Functional Principal Component Analysis

— Advanced Statistics —

Daniel López Montero

MSc Mathematics and Applications
Universidad Autónoma de Madrid

May 15, 2025

Abstract

In this manuscript, we explore the application of dimensionality reduction algorithms to real-world datasets within the context of functional data analysis. We establish several theoretical results related to Principal Component Analysis (PCA) for functional data and introduce a novel variation, *Fourier PCA*, inspired by Fourier theory. Additionally, we extend Kernel PCA to the functional data setting by proposing new kernels, adapted from well-known finite-dimensional counterparts, and provide theoretical foundations for their use. Finally, we evaluate and compare the performance of these methods. All code associated with this study is available in a *GitHub repository*[1].

Keywords — Functional Data Analysis, Functional PCA, PCA for functional data, Fourier PCA, Kernel PCA

1 Introduction

For many years, principal component analysis has been a key dimension reduction tool for multivariate data in classical statistics and machine learning. Lately, it has been extended to functional data and termed functional principal component analysis (FPCA). FPCA has taken off to become the most prevalent tool in functional data analysis. This is partly because FPCA facilitates the conversion of inherently infinite-dimensional functional data to a finite-dimensional vector. Under mild assumptions, the underlying stochastic process can be expressed as a countable sequence of uncorrelated random variables, which are then truncated to a finite vector. Then the tools of multivariate data analysis can be readily applied to infinite-dimensional data without the hassle of handling lower-dimensional representations of the data.

Specifically, the dimension reduction is achieved through the Karhunen-Loève expansion, which is a well-known result in stochastic processes. It decomposes observed random trajectories $X_i(t)$ in a functional basis that consists of the eigenvectors of the covariance operator of the process X, i.e., $K(s,t) := \mathrm{Cov}(X_s, X_t)$.

The extension of FPCA to multivariate functional data is hence of high practical relevance. Existing approaches for multivariate functional principal component analysis (mFPCA) are based on the multivariate functional Karhunen-Loève expansion to compress the data into a finite-dimensional space (Berrendero et al. 2011; Chiou et al. 2014; Jacques et al. 2014; Ramsay et

[1] GitHub repository: `https://github.com/dani2442/Functional-Data-Analysis`

real numbers	$\alpha, \beta, \gamma, \dots$	Greek characters
integers	i, j, k, m, n	
vector spaces	$\mathcal{X}, \mathcal{Y}, \mathcal{H}, \dots$	Calligraphic letters
subsets of the real plain	$\Omega, \Lambda, \Gamma, \dots$	capital Greek characters
functions	$x, y, f, \dots$	small Latin characters
vector of functions (or vectors)	$\mathbf{u}, \mathbf{v}, \mathbf{w}, \dots$	small bold Latin characters
operators	$A, B, K, \dots$	capital Latin characters
matrices	$\mathbf{A}, \mathbf{B}, \mathbf{K}, \dots$	capital bold Latin characters
random variables	X, Y, Z	capital Latin characters
convolution operator	$*$	$f * g$ convolution operator

Table 1: Notations used in this paper.

al. 2005). However, (Berrendero et al. 2011) proposed a PCA-like method to compress a multivariate function space $L^2(\mathbb{R}, \mathbb{R}^p)$ to a lower-dimensional multivariate function space $L^2(\mathbb{R}, \mathbb{R}^k)$ with $k \ll p$, using the principles of standard finite-dimensional PCA. We have explored some theoretical properties of this method, and we have proposed a similar method that uses the Fourier Theory to prove that this new method is translation equivariant. Therefore, this method is able to handle continuous streams of data.

Lastly, we explored Kernel PCA (Schölkopf et al. 1997), a method that compresses data through a non-linear projection, offering greater expressiveness than linear techniques. We proposed an extension of this approach to functional data by generalizing the kernel functions. Furthermore, we demonstrated that these generalized kernels are reproducing kernels and examined their universality. Universality is a foundational concept in machine learning. For instance, in deep learning, the Universal Approximation Theorem (Cybenko 1989) establishes that neural networks can approximate any continuous function. This property also plays a critical role in kernel-based methods.

In Table 1, we have summarized the notation used throughout this manuscript. We will denote by $\mathcal{X}$ the space of our input space; in our case, it will be $L^2(\mathbb{R}, \mathbb{R}^d)$. We will denote by $\mathcal{Z}$ the space of the *principal component space* or *latent* space; it can be a finite or infinite-dimensional space, e.g., $\mathbb{R}^n$ or $L^2(\mathbb{R}, \mathbb{R}^k)$, see Table 2. The compression method has two functions, the projection that transforms the input into the lower dimensional space, and the functions that *reverse* or

$$\Phi : \mathcal{X} \to \mathcal{Z} \tag{Transform}$$

$$\Phi^* : \mathcal{Z} \to \mathcal{X} \tag{Inverse Transform}$$

Let us denote by P the projection onto the subspace spanned by the principal components, i.e.,

$$P : \mathcal{X} \to \mathcal{X} \qquad P = \Phi^* \circ \Phi$$

Let X be a random variable taking values in the space $\mathcal{X}$. The principal component analysis problem can be interpreted from two equivalent perspectives: maximizing the variance of the projections or minimizing the mean squared reconstruction error:

$$\max_{\Phi} \operatorname{Var}(\Phi(X)) \qquad \Leftrightarrow \qquad \min_{\Phi} \mathbb{E}\|X - P_\Phi X\|^2_{\mathcal{X}}$$

Table 2 summarizes some of the most common PCA methods and their corresponding optimal solutions.

	Compress ($\mathcal{X} \longrightarrow \mathcal{Z}$)	**Optimal Value**
PCA (Pearson 1901)	$\mathbb{R}^p \longrightarrow \mathbb{R}^k$ $\mathbf{x} \longmapsto \mathbf{V}^\top \mathbf{x}$	$\boldsymbol{\Sigma}\mathbf{v}_i = \lambda_i \mathbf{v}_i$
Kernel PCA (Schölkopf et al. 1997)	$\mathbb{R}^p \longrightarrow \mathbb{R}^k$ $\mathbf{x} \longmapsto \left\{\sum_{j=1}^n \boldsymbol{\alpha}_{ij} K(\mathbf{x}_j, \mathbf{x})\right\}_{i=1}^k$	$\mathbf{K}^2 \boldsymbol{\alpha}_i = \lambda_i \boldsymbol{\alpha}_i$
Functional PCA (Karhunen 1946; Loève 1946)	$L^2(\mathbb{R}) \longrightarrow \mathbb{R}^k$ $x(t) \longmapsto \left\{\int x(t) e_j(t) dt\right\}_{j=1}^k$	$\int K(t,s) e_i(t) ds = \lambda_i e_i(s)$
PCA for Functional Data (Berrendero et al. 2011)	$L^2(\mathbb{R}, \mathbb{R}^p) \longrightarrow L^2(\mathbb{R}, \mathbb{R}^k)$ $\mathbf{x}(t) \longmapsto \mathbf{V}(t)^\top \mathbf{x}(t)$	$\boldsymbol{\Sigma}(t)\mathbf{v}_i(t) = \lambda_i(t)\mathbf{v}_i(t)$
Fourier PCA — *Ours* —	$L^2(\mathbb{R}, \mathbb{R}^p) \longrightarrow L^2(\mathbb{R}, \mathbb{R}^k)$ $\mathbf{x}(t) \longmapsto (\mathbf{V}^\top * \mathbf{x})(t)$	$\widehat{\boldsymbol{\Sigma}}(\xi)\widehat{\mathbf{v}}_i(\xi) = \lambda_i(\xi)\widehat{\mathbf{v}}_i(\xi)$
Kernel Functional PCA — *Ours* —	$L^2(\mathbb{R}) \longrightarrow \mathbb{R}^k$ $\mathbf{x} \longmapsto \{\sum_{i=1}^n \boldsymbol{\alpha}_{ij} K(x_i, x)\}_{j=1}^k$	$\mathbf{K}^2 \boldsymbol{\alpha}_i = \lambda_i \boldsymbol{\alpha}_i$

Table 2: Principal Component Analysis Methods.

2 PCA for Functional Data

Let $\mathbf{X}(t)$ be a $p-$dimensional stochastic process, i.e., $\mathbf{X}(t) = (X_1(t), \ldots, X_p(t))'$ defined on a probability space $(\Omega, \mathcal{F}, P)$. We will assume that $t \in \mathbb{R}$. Moreover, we assume that the random vector $\mathbf{X}(t)$ has mean vector $\boldsymbol{\mu}(t) = 0$, otherwise, we consider $\tilde{\mathbf{X}}(t) := \mathbf{X}(t) - \boldsymbol{\mu}(t)$ instead of $\mathbf{X}(t)$. Let us define the point-wise covariance matrix $\boldsymbol{\Sigma}(t) = \mathbb{E}[\mathbf{X}(t)\mathbf{X}(t)']$. Note that $\Sigma(t)$ is positive definite and symmetric.

We seek a linear function of $\mathbf{X}$'s components that accounts for most of its information. In other words, we aim to find $\mathbf{v}(t) \in \mathbb{R}^p$ that maximizes the variance of the projection of $\mathbf{X}(t)$ in $\mathbf{v}(t)$. It can be expressed as $\max_{\mathbf{v}(t)} \operatorname{Var}(\mathbf{X}(t)'\mathbf{v}(t))$. We notice that $\operatorname{Var}(\mathbf{X}(t)'\mathbf{v}(t)) = \mathbf{v}(t)'\mathbb{E}[\mathbf{X}(t)\mathbf{X}(t)']\mathbf{v}(t) = \mathbf{v}(t)'\boldsymbol{\Sigma}(t)\mathbf{v}(t)$. The criterion used in (Berrendero et al. 2011) is to consider the measurement of the integrated variance so that the weighting function $\mathbf{v}(t) : \mathbb{R} \to \mathbb{R}^p$ is defined as the function maximizing

$$\underset{\|\mathbf{v}(t)\|=1}{\arg\max} \int \operatorname{Var}(\mathbf{v}(t)'\mathbf{X}(t))dt = \underset{\|\mathbf{v}(t)\|=1}{\arg\max} \int \mathbf{v}(t)'\boldsymbol{\Sigma}(t)\mathbf{v}(t)dt \tag{1}$$

where $\| \cdot \|$ denotes the usual euclidean norm. The restriction on the norm of $\mathbf{v}$ is needed to reach a unique solution for each t, except for the sign.

Proposition 1. The maximum value of the objective in (1) is achieved when $u(t)$ satisfies

$$\boldsymbol{\Sigma}(t)\mathbf{v}(t) = \lambda(t)\mathbf{v}(t)$$

where $\lambda(t)$ denotes the largest eigenvalue of $\boldsymbol{\Sigma}(t)$ at each time t. In this case, the maximal value of the objective is given by

$$\int \mathrm{Var}(\mathbf{v}(t)'\mathbf{X}(t))dt = \int \lambda(t)dt.$$

On the other hand, the minimum value is achieved in the minimum eigenvalue.

Proof. The equation (1) is similar to the Rayleigh quotient. This problem can be solved in a similar fashion to the finite-dimensional case. We will use the Lagrange multiplier formulation.

$$\mathcal{L}(\mathbf{v}) := \int \mathbf{v}(t)'\mathbf{\Sigma}(t)\mathbf{v}(t)dt - \int \lambda(t)(\mathbf{v}(t)'\mathbf{v}(t) - 1)dt$$

We can calculate the Gateaux derivative of $\mathcal{L}(\mathbf{v})$ with respect to $\mathbf{v}(t)$

$$\begin{aligned} D_{\mathbf{v}}\mathcal{L}(\mathbf{v})(\mathbf{h}) &= \lim_{\epsilon \to 0} \frac{\mathcal{L}(\mathbf{v} + \epsilon\mathbf{h}) - \mathcal{L}(\mathbf{v})}{\epsilon} \\ &= \lim_{\epsilon \to 0} \frac{\int \epsilon^2 \mathbf{h}'\mathbf{\Sigma}(t)\mathbf{h} + 2\epsilon\mathbf{h}'\mathbf{\Sigma}(t)\mathbf{v}dt - \int 2\lambda(t)\epsilon\mathbf{h}'\mathbf{v} + \epsilon^2\lambda(t)\mathbf{h}'\mathbf{h}dt}{\epsilon} \\ &= 2\int \mathbf{h}(t)'(\mathbf{\Sigma}(t)\mathbf{v}(t) - \lambda(t)\mathbf{v}(t))dt \end{aligned}$$

We now that $\mathcal{L}(\mathbf{v})$ is a critical points if $D_{\mathbf{v}}f(\mathbf{v})(\mathbf{h}) = 0$ for all $\mathbf{h} \in L^2(\mathbb{R})$ and

$$D_{\mathbf{v}}\mathcal{L}(\mathbf{v}) \equiv 0 \Leftrightarrow \mathbf{\Sigma}(t)\mathbf{v}(t) = \lambda(t)\mathbf{v}(t)$$

On the other hand, we can substitute in (1) assuming that $\mathbf{v}(t)$ is an eigenvector of $\mathbf{\Sigma}(t)$ and that $\|\mathbf{v}(t)\| = 1$, we obtain that the variance is given by

$$\int \mathrm{Var}(\mathbf{X}(t)'\mathbf{v}(t))dt = \int \mathbf{v}(t)'\mathbf{\Sigma}(t)\mathbf{v}(t)dt = \int \mathbf{v}(t)'\lambda(t)\mathbf{v}(t)dt = \int \lambda(t)dt$$

We will denote by $\lambda_1(t) := \max_{1 \le j \le p} \lambda_j(t)$, and we notice that it is the one that maximizes the variance. And, we denote by $\mathbf{v}_1(t)$ the normalized eigenvectors of $\lambda_1(t)$ and name it the first principal component. □

If we order the eigenvectors of $\mathbf{\Sigma}(t)$ in descending order with respect to the eigenvalues, we will denote $\lambda_j(t) \in \mathbb{R}$ and $\mathbf{v}_j(t)$ for $j = 1, \dots, p$. We have proven that the principal component for the PCA method when $k = 1$ is given by the 1st eigenvector each time t. Next, we will define the PCA problem for a general number of principal components.

Definition 1. Let $k \geq 1$ be the number of principal components, the PCA problem is to find the functions $\{\mathbf{v}_i(t)\}_{i=1}^k \subseteq L^2(\mathbb{R}, \mathbb{R}^p)$ that maximize the variance of the sum of the projections constrained to orthogonal principal components, i.e.,

$$\sup_{\substack{\|\mathbf{v}_i(t)\|=1 \; 1 \le i \le k \\ \mathbf{v}_i(t) \perp \mathbf{v}_j(t) \; i \neq j}} \int \mathrm{Var}\left(\sum_{i=1}^k \mathbf{v}_i(t)'\mathbf{X}(t)\right) dt = \sup_{\substack{\|\mathbf{v}_i(t)\|=1 \; 1 \le i \le k \\ \mathbf{v}_i(t) \perp \mathbf{v}_j(t) \; i \neq j}} \sum_{i=1}^k \int \mathbf{v}_i(t)'\mathbf{\Sigma}(t)\mathbf{v}_i(t)dt \tag{2}$$

Notice that when $k = 1$, the same problem is solved above. The second expression is equivalent because $\{\mathbf{v}_i(t)\}_{i=1}^k$ are orthogonal, and thus, we can get the summation outside of the variance.

We can use the Fatou Lemma to transform this problem to the well know finite-dimensional case

$$\sup_{\substack{\|\mathbf{v}_i(t)\|=1 \;\; 1\le i\le k \\ \mathbf{v}_i(t)\perp\mathbf{v}_j(t) \;\; i\neq j}} \sum_{i=1}^{k} \int \mathbf{v}_i(t)'\mathbf{\Sigma}(t)\mathbf{v}_i(t)dt \leq \int \sup_{\substack{\|\mathbf{v}_i\|=1 \;\; 1\le i\le k \\ \mathbf{v}_i\perp\mathbf{v}_j \;\; i\neq j}} \sum_{i=1}^{k} \mathbf{v}_i'\mathbf{\Sigma}(t)\mathbf{v}_i dt$$

$$= \int \sup_{\mathbf{v}_i\perp\mathbf{v}_j \;\; i\neq j} \sum_{i=1}^{k} \frac{\mathbf{v}_i'\mathbf{\Sigma}(t)\mathbf{v}_i}{\mathbf{v}_i'\mathbf{v}_i} dt$$

Using the Courant-Fischer Theorem 1, we know that the vectors $\{\mathbf{v}_i\}$ that maximize the Rayleigh, i.e., $\frac{\mathbf{v}_i'\mathbf{\Sigma}(t)\mathbf{v}_i}{\mathbf{v}_i'\mathbf{v}_i}$ are exactly the eigenvectors of the largest eigenvalues of the matrix $\mathbf{\Sigma}(t)$ at each time t. Therefore, we can denote the eigenvalues $\lambda_1(t) \geq \cdots \geq \lambda_k(t)$ and its corresponding normalized eigenvectors $\mathbf{v}_1(t), \ldots, \mathbf{v}_k(t)$.

Theorem 1 (Courant-Fischer Theorem)**.** *Let A be a compact self-adjoint operator on a Hilbert space $\mathcal{H}$, whose positive eigenvalues are listed in decreasing order $\lambda_1 \geq \lambda_2 \geq \ldots$ and whose corresponding eigenvalues are $u_1, u_2, \ldots$. Then,*

$$\max_{S_k} \min_{x\in S_k, \|x\|=1} \langle Ax, x\rangle = \lambda_k$$

$$\min_{S_{k-1}} \max_{x\in S_{k-1}^{\perp}, \|x\|=1} \langle Ax, x\rangle = \lambda_k$$

where $S_k \subseteq \mathcal{H}$ such that for all $x \in S_k$ is a $k-$dimensional subspace.

Let $P_{\mathbf{V}(t)} : \mathcal{X} \to \mathcal{X}$ be the projection to the lower dimensional space spanned by $\mathbf{V}(t)$ in $\mathcal{X}$, i.e.,

$$P_{\mathbf{V}(t)}(\mathbf{x}) = \mathbf{V}(t)\mathbf{V}(t)^{\top}\mathbf{x}$$

The next result states that the solution that minimizes the mean square error of the projection is exactly the solution that maximizes the variances. This provides a different lens for seeing the Principal Component Analysis.

Theorem 2. *The k principal components obtained from the eigen-problem $\mathbf{\Sigma}(t)\mathbf{v}(t) = \lambda(t)\mathbf{v}(t)$ form a orthonormal basis $\mathbf{V}(t) := \{\mathbf{v}_i(t)\}_{i=1}^{k}$ that minimizes the square error of the reconstruction, i.e.*

$$\inf_{\mathbf{V}(t)'\mathbf{V}(t)=I} \mathbb{E}\|\mathbf{X}(t) - P_{\mathbf{V}(t)}\mathbf{X}(t)\|_2^2$$

where $\|\mathbf{v}(t)\|_2^2 := \int \mathbf{v}(t)'\mathbf{v}(t)dt$ In other words, it is equivalent to maximizing the variance rather than minimizing the reconstruction error.

We will prove some properties regarding the eigenvalues and eigenvectors when seen as a continuously differentiable function $\mathbf{\Sigma}(t) \in \mathcal{C}^2(\mathbb{R}, \mathbb{R}^{p\times p})$. The next result uses this fact to build some intuition in the solutions of the eigenvalue problem.

Proposition 2. *Assume that $\mathbf{\Sigma}(t)$ is continuously differentiable. and that $\mathbf{\Sigma}(t)$ has simple eigenvalues at $t = t^*$. Let us denote $\lambda_1(t^*) > \cdots > \lambda_p(t^*)$ the eigenvalues. Then, for $k = 1, \ldots, p$ and a sufficiently small neighborhood of t^*, i.e., $t \in (t^* - \epsilon, t^* + \epsilon)$, there exist functions $\lambda_k(t)$ and $\mathbf{v}_k(t)$ differentiable in a neighborhood of t^* verifying that $\lambda_k(t)$ passes through $\lambda_k(t^*)$ and $\|\mathbf{v}_k(t)\| = 1$ and satisfying $\mathbf{\Sigma}(t)\mathbf{v}_k(t) = \lambda_k(t)\mathbf{v}_k(t)$. Furthermore, these functions must hold the following properties*

(i) The derivatives of $\lambda_k(t)$ in a neighborhood of t^* are given by

$$\dot{\lambda}_k(t) = \mathbf{v}_k^*(t)\dot{\mathbf{\Sigma}}(t)\mathbf{v}_k(t)$$
$$\ddot{\lambda}_k(t) = \mathbf{v}_k^*(t)\ddot{\mathbf{\Sigma}}(t)\mathbf{v}_k(t) + 2\sum_{j\neq k}\frac{|\mathbf{v}_k^*(t)\dot{\mathbf{\Sigma}}(t)\mathbf{v}_j(t)|^2}{\lambda_k(t)-\lambda_j(t)}$$

(ii) The derivative of $\mathbf{v}_k(t^*)$ is given by

$$\dot{\mathbf{v}}_k(t) = \sum_{j\neq k}\frac{\mathbf{v}_j^*(t)\dot{\mathbf{\Sigma}}(t)\mathbf{v}_j(t)}{\lambda_k(t)-\lambda_j(t)}\mathbf{v}_k(t)$$

We can interpret both summation terms as *forces* acting on the eigenvalue $\lambda_k(t)$; on the left-hand side, we see that the original covariance matrix $\mathbf{\Sigma}(t)$ applies force, while all the other eigenvalues λ_j provide a repulsive force of the order of $\frac{1}{\lambda_k-\lambda_j}$, so it becomes stronger the closer they become. Hence, we expect the eigenvalues not to intersect under normal circumstances. Furthermore, they are constrained by the trace of the matrix as $\mathrm{tr}(\mathbf{\Sigma}(t)) = \sum_i \lambda_i(t)$.

This effect is consistent with the smallness of the class of hermitian matrices that have repeated eigenvalues. Consider the space of Hermitian $n \times n$ matrices. We see this space as a manifold embedded in the manifold of matrices. Considering the space of Hermitian matrices with simple eigenvalues, we can see that its dimension is n^2. Indeed, any Hermitian matrix can be expressed $\mathbf{A} = \mathbf{Q\Lambda Q}^\dagger$ where $\mathbf{\Lambda}$. The orthogonal matrix gives us $2(n-1) + 2(n-2) + \cdots + 2 = n(n-1)$ degrees of freedom because we can choose the first direction as we please, but the following have to verify that it is orthogonal to the previous ones, reducing the dimension. The diagonal gives us n degrees of freedom because it has to belong to $\mathbb{R}$. On the other hand, the class of Hermitian matrices with at least two equal eigenvalues has dimension less than or equal to $n^2 - 3$, which is significantly smaller (Lebesgue measure zero). Indeed, assume that the two eigenvalues are equal. Then, we reduce a degree of freedom by one on the diagonal and reduce 2 more because if we express the 2x2 matrix $\mathbf{Q}_2\mathbf{\Lambda}_2\mathbf{Q}_2^\dagger = \lambda\mathbf{Q}_2\mathbf{Q}_2^\dagger = \lambda\mathbf{I}$.

3 Fourier PCA

In most applications, functional data is acquired in a continuous stream with no notion of start and end. Most methods used in functional data analysis are not prepared to deal with this kind of data. We propose a method based on the Fourier transform that allows the processing of multivariate signals for PCA. The goal is to find a PCA-like method that is robust to translation, i.e., shifting in the time axis; in fact, we aim for a method that is translation invariant. Let us formulate the problem. let $\mathbf{X}(t) = (X_1(t), \dots, X_p(t))'$ be a $p-$dimensional stochastic process defined for $t \in \mathbb{R}$. We assume that $\mathbf{X}(t)$ is mean-centered, otherwise, we consider $\tilde{\mathbf{X}}(t) := \mathbf{X}(t) - \boldsymbol{\mu}(t)$. For simplicity, we consider the case where $t \in \mathbb{R}$, however, this method can be used similarly for any interval.

Let us build the intuition of the proposed method. Firstly, consider the standard PCA for functional data presented in the previous section. Assume that we want to compress the p different signals into only k where $k \ll p$. Then, following the previous section procedure, we select the eigenvectors $\mathbf{V}(t) := \{v_j(t)\}_{j=1}^k \in \mathbb{C}^{p\times k}$ whose eigenvalues are the greatest. Then, the z-scores are given by

$$\Phi_j : \mathbf{X}(t) \mapsto \sum_{j=1}^{p} \overline{v}_{ij}(t) X_i(t) \qquad j = 1, \dots, k$$

Let us denote $\mathbf{\Phi} := (\Phi_1, \dots, \Phi_k)$ and $\mathbf{Z}(t) := \Phi(\mathbf{X}(t)) = \mathbf{V}(t)^\dagger \mathbf{X}(t)$. On the other hand, the reconstructed signal is calculated as follows

$$\Phi_i^* : \mathbf{Z}(t) \mapsto \mathbf{v}_j(t)\mathbf{Z}(t) = \sum_{i=1}^{k} v_{ij}(t) Z_i(t) \qquad i = 1, \dots, p$$

We denote $\mathbf{\Phi}^* = (\Phi_1^*, \dots, \Phi_p^*)$ and $\widetilde{\mathbf{X}}(t) := \mathbf{\Phi}^*(\mathbf{Z}(t)) = (\mathbf{\Phi}^* \circ \mathbf{\Phi})(\mathbf{X}(t)) = \mathbf{V}(t)\mathbf{V}(t)^\dagger \mathbf{X}(t)$ which approximately verifies $\widetilde{\mathbf{X}}(t) \approx \mathbf{X}(t)$ because $\mathbf{V}(t)$ is *almost* an unitary matrix. The notation $\mathbf{\Phi}$ and $\mathbf{\Phi}^*$ is not a coincidence; we see that $\mathbf{\Phi}^*$ is the adjoint operation of $\mathbf{\Phi}$.

We notice that the operation $\mathbf{\Phi}$ does not verify symmetry by translation. However, if we use the convolution operation instead of the standard multiplication, we obtain the property we are aiming for. Let $\tau_s f(t) := f(t-s)$ be the translation operation, and let us define the modified operations $\mathbf{\Psi}$ and $\mathbf{\Psi}^*$ as follows

$$\Psi_j : \mathbf{X}(t) \mapsto \sum_{i=1}^{p} (\overline{v}_{ij}^{\text{rev}} * X_i)(t) \qquad j = 1 \dots, k$$

$$\Psi_i^* : \mathbf{Z}(t) \mapsto \sum_{j=1}^{k} (v_{ij} * Z_j)(t) \qquad i = 1 \dots, p$$

where $*$ denotes the convolution operator between functions, i.e, $(f * g)(t) = \int f(t-s)g(s)ds$. And $v_{ij}^{\text{rev}}(t) := v_{ij}(-t)$, i.e., is the reversal. Using this formulation, we obtain the following result.

Lemma 1. Let $\mathbf{\Psi}$ and $\mathbf{\Psi}^*$ be the same as in the previous definition. Then for any $\mathbf{X} : \mathbb{R} \to \mathbb{R}^p$

$$(\tau_s \circ \mathbf{\Psi})(\mathbf{X}(t)) = (\mathbf{\Psi} \circ \tau_s)(\mathbf{X}(t))$$
$$(\tau_s \circ \mathbf{\Psi}^*)(\mathbf{X}(t)) = (\mathbf{\Psi}^* \circ \tau_s)(\mathbf{X}(t))$$

where the translation τ_s is applied to the vectors for each component, i.e., $\tau_s \mathbf{X}(t) = \mathbf{X}(t-s)$. Hence,

$$\tau_s \circ (\mathbf{\Psi}^* \circ \mathbf{\Psi})(\mathbf{X}(t)) = \mathbf{\Psi}^* \circ (\tau_s \circ \mathbf{\Psi})(\mathbf{X}(t)) = (\mathbf{\Psi}^* \circ \mathbf{\Psi})(\tau_s \mathbf{X}(t)) \tag{3}$$

Proof. We can prove that this new definition verifies the following property for $j = 1, \dots, k$

$$\tau_s \boldsymbol{\Psi}_j(\mathbf{X}(t)) = \tau_s \left(\sum_{i=1}^{k} (\overline{v}_{ij}^{\text{rev}} * X_i)(t) \right) = \sum_{i=1}^{k} (\overline{v}_{ij}^{\text{rev}} * (\tau_s X_i))(t) = \boldsymbol{\Psi}_j(\tau_s \mathbf{X}(t)) \tag{4}$$

On the other hand, for $i = 1, \dots, p$

$$\tau_s \boldsymbol{\Psi}_i^*(\mathbf{Z}(t)) = \tau_s \left(\sum_{j=1}^{k} (v_{ij} * Z_j)(t) \right) = \sum_{j=1}^{k} (v_{ij} * (\tau_s Z_j))(t) = \Psi_i^*(\tau_s \mathbf{Z}(t)) \tag{5}$$

Therefore, combining (4) and (5) yields

$$\tau_s \circ (\boldsymbol{\Psi}^* \circ \boldsymbol{\Psi})(\mathbf{X}(t)) = \boldsymbol{\Psi}^* \circ (\tau_s \circ \boldsymbol{\Psi})(\mathbf{X}(t)) = (\boldsymbol{\Psi}^* \circ \boldsymbol{\Psi})(\tau_s \mathbf{X}(t))$$

□

We will see that this new definition of PCA verifies $\tilde{\mathbf{X}}(t) := (\boldsymbol{\Psi}^* \circ \boldsymbol{\Psi})(\mathbf{X}(t)) \approx \mathbf{X}(t)$, for a specific choice of $\mathbf{V}(t)$. Then, This is usually named by translation equivariance or τ_s-equivariance.

Let us prove that this new definition of PCA makes sense for a similar formulation of the problem. We will denote the new matrix multiplication with the convolution operator as follows: let $\mathbf{A}(t) \in \mathbb{R}^{n \times m}$ and $\mathbf{B}(t) \in \mathbb{R}^{m \times p}$. Then, $(\mathbf{A} * \mathbf{B})(t) := \int \mathbf{A}(s)\mathbf{B}(t-s)ds \equiv \mathbf{C}(t)$. This is equivalent to $c_{ij}(t) = \sum_{k=1}^{m} (a_{ik} * b_{kj})(t)$. In contrast with the standard multiplication that results in $c_{ij}(t) = \sum_{k=1}^{m} a_{ik}(t) b_{kj}(t)$.

Assume that $\mathbf{X}(t) \in L^2(\mathbb{R}, \mathbb{R}^p) \cap L^1(\mathbb{R}, \mathbb{R}^p)$. Consider the Fourier transform of $\mathbf{X}(t)$ denoted by $\widehat{\mathbf{X}}(\xi) = (\widehat{X}_1(\xi), \dots, \widehat{X}_p(\xi))'$ for $\xi \in \mathbb{R}$. Notice that we have considered the Fourier transform applied element-wise, and the same will hold henceforth. Then, the covariance matrix is given by $\widehat{\boldsymbol{\Sigma}}(\xi) := \mathbb{E}[\widehat{\mathbf{X}}(\xi)\widehat{\mathbf{X}}(\xi)^\dagger]$ and due to the Convolution theorem

$$\widehat{\boldsymbol{\Sigma}}(\xi) = \mathbb{E}[\widehat{\mathbf{X}}(\xi)\widehat{\mathbf{X}}(\xi)^\dagger] = \mathbb{E}[\mathcal{F}\{\mathbf{X}(t) * \mathbf{X}(t)^\dagger\}(\xi)]$$

This is a very useful property in signal processing and system analysis because convolution in the time domain (which is computationally expensive) becomes multiplication in the frequency domain (which is faster with tools like FFT). Next, we compute the eigenvalues and eigenvectors of $\widehat{\boldsymbol{\Sigma}}(\xi)$ as in the standard mFPCA.

$$\widehat{\boldsymbol{\Sigma}}(\xi)\hat{\mathbf{v}}(\xi) = \hat{\lambda}(\xi)\hat{\mathbf{v}}(\xi)$$

Let us denote by $\widehat{\mathbf{V}}(\xi) \in \mathbb{C}^{p \times k}$ the first $k \ll p$ eigenvectors ordered by their respective eigenvalues. Note that $\boldsymbol{\Sigma}(\xi)$ is Hermitian, hence by the Spectral Theorem, the eigenvalues are real numbers. We specify the eigenvalue basis, which is used to apply $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ by adding it to the right as follows: $\boldsymbol{\Phi}(\mathbf{X}(t); \mathbf{V}(t))$ and $\boldsymbol{\Psi}(\mathbf{X}(t); \mathbf{V}(t))$. Then, the z-scores for $\widehat{\mathbf{X}}(\xi)$ are given for $j = 1, \dots, k$ by

$$\begin{aligned} \widehat{\mathbf{Z}}_j(\xi) := \Phi_j(\widehat{\mathbf{X}}(\xi); \widehat{\mathbf{V}}(\xi)) &= \sum_{i=1}^{p} \overline{\widehat{v}_{ij}(\xi)} \widehat{X}_i(\xi) = \sum_{i=1}^{p} \mathcal{F}\{\overline{v}_{ij}^{\text{rev}} * X_i(t)\}(\xi) \\ &= \mathcal{F}\{\Psi_j(X(t); \mathbf{V}(t))\}(\xi) \end{aligned}$$

The first step is due to $\mathcal{F}^{-1}\left\{\overline{\widehat{v}_{ij}(\xi)}\right\}(t) = \overline{v}_{ij}(-t) = \overline{v}_{ij}^{\text{rev}}$. Therefore, $\mathbf{Z}(t) = \Psi(\mathbf{X}(t); \mathbf{V}(t))$. Next, the reconstruction step for $i = 1, \dots, p$ is given by

$$\begin{aligned} \widetilde{\widehat{X}_i}(\xi) := \Phi_i^*(\widehat{\mathbf{Z}}(\xi); \widehat{\mathbf{V}}(\xi)) = \sum_{j=1}^{k} \widehat{v}_{ij}(\xi)\widehat{Z}_j(\xi) &= \sum_{j=1}^{k} \mathcal{F}\{(v_{ij} * Z_j)(t)\}(\xi) = \mathcal{F}\{\Psi_i^*(\mathbf{Z}(t); \mathbf{V}(t))\} \\ &= \mathcal{F}\{\Psi_i^*(\boldsymbol{\Psi}(\mathbf{X}(t); \mathbf{V}(t)); \mathbf{V}(t))\}(\xi) = \mathcal{F}\{(\Psi_i^* \circ \boldsymbol{\Psi})(\mathbf{X}(t); \mathbf{V}(t))\}(\xi) \end{aligned}$$

Let us define,

$$\widetilde{\mathbf{X}}(t) := (\boldsymbol{\Psi}^* \circ \boldsymbol{\Psi})(\mathbf{X}(t); \mathbf{V}(t))$$

The following diagram shows the two ways to calculate the same thing.

$$\begin{tikzcd}
\mathbf{X}(t) \arrow[r, "\boldsymbol{\Psi}(\;\cdot\;; \mathbf{V}(t))"] \arrow[d, "\mathcal{F}"'] & \mathbf{Z}(t) \arrow[r, "\boldsymbol{\Psi}^*(\;\cdot\;; \mathbf{V}(t))"] & \widetilde{\mathbf{X}}(t) \\
\widehat{\mathbf{X}}(\xi) \arrow[r, "\boldsymbol{\Phi}(\;\cdot\;; \widehat{\mathbf{V}}(\xi))"'] & \widehat{\mathbf{Z}}(\xi) \arrow[r, "\boldsymbol{\Phi}^*(\;\cdot\;; \widehat{\mathbf{V}}(\xi))"'] & \widetilde{\widehat{\mathbf{X}}}(\xi) \arrow[u, "\mathcal{F}^{-1}"]
\end{tikzcd}$$

Then, due to the Plancherel Theorem, we obtain

$$\mathbb{E}\|\widehat{\mathbf{X}}(\xi) - \widetilde{\widehat{\mathbf{X}}}(\xi)\|_{L^2}^2 = \mathbb{E}\|\mathcal{F}\{\mathbf{X}(t) - \widetilde{\mathbf{X}}(t)\}\|_{L^2}^2 = \mathbb{E}\|\mathbf{X}(t) - \widetilde{\mathbf{X}}(t)\|_{L^2}^2$$

Hence, we approximate $\widetilde{\widehat{\mathbf{X}}}(\xi) \approx \widehat{\mathbf{X}}(\xi)$ well if and only if $\widetilde{\mathbf{X}}(t) \approx \mathbf{X}(t)$ is well approximated as well. And, since $\widetilde{\widehat{\mathbf{X}}}(\xi) \approx \widehat{\mathbf{X}}(\xi)$ in the best linear approximation by Theorem 2, we expect $\widetilde{\mathbf{X}}(t)$ to be a good approximation as well.

4 Kernel Functional PCA

Let us consider $x_1(t), \dots, x_n(t) \in \mathcal{X}$ samples of our data where the input space $\mathcal{X} = L^2(\mathbb{R})$. Henceforward, we will denote them simply as $x_1, \dots, x_n$, taking for granted that they depend on time. We want to map our input space to a usually higher-dimensional space, i.e, $\mathbb{K}-$Hilbert space $\mathcal{H}$, often referred to as *feature space*. We will consider the Hilbert space over the field $\mathbb{C}$ or $\mathbb{R}$. The map $\varphi : \mathcal{X} \to \mathcal{H}$ is usually called the *feature map*. Our goal is to avoid using φ and instead do the operations using a *kernel* $K : \mathcal{X} \times \mathcal{X} \to \mathbb{K}$ even without knowing explicitly φ. Let us introduce a couple of definitions.

Definition 2 (Positive definite function)**.** A function $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$ where $\mathcal{Y}$ is a complex Hilbert space is called a *positive definite function* if for all $n \in \mathbb{N}$, $\alpha_1, \dots, \alpha_n \in \mathbb{C}$ and $x_1, \dots, x_n \in \mathcal{X}$

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \langle K(x_i, x_j) y_j, y_i \rangle \geq 0$$

We will only consider the case where $\mathcal{Y} = \mathbb{C}$ and therefore $\mathcal{L}(\mathcal{Y}) = \mathbb{C}$.

Firstly, let us define a *reproducing kernel*.

Definition 3 (Reproducing Kernel)**.** A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{C}$ is a *reproducing kernel* of a Hilbert space $\mathcal{H}$ if and only if

(i) $K(\cdot, x) \in \mathcal{H} \qquad \forall x \in \mathcal{X}$

(ii) $\langle \varphi, K(\cdot, x) \rangle = \varphi(x) \qquad \forall x \in \mathcal{X} \quad \forall \varphi \in \mathcal{H}$ (*Reproducing property*)

To characterize the space of RKHS we will usually use the following theorem.

Theorem 3 (Moure-Aronszajn Theorem)**.** Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{C}$ is a positive definite function. Then, there exists a unique Hilbert space $\mathcal{H}$ of functions on $\mathcal{X}$. Moreover, the subspace $\mathcal{H}_0 \subseteq \mathcal{H}$ spanned by $\{K(\cdot, x)\}_{x \in \mathcal{X}}$ is dense in $\mathcal{H}$ and $\mathcal{H}$ is the set of functions on $\mathcal{X}$ which are pointwise limits of Cauchy sequences in $\mathcal{H}_0$ with the inner product

$$\langle f, g \rangle_{\mathcal{H}_0} := \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \bar{\beta}_j K(y_j, x_i)$$

where $f = \sum_{i=1}^{n} \alpha_i K(\cdot, x_i)$ and $g = \sum_{j=1}^{m} \beta_j K(\cdot, y_j)$.

Another useful characterization of an RKHS is given by the following result. This result gives us the necessary tools to build the Kernel Functional PCA.

Lemma 2. Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{C}$ is a reproducing kernel if and only if there exists a mapping $\varphi : \mathcal{X} \to \ell^2(E)$ such that

$$K(x, y) = \langle \varphi(x), \varphi(y) \rangle_{\ell^2(E)} = \sum_{\alpha \in E} (\varphi(x))_\alpha (\overline{\varphi(y)})_\alpha$$

4.1 Kernel PCA

Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive definite kernel and $\mathcal{H}$ be its RKHS. Due to Lemma 2, we assume that there exists a mapping $\varphi : \mathcal{X} \to \ell^2(E) \equiv \mathcal{H}$ such that $K(x, y) = \langle \varphi(x), \varphi(y) \rangle_{\ell^2(E)}$. Let us assume that the data is centered, i.e., $\frac{1}{n} \sum_{i=1}^{n} \varphi(x_i) = 0$. Otherwise, we can center the kernel using the following transformation $K \leftarrow K - \mathbf{1}_n K - K\mathbf{1}_n + \mathbf{1}_n K \mathbf{1}_n$ where $\mathbf{1}_n$ is the matrix filled with $1/n$ entries.

The orthogonal projection into a direction $f \in \mathcal{H}$ is the function $h_f : \mathcal{X} \to \mathbb{C}$ defined by

$$h_f(x) = \left\langle \varphi(x), \frac{f}{\|f\|_{\mathcal{H}}} \right\rangle_{\mathcal{H}}$$

Using the fact that $\varphi(x) = K(\cdot, x)$, then for all $f \in \mathcal{H}$ we have that, by the reproducing property, $\langle \varphi(x), f \rangle = f(x)$. Therefore, the empirical variance $\widehat{\mathrm{Var}}(h_f)$ can be expressed by

$$\widehat{\mathrm{Var}}(h_f) = \frac{1}{n} \sum_{i=1}^{n} \frac{\langle \varphi(x_i), f \rangle_{\mathcal{H}}^2}{\|f\|_{\mathcal{H}}^2} = \frac{1}{n} \sum_{i=1}^{n} \frac{f(x_i)^2}{\|f\|_{\mathcal{H}}^2}$$

Then, our goal is to find the variance of the i-th principal direction $f_j \in \mathcal{H}$ such that

$$f_j \in \underset{f \perp \{f_1, \ldots, f_{j-1}\}}{\arg\max}\ \widehat{\mathrm{Var}}(h_f) \qquad \text{where } \|f\|_{\mathcal{H}} = 1 \tag{6}$$

Due to the Representer theorem, we know that there exists $f_j \in \mathcal{H}$ that belongs to (6) for each $j = 1, 2, \ldots, n$ and admits a representation of the form

$$f_j = \sum_{k=1}^{n} \alpha_{j,k} K(\cdot, x_k) \tag{7}$$

with $\boldsymbol{\alpha}_j = (\alpha_{j,1}, \ldots, \alpha_{j,n})^\top \in \mathbb{R}^n$. Let us define the matrix $\mathbf{K} := \{K(x_i, x_j)\}_{i,j=1}^n \in \mathbb{C}^{n \times n}$ that is Hermitian and positive definite. On the other hand, we can prove for each f_j the following properties

$$\|f_j\|_{\mathcal{H}}^2 = \sum_{r,s=1}^{n} \bar{\alpha}_{j,r} \alpha_{j,s} K(x_r, x_s) = \boldsymbol{\alpha}_j^\dagger \mathbf{K} \boldsymbol{\alpha}_j$$

$$\sum_{k=1}^{n} f_j(x_k)^2 = \boldsymbol{\alpha}_j^\dagger \mathbf{K}^\dagger \mathbf{K} \boldsymbol{\alpha}_j = \boldsymbol{\alpha}_j^\dagger \mathbf{K}^2 \boldsymbol{\alpha}_j \tag{8}$$

$$\langle f_i, f_j \rangle_{\mathcal{H}} = \boldsymbol{\alpha}_i^\dagger \mathbf{K} \boldsymbol{\alpha}_j \tag{9}$$

The optimization problem (6) can be equivalently expressed using (8) and (9) as follows

$$\begin{aligned} &\underset{\boldsymbol{\alpha} \in \mathbb{R}^n}{\arg\max}\, \boldsymbol{\alpha}^\dagger \mathbf{K}^2 \boldsymbol{\alpha} \\ \text{such that} \quad &\boldsymbol{\alpha}_i^\top \mathbf{K} \boldsymbol{\alpha} = 0 \qquad i = 1, \ldots, j-1 \\ &\boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha} = 1 \end{aligned}$$

This problem is similar to the usual Principal Component Analysis. In fact, using the eigendecomposition of $\mathbf{K}$, i.e., $\mathbf{K} = \mathbf{V} \boldsymbol{\Lambda} \mathbf{V}^\dagger$, we can define $\boldsymbol{\beta} := \mathbf{K}^{1/2} \boldsymbol{\alpha}$ and solve the PCA problem

$$\begin{aligned} &\underset{\boldsymbol{\beta} \in \mathbb{R}^n}{\arg\max}\, \boldsymbol{\beta}^\dagger \mathbf{K} \boldsymbol{\beta} \\ \text{such that} \quad &\boldsymbol{\beta}_i^\dagger \boldsymbol{\beta} = 0 \quad i = 1, \ldots, j-1 \\ &\boldsymbol{\beta}^\dagger \boldsymbol{\beta} = 1 \end{aligned}$$

	Finite-dimensional	**Infinite-dimensional**
Exponential kernel	$\exp(\langle x, x' \rangle_2)$	$\exp\left(\langle x, x' \rangle_{\ell^2(\mathbb{Z})}\right)$
Gaussian RBF kernel	$\exp\left(-\sigma^{-2}\|x - x'\|_2^2\right)$	$\exp\left(-\sigma^{-2}\|x - x'\|_{L^2(\mathbb{R})}^2\right)$
Polynomial kernel	$(\langle x, x' \rangle_2)^p$	$\left(\langle x, x' \rangle_{H^k(\Omega)}\right)^p$

Table 3: Examples of functional reproducing kernels $K(x, x')$.

Therefore, a solution is given by $\boldsymbol{\alpha}_j = \mathbf{K}^{-1/2}\boldsymbol{\beta}_j = \frac{1}{\sqrt{\lambda_j}}\boldsymbol{v}_j$ where $\boldsymbol{v}_j$ is the $j-$th eigenvector of $\mathbf{K}$ and λ_j its eigenvalue. On the other hand, using (7), the z-scores are given by

$$f_j(x) = \frac{1}{\sqrt{\lambda_j}} \sum_{k=1}^{n} v_{j,k} K(x, x_k) \qquad \forall x \in \mathcal{X}$$

4.2 Functional Kernels

Next, we will consider a variation of some of the most commonly used kernels for functional spaces.

Example 1 (*Polynomial kernels*)**.** Let us consider the case where $\mathcal{X} = L^2([-\pi, \pi], \mathbb{R})$ and the family of kernels given by

$$K(x, x') := (\langle x, x' \rangle_{L^2(-\pi,\pi)})^d$$

where $d \geq 1$. Firstly, let us prove that $K(x, x) = \langle x, x \rangle_{L^2(-\pi,\pi)}$ is a reproducing kernel. We consider the feature map $\varphi : \mathcal{X} \to \ell^2(\mathbb{Z})$ given by $x \mapsto \{\hat{x}(m)\}_{m \in \mathbb{Z}}$ where $\hat{x}(m) := \frac{1}{2\pi}\int_{-\pi}^{\pi} x(t)e^{-int}dt$. Therefore, by the Plancherel identity, we obtain

$$\langle \varphi(x), \varphi(y) \rangle_{\ell^2(\mathbb{Z})} = \sum_{m \in \mathbb{Z}} \hat{x}(m)\overline{\hat{y}(m)} = \langle x, y \rangle_{L^2(-\pi,\pi)} = K(x, y)$$

Hence, Lemma 2 assures us that $K(x, y)$ is an RKHS. Let $f(z) = z^d$ be a complex holomorphic function; then Lemma 4.7 in (Christmann et al. 2008) assures us that the composition of a kernel with f must be a kernel.

Example 2 (Gaussian Radial Basis Function)**.** Consider again the function space $\mathcal{X} = C([-\pi, \pi], \mathbb{R})$. The *Gaussian radial basis function* can be defined as follows

$$K(x, y) = \exp\{-\|x - y\|_{L^2(-\pi,\pi)}^2/(2\sigma^2)\}$$

where $\sigma^2 > 0$. Decomposing $K(x, y)$ into

$$K(x, y) = \frac{\exp\{\sigma^{-2}\langle x, y \rangle_{L^2(-\pi,\pi)}\}}{\exp\{\sigma^{-2}/2\langle x, x \rangle_{L^2(-\pi,\pi)}\}\exp\{\sigma^{-2}/2\langle y, y \rangle_{L^2(-\pi,\pi)}\}} \tag{10}$$

We know that the product of kernels is a kernel by Lemma 4.6 in (Christmann et al. 2008). We already know that the numerator is a kernel by the previous exercise. Moreover, Lemma 4.8 in (Christmann et al. 2008) assures us that the denominator is a positive definite kernel because we can find a feature map $\varphi_2 : x \mapsto \exp\{-\sigma^{-2}/2\langle x, x \rangle_{L^2(-\pi,\pi)}\}$.

4.3 Universality of Kernels

Let us introduce a new concept that will be useful to characterize the *size* of the Hilbert space spanned by each kernel. But first let us introduce the notation for different kernels.

Definition 4. A reproducing kernel $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$ is called

(i) *Mercer* provided that $\mathcal{H}_K \subseteq C(\mathcal{X}, \mathcal{Y})$.

(ii) C_0-*Kernel* provided that $\mathcal{H}_K \subseteq C_0(\mathcal{X}, \mathcal{Y})$.

We will use the C_0-kernel in the case of using L^p families because given a probability measure μ, $C_0(\mathcal{X}, \mathcal{Y}) \subseteq L^p(\mathcal{X}, \mu; \mathcal{Y}) \subseteq L^q(\mathcal{X}, \mu; \mathcal{Y})$ for all $1 \leq p < q \leq \infty$.

Definition 5. Let $K : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{Y})$ be a reproducing kernel.

(i) Mercer kernel is called *compact universal* if $\mathcal{H}_K$ is dense in $L^2(\mathcal{X}, \mu; \mathcal{Y})$ for each probability measure μ with compact support.

(ii) C_0-Kernel K is called *universal* if $\mathcal{H}_K$ is dense in $L^2(\mathcal{X}, \mu, \mathcal{Y})$ for each probability measure μ.

In practice, *(ii)* assumption is equivalent to $\mathcal{H}_K$ dense in $C_0(\mathcal{X}, \mathcal{Y})$ by the following Theorem 1 described in (Carmeli et al. 2008).

Theorem 4. Suppose K is a $\mathcal{C}_0$-kernel. The following facts are equivalent.

(i) The kernel K is *universal.*

(ii) The space $\mathcal{H}_K$ is dense in $\mathcal{C}_0(X; \mathcal{Y})$.

(iii) There is $1 \leq p < \infty$ such that $\mathcal{H}_K$ is dense in $L^p(\mathcal{X}, \mu; \mathcal{Y})$ for all probability measures μ on $\mathcal{X}$.

For the next result, we need some preliminary facts. First of all, K is a Mercer kernel and μ is a probability measure on $\mathcal{X}$, the space $\mathcal{H}_K$ is a subspace of $L^2(\mathcal{X}, \mu; \mathcal{Y})$, provided that $\|K(x, x)\|$ is bounded on the support of μ. This last condition is always satisfied if μ has compact support. If $\mathcal{H}_K$ is a subspace of $L^2(\mathcal{X}, \mu; \mathcal{Y})$, we denote the canonical inclusion by

$$i_\mu : \mathcal{H}_K \hookrightarrow L^2(\mathcal{X}, \mu; \mathcal{Y}).$$

Proposition 3. Let K be a Mercer kernel and μ a probability measure such that K is bounded on the support of μ. The inclusion i_μ is a bounded operator, its adjoint $i_\mu^* \colon L^2(\mathcal{X}, \mu; \mathcal{Y}) \longrightarrow \mathcal{H}_K$ is given by

$$(i_\mu^* f)(x) = \int_X K(x, y) f(y) \, d\mu(y), \tag{11}$$

where the integral converges in norm, and the composition $i_\mu i_\mu^* = L_\mu$ is the integral operator on $L^2(\mathcal{X}, \mu; \mathcal{Y})$ with kernel K

$$(L_\mu f)(x) = \int_X K(x, y) f(y) \, d\mu(y).$$

In particular, if $K(x, x)$ is a compact operator for all $x \in \mathcal{X}$, then L_K is a compact operator.

Finally, we arrive at a theorem that can be used to verify that a Kernel is universal.

Theorem 5. Suppose K is a $\mathcal{C}_0$-kernel. Then the following facts are equivalent.

(i) The kernel K is *universal.*

(ii) The operator $i_\mu^* : L^2(\mathcal{X}, \mu; \mathcal{Y}) \to \mathcal{H}_K$ is an injective operator for all probability measures μ on $\mathcal{X}$.

(iii) The integral operator $L_\mu : L^2(\mathcal{X}, \mu; \mathcal{Y}) \to L^2(\mathcal{X}, \mu; \mathcal{Y})$ is injective for all probability measures μ on $\mathcal{X}$.

The following result in (Christmann et al. 2008) is the one I want to generalize.

Theorem 6 (Universality of Gaussian Kernel)**.** Let μ be a finite measure on $\mathbb{R}^d$ or a Lebesgue measure on $\mathbb{R}^d$, and $p \in (1, \infty)$. Moreover, let K be the Gaussian RBF kernel with $\sigma > 0$. Then the operator $i_\mu^* : L^p(\mathbb{R}^d, \mu) \to \mathcal{H}_K$ defined by (11) is injective.

Therefore, the Gaussian RBF Kernel is universal in a finite-dimensional space.

5 Experiments

5.1 Correlated $n-$dimensional Brownian Motion

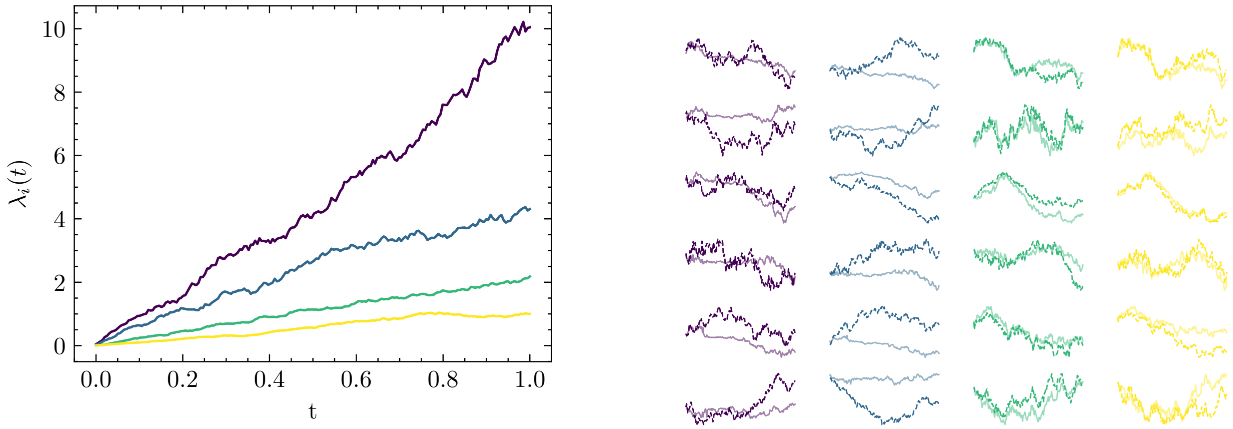
We simulate a multidimensional Brownian motion with covariance given by the product of $\Sigma = CC^T$ (it needs to be symmetric) where C is a random matrix. Then, the simulated Brownian motion is given by

$$X_{t+1} = X_t + \sqrt{\Delta t}\mathcal{N}(0, \mathbf{\Sigma})$$

For this experiment, we use 5 dimensions, i.e., $p = 5$, and our goal is to compress each observation into one ($k = 1$). The signals are correlated, so it should be possible with great error despite the noisy signal. The results for 100 experiments can be found in Figure 1. Observe that both algorithms perform identically in this regime. This is an interesting result and may require further consideration.

	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$
FPCA	1.03 ± 0.28	0.42 ± 0.17	0.10 ± 0.07	0.01 ± 0.02	0.00
Fourier PCA	1.03 ± 0.28	0.42 ± 0.17	0.10 ± 0.07	0.01 ± 0.02	0.00

Table 4: Mean Square Error for different values of k.



(a) Eigenvalues for the 5 different signals. The eigenvalues grows linearly as the variance does.

(b) Reconstruction error for $k = 1$ and $p = 5$. The real Brownian Motion is the dashed line, and the reconstructed is the diffused continuous line.

Figure 1: Experiment with the n-dimensional Brownian Motion.

5.2 Anomaly Detection for satellite telemetry

ESA released a large-scale, real-life satellite telemetry anomaly dataset[2]. Solar arrays power regulators switch off, video processing unit reset, attitude disturbances...Throughout its life after launch, a spacecraft is subject to several unexpected behaviours that can lead to loss of scientific data, inadequate performance, and sometimes triggering of spacecraft safe mode, which typically entails a challenging recovery. The dataset, derived from three missions and totaling 31 GB, is curated

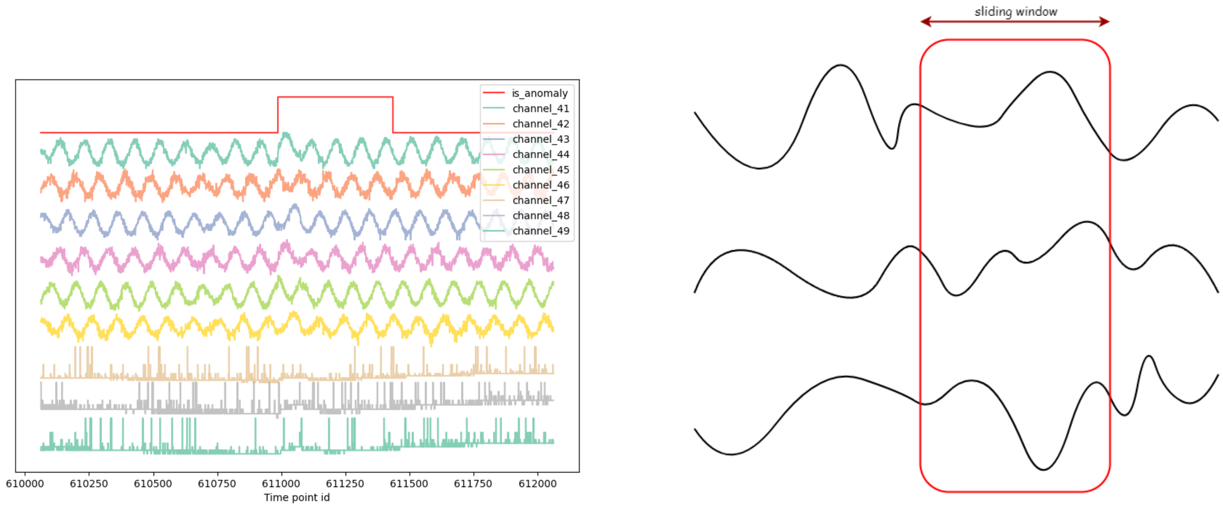
[2] `https://esoc.esa.int/esa-releases-building-block-open-database-satellite-anomalies`

and annotated to aid the development of AI models for anomaly detection. It consists of over 50 channels and 15 telecommands with over 50M data instances. Figure 2 shows a sample of the data within the spacecraft dataset. We have discarded those channels that spark and contain discontinuities and do not have a functional structure. The way to process the signal is to create a sliding window to process the data. This allows multiple samples of the data and makes it computationally feasible. However, we need to configure the sliding windows length parameter.

	Mean	Median
FPCA	0.082	0.084
Fourier PCA	0.065	0.053

Table 5: Mean and Median values for FPCA and Fourier PCA

The way we detect the anomalies is when the reconstruction error output by the PCA is above a certain threshold. For example, if it surpasses the 95% quantile errors. We notice in Figure 3 that most of the anomaly spikes an error in the reconstruction error as we expect.



(a) Dataset sample of the input data for different channels and anomaly ground truth for each timestep.

(b) The sliding window approach is widely used in signal processing for continuous data streams.

Figure 2: Dataset sample and sliding window used for processing.

5.3 Kernel FPCA vs FPCA

We create a set of functions that we are not able to separate using any linear projection; hence, Functional PCA is not able to separate both classes. We create three different Gaussian process distributions. We use a periodic covariance, specifically, we use the Exponential Sine Square kernel that is given by

$$k(x, x') = \sigma^2 \exp\left(-\frac{2}{\ell^2} \sin^2\left(\pi \frac{|x - x'|}{p}\right)\right)$$

where σ^2 is the overall variance (σ is also known as amplitude), ℓ is the lengthscale, and p the period, which is the distance between repetitions. We choose $p = 1$ and we subtract the resulting

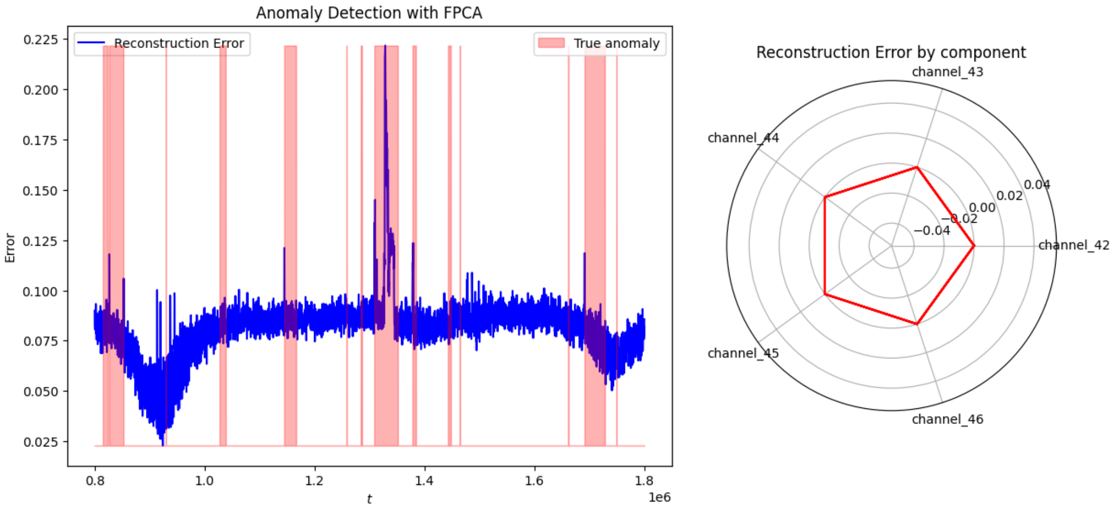


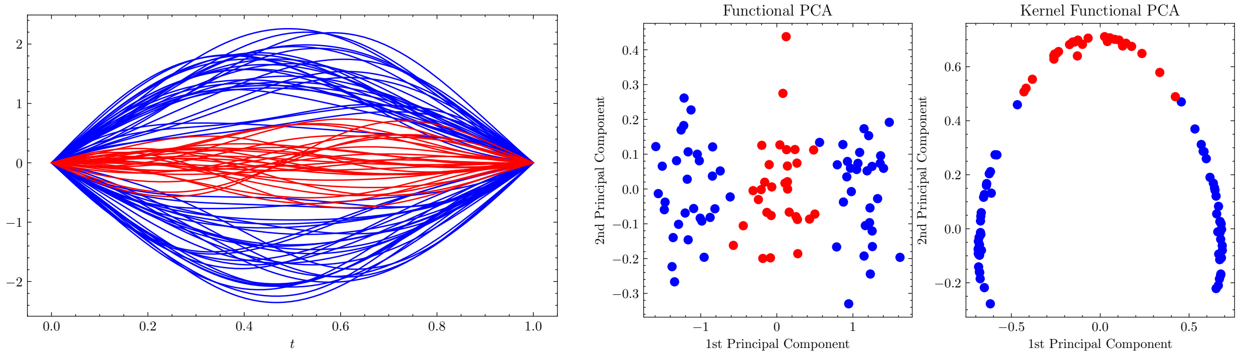
Figure 3: Comparison between reconstruction error and the real true anomaly. The right figure should show the individual reconstruction errors that contribute to each channel.

sampling to the first value so that they start at the same value. We create the first class with mean $\mu(t) = \pm\sin(t\pi)$ and the second class with zero mean (Figure 4a). Moreover, we use the Kernel of the RBF for the PCA kernel, i.e.,

$$K(x, y) := \exp\{-\|x - y\|^2_{\mathcal{H}}/(2\sigma^2)\}$$

where we use $\sigma^2 = 1$ as the kernel.

Finally, we see that using only the 2nd principal component we are able to separate both classes (4b). We can see that, alternatively, the standard Functional PCA is not able to separate both categories of functions. In fact, it can be proven that using a single principal component, we are not able to separate the classes.



(a) Dataset created using a periodic exponential sinusoidal kernel for the Gaussian process.

(b) Reduction of the functional data in the first 2 principal components for FPCA and kFPCA.

Figure 4: Comparison between kFPCA and FPCA.

A Appendix

Proof of Theorem 1. Notice that A is compact. Let $S' = \overline{\text{span}}\{u_k, u_{k+1}, \dots\}$. The subspace S' has codimension $k-1$. Therefore, $S' \cap S_k$ has strictly positive dimension and thus, $\exists x \in S' \cap S_k$ with $\|x\| = 1$. Since $x \in S'$, $\langle Ax, x\rangle \leq \lambda_k$. Therefore,

$$\inf_{x \in S_k, \|x\|=1} \langle Ax, x\rangle \leq \lambda_k.$$

But A is compact, therefore $f : x \mapsto \langle Ax, x\rangle$ is weakly continuous. Furthermore, for any bounded set in $\mathcal{H}$ is weakly compact. This lets us replace the infimum by the minimum. So

$$\sup_{S_k} \min_{x \in S_k,\, \|x\|=1} \langle Ax, x\rangle \leq \lambda_k$$

The equality is achieved when $S_k := \text{span}\{u_1, \dots, u_k\}$,

$$\max_{S_k} \min_{x \in S_k,\, \|x\|=1} \langle Ax, x\rangle = \lambda_k.$$

Analagously, consider a $(k-1)-$dimensional subspace S_{k-1}, whose orthogonal complement is denoted by $S_{k-1}^{\perp}$. If $S' = \text{span}\{u_1, \dots, u_k\}$, $S' \cap S_{k-1}^{\perp} \neq 0$. So, there exists $x \in S_{k-1}^{\perp}$ with $\|x\| = 1$ such that $\langle Ax, x\rangle \geq \lambda_k$. This implies

$$\max_{S_{k-1}^{\perp},\, \|x\|=1} \langle Ax, x\rangle \geq \lambda_k$$

Similarly, the compactness of A is applied. Therefore,

$$\inf_{S_{k-1}} \max_{x \in S_{k-1}^{\perp},\, \|x\|=1} \geq \lambda_k.$$

The infimum is achieved for $S_{k-1} := \text{span}\{u_1, \dots, u_{k-1}\}$ and we deduce

$$\min_{S_{k-1}} \max_{x \in S_{k-1}^{\perp},\, \|x\|=1} \langle Ax, x\rangle = \lambda_k.$$

□

Proof of Theorem 2. Using Fatou's Lemma, we can transform again the infinite dimensional problem in the well-known discrete case

$$\begin{aligned}\inf_{\mathbf{V}(t)^\top \mathbf{V}(t) = I} \mathbb{E} \int \|\mathbf{X}(t) - \mathbf{V}(t)\mathbf{V}(t)^\top \mathbf{X}(t)\|_2^2 dt &= \inf_{\mathbf{V}(t)^\top \mathbf{V}(t) = I} \int \mathbb{E}\|\mathbf{X}(t) - \mathbf{V}(t)\mathbf{V}(t)'\mathbf{X}(t)\|_2^2 dt \\ &\geq \int \inf_{\mathbf{V}^\top \mathbf{V} = I} \mathbb{E}\|\mathbf{X}(t) - \mathbf{V}\mathbf{V}^\top \mathbf{X}(t)\|_2^2 dt\end{aligned}$$

We claim that the best rank$-k$ approximation to $\mathbf{X}(t)$ in the spectral norm.

$$\mathbb{E}\|\mathbf{X}(t) - \mathbf{V}\mathbf{V}^\top \mathbf{X}(t)\|_2^2 = \mathbb{E}[\mathbf{X}(t)^\top (\mathbf{I} - \mathbf{V}\mathbf{V}^\top)^\top (\mathbf{I} - \mathbf{V}\mathbf{V}^\top)\mathbf{X}(t)]$$

Let us denote $\mathbf{M} := (\mathbf{I} - \mathbf{V}\mathbf{V}^\top)^2$ and, owing to its symmetry, we can calculate its eigen-decomposition, i.e., $\mathbf{M} = \tilde{\mathbf{Q}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{Q}}'$. Let us denote $\mathbf{Y}(t) := \tilde{\mathbf{Q}}^\top \mathbf{X}(t)$. Then,

$$\begin{aligned}\mathbb{E}\|\mathbf{X}(t) - \mathbf{V}\mathbf{V}^\top \mathbf{X}(t)\|_2^2 &= \mathbb{E}[\mathbf{X}(t)^\top \mathbf{M}\mathbf{X}(t)] = \mathbb{E}[\mathbf{Y}(t)^\top \tilde{\mathbf{\Lambda}}\mathbf{Y}(t)] \\ &= \sum_i \tilde{\lambda}_i \mathbb{E}[Y_i(t)^2] = \sum_i \tilde{\lambda}_i \mathbb{E}[\tilde{\mathbf{Q}}^\top \mathbf{X}(t)\mathbf{X}(t)^\top \tilde{\mathbf{Q}}] = \sum_i \tilde{\lambda}_i \tilde{\mathbf{Q}}^\top \mathbf{\Sigma}(t)\tilde{\mathbf{Q}} \\ &= \text{tr}(\tilde{\mathbf{\Lambda}}\tilde{\mathbf{Q}}^\top \mathbf{\Sigma}(t)\tilde{\mathbf{Q}}) = \text{tr}(\tilde{\mathbf{Q}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{Q}}^\top \mathbf{\Sigma}(t)) = \text{tr}(\mathbf{M}\mathbf{\Sigma}(t))\end{aligned}$$

We know that $\mathbf{\Sigma}(t) = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ is positive definite symmetric, then there exists $\mathbf{\Sigma}(t)^{1/2} = \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{U}^\top$. And, we know that $\|\mathbf{A}\|_F := \sqrt{\mathrm{tr}(\mathbf{A}^\top\mathbf{A})}$. Therefore,

$$\mathrm{tr}(\mathbf{M}\mathbf{\Sigma}(t)) = \|(\mathbf{I} - \mathbf{V}\mathbf{V}^\top)\mathbf{\Sigma}^{1/2}(t)\|_F^2 = \|\mathbf{\Sigma}^{1/2}(t) - \mathbf{V}\mathbf{V}^\top\mathbf{\Sigma}^{1/2}(t)\|_F^2$$

By the Low-rank approximation theorem, the Frobenius norm of $\|\mathbf{\Sigma}^{1/2}(t) - \tilde{\mathbf{A}}\|_F^2$ is minimized when $\tilde{\mathbf{A}} = \mathbf{U}_k\mathbf{\Lambda}_k\mathbf{U}_k^\top$ where $\mathbf{U}_k$ are the first k columns of $\mathbf{U}$ where is given by the eigen-decomposition $\mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{U}^\top = \mathbf{\Sigma}^{1/2}(t)$. And, the error is given by $\sum_{i=k+1}^n \lambda_i$. We can prove that the equation

$$\mathbf{V}\mathbf{V}^\top\mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{U}^\top = \mathbf{U}_k\mathbf{\Lambda}_k\mathbf{U}_k^\top$$

is satisfied when $\mathbf{V} = \mathbf{U}_k$, which completes the proof.

Notice that the expected total square error must be

$$\mathbb{E}\|\mathbf{X}(t) - \mathbf{V}(t)\mathbf{V}(t)^\top\mathbf{X}(t)\|_2^2 = \sum_{i=k+1}^{p} \int \lambda_i(t)dt$$

□

Proof of Proposition 2. The characteristic polynomial of $\mathbf{\Sigma}(t)$ is

$$\mathfrak{F}(t,\lambda) := \det(\mathbf{\Sigma}(t) - \lambda\mathbf{I})$$

a polynomial of degree n in λ whose coefficients are differentiable functions of t, i.e.,

$$\mathfrak{F}(t,\lambda) = a_n(t)\lambda^n + \cdots + a_1(t)\lambda + a_0(t)$$

where $a_i(t) \in \mathcal{C}^\infty(\mathbb{R},\mathbb{R})$ for $i = 0,\ldots,n$. The assumption that at t^* every root λ_i is a simple root of $\mathbf{\Sigma}(t^*)$ means that

$$\mathfrak{F}(t^*,\lambda_i) = 0, \qquad \frac{\partial}{\partial\lambda}\mathfrak{F}(t^*,\lambda)|_{\lambda=\lambda_i} \neq 0$$

Then, according to the Implicit Function Theorem, under these conditions, the equation $\mathfrak{F}(t,\lambda) = 0$ has a solution $\lambda = \lambda_i(t)$ in a neighborhood of $t = t^*$ that depends differentiably on t. Moreover,

$$\mathfrak{F}_t(t,\lambda(t)) + \dot{\lambda}(t)\mathfrak{F}_\lambda(t,\lambda(t)) = 0$$

Invoking Theorem 8, Chapter 9 in Lax (Lax et al. 2007), for every eigenvector $\mathbf{v}_i(t^*)$ with multiplicity 1, we can choose an eigenvector $\mathbf{v}_i(t)$ of $\mathbf{\Sigma}(t)$ pertaining to the eigenvalue $\lambda_i(t)$ to depend differentiably on t. Moreover, we can assume that $\|\mathbf{v}_i(t)\| = 1$ without loss of generality. The eigenvector equations give us two equivalent formulations

$$\mathbf{\Sigma}(t)\mathbf{v}_k(t) = \lambda_k(t)\mathbf{v}_k(t) \Leftrightarrow \mathbf{v}_k^*(t)\mathbf{\Sigma}(t) = \lambda_k(t)\mathbf{v}_k^*(t) \tag{12}$$

where every element is differentiable. Therefore, differentiating and using the product rule to obtain

$$\dot{\mathbf{\Sigma}}(t)\mathbf{v}_k(t) + \mathbf{\Sigma}(t)\dot{\mathbf{v}}_k(t) = \dot{\lambda}_k(t)\mathbf{v}_k(t) + \lambda_k(t)\dot{\mathbf{v}}_k(t) \tag{13}$$

Left multiplying (13) by the vector $\mathbf{v}_k^*(t)$ and using (12), we cancel some terms and obtain

$$\dot{\lambda}_k(t) = \mathbf{v}_k^*(t)\dot{\mathbf{\Sigma}}(t)\mathbf{v}_k(t) \tag{14}$$

This formula is usually called the *Hadamard first variation formula*. On the other hand, if we left multiply (13) by any other vector $\mathbf{v}_j^*(t)$ for $j \neq k$, we obtain

$$\mathbf{v}_j^*(t)\dot{\mathbf{\Sigma}}(t)\mathbf{v}_k(t) = (\lambda_k(t) - \lambda_j(t))\mathbf{v}_j^*(t)\dot{\mathbf{v}}_k(t) \Rightarrow \mathbf{v}_j^*(t)\dot{\mathbf{v}}_k(t) = \frac{\mathbf{v}_j^*(t)\dot{\mathbf{\Sigma}}(t)\mathbf{v}_k(t)}{\lambda_k(t) - \lambda_j(t)}$$

Using the fact that $\mathbf{v}_j^*(t)$ is a basis, i.e., $\dot{\mathbf{v}}_k(t) = \sum_{j=1}^{p}(\mathbf{v}_j^*(t)\dot{\mathbf{v}}_k(t))\mathbf{v}_j(t)$ for all $\mathbf{v}(t)$. Moreover, since $\|\mathbf{v}_k(t)\| = 1$, then $\dot{\mathbf{v}}_k(t) \perp \mathbf{v}_k(t)$. Hence,

$$\begin{aligned}\dot{\mathbf{v}}_k(t) = \sum_j(\mathbf{v}_j^*(t)\dot{\mathbf{v}}_k(t))\mathbf{v}_j(t) &= \sum_{j\neq k}\frac{\mathbf{v}_j^*(t)\dot{\mathbf{\Sigma}}(t)\mathbf{v}_k(t)}{\lambda_k - \lambda_j}\mathbf{v}_j(t) + \mathbf{v}_k^*(t)\dot{\mathbf{v}}_k(t)\mathbf{v}_k(t)\\ &= \sum_{j\neq k}\frac{\mathbf{v}_j^*(t)\dot{\mathbf{\Sigma}}(t)\mathbf{v}_j(t)}{\lambda_k(t) - \lambda_j(t)}\mathbf{v}_k(t) \qquad (15)\end{aligned}$$

We call this equation the *first variation formula for eigenvectors*. Analogously, we obtain an equivalent formulation for the conjugates

$$\dot{\mathbf{v}}_k^*(t) = \sum_{j\neq k}\frac{\mathbf{v}_k^*(t)\dot{\mathbf{\Sigma}}(t)\mathbf{v}_k(t)}{\lambda_k(t) - \lambda_j(t)}\mathbf{v}_k^*(t) \qquad (16)$$

Let us calculate a further step. We can differentiate (14), obtaining

$$\ddot{\lambda}_k(t) = \dot{\mathbf{v}}_k^*(t)\dot{\Sigma}(t)\mathbf{v}_k(t) + \mathbf{v}_k^*(t)\ddot{\mathbf{\Sigma}}(t)\mathbf{v}_k(t) + \mathbf{v}_k^*(t)\dot{\mathbf{\Sigma}}(t)\dot{\mathbf{v}}_k(t)$$

Applying (15) and (16), we conclude the *Hadamard second variation formula*

$$\begin{aligned}\ddot{\lambda}_k(t) &= \mathbf{v}_k^*(t)\ddot{\mathbf{\Sigma}}(t)\mathbf{v}_k(t) + 2\sum_{j\neq k}\frac{(\mathbf{v}_k^*\dot{\mathbf{\Sigma}}(t)v_j(t))(\mathbf{v}_j^*(t)\dot{\Sigma}(t)\mathbf{v}_k(t))}{\lambda_k(t) - \lambda_j(t)}\\ &= \mathbf{v}_k^*(t)\ddot{\mathbf{\Sigma}}(t)\mathbf{v}_k(t) + 2\sum_{j\neq k}\frac{|\mathbf{v}_k^*(t)\dot{\mathbf{\Sigma}}(t)\mathbf{v}_j(t)|^2}{\lambda_k(t) - \lambda_j(t)}\end{aligned}$$

□

References

Berlinet, Alain et al. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. en. Boston, MA: Springer US. ISBN: 978-1-4613-4792-7 978-1-4419-9096-9. DOI: 10.1007/978-1-4419-9096-9. URL: http://link.springer.com/10.1007/978-1-4419-9096-9 (visited on 05/09/2025).

Berrendero, J. R. et al. (Sept. 2011). "Principal components for multivariate functional data". In: *Computational Statistics & Data Analysis* 55.9, pp. 2619–2634. ISSN: 0167-9473. DOI: 10.1016/j.csda.2011.03.011. URL: https://www.sciencedirect.com/science/article/pii/S0167947311001022 (visited on 04/10/2025).

Carmeli, C. et al. (July 2008). *Vector valued reproducing kernel Hilbert spaces and universality*. arXiv:0807.1659 [math]. DOI: 10.48550/arXiv.0807.1659. URL: http://arxiv.org/abs/0807.1659 (visited on 05/11/2025).

Chiou, Jeng-Min et al. (2014). “Multivariate Functional Principal Component Analysis: A Normalization Approach”. In: *Statistica Sinica* 24.4. Publisher: Institute of Statistical Science, Academia Sinica, pp. 1571–1596. ISSN: 1017-0405. URL: https://www.jstor.org/stable/24310959 (visited on 04/10/2025).

Christmann, Andreas et al. (2008). *Support Vector Machines*. en. Information Science and Statistics. New York, NY: Springer New York. ISBN: 978-0-387-77241-7 978-0-387-77242-4. DOI: 10.1007/978-0-387-77242-4. URL: https://link.springer.com/10.1007/978-0-387-77242-4 (visited on 05/09/2025).

Cybenko, G. (Dec. 1989). “Approximation by superpositions of a sigmoidal function”. en. In: *Mathematics of Control, Signals and Systems* 2.4, pp. 303–314. ISSN: 1435-568X. DOI: 10.1007/BF02551274. URL: https://doi.org/10.1007/BF02551274 (visited on 05/14/2025).

Ifantis, Evangelos K. (Jan. 1988). “A theorem concerning differentiability of eigenvectors and eigenvales with some applications”. EN. In: *Applicable Analysis*. Publisher: Gordon and Breach Science Publishers Ltd. DOI: 10.1080/00036818808839766. URL: https://www.tandfonline.com/doi/abs/10.1080/00036818808839766 (visited on 04/16/2025).

Jacques, Julien et al. (2014). “Model-based clustering for multivariate functional data”. In: *Computational Statistics & Data Analysis* 71.C. Publisher: Elsevier, pp. 92–106. ISSN: 0167-9473. DOI: 10.1016/j.csda.2012.12.004. URL: https://EconPapers.repec.org/RePEc:eee:csdana:v:71:y:2014:i:c:p:92-106 (visited on 04/19/2025).

Karhunen, K. (1946). *Zur Spektraltheorie stochastischer Prozesse*. Suomalainen Tiedeakatemia.

Lax, Peter D. et al. (2007). *Linear algebra and its applications*. 2nd ed. Pure and applied mathematics. OCLC: ocn138342469. Hoboken, N.J: Wiley-Interscience. ISBN: 978-0-471-75156-4.

Loève, M. (1946). “Fonctions aléatoires à décomposition orthogonale exponentielle”. In: *La Revue Scientifique* 84, pp. 159–162.

Pearson, Karl (1901). “Principal components analysis”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 6.2, p. 559.

Ramsay, J. O. et al. (2005). *Functional Data Analysis*. en. Springer Series in Statistics. New York, NY: Springer. ISBN: 978-0-387-40080-8 978-0-387-22751-1. DOI: 10.1007/b98888. URL: http://link.springer.com/10.1007/b98888 (visited on 04/19/2025).

Schölkopf, Bernhard et al. (1997). “Kernel principal component analysis”. en. In: *Artificial Neural Networks — ICANN’97*. Ed. by Wulfram Gerstner et al. Berlin, Heidelberg: Springer, pp. 583–588. ISBN: 978-3-540-69620-9. DOI: 10.1007/BFb0020217.

Tao, Terence (Oct. 2008). *When are eigenvalues stable?* en. URL: https://terrytao.wordpress.com/2008/10/28/when-are-eigenvalues-stable/ (visited on 04/16/2025).