Kernel methods for exploratory analysis

Internship Report

Master 2: "Research and Innovation" of Applied and Pure Mathematics

Joséphine MARTIN

2023

Supervisors: Marie-Laure Martin and Nathalie Vialaneix

Contents

Abstract

My internship is part of the INRAE project "PeerSim". The general motivation of this internship is twofold. The first one is to study the impact of combined stresses, like CO_2 rate and temperature, on the plant *Arabidopsis thaliana*. The second one is to investigate the impact of the sample size on the biological interpretation on statistical results. This second point aims to identify a minimum number of experimental replicates needed to properly estimate the impact of stresses.

The internship is organized into two parts, the theoretical part and the practical part. The theoretical part justifies the tools used in the practical part.

This report investigates the theory of kernel methods for exploratory analysis. The first section deals with the Principal Component Analysis (PCA). Then, the second section focuses on reproducing kernel Hilbert space which allows the extension of PCA to non-numerical data comparison. The extension of PCA is called Kernel PCA. In the context of the "PeerSim" project, Kernel PCA offers the advantages of comparing non-numerical data, especially comparing gene networks.

Part I
Principal Component Analysis

1 Introduction to PCA

Principal Component Analysis (PCA) is a mathematical method used in exploratory analyses. The goal is to analyse p variables on a population of n individuals in a simple way. The idea is then to reduce the dimension (usually 2 or 3) of the space of interest without losing too much information.

To illustrate the intuitive idea behind PCA, consider the following Figure 1 where we aim to project a three-dimensional fish onto a two-dimensional space.

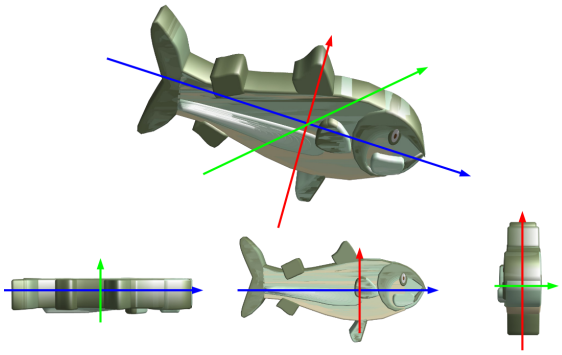

Figure 1: 3-dimensional fish onto a 2-dimensional space

The projection that best captures the overall shape of the fish is achieved by projecting onto the blue and red axes. Among all the possible projections, this specific one allows the fish to take up the most space in the projected dimension. This is the principle of PCA: finding the axes that maximize the variability of the projection.

From a mathematical standpoint, we seek the best projection plane, in a least squares sense, to obtain the most faithful representation of the data.

In this section, we explore the theoretical aspects of PCA on a dataset with n individuals and p variables. We aim to find the axes that maximize the projection. To achieve this, we use the variance-covariance matrix to analyze the dispersion of the given data. By doing so, we can obtain the coordinates for both projected individuals and variables.

1.1 Notation and statistical model

Consider p statistical variables, $X^j \in \mathbb{R}^n$ for $j \in [\![1, p]\!]$ on n individuals. Let x_i^j denote the measure of X^j for the ith individual. We obtain the following matrix

$$X = \begin{pmatrix} x_1^1 & \dots & x_1^j & \dots & x_1^p \\ \vdots & \dots & \vdots & \dots & \vdots \\ x_i^1 & \dots & x_i^j & \dots & x_i^p \\ \vdots & \dots & \vdots & \dots & \vdots \\ x_n^1 & \dots & x_n^j & \dots & x_n^p \end{pmatrix}.$$

1. For each individual i, we consider the vector $x_i \in \mathbb{R}^p$ which is the i-th row of X. This element is in the vector space E called the **individual space**. This space is endowed with the canonical basis and a metric M making it an Euclidien space. In this document, the particular case where $M := I_p$ is studied.

2. For each variable j, we consider the vector $x^j \in \mathbb{R}^n$ which is the j-th column of X. This element is in the vector space F called the **variable space** . This space is endowed with the canonical basis and a diagonal metric $\frac{1}{n} I_n$ allowing it to be an Euclidean space.

   Each variable X^j is associated to $\tilde{x}^j = x^j - \bar{x}^j \mathbb{1}_n \in \mathbb{R}^n \subseteq F$ where $\mathbb{1}_n := (1 \dots 1)^\top \in \mathbb{R}^n$ and $\bar{x}^j$ is the empirical mean of X^j defined by $\bar{x}^j = \frac{1}{n} \sum_{i=1}^n x_i^j$.

We also consider the following definitions:

- **Individual barycenter**: $g = \frac{1}{n} X^\top \mathbb{1}_n = (\bar{x}^1 \dots \bar{x}^p)^\top \in \mathbb{R}^p$;
- **Matrix of the centered data**: $\bar{X} = X - \mathbb{1}_n g^\top$;
- **Empirical standard deviation**: $\forall\, j \in [\![1, p]\!]$, $\sigma_j = \left(\frac{1}{n} \tilde{x}^{j\top} \tilde{x}^j\right)^{1/2} = \frac{1}{\sqrt{n}} \|\tilde{x}^j\|$;
- **Matrix of reduced data** $\tilde{X} = \bar{X} \Sigma^{-1/2}$ with $\Sigma = \mathrm{diag}(\sigma_1^2, \dots, \sigma_p^2)$ (this definition assumes that $\forall\, j \in [\![1, p]\!]$, $\sigma_j \neq 0$);
- **Empirical covariance of** X^j **and** $X^{j'}$: $\forall\, j, j' \in [\![1, p]\!]$, $\frac{1}{n} \tilde{x}^{j\top} \tilde{x}^{j'} = \frac{1}{n} \langle \tilde{x}^j, \tilde{x}^{j'} \rangle$;
- **Variance/covariance matrix**: $V = \frac{1}{n} \bar{X}^\top \bar{X} \in \mathcal{M}_p(\mathbb{R})$;
- **Correlation of** $X^j, X^{j'}$: $\cos(\theta(x^j, x^{j'})) = \frac{\langle x^j, x^{j'} \rangle}{\|x^j\| \|x^{j'}\|}$.

1.2 Inertia

1.2.1 Global inertia

In statistic studies, an observation is represented as a point with the following coordinates $(x_i^1, \dots, x_i^p)$ in a p-dimensional space. The data dispersion is measured thanks to the inertia. The global inertia of data is the mean of the squared distances between the data and their barycenter:

$$I_E = \frac{1}{n} \sum_{i=1}^n (x_i - g)(x_i - g)^\top = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^p (x_i^j - \bar{x}^j)^2 = \frac{1}{n} \mathrm{Tr}(\bar{X}^\top \bar{X}) = \mathrm{Tr}(V).$$

1.2.2 Projected data inertia

Consider E_q, a q-dimensional subspace of E. Let $\mathcal{B}_q = (e_1, \dots e_q)$ be an orthonormal basis of this subspace, the orthogonal projection onto E_q is defined by $p_{E_q}(x) = \sum_{j=1}^q \langle x, e_j \rangle e_j$ for all $x \in E$. The orthogonal projector is also written $p_{E_q}(x) = P_{E_q} x$ with P_{E_q} a $p \times p$-dimensional symmetric matrix such that

$$P_{E_q} = (e_1, \dots e_q)(e_1, \dots e_q)^\top.$$

Furthermore, since the orthogonal projection onto E_q of a point $x \in E_q$ is x, we have $P_{E_q}^2 = P_{E_q}$.

Hence, the projected data is associated to the data matrix $\bar{X}P_{E_q}^\top$ since each individual is projected onto E_q using column vector $P_{E_q}x_i$ or a row vector $x_i^\top P_{E_q}$. Since the covariance-variance matrix of the data matrix $\bar{X}P_{E_q}^\top$ is

$$V_{E_q} = \frac{1}{n}(\bar{X}P_{E_q}^\top)^\top(\bar{X}P_{E_q}^\top) = P_{E_q}VP_{E_q}^\top, \quad P_{E_q}^\top = P_{E_q} \quad \text{and} \quad P_{E_q}^2 = P_{E_q}$$

we have

$$\mathrm{Tr}(V_{E_q}) = \mathrm{Tr}(P_{E_q}VP_{E_q}^\top) = \mathrm{Tr}(VP_{E_q}^\top P_{E_q}) = \mathrm{Tr}(VP_{E_q}^2) = \mathrm{Tr}(VP_{E_q}).$$

Thus, the inertia of the projected data is equal to $\mathrm{Tr}(VP_{E_q})$. Therefore, the goal is to find a projector that maximizes the trace of VP_{E_q} because this will retain the maximum variability of the initial data.

1.3 Principal axes and principal components

In this section, we focus on the axes that maximize the variability of the projection. Then we can introduce components that are linear combinaison of the original variables in order to capture the most significant information in the considered dataset.

1.3.1 Principal axes

First, we will investigate the case where $q = 1$.

Lemma 1. The vector $a_1 \in \mathbb{R}^p$ that maximizes the inertia of the projected data is the eigenvector of the matrix V associated with its largest eigenvalue.

Proof. Let E_1 be the line spanned by the vector a_1. Without loss of generality, we assume that a_1 is normalized, i.e., $\|a_1\|^2 = 1$. The matrix of the orthogonal projection of E onto E_1 is then given by

$$P_{E_1} = a_1 a_1^\top.$$

As seen in Section 1.2.2, the inertia of the data projected onto this line is

$$\mathrm{Tr}(VP_1) = \mathrm{Tr}(Va_1a_1^\top) = \mathrm{Tr}\ (a_1^\top Va_1) = a_1^\top Va_1$$

since $a_1^\top Va_1$ is a scalar. Since V is positive semidefinite and symmetric, by the spectral theorem, it is diagonalizable in an orthonormal basis, and its eigenvalues, $\lambda_1, \ldots, \lambda_n$, are non-negative. Thus, we write $V = QDQ^\top$ with Q being a rotation matrix (i.e., $QQ^\top = \mathbb{I}_p$) made up of the orthonormal vectors $q_1, \ldots, q_p$, and $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$. Without loss of generality, we assume that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$. We then have

$$a_1^\top Va_1 = a_1^\top QDQ^\top a_1 = u_1^\top Du_1$$

where $u_1 := Q^\top a_1 \in \mathbb{R}^p$ is a non-zero vector. But,

$$u_1^\top Du_1 = \sum_{j=1}^{p} \lambda_j u_{1j}^2 \leq \lambda_1 \sum_{j=1}^{p} {u_{1j}}^2 = \lambda_1$$

since λ_1 is the largest eigenvalue and $\|u_1\|^2 = a_1^\top QQ^\top a_1 = \|a_1\|^2 = 1$ by definition. The projected inertia is maximum when $u_1^\top Du_1$ is maximum, which is achieved when $u_1 = (1\ 0 \ldots 0)^\top$, i.e.,

$$a_1 = Qu_1 = q_1.$$

Finally, the vector maximizing the inertia of the data is the eigenvector associated with the largest eigenvalue of V, and the inertia of the projected data is then equal to:

$$\mathrm{Tr}(VP_1) = a_1^\top V a_1 = q_1^\top V q_1 = \lambda_1 q_1^\top q_1 = \lambda_1.$$

□

Then, we construct a subspace such that the projected inertia onto this subspace is maximized.

Proposition 1. The q-dimensional subspace E_q is spanned by the q eigenvectors, a_1, ..., a_q, of V associated with the q largest eigenvalues.
The eigenvectors of V with norm 1, $(a_k)_{k=1,\dots,p}$, are called **principal axes**, and they are orthonormal.

Proof. **Step 1:** First, we demonstrate that, if H and G are two orthogonal subspaces then

$$\mathrm{Tr}(VP_{H\oplus G}) = \mathrm{Tr}(VP_H) + \mathrm{Tr}(VP_G).$$

This comes from the fact that the projector associated with the direct sum of two orthogonal subspaces is the sum of the projectors associated with each of these subspaces. Indeed, each $x \in E$ can be rewritten as:

$$x = x_H + x_{H^\perp}.$$

If $(h_1, \dots, h_k)$ is an orthonormal basis of H, by the incomplete basis theorem, $(h_{k+1}, \dots, h_p)$ is an orthonormal basis of $H^\perp \supseteq G$. Without loss of generality, let $(h_{k+1}, \dots, h_l)$ with $l \leq p$ be a basis of G. Thus,

$$p_{H\oplus G}(x) = \sum_{j=1}^{l} \langle x, h_j\rangle h_j = \sum_{j=1}^{k} \langle x, h_j\rangle h_j + \sum_{j=k+1}^{l} \langle x, h_j\rangle h_j = p_H(x) + p_G(x).$$

But, $P_H x = (h_1 \dots h_k)(h_1 \dots h_k)^\top x = \sum_{j=1}^{l} (h_j^\top x) h_j = \sum_{i=1}^{k} \langle x, h_i\rangle h_i$ (and the same for P_G), which concludes this part of the proof.
Step 2: We prove the following proposition by induction

> The q-dimensional subspace E_q having the maximum inertia is the direct sum of 1-dimensional orthonormal subspaces spanned by the eigenvectors associated with the q largest eigenvalues of V.

$q = 1$**:** As we saw in Lemma 1, the subspace E_1 that has the maximum inertia is spanned by the eigenvector associated with the largest eigenvalue λ_1 of V. Therefore, the proposition is satisfied for a 1-dimensional subspace.
Induction step: Assume the statement holds for $q < p$. Let us first show that the subspace E_{q+1} of maximum inertia is the direct sum of E_q and a vector in $E_q^\perp$: let G_{q+1} be a $q+1$-dimensional subspace of E. Since $\dim(E_q^\perp) = p - q$, we have $\dim(E_q^\perp) + \dim(G_{q+1}) = p + 1 > p = \dim(E)$, which means that $\dim(E_q^\perp \cap G_{q+1}) \geq 1$ and therefore there exists a vector $v \in E_q^\perp \cap G_{q+1}$. We can write $G_{q+1} = v \oplus G_q$ where G_q and v are orthogonal in G_{q+1} and we also denote $\tilde{E} = v \oplus E_q$. Thus, $\dim(G_q) = q$ and by Step 1,

$$\mathrm{Tr}(VP_{G_{q+1}}) = \mathrm{Tr}(VP_{\{v\}}) + \mathrm{Tr}(VP_{G_q})$$

$$\mathrm{Tr}(VP_{\tilde{E}}) = \mathrm{Tr}(VP_{\{v\}}) + \mathrm{Tr}(VP_{E_q})$$

Since E_q is the maximal inertia subspace, we have, $\mathrm{Tr}(PV_{G_q}) \leq \mathrm{Tr}(PV_{E_q})$, so $\mathrm{Tr}(PV_{G_{q+1}}) \leq \mathrm{Tr}(PV_{\tilde{E}})$ for every G_{q+1}. By choosing v such that the inertia of the projected data is the largest in $E_q^\top$, the $(q+1)$-dimensional subspace such that the inertia is maximum is thus $\tilde{E}$.
Let us show that the subspace spanned by v is actually spanned by the eigenvector associated with the $(q+1)$-th largest eigenvalue of V. Without loss of generality, we assume that $\|v\|^2 = 1$ and that $\lambda_1 \geq \ldots \geq \lambda_{q+1} \geq \cdots \geq \lambda_p$. We define $u_{q+1} := Q^\top v$.
By the induction hypothesis, the orthonormal vectors $a_k = Qu_k$ span E_q for $k \in [\![1, q]\!]$, where $u_k = e_k$ is the k-th vector of the canonical basis of E. Since v is orthogonal to E_q, we have

$$\forall k \in [\![1, q]\!], \quad 0 = v^\top a_k = (Qu_{q+1})^\top Qu_k = u_{q+1}^\top u_k = u_{q+1,k}.$$

That leads to the following upper bound

$$\mathrm{Tr}(VP_{\{v\}}) = v^\top Vv = {u_{q+1}}^\top Du_{q+1} = \sum_{j=1}^{p} \lambda_j {u_{q+1,j}}^2 = \sum_{j=q+1}^{p} \lambda_j {u_{q+1,j}}^2 \leq \sum_{j=q+1}^{p} \lambda_{q+1} {u_{j,q+1}}^2 = \lambda_{q+1}$$

since $\sum_{j=q+1}^{p} {u_{q+1,j}}^2 = u_{q+1}^\top u_{q+1} = v^\top QQ^\top v = \|v\|^2 = 1$.
Finally, the projected inertia is maximum when ${u_{q+1}}^\top Du_{q+1} = \lambda_{q+1}$, which is achieved when $u_{q+1} = e_{q+1}$, i.e.,

$$v = Qu_{q+1} = a_{q+1}.$$

This completes the proof by definition of Q. □

Remark: The proposition allowed us to establish a link between the eigenvalues of V and the inertia of the data. Step 2 allows us to construct the subspace by proceeding step by step. That is, for each dimension of projection, q, we choose a vector v orthogonal to E_q and itself such that its projection has maximum inertia in $E_q^\perp$.

1.3.2 Principal components

The **principal components** are the vectors $c_k \in \mathbb{R}^n$ defined by:

$$\forall k \in [\![1, p]\!], \qquad c_k = \bar{X} a_k.$$

Proposition 2. The variance of a principal component is equal to the eigenvalue λ_k of the matrix V corresponding to the eigenvector a_k.

$$\mathrm{Var}(c_k) = \lambda_k.$$

In addition, the principal components are the eigenvectors of the matrix

$$\frac{1}{n} \bar{X} \bar{X}^\top \in \mathcal{M}_n(\mathbb{R}).$$

Proof. On one hand, by definition of a_k, we compute $\text{Var}(c_k)$:

$$\text{Var}(c_k) = \frac{1}{n} c_k^\top c_k = \frac{1}{n} a_k^\top \bar{X}^\top \bar{X} a_k = a_k^\top V a_k = \lambda_k.$$

On the other hand, as we saw in Lemma 1, a_k is an eigenvector of the matrix V. Hence,

$$\begin{aligned} V a_k = \lambda_k a_k \quad &\Leftrightarrow \quad \frac{1}{n} \bar{X}^\top \bar{X} a_k = \lambda_k a_k \\ &\Leftrightarrow \quad \frac{1}{n} \bar{X} \bar{X}^\top \bar{X} a_k = \lambda_k \bar{X} a_k \\ &\Leftrightarrow \quad \frac{1}{n} \bar{X} \bar{X}^\top c_k = \lambda_k c_k. \end{aligned}$$

□

Remark: If we denote A the column matrix of the principal axes $(a_k)_{k=1,\dots,p}$ and C the one of the principal components $(c_k)_{k=1,\dots,p}$, we have $C = \bar{X}A$.

Summary:

1. Principal axes $(a_k)_k$: $Va_k = \lambda_k a_k$ and $(a_k)_k$ are orthonormal;
2. Principal components $(c_k)_k$: $\frac{1}{n}\bar{X}\bar{X}^\top c_k = \frac{1}{n}\bar{X}\bar{X}^\top c_k = \lambda_k c_k$ and $(c_k)_k$ are orthogonal with norm $\sqrt{n\lambda_k}$;
3. $C = \bar{X}A$;
4. We can also introduce the factorial axes, which are the eigenvectors of $\frac{1}{n}\bar{X}\bar{X}^\top$ with norm $\sqrt{n}$, $(u_k)_k$. They are obtained as $u_k = \frac{\sqrt{n} \times c_k}{\|c_k\|}$.

2 Graphs

As seen in the previous section, the principal components capture and represent the most significant information in a dataset. We can then define the coordinates of projected individuals and variables onto a lower subspace using these components and the principal axes. Therefore data analysis and the extraction of insights are made easier.

2.1 Individuals

Graphs are used to help the interpretation of PCA. The obtained graphs allow to represent "as best as possible" the Euclidean distances between individuals.

Proposition 3. The coordinates of the orthogonal projection of $x_i - g$ onto E_q are the q first elements of the i-th row of matrix C.

Proof. Each individual x_i is represented by its orthogonal projection onto the subspace $E_q = \text{Span}(a_1, \dots a_q)$. Hence, the coordinates of the individual i onto a_k are given by

$$\langle x_i - g, a_k \rangle = (x_i - g)^\top a_k = e_i^\top \bar{X} a_k = c_{i,k}$$

with e_i being a vector of the E-canonical basis. □

Definition. The **overall quality** is explained by the proportion of dispersion retained by the project, as measured by inertia:

$$\chi_q := \frac{\mathrm{Tr}(VP_{E_q})}{\mathrm{Tr}(V)} = \frac{\sum_{k=1}^{q}\lambda_k}{\sum_{k=1}^{p}\lambda_k}.$$

Proof. The second equality comes from the definition of the trace, $\mathrm{Tr}(V) = \sum_{k=1}^{p}\lambda_k$. In addition, we have,

$$I_{E_q} = \mathrm{Tr}(VP_{E_q}) = \sum_{k=1}^{q}\mathrm{Tr}(VP_{\{a_k\}}) = \sum_{k=1}^{q}\lambda_k.$$

thanks to the proofs of Lemma 1 and Proposition 2. □

Remark: The goal is to have χ_q the closest to 1.

Definition. The **quality of the representation** of each x_i is given by the squared cosine of the angle between x_i and its projection

$$\cos^2\left(\theta\left(x_i - g, P_{E_q}(x_i - g)\right)\right) = \frac{\|P_{E_q}(x_i - g)\|^2}{\|x_i - g\|^2} = \frac{\sum_{k=1}^{q}(c_{i,k})^2}{\sum_{k=1}^{p}(c_{i,k})^2}.$$

Proof. Since

$$\begin{aligned}
\cos\left(\theta\left(x_i - g, P_{E_q}(x_i - g)\right)\right) &= \frac{\langle x_i - g, P_{E_q}(x_i - g)\rangle}{\|x_i - g\|\|P_{E_q}(x_i - g)\|} \\
&= \frac{\langle P_{E_q}(x_i - g) + P_{E_q^\top}(x_i - g), P_{E_q}(x_i - g)\rangle}{\|x_i - g\|\|P_{E_q}(x_i - g)\|} \\
&= \frac{\langle P_{E_q}(x_i - g), P_{E_q}(x_i - g)\rangle}{\|x_i - g\|\|P_{E_q}(x_i - g)\|} \\
&= \frac{\|P_{E_q}(x_i - g)\|^2}{\|x_i - g\|\|P_{E_q}(x_i - g)\|} \\
&= \frac{\|P_{E_q}(x_i - g)\|}{\|x_i - g\|},
\end{aligned}$$

we find the first equality by applying squared to the previous equality.
For the second equality, we use

1. $\langle x_i - g, a_k\rangle = c_{i,k}$
2. $P_{E_q}(x_i - g) = \sum_{k=1}^{q}\langle x_i - g, a_k\rangle a_k$
3. $x_i - g = \sum_{k=1}^{p}\langle x_i - g, a_k\rangle a_k$.

Hence, by orthonormality of the principal axis, we have

$$\begin{aligned}\|P_{E_q}(x_i - g)\|^2 &= \langle P_{E_q}(x_i - g), P_{E_q}(x_i - g)\rangle \\ &= \sum_{l,k=1}^{q} \langle x_i - g, a_k\rangle\langle x_i - g, a_l\rangle\langle a_k, a_l\rangle \\ &= \sum_{k=1}^{q} \langle x_i - g, a_k\rangle^2 \quad \text{since } (a_k)_k \text{ are orthogonal} \\ &= \sum_{k=1}^{q} (c_{i,k})^2.\end{aligned}$$

Similarly, $\|x_i - g\|^2 = \sum_{k=1}^{p}(c_{i,k})^2$, which completes the proof. □

Definition. The **contribution of each individual to the data inertia** is

$$\gamma_i := \frac{\|x_i - g\|^2}{n\mathrm{Tr}(V)} = \frac{\sum_{k=1}^{p}(c_{i,k})^2}{n\sum_{k=1}^{p}\lambda_k}.$$

In particular, the contribution of the ith individual to component c_k is

$$\gamma_i^k := \frac{(c_{i,k})^2}{n\lambda_k}.$$

This allows to find the most influential observations and eventually outliers. In this case, we can remove outliers for a new analysis and still represent them (as an external data) with respect to the principal axis.

2.2 Variables

The obtained graphs allow to represent the correlations between the components and the initial variables.

Proposition 4. The coordinates of the orthogonal projection of $\tilde{x}^j$ onto F_q with respect to the norm induced by $\frac{1}{n}\mathbb{I}_n$ are the q first elements of the jth row of the matrix $AD^{1/2}$, where $D^{1/2} := \mathrm{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_p})$

Proof. Without loss of generality, we consider only the projections onto the axes corresponding to the positive eigenvalues, the others being irrelevant for the analysis. The orthogonal projection $p_{F_q}(\tilde{x}^j)$ onto the subspace F_q, spanned by the q first factorial axes, represent the variable X^j. Since V is symmetric and positive, its eigenvalues are all positives. Consider the vectors u_k for $k \in [\![1, r]\!]$ with $r \leq p$ such that c_r is the last component for which $\lambda_r \neq 0$. The coordinates of the variables with respect to the norm induced by $\frac{1}{n}\mathbb{I}_n$ are the coordinates of the orthogonal projection of $\tilde{x}^j$ onto u_k (because it has norm 1 for $\frac{1}{n}\mathbb{I}_n$):

$$\frac{\langle \tilde{x}^j, u_k\rangle}{\|u_k\|^2} = \frac{1}{n}\tilde{x}^{j\top}u_k = \frac{1}{n\sqrt{\lambda_k}}\tilde{x}^{j\top}c_k = \frac{1}{n\sqrt{\lambda_k}}e^{j\top}\bar{X}^\top c_k = \frac{1}{n\sqrt{\lambda_k}}e^{j\top}\bar{X}^\top\bar{X}a_k = \frac{1}{\sqrt{\lambda_k}}e^{j\top}Va_k = \sqrt{\lambda_k}\,a_{j,k}.$$

Here, e^j is the j-th vector of the canonical basis of F. This completes the proof. □

Definition. The **quality of the representation** of each $\tilde{x}^j$ is given by the squared cosine of the angle between $\tilde{x}^j$ and its projection.

$$\cos^2\left(\theta\left(\tilde{x}^j, P_{F_q}(\tilde{x}^j)\right)\right) = \frac{\|P_{F_q}(\tilde{x}^j)\|^2}{\|\tilde{x}^j\|^2} = \frac{\sum_{k=1}^{q}\lambda_k(a_{j,k})^2}{\sum_{k=1}^{p}\lambda_k(a_{j,k})^2}.$$

Remark: The proof is similar to the one on the quality of the representation for individuals. The variable-factor correlation index allows the interpretation of the factor axes thanks to a correlation factor between principal variables c_k and initial variables $\tilde{x}^j$. Thus

$$\cos\left(\theta\left(\tilde{x}^j, c_k\right)\right) = \cos\left(\theta\left(\tilde{x}^j, u_k\right)\right) = \frac{\langle\tilde{x}^j, u_k\rangle}{\|\tilde{x}^j\|\|u_k\|} = \frac{\langle\tilde{x}^j, u_k\rangle}{\|\tilde{x}^j\|\sqrt{n}} = \frac{n\sqrt{\lambda_k}a_{j,k}}{\|\tilde{x}^j\|\sqrt{n}} = \frac{\sqrt{\lambda_k}a_{j,k}}{\sigma_j}.$$

because c_k and u_k are co-linear and the norm of u_k is $\sqrt{n}$. The correlation (unit) circle, $\mathcal{S}_n \cap F_2$, allows to determine the representation quality of a variable. The reduced variables $\tilde{\tilde{x}}^j = \frac{\tilde{x}^j}{\sigma_j}$ are on the sphere $\mathcal{S}_n$ of radius 1 in F. Since the projections $\tilde{x}^j$ and $\tilde{\tilde{x}}^j$ are colinear and $\|P_{F_q}(\tilde{\tilde{x}}^j)\| = \cos\left(\theta\left(\tilde{\tilde{x}}^j, P_{F_q}(\tilde{\tilde{x}}^j)\right)\right) \leq 1$, the closer $P_{F_q}(\tilde{\tilde{x}}^j)$ is to the unit circle, the better the quality of the representation of the variables is.

3 Choice of the dimension

The quality of the data-interpretation depends essentially on the dimension q of the representation subspaces. There exist several criteria that allow to choose q:

1. We can first choose q so that the overall quality, measured by the proportion of explained inertia,

   $$r_q = \frac{\sum_{k=1}^{q}\lambda_k}{\sum_{k=1}^{p}\lambda_k}$$

   is higher than a fixed threshold value.

2. Another method is to use the Kaiser's rule. This consists in considering only the eigenvalues that are larger than their mean. In particular, in the case of reduced data matrices, only eigenvalues with a value larger than 1 are considered.

3. It is also possible to use the scree plot and look for an "elbow" in the graph, and to retain the eigenvalues up to this elbow. Catell's *scree-test* is the analytical version of this. For $k \in [\![1, p]\!]$, it computes the first differences

   $$\lambda_k - \lambda_{k+1} = \varepsilon_k$$

   and then, the second differences

   $$\varepsilon_k - \varepsilon_{k+1} = \delta_k.$$

   The selected eigenvalues λ_1, λ_2, ..., λ_k, λ_{k+1} are those such that δ_k is positive.

4 Application

Analysis for the project

To illustrate the theory, we will use the data made available within the context of the "PeerSim" project.

To analyze combined stresses on the plant *Arabidopsis thaliana*, we are investigating how genes are expressed under different culture conditions. These conditions include ambient temperature with either elevated or ambient CO2 concentration, as well as high temperature with either elevated or ambient CO2 concentration. We have gene expression data from 6 replicates for each condition. On the whole we study 24 replicates.

From a mathematical standpoint, the replicates refer to the 24 individuals in the dataset, while the genes refer to the variables being studied.

When performing PCA on the gene expression matrix, we obtain graphs of individuals and graphs of variables.

On the individual graphs:

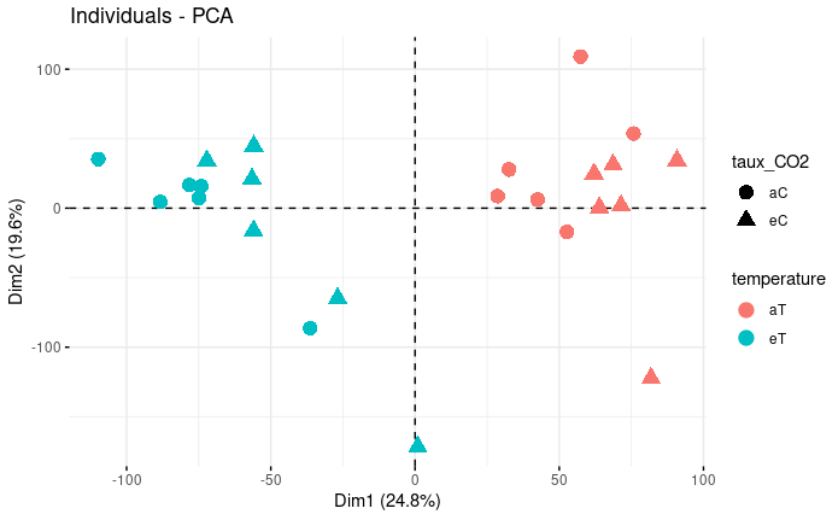

Figure 2: Individual graph

In Figure 2, the graph represents the projected individuals onto the principal axes 1 and 2.

Axis 1 explains 24.8% of the total variance, and axis 2 explains 19.6% of the total variance. Thus, using axes 1 and 2 we can explain 44.4% of the variance in the dataset. In this graph of individuals, the color distinguishes the temperature conditions. Here, axis 1 plays a discriminating role by separating the replicates studied under high-temperature conditions from the replicates studied under ambient temperature conditions.

On the variable graphs:

In Figure 3, the graph represents the variables projected onto the axes 1 and 2. This represent the correlation between the principal component and the genes. Here, we represent the genes with a contribution to the axes higher than 0.96.

The graph of variables helps extract information from the graph of individuals, revealing additional insights. For example if genes show a correlation with axis 1 in the variable graph, it indicates a statistical relationship with the positively represented individuals along axis 1 in the individual graph. Nevertheless, further investigation is needed to determine the biological or functional implications of this association.

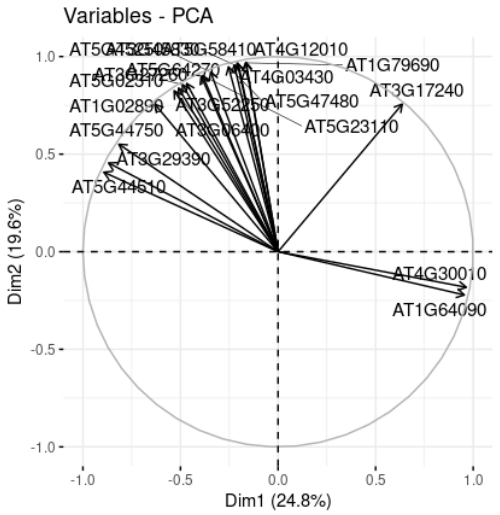

Figure 3: Variable graph

Part II
Kernel Principal Component Analysis

Kernel Principal Component Analysis (Kernel PCA) is an extension of the PCA. To motivate the interest in kernel PCA, let us talk about the "PeerSim" project. The initial objective of my internship was to determine the impact of the number of replicates on the biological interpretation of statistical results. To achieve this, we can construct multiple graphs by varying the number of studied replicates while keeping the same number of replicates for each condition. These graphs are constructed based on the relationships between genes, derived from the gene expression matrix. From there, the goal is to compare these graphs. Since kernel PCA offers the advantage of comparing non-numerical data, we can compare in particular gene networks.

We consider n observations $(x_i)_{i=1,\dots,n}$ that take their values in an arbitrary space $\mathcal{X}$.

1 Reproducing Kernel Hilbert spaces

In this section, we first introduce Reproducing Kernel Hilbert Spaces and properties in order to define Kernel PCA.

1.1 Some reminders on Hilbert spaces

Definition. A pre-Hilbert space $\mathcal{H}$ is a vector space endowed with an inner product $\langle ., . \rangle_{\mathcal{H}}$.

Definition. A Cauchy sequence (f_n) is a sequence such that

$$\lim_{N \to \infty} \sup_{n,m \geq N} \|f_n - f_m\|_{\mathcal{H}} = 0.$$

Definition. A Hilbert space is a complete pre-Hilbert space for the norm $\|.\|_{\mathcal{H}}$, meaning that every Cauchy sequence in $\mathcal{H}$ converges in $\mathcal{H}$.

1.2 RKHS and kernel

Definition. Let $\mathcal{X}$ be an arbitrary space, $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a function and $\mathcal{H}$ be a Hilbert space endowed with an inner product $\langle ., . \rangle_{\mathcal{H}}$. For every $x \in \mathcal{X}$, we denote by κ_x the function defined by $\kappa_x : t \in \mathcal{X} \to K(x, t)$. The function K is called **Reproducing Kernel** if

- $\forall x$, κ_x is an element of $\mathcal{H}$;
- $\forall x \in \mathcal{X}$ and $\forall f \in \mathcal{H}$, $f(x) = \langle f, \kappa_x \rangle_{\mathcal{H}}$ (reproducing property).

If such a reproducing kernel exists, $\mathcal{H}$ is called a **Reproducing Kernel Hilbert Space** (RKHS).

Proposition 5. If $\mathcal{H}$ is a RKHS, then the reproducing kernel is unique. Conversely, a function K can be the reproducing kernel of at most one RKHS.

Proof. Let $\mathcal{H}$ be a RKHS with K and K' two reproducing kernels of this space. For every $x \in \mathcal{X}$, we have

$$\begin{aligned} \|\kappa_x - \kappa'_x\|^2_{\mathcal{H}} &= \langle \kappa_x - \kappa'_x, \kappa_x - \kappa'_x \rangle_{\mathcal{H}} \\ &= \langle \kappa_x, \kappa_x \rangle_{\mathcal{H}} - \langle \kappa'_x, \kappa_x \rangle_{\mathcal{H}} - \langle \kappa_x, \kappa'_x \rangle_{\mathcal{H}} + \langle \kappa'_x, \kappa'_x \rangle_{\mathcal{H}} \\ &= \kappa_x(x) - \kappa'_x(x) - \kappa_x(x) + \kappa'_x(x) = 0. \end{aligned}$$

Thus, we have

$$\begin{aligned} \forall\, x \in \mathcal{X}, \qquad \kappa_x = \kappa'_x &\Leftrightarrow \kappa_x(t) = \kappa'_x(t) \qquad \forall t \in \mathcal{X} \\ &\Leftrightarrow K(x, t) = K'(x, t) \qquad \forall t \in \mathcal{X} \\ &\Leftrightarrow K = K'. \end{aligned}$$

Conversely, we now assume that K is the reproducing kernel of two Hilbert spaces $\mathcal{H}$ and $\mathcal{H}'$. By definition of the reproducing kernel, we know that all the functions κ_x for $x \in \mathcal{X}$ are in $\mathcal{H}$. Therefore,

$$\mathcal{H}_0 = \text{Span}\left\{ \sum_{i=1}^{n} \alpha_i \kappa_{x_i}, \forall i \in [\![1, n]\!] \quad \alpha_i \in \mathbb{R}, x_i \in \mathcal{X} \right\}$$

is a subspace of $\mathcal{H}$.
If $f \in \mathcal{H}$ is orthogonal to $\mathcal{H}_0$ then it is in particular orthogonal to κ_x for any $x \in \mathcal{X}$ which implies

$$\forall x \in \mathcal{X} \quad f(x) = \langle f, \kappa_x \rangle_{\mathcal{H}} = 0 \text{ i.e. } f = 0.$$

Then $\mathcal{H}_0^{\perp} = \{0\}_{\mathcal{H}}$. Since $\mathcal{H}$ is a Hilbert space $(\mathcal{H}_0^{\perp})^{\perp} = \overline{\mathcal{H}_0}$ thus

$$\overline{\mathcal{H}_0} = (\mathcal{H}_0^{\perp})^{\perp} = \{0\}_{\mathcal{H}}^{\perp} = \mathcal{H}.$$

In other words, $\mathcal{H}_0$ is dense in $\mathcal{H}$.
Moreover for $f \in \mathcal{H}_0$ we can write $f := \sum_{i=1}^{n} \alpha_i \kappa_{x_i}$.
The $\mathcal{H}$-norm for functions in $\mathcal{H}_0$ only depends on the reproducing kernel K because

$$\begin{aligned} \|f\|_{\mathcal{H}}^2 &= \sum_{i,j=1}^{n} \alpha_i \alpha_j \langle \kappa_{x_i}, \kappa_{x_j} \rangle_{\mathcal{H}} \\ &= \sum_{i,j=1}^{n} \alpha_i \alpha_j K(x_i, x_j). \end{aligned}$$

Suppose now that $\mathcal{H}'$ is also a RKHS that admits K as reproducing kernel. By the same argument $\mathcal{H}_0$ is dense in $\mathcal{H}'$ and the $\mathcal{H}'$-norm in $\mathcal{H}_0$ is also given by $\|f\|_{\mathcal{H}'}^2 = \sum_{i,j=1}^{n} \alpha_i \alpha_j K(x_i, x_j)$.
In particular, for any $f \in \mathcal{H}_0$

$$\|f\|_{\mathcal{H}} = \|f\|_{\mathcal{H}'},$$

and by the reproducing property

$$\forall x \in \mathcal{X} \quad \langle f, \kappa_x \rangle_{\mathcal{H}} = f(x) = \langle f, \kappa_x \rangle_{\mathcal{H}'}.$$

Then, let us prove that the two RKHS are equal.
Since $\mathcal{H}_0$ is dense in $\mathcal{H}$, for any $f \in \mathcal{H}$, there exists $(f_n)_n \in \mathcal{H}_0$ such that

$$\|f_n - f\|_{\mathcal{H}} \to_{n \to \infty} 0$$

i.e. $\forall \epsilon > 0 \quad \exists N \in \mathbb{N}, \forall n \geq N \quad \|f_n - f\|_{\mathcal{H}} \leq \epsilon/2$.
This sequence is Cauchy for the $\mathcal{H}$-norm because for $n, m > N$

$$\|f_n - f_m\|_{\mathcal{H}} \leq \|f_n - f\|_{\mathcal{H}} + \|f_m - f\|_{\mathcal{H}} \leq \epsilon.$$

Since the norms coincide for each element of $\mathcal{H}_0$, the sequence $(f_n)_n$ is also Cauchy for the $\mathcal{H}'$-norm.
Moreover, as $\mathcal{H}'$ is complete, the Cauchy sequence converges towards a function $g \in \mathcal{H}'$.
Using the reproducing property and that the inner-product coincide for any element of $\mathcal{H}_0$ we have

$$\forall x \in \mathcal{X} \quad f(x) = \langle f, \kappa_x \rangle_{\mathcal{H}} = \lim_{n \to \infty} \langle f_n, \kappa_x \rangle_{\mathcal{H}} = \lim_{n \to \infty} \langle f_n, \kappa_x \rangle_{\mathcal{H}'} = \langle g, \kappa_x \rangle_{\mathcal{H}'} = g(x).$$

In other words $f = g$, therefore $f \in \mathcal{H}'$ i.e; $\mathcal{H} \subseteq \mathcal{H}'$. By symmetry of the argument $\mathcal{H} = \mathcal{H}'$.
Hence, we have also shown, by the reproducing property, that the inner product in $\mathcal{H}$ and in $\mathcal{H}'$ coincide for each f which completes the proof. □

Theorem 1. The Hilbert space $\mathcal{H}$ is a RKHS if and only if the linear map

$$\forall x \in \mathcal{X}, \qquad \begin{array}{rccc} F: & \mathcal{H} & \longrightarrow & \mathbb{R} \\ & f & \longmapsto & f(x) \end{array}$$

is continuous.

Proof. $\Rightarrow$ We assume that the reproducing kernel K exists. Since $F(f) = \langle f, \kappa_x \rangle_{\mathcal{H}}$, the map F is linear. Moreover, for every fixed $x \in \mathcal{X}$ and every $f \in \mathcal{H}$, we have

$$\begin{array}{rcll} |F(f)| = |f(x)| = |\langle f, \kappa_x \rangle_{\mathcal{H}}| & \leq & \|f\|_{\mathcal{H}} \|\kappa_x\|_{\mathcal{H}} & \text{by Cauchy-Schwarz} \\ & \leq & \|f\|_{\mathcal{H}} \langle \kappa_x, \kappa_x \rangle_{\mathcal{H}}^{1/2} = \|f\|_{\mathcal{H}} K(x,x)^{1/2}. & \end{array}$$

The map $f \longmapsto f(x)$ is therefore Lipschitz with the Lipschitz constant equal to $K(x,x)^{1/2} \geq 0$, and hence, it is continuous.
$\Leftarrow$ Conversely, let $\mathcal{H}$ be a Hilbert space, assume for every $x \in \mathcal{X}$, the linear map F continuous. By Riesz's representation theorem, there exists an unique $g_x \in \mathcal{H}$ such that

$$f(x) = \langle f, g_x \rangle_{\mathcal{H}}.$$

The function $g_x \in \mathcal{H}$ defines a function $K : (x,t) \longmapsto g_x(t)$. Thus defined, the function K verifies the definition of the reproducing kernel. Hence, $\mathcal{H}$ is a RKHS. □

Remark: By the linearity of F, if there exists a sequence $f_n \in \mathcal{H}$ that converges towards 0, then $f_n(x)$ converges towards 0 for each $x \in \mathcal{X}$.

Definition. A **kernel** is a function

$$K : \mathcal{X} \times \mathcal{X} \to \mathbb{R},$$

that is symmetric and positive definite (i.e., $\forall k \in \mathbb{N}$, $\forall i \in [\![1,k]\!]$, $\alpha_i \in \mathbb{R}$, and $x_i \in \mathcal{X}$, $\sum_{i,j=1}^{k} \alpha_i, \alpha_j K(x_i, x_j) \geq 0$.) For every set $\{x_1, \ldots, x_n\} \subset \mathcal{X}$, we can also define a positive semi-definite symmetric kernel matrix

$$\mathbf{K} := (\kappa_{x_i x_{i'}})_{i,i'=1,\ldots,n} \in \mathcal{S}_n^+(\mathbb{R}),$$

where $\kappa_{x_i x_{i'}} := K(x_i, x_{i'})$. This matrix $\mathbf{K}$ is called the **Gram matrix**.

The following proposition allows us to define a RKHS solely based on the kernel definition.

Proposition 6. Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a function. The function K defines a kernel if and only if K is the reproducing kernel of a Hilbert space $\mathcal{H}$ endowed with the inner product $\langle ., . \rangle_{\mathcal{H}}$.

Proof. $\Leftarrow$ We assume that K is a reproducing kernel. Let us show that K is a kernel. It is symmetric because

$$\forall (x,y) \in \mathcal{X}^2, K(x,y) = \kappa_x(y) = \langle \kappa_x, \kappa_y \rangle_{\mathcal{H}} = \langle \kappa_y, \kappa_x \rangle_{\mathcal{H}} = K(y,x).$$

Moreover, it is positive because for every $N \in \mathbb{N}, (x_1, \dots, x_N) \in \mathcal{X}^N$ and $(a_1, \dots a_N) \in \mathbb{R}^N$

$$\sum_{i,j=1}^{N} a_i a_j K(x_i, x_j) = \sum_{i,j=1}^{N} a_i a_j \langle \kappa_{x_i}, \kappa_{x_j} \rangle_{\mathcal{H}} = \left\| \sum_{i=1}^{N} a_i \kappa_{x_i} \right\|_{\mathcal{H}}^2 \geq 0.$$

$\Rightarrow$ We now assume that K is a kernel, let us show that K is the reproducing kernel of a Hilbert space $\mathcal{H}$.
To do this, we consider the space $\mathcal{H}_0$ spanned by the functions $(\kappa_x)_{x \in \mathcal{X}}$. Thus, for every $f, g \in \mathcal{H}_0$, we can write $f = \sum_{i=1}^{m} a_i \kappa_{x_i}$ and $g = \sum_{j=1}^{n} b_j \kappa_{x_j}$. We endow this space with the following symmetric bilinear form

$$\langle f, g \rangle_{\mathcal{H}_0} := \sum_{i,j} a_i b_j K(x_i, x_j).$$

Due to the symmetry of K and for every $x \in \mathcal{X}$ and $f \in \mathcal{H}_0$:

$$\langle f, \kappa_x \rangle_{\mathcal{H}_0} = \sum_{i=1}^{m} a_i K(x_i, x) = \sum_{i=1}^{m} a_i \kappa_{x_i}(x) = f(x).$$

Also, since K is positive definite,

$$\|f\|_{\mathcal{H}_0}^2 = \sum_{i,j=1}^{m} a_i a_j K(x_i, x_j) \geq 0.$$

Thus, f is a positive semi-definite and symmetric bilinear form and we can apply Cauchy-Schwarz inequality,

$$\forall x \in \mathcal{X}, \qquad |f(x)| = |\langle f, \kappa_x \rangle_{\mathcal{H}_0}| \leq \|f\|_{\mathcal{H}_0} K(x, x)^{1/2}.$$

Therefore if $\|f\|_{\mathcal{H}_0} = 0$ then for all $x \in \mathcal{X}$ $f(x) = 0$ thus $f = 0$, which is sufficient to conclude that $\langle ., . \rangle_{\mathcal{H}_0}$ defines an inner product and thus that $(\mathcal{H}_0, \langle ., . \rangle_{\mathcal{H}_0})$ is a prehilbertian space.
In addition, since, for all $x \in \mathcal{X}$, $F(f) = \langle f, \kappa_x \rangle_{\mathcal{H}_0}$ where $\langle ., . \rangle_{\mathcal{H}_0}$ is a bilinear form, we have $F(f) \leq \|f\|_{\mathcal{H}_0} K(x, x)^{1/2}$ for all $f \in \mathcal{H}_0$. This implies that $F : f \in \mathcal{H}_0 \to f(x) \in \mathbb{R}$ is a continuous linear map.
Now, let $\mathcal{H}$ be the smallest space containing $\mathcal{H}_0$ such that any Cauchy sequence $(f_n)_n \in \mathcal{H}_0$ converges to an element $f \in \mathcal{H}$. We will show that $\mathcal{H}$ is a Hilbert space and that F is continuous on $\mathcal{H}$. Then, by Theorem 1, we will be able to conclude that $\mathcal{H}$ is a RKHS with kernel K.
First, note that $\mathcal{H}_0$ is dense in $\mathcal{H}$ by definition of $\mathcal{H}$. Then, for every $f, g \in \mathcal{H}$, there exist Cauchy sequences $(f_n)_n$ and $(g_n)_n$ in $\mathcal{H}_0$ which converge to f and g in $\mathcal{H}$ respectively. We will show that $(\langle f_n, g_n \rangle_{\mathcal{H}_0})_n$ is a Cauchy sequence: For every $n, m \in \mathbb{N}$,

$$\begin{aligned} |\langle f_n, g_n \rangle_{\mathcal{H}_0} - \langle f_m, g_m \rangle_{\mathcal{H}_0}| &= |\langle f_n - f_m, g_n \rangle_{\mathcal{H}_0} + \langle f_n, g_n - g_m \rangle_{\mathcal{H}_0}| \\ &\leq \|f_n - f_m\|_{\mathcal{H}_0} \|g_n\|_{\mathcal{H}_0} + \|f_m\|_{\mathcal{H}_0} \|g_n - g_m\|_{\mathcal{H}_0} \end{aligned}$$

by Cauchy-Schwarz inequality. Now, by definition of Cauchy sequences, for every $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that, for every $n, m \geq N$, $\|f_n - f_m\|_{\mathcal{H}_0} \leq \varepsilon$ and $\|g_n - g_m\|_{\mathcal{H}_0} \leq \varepsilon$. Moreover, each Cauchy sequence is bounded then we can define a constant $M := \max(m_f, m_g)$ with $m_f = \sup(\{\|f_n\|_{\mathcal{H}_0}\}_{n \in \mathbb{N}})$ and $m_g = \sup(\{\|g_n\|_{\mathcal{H}_0}\}_{n \in \mathbb{N}})$. Thus, for every $n, m \geq N$, we obtain:

$$|\langle f_n, g_n \rangle_{\mathcal{H}_0} - \langle f_m, g_m \rangle_{\mathcal{H}_0}| \leq 2M\varepsilon.$$

Since M is a constant and since this inequality holds for any $\varepsilon > 0$, we deduce that $(\langle f_n, g_n\rangle_{\mathcal{H}_0})_n$ is a Cauchy sequence. This Cauchy sequence is in $\mathbb{R}$ and therefore converges.
To show that the limit of the Cauchy sequence $(\langle f_n, g_n\rangle_{\mathcal{H}_0})_n$ only depends on f and g, we consider two other Cauchy sequences (f'_n) and (g'_n) which converge to f and g respectively. For every $n \in \mathbb{N}$, we have

$$|\langle f_n, g_n\rangle_{\mathcal{H}_0} - \langle f'_n, g'_n\rangle_{\mathcal{H}_0}| \leq \|f_n - f'_n\|_{\mathcal{H}_0}\|g_n\|_{\mathcal{H}_0} + \|f'_n\|_{\mathcal{H}_0}\|g_n - g'_n\|_{\mathcal{H}_0}$$

$(f_n - f'_n)_n$ and $(g_n - g'_n)_n$ are Cauchy sequences in $\mathcal{H}_0$ converging to 0. It follows that the inner products $(\langle f_n, g_n\rangle_{\mathcal{H}_0})_n$ and $(\langle f'_n, g'_n\rangle_{\mathcal{H}_0})_n$ have the same limits. This enables us to to define a positive semi-definite symmetric bilinear form on $\mathcal{H}$ by setting

$$\langle f, g\rangle_{\mathcal{H}} = \lim_n \langle f_n, g_n\rangle_{\mathcal{H}_0}.$$

To demonstrate that this symmetric bilinear form is an inner product, we finally need to show that: $\|f\|_{\mathcal{H}} = 0 \implies f = 0$.
Let $f \in \mathcal{H}$ such that $\|f\|_{\mathcal{H}} = 0$. By definition of $\mathcal{H}$, there exists a Cauchy sequence $(f_n)_n$ in $\mathcal{H}_0$ which converges to f. By definition of the symmetric bilinear form on $\mathcal{H}$, we have

$$\lim_n \langle f_n, f_n\rangle_{\mathcal{H}_0} = \langle f, f\rangle_{\mathcal{H}} = 0.$$

Since $\langle ., .\rangle_{\mathcal{H}_0}$ is an inner product, by Cauchy-Schwarz inequality we have:

$$\forall x \in \mathcal{X}, \quad \forall n \in \mathbb{N}, \quad |f_n(x)| \leq \|f_n\|_{\mathcal{H}_0} K(x, x)^{1/2}$$

and thus, as $\lim_n \|f_n\|_{\mathcal{H}_0} = 0$, we obtain $\lim_n |f_n(x)| = 0$ which proves that $f(x) = \lim_n f_n(x) = 0$ for all $x \in \mathcal{X}$. We finally conclude that $f = 0$, that $\langle ., .\rangle_{\mathcal{H}}$ is an inner product, and that $(\mathcal{H}, \langle ., .\rangle_{\mathcal{H}})$ is a pre-Hilbert space.
To finally show that $(\mathcal{H}, \langle ., .\rangle_{\mathcal{H}})$ is a Hilbert space, we thus just need to show that it is a complete space. Let $(f_n)_n \in \mathcal{H}$ be a Cauchy sequence which converges towards a function f. We want to prove that f belongs to $\mathcal{H}$.
Since $\mathcal{H}_0$ is dense in $\mathcal{H}$, we have

$$\forall \varepsilon > 0 \quad \forall n \in \mathbb{N}^*, \quad \exists g_n \in \mathcal{H}_0 : \quad \|f_n - g_n\|_{\mathcal{H}} \leq \varepsilon.$$

Let us show that $(g_n)_n$ is a Cauchy sequence: $(f_n)_n$ is a Cauchy sequence, which means that, for all $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that for every $n, m > N$, $\|f_n - f_m\|_{\mathcal{H}} < \varepsilon$ then

$$\|g_n - g_m\|_{\mathcal{H}_0} = \|g_n - g_m\|_{\mathcal{H}} \leq \|g_n - f_n\|_{\mathcal{H}} + \|f_n - f_m\|_{\mathcal{H}} + \|f_m - g_m\|_{\mathcal{H}} \leq 3\varepsilon$$

Hence, $(g_n)_n$ is a also a Cauchy sequence. Thus, by definition of $\mathcal{H}$, the sequence $(g_n)_n$ converges towards $g \in \mathcal{H}$, i.e. $\forall \varepsilon \quad \exists N > 0, \quad \forall n, m > N, \|g_n - g\|_{\mathcal{H}} < \varepsilon$. Let us prove that the sequence $(f_n)_n$ also converges towards $g \in \mathcal{H}$:

$$\|f_n - g\|_{\mathcal{H}} \leq \|f_n - g_n\|_{\mathcal{H}} + \|g_n - g\|_{\mathcal{H}} \leq \varepsilon + \varepsilon = 2\varepsilon.$$

Therefore, $(f_n)_n$ converges to g in $\mathcal{H}$, which proves that $f = g \in \mathcal{H}$ and that $\mathcal{H}$ is complete.
Finally, it only remains to prove that the continuous linear map $F : \mathcal{H}_0 \to \mathbb{R}$ extends to a continuous linear map on $\mathcal{H}$. The linearity of F on $\mathcal{H}$ arises directy from the bilinearity of the inner product $\langle ., .\rangle_{\mathcal{H}}$ because, for a given $x \in \mathcal{X}$, we have:

$$\forall\, f \in \mathcal{H}, \quad F(f) = f(x) = \lim_n f_n(x) = \lim_n \langle f_n, \kappa_x\rangle_{\mathcal{H}_0} = \langle f, \kappa_x\rangle_{\mathcal{H}}.$$

Moreover, by Cauchy-Schwarz inequality on the inner product $\langle ., . \rangle_{\mathcal{H}}$, we have

$$\forall f \in \mathcal{H}, \quad |F(f)| = |\langle f, \kappa_x \rangle_{\mathcal{H}}| \leq \|f\|_{\mathcal{H}} K(x,x)^{1/2}$$

which proves the continuity of F on the Hilbert space $\mathcal{H}$.
Finally, by Theorem 1, $\mathcal{H}$ is a RKHS. □

In addition to the previous proposition, the following theorem shows that the RKHS is defined uniquely based on a kernel and the same kernel enables the definition of a mapping function. This will be particularly useful in kernel PCA because we prefer to define a kernel rather than an explicit RKHS and a mapping function, which will remain implicit.

Theorem 2. (**Moore-Aronszajn**) The function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defines a kernel if and only if there exists an unique reproducing kernel Hilbert space $\mathcal{H}$, endowed with the inner product $\langle ., . \rangle_{\mathcal{H}}$, and a function $\phi : \mathcal{X} \to \mathcal{H}$ such that

$$\forall x, x' \in \mathcal{X} \quad K(x,x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}.$$

This function is called the mapping function.

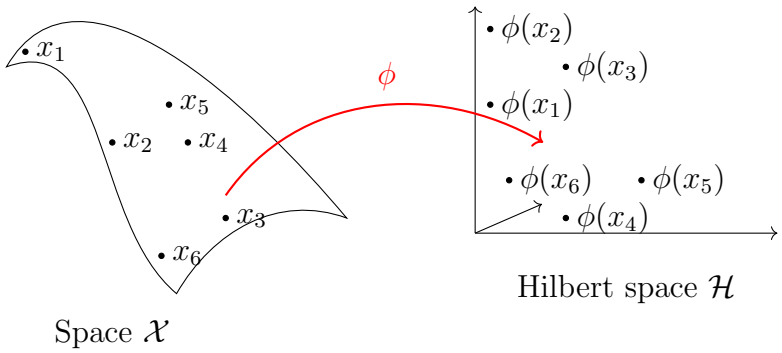

Figure 4: Information embedded in a Hilbert space

Proof. $\Rightarrow$ We assume that K defines a kernel then, by Proposition 6, K is a reproducing kernel of the Hilbert space $\mathcal{H}$ endowed with the inner product $\langle ., . \rangle_{\mathcal{H}}$.
By definition of the RKHS, for each $x \in \mathcal{X}$, we define $\kappa_x : t \in \mathcal{X} \to K(t,x) \in \mathbb{R}$ such that $\kappa_x \in \mathcal{H}$. Thus κ_x allows to define a mapping function $\phi : \mathcal{X} \to \mathcal{H}$ such that

$$\forall x \in \mathcal{X}, \qquad \phi(x) = \kappa_x,$$

which satisfies:

$$\forall (x,y) \in \mathcal{X}^2, \qquad \langle \phi(x), \phi(y) \rangle_{\mathcal{H}} = \langle \kappa_x, \kappa_y \rangle_{\mathcal{H}} = K(x,y),$$

by the kernel reproducing property.
$\Leftarrow$ Conversely, assume that K is a reproducing kernel of the Hilbert space $\mathcal{H}$, and ϕ a function

such that $K(x,x') = \langle \phi(x), \phi(x') \rangle$ for every $x, x' \in \mathcal{X}$. By the previous proposition K is indeed symmetric and positive.
Moreover, the mapping function $\phi(x)$ corresponds to κ_x.
Indeed, for every $x \in \mathcal{X}$, since K is a reproducing kernel in a Hilbert space $\mathcal{H}$, we can define for every $t \in \mathcal{X}$, $\kappa_x : t \to K(x,t) \in \mathbb{R}$. In particular, by assumption

$$\langle \kappa_x, \kappa_t \rangle = K(x,t) = \langle \phi(x), \phi(t) \rangle$$

By Riesz's representation theorem, $\phi(x) \in \mathcal{H}$ is unique for every $\phi(t) \in \mathcal{H}$ i.e. $\forall t \in \mathcal{X}$

$$\langle \kappa_x, \kappa_t \rangle = \langle \phi(x), \kappa_t \rangle = \langle \phi(x), \phi(t) \rangle \Leftrightarrow \langle \phi(x), \kappa_t - \phi(t) \rangle = 0$$

In particular we have,

$$\langle \phi(x), \kappa_x - \phi(x) \rangle = 0 \Leftrightarrow \kappa_x - \phi(x) = 0 \Leftrightarrow \phi(x) = \kappa_x.$$

□

2 Kernel trick

We use kernel structures to transform data into a space with more interresting properties. For example, we can map a non-linear model in $\mathcal{X}$ to a linear model in $\mathcal{H}$, $f(x) = \langle \phi(x), f \rangle_{\mathcal{H}}$.

Usually, $\mathcal{H}$ and ϕ are not explicitly given but are used implicitly through the kernel: this is called the **kernel trick**. It consists of using the mapping of $\mathcal{X}$ into $\mathcal{H}$ by expressing inner products and distances in $\mathcal{H}$ thanks to the kernel values. For example, the distance between two elements x_1, x_2 of $\mathcal{X}$ can be expressed solely with the kernel as the distance between their respective images in $\mathcal{H}$, i.e.,

$$d^2_{\mathcal{H}}(x_1, x_2) = \|\phi(x_1) - \phi(x_2)\|^2_{\mathcal{H}}.$$

With the kernel trick, we obtain the following relation

Proposition 7. Let x_1 and x_2 be two elements of $\mathcal{X}$, then

$$d_{\mathcal{H}}(x_1, x_2) = \sqrt{K(x_1,x_1) + K(x_2,x_2) - 2K(x_1,x_2)}$$

Proof.

$$\begin{aligned} d^2_{\mathcal{H}}(x_1, x_2) = \|\phi(x_1) - \phi(x_2)\|^2_{\mathcal{H}} &= \langle \phi(x_1) - \phi(x_2), \phi(x_1) - \phi(x_2) \rangle_{\mathcal{H}} \\ &= \langle \phi(x_1), \phi(x_1) \rangle_{\mathcal{H}} + \langle \phi(x_2), \phi(x_2) \rangle_{\mathcal{H}} - 2\langle \phi(x_1), \phi(x_2) \rangle_{\mathcal{H}} \\ &= K(x_1,x_1) + K(x_2,x_2) - 2K(x_1,x_2) \end{aligned}$$

□

Examples: 1. The Gaussian kernel defined by

$$K(x,y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}} \ \forall \sigma \in \mathbb{R}^d$$

induces the distance

$$d_{\mathcal{H}}(x,y) = \sqrt{2\left(1 - e^{-\frac{\|x-y\|^2}{2\sigma^2}}\right)}.$$

2. If $\mathcal{S}$ is a set of points $\{x_1, \dots x_n\}$ in $\mathcal{X}$, we can compute the distance in $\mathcal{H}$ between this set and a point of $\mathcal{X}$ using the barycenter. We speak of similarity between a point in $\mathcal{X}$ and a set $\mathcal{S}$. To do this, we map all the points of $\mathcal{S}$ in the feature space, we *summarize* $\mathcal{S}$ by its barycenter

$$\mu := \frac{1}{n}\sum_{i=1}^{n} \phi(x_i)$$

which leads to the distance between x and $\mathcal{S}$

$$d_{\mathcal{H}}(x,\mu) = \|\phi(x) - \mu\|_{\mathcal{H}} = \sqrt{K(x,x) - \frac{2}{n}\sum_{i=1}^{n} K(x,x_i) + \frac{1}{n^2}\sum_{i,j=1}^{n} K(x_i,x_j)}.$$

Definition. Let $\mathcal{S}$ be a set of points $\{x_1, \dots x_n\}$ in $\mathcal{X}$. The mapping function corresponding to the centered data of $\phi : \mathcal{S} \to \mathcal{H}$ in the feature space is defined by

$$\forall x \in \mathcal{S}, \quad \tilde{\phi}(x) = \phi(x) - \mu.$$

where μ is the barycenter of the elements of $\mathcal{S}$ in the feature space.

Proposition 8. Let $\mathcal{S} = \{x_1, \dots, x_n\} \in \mathcal{X}$ be a set of points, and let K be a kernel defined on $\mathcal{X}$. If $\mathbf{K}$ is the symmetric Gram matrix of size $n \times n$, whose entries are $K(x_i,x_j) = \kappa_{x_i x_j}$ then the Gram matrix $\tilde{\mathbf{K}}$ associated with the centered data $\tilde{K}(x_i,x_j) = \tilde{\kappa}_{x_i x_j}$ has entries

$$\tilde{\kappa}_{x_i x_j} = \langle \tilde{\phi}(x_i), \tilde{\phi}(x_j) \rangle_{\mathcal{H}} = \kappa_{x_i x_j} - \frac{1}{n}\sum_{l=1}^{n}(\kappa_{x_i x_l} + \kappa_{x_j x_l}) + \frac{1}{n^2}\sum_{l,l'=1}^{n} \kappa_{x_l x_{l'}}$$

which can be rewritten in a matrix form:

$$\tilde{\mathbf{K}} = (I_n - \frac{1}{n}\mathbb{1}_{n\times n})\mathbf{K}(I_n - \frac{1}{n}\mathbb{1}_{n\times n}),$$

where $\mathbb{1}_{n\times n} \in \mathcal{M}_n(\mathbb{R})$ has entries 1.
The centered function $\tilde{K}(x,y) = \langle \phi(x) - \mu, \phi(y) - \mu \rangle$, for $x, y \in \mathcal{S}$, defines a centered kernel.

Proof. We define for every $0 \le i,j \le n$

$$\begin{aligned}\tilde{\kappa}_{x_i x_j} = \langle \tilde{\phi}(x_i), \tilde{\phi}(x_j) \rangle_{\mathcal{H}} &= \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{H}} - \langle \mu, \phi(x_i) + \phi(x_j) \rangle_{\mathcal{H}} + \langle \mu, \mu \rangle_{\mathcal{H}} \\ &= \kappa_{x_i x_j} - \frac{1}{n}\sum_{l=1}^{n}(\kappa_{x_i x_l} + \kappa_{x_j x_l}) + \frac{1}{n^2}\sum_{l,l'=1}^{n} \kappa_{x_l x_{l'}}.\end{aligned}$$

This can be rewritten in matrix form as

$$\tilde{\mathbf{K}} = \mathbf{K} - \frac{1}{n}\mathbb{1}_{n\times n}\mathbf{K} - \frac{1}{n}\mathbf{K}\mathbb{1}_{n\times n} + \frac{1}{n^2}\mathbb{1}_{n\times n}\mathbf{K}\mathbb{1}_{n\times n} = (I_n - \frac{1}{n}\mathbb{1}_{n\times n})\mathbf{K}(I_n - \frac{1}{n}\mathbb{1}_{n\times n}).$$

Finally, we show that $\tilde{K}$ is indeed a kernel.

$$\begin{aligned}
\tilde{K}(x,y) &= \langle \phi(x)-\mu, \phi(y)-\mu\rangle = \langle \phi(x),\phi(y)\rangle_{\mathcal{H}} - \langle \mu, \phi(x)+\phi(y)\rangle_{\mathcal{H}} + \langle \mu,\mu\rangle_{\mathcal{H}} \\
&= \langle \phi(x),\phi(y)\rangle_{\mathcal{H}} - \frac{1}{n}\sum_{i=1}^{n}\left(\langle \phi(x_i),\phi(x)\rangle_{\mathcal{H}} + \langle \phi(x_i),\phi(y)\rangle_{\mathcal{H}}\right) + \frac{1}{n^2}\sum_{l,l'=1}^{n}\langle \phi(x_l),\phi(x_{l'})\rangle_{\mathcal{H}} \\
&= K(x,y) - \frac{1}{n}\sum_{i=1}^{n}(K(x_i,x)+K(x_i,y)) + \frac{1}{n^2}\sum_{l,l'=1}^{n}K(x_l,x_{l'})
\end{aligned}$$

For every $x, y \in \mathcal{X}$, $\tilde{K}(x,y) \in \mathbb{R}$ because $\tilde{K}(x,y)$ is a sum of element of $\mathbb{R}$. Moreover, due to the symmetry of K,

$$\begin{aligned}
\tilde{K}(x,y) &= \langle \phi(x),\phi(y)\rangle_{\mathcal{H}} - \frac{1}{n}\sum_{i=1}^{n}\left(\langle \phi(x_i),\phi(x)\rangle_{\mathcal{H}} + \langle \phi(x_i),\phi(y)\rangle_{\mathcal{H}}\right) + \frac{1}{n^2}\sum_{l,l'}\langle \phi(x_l),\phi(x_{l'})\rangle_{\mathcal{H}} \\
&= \langle \phi(y),\phi(x)\rangle_{\mathcal{H}} - \frac{1}{n}\sum_{i=1}^{n}\left(\langle \phi(x_i),\phi(y)\rangle_{\mathcal{H}} + \langle \phi(x_i),\phi(x)\rangle_{\mathcal{H}}\right) + \frac{1}{n^2}\sum_{l,l'}\langle \phi(x_l),\phi(x_{l'})\rangle_{\mathcal{H}} \\
&= \tilde{K}(y,x)
\end{aligned}$$

To conclude, we show the positivity of $\tilde{K}$. For every $(\alpha_i)_{i\in[\![1,n]\!]} \in \mathbb{R}$ and $(x_i)_i \in \mathcal{X}$, we have:

$$\sum_{i,j=1}^{n}\alpha_i\alpha_j\tilde{K}(x_i,x_j) = \sum_{i,j=1}^{n}\langle \alpha_i\tilde{\phi}(x_i), \alpha_j\tilde{\phi}(x_j)\rangle_{\mathcal{H}} = \left\|\sum_{i=1}^{n}\alpha_i\tilde{\phi}(x_i)\right\|_{\mathcal{H}}^2 \geq 0.$$

□

The RKHS is a space of potentially non-linear functions and the norm of f measures the smoothness of f. In the following theorem, we seek a form of f that minimizes a regularization function g. In other words, the goal is to limit the complexity of the obtained solution.

Theorem 3. (**Representer theorem**) Let $\mathcal{X}$ be a space, $K : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ be a kernel on $\mathcal{X}$, and $\mathcal{H}$ be the corresponding RKHS. Let $\mathcal{S} = \{x_1, \dots x_n\} \subseteq \mathcal{X}$, and let g be a $(n+1)$-variable function defined as:

$$\begin{array}{rccc} g: & \mathbb{R}^{n+1} & \longrightarrow & \mathbb{R} \\ & (f(x_1),\dots,f(x_n),\|f\|_{\mathcal{H}}) & \longmapsto & q((f(x_1),\dots,f(x_n)) + \lambda\tilde{g}(\|f\|_{\mathcal{H}}) \end{array}$$

where $q(.)$ is a cost function that measures the *goodness-of-fit* of f to a given problem, $\lambda > 0$ is a scalar, and $\tilde{g}$ is a strictly increasing function.
If $f \in \mathcal{H}$ minimizes $g(f(x_1),\dots,f(x_n),\|f\|_{\mathcal{H}})$, then f belongs to $\mathrm{Span}(\kappa_{x_1},\dots\kappa_{x_n})$.

Proof. Let $\mathcal{H}_L$ be a finite dimensional subspace of $\mathcal{H}$. Thus, $f \in \mathcal{H}$ can be uniquely written as

$$f = f_L + f_{L^\perp}$$

with $f_L \in \mathcal{H}_L$ and $f_{L^\perp}$ in the orthogonal complement $\mathcal{H}_{L^\perp}$ of $\mathcal{H}_L$. By the Pythagorean theorem in

$\mathcal{H}$, we have

$$\|f\|_{\mathcal{H}}^2 = \|f_L\|_{\mathcal{H}}^2 + \|f_{L^\perp}\|_{\mathcal{H}}^2.$$

Thus, by the monotonicity of $\tilde{g}$, we have, for every $x_i \in \mathcal{S}$:

$$\begin{array}{rcl} \tilde{g}(\|f\|_{\mathcal{H}}) & \geq & \tilde{g}(\|f_L\|_{\mathcal{H}}) \\ & \Leftrightarrow & \\ g(f(x_1), \ldots, f(x_n), \|f\|_{\mathcal{H}}) & \geq & g(f(x_1), \ldots, f(x_n), \|f_L\|_{\mathcal{H}}). \end{array}$$

Now, $g(f(x_1), \ldots, f(x_n), \|f\|_{\mathcal{H}})$ is the minimum of g, so $g(f(x_1), \ldots, f(x_n), \|f\|_{\mathcal{H}}) = g(f(x_1), \ldots, f(x_n), \|f_L\|_{\mathcal{H}})$, and the minimum of g is reached when $\tilde{g}(\|f_L\|) = \tilde{g}(\|f\|)$, i.e., when $\|f_{L^\perp}\|_{\mathcal{H}} = 0$ because $\tilde{g}$ is strictly increasing. This implies that the minimum of g is reached when $f_{L^\perp} = 0$. Since $\mathcal{H}$ is a RKHS, by the reproducing property, we deduce that the minimum of g is reached when $f \in \text{Vect}(\kappa_{x_1}, \ldots, \kappa_{x_n})$ for every $x_i \in \mathcal{S}$.
This completes the proof. □

Remark/Consequences:

1. When the representer theorem holds, we know that we can look for a solution of the form

$$f = \sum_{i=1}^{n} c_i \kappa_{x_i} \qquad \text{for some } c \in \mathbb{R}^n.$$

and the norm of f is then

$$\|f\|_{\mathcal{H}}^2 = \left\| \sum_{i=1}^{n} c_i \kappa_{x_i} \right\|_{\mathcal{H}}^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j K(x_i, x_j) = c^\top \mathbf{K} c.$$

2. Therefore, seeking f that minimizes $g(f(x_1), \ldots, f(x_n), \|f\|_{\mathcal{H}})$ amounts to seeking $c \in \mathbb{R}^n$ that minimizes $g([\mathbf{K}c]_1, \ldots, [\mathbf{K}c]_n, c^\top \mathbf{K} c)$. This latter consequence is used in the next section.

3 From PCA to Kernel PCA

We consider $\mathcal{X}$ an arbitrary space containing n data points $(x_i)_{i \in [\![1,n]\!]}$. In order to extend PCA to such data, we introduce Kernel PCA which enables the comparison of data that may not be linearly separable or even numerical in nature. In this section, we will show how to adapt PCA to RKHS. Note that, since we no longer always have numerical variables, we should work on the dual version of the PCA (*i.e.*, computing the principal components in analogy with the standard case and then linking them to principal axes that are never explicitly computed). Even when $\mathcal{X}$ is the standard $\mathbb{R}^p$, Kernel PCA can present some advantages over PCA: in cases where data exhibit a non-linear separability, Kernel PCA provides enhanced separation capabilities and facilitates the detection of non-linear relationships.

To achieve this, we consider a positive kernel function, denoted as $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. As discussed in the previous section, the kernel function allows us to define an implicit Hilbert space $\mathcal{H}$ and an associated mapping function $\phi : \mathcal{X} \to \mathcal{H}$. For the purpose of our analysis, we consider centered data and a modified kernel $\tilde{K}$ defined as $\tilde{K}(x_i, x_j) = \langle \tilde{\phi}(x_i), \tilde{\phi}(x_j) \rangle_{\mathcal{H}}$, where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ denotes the inner product in the implicit reproducing kernel Hilbert space $\mathcal{H}$, and $\tilde{\phi} : \mathcal{X} \to \mathcal{H}$ represents the centered mapping function.

Furthermore, we denote the centered Gram matrix associated with the kernel as $\tilde{\mathbf{K}} = (\tilde{K}(x_i, x_j))_{i,j\in[\![1,n]\!]}$. This matrix is also called similarity matrix. This matrix captures the pairwise similarities between the centered data points using the kernel function. Once defined, we can use the same tools on this matrix as in standard PCA.

Definition. We define the global inertia of the data by

$$I_{\mathcal{H}}(x_1, \dots, x_n) := \sum_{i=1}^{n} \left\| \tilde{\phi}(x_i) \right\|_{\mathcal{H}}^2 = \sum_{i=1}^{n} \tilde{K}(x_i, x_i) = \mathrm{Tr}(\tilde{\mathbf{K}})$$

In addition, for $a_k \in \mathcal{H}$, we define the inertia of data projected onto a_k:

$$\begin{aligned} I_{\mathcal{H}}(P_{a_k}(x_1), \dots, P_{a_k}(x_n)) &:= \sum_{i=1}^{n} \|P_{a_k}(\tilde{\phi}(x_i))\|_{\mathcal{H}}^2 = \sum_{i=1}^{n} \left\| \frac{\langle \tilde{\phi}(x_i), a_k \rangle_{\mathcal{H}}}{\|a_k\|_{\mathcal{H}}^2} a_k \right\|_{\mathcal{H}}^2 \\ &= \sum_{i=1}^{n} \frac{1}{\|a_k\|_{\mathcal{H}}^4} \left\langle \langle \tilde{\phi}(x_i), a_k \rangle_{\mathcal{H}} a_k, \langle \tilde{\phi}(x_i), a_k \rangle_{\mathcal{H}} a_k \right\rangle_{\mathcal{H}} = \sum_{i=1}^{n} \frac{\langle \tilde{\phi}(x_i), a_k \rangle_{\mathcal{H}}^2}{\|a_k\|_{\mathcal{H}}^2} \end{aligned}$$

where $P_{a_k}(\tilde{\phi}(x_i))$ is the projection of $\tilde{\phi}(x_i)$ onto a_k.
Furthermore, when a_k has a norm of 1, $I_{\mathcal{H}}(P_{a_k}(x_1), \dots, P_{a_k}(x_n)) = \sum_{i=1}^{n} \langle \tilde{\phi}(x_i), a_k \rangle_{\mathcal{H}}^2$.

Here, the inertia of projected data onto a_k is defined by the inner product that induced the RKHS $\mathcal{H}$. As we want to adapt PCA to the RKHS, we focus on the dual version of the PCA using the Gram matrix $\tilde{\mathbf{K}}$. This will allow us to define inertia based on $\tilde{\mathbf{K}}$.

Since $\tilde{\mathbf{K}}$ is a positive and symmetric matrix, we can write its spectral decomposition as $\tilde{\mathbf{K}} = UDU^\top$ with $D = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$, where the eigenvalues are positive and arranged in decreasing order, and U is the rotation matrix containing orthonormal eigenvectors $(u_1, \dots, u_n)$ as its columns.

Definition. We define the principal components c_k as

$$\forall k \in [\![1, n]\!], \qquad c_k := \sqrt{\lambda_k} u_k.$$

Proposition 9. Let $k \in [\![1, n]\!]$. The principal components are orthogonal for $\tilde{\mathbf{K}}$ and the norm c_k induced by the matrix $\tilde{\mathbf{K}}$ is λ_k^2.

Proof. Since the column vectors, $(u_k)_k$ of the rotation matrix are orthonormal, we have

$$\forall k \in [\![1, n]\!] \qquad c_k^\top \tilde{\mathbf{K}} c_{k'} = \sqrt{\lambda_k \lambda_{k'}} u_k^\top \tilde{\mathbf{K}} u_{k'} = \lambda_{k'} \sqrt{\lambda_k \lambda_{k'}} u_k^\top u_{k'} = \begin{cases} \lambda_k^2 & \text{if} \quad k = k' \\ 0 & \text{otherwise} \end{cases}$$

which proves that $(c_k)_k$ are orthogonal for $\tilde{\mathbf{K}}$ and the induced norm of c_k is λ_k^2 for the norm induced by the matrix $\tilde{\mathbf{K}}$. □

Remark: The principal components are straightforwardly eigenvectors of the matrix $\tilde{\mathbf{K}}$ associated with the eigenvalues $(\lambda_k)_k$.

To establish a link between the Gram matrix and the inertia on the projected data, we define the covariance operator and the principal axes.

Definition. The covariance operator of $\tilde{\phi}(x_i)$ for $i \in [\![1, n]\!]$ is defined by

$$\Gamma = \sum_{i=1}^{n} \langle \tilde{\phi}(x_i), . \rangle_{\mathcal{H}} \tilde{\phi}(x_i).$$

Definition. The principal axes are defined by

$$\forall k \in [\![1, r]\!], \quad a_k := \sum_{j=1}^{n} \frac{1}{\lambda_k} c_{k,j} \tilde{\phi}(x_j).$$

Proposition 10. The principal axes $(a_k)_{k=1,\dots,r}$ are the eigenfunctions of Γ associated with the eigenvalues $(\lambda_k)_{k=1,\dots,r}$. They are orthonormal.

Proof. First, let us show that a_k is the eigenfunction of Γ associated with λ_k.

$$\begin{aligned}
\Gamma a_k &= \sum_{i=1}^{n} \langle \tilde{\phi}(x_i), a_k \rangle_{\mathcal{H}} \tilde{\phi}(x_i) = \sum_{i=1}^{n} \tilde{\phi}(x_i) \sum_{j=1}^{n} \frac{1}{\lambda_k} c_{k,j} \langle \tilde{\phi}(x_i), \tilde{\phi}(x_j) \rangle_{\mathcal{H}} \\
&= \sum_{i=1}^{n} \tilde{\phi}(x_i) \sum_{j=1}^{n} \frac{1}{\lambda_k} c_{k,j} \tilde{K}(x_i, x_j) = \sum_{i=1}^{n} \tilde{\phi}(x_i) \frac{1}{\lambda_k} e_i^\top \tilde{\mathbf{K}} c_k \\
&= \lambda_k \sum_{i=1}^{n} \frac{1}{\lambda_k} c_{k,i} \tilde{\phi}(x_i) = \lambda_k a_k
\end{aligned}$$

where $e_i \in \mathbb{R}^n$ denotes the i-th vector of the canonical basis.
On the other hand, since the vectors $(c_k)_k$ are orthogonal for $\tilde{\mathbf{K}}$, and their norm induced by the matrix $\tilde{\mathbf{K}}$ is λ_k^2, we have

$$\begin{aligned}
\langle a_k, a_{k'} \rangle_{\mathcal{H}} &= \sum_{i,j=1}^{n} \frac{1}{\lambda_{k'}} \frac{1}{\lambda_k} c_{k',i} c_{k,j} \langle \tilde{\phi}(x_i), \tilde{\phi}(x_j) \rangle_{\mathcal{H}} \\
&= \sum_{i,j=1}^{n} \frac{1}{\lambda_{k'}} \frac{1}{\lambda_k} c_{k',i} c_{k,j} \tilde{K}(x_i, x_j) = \frac{1}{\lambda_{k'}} \frac{1}{\lambda_k} c_k^\top \tilde{\mathbf{K}} c_{k'} = \delta_{k,k'}
\end{aligned}$$

which proves the orthonormality of $(a_k)_k$. □

Proposition 11. The inertia of data projected onto a principal axis a_k can be rewritten with $\tilde{\mathbf{K}}$ as

$$I_{\mathcal{H}}(P_{a_k}(x_1), \dots, P_{a_k}(x_n)) = \sum_{i=1}^{n} \langle \tilde{\phi}(x_i), a_k \rangle_{\mathcal{H}}^2 = \frac{1}{\lambda_k^2} c_k^\top \tilde{\mathbf{K}}^2 c_k$$

We note also that the inertia of the data projected onto a_k is equal to the eigenvalue λ_k associated to a_k.

Proof. Indeed, we have

$$\begin{aligned}\sum_{i=1}^{n}\langle\tilde{\phi}(x_i),a_k\rangle_{\mathcal{H}}^2 &= \sum_{i=1}^{n}\left(\sum_{j=1}^{n}\frac{1}{\lambda_k}c_{k,j}\langle\tilde{\phi}(x_i),\tilde{\phi}(x_j)\rangle_{\mathcal{H}}\right)^2\\ &= \frac{1}{\lambda_k^2}\sum_{i=1}^{n}\left(\sum_{j=1}^{n}c_{k,j}\tilde{K}(x_i,x_j)\right)^2\\ &= \frac{1}{\lambda_k^2}(\tilde{\mathbf{K}}c_k)^\top(\tilde{\mathbf{K}}c_k)\\ &= \frac{1}{\lambda_k^2}c_k^\top\tilde{\mathbf{K}}^2c_k.\end{aligned}$$

Moreover, since c_k is an eigenvector of $\tilde{\mathbf{K}}$ and $c_k^\top\tilde{\mathbf{K}}c_k=\lambda_k^2$, we have,

$$\frac{1}{\lambda_k^2}c_k^\top\tilde{\mathbf{K}}^2c_k=\frac{1}{\lambda_k^2}\times\lambda_k c_k^\top\tilde{\mathbf{K}}c_k=\lambda_k$$

which completes the proof. □

Proposition 12. The coordinates of the projection of $\tilde{\phi}(x_i)$ onto the principal axis a_k is

$$\langle\tilde{\phi}(x_i),a_k\rangle_{\mathcal{H}}=c_{k,i}$$

Proof. Indeed, for $(i,k)\in[\![1,n]\!]\times[\![1,r]\!]$,

$$\begin{aligned}\langle\tilde{\phi}(x_i),a_k\rangle_{\mathcal{H}} &= \frac{1}{\lambda_k}\sum_{j=1}^{n}c_{k,j}\langle\tilde{\phi}(x_i),\tilde{\phi}(x_j)\rangle_{\mathcal{H}}\\ &= \frac{1}{\lambda_k}e_i^\top\tilde{\mathbf{K}}c_k=\frac{1}{\lambda_k}\lambda_k c_{k,i}=c_{k,i}\end{aligned}$$

where $c_{k,i}$ is the i-th term of the k-th principal component. □

As in the standard PCA, these coordinates are usefull to obtain a low-dimensional representation of the sample that emphasizes its underlying structure.

As the observations are described though their similarities thanks to the kernel rather than by numerical values, the principal components are then more difficult to interpret. In particular, Kernel PCA does not allow the representation of variables, unlike standard PCA. However, as standard PCA, Kernel PCA representation corresponds to the projection onto the subspace of $\mathcal{H}$ spanned by orthonormal functions that maximize the inertia.

Lemma 2. We consider the positive eigenvalues λ_k for $k \in [\![1, r]\!]$ with $r \leq n$ where r is the last index where $\lambda_r \neq 0$.
The vectors $\{b_1, \ldots, b_r\} \subset \mathbb{R}^n$ solutions of the problem

$$\text{argmax} \quad b_k^\top \tilde{\mathbf{K}}^2 b_k \tag{1}$$

under the contraints, for $k \in [\![1, r]\!]$

$$\begin{cases} b_k^\top \tilde{\mathbf{K}} b_k = 1 \\ b_k^\top \tilde{\mathbf{K}} b_{k'} = 0 \text{ for } k' \in [\![1, k-1]\!] \end{cases}$$

are given by, $\forall\, k \in [\![1, r]\!], \quad b_k = \frac{1}{\sqrt{\lambda_k}} u_k$.

Proof. Let $e_i \in \mathbb{R}^n$ denote the i-th vector of the canonical basis and D be the diagonal matrix $D = \text{diag}(\lambda_1, \ldots, \lambda_n)$. We proceed by induction.
Initialization ($k = 1$): Assuming $\lambda_1 \neq 0$. Let $b_1 \in \mathbb{R}^n$ be a vector such that $b_1^\top \tilde{\mathbf{K}} b_1 = 1$ and that maximizes the quantity $b^\top \tilde{\mathbf{K}}^2 b$. Using the spectral decomposition of $\tilde{\mathbf{K}}$, we have $\tilde{\mathbf{K}}^2 = U D^2 U^\top$ and then

$$b_1^\top \tilde{\mathbf{K}}^2 b_1 = b_1^\top U D^2 U^\top b_1 = q_1^\top D q_1.$$

with $q_1 = D^{1/2} U^\top b_1$. Since $q_1^\top q_1 = b_1^\top U D^{1/2} D^{1/2} U^\top b_1 = b_1^\top \tilde{\mathbf{K}} b_1 = 1$, we have,

$$q_1^\top D q_1 = \sum_{i=1}^n \lambda_i q_{1,i}^2 \leq \lambda_1 \sum_{i=1}^n q_{1,i}^2 = \lambda_1.$$

The quantity $b_1^\top \tilde{\mathbf{K}}^2 b_1$ is maximized if and only if $q_1^\top D q_1$ is maximized, which is achieved when $q_1 = (1\ 0\ \ldots\ 0)^\top$. Let us consider the system resulting from the expression $q_1 = D^{1/2} U^\top b_1$:

$$\begin{cases} \sqrt{\lambda_1} u_1^\top b_1 = 1 \\ \sqrt{\lambda_k} u_k^\top b_1 = 0 \quad \text{for} \quad k \in [\![2, n]\!]. \end{cases}$$

Thus, we can deduce, thanks to the orthonormality of the $(u_k)_k$ that b_1 and u_1 are co-linear, i.e. there exists $\alpha \in \mathbb{R}^*$ such that $b_1 = \alpha u_1$, then

$$\sqrt{\lambda_1} u_1^\top b_1 = 1 \Leftrightarrow \sqrt{\lambda_1} \alpha u_1^\top u_1 = 1 \Leftrightarrow \alpha = \frac{1}{\sqrt{\lambda_1}}$$

Hence, for the case $k = 1$, $b_1 = \frac{1}{\sqrt{\lambda_1}} u_1$ maximizes the quantity $b_1^\top \tilde{\mathbf{K}}^2 b_1$ under the contraint $b_1^\top \tilde{\mathbf{K}} b_1 = 1$.
Induction step ($k \geq 2$)**:** The vectors $\{b_1, \ldots, b_{k-1}\} \subset \mathbb{R}^n$ solutions of the problem

$$\text{argmax} \quad b_{k'}^\top \tilde{\mathbf{K}}^2 b_{k'}$$

under the contraints, for $k' \in [\![1, k-1]\!]$

$$\begin{cases} b_{k'}^\top \tilde{\mathbf{K}} b_{k'} = 1 \\ b_{k'}^\top \tilde{\mathbf{K}} b_i = 0 \text{ for } i \in [\![1, k'-1]\!] \end{cases}$$

are given by $b_{k'} := \frac{1}{\sqrt{\lambda_{k'}}} u_{k'}$ for $k' \in [\![1, k-1]\!]$.

We want to show that the vector that satisfies argmax $b_k^\top \tilde{\mathbf{K}}^2 b_k$ under the contraint $b_k^\top \tilde{\mathbf{K}} b_k = 1$ and $b_k^\top \tilde{\mathbf{K}} b_{k'} = 0$ for $k' \in [\![1, k-1]\!]$ is $\frac{1}{\sqrt{\lambda_k}} u_k$.

We assume $\lambda_{k'} \neq 0$ for $k' \in [\![1, k]\!]$. We seek a vector b_k such that for any vector $b_{k'} \in C_{k-1}$, $b_k^\top \tilde{\mathbf{K}} b_{k'} = 0$ and $b_k^\top \tilde{\mathbf{K}} b_k = 1$ and that maximizes the quantity $b_k^\top \tilde{\mathbf{K}}^2 b_k$. Using the spectral decomposition of $\tilde{\mathbf{K}}$ we have,

$$b_k^\top \tilde{\mathbf{K}}^2 b_k = b_k^\top U D^2 U^\top b_k = q_k^\top D q_k$$

where $q_k := D^{1/2} U^\top b_k$. Moreover,

$$\begin{aligned} \forall k' \in [\![1, k-1]\!], \quad q_{k,k'} &= q_k^\top e_{k'} = \frac{1}{\sqrt{\lambda_{k'}}} q_k^\top D^{1/2} e_{k'} \\ &= \frac{1}{\sqrt{\lambda_{k'}}} b_k^\top U D^{1/2}\ D^{1/2} U^\top U e_{k'} = \frac{1}{\sqrt{\lambda_{k'}}} b_k^\top \tilde{\mathbf{K}} U e_{k'} = b_k^\top \tilde{\mathbf{K}} b_{k'} = 0 \end{aligned}$$

which implies $q_k^\top D q_k = \sum_{i=1}^n \lambda_i q_{k,i}^2 = \sum_{i=k}^n \lambda_i q_{k,i}^2$. Since $q_k^\top q_k = b_k^\top \tilde{\mathbf{K}} b_k = 1$, we have the inequality

$$b_k^\top \tilde{\mathbf{K}}^2 b_k = q_k^\top D q_k = \sum_{i=k}^n \lambda_i q_{k,i}^2 \leq \sum_{i=k}^n \lambda_k q_{k,i}^2 = \lambda_k.$$

Thus, the maximum is reached when $q_k^\top D q_k = \lambda_k$, which is achieved when $q_k = e_k$. Since $q_k := D^{1/2} U^\top b_k$ we find the expression of b_k by solving the system

$$\begin{cases} \sqrt{\lambda_k} u_k^\top b_k = 1 \\ \sqrt{\lambda_j} u_j^\top b_k = 0 \quad \text{for} \quad j \in [\![1, n]\!] \backslash \{k\}. \end{cases}$$

In other words, b_k is co-linear to the vector u_k since the vectors $(u_k)_k$ are orthonormal. i.e. there exists a scalar α such that $b_k = \alpha u_k$ then

$$\sqrt{\lambda_k} u_k^\top b_k = 1 \Leftrightarrow \sqrt{\lambda_k} \alpha u_k^\top u_k = 1 \Leftrightarrow \alpha = \frac{1}{\sqrt{\lambda_k}}$$

Hence, the vector $b_k = \frac{1}{\sqrt{\lambda_k}} u_k$ maximizes the quantity argmax $b_k^\top \tilde{\mathbf{K}}^2 b_k$ under the contraint $b_k^\top \tilde{\mathbf{K}} b_k = 1$ and $b_k^\top \tilde{\mathbf{K}} b_{k'} = 0$ for $k' \in [\![1, k-1]\!]$.

This completes the proof. □

Theorem 4. We consider the positive eigenvalues λ_k for $k \in [\![1, r]\!]$ with $r \leq n$ where r is the last index where $\lambda_r \neq 0$. The functions $\{a_1, \ldots, a_r\}$ are the solutions of the problem

$$\begin{cases} \text{argmax} \quad I_{\mathcal{H}}(P_{a_1}(x_1), \ldots, P_{a_1}(x_n)) \\ \text{argmax}_{a_k \perp (a_1, \ldots, a_{k-1})} I_{\mathcal{H}}(P_{a_k}(x_1), \ldots, P_{a_k}(x_n)), \quad \text{if } k \in [\![2, r]\!] \end{cases} \tag{2}$$

under the contraints $\|a_k\|_{\mathcal{H}} = 1$.

Proof. We seek for orthonormal functions f_k of $\mathcal{H}$, for $k \in [\![1, r]\!]$, that are solution of the Problem (2).
Let $f_k \in \mathcal{H}$, for $k \in [\![1, r]\!]$, be functions with norm 1, then we have

$$I_{\mathcal{H}}(P_{f_k}(x_1), \dots, P_{f_k}(x_n)) = \sum_{i=1}^{n} \langle \tilde{\phi}(x_i), f_k \rangle_{\mathcal{H}}^2.$$

Let us show that minimizing the reconstruction error $\sum_{k=1}^{r} \sum_{i=1}^{n} \|\tilde{\phi}(x_i) - P_{f_k}(x_i)\|^2$ is equivalent to maximizing $\sum_{k=1}^{r} I_{\mathcal{H}}(P_{f_k}(x_1), \dots, P_{f_k}(x_n))$. We have

$$\begin{aligned}
\sum_{i=1}^{n} \|\tilde{\phi}(x_i) - P_{f_k}(x_i)\|^2 &= \sum_{i=1}^{n} \|\tilde{\phi}(x_i) - \langle \tilde{\phi}(x_i), f_k \rangle_{\mathcal{H}}\|^2 \\
&= \sum_{i=1}^{n} (\|\tilde{\phi}(x_i)\|^2 - 2\langle \phi(x_i), f_k \rangle_{\mathcal{H}}^2 + (\langle \tilde{\phi}(x_i), f_k \rangle_{\mathcal{H}} \|f_k\|_{\mathcal{H}})^2) \\
&= \sum_{i=1}^{n} (\|\tilde{\phi}(x_i)\|^2 - \langle \phi(x_i), f_k \rangle_{\mathcal{H}}^2) \\
&= I_{\mathcal{H}}(x_1, \dots, x_n) - I_{\mathcal{H}}(P_{f_k}(x_1), \dots, P_{f_k}(x_n)).
\end{aligned}$$

Then, since the global inertia is a constant, we have

$$\begin{aligned}
\operatorname{argmin} \sum_{k=1}^{r} \sum_{i=1}^{n} \|\tilde{\phi}(x_i) - P_{f_k}(x_i)\|^2 &= \operatorname{argmin} \sum_{k=1}^{r} I_{\mathcal{H}}(x_1, \dots x_n) - I_{\mathcal{H}}(P_{f_k}(x_1), \dots, P_{f_k}(x_n)) \\
&= -\operatorname{argmin} \sum_{k=1}^{r} I_{\mathcal{H}}(P_{f_k}(x_1), \dots, P_{f_k}(x_n)) \\
&= \operatorname{argmax} \sum_{k=1}^{r} I_{\mathcal{H}}(P_{f_k}(x_1), \dots, P_{f_k}(x_n)).
\end{aligned}$$

We will now complete the proof by induction.
Initialization ($k = 1$): Let us prove that the function f_1 that maximizes $I_{\mathcal{H}}(P_{f_1}(x_1), \dots, P_{f_1}(x_n))$ under the constraint $\|f_1\|_{\mathcal{H}} = 1$ is a_1. Since the maximization of the projected inertia is equivalent to minimizing $\sum_{i=1}^{n} \|\tilde{\phi}(x_i) - P_{f_1}(x_i)\|^2$, the Representer Theorem 3 shows that f_1 can be written as $f_1 = \sum_{i=1}^{n} b_i \tilde{\phi}(x_i)$ for some $b_i \in \mathbb{R}$. Therefore, for $b = (b_1, \dots, b_n) \in \mathbb{R}^n$,

$$\|f_1\|_{\mathcal{H}}^2 = b^\top \tilde{K} b$$

which leads to the constraint $b^\top \tilde{K} b = 1$. In addition,

$$\forall\, i = 1, \dots, n, \qquad P_{f_1}(x_i) = \sum_{i'=1}^{n} b_{i'} \tilde{K}(x_i, x_{i'})$$

which leads to show that

$$I_{\mathcal{H}}(P_{f_1}(x_1), \dots, P_{f_1}(x_n)) = \sum_{i=1}^{n} \left(\sum_{i'=1}^{n} b_{i'} K(x_i, x_{i'}) \right)^2 = b^\top \tilde{K}^2 b.$$

Hence, by Lemma 2, $f_1 = \sum_{i=1}^{n} \frac{1}{\sqrt{\lambda_1}} u_{1i} \tilde{\phi}(x_i) = \sum_{i=1}^{n} \frac{1}{\lambda_1} c_{1i} \tilde{\phi}(x_i) = a_1$, which concludes the case $k = 1$.

Induction step ($k \geq 2$): Suppose that $a_1, \ldots, a_{k-1}$ are the orthogonal functions that, for all $k' \leq k-1$ maximizes the projected inertia $\sum_{l=1}^{k'} I_{\mathcal{H}}(P_{f_l}(x_1), \ldots, P_{f_l}(x_n))$ under the constraints that $\|f_{k'}\|_{\mathcal{H}} = 1$ and $\langle f'_k, f_i \rangle_{\mathcal{H}} = 0$ for all $i \leq k' - 1$.

Let us prove that a_k is the function, orthogonal to all $(a_{k'})_{k' \leq k-1}$ and with norm 1, that maximizes $\sum_{l=1}^{k-1} I_{\mathcal{H}}(P_{a_l}(x_1), \ldots, P_{a_l}(x_n)) + I_{\mathcal{H}}(P_{f_k}(x_1), \ldots, P_{f_k}(x_n))$. This question is equivalent to maximizing only $I_{\mathcal{H}}(P_{f_k}(x_1), \ldots, P_{f_k}(x_n))$ under the constraints $\|f_k\|_{\mathcal{H}} = 1$ and $\langle f_k, a_{k'} \rangle_{\mathcal{H}} = 0$ for all $k' \leq k-1$. Using the fact that maximizing $I_{\mathcal{H}}(P_{f_k}(x_1), \ldots, P_{f_k}(x_n))$ is equivalent to minimizing $\sum_{i=1}^{n} \|\tilde{\phi}(x_i) - P_{f_k}(x_i)\|^2$, we again show that f_k can be written as $f_k = \sum_{i=1}^{n} b_i \tilde{\phi}(x_i)$ for some $b_i \in \mathbb{R}$. With the same arguments that for the case $k = 1$, we show that b maximizes

$$b^\top \tilde{K}^2 b$$

under the constraint that $b^\top \tilde{K} b = 1$ and the additional constraints

$$\forall\, k' = 1, \ldots, k-1, \qquad \left\langle \sum_{i=1}^{n} b_i \tilde{\phi}(x_i), \sum_{i=1}^{n} \frac{1}{\lambda_{k'}} c_{k'i} \tilde{\phi}(x_i) \right\rangle_{\mathcal{H}} = 0 \quad \Leftrightarrow b^\top \tilde{K} c_{k'} = 0.$$

Again using Lemma 2, we conclude that $f_k = \sum_{i=1}^{n} \frac{1}{\sqrt{\lambda_k}} u_{ki} \tilde{\phi}(x_i) = a_k$, which concludes the proof. □

Remark: In order to compare kernel PCA to standard PCA, we consider the kernel defined by $K(x, y) = x^\top y$. Thus $\tilde{\mathbf{K}} = X' X'^\top$, where $X' = \frac{1}{\sqrt{n}} \bar{X}$. Since the principal components are eigenvectors of $\frac{1}{n} \bar{X} \bar{X}^\top$ and c_k is an eigenvector of $\tilde{\mathbf{K}}$, we identify the vectors c_k to the principal components of the standard PCA.

4 Application

As mentioned earlier, performing kernel PCA on numerical data can potentially lead to better separability of the data. In standard PCA the graph representing the individuals on the first two axes shows a good separability of the individuals.

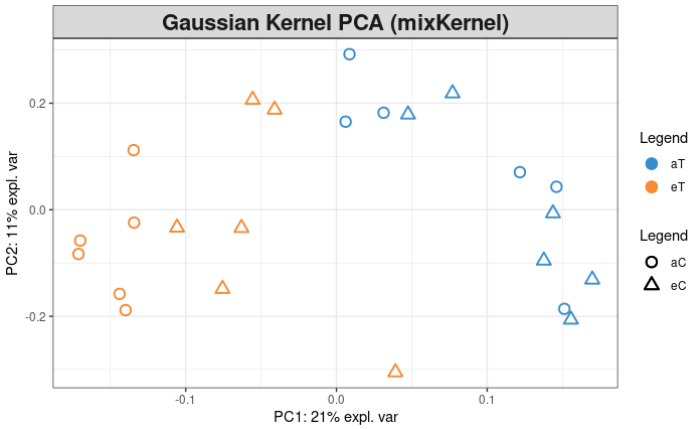

Figure 5: Gaussian Kernel PCA

Similarly, when applying Gaussian kernel PCA, we can expect to see a similar interpretation, where the individuals are well-separated in the transformed feature space defined by the Gaussian kernel.

Conclusion

To conclude, my internship within the INRAE project "PeerSim" has provided a theoretical understanding of kernel methods for exploratory analysis. While the focus was primarily on theory, practical aspects were also incorporated to complement the understanding of these methods.

The theoretical part of the internship allowed me to discover and deepen the PCA and its extension on RKHS, kernel PCA. In the context of the project, this extension would enable the comparison of gene networks.

By exploring the theory behind PCA, I have gained mathematical understanding of the concept of "optimal dimension reduction" and how to effectively reduce the dimension of a space of interest while preserving "essential" information of a dataset. This understanding allows meaningful exploratory analysis of multivariate data by maximizing the variability of the projection. This is achieved through the eigendecomposition of the variance-covariance matrix.

The extension kernel PCA expanded the capabilities of PCA, allowing for the comparison and analysis of non-numerical data, such as gene networks. Thus, by exploring the theory behind kernel PCA, I gained an understanding of how to compare graphs using similarity matrix. Such matrix is defined by kernels which measure the similarities between each graphs. Thus, to understand how to adapt PCA to RKHS, I focused on the theoretical aspects of kernels. First the theoretical part on RKHS shows that defining a kernel directly induces an implicit RKHS and an mapping function. Building on the theoretical knowledge of kernels and the tools of PCA, we demonstrated how to adapt PCA to RKHS with Kernel PCA. Specifically, we explored the dual version of PCA using the similarity matrix. Through this similarity matrix, the induced mapping function and the induced Hilbert space, we determined the objects in kernel PCA which correspond to objects in PCA and we obtained the coordinates of the projected individuals. These coordinates are solely determined by the principal components which are the eigenvectors of the Gram matrix. Thus, there is no need to know the explicit form of the Hilbert space and mapping function to achieve dimension reduction. Instead, focusing on studying the similarity matrix and its eigendecomposition allows for graph comparison.

Until now, our interest has been focused on the graph comparison using an arbitrary kernel. However, a natural question arises: what would be a good kernel for determining similarities between gene networks? Additionally, as my internship initially aimed to explore the impact of the number of replicates on the biological interpretation of statistical results, it would be necessary to construct graphs capturing gene relationships from gene expression matrices with varying numbers of replicates in order to compare and analyze them. This is left for future work in the project.

References

[1] Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Space in Probability and Statistics*. Jan. 2004. ISBN: 978-1-4613-4792-7. DOI: 10.1007/978-1-4419-9096-9.

[2] Thomas Gärtner, Peter Flach, and Stefan Wrobel. "On Graph Kernels: Hardness Results and Efficient Alternatives". In: *Learning Theory and Kernel Machines*. Ed. by Bernhard Schölkopf and Manfred K. Warmuth. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 129–143. ISBN: 978-3-540-45167-9.

[3] Jérôme J. Mariette and Nathalie Vialaneix. "Des noyaux pour les omiques". In: *Intégration de Données Biologiques*. July 2022. URL: https://hal.inrae.fr/hal-03809784.

[4] Jan Ramon and Thomas Gartner. "Expressivity versus Eciency of Graph Kernels". In: (Aug. 2004).

[5] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004. DOI: 10.1017/CBO9780511809682.

[6] S.V.N. Vishwanathan et al. "Graph Kernels". In: *Journal of Machine Learning Research* 11.40 (2010), pp. 1201–1242. URL: http://jmlr.org/papers/v11/vishwanathan10a.html.

[7] Wikistat. *Analyse en Composantes Principales— Wikistat*. [En ligne; Page disponible le 21-janvier-2016]. 2016. URL: http://wikistat.fr/pdf/st-m-explo-acp.pdf.