \phi_i\rangle\right)^2 = \sup_{i=1...r} \mathbb{E}\langle\tilde{X}, \hat{\phi}_i - \phi_i\rangle^2 \leq \sup_{i=1...r} \mathbb{E}\|\tilde{X}\|^2\|\hat{\phi}_i - \phi_i\|^2 \xrightarrow{N\to\infty} 0.$$

## 1.4 Discrete observations and smoothing

In practice, curves are never observed on their full domain, they are observed on a finite discrete subset $\{t_1, ..., t_k\}$. Moreover, each observation can be subject to measurement errors. Therefore, in practice the observations are of the form:

$$X_{ij} = X_i(t_j) + \epsilon_{ij},\ i = 1, ..., N,\ j = 1, ..., K,$$

where the $\epsilon_{ij}$ are i.i.d, of mean 0 and variance $\sigma^2$ and independent of the $X_i$'s.
To be able to work with functional data, the full curves $X_i$ need to be approximated. This can be done by smoothing either the observations themselves or the covariance kernel.
To smooth the observations, a popular approaches is the penalized smoothing method. One first needs to choose a function space $\mathbb{W}$ in which we want our approximations to live, and a penalty function $P : \mathbb{W} \to \mathbb{R}_+$ that penalizes unwanted behaviors. Then, we approximate $X_i$ with

$$\tilde{X}_i = \arg\min_{f \in \mathbb{W}} \left[ \sum_{j=1}^{K} (X_i(t_j) - f(t_j))^2 + \lambda P(f) \right],$$

where $\lambda > 0$ is a tuning parameter. A popular choice for $\mathbb{W}$ and $P$ are

$$\mathbb{W} = C^2[0, 1] \text{ and } P(f) = \int_0^1 \left| \frac{d^2 f}{dt^2}(t) \right|^2 dt.$$

With this choice, we ask for our approximation to be twice differentiable and not to be too rough. Moreover, one can show that an explicit solution can be found in terms of cubic splines. For more details, see [Ramsay and Silverman, 2005, Chapter 5]. Using the smooth estimates $\tilde{X}_1, ..., \tilde{X}_N$, the mean and covariance kernel can be estimated, as explained in the previous section.
When the observation grid $\{t_1, ..., t_K\}$ is sparse, the penalized smoothing approach will typically yield biased estimates. In that case, a popular approach is the PACE method described in [Yao et al., 2005]. In this method, one starts by estimating the covariance kernel. This is done by smoothing the covariance matrix of the observations $X_{ij}$, $i = 1, ..., N$, $j = 1, ..., K$, where the diagonal is removed to take care of the noise $\epsilon_{ij}$. Then, eigenpairs of the covariance operator are estimated by $\{\hat{\lambda}_k, \hat{\phi}_k\}$, the eigenpairs of the estimated smoothed covariance kernel. Finally, the components scores $\xi_{ik} = \langle X_i, \phi_k \rangle$ are estimated using the explicit formulat for the conditional expectation in the Gaussian case:

$$\hat{\xi}_{ik} = \widehat{\mathbb{E}}[\xi_{ik} \mid X_i(t_1), ..., X_i(t_K)].$$

The full curves $X_i$ are then estimated using the FPCA method:

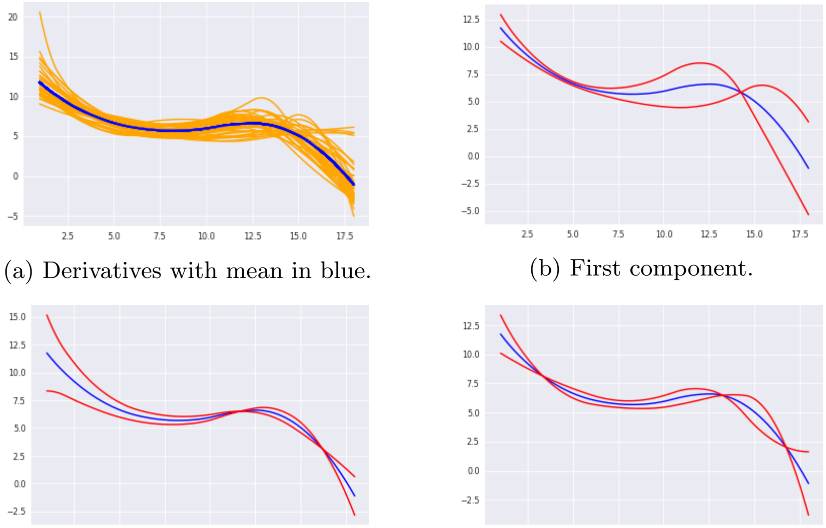
$$X_i = \hat{\mu} + \sum_{k=1}^{r} \hat{\xi}_{ik} \hat{\phi}_k.$$

This method is shown to have better asymptotic properties when the observations are sparse.

*Example* 2. We illustrate the methods above with the growth data set of [Tuddenham and Snyder, 1954], shown in Figure 1.1. We focus only on boys, for which there are 39 samples observed at 95 regularly sampled ages. We start by smoothing the observations using cubic splines. We then compute the derivatives of these smooth estimates. These derivatives are shown in the top left of Figure 1.4. We then compute the principal components using the eigenpairs of the estimated covariance kernel. The first three principal components explain 85% of the total variance. In the top right and bottom of Figure 1.4, we show the mean derivative in blue, along with the curves

$$\hat{\mu} \pm 2\sqrt{\lambda_i}\phi_i,\ i = 1, 2, 3$$

in red. We see that the first principal component explains two main modes of variation, one between 8 and 13, and the other between 16 and 18. Individuals having a high positive score in this component typically experience their puberty earlier and have a big spike in their growth around 12 years old. On the other hand, individuals having a high negative score in this component will experience puberty later, around 15 years old. The second component deals mainly with the variation close to birth. The third component is more difficult to interpret and is similar to the first component.



(a) Derivatives with mean in blue.

(b) First component.

(c) Second component.

(d) Third component.

Figure 1.4: Example of FPCA on the growth data set.

This example shows how FPCA can be used to analyze the variation of a population of curves. We note that our analysis is rather shallow since it does not take care of the time-registration problem. A thorough analysis of this dataset is performed in [Ramsay and Silverman, 2013, Chapter 6]

# Chapter 2

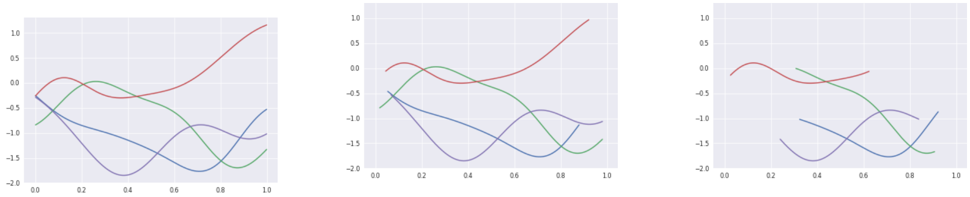
# Censored functional data

## 2.1 Functional fragments framework

*Censored functional data* or *functional fragments* are functional data that are not observed in the full domain on which they are defined. If the data live in $\mathbb{L}^2(\mathcal{I})$ for some interval $\mathcal{I} \subset \mathbb{R}$, an example of functional fragment is a function $f \in \mathbb{L}^2(\mathcal{J})$ for some interval $\mathcal{J} \subset \mathcal{I}$.

By the *functional fragments framework*, we mean the statistical framework in which some or all of the observed data are in the form of fragments of some underlying, unobservable, functional data. In this case, the data at hand are pairs $\{(X_i, \mathcal{O}_i)\}_{i=1}^n$, where the $X_i$'s are random functions in $\mathbb{L}^2(\mathcal{O}_i)$ for some subintervals $\mathcal{O}_i$. We will often assume that the subintervals $\{\mathcal{O}_i\}_{i=1}^n$ are themselves random, in order to make the asymptotic theory more tractable. This framework often arises in practice when an observation is unavailable before or after a certain time.
The main issue in the functional fragments framework is to know to what extent one can recover precise information on the underlying population from the observed fragments. For example, how precisely can we estimate the mean and covariance when no curve is fully observed.
Following [Descary and Panaretos, 2017], we distinguish between two ways in which the intervals $\{\mathcal{O}_i\}_{i=1}^n$ are distributed, see Figure 2.1:

**1.** A "blanket " regime, where the curves are typically observed on most or all of the domain. Then the number of observations at a given point of the domain is close to the total number of observations.
**2.** A "banded" regime, where the lengths of the $\mathcal{O}_i$ are bounded by some value $\delta > 0$. Then, we have no explicit information on the covariance of points that are at distance larger than $\delta$.



(a) Full curves

(b) Fragments in the blanket regime

(c) Fragments in the banded regime

Figure 2.1: Example of functional fragments

In the rest of this chapter, we will present various methods used in the literature dealing with those issues.

## 2.2 Naive estimations

We present here the methods presented in [Kraus, 2015]. The data are i.i.d curves $X_i$ in $\mathbb{L}^2[0,1]$ observed only on a random interval $\mathcal{O}_i \subset [0,1]$, $i = 1, ..., n$. To estimate the population mean

$\mu = \mathbb{E}[X_1]$ and covariance operator $\mathcal{K} = \mathbb{E}[(X_1 - \mu) \otimes (X_1 - \mu)]$, the unobserved parts of the curves are ignored and the sample estimators are created naively as follows. The sample mean $\hat{\mu}$ is found by taking the mean of the pointwise observed values:

$$\hat{\mu}(t) = \frac{\mathbf{1}[\exists \mathcal{O}_i \ni t]}{\sum_{i=1}^{n} \mathbf{1}[t \in \mathcal{O}_i]} \sum_{i=1}^{n} \mathbf{1}[t \in \mathcal{O}_i] \cdot X_i(t). \tag{2.1}$$

The covariance operator is estimated in the same fashion via its associated covariance kernel $K(\cdot, \cdot)$. The sample kernel is given by:

$$\hat{K}(s,t) = \frac{\mathbf{1}[\exists \mathcal{O}_i \ni s, t]}{\sum_{i=1}^{n} \mathbf{1}[s, t \in \mathcal{O}_i]} \sum_{i=1}^{n} \mathbf{1}[s, t \in \mathcal{O}_i] \cdot \{X_i(s) - \hat{\mu}_{st}(s)\} \{X_i(t) - \hat{\mu}_{st}(t)\}, \tag{2.2}$$

where $\hat{\mu}_{st}$ is an estimation of the mean using only the curves observed at $s$ and $t$:

$$\hat{\mu}_{st}(s) = \frac{\mathbf{1}[\exists \mathcal{O}_i \ni s, t]}{\sum_{i=1}^{n} \mathbf{1}[s, t \in \mathcal{O}_i]} \sum_{i=1}^{n} \mathbf{1}[s, t \in \mathcal{O}_i] \cdot X_i(s).$$

The sample covariance operator $\hat{\mathcal{K}}$ is then defined by

$$\hat{\mathcal{K}} f(t) = \int_0^1 \hat{K}(s,t) f(s) \, \mathrm{d}s.$$

We note that this operator need not be positive-definite. This can be dealt with by clipping the negative eigenvalues to zero.
The following proposition, proved in [Kraus, 2015, Prop. 1], shows that under some assumptions on the random intervals $\{\mathcal{O}_i\}_{i=1}^{n}$, the above estimates enjoy the same asymptotic convergence rate as their counterparts when the curves are fully observed.

**Proposition 2.**

1. Suppose that $\mathbb{E}\|X_1\|^2 < \infty$ and the $\mathcal{O}_i$'s are i.i.d with $\inf_{t \in [0,1]} \mathbb{P}[t \in \mathcal{O}_1] > 0$. Then

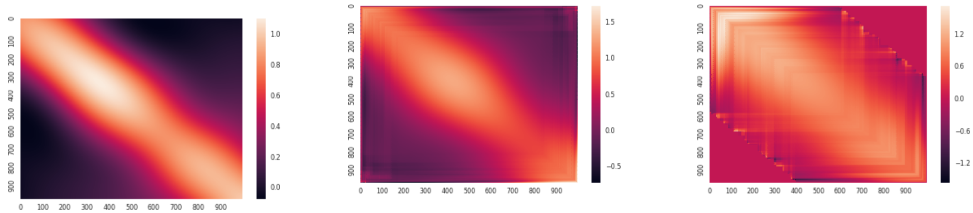
$$\mathbb{E}\|\hat{\mu} - \mu\|^2 = O(n^{-1}) \text{ as } n \to \infty.$$

2. Suppose further that $\mathbb{E}\|X_1\|^4 < \infty$ and that $\inf_{s,t \in [0,1]} \mathbb{P}[s, t \in \mathcal{O}_1] > 0$. Then

$$\mathbb{E}\|\hat{\mathcal{K}} - \mathcal{K}\|_{HS}^2 = O(n^{-1}) \text{ as } n \to \infty.$$

□

For the covariance operator, even though the theoretical convergence rate is good, in practice the estimate is not adequate in the "banded" regime (see Section 2.1). The problem is that in this regime, the estimated kernel is necessarily zero in the region $\{(s,t) \in [0,1]^2 \mid |s - t| > \delta\}$. This problem that is not present in the "blanket" regime, as soon as one full curve is observed, see Figure .



(a) Full curves

(b) Fragments in the blanket regime

(c) Fragments in the banded regime

Figure 2.2: Sample estimates of the covariance kernel in the different regimes

## 2.3 Curve extension

In [Delaigle and Hall, 2013], the authors take the approach of manually extend the fragments by gluing some of their parts. This procedure is carried on in the context of classification, but it can readily be expended to the plain estimation of the population mean and covariance.
The extension of a fragment to the right is done by iteratively gluing a small section of another randomly chosen nearby fragment to the right endpoint of the original fragment. Likewise for the extension to the left. A few steps of this procedure are shown in Figure 2.3.

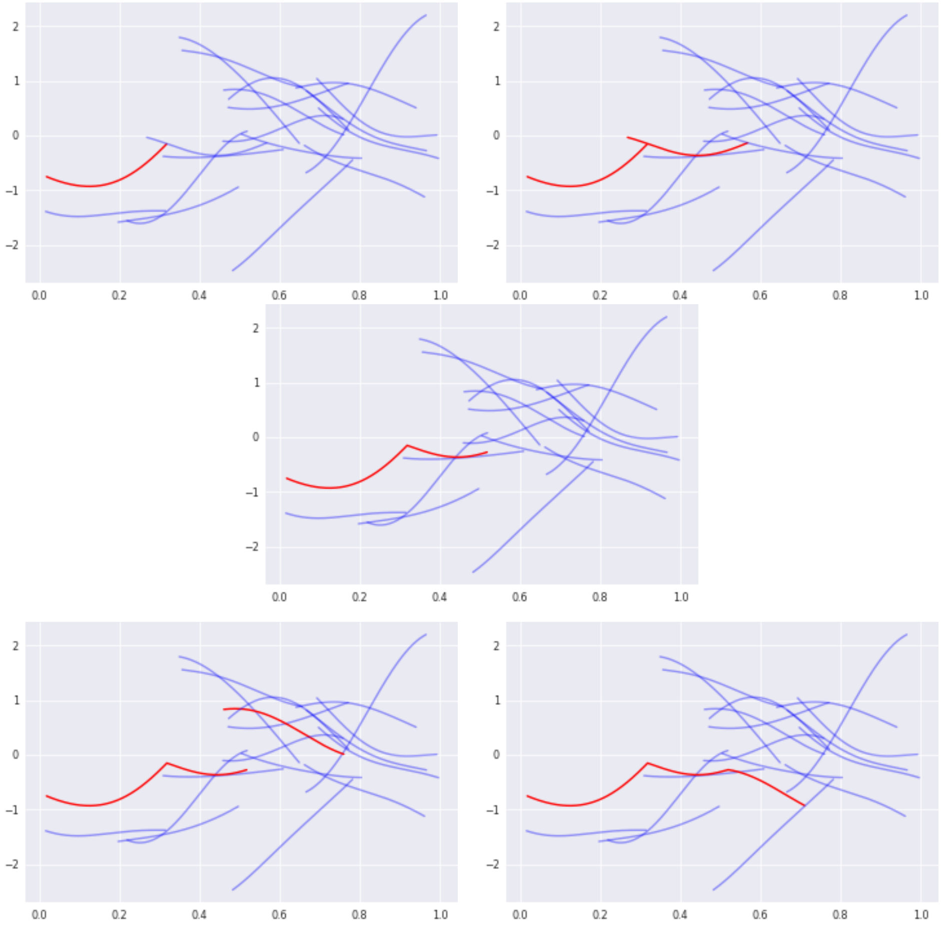
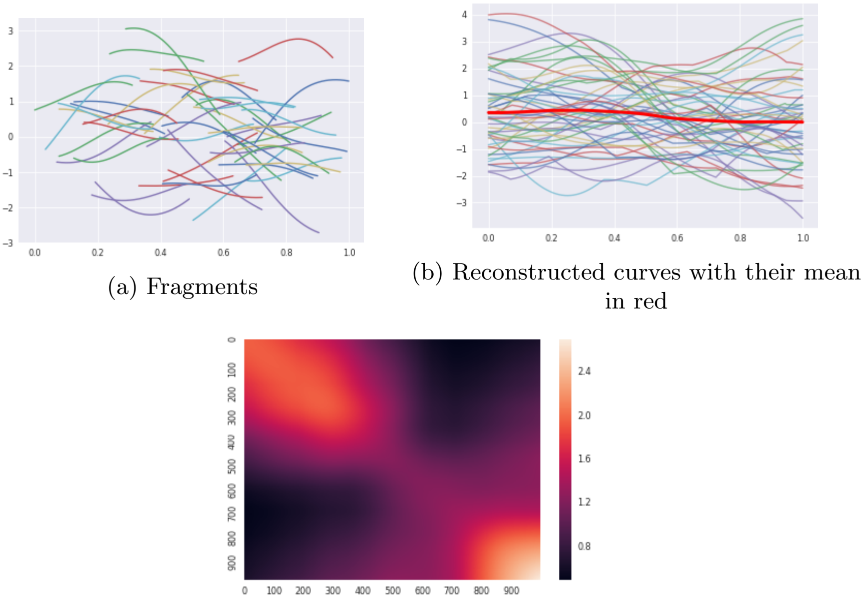


Figure 2.3: Illustration of a few steps of the gluing procedure

In this way, for each fragment $(X_i, \mathcal{O}_i)$, one gets a curve $\tilde{X}_i$ defined on the entire domain. From these full curves, the sample mean and covariance can be estimated. This estimation method works both in the "blanket" and the "banded" regime, since in both cases full curves are constructed. Examples of those estimates are shown in Figure 2.4.
To the best of our knowledge, no theoretical convergence rates for these estimates have been found. We also note that another function extension method for functional fragments is presented in [Delaigle and Hall, 2016] by the same authors. This method assumes that discretized versions of the curves are Markov processes and then extend the fragments using this assumption.

## 2.4 Covariance recovery

In this section, we present the covariance recovery method of [Descary and Panaretos, 2017]. In this paper, the authors show that under some smoothness assumptions, one can use a matrix completion method to recover the covariance matrix of a process from the observations of fragments in the "banded" regime.



(a) Fragments

(b) Reconstructed curves with their mean in red

(c) Estimated covariance kernel form the reconstructed curves

Figure 2.4: Example of mean and covariance estimation using the gluing procedure

### 2.4.1 Setup

We consider a continuous stochastic process $X \in \mathbb{L}^2[0,1]$ with mean function $\mathbb{E}[X] = \mu$ and covariance kernel $r(s,t) = \text{cov}\{X(s), X(t)\}$. Suppose we observe i.i.d. fragments $(X_i, \mathcal{O}_i)$ of length $\delta \in (0,1)$. Then, one can compute the estimates $\tilde{\mu}_n$ and $\tilde{r}_n$ given by Equations 2.1 and 2.2. Suppose further that we observe these fragments only on a grid of size $K$

$$(t_1, ..., t_K) \in \mathcal{T}_K = \{(x_1, ..., x_K) \in \mathbb{R}^K \mid x_i \in I_{i,K}\},$$

where $\{I_{i,K}\}_{i=1}^K$ is the regular partition of $[0,1]$ in intervals of length $1/K$. From these discrete observations, one can compute the mean vector $\tilde{\mu}_n^K = (\tilde{\mu}_n(t_i))_{i=1}^K$ and the covariance matrix $\tilde{R}_n^K = \{\tilde{r}_n(t_i, t_j)\}_{i,j=1}^K$. As noted above, this matrix will have a banded structure. Indeed, each fragment $X_i$ is observed only on $\mathcal{O}_i \cap \{t_i\}_{i=1}^K$, which has between $\lfloor K\delta \rfloor - 1$ and $\lceil K\delta \rceil + 1$ points. Thus, the matrix $\tilde{R}_n^K$ is guaranteed to have non-zero values only on the band $\{(i,j) \mid |i-j| < \lfloor K\delta \rfloor - 1\}$. Denote by $R^K$ the true covariance matrix $\{r(t_i, t_j)\}_{i,j=1}^K$ and by $P_\delta^K \in \mathbb{R}^{K \times K}$ the band indicator matrix

$$P_\delta^K(i,j) = \mathbf{1}\left[\,|i-j| < \lfloor K\delta \rfloor - 1\right].$$

Then $\tilde{R}_n^K$ can realistically be seen only as an estimator of $P_\delta^K \circ R^K$, where "$\circ$" denotes the Hadamard product.
The question now is, under what non-parametric conditions on the process $X$ one can hope to efficiently recover the full covariance matrix $R^K$ from its banded version $P_\delta^K \circ R^K$? The estimation of the covariance in this setup can thus be seen as a matrix completion problem.

### 2.4.2 Identifiability and estimation

The main result of [Descary and Panaretos, 2017] is that under simple non-parametric assumptions given below, one can exactly recover the covariance matrix from its banded version. Moreover, they show that these condtions are, in some sense, necessary. These assumptions are the following:

1. The covariance kernel $r(s,t)$ has finite rank $q$, i.e., it admits a Mercer decomposition

$$r(s,t) = \sum_{j=1}^{q} \lambda_j \phi_j(s) \phi_j(t).$$

2. The eigenfunctions $\{\phi_1, ..., \phi_q\}$ are all real analytic.

The following results, shown in [Descary and Panaretos, 2016, Prop. 1], indicate that the second condition is not as strict as it seems.

**Proposition 3.** The set of trace class covariance operators of rank at most $q$ with analytic eigenfunctions is dense in the space of rank $q$ trace class covariance operators with the nuclear norm. In particular, for any process $X \in \mathbb{L}^2[0,1]$ with finite rank $q$ trace class operator and any $\epsilon > 0$, there exists a process $Y$, satisfying the conditions above, such that

$$\mathbb{E}\|X - Y\|_{\mathbb{L}^2}^2 < \epsilon.$$

□

The main property of finite rank analytic kernel we will use is the one known as analytic continuation. We will need it in the following form, which is a special case of [Krantz and Parks, 2002, Corollary 1.2.6]:

**Proposition 4.** Let $r, l : [0,1]^2 \to \mathbb{R}$ be two real analytic finite rank kernels. Suppose that there exists an open set $U \subseteq [0,1]^2$ such that $r(s,t) = l(s,t)$ for all $(s,t) \in U$. Then $r(s,t) = l(s,t)$ for all $(s,t) \in [0,1]^2$, i.e., $r$ and $l$ are equal. □

In our setup, this implies that knowing the covariance kernel on the band $\{(s,t) \in [0,1]^2 \mid |s-t| < \delta\}$ allows us to recover the kernel on the whole domain $[0,1]^2$. This shows that the model is identifiable in the sense that one can recover the true covariance from the observations of fragments.
In the discretized setup, this translates into the following identifiability result:

**Theorem 10** ([Descary and Panaretos, 2017])**.** Suppose that the covariance kernel $r(s,t)$ is analytic of finite rank $q$. If $\delta$ and $K$ are such that

$$K > \frac{q+2}{\delta}$$

then, for almost all $(t_1, ..., t_K) \in \mathcal{T}_K$, then the covariance matrix $R^K$ is the unique solution of

$$\min_{M \in \mathbb{R}^{K \times K}} \operatorname{rank}(M) \quad \text{subject to} \quad \|P_\delta^K \circ (R^K - M)\|_{Frob} = 0. \tag{2.3}$$

Equivalently, introducing a Lagrange multiplier, for all $\tau > 0$ sufficiently small,

$$R^K = \operatorname*{argmin}_{M \in R^{K \times K}} \left\{ \|P_\delta^K \circ (R^K - M)\|_{Frob}^2 + \tau \operatorname{rank}(M) \right\}$$

almost everywhere on $\mathcal{T}_K$.

*Proof.* By [Descary and Panaretos, 2016, Theorem 4], all $q \times q$ minors of $R^K$ are non-zero. Moreover, the condition $K > (q+2)/\delta$ implies that $R^K$ and $P_\delta^K \circ R^K$ share at least one common $q \times q$ submatrix. Combining these two results yields that $R^K$ has rank $q$ and that $P_\delta^K \circ R^K$ has rank at least $q$. Therefore $R^K$ is a solution of the matrix completion problem 2.3. It remains to show that it is the unique solution thereof. Let $R^*$ be another solution. Then, by design, $R^*$ has rank $q$ and is equal to $R^K$ on the band $B_\delta = \{(i,j) \mid |i-j| < \lfloor K\delta \rfloor - 1\}$. Since $K > (q+2)/\delta$, we can find a $(q+1) \times (q+1)$ submatrix $A$ of $R^*$ with exactly one entry $x^*$ outside the band $B_\delta$, see Figure 2.5. The determinant of $A$ can then be written as $ax^* + b$ where $a$ is the determinant of a $q \times q$ submatrix contained in $B_\delta$, and is thus non-zero, and $b$ depends only on the values of $R^*$ in $B_\delta$. Since $R^*$ has rank $q$,

$$\det(A) = ax^* + b = 0.$$

Therefore, $x^*$ is uniquely determined by the values of $R^*$ in $B_\delta$, and they are themselves fixed to be equal to the values of $R^K$ in that band. Therefore, all solutions of 2.3 are equal on the entry $x^*$.
One can iterate this procedure to reconstruct entirely the matrix $R^K$. The solution of 2.3 is thus unique and given by $R^K$. □

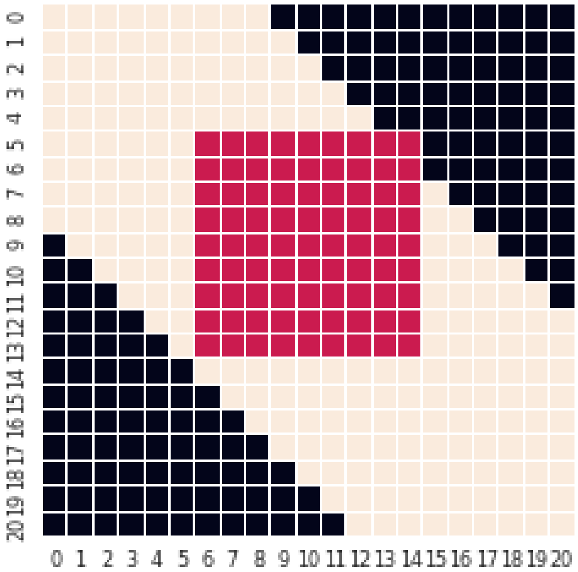


Figure 2.5: Illustration of the matrix completion method with $K = 21$, $q = 8$, $\delta = 0.5$. The band $B_\delta$ is colored in white and a possible $(q+1) \times (q+1)$ submatrix $A$ is colored in red.

By pluging-in the banded estimator $\tilde{R}_n$ defined in Section 2.4.1, the theorem naturally leads to the following estimator of the full matrix $R^K$:

**Definition 1.** Define the estimator $\hat{R}_n^K$ of $R^K$ as a minimum of

$$\underset{0 \preceq M \in R^{K \times K}}{\operatorname{argmin}} \left\{ \|\tilde{R}_n^K - P_\delta^K \circ M\|_{Frob}^2 + \tau_n \operatorname{rank}(M) \right\}, \tag{2.4}$$

where $\tau_n$ is a parameter that should converge to 0 as $n$ goes to infinity. From this estimator of the covariance matrix, one can estimate the covariance kernel by the step function

$$\hat{r}_n(s,t) = \hat{R}_n^K(i,j) \quad \text{if} \quad (s,t) \in I_{i,K} \times I_{j,K}. \tag{2.5}$$

□

We have the following consistency result.

**Theorem 11.** Assume $\mathbb{E}\|X_i\|_{\mathbb{L}^2}^4 < \infty$, the intervals $\{\mathcal{O}_1, ..., \mathcal{O}_n\}$ are i.i.d. independent of the $X_i$'s and that $\inf_{s,t \in [0,1]} \mathbb{P}[s,t \in \mathcal{O}_1] > 0$. Let $K^* = \lfloor (q+2)/\delta \rfloor + 1$ be the critical resolution. Then, if $\tau_n \to 0$, for almost any grid in $\mathcal{T}_K$,

$$\int\int_{[0,1]^2} \left( \hat{r}_n^K(x,y) - r(x,y) \right)^2 dxdy \le O_{\mathbb{P}}(1/n) + 4K^{-2} \sup_{x,y \in [0,1]} \|\nabla r(x,y)\|_2^2, \tag{2.6}$$

uniformly in $K$, for any refinement $K = m \times K^*$. □

### 2.4.3 Numerical implementation

The implementation of the estimator $R_n^K$ is explained in detail in [Descary and Panaretos, 2017, Section 5]. Roughly, it consists of two steps:

1. Find an approximate minimum $M_i$ of
$$\min_{0 \preceq M \in \mathbb{R}^{K \times K}} \|\tilde{R}_n^K - P_\delta^K \circ M\|_{Frob}^2 \text{ subject to } \operatorname{rank}(M) \le i$$
for $i \in \{1, ..., \lceil K\delta \rceil - 3\}$. This can be done by writing $M_i = \gamma\gamma^T$ with $\gamma \in \mathbb{R}^{K \times i}$ and optimizing on $\gamma$ using general purpose optimization algorithm, e.g., BFGS. Note that the objective function is not convex in $\gamma$. However, numerical experiments indicate that local minima are good enough.

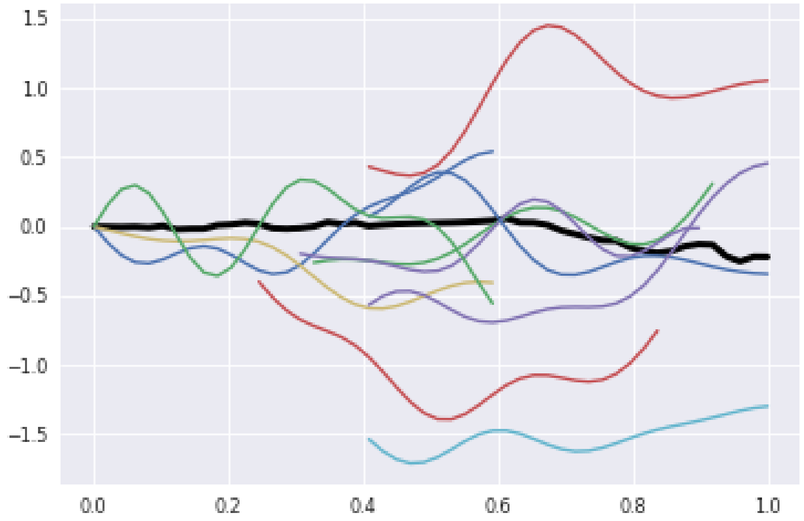
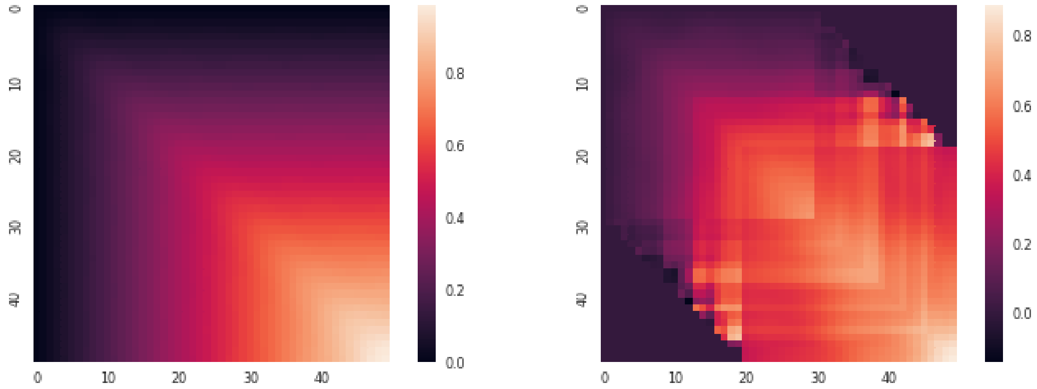


Figure 2.6: Plot of 10 fragments (out of the total 100), and the estimated mean in black.

2. Compute the objective value $f(i)$ of $M_i$ for all $i \in \{1, ..., \lceil K\delta \rceil - 3\}$ and take $\hat{R}_n^K = M_j$ for the rank $j$ after which the function $f$ does not decrease much more.

### 2.4.4 Example

We present here a numerical example of the covariance recovery method. Let $r'(s,t) = \min(s,t)$ be the covariance kernel of the brownian motion and $r(s,t)$ be its projection on its first 10 eigenfunctions. Then $r(s,t)$ is a rank 10 analytic kernel. We take $\mu \equiv 0$, $n = 100$, $K = 50$, $\delta = 0.6$ in the following. Example of fragments along with the estimated mean $\tilde{\mu}_n^K$ are shown in Figure 2.6. The true covariance matrix $R^K$ along with the banded estimator $\tilde{R}_n^K$ are shown in Figure 2.7.



(a) True covariance matrix

(b) Banded estimator

Figure 2.7

The relative error $\|R^K - \tilde{R}_n^K\| / \|R^K\|$ is 27.2% here. We now perform the first step of the implementation, as described in the previous section. A plot of the objective values for $i \in \{1, ..., 10\}$ is shown in Figure 2.8. From this plot, we decide to choose the rank of the estimator to be $i = 5$. The resulting estimator $\hat{R}_n^K = M_5$ is shown in Figure 2.9. The relative error with this estimate is $\|R^K - \hat{R}_n^K\| / \|R^K\| = 21.3\%$. By using the matrix completion method, we have thus achieved a gain of 28% in relative error compared to the naive estimator $\tilde{R}_n^K$ described in Section 2.2.

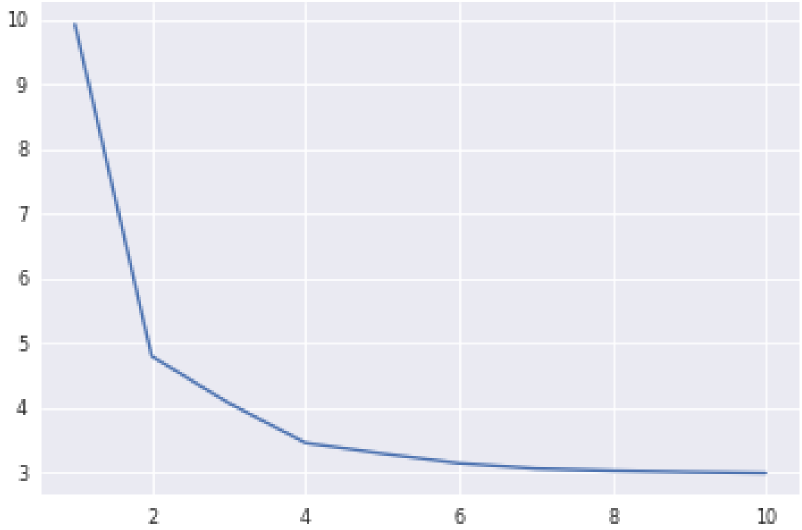


Figure 2.8: Objective value in function of the rank.

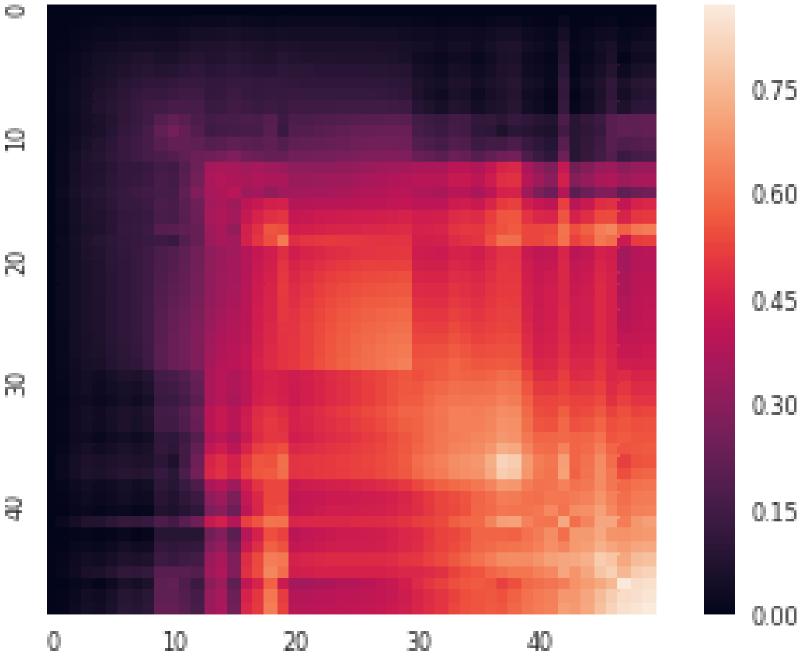


Figure 2.9: Estimated covariance matrix $\hat{R}_n^K$.

# Chapter 3

# Gaussian measures in Hilbert space

## 3.1 Gaussian measures in finite dimensions

Before exploring the construction and properties of infinite dimensional Gaussian measures, we start by going through the basic definitions and properties of finite dimensional Gaussian measures in ways that are easily generalizable to the infinite dimensional case.
We start with one-dimensional Hilbert spaces, i.e., the real line.

**Definition 2.** Let $a \in \mathbb{R}$ and $\sigma^2 > 0$ be parameters. A measure $\mu$ on the Borel $\sigma$-algebra $\mathcal{B}(\mathbb{R})$ is said to be a *Gaussian measure* with mean $a$ and variance $\sigma^2$ if

$$\mu(A) = \int_A \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-a)^2}{2}} dx, \quad \forall A \in \mathcal{B}(\mathbb{R}), \tag{3.1}$$

where the integral is with respect to the Lebesgue measure on $\mathbb{R}$. We denote this measure $N_{a,\sigma^2}$. □

We have the following elementary properties:

**Properties.**

1. The measure $N_{a,\sigma^2}$ indeed has mean $a$ and variance $\sigma^2$, i.e.,

$$\int_{\mathbb{R}} x N_{a,\sigma^2}(dx) = a \quad \text{and} \quad \int_{\mathbb{R}} (x-a)^2 N_{a,\sigma^2}(dx) = \sigma^2.$$

2. The measure $N_{a,\sigma^2}$ is equivalent to the Lebesgue measure $\lambda$ on $\mathcal{B}(\mathbb{R})$, and its Radon-Nykodim derivative is given by

$$\frac{dN_{a,\sigma^2}}{d\lambda}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(x-a)^2}.$$

3. The Fourier transform of $N_{a,\sigma^2}$ is

$$\widehat{N_{a,\sigma^2}}(y) = \int_{\mathbb{R}} e^{iyx} N_{a,\sigma^2}(dx) = e^{iay - \frac{1}{2}\sigma^2 y^2}$$

4. If $X$ is a random variable with distribution $N_{a,\sigma^2}$ and $\alpha, \beta \in \mathbb{R}$, then $\alpha X + \beta$ has distribution $N_{\alpha a+\beta, \alpha^2\sigma^2}$.

□

Note that one can also consider the limiting case $\sigma^2 = 0$, in which case the measure $N_{a,0}$ is the Dirac measure concentrated in $a$.
The definition of a Gaussian measure in arbitrary finite dimension relies on the following intuition. We would like these measures to "look like" one-dimensional Gaussian measure in every direction. Formally, this translates to:

**Definition 3.** For $f \in \mathbb{R}^d$, let $f^*$ denote the linear functional $f^*(x) = \langle f, x\rangle$. A measure $\mu$ on $\mathbb{R}^d$ is said *Gaussian* if for all $f \in \mathbb{R}^d$, the pushforward measure $f^*_\sharp \mu$ on $\mathbb{R}$ is Gaussian. □

For $\mu$ and $f$ as above, call $x_f$ the mean of $f^*_\sharp\mu$. Since the mean is linear, the function $f \mapsto x_f$ is linear, and thus, there exists some $a \in \mathbb{R}^d$ with $x_f = \langle a, f\rangle$. Then,

$$\int_{\mathbb{R}^d} x\mu(dx) = a.$$

Now, consider the function

$$\mathbb{R}^d \times \mathbb{R}^d \ni (f,g) \mapsto \beta(f,g) = \int_{\mathbb{R}^d} \langle f, x-a\rangle\langle g, x-a\rangle\mu(dx) \in \mathbb{R}.$$

It is easy to see that this gives a symmetric positive semidefinite bilinear form on $\mathbb{R}^d$ and thus has a representation $\beta(f,g) = \langle Qf, g\rangle$ for some p.s.d. matrix $Q$. Moreover, $\beta(f,f) = \langle Qf, f\rangle$ is the variance of $f^*_\sharp\mu$.
Finally, we note that the measure is uniquely determined by these two parameters $a \in \mathbb{R}^d$ and $Q \in \mathbb{R}^{d\times d}_{\succeq 0}$, since the Fourier transform of $\mu$ depends solely on them:

$$\widehat{\mu}(f) = \int_{\mathbb{R}^d} e^{i\langle f,x\rangle}\mu(dx) = \int_{\mathbb{R}} e^{iy} f^*_\sharp\mu(dy) = e^{i\langle f,a\rangle - \frac{1}{2}\langle Qf,f\rangle}. \tag{3.2}$$

Finally, we prove that this correspondence between Gaussian measure on $\mathbb{R}^d$ and pairs $(a, Q)$ is bijective.

**Proposition 5.** For any $a \in \mathbb{R}^d$ and any p.s.d. matrix $Q \in \mathbb{R}^{d\times d}_{\succeq 0}$, there exists a unique Gaussian measure $\mu$ on $\mathbb{R}^d$ having Fourier transform given by Equation 3.2. This measure is denoted $N_{a,Q}$.

*Proof.* The uniqueness is immediate. For the existence, consider the measure $\nu$ on $\mathcal{B}(\mathbb{R}^d)$ given by the product measure

$$\nu = \bigtimes_{i=1}^{d} N_{0,1}.$$

Let $L : \mathbb{R}^d \to \mathbb{R}^d$ be defined by $L(x) = a + Q^{1/2}x$ and define $\mu$ as $L_\sharp\nu$. Then,

$$\widehat{\mu}(f) = \int_{\mathbb{R}^d} e^{i\langle f,x\rangle}\mu(dx) = \int_{\mathbb{R}^d} e^{i\langle f, a+Q^{1/2}y\rangle}\nu(dy) = e^{i\langle f,a\rangle}\int_{\mathbb{R}^d} e^{i\langle Q^{1/2}f,y\rangle}\nu(dy) = e^{i\langle f,a\rangle - \frac{1}{2}\langle Qf,f\rangle},$$

where the last equality is given by Fubini's theorem and $\widehat{N_{0,1}}(t) = e^{-\frac{1}{2}t^2}$. □

We have the following properties:

**Properties.**

1. The measure $N_{a,Q}$ indeed has covariance matrix $Q$, i.e., if $X \sim N_{a,Q}$
$$Q_{ij} = cov\left(X_i, X_j\right).$$
2. The support of $N_{a,Q}$ is the image of $Q$.
3. If $Q$ is non-singular, the measure $N_{a,Q}$ is equivalent to the Lebesgue measure $\lambda^n$ on $\mathcal{B}(\mathbb{R}^n)$, and its Radon-Nykodim derivative is given by
$$\frac{dN_{a,Q}}{d\lambda^n}(x) = \frac{1}{\sqrt{|2\pi Q|}}e^{-\frac{1}{2}(x-a)^T Q^{-1}(x-a)}.$$
4. If $X$ is a random variable with distribution $N_{a,Q}$ and $\beta \in \mathbb{R}^m$, $A \in \mathbb{R}^{m\times n}$, then $AX + \beta$ has distribution $N_{Aa+\beta, AQA^T}$.

□

## 3.2 Gaussian measures in infinite dimensions

We now investigate the construction and properties of Gaussian measures in infinite dimensional space, with a focus on separable Hilbert spaces. Our exposition follows those of [Da Prato, 2006] and [Da Prato and Zabczyk, 2008].
The following results demonstrates the main reason why measure theory is much more complicated in infinite dimension: there is no equivalent of the Lebesgue measure in any infinite dimensional normed vector space.

**Proposition 6.** Let $(V, \|\cdot\|)$ be an infinite dimensional separable normed vector space. If $\mu$ is a measure on $V$ that is locally finite and translation invariant, then $\mu \equiv 0$.

*Proof.* Suppose such a measure $\mu$ exists. Let $B$ be an open ball with finite measure. By Riesz's lemma, $B$ contains the countable disjoint union of open balls $\{B_n\}_{n=1}^{\infty}$ each with same radius $\delta$. Since $\mu$ is translation invariant, these balls all have the same measure. Then we get the following inequality

$$\mu\left(\sqcup_{n=1}^{\infty} B_n\right) = \sum_{n=1}^{\infty} \mu(B_n) = \sum_{n=1}^{\infty} \mu(B_1) \leq \mu(B) < \infty.$$

Thus, $\mu(B_1) = 0$ and the same holds for all open balls of radius $\delta$. Since $V$ is separable, it can be covered by a countable union of such open balls. Thus $\mu(V) = 0$. □

Thus, in contrast with the finite dimensional case, there is no standard measure with respect to which one can define Radon-Nikodym densities. In some sense, this explains why most multivariate statistics tool do not apply in functional data analysis. The reason is that the concept of likelihood is ill-defined in infinite dimensions.
Luckily, our definition of multivariate Gaussian measures does not depend on densities with respect to the Lebesgue measure. It can therefore be generalized easily.

**Definition 4.** Let $(E, \|\cdot\|)$ be a Banach space. A measure $\mu$ on $\mathcal{B}(E)$ is said *Gaussian* if for all $f \in E^*$, the pushforward measure $f_\sharp \mu$ is Gaussian on $\mathcal{B}(\mathbb{R})$. □

For Hilbert spaces, the dual space is isomorphic to the original space. Thus the definition simplifies to:

**Definition 5.** For $f \in H$, let $f^*$ denote the linear functional $f^*(x) = \langle f, x\rangle$. A measure $\mu$ on $H$ is said *Gaussian* if for all $f \in H$, the pushforward measure $f^*_\sharp \mu$ on $\mathbb{R}$ is Gaussian. □

This last definition is exactly the same as that for Gaussian measure in finite dimensions. We will thus use the same strategy as in the finite dimensional case to find parameters characterizing a given measure. We will need the following technical lemma, given in [Da Prato and Zabczyk, 2008, Lemma 2.15].

**Lemma 1.** Let $\mu$ be a probability measure on a separable Hilbert space $H$, such that there exists $k \in \mathbb{N}$ with

$$\int_H |\langle h, x\rangle|^k \mu(dx) < \infty, \quad \forall h \in H.$$

Then, there exists some constant $c > 0$ such that for all $h_1, h_2, ..., h_k \in H$,

$$\left| \int_H \langle h_1, x\rangle \langle h_2, x\rangle ... \langle h_k, x\rangle \mu(dx) \right| \leq c\|h_1\|\|h_2\|...\|h_k\|.$$

*Proof.* Consider the family of closed sets

$$U_n = \{h \in H \mid \int_H |\langle h, x\rangle|^k \mu(dx) \leq n\}, \; n \in \mathbb{N}.$$

By assumtion, the union of the $U_n$ is all of $H$. By Baire's category theorem, this implies that one of them, say $U_{n_0}$ has non-empty interior. This implies that for some $z \in U_{n_0}$ and some $r > 0$, the open ball $B(z, r)$ is contained in $U_{n_0}$. In other words, for all $y \in H$ with $\|y\| < r$, it holds that

$$\int_H |\langle z + y, x\rangle|^k \mu(dx) \leq n_0.$$

Now, using the general inequality (provable using Holder's inequality)

$$|a-b|^k \leq 2^k |a|^k + 2^k |b|^k, \ a, b \in \mathbb{R},$$

we get

$$\int_H |\langle y, x\rangle|^k \mu(dx) \leq 2^k \int_H |\langle z+y, x\rangle|^k \mu(dx) + 2^k \int_H |\langle z, x\rangle|^k \mu(dx) \leq 2^{k+1} n_0.$$

For all $z \in H$, $\frac{r}{2\|z\|} z$ has norm less than $r$. Therefore, for all $z \in H$,

$$\int_H |\langle z, x\rangle|^k \mu(dx) \leq 2^{2k+1} r^{-k} n_0 \|z\|^k := c\|z\|^k.$$

Finally, using Holder's inequality, we get

$$\begin{aligned} \left| \int_H \langle h_1, x\rangle \langle h_2, x\rangle ... \langle h_k, x\rangle \mu(dx) \right| &\leq \int_H |\langle h_1, x\rangle \langle h_2, x\rangle ... \langle h_k, x\rangle| \, \mu(dx) \\ &\leq \left( \int_H |\langle h_1, x\rangle|^k \mu(dx) \right)^{1/k} ... \left( \int_H |\langle h_k, x\rangle|^k \mu(dx) \right)^{1/k} \\ &\leq c\|h_1\|\|h_2\|...\|h_k\|. \end{aligned}$$

□

Now consider the linear functional

$$H \ni f \mapsto \alpha(f) = \int_H \langle f, x\rangle \mu(dx) \in \mathbb{R}.$$

By the previous lemma, it is continuous. Thus, by Riesz's representation theorem, there exists some $a \in H$ with $\alpha(f) = \langle f, a\rangle$ for all $f \in H$. We call $a$ the *mean* of $\mu$. Next, consider the symmetric bilinear form

$$H \times H \ni (f, g) \mapsto \beta(f, g) = \int_H \langle f, x-a\rangle \langle g, x-a\rangle \mu(dx) \in \mathbb{R}.$$

By the lemma, it is bounded. Therefore, by Riesz's representation theorem for bilinear forms (see e.g. [Debnath and Mikusinski, 2005, Theorem 4.3.13]), there exists a bounded self-adjoint operator $Q$ on $H$ such that $\beta(f, g) = \langle Qf, g\rangle$ for all $f, g \in H$. We call $Q$ the *covariance operator* of $\mu$. From

$$\langle Qf, f\rangle = \int_H \langle f, x-a\rangle^2 \mu(dx) \geq 0,$$

we also get that $Q$ is non-negative. If $Q$ were compact, one could use the nice spectral theory of self-adjoint compact operators to study $Q$. In fact, the following stronger results holds:

**Theorem 12.** The operator $Q$ defined above is trace class.

*Proof.* See [Da Prato and Zabczyk, 2008, Prop. 2.16] □

**Corollary 1.** The operator $Q$ is Hilbert-Schmidt and compact. □

Moreover, one can show that

$$\mathrm{Tr}(Q) = \int_H \|x-a\|^2 \mu(dx).$$

This is analogous to the result in finite dimensions that the trace of the covariance matrix is equal to the total variance of the measure.
The Fourier transform of $\mu$ can easily be found using the results above:

$$\widehat{\mu}(f) = e^{i\langle h, a\rangle - \frac{1}{2}\langle Qf, f\rangle}.$$

We will need the following result, proved in [Da Prato and Zabczyk, 2008, Prop. 2.5]

**Proposition 7.** Let $H$ be a separable Hilbert space and $\mu, \nu$ be probability measures on $\mathcal{B}(H)$. If the Fourier transform $\hat{\mu}, \hat{\nu}$ are equal on $H$, then $\mu = \nu$ on $\mathcal{B}(H)$.

We are now ready to prove the analogue of Proposition 5 in infinite dimensions.

**Theorem 13.** Let $a$ be an element of $H$ and $Q$ a non-negative trace class operator on $H$. Then, there exists a unique Gaussian measure with mean $a$ and covariance operator $Q$.

*Proof.* Uniqueness follows from the previous proposition. For the existence, let $\{\lambda_i, \phi_i\}_{i=1}^{\infty}$ be a complete orthogonal system of eigenvectors of $Q$, i.e., $\{\phi_i\}_{i=1}^{\infty}$ forms a complete orthonormal basis of $H$ and

$$Q = \sum_{i=1}^{\infty} \lambda_i \; \phi_i \otimes \phi_i.$$

Let $\xi_i{}_{i=1}^{\infty}$ be a sequence of i.i.d. standard Gaussian random variables in $\mathbb{R}$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $X$ be the random element of $H$ defined by

$$X = a + \sum_{i=1}^{\infty} \sqrt{\lambda_i} \xi_i \phi_i.$$

The sum converges in $\mathbb{L}^2(\Omega, H)$ since $\mathrm{Tr}(Q) = \sum_{i=1}^{\infty} \lambda_i < \infty$ and

$$\mathbb{E} \| \sum_{i=m}^{\infty} \sqrt{\lambda_i} \xi_i \phi_i \|^2 = \mathbb{E} \left[ \sum_{i=m}^{\infty} \lambda_i \xi_i^2 \right] = \sum_{i=m}^{\infty} \lambda_i \overset{m \to \infty}{\longrightarrow} 0.$$

Thus $X$ is a well defined element of $\mathbb{L}^2(\Omega, H)$. Let $\mu$ be its associated measure on $H$. Its Fourier transform is given by

$$\begin{aligned}
\widehat{\mu}(f) &= \int_H e^{i\langle f,h \rangle} \mu(dh) && (3.3)\\
&= \int_\Omega e^{i\langle f,X \rangle} \mathbb{P}(dX) && (3.4)\\
&= \int_\Omega e^{i\langle f,a \rangle + i \sum_{i=1}^{\infty} \sqrt{\lambda_i} \xi_i \langle f, \phi_i \rangle} \mathbb{P}(dX) && (3.5)\\
&= e^{i\langle f,a \rangle} \prod_{i=1}^{\infty} \int_\Omega e^{i\sqrt{\lambda_i}\langle f,\phi_i \rangle \xi_i} \mathbb{P}(dX) && (3.6)\\
&= e^{i\langle f,a \rangle - \frac{1}{2} \sum_{i=1}^{\infty} \lambda_i \langle f, \phi_i \rangle^2} = e^{i\langle f,a \rangle - \frac{1}{2}\langle Qf,f \rangle} && (3.7)
\end{aligned}$$

Therefore, the law of $X$ is Gaussian with mean $a$ and covariance operator $Q$. □

We denote by $N_{a,Q}$ this measure. It has the following property.

**Proposition 8.** Let $H, K$ be separable Hilbert spaces and $B : H \to K$ a continuous linear map. Let $N_{a,Q}$ be a Gaussian measure on $H$. Then, the induced measure on $K$ is Gaussian with

$$B_\sharp N_{a,Q} = N_{Ba, BQB^*},$$

where $B^* : K \to H$ is the adjoint of $B$.

*Proof.* For any $k \in K$, we have

$$\begin{aligned}
\widehat{B_\sharp N_{a,Q}}(k) &= \int_K e^{i\langle k,y \rangle} B_\sharp N_{a,Q}(dy) = \int_H e^{i\langle k,Bx \rangle} N_{a,Q}(dx) && (3.8)\\
&= \int_H e^{i\langle B^*k,x \rangle} N_{a,Q}(dx) = \widehat{N_{a,Q}}(B^*k) && (3.9)\\
&= e^{i\langle B^*k,a \rangle - \frac{1}{2}\langle QB^*k, B^*k \rangle} = e^{i\langle k,Ba \rangle - \frac{1}{2}\langle BQB^*k,k \rangle} && (3.10)
\end{aligned}$$

□

## 3.3 The Feldman-Hajek theorem

Let $N_{a_1,Q_1}$ and $N_{a_2,Q_2}$ be Gaussian measures. We want to investigate the conditions one can impose on $a_i, Q_i$, $i = 1, 2$, that make the two measures either singular or equivalent.
We recall that two measures $\mu, \nu$ on $(\Omega, \mathcal{F})$ are said *singular* if there exists disjoint sets $A, B \in \mathcal{F}$ whose union is $\Omega$, with $\mu(B) = \nu(A) = 0$. We say that $\mu$ is absolutely continuous with respect to $\nu$, denoted $\mu << \nu$ if

$$\forall A \in \mathcal{F}, \quad \nu(A) = 0 \implies \mu(A) = 0.$$

Finally, we say that $\mu$ and $\nu$ are equivalent if we have both $\mu << \nu$ and $\nu << \mu$.
In finite dimensions, the problem is easy. Let $N_{a_1,Q_1}$ and $N_{a_2,Q_2}$ be Gaussian measures on $\mathbb{R}^d$, then

**Case 1** If $Q_1$ and $Q_2$ have the same range and $a_2 - a_1$ belongs to this range, then $N_{a_1,Q_1}$ and $N_{a_2,Q_2}$ are equivalent.

**Case 2** Otherwise, they are singular.

It is a remarkable result that roughly the same conclusion holds in infinite dimensions. Indeed, we will show next that two Gaussian measures on a separable Hilbert space are either equivalent or singular. Moreover, the conditions for them to be equivalent are quite restrictive, as opposed to the finite-dimensional case, where it suffices for the covariance matrices to be non-singular.
First, we need to study the notion of *Hellinger integral* and *Hellinger distance.* The latter defines a distance on probability measures giving information on their mutual singularity.

**Definition 6.** Let $\mu, \nu$ be probability densities on $(\Omega, \mathcal{F})$. The *Hellinger integral* $H(\mu, \nu)$ is defined by

$$H(\mu, \nu) = \int_\Omega \sqrt{\frac{d\mu}{d\lambda}\frac{d\nu}{d\lambda}}\, d\lambda,$$

where $\lambda$ is some probability measure on $(\Omega, \mathcal{F})$ with respect to which both $\mu$ and $\nu$ are absolutely continuous. The *Hellinger distance* $d(\mu, \nu)$ is defined by

$$d^2(\mu, \nu) = 1 - H(\mu, \nu) = \frac{1}{2}\int_\Omega \left(\sqrt{\frac{d\mu}{d\lambda}} - \sqrt{\frac{d\nu}{d\lambda}}\right)^2 d\lambda.$$

We have the following properties:

**Properties.**

1. A probability measure $\lambda$ with $\mu, \nu << \lambda$ always exists.
2. The definitions do not depend on the measure $\lambda$.
3. For any $\mu, \nu$, the Hellinger integral, and thus the Hellinger distance, are bounded between 0 and 1.
4. The measures $\mu$ and $\nu$ are singular if and only if $H(\mu, \nu) = 0$.

*Proof.* 1. Take $\lambda = \frac{1}{2}(\mu + \nu)$.

2. Let $\lambda'$ be another measure with $\mu, \nu << \lambda'$. Then,

$$\mu, \nu << \lambda, \lambda' << \chi := \frac{1}{2}(\lambda + \lambda')$$

and

$$\int_\Omega \sqrt{\frac{d\mu}{d\lambda}\frac{d\nu}{d\lambda}}\, d\lambda = \int_\Omega \sqrt{\frac{d\mu}{d\lambda}\frac{d\nu}{d\lambda}}\, \frac{d\lambda}{d\chi} d\chi = \int_\Omega \sqrt{\frac{d\mu}{d\lambda}\frac{d\lambda}{d\chi}\frac{d\nu}{d\lambda}\frac{d\lambda}{d\chi}}\, d\chi = \int_\Omega \sqrt{\frac{d\mu}{d\chi}\frac{d\nu}{d\chi}}\, d\chi$$

and the same chain of equalities holds with $\lambda'$. Thus,

$$\int_\Omega \sqrt{\frac{d\mu}{d\lambda}\frac{d\nu}{d\lambda}}\, d\lambda = \int_\Omega \sqrt{\frac{d\mu}{d\chi}\frac{d\nu}{d\chi}}\, d\chi = \int_\Omega \sqrt{\frac{d\mu}{d\lambda'}\frac{d\nu}{d\lambda'}}\, d\lambda'.$$

3. Using the Cauchy-Schwartz inequality, this follows from
$$0 \leq H(\mu,\nu) = \int_\Omega \sqrt{\frac{d\mu}{d\lambda}\frac{d\nu}{d\lambda}}\, d\lambda \leq \left(\int_\Omega \frac{d\mu}{d\lambda} d\lambda\right)^{1/2} \left(\int_\Omega \frac{d\nu}{d\lambda} d\lambda\right)^{1/2} = 1.$$

4. Suppose $\mu \perp \nu$. Then, there exists disjoint sets $A, B \in \mathcal{F}$ such that $\Omega = A \cup B$ with $\mu(A) = 1 = \nu(B)$. Then
$$H(\mu,\nu) = \int_A \sqrt{\frac{d\mu}{d\lambda}\frac{d\nu}{d\lambda}}\, d\lambda + \int_B \sqrt{\frac{d\mu}{d\lambda}\frac{d\nu}{d\lambda}}\, d\lambda = 0.$$
Conversely, suppose $H(\mu,\nu) = 0$ and let $f = \frac{d\mu}{d\lambda}, g = \frac{d\nu}{d\lambda}$. Then, we have
$$\int_\Omega \sqrt{fg}\, d\lambda = 0.$$
Therefore, $\lambda(\{fg = 0\}) = 1$, and since $\mu, \nu << \lambda$, it follows that
$$\mu(\{fg = 0\}) = 1 = \nu(\{fg = 0\}).$$
Now, let $A = \{g = 0\}$,$B = \{f = 0\}$ and $C = \{fg = 0\}$ so that $\mu(B) = \int_B f d\lambda = 0$ and $\nu(A) = \int_A g d\lambda = 0$. Since $C = A \cup B$, defining $\tilde{A} = (C \backslash B) \sqcup (\Omega \backslash C)$ yields $\Omega = \tilde{A} \sqcup B$ and
$$\mu(\tilde{A}) = \mu(C) - \mu(B) + (1 - \mu(C)) = 1 \text{ and } \nu(B) = \nu(C) - \nu(A \backslash B) = 1.$$
Thereby, $\mu$ and $\nu$ are singular.

□

The reason we want to study the Hellinger integral in the context of measures on Hilbert spaces is that it behaves well with respect to infinite products. Let us start with finite products.

**Proposition 9.** Let $\mu_1, \mu_2, \nu_1, \nu_2$ be probability measures on $(\Omega, \mathcal{F})$. Consider the product measures $\mu_1 \times \mu_2$ and $\nu_1 \times \nu_2$ on the product space $(\Omega \times \Omega, \mathcal{F} \otimes \mathcal{F})$. Then,
$$H(\mu_1 \times \mu_2, \nu_1 \times \nu_2) = H(\mu_1, \nu_1) H(\mu_2, \nu_2).$$

*Proof.* Let $\lambda_1, \lambda_2$ be measures such that $\mu_1, \nu_1 << \lambda_1$, and $\mu_2, \nu_2 << \lambda_2$. Then, for $A_1, A_2 \in \mathcal{F}$,
$$\begin{aligned}(\lambda_1 \times \lambda_2)(A_1 \times A_2) = \lambda_1(A_1)\lambda_2(A_2) = 0 \quad &\Longrightarrow \quad \mu_1(A_1)\mu_2(A_2) = 0 \text{ and } \nu_1(A_1)\nu_2(A_2) = 0 \\ &\Longrightarrow \quad (\mu_1 \times \mu_2)(A_1 \times A_2), (\nu_1 \times \nu_2)(A_1 \times A_2) = 0.\end{aligned}$$
Since rectangular sets like $A_1 \times A_2$ generate $\mathcal{F} \otimes \mathcal{F}$, we get that $(\mu_1 \times \mu_2), (\nu_1 \times \nu_2) << (\lambda_1 \times \lambda_2)$. Now, define $f_i = \frac{d\mu_i}{d\lambda_i}$, $g_i = \frac{d\nu_i}{d\lambda_i}$. Then, by Fubini's theorem,
$$\int_{A_1 \times A_2} d(\mu_1 \times \mu_2) = \int_{A_1} d\mu_1 \int_{A_2} d\mu_2 = \int_{A_1} f_1 d\lambda_1 \int_{A_2} d\lambda_2 = \int_{A_1 \times A_2} f_1 f_2 d(\lambda_1 \times \lambda_2).$$
Thus, $f_1 f_2 = \frac{d(\mu_1 \times \mu_2)}{d(\lambda_1 \times \lambda_2)}$ and similarly, $g_1 g_2 = \frac{d(\nu_1 \times \nu_2)}{d(\lambda_1 \times \lambda_2)}$. Finally, we conclude with
$$H(\mu_1 \times \mu_2, \nu_1 \times \nu_2) = \int_{\Omega \times \Omega} \sqrt{f_1 f_2 g_1 g_2} d(\lambda_1 \times \lambda_2) = \int_\Omega \sqrt{f_1 g_1} d\lambda_1 \int_\Omega \sqrt{f_2 g_2} d\lambda_2 = H(\mu_1, \nu_1) H(\mu_2, \nu_2).$$

□

**Corollary 2.** Let $\{\mu_i\}_{i=1}^\infty, \{\nu_i\}_{i=1}^\infty$ be measures on $(\Omega, \mathcal{F})$. Then, for all $N \in \mathbb{N}$,
$$H\left(\bigtimes_{i=1}^N \mu_i, \bigtimes_{i=1}^N \nu_i\right) = \prod_{i=1}^N H(\mu_i, \nu_i),$$
and consequently
$$H\left(\bigtimes_{i=1}^\infty \mu_i, \bigtimes_{i=1}^\infty \nu_i\right) = \prod_{i=1}^\infty H(\mu_i, \nu_i).$$

From Property 4 above, we see that $H(\mu, \nu) > 0$ is a necessary condition for $\mu << \nu$ or $\nu << \mu$. It turns out that it is also sufficient for infinite product measures.

**Theorem 14** (Kakutani)**.** Let $\{\mu_i\}_{i=1}^{\infty}, \{\nu_i\}_{i=1}^{\infty}$ be measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, with $\mu_i << \nu_i$ for all $i$. Let $\mu = \bigtimes_{i=1}^{\infty} \mu_i$ and $\nu = \bigtimes_{i=1}^{\infty} \nu_i$ be the product measures on $(\mathbb{R}^\infty, \mathcal{B}(\mathbb{R}^\infty))$. If

$$H(\mu, \nu) = \prod_{i=1}^{\infty} H(\mu_i, \nu_i) > 0,$$

then $\mu << \nu$. Moreover, in that case the sequence

$$f_N(x) = \prod_{i=1}^{N} \frac{d\mu_i}{d\nu_i}(x_i), \; x \in \mathbb{R}^\infty$$

converges to $\frac{d\mu}{d\nu}$ in $\mathbb{L}^1(\mathbb{R}^\infty, \mathcal{B}(\mathbb{R}^\infty), \nu)$.

*Proof.* To show that $\{f_N\}$ converges in $\mathbb{L}^1$, it is sufficient to show that $\{\sqrt{f_N}\}$ converges in $\mathbb{L}^2$. Now, for $N, m \in \mathbb{N}$, we have

$$\begin{aligned} \int_{\mathbb{R}^\infty} |\sqrt{f_{N+m}} - \sqrt{f_N}|^2 d\nu &= \int_{\mathbb{R}^\infty} \prod_{i=1}^{N} \frac{d\mu_i}{d\nu_i} \left| \sqrt{\prod_{i=N+1}^{N+m} \frac{d\mu_i}{d\nu_i}} - 1 \right|^2 d\nu \\ &= \int_{\mathbb{R}^\infty} \left| \sqrt{\prod_{i=N+1}^{N+m} \frac{d\mu_i}{d\nu_i}} - 1 \right|^2 d\nu \\ &= 2 - 2 \prod_{i=N+1}^{N+m} \int_{\mathbb{R}^\infty} \sqrt{\frac{d\mu_i}{d\nu_i}} d\nu = 2 - \prod_{i=N+1}^{N+m} H(\mu_i, \nu_i). \end{aligned}$$

Now, by assumption, $\prod_{i=1}^{\infty} H(\mu_i, \nu_i) > 0$. Equivalently, $-\sum_{i=1}^{\infty} \log H(\mu_i, \nu_i) < \infty$. Therefore, for all $\epsilon > 0$, there exists some $N$ sufficiently large with $-\sum_{i=N}^{\infty} \log H(\mu_i, \nu_i) < \epsilon$. It follows that

$$\int_{\mathbb{R}^\infty} |\sqrt{f_{N+m}} - \sqrt{f_N}|^2 d\nu = 2 - 2 \prod_{i=N+1}^{N+m} H(\mu_i, \nu_i) \leq 2 - 2 \prod_{i=N+1}^{\infty} H(\mu_i, \nu_i) \leq 2 - 2e^{-\epsilon} \overset{\epsilon \to 0}{\longrightarrow} 0,$$

and thus $\{f_N\}$ converges in $\mathbb{L}^1$ to some $f$. We now show that $f$ is the Radon-Nikodym derivative of $\mu$ with respect to $\nu$, which in turn implies that $\mu << \nu$.
Let $\alpha(x)$ be a measurable function on $(\mathbb{R}^\infty, \mathcal{B}(R^\infty))$ depending only on $x_1, ..., x_k$ for some $k \in \mathbb{N}$. Then, for $n \geq k$,

$$\int_{\mathbb{R}^\infty} \alpha(x) d\mu = \int_{\mathbb{R}^\infty} \alpha(x) f_n(x) d\nu \overset{n \to \infty}{\longrightarrow} \int_{\mathbb{R}^\infty} \alpha(x) f(x) d\nu.$$

Since cylindrical sets generate $\mathcal{B}(\mathbb{R}^\infty)$, we get for all measurable $\alpha(x)$ the identity

$$\int_{\mathbb{R}^\infty} \alpha(x) d\mu = \int_{\mathbb{R}^\infty} \alpha(x) f(x) d\nu$$

which proves that $f$ indeed is the Radon-Nikodym derivative of $\mu$ with respect to $\nu$. □

**Corollary 3.** Let $\{\mu_i\}_{i=1}^{\infty}, \{\nu_i\}_{i=1}^{\infty}, \mu, \nu$ be as in the previous theorem. Suppose further that $\nu_i << \mu_i$ for all $i$, so that the measures $\mu_i, \nu_i$ are equivalent for all $i$. Then, depending on whether $H(\mu, \nu)$ is non-zero, the measures $\mu$ and $\nu$ are either equivalent or singular.

We now want to apply these result in the Gaussian case. For that purpose, we will need the following lemma.

**Lemma 2.** Let $N_{a,Q}$ be a Gaussian measure on a separable Hilbert space $H$. Let $\{\lambda_i, \phi_i\}_{i=1}^{\infty}$ be a complete orthogonal system of eigenvectors of $Q$. Consider the isomorphism

$$\Xi : \ell^2(\mathbb{R}) \to H : \{x_k\} \mapsto \sum_{k=1}^{\infty} x_k \phi_k.$$

Let $\nu = \bigtimes_{k=1}^{\infty} N_{a_k, \lambda_k}$ be the product measure on $\mathbb{R}^\infty$, where $a_k = \langle a, \phi_k \rangle$. Then $\nu$ is concentrated on $\ell^2(\mathbb{R})$ and $\Xi_\sharp \nu = \mu$.

*Proof.* We first show that $\nu$ is concentrated on $\ell^2(\mathbb{R})$, i.e., that $\nu\left(\{x \in \mathbb{R}^\infty \mid \|x\|_{\ell^2} < \infty\}\right) = 1$. To see that, we use the equalities

$$\int_{\mathbb{R}^\infty} \|x\|_{\ell^2}^2 \nu(dx) = \int_{\mathbb{R}^\infty} \sum_{k=1}^{\infty} x_k^2 \nu(dx) = \sum_{k=1}^{\infty} \int_{\mathbb{R}} x_k^2 N_{a_k,\lambda_k}(dx_k) = \sum_{k=1}^{\infty} (\lambda_k + a_k^2) = \operatorname{Tr}(Q) + \|a\|^2 < \infty.$$

Therefore, $\|x\|_{\ell^2}^2$ is finite $\nu$-almost everywhere. By essentially the same proof as in Theorem 13, one can show that $\nu$ is Gaussian, with mean $\tilde{a} = \{a_k\}$ and covariance operator

$$\tilde{Q} = \sum_{k=1}^{\infty} \lambda_k e_k \otimes e_k.$$

By Proposition 8, $\Xi_\sharp \nu$ is Gaussian with mean $\Xi(\tilde{a}) = a$ and covariance operator

$$\Xi \tilde{Q} \Xi^*(x) = \Xi \tilde{Q} \{\langle x, \phi_k \rangle\} = \Xi \{\lambda_k \langle x, \phi_k \rangle\} = \sum_{k=1}^{\infty} \lambda_k \langle x, \phi_k \rangle \phi_k = Qx.$$

Thus, $\Xi_\sharp \nu = \mu$. □

As a first application of Theorem 14, we look at Gaussian measures having the same covariance. As we saw earlier, this case is very simple in finite dimensions. There, two Gaussian measures with the same covariance matrix are equivalent if and only if the difference of their mean is in the range of the (square root of the) covariance matrix. The next theorem shows that the exact same conclusion is true in a separable Hilbert space $H$.

**Theorem 15.** Let $N_{a_1,Q}$ and $N_{a_2,Q}$ be Gaussian measures on $H$. Then,

1. They are either equivalent or singular.
2. They are equivalent if and only if $a_1 - a_2 \in Q^{1/2}(H)$. In that case the Radon-Nikodym derivative is given by

$$\frac{dN_{a_1,Q}}{dN_{a_2,Q}}(x) = \exp\left[\langle Q^{-1/2}(a_1 - a_2), Q^{-1/2}(x - a_2)\rangle - \frac{1}{2}\|Q^{-1/2}(a_1 - a_2)\|^2\right]$$

   for $N_{a_2,Q}$-almost all $x \in H$.

*Proof.* By restricting the measures to their support, we can assume that $Q$ is non-singular, i.e., $\lambda_k \neq 0$ for all $k$, where $\{\lambda_i, \phi_i\}_{i=1}^{\infty}$ is a complete orthogonal system of eigenvectors of $Q$. Then, 1. is a direct consequence of Corollary 3, since non-degenerate Gaussian measures on $\mathbb{R}$ are always equivalent.

For 2., let us clarify the notations first. For $Q$ non-singular, $Q^{-1/2}$ denotes the pseudo-inverse of $Q^{1/2}$. The random variable $\langle Q^{-1/2}(a_1 - a_2), Q^{-1/2}(x - a_2)\rangle$ makes sense when defined as follows

$$\langle Q^{-1/2}(a_1 - a_2), Q^{-1/2}(x - a_2)\rangle = \sum_{k=1}^{\infty} \frac{\langle a_1 - a_2, \phi_k \rangle \langle x - a_2, \phi_k \rangle}{\lambda_k}.$$

Indeed, the sum converges in $\mathbb{L}^2(H, \mathcal{B}(H), N_{a_2,Q})$. Indeed, we have

$$\begin{aligned}
\int_H \sum_{k=1}^{\infty} \left( \frac{\langle a_1 - a_2, \phi_k \rangle \langle x - a_2, \phi_k \rangle}{\lambda_k} \right)^2 N_{a_2,Q}(dx) &= \sum_{k=1}^{\infty} \frac{\langle a_1 - a_2, \phi_k \rangle^2}{\lambda_k^2} \int_H \langle x - a_2, \phi_k \rangle^2 N_{a_2,Q}(dx) \\
&= \sum_{k=1}^{\infty} \frac{\langle a_1 - a_2, \phi_k \rangle^2}{\lambda_k^2} \langle Q\phi_k, \phi_k \rangle \\
&= \sum_{k=1}^{\infty} \frac{\langle a_1 - a_2, \phi_k \rangle^2}{\lambda_k} = \|Q^{-1/2}(a_1 - a_2)\|^2.
\end{aligned}$$

By the previous lemma, we can identify the two Gaussian measures on $H$ with the measures on $\mathbb{R}^\infty$

$$\mu = \bigtimes_{k=1}^{\infty} N_{a_{1k},\lambda_k} \quad \text{and} \quad \nu = \bigtimes_{k=1}^{\infty} N_{a_{2k},\lambda_k},$$

where $a_{ik} = \langle a_i, \phi_k \rangle$, $i = 1, 2$. Now, for a given coordinate $k$, the Radon-Nikodym derivative $\frac{dN_{a_{1k},\lambda_k}}{dN_{a_{2k},\lambda_k}}$ are simply given by the ratio of their densities with respect to the Lebesgue measure:

$$\begin{aligned}
\frac{dN_{a_{1k},\lambda_k}}{dN_{a_{2k},\lambda_k}}(x_k) &= \exp\left[ \frac{1}{2\lambda_k} \left( (x_k - a_{2k})^2 - (x_k - a_{1k})^2 \right) \right] \\
&= \exp\left[ \frac{1}{2\lambda_k} \left( 2(a_{1k} - a_{2k})(x_k - a_{2k}) - (a_{2k} - a_{1k})^2 \right) \right].
\end{aligned}$$

By simple integration, we get

$$H(N_{a_{1k},\lambda_k}, N_{a_{2k},\lambda_k}) = \int_{\mathbb{R}} \sqrt{\frac{dN_{a_{1k},\lambda_k}}{dN_{a_{2k},\lambda_k}}(x)}\; N_{a_2,Q}(dx) = \exp\left[ -\frac{(a_{1k} - a_{2k})^2}{8\lambda_k} \right].$$

Then,

$$H(\mu, \nu) = \prod_{k=1}^{\infty} H(N_{a_{1k},\lambda_k}, N_{a_{2k},\lambda_k}) = \exp\left[ -\frac{1}{8} \sum_{k=1}^{\infty} \frac{(a_{1k} - a_{2k})^2}{\lambda_k} \right] = \exp\left[ -\frac{1}{8} \|Q^{-1/2}(a_1 - a_2)\|^2 \right].$$

So $\mu$ and $\nu$ are equivalent if and only if $H(\mu, \nu) > 0$ if and only if $a_1 - a_2 \in Q^{1/2}(H)$. Finally, we compute the Radon-Nikodym derivative $\frac{d\mu}{d\nu}$ using Kakutani's theorem:

$$\begin{aligned}
\frac{d\mu}{d\nu}(x) &= \prod_{k=1}^{\infty} \frac{dN_{a_{1k},\lambda_k}}{dN_{a_{2k},\lambda_k}}(x_k) = \prod_{k=1}^{\infty} \exp\left[ \frac{1}{2\lambda_k} \left( 2(a_{1k} - a_{2k})(x_k - a_{2k}) - (a_{2k} - a_{1k})^2 \right) \right] \\
&= \exp\left[ \sum_{k=1}^{\infty} \frac{(x_k - a_{2k})(a_{1k} - a_{2k})}{\lambda_k} - \frac{1}{2} \sum_{k=1}^{\infty} \frac{(a_{2k} - a_{1k})^2}{\lambda_k} \right] \\
&= \exp\left[ \langle Q^{-1/2}(a_1 - a_2), Q^{-1/2}(x - a_2) \rangle - \frac{1}{2} \|Q^{-1/2}(a_1 - a_2)\|^2 \right].
\end{aligned}$$

□

We finally state the main theorem of this chapter, the Feldman-Hajek theorem. It gives a complete characterization of whether Gaussian measures are singular, equivalent or neither. The proof of this theorem is somewhat involved and not so enlightening. Therefore, we do not write it here. It can be found in [Da Prato, 2006, Theorem 2.25].

**Theorem 16** (Feldman-Hajek)**.** Let $N_{a_1,Q_1}$ and $N_{a_2,Q_2}$ be Gaussian measures on a Hilbert space $H$. Then,

1. They are either equivalent or singular.

2. They are equivalent if and only if the following three conditions hold:

(a) $Q_1^{1/2}(H) = Q_2^{1/2}(H)$

(b) $a_1 - a_2 \in Q_i^{1/2}(H)$.

(c) The operator $(Q_1^{-1/2}Q_2^{1/2})(Q_1^{-1/2}Q_2^{1/2})^* - I$ is Hilbert-Schmidt on $\overline{Q_i^{1/2}(H)}$.

The conditions (a) and (b) above are not very interesting since they are also required in the finite dimensional case. Condition (c) however is much more restrictive and proper the infinite dimensional case. It essentially says that the isomorphism

$$Q_1^{-1/2}Q_2^{1/2} : \overline{Q_i^{1/2}(H)} \to \overline{Q_i^{1/2}(H)}$$

must not be too far from the identity in order for the measure to be equivalent. In other words, it requires that the operators $Q_1^{1/2}$ and $Q_2^{1/2}$ should not be too different in an infinite number of dimensions. To make this concrete, consider the following example.
Let $Q_1$ be any trace class operator on $H$, and let $Q_2 = \alpha Q_1$ for some scalar $0 < \alpha \neq 1$. Then, $Q_1$ and $Q_2$ differ by the same constant amount in every dimension. Thus, we have

$$(Q_1^{-1/2}Q_2^{1/2})(Q_1^{-1/2}Q_2^{1/2})^* - I = (\sqrt{\alpha}I)(\sqrt{\alpha}I)^* - I = (\alpha - 1)I,$$

which is obviously not Hilbert-Schmidt.
This last example emphasizes the difference between the finite and the infinite dimensional case. In the former, measures $N_{0,\Sigma}, N_{0,\alpha\Sigma}$ are always equivalent and can actually become arbitrarily close as $\alpha \to 1$. In infinite dimensions, the measures are equivalent if and only if $\alpha = 1$.

# Chapter 4

# Classification of functional data

## 4.1 Introduction

In the last section, we have shown that Gaussian measures can often be singular. In the context of classification, this translates to the result that one can perform perfect classification on sample from these measures. Concretely, suppose we have probability measures $\mu, \nu$ on $(\Omega, \mathcal{F})$ that are singular. Let $A, B \in \mathcal{F}$ with $\Omega = A \sqcup B$ and

$$\mu(A) = 1,\ \nu(B) = 1,\ \mu(B) = 0,\ \nu(A) = 0.$$

Suppose now that we want to classify samples $\omega \in \Omega$ as coming from measure $\mu$ or $\nu$. We can simply use the classifier

$$C(\omega) = \begin{cases} \mu & \text{if } \omega \in A \\ \nu & \text{if } \omega \in B. \end{cases}$$

This rule will perform perfect classification, meaning that if $\omega$ is indeed sampled from either $\mu$ or $\nu$, the classifier will output the true measure with probability one.
The problem is that the sets $A$ and $B$ are not known in general, so this classifier cannot be used in practice. However, we will show that in the case of singular Gaussian measures on Hilbert spaces, one can construct a classifier that will achieve this perfect classification.

Suppose now that we have two equivalent measures $\mu_0$ and $\mu_1$ on $(\Omega, \mathcal{F})$. Define the measure $\kappa$ on $\Omega \times \{0, 1\}$ by

$$\kappa(A) = \kappa\left(A_0 \sqcup A_1\right) = p \cdot \mu_0(A_0) + (1-p) \cdot \mu_1(A_1),\ A_i = A \cap (\Omega \times \{i\}),\ p \in [0, 1].$$

Concretely, $\kappa$ is the measure that is $\mu_0$ with probability $p$ and $\mu_1$ with probability $1 - p$. The classification task is then to estimate $y$ from $\omega$ for $(\omega, y) \sim \kappa$. It is well-known that the classifier achieving the lowest misclassification probability is the Bayes rule $\mathbf{1}[C^*(\omega) > 1]$ with

$$C^*(\omega) = \frac{\mathbb{P}_{(X,Y)\sim\kappa}[Y = 1 \mid X = \omega]}{\mathbb{P}_{(X,Y)\sim\kappa}[Y = 0 \mid X = \omega]} = \frac{1-p}{p}\frac{d\mu_1}{d\mu_0}(\omega), \tag{4.1}$$

see [Devroye et al., 2013, Theorem 2.1]. The problem this time is that the Radon-Nikodym derivative $\frac{d\mu_1}{d\mu_0}$ is not known in general. However, using the results shown in the previous chapter, for Gaussian measures the derivative can be explicitly written in terms of the parameters of the measures. We will see in this chapter how to use these expressions to design classifiers.

We will only consider the *linear* and *quadratic discriminant analysis* methods for binary classification. These are well-know classifier in the finite dimensional case that can easily be extended to infinite dimensions. We note that there are many other classifiers on functional data that have been developed, e.g., nearest neighbors, kernel and depth classifiers. An overview of these techniques can be found in [Ferraty and Romain, 2010, Chapter 10].

## 4.2 Linear and quadratic discriminant analysis for functional data

### 4.2.1 Finite dimensional case

Linear and quadratic discriminant analysis (LDA and QDA) in $\mathbb{R}^d$ are two of the most basic classification methods that can be traced back to the work of Fischer in the thirties. Despite their simplicity, their performances are often close to those of more complex classifiers. The idea behind them is essentially that of maximum likelihood estimation.

Concretely, we suppose that we have two populations on $\mathbb{R}^d$, the first one $\Pi_0$ following a Gaussian distribution with mean $\mu_0$ and covariance $\Sigma_0$ and the second $\Pi_1$ following a Gaussian distribution with mean $\mu_1$ and covariance $\Sigma_1$. Denote by $f_i(x)$ the probability density function of $\Pi_i$. Then, the QDA classifier classifies an observation $x \in \mathbb{R}^d$ as coming from population $\Pi_1$ if and only if

$$1 < \frac{(1-p)\cdot f_1(x)}{p \cdot f_0(x)} = \frac{|\Sigma_0|^{1/2}}{|\Sigma_1|^{1/2}} \exp\left[-\frac{1}{2}\left\{(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) - (x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0)\right\}\right],$$

or equivalently if

$$2\log\left(\frac{p}{1-p}\right) < \log|\Sigma_0| - \log|\Sigma_1| - (x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) + (x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0), \tag{4.2}$$

where $p = \mathbb{P}[x \in \Pi_0]$ is a constant often taken to be $1/2$ without any a priori knowledge.

Linear discriminant analysis is used in the same framework where we further assume that $\Sigma_0 = \Sigma_1 := \Sigma$. In that case, the classification rule simplifies to

$$2\log\left(\frac{p}{1-p}\right) < 2(x-\mu_1)^T\Sigma^{-1}(\mu_1-\mu_0) + (\mu_1-\mu_0)^T\Sigma^{-1}(\mu_1-\mu_0)$$

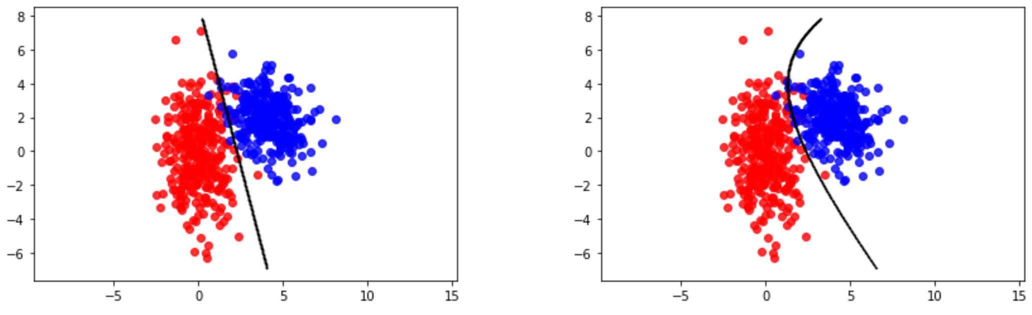
Defining

$$w = 2\Sigma^{-1}(\mu_1-\mu_0) \quad \text{and} \quad c = 2\log\left(\frac{p}{1-p}\right) - (\mu_1-\mu_0)^T\Sigma^{-1}(\mu_1-\mu_0) + w^T\mu_1,$$

the LDA rule classifies $x$ to population $\Pi_1$ if and only if

$$w^Tx > c.$$

Thus, it is a *linear classifier*, meaning that it projects the data in the direction given by $w$ and then classifies using only these projections. This yields a simple and fast algorithm, yet powerful. An example is shown in Figure 4.1.



(a) LDA (b) QDA

Figure 4.1: Example of LDA and QDA where the decision boundary is shown in black.

We note that both LDA and QDA are equivalent to the Bayes rule 4.1 when the assumptions of normality and homoscedasticity (for the LDA) are satisfied. Indeed, the likelihood ratio $f_1(x)/f_0(x)$

is equal to the Radon-Nikodym derivative of the two induced Gaussian measures in this case. Thus, under these assumptions, LDA and QDA are optimal classifiers.

In practice, none of the parameters $\mu_i, \Sigma_i$ are known. They can be estimated in the usual way by the empirical estimators $\hat{\mu}_i, \hat{\Sigma}_i$. These can then be plugged into either the LDA or QDA rules. Since these estimators converge to their associated real values, LDA and QDA using these estimates are asymptotically optimal.

Without the Gaussian assumption, LDA can still be seen as the optimal linear classifier in the following sense. Suppose here that $p = 1/2$, then the direction $v$ with $\|v\| = 1$ maximizing the variance of $v^T X$ is $w$ given by the LDA rule, where $X \sim \frac{1}{2}\Pi_0 + \frac{1}{2}\Pi_1$. Indeed, we have

$$\begin{aligned}
\mathrm{Var}[v^T X] &= \mathbb{E}[\mathrm{Var}[v^T X \mid \Pi]] + \mathrm{Var}[\mathbb{E}[v^T X \mid \Pi]] = v^T \Sigma v + \left(v^T \left(\frac{\mu_0 - \mu_1}{2}\right)\right)^2 \\
&= v^T \left(\Sigma + \left(\frac{\mu_0 - \mu_1}{2}\right)\left(\frac{\mu_0 - \mu_1}{2}\right)^T\right) v \\
&= (\Sigma^{1/2} v)^T \left(I + \Sigma^{-1/2}\left(\frac{\mu_0 - \mu_1}{2}\right)\left(\frac{\mu_0 - \mu_1}{2}\right)^T \Sigma^{-1/2}\right)(\Sigma^{1/2} v)
\end{aligned}$$

The maximum of this last expression, conditioned on $\|v\| = 1$, is attained when $\Sigma^{1/2} v$ is the first eigenvector of the matrix in the middle. This eigenvector is easily seen to be $\Sigma^{-1/2}(\frac{\mu_0 - \mu_1}{2})$, giving

$$v = \Sigma^{-1}\left(\frac{\mu_0 - \mu_1}{2}\right) \propto w.$$

Intuitively, the direction maximizing $\mathrm{Var}[v^T X]$ is the one on which the projections of the two populations are the most well-separated.

### 4.2.2 Infinite dimensional case: Equivalent Gaussian measures

In the infinite dimensional case, densities do not exist (as there is no Lebesgue measure, see Prop. 6). However, for separable Hilbert spaces and Gaussian random element on them, the theory developed in the last chapter allows us to construct powerful classifiers.

For example, suppose we have to populations on a separable Hilbert space $H$, $\Pi_0, \Pi_1$, that follow have Gaussian distributions with mean $\mu_0, \mu_1$ and same covariance operator $Q$. Suppose further that $\mu_1 - \mu_0 \in Q^{1/2}(H)$. Then, using Theorem 15, one can construct the Bayes classifier $\mathbf{1}[C^*(x) > 1]$ with (assume $p = 1/2$)

$$C^*(x) = \frac{dN_{\mu_1,Q}}{dN_{\mu_0,Q}}(x) = \exp\left[\langle Q^{-1/2}(\mu_1 - \mu_0), Q^{-1/2}(x - \mu_0)\rangle - \frac{1}{2}\|Q^{-1/2}(\mu_1 - \mu_0)\|^2\right].$$

Equivalently, it classifies to $\Pi_1$ if

$$\begin{aligned}
& \frac{1}{2}\|Q^{-1/2}(\mu_1 - \mu_0)\|^2 < \langle Q^{-1/2}(\mu_1 - \mu_0), Q^{-1/2}(x - \mu_0)\rangle \\
\iff & \frac{1}{2}\|Q^{-1/2}(\mu_1 - \mu_0)\|^2 + \langle Q^{-1}(\mu_1 - \mu_0), \mu_0\rangle < \langle Q^{-1}(\mu_1 - \mu_0), x\rangle.
\end{aligned}$$

This rule is, mutatis mutandis, the same as the LDA rule in finite dimension. We therefore also call it the LDA rule. As in finite dimension, the classification depends only on the projection of $x$ on the 1-dimensional subspace spanned by $Q^{-1}(\mu_1 - \mu_0)$. Therefore, the LDA rule reduces the classification of infinite dimensional objects, like functions, to a one dimensional problem.

As in the finite dimensional case, such a nice linear behavior is not present when the covariance operator of the two populations is not the same. In that case, one can in general still write down the Radon-Nikodym derivative, see [Bogachev, 2015, Coro. 6.4.11], and thus use the Bayes classifier as above. This yields a generalization of the QDA rule. See also Section 4.2.4 for a more direct approach.

The proof that LDA is the variance-maximizer linear classifier is still valid in the Hilbert space framework. One needs just one the inner product in $H$ instead of that of $\mathbb{R}^d$. The LDA rule is also optimal in another way: it is the best linear centroid classifier, see [Delaigle and Hall, 2012, Section 2.2]. A centroid classifier classifies a sample $x$ to population $\Pi_1$ if and only if $d(x, \mu_1) < d(x, \mu_0)$, where $d$ is a (pseudo)metric on $H$. When the distance used is

$$d_v(x, \mu_i) = |\langle x, v\rangle - \langle \mu_i, v\rangle|,$$

we call it the linear centroid classifier with direction $v$. One easily sees that the LDA rule is equivalent to

$$0 < d_v^2(x, \mu_0) - d_v^2(x, \mu_1),$$

with $v = Q^{-1}(\mu_1 - \mu_0)$. Thus, it is a linear centroid classifier. The next proposition shows that it is optimal.

**Proposition 10.** Let $\Pi_0, \Pi_1$ have a Gaussian distribution with mean $\mu_0, \mu_1$ and covariance $Q$. Suppose $\mu_1 - \mu_0 \in Q^{1/2}(H)$. Then, the linear centroid classifier having the lowest misclassification probability is the one with direction $Q^{-1}(\mu_1 - \mu_0)$.

*Proof.* Lets compute the probability of misclassifying an element from population $\Pi_1$ in population $\Pi_0$:

$$\begin{aligned}\mathbb{P}_{\Pi_1}\left[d_v(x, \mu_1) \geq d_v(x, \mu_0)\right] &= \mathbb{P}_{\Pi_1}\left[(\langle x, v\rangle - \langle \mu_1, v\rangle)^2 \geq (\langle x, v\rangle - \langle \mu_0, v\rangle)^2\right] \\ &= \mathbb{P}_{\Pi_1}\left[2\langle x, v\rangle\langle \mu_0 - \mu_1, v\rangle + \langle \mu_1, v\rangle^2 - \langle \mu_0, v\rangle^2 \geq 0\right] \\ &= \mathbb{P}_{\Pi_1}\left[\langle x, v\rangle \geq \frac{\langle \mu_0 + \mu_1, v\rangle}{2}\right] = 1 - \Phi\left(\frac{\langle \mu_0 - \mu_1, v\rangle}{2\sqrt{\langle Qv, v\rangle}}\right)\end{aligned}$$

A quick computation shows that $\mathbb{P}_{\Pi_1}\left[d_v(x, \mu_1) \geq d_v(x, \mu_0)\right] = \mathbb{P}_{\Pi_0}\left[d_v(x, \mu_1) < d_v(x, \mu_0)\right]$, so that the overall probability of misclassification is

$$\frac{1}{2}\mathbb{P}_{\Pi_1}\left[d_v(x, \mu_1) \geq d_v(x, \mu_0)\right] + \frac{1}{2}\mathbb{P}_{\Pi_0}\left[d_v(x, \mu_1) < d_v(x, \mu_0)\right] = 1 - \Phi\left(\frac{\langle \mu_0 - \mu_1, v\rangle}{2\sqrt{\langle Qv, v\rangle}}\right). \tag{4.3}$$

The problem of minimizing the misclassification probability is thus equivalent to the maximization problem

$$\max_{v \in H} \frac{\langle \mu_0 - \mu_1, v\rangle}{\sqrt{\langle Qv, v\rangle}} \iff \max_{v \in H} \frac{\langle \mu_0 - \mu_1, v\rangle^2}{\langle Qv, v\rangle}.$$

Now, by the Cauchy-Schwarz inequality,

$$\frac{\langle \mu_0 - \mu_1, v\rangle^2}{\langle Qv, v\rangle} = \frac{\langle Q^{-1/2}(\mu_0 - \mu_1), Q^{1/2}v\rangle^2}{\langle Q^{1/2}v, Q^{1/2}v\rangle} \leq \frac{\|Q^{-1/2}(\mu_0 - \mu_1)\|^2 \, \|Q^{1/2}v\|^2}{\|Q^{1/2}v\|^2} = \|Q^{-1/2}(\mu_0 - \mu_1)\|^2,$$

where the inequality is an equality if and only if

$$v \propto Q^{-1}(\mu_0 - \mu_1).$$

In particular, the LDA direction $v = Q^{-1}(\mu_1 - \mu_0)$ achieves this maximum and is thus optimal. □

The proof of the previous proposition shows that the misclassification probability of LDA is $1 - \Phi\left(\|Q^{-1/2}(\mu_1 - \mu_0)\|/2\right)$. Thus, the larger $\mu_1 - \mu_0$ is in the RKHS of $Q$, the better the classification is.

We note that although the analysis of these classifiers is performed with the assumptions that the distributions are Gaussian, they also perform well when this assumption is not satisfied.

### 4.2.3 Infinite dimensional case: Singular Gaussian measures

A natural question now is: what happens when $\mu_1 - \mu_0 \notin Q^{1/2}(H)$? In that case, by Theorem 15, the distributions of $\Pi_0$ and $\Pi_1$ are singular. Let $A, B \in \mathcal{B}(H)$ form a partition of $H$ with $N_{\mu_0,Q}(B) = N_{\mu_1,Q}(A) = 0$. Then, a perfect classifier would be $\mathbf{1}[x \in B]$, see Section 4.1. The sets $A$ and $B$ cannot be written down explicitly. However, using the LDA approach, one can construct sequences of classifiers $C_n(x)$ that are asymptotically perfect, i.e.,

$$\lim_{n\to\infty} \left( \frac{1}{2}\mathbb{P}_{\Pi_0}[C_n(x) = 1] + \frac{1}{2}\mathbb{P}_{\Pi_1}[C_n(x) = 0] \right) = 0.$$

Following [Kraus and Stefanucci, 2017], we will present three techniques for constructing such sequences based on the LDA classifier.

The first one is just to project the problem on the first few principal components of the data. More precisely, let $\{\lambda_i, \phi_i\}_{i=1}^{\infty}$ be a complete orthonormal system of eigenvectors of $Q$. Then, since $\mu_1 - \mu_0 \notin Q^{1/2}(H)$, we must have

$$\sum_{i=1}^{\infty} \frac{\langle \mu_1 - \mu_0, \phi_i \rangle^2}{\lambda_i} = \infty. \tag{4.4}$$

The LDA direction $w = Q^{-1}(\mu_1 - \mu_0)$ does not belong to $H$. However, one can 'approximate' it in the subspace spanned by the first $m$ principal components $\{\phi_i\}_{i=1}^{m}$:

$$w_m = \sum_{i=1}^{m} \frac{\langle \mu_1 - \mu_0, \phi_i \rangle}{\lambda_i} \phi_i.$$

By Equation 4.3, the misclassification probability of the linear centroid classifier with direction $w_m$ is given by $1 - \Phi(\Upsilon_m)$ with

$$\Upsilon_m = \frac{\langle \mu_1 - \mu_0, w_m \rangle}{2\sqrt{\langle Q w_m, w_m \rangle}} = \frac{\langle \mu_1 - \mu_0, w_m \rangle}{2\|Q^{1/2} w_m\|},$$

and

$$Q^{1/2} w_m = \sum_{i=1}^{\infty} \sqrt{\lambda_i} \langle w_m, \phi_i \rangle \phi_i = \sum_{i=1}^{m} \frac{\langle \mu_1 - \mu_0, \phi_i \rangle}{\sqrt{\lambda_i}} \phi_i, \ \ \|Q^{1/2} w_m\|^2 = \sum_{i=1}^{m} \frac{\langle \mu_1 - \mu_0, \phi_i \rangle^2}{\lambda_i}$$

so that

$$\Upsilon_m = \frac{1}{2}\|Q^{1/2} w_m\| \overset{m\to\infty}{\longrightarrow} \infty \ \ \text{(by Equation 4.4).}$$

Therefore, the misclassification probability of the linear centroid classifier with direction $w_m$ goes to 0 as $m$ goes to infinity. This sequence of classifiers is thus asymptotically perfect.

The sequence of directions $w_m$ above is gotten by projecting the LDA direction on the subspaces $\text{span}\{\phi_1, ..., \phi_m\}$. We note that these subspaces do not depend on $\mu_0$ nor $\mu_1$. It should therefore be expected that there exist better sequences of subspaces on which one can project the LDA direction. One such sequence is the sequence of Krylov subspaces

$$K_m = \text{span}\left\{\mu_1 - \mu_0, ..., Q^{m-1}(\mu_1 - \mu_0)\right\}.$$

Projecting an element on these subspaces yields the conjugate gradient algorithm presented in [Kraus and Stefanucci, 2017, Algo. 1]. These subspaces depend on $\mu_0$ and $\mu_1$ and the next proposition shows that they give better classifiers than the principal component subspaces.

**Proposition 11.** [Kraus and Stefanucci, 2017, Prop. 2] For any $m$, projecting the LDA solution on $K_m$ gives a smaller misclassification probability than projecting it on the first $m$ principal components.

In particular, it is asymptotically perfect.

Finally, the third approach is to project the LDA direction on balls of finite radii. This is done by using the directions

$$w_\alpha = (Q + \alpha I)^{-1}(\mu_1 - \mu_0) \text{ with } \alpha \to 0_+.$$

This last expression makes sense:

$$\sum_{i=1}^{\infty} \frac{\langle \mu_1 - \mu_0, \phi_i \rangle^2}{(\lambda_i + \alpha)^2} = \frac{1}{\alpha^2} \sum_{i=1}^{\infty} \frac{\langle \mu_1 - \mu_0, \phi_i \rangle^2}{(\lambda_i/\alpha + 1)^2} \leq \frac{1}{\alpha^2} \sum_{i=1}^{\infty} \langle \mu_1 - \mu_0, \phi_i \rangle^2 = \frac{1}{\alpha^2} \|\mu_1 - \mu_0\|^2 < \infty.$$

The misclassification probability is given by $1 - \Phi(\Upsilon_\alpha)$ with

$$\Upsilon_\alpha = \frac{\langle \mu_0 - \mu_1, w_\alpha \rangle}{2\sqrt{\langle Q w_\alpha, w_\alpha \rangle}}.$$

Using

$$\sum_{i=1}^{\infty} \frac{\lambda_i \langle \mu_1 - \mu_0, \phi_i \rangle^2}{(\lambda_i + \alpha)^2} \leq \sum_{i=1}^{\infty} \frac{(\lambda_i + \alpha)\langle \mu_1 - \mu_0, \phi_i \rangle^2}{(\lambda_i + \alpha)^2} = \sum_{i=1}^{\infty} \frac{\langle \mu_1 - \mu_0, \phi_i \rangle^2}{\lambda_i + \alpha},$$

we get

$$\Upsilon_\alpha = \frac{\sum_{i=1}^{\infty} \frac{\langle \mu_1 - \mu_0, \phi_i \rangle^2}{\lambda_i + \alpha}}{\left( \sum_{i=1}^{\infty} \frac{\lambda_i \langle \mu_1 - \mu_0, \phi_i \rangle^2}{(\lambda_i + \alpha)^2} \right)^{1/2}} \geq \left( \sum_{i=1}^{\infty} \frac{\langle \mu_1 - \mu_0, \phi_i \rangle^2}{\lambda_i + \alpha} \right)^{1/2} \overset{\alpha \to 0_+}{\longrightarrow} \infty \text{ by Equation 4.4.}$$

Therefore, using $\alpha = 1/m$, we get a sequence of classifiers that is asymptotically perfect.

All three methods presented above work in the same way. They project the LDA direction on various sequences of subsets of $H$: the principal component subspaces, the Krylov subspaces or balls of increasing radii. These give sensible approximation of the non-existing $Q^{-1}(\mu_1 - \mu_0)$. However, when the latter does exist, it still makes sense to use these methods and see them as regularization techniques. As all three sequences converge to the range of $Q$, it is easy to prove that the induced sequences of direction $w_m$ converge to the LDA direction $Q^{-1}(\mu_1 - \mu_0)$ when it exists.

### 4.2.4 QDA for infinite dimensional data

Consider now the classification framework where two populations $\Pi_0$ and $\Pi_1$ have means $\mu_0$ and $\mu_1$ and covariance operators $Q_0$ and $Q_1$. Then, we can adapt the QDA rule 4.2 in this setting. This cannot be done directly, since the determinant of an operator is not well-defined and since the inverses $Q_0^{-1}, Q_1^{-1}$ do not exist in general. However, one can get around these issues by truncating these operations at order $m$. More precisely, let $\{\lambda_i, \phi_i\}_{i=1}^{\infty}$ and $\{\nu_i, \psi_i\}_{i=1}^{\infty}$ be complete orthonormal system of eigenvectors of $Q_0$ and $Q_1$ respectively. Then, one can use the following approximations

$$\log |Q_0| \rightsquigarrow \sum_{i=1}^{m} \log \lambda_i \quad \text{and} \quad Q_0^{-1} \rightsquigarrow \sum_{i=1}^{m} \frac{1}{\lambda_i} \phi_i \otimes \phi_i,$$

and similarly for $Q_1$. These yield the following QDA classifier: we classify an observation $x$ to population $\Pi_1$ if and only if

$$2 \log \left( \frac{p}{1-p} \right) < \sum_{i=1}^{m} \left[ \log \left( \frac{\lambda_i}{\nu_i} \right) + \frac{\langle x - \mu_0, \phi_i \rangle^2}{\lambda_i} - \frac{\langle x - \mu_1, \psi_i \rangle^2}{\nu_i} \right].$$

In [Delaigle and Hall, 2013, Theorem 1], it is shown that under various conditions, this classifier is asymptotically perfect as $m$ goes to infinity. We are, however, still looking for a clean proof that this classifier is asymptotically perfect when the distributions of $\Pi_0$ and $\Pi_1$ are singular.

## 4.3 Classification of functional fragments

We now investigate the problem of classifying functional fragments. We use the same framework and notations as in Chapter 2. We will work in the following classification framework: we assume $H = \mathbb{L}^2[0,1]$, there are two populations $\Pi_0$ and $\Pi_1$ on $H$ and a probability distribution $\upsilon$ on the set of subintervals of $[0,1]$ that is independent of $\Pi_0$ and $\Pi_1$. Given a new observation $(X, [a,b])$, we want to decide whether $X$ comes from population $\Pi_0$ or $\Pi_1$.
In that form, this framework is (intentionally) too vague. For example, here are some possibilities:

1. We have full information on the distribution of $\Pi_0$ and $\Pi_1$ and need to classify fragments.
2. We observe fragments and need to classify fragments.
3. We observe fragments and need to classify full curves.
4. We observe full curves and need to classify fragments.

To the best of our knowledge, only two papers have been written exclusively on that subject in the FDA literature, [Delaigle and Hall, 2013] and [Kraus and Stefanucci, 2017]. We will next present the methods given in these papers. We also note that in the case where the functions are observed only on a finite grid, which is always the case in practice, methods from machine learning, in particular the use of recurrent neural networks, have been developed to solve these classification tasks.

First, we describe a general approach to the classification of functional fragments. The "training" procedure yields information on the populations $\Pi_0$ and $\Pi_1$, typically in the form of estimates of the means $\hat{\mu}_0, \hat{\mu}_1$ and covariance kernels $\hat{K}_0, \hat{K}_1$ of $\Pi_0$ and $\Pi_1$. We can also assume homoscedasticity and have an estimate $\hat{K}$ of the common covariance kernel. It can also be, as in case 1. above, that the true values of these parameters are given. Then, given a new observation $(X, [a,b])$, we use the functional classification techniques, e.g. LDA and QDA, on this observation by restricting the problem to $[a,b]$. For example, we use LDA on $X \in \mathbb{L}^2[a,b]$ with parameters

$$\hat{\mu}_0 \,|_{[a,b]}, \ \hat{\mu}_1 \,|_{[a,b]} \ \text{ and } \hat{K} \,|_{[a,b]\times[a,b]} \ .$$

This gives a general method for the classification of functional fragments. We will see next how one can estimate $\hat{\mu}_j$, $\hat{K}_j$ in clever ways to achieve better classification accuracy.

In [Delaigle and Hall, 2013], the authors use their curve extension method, described in Section 2.3, to get more information on the population $\Pi_0$ and $\Pi_1$ and to then achieve a better classification accuracy. More precisely, it is assumed that fragments

$$(X_{0,1}, \mathcal{O}_{0,1}), ..., (X_{0,n_0}, \mathcal{O}_{0,n_0}) \overset{i.i.d}{\sim} \Pi_0 \otimes \upsilon \ \text{ and } \ (X_{1,1}, \mathcal{O}_{1,1}), ..., (X_{1,n_1}, \mathcal{O}_{1,n_1}) \overset{i.i.d}{\sim} \Pi_1 \otimes \upsilon$$

are observed. Then, a new observation $(X, [a,b])$ is classified in the following way: the observed fragments are all extended to $\tilde{X}_{j,i}$ on an interval

$$\mathcal{I} \supset \bigcup_{i=1}^{n_0} \mathcal{O}_{0,i} \ \cup \ \bigcup_{i=1}^{n_1} \mathcal{O}_{1,i} \ \cup \ [a,b]$$

using the gluing method. Weights $w_{j,i}$ quantifying how far the interval $\mathcal{O}_{j,k}$ is from the interval $\mathcal{I}$ are computed for $j = 0,1$, $i = 1, ..., n_j$. The means and covariance kernels of both populations are are computed by

$$\hat{\mu}_j(t) = \frac{1}{n_j} \sum_{i=1}^{n_j} w_{j,i} \tilde{X}_{j,i}(t) \ \text{ and } \ \hat{K}_j(s,t) = \frac{1}{n_j} \sum_{i=1}^{n_j} w_{j,i} \left[ (\tilde{X}_{j,i}(s) - \hat{\mu}_j(s))(\tilde{X}_{j,i}(t) - \hat{\mu}_j(t)) \right],$$

for $j = 0,1$, $s,t \in \mathcal{I}$. Finally, the curve $X$ on $\mathcal{I}$ is classified using the QDA method described in Section 4.2.4. The LDA rule can also be used if all the observations are aggregated when estimating the covariance kernel. In [Delaigle and Hall, 2013, Section 5.2], it is shown that under some conditions this classifier performs better than using QDA (or LDA) without curve extensions.

In [Kraus and Stefanucci, 2017], the framework is the same as above. Their approach to the classification of fragments is, however, more pragmatic. They consider the smallest common domain of all curves

$$\mathcal{I}_0 = \bigcap_{i=1}^{n_0} \mathcal{O}_{0,i} \ \cap \ \bigcap_{i=1}^{n_1} \mathcal{O}_{1,i} \ \cap \ [a,b]$$

and the largest subdomain $\mathcal{I}^{max}$ of $[a,b]$ where enough curves have been observed. They then consider a chain of domains

$$\mathcal{I}_0 \subset \mathcal{I}_1 \subset ... \subset \mathcal{I}_K = \mathcal{I}^{max} \subseteq [a,b].$$

For all such subdomains $\mathcal{I}_j$, the estimated means $\hat{\mu}_0^j, \hat{\mu}_1^j$ and common covariance $\hat{K}^j$ are computed using the naive estimators given in Section 2.2. They then perform cross-validation on all the $\mathcal{I}_j$'s to estimate the misclassification rate on each using LDA with one of the regularization techniques presented in the last section. Finally, they classify the observation $(X,[a,b])$ on the subdomain having the lowest cross-validation misclassification rate. The idea behind this approach is that performing classification only on $\mathcal{I}_0$ might be too restrictive since a big fraction of the curves will be ignored. On the other hand, using $\mathcal{I}^{max}$ might be too optimistic since many curves may be unobserved on this domain and so the estimations of the means and covariance may be too biased. Thus, this algorithm looks for the right balance between these two extreme subdomains.
We note that this approach is designed for when the fragments are observed in the "blanket" regime (see Section 2.1). Indeed, in that regime the smallest common domain $\mathcal{I}_0$ is already quite large and so classification can be performed meaningfully on it. In the "banded" regime, $\mathcal{I}_0$ is typically small, even empty when $\delta < 0.5$.

Finally, we present a novel approach based on the covariance recovery of [Descary and Panaretos, 2017]. We use the same framework as above and assume further that we are in the "banded" regime, i.e., that all observations are of length $\delta \in (0,1)$. The naive estimates of the means $\hat{\mu}_0, \hat{\mu}_1$ and covariance kernels $\tilde{K}_0, \tilde{K}_1$ are computed. As before, we can also assume homoscedasticity and get an estimate $\tilde{K}$ of the common covariance kernel. Then, using the matrix completion method, we get estimates $\hat{K}_0, \hat{K}_1$ of rank $r_0, r_1$. Finally, a new observation $(X,[a,b])$ is classified using QDA with parameters

$$\hat{\mu}_0\,|_{[a,b]},\ \hat{\mu}_1\,|_{[a,b]},\ \hat{K}_0\,|_{[a,b]\times[a,b]} \ \text{ and } \hat{K}_1\,|_{[a,b]\times[a,b]},$$

or LDA with $\hat{K}\,|_{[a,b]\times[a,b]}$. In the case where $b-a \leq \delta$, this can be seen as a regularization technique on the covariance kernel estimation by intentionally reducing the rank of the naive estimates. In the case where $b-a > \delta$, the matrix completion method recovers new meaningful information on the kernels outside of the band $\{|s-t| \leq \delta\}$. This can be used to classify longer fragments with more accuracy.

## 4.4 A probabilistic approach to classification

### 4.4.1 Explain the perfect classification phenomenon through the F-H theorem. If I manage to find something, also the classification of fragments.

# Chapter 5

# Numerical experiments

# Bibliography

[Bogachev, 2015] Bogachev, V. (2015). *Gaussian Measures:*. Mathematical Surveys and Monographs. American Mathematical Society.

[Da Prato, 2006] Da Prato, G. (2006). *An Introduction to Infinite-Dimensional Analysis*. Universitext. Springer Berlin Heidelberg.

[Da Prato and Zabczyk, 2008] Da Prato, G. and Zabczyk, J. (2008). *Stochastic Equations in Infinite Dimensions*. Encyclopedia of Mathematics and its Applications. Cambridge University Press.

[Debnath and Mikusinski, 2005] Debnath, L. and Mikusinski, P. (2005). *Introduction to Hilbert Spaces with Applications*. Elsevier Science.

[Delaigle and Hall, 2012] Delaigle, A. and Hall, P. (2012). Achieving near perfect classification for functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):267–286.

[Delaigle and Hall, 2013] Delaigle, A. and Hall, P. (2013). Classification using censored functional data. *Journal of the American Statistical Association*, 108(504):1269–1283, https://doi.org/10.1080/01621459.2013.824893.

[Delaigle and Hall, 2016] Delaigle, A. and Hall, P. (2016). Approximating fragmented functional data by segments of markov chains. *Biometrika*, 103(4):779–799.

[Descary and Panaretos, 2016] Descary, M.-H. and Panaretos, V. M. (2016). Functional data analysis by matrix completion. arXiv:1609.00834.

[Descary and Panaretos, 2017] Descary, M.-H. and Panaretos, V. M. (2017). Recovering covariance from functional fragments. arXiv:1708.02491.

[Devroye et al., 2013] Devroye, L., Györfi, L., and Lugosi, G. (2013). *A Probabilistic Theory of Pattern Recognition*. Stochastic Modelling and Applied Probability. Springer New York.

[Ferraty and Romain, 2010] Ferraty, F. and Romain, Y. (2010). *The Oxford Handbook of Functional Data Analysis*. Oxford Handbooks in Mathematics. OUP Oxford.

[Horváth and Kokoszka, 2012] Horváth, L. and Kokoszka, P. (2012). *Inference for Functional Data with Applications*. Springer Series in Statistics. Springer New York.

[Hsing and Eubank, 2015] Hsing, T. and Eubank, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. Wiley Series in Probability and Statistics. Wiley.

[Krantz and Parks, 2002] Krantz, S. and Parks, H. (2002). *A Primer of Real Analytic Functions*. A Primer of Real Analytic Functions. Birkhäuser Boston.

[Kraus, 2015] Kraus, D. (2015). Components and completion of partially observed functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(4):777–801.

[Kraus and Stefanucci, 2017] Kraus, D. and Stefanucci, M. (2017). Classification of functional fragments by regularized linear classifiers with domain selection. arXiv:1708.08257.

[Ramsay and Silverman, 2005] Ramsay, J. and Silverman, B. (2005). *Functional Data Analysis.* Springer Series in Statistics. Springer.

[Ramsay and Silverman, 2013] Ramsay, J. and Silverman, B. (2013). *Applied Functional Data Analysis: Methods and Case Studies.* Springer Series in Statistics. Springer New York.

[Tuddenham and Snyder, 1954] Tuddenham, R. D. and Snyder, M. M. (1954). Physical growth of california boys and girls from birth to eighteen years. *University of California Publications in Child Development.*

[Yao et al., 2005] Yao, F., Müller, H.-G., and Wang, J.-L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590, https://doi.org/10.1198/016214504000001745.