Lecture 7. Functional PCA Computation

Functional PCA, Kernel PCA, Nonlinear Functional PCA

Jun Song

Department of Statistics
Korea University

Table of contents

Coordinate representation

Introduction

- We will assume that $X_1, \ldots, X_N \in \mathcal{H}$ are fully observed.
- $\mathcal{H}$ has a basis (finite).

Coordinate representation

- Let $\mathcal{H}_1$ be a vector space with basis $\mathcal{B}=\{b_1, \ldots, b_n\}$.
- For each $f \in \mathcal{H}_1$, there is a vector $\alpha = (\alpha_1, \ldots, \alpha_n)^\mathsf{T}$ such that $f = \sum_{i=1}^{n} \alpha_i b_i$.
- $[f]_\mathcal{B} = \alpha$ is a coordinate representation of f w.r.t. $\mathcal{B}$
- Let $\mathcal{H}_2$ be another Hilbert spaces, spanned by $\mathcal{C} = \{c_1, \ldots, c_m\}$
- If $A : \mathcal{H}_1 \to \mathcal{H}_2$ is a linear operator. Then, it can be shown that, for any $f \in \mathcal{H}_1$,

$$[Af]_\mathcal{C} = ({}_\mathcal{C}[A]_\mathcal{B})[f]_\mathcal{B},$$

where

$${}_\mathcal{C}[A]_\mathcal{B} = ([Ab_1]_\mathcal{C}, \ldots, [Ab_n]_\mathcal{C})$$

Coordinate representation

- Let $\mathcal{H}_1$ be a vector space with basis $\mathcal{B}=\{b_1, \ldots, b_n\}$.
- Let $\mathcal{H}_2$ be another Hilbert spaces, spanned by $\mathcal{C} = \{c_1, \ldots, c_m\}$
- Let $\mathcal{H}_3$ be another Hilbert spaces, spanned by $\mathcal{D} = \{d_1, \ldots, d_\ell\}$
- If $A : \mathcal{H}_1 \to \mathcal{H}_2$, $B : \mathcal{H}_2 \to \mathcal{H}_3$ are linear operators. Then, it can be shown that, for any $f \in \mathcal{H}_1$,

$$[BAf]_{\mathcal{C}} = ({}_{\mathcal{D}}[B]_{\mathcal{C}})({}_{\mathcal{C}}[A]_{\mathcal{B}})[f]_{\mathcal{B}},$$

- Subscripts omitted if it is trivial.

Coordinate Representation Toolkit

Let G be a gram matrix for $\mathcal{H}_1$ w.r.t. $\mathcal{B}$. i.e., $G \in \mathbb{R}^{n \times n}$ such that $(G)_{ij} = \langle b_i, b_j \rangle_{\mathcal{H}_1}$.
Suppose that $x_1, \ldots, x_N \in \mathcal{H}_1$

1. $\langle x_i, x_j \rangle = [x_i]^\intercal G [x_j]$
2. For $g \in \mathcal{H}_1$, $f \in \mathcal{H}_2$, ${}_c[g \otimes f]_{\mathcal{B}} = [g]_c [f]_{\mathcal{B}}^\intercal G$

   Let $Q_N = I_N - N^{-1} 1_N 1_N^\intercal$ and $[x_{1:N}] \in \mathbb{R}^{n \times N}$ whose columns are $[x_i]$.
   $\hat{C} = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x}) \otimes (x_i - \bar{x})$
3. $[\hat{C}] = N^{-1} [x_{1:N}] Q_N [x_{1:N}]^\intercal G$
4. $[\hat{C}^\alpha] = G^{-1/2} (N^{-1} G^{1/2} ([x_{1:N}] Q_N [x_{1:N}] [x_{1:N}]^\intercal G^{1/2})^\alpha G^{1/2}$

Coordinate Representation Toolkit: Proof 1-2

Let G be a gram matrix for $\mathcal{H}_1$ w.r.t. $\mathcal{B}$. i.e., $G \in \mathbb{R}^{n \times n}$ such that $(G)_{ij} = \langle b_i, b_j \rangle_{\mathcal{H}_1}$.
Suppose that $x_1, \ldots, x_N \in \mathcal{H}_1$

1. $\langle x_i, x_j \rangle = [x_i]^\mathsf{T} G [x_j]$
   $\langle x_i, x_j \rangle = \langle \sum_k [x_i]_k b_k, \sum_k [x_j]_\ell b_\ell \rangle = \sum_k [x_i]_k \sum_k [x_j]_\ell \langle b_k, b_\ell \rangle = [x_i]^\mathsf{T} G [x_j]$
2. For $g \in \mathcal{H}_1$, $f \in \mathcal{H}_2$, ${}_c[g \otimes f]_{\mathcal{B}} = [g]_c [f]_{\mathcal{B}}^\mathsf{T} G$
   The i-th column vector of $[g \otimes f]$ is $[(g \otimes f) b_i]$.

$$[(g \otimes f) b_i] = [g \langle f, b_i \rangle] = [g] \cdot [f]^\mathsf{T} G [b_i],$$

Note that $[b_i] = e_i$ whose i-th element is one and the other entries are zero. Thus $G[b_i]$ is i-th column of G. Stacking them over $i = 1, \ldots, n$,

$$[g \otimes f] = [g][f]^\mathsf{T} G$$

Coordinate Representation Toolkit: Proof 3

Let G be a gram matrix for $\mathcal{H}_1$ w.r.t. $\mathcal{B}$. i.e., $G \in \mathbb{R}^{n \times n}$ such that $(G)_{ij} = \langle b_i, b_j \rangle_{\mathcal{H}_1}$. Suppose that $x_1, \ldots, x_N \in \mathcal{H}_1$. Let $Q_N = I_N - N^{-1} 1_N 1_N^\intercal$ and $[x_{1:N}] \in \mathbb{R}^{n \times N}$ whose columns are $[x_i]$. $\hat{C} = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x}) \otimes (x_i - \bar{x})$

3. $[\hat{C}] = N^{-1} [x_{1:N}] Q_N [x_{1:N}]^\intercal G$

$$
\begin{aligned}
[\hat{C}] &= \left[\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x}) \otimes (x_i - \bar{x}) \right] = \left[\frac{1}{N} \sum_{i=1}^N (x_i \otimes x_i) - \bar{x} \otimes \bar{x} \right] \\
&= \frac{1}{N} \sum_{i=1}^N [x_i \otimes x_i] - [\bar{x} \otimes \bar{x}] = \frac{1}{N} \sum_{i=1}^N [x_i][x_i]^\intercal G - [\bar{x}][\bar{x}]^\intercal G \\
&= \frac{1}{N} [x_{1:n}][x_{1:n}]^\intercal G - N^{-2} [x_{1:N}] 1_N 1_N^\intercal [x_{1:N}]^\intercal G \\
&= N^{-1} [x_{1:N}] Q_N [x_{1:N}]^\intercal G
\end{aligned}
$$

Coordinate Representation Toolkit: Proof 4

3. $[\hat{C}] = N^{-1}[x_{1:N}]Q_N[x_{1:N}]^\mathsf{T}G$
4. $[\hat{C}^\alpha] = G^{-1/2}(N^{-1}G^{1/2}([x_{1:N}]Q_N[x_{1:N}][x_{1:N}]^\mathsf{T}G^{1/2})^\alpha G^{1/2}$
   - Spectral decomposition of $\hat{C} = \sum_i^n \lambda_i v_i \otimes v_i$, where $\hat{C}v_i = \lambda_i v_i$ and $\langle v_i, v_j \rangle = [v_i]^\mathsf{T}G[v_i] = \delta_{ij}$.
   - $[\hat{C}v_i] = N^{-1}[x_{1:N}]Q_N[x_{1:N}]^\mathsf{T}G[v_i] = \lambda_i[v_i]$
   - $\underbrace{N^{-1}G^{1/2}[x_{1:N}]Q_N[x_{1:N}]^\mathsf{T}G^{1/2}}_{A}\underbrace{(G^{1/2}[v_i])}_{w_i} = \lambda_i \underbrace{(G^{1/2}[v_i])}_{w_i}$, $\underbrace{(G^{1/2}[v_i])^\mathsf{T}}_{w_i^\mathsf{T}}\underbrace{(G^{1/2}[v_j])}_{w_j} = \delta_{ij}$
   - $A = \sum_i \lambda_i w_i w_i^\mathsf{T} \implies A^\alpha = \sum_i \lambda_i^\alpha w_i w_i^\mathsf{T}$
   - $[\hat{C}^\alpha] = [\sum_i^n \lambda_i^\alpha v_i \otimes v_i] = \sum_i^n \lambda_i^\alpha [v_i][v_i]^\mathsf{T}G = G^{-1/2}\underbrace{\sum_i \lambda_i^\alpha (G^{1/2}[v_i])(G^{1/2}[v_i])^\mathsf{T}}_{A^\alpha} G^{1/2}$

Functional PCA

Principal component analysis (PCA)

- Let X be a random element in $\mathcal{H}$. Let $v_1, \ldots, v_d$ be the first d principal loadings of X.
- Now, suppose that we want to find a subspace S_d whose dimension is d such that

$$S_d = \underset{S \leq \mathcal{H}, dim(S)=d}{\text{argmin}} \; E\|X - P_S X\|^2$$

- Then the solution is $S_d = \text{span}(\{v_1, \ldots, v_d\})$. (Why?)
- Dimension reduction: find **a lower dimensional space while preserving as much information** as possible.

Principal component analysis (PCA)

Unsupervised dimension reduction

Setting : A random element $X \in \mathcal{H}$, no response variable Y.

Goal: Find the functional direction $\beta_1 \in \mathcal{H}$, such that

$$\beta_1 = \text{argmax}_{\beta \in \mathcal{H}} \text{var}(\langle \beta, X \rangle), \quad \|\beta\| = 1.$$

Then subsequently, find $\beta_2, \ldots, \beta_d$ that maximizes the variance of linear combination of X, in which $\langle \beta_i, X \rangle$'s are uncorrelated. Finally,

$$\text{replace } X \in \mathcal{H} \text{ with } (\langle \beta_1, X \rangle, \ldots, \langle \beta_d, X \rangle)^\intercal \in \mathbb{R}^d$$

Solution: Let $(\lambda_1, v_1), \ldots, (\lambda_m, v_m)$ are the pairs of eigenvalue-eigenfunction of $C = E[(X - EX) \otimes (X - EX)]$ with $\lambda_1 \geq \lambda_2 \geq \ldots$ Then $\beta_k = v_k$.

Eigenfunction problem

$$\hat{C} = E_n[(X - E_n X) \otimes (X - E_n X)] = \frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X}) \otimes (X_i - \bar{X})$$

- The eigenfunction problem is

$$\begin{aligned} \text{maximize} \quad & \langle v_i, \hat{C} v_i \rangle = \text{var}_n(\langle v, X \rangle) \\ \text{subject to} \quad & v_i \in \mathcal{H}, \langle v_i, v_j \rangle = \delta_{ij}, i, j = 1, \ldots, d \end{aligned}$$

- We can use the toolkit.
- Suppose that $\mathcal{H}$ has a basis $\mathcal{B} = \{b_1, \ldots, b_n\}$.

$$\begin{aligned} \text{maximize} \quad & \langle v_i, \hat{C} v_i \rangle = [v_i]^\intercal G[\hat{C}][v_i] = N^{-1} [v_i]^\intercal G[x_{1:N}] Q_N [x_{1:N}]^\intercal G[v_i] \\ \text{subject to} \quad & v_i \in \mathcal{H}, \langle v_i, v_j \rangle = [v_i]^\intercal G[v_j] = \delta_{ij}, i, j = 1, \ldots, d \end{aligned}$$

Eigenfunction problem $\rightarrow$ Eigenvector problem

- The problem is a generalized eigenvalue problem.

$$\begin{aligned} \text{maximize} \quad & \langle v_i, \hat{C} v_i \rangle = [v_i]^\mathsf{T} G[\hat{C}][v_i] = N^{-1} [v_i]^\mathsf{T} G[x_{1:N}] Q_N [x_{1:N}]^\mathsf{T} G[v_i] \\ \text{subject to} \quad & v_i \in \mathcal{H}, \langle v_i, v_j \rangle = [v_i]^\mathsf{T} G[v_j] = \delta_{ij}, i, j = 1, \ldots, d \end{aligned}$$

- Let $w_i \in \mathbb{R}^n$ such that $w_i = G^{1/2}[v_i]$ then $[v_i] = G^{-1/2} w_i$. Then the above problem is

$$\begin{aligned} \text{maximize} \quad & w_i^\mathsf{T} G^{1/2} [x_{1:N}] Q_N [x_{1:N}]^\mathsf{T} G^{1/2} w_i \\ \text{subject to} \quad & w_i \in \mathbb{R}^n, w_i^\mathsf{T} w_i = \delta_{ij}, i, j = 1, \ldots, d \end{aligned}$$

- Eigenvector problem of $G^{1/2}[x_{1:N}] Q_N [x_{1:N}]^\mathsf{T} G^{1/2}$ in a Euclidean space.

Generalized Eigenvalue Problem

- The previous algorithm can be considered as a generalized eigenvalue problem.

$$Av = \lambda Gv$$

- If we let $w = G^{1/2}v$, the above equation becomes

$$AG^{-1/2}w = \lambda G^{1/2}w, \quad \Longrightarrow \quad G^{-1/2}AG^{-1/2}w = \lambda w$$

- Note that the eigenvalues do not change.
- It is also used in the following problem. Let A, Σ be a matrix, S be a subspace. If you have that $col(A) \subseteq \Sigma S$. Then the elements in S can be found by the equation $Av = \lambda \Sigma v$. (Sufficient dimension reduction)

Functional PCA: Computatoin

Given a dataset, $\{x_i(t_{ij} : i = 1, \ldots, N, j = 1, \ldots, J_i\}$, choose a basis $\mathcal{B} = \{b_1, \ldots, b_n\}$.

1. Convert the dataset to functional objects and get the coordinates $[x_{1:N}] \in \mathbb{R}^{n \times N}$.
2. Compute the gram matrix G, where $G_{ij} = \langle b_i, b_j \rangle$.
3. Compute eigenvalue-eigenvector of $G^{1/2}[x_{1:N}]Q_N[x_{1:N}]^\mathsf{T}G^{1/2}$, say $(\lambda_1, w_1), \ldots, (\lambda_d, w_d)$.
4. $[v_i] = G^{-1/2}w_i$.

Let $b(\cdot) = (b_1(\cdot), \ldots, b_n(\cdot))^\mathsf{T}$

- **i-th FPC Loading:** $v_i(\cdot) = [v_i]^\mathsf{T}b(\cdot)$.
- **i-th FPC Scores/Predictors:**: $\langle v_i, x_{1:N} \rangle = [x_{1:N}]^\mathsf{T}G[v_i]$

Kernel PCA

RKHS

- Reproducing kernel Hilbert space is a Hilbert space having a reproducing kernel.
- There are equivalent definitions in a more rigorous way out there. (out of scope)

Book: Alain Berlinent and Christine Thomas-Agnan (2004). Reproducing Kernel Hilbert Space in Probability and Statistics.

RKHS: Reproducing kernel

Definition (Reproducing Kernel)

A bivariate function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a **reproducing kernel** of a Hilbert space $\mathcal{H}$ if for all $f \in \mathcal{H}$,

$$f(x) = \langle \kappa(x, \cdot), f \rangle_{\mathcal{H}}$$

- For $x \in \mathcal{X}$, $\kappa(x, \cdot) \in \mathcal{H}$.
- For $x, y \in \mathcal{X}$, $\langle \kappa(x, \cdot), \kappa(y, \cdot) \rangle = \kappa(x, y)$

RKHS: Reproducing kernel

Definition (Kernel)

A bivariate function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a **kernel** if

1. $\kappa(\cdot, \cdot)$ is symmetric. i..e, $\kappa(x, y) = \kappa(y, x)$.
2. κ is positive semi-definite. i.e., for any $n \in \mathbb{N}$ and choice of $x_1, \ldots, x_n \in \mathcal{X}$ and $c_1, \ldots, c_n \in \mathbb{R}$,

$$\sum_{i,j} c_i c_j \kappa(x_i, x_j) \geq 0$$

- You can consider $f = \sum_i c_i \kappa(x_i, \cdot) \in \mathcal{H}$. Then

$$\langle f, f \rangle_{\mathcal{H}} = \langle \sum_i c_i \kappa(x_i, \cdot), \sum_j c_j \kappa(x_j, \cdot) \rangle_{\mathcal{H}} = \sum_{i,j} c_i c_j \kappa(x_i, x_j)$$

- Or, the **gram matrix**, $K \in \mathbb{R}^{n \times n}$ such that $K_{ij} = \langle \kappa(x_i, \cdot), \kappa(x_j, \cdot) \rangle_{\mathcal{H}} = \kappa(x_i, x_j)$, is positive semi-definite for every choice of $n \in \mathbb{N}$, $x_i \in \mathcal{X}$ for $i = 1, \ldots, n$

RKHS: Reproducing kernel

- Given a reproducing kernel $\kappa(\cdot, \cdot)$, can we build an RKHS? Is it unique?
- **Moore-Aronszajn theorem**

RKHS: Moore-Aronszajn theorem

Theorem

Suppose κ is a symmetric, positive definite kernel on a set $\mathcal{X}$. Then there is a unique Hilbert space of functions on $\mathcal{X}$ for which κ is a reproducing kernel.

Sketch of proof

1. Build an RKHS given a p.d. kernel κ.
2. Show that it's unique. i.e., if there is another RKHS for which κ is a reproducing kernel, then the space is the same as the space constructed in step 1.

Build an RKHS over $\mathcal{X}$ using a p.d. kernel: 1

Define

$$\mathcal{H}_0 = \text{span}\{\kappa(x, \cdot) : x \in \mathcal{X}\}.$$

Notice that $\mathcal{H}_0$ is a vector space. If $f \in \mathcal{H}_0$, then $f : \mathcal{X} \to \mathbb{R}$. In addition, there exists $[f] \in \mathbb{R}^m$ for some $m \in \mathbb{N}$, such that

$$f = \sum_{i=1}^{m} [f]_i \kappa(x_i, \cdot), \quad f(x) = \sum_{i=1}^{m} [f]_i \kappa(x_i, x), \quad \text{for } x \in \mathcal{X}$$

Hilbert space.. need an inner product.

Build an RKHS over $\mathcal{X}$ using a p.d. kernel: 2

Define

$$\mathcal{H}_0 = \text{span}\{\kappa(x, \cdot) : x \in \mathcal{X}\}.$$

Define the inner product in $\mathcal{H}_0$ by $\langle \kappa(x, \cdot), \kappa(y, \cdot) \rangle_{\mathcal{H}_0} = \kappa(x, y)$ for $x, y \in \mathcal{X}$.

- For $f, g, \in \mathcal{H}_0$, we can represent them by $f = \sum_{i=1}^{m} [f]_i \kappa(x_i, \cdot)$, $g = \sum_{j=1}^{n} [g]_j \kappa(x_j, \cdot)$.
- The inner product between f and g is then

$$\begin{aligned}
\langle f, g \rangle_{\mathcal{H}_0} &= \langle \sum_{i=1}^{m} [f]_i \kappa(x_i, \cdot), \sum_{j=1}^{n} [g]_j \kappa(x_j, \cdot) \rangle_{\mathcal{H}_0} \\
&= \sum_{i=1}^{m} \sum_{j=1}^{n} [f]_i [g]_j \langle \kappa(x_i, \cdot), \kappa(x_j, \cdot) \rangle_{\mathcal{H}_0} \\
&= \sum_{i=1}^{m} \sum_{j=1}^{n} [f]_i [g]_j \kappa(x_i, x_j)
\end{aligned}$$

Build an RKHS over $\mathcal{X}$ using a p.d. kernel: 3

Define

$$\mathcal{H}_0 = \mathrm{span}\{\kappa(x, \cdot) : x \in \mathcal{X}\}.$$

Define the **inner product** in $\mathcal{H}_0$ by $\langle \kappa(x, \cdot), \kappa(y, \cdot) \rangle_{\mathcal{H}_0} = \kappa(x, y)$ for $x, y \in \mathcal{X}$.
Next, is this space **complete**?

- Define $\mathcal{H}$ be the completion of $\mathcal{H}_0$, a smallest complete space including $\mathcal{H}_0$.

Reproducing property?

- $f \in \mathcal{H}$ can be represented as $f(x^*) = \sum_{i=1}^{\infty} [f]_i \kappa(x_i, x^*)$.
- For $x \in \mathcal{X}$, $\kappa(\cdot, x)$ is a form of evaluation functional.

$$\langle \kappa(x, \cdot), f \rangle_{\mathcal{H}} = \langle \kappa(x, \cdot), \sum_{i=1}^{\infty} [f]_i \kappa(x_i, \cdot) \rangle_{\mathcal{H}} = \sum_{i=1}^{\infty} [f]_i \kappa(x_i, x) = f(x)$$

Example of kernel

- Gaussian radial basis function (Gaussian RBF)

$$\kappa(x, y) = \exp(-\gamma \|x - y\|_{\mathcal{X}}^2), \quad \gamma > 0$$

- Polynomial kernel

$$\kappa(x, y) = (\langle x, y \rangle_{\mathcal{X}} + c)^d$$

$$\langle \kappa(x, \cdot), \kappa(y, \cdot) \rangle_{\mathcal{H}} = \kappa(x, y)$$

Application: Kernel Trick

Feature Mapping

- Feature map: $\phi : \mathcal{X} \to \mathcal{H}$
- For each data point $x \in \mathcal{X}$, we map it to $\phi(x) \in \mathcal{H}$. e.g., given a data point $x_i \in \mathcal{X}$, $i = 1, \ldots, n$

$$\phi(x_i) = \kappa(x_i, \cdot) \in \mathcal{H}, \quad i = 1, \ldots, n$$

- If X is a random element in $\mathcal{X}$, then $\phi(X)$ is a random element in $\mathcal{H}$.
- $X_1, \ldots, X_n$ are n random copies of X. Then $\phi(X_1), \ldots, \phi(X_n)$ are n random copies of $\phi(X)$.

Kernel PCA: Introduction

- **PCA**: the fist PC is

$$\beta_1 = \text{argmax}_{\beta \in \mathbb{R}^p} \text{var}(\beta^\intercal X), \quad \|\beta\| = 1.$$

- i.e., considering it as a linear map, $L_1(x) = \beta_1^\intercal x$.

$$L_1 = \underset{L \text{ is linear}: \mathbb{R}^p \to \mathbb{R}}{\text{argmax}} \text{var}(L(X))$$

- **Kernel PCA** is seeking

$$f = \underset{f: \mathbb{R}^p \to \mathbb{R}}{\text{argmax}} \text{var}(f(X)), \quad f \in \mathcal{H}.$$

Be careful: Many of the kernel PCA documents out there are poorly written...

Kernel PCA: Sample level-ready

- Let $x_1, \ldots, x_n \in \mathbb{R}^p$.
- Define $\mathcal{H}_n = \text{span}\{\kappa(\cdot, x_i) : i = 1, \ldots, n\}$.
- Define $K \in \mathbb{R}^{n \times n}$ s.t.

$$K_{ij} = \langle \kappa(x_i, \cdot), \kappa(x_j, \cdot) \rangle_{\mathcal{H}_n} = \kappa(x_i, x_j)$$

  which is called Gram matrix.
- For any $f \in \mathcal{H}_n$, $f(x^*) = \sum_{j=1}^n [f]_j \kappa(x_j, x^*)$ for $[f] \in \mathbb{R}^n$.
- $\langle f, g \rangle = \langle \sum_{j=1}^n [f]_j \kappa(x_j, \cdot), \sum_{k=1}^n [g]_k \kappa(x_k, \cdot) \rangle = [f]^\intercal K [g]$
- Let's compute

$$\text{var}_n(f(X)) = E_n[(f(X) - E_n(f(X)))^2]$$

Kernel PCA: Sample level-mean

- For any $f \in \mathcal{H}_n$, $f = \sum_{j=1}^n [f]_j \kappa(x_j, \cdot)$ for $[f] \in \mathbb{R}^n$.

$$
\begin{aligned}
(f(x_1), \ldots, f(x_n))^\mathsf{T} &= (\textstyle\sum_{j=1}^n [f]_j \kappa(x_1, x_j), \ldots, \sum_{j=1}^n [f]_j \kappa(x_n, x_j))^\mathsf{T} \\
&= K[f]
\end{aligned}
$$

$$
\begin{aligned}
E_n(f(X)) &= n^{-1} \sum_{i=1}^n f(x_i) = n^{-1} \sum_{i=1}^n \sum_{j=1}^n [f]_j \kappa(x_i, x_j) \\
&= n^{-1} \sum_{i=1}^n \sum_{j=1}^n [f]_j \kappa(x_j, x_i) \\
&= n^{-1} \sum_{i=1}^n \sum_{j=1}^n \kappa(x_i, x_j) [f]_j \\
&= n^{-1} 1_n^\mathsf{T} K[f]
\end{aligned}
$$

Kernel PCA: Sample level-variance

- Define $Q_n = I_n - n^{-1} 1_n 1_n^\intercal$

$$f(x_i) - E_n(f(X)) = (K[f])_i - n^{-1} 1_n^\intercal K[f] = i\text{-th row of } QK[f]$$

$$\begin{aligned}
\mathrm{var}_n(f(X)) &= E_n[(f(X) - E_n(f(X)))^2] \\
&= n^{-1} \sum_{i=1}^{n} (f(x_i) - E_n(f(X)))^2 \\
&= n^{-1} (QK[f])^\intercal (QK[f]) \\
&= n^{-1} [f]^\intercal KQK[f]
\end{aligned}$$

Note: if we construct $\mathcal{H}_n = \mathrm{span}\{\kappa(\cdot, x_1) - E_n(\kappa(\cdot, X)), \ldots, \kappa(\cdot, x_1) - E_n(\kappa(\cdot, X))\}$, then

$$\mathrm{var}_n(f(X)) = [f]^\intercal G^2 [f],$$

where $G = QKQ$

Kernel PCA: Sample level-computation

Note: $\|f\|^2_{\mathcal{H}_n} = \langle f, f \rangle_{\mathcal{H}_n} = [f]^\mathsf{T} K [f]$

- The optimization problem:

$$\begin{aligned} \text{minimize} \quad & \operatorname{var}_n(f(X)) \\ \text{subject to} \quad & f \in \mathcal{H}_n, \quad \|f\|_{\mathcal{H}_n} = 1 \end{aligned}$$

- is equivalent to

$$\begin{aligned} \text{minimize } & [f]^\mathsf{T} K Q K [f] \\ \text{subject to } & [f] \in \mathbb{R}^n, \quad [f]^\mathsf{T} K [f] = 1 \end{aligned}$$

- Generalized eigenvalue problem. Let $v = K^{1/2}[f]$, $[f] = K^{-1/2} v$

$$\begin{aligned} \text{minimize } & v^\mathsf{T} K^{1/2} Q K^{1/2} v \\ \text{subject to } & v \in \mathbb{R}^n, \quad \|v\|_2 = 1 \end{aligned}$$

Kernel PCA: Sample level-computation

Note: $\|f\|^2_{\mathcal{H}_n} = \langle f, f \rangle_{\mathcal{H}_n} = [f]^\intercal K [f]$

- Let $v = K^{1/2}[f]$, $[f] = K^{-1/2} v$

$$\begin{aligned} &\text{minimize } v^\intercal K^{1/2} Q K^{1/2} v \\ &\text{subject to } v \in \mathbb{R}^n, \quad \|v\|_2 = 1 \end{aligned}$$

- It's an eigenvalue problem of $K^{1/2} Q K^{1/2}$. Let $v_1, \ldots, v_d$ are the first d-eigenvector of the matrix.
- $[f_k] = K^{-1/2} v_i$, $k = 1, \ldots, d$.
- $f_k(x) = \sum_{j=1}^n [f_k]_j \kappa(x_j, x)$ a linear combination of nonlinear functions of x.
- **Nonlinear PCA**

Kernel PCA: Sample level-computation

Note

$$\begin{aligned}(f(x_1),\ldots,f(x_n))^\mathsf{T} &= (\textstyle\sum_{j=1}^n [f]_j \kappa(x_1,x_j),\ldots,\sum_{j=1}^n [f]_j \kappa(x_n,x_j))^\mathsf{T} \\ &= K[f]\end{aligned}$$

- $[f_k] = K^{-1/2} v_i$, $k = 1,\ldots,d$.
- $f_k(x) = \sum_{j=1}^n [f_k]_j \kappa(x_j, x)$ a linear combination of nonlinear functions of x.
- The k-th nonlinear principal components are

$$\begin{aligned}(f_k(x_1),\ldots,f_k(x_n))^\mathsf{T} &= (\textstyle\sum_{j=1}^n [f_k]_j \kappa(x_1,x_j),\ldots,\sum_{j=1}^n [f_k]_j \kappa(x_n,x_j))^\mathsf{T} \\ &= K[f_k]\end{aligned}$$

Understand Kernel PCA (Advanced)

- The previous method directly extracts $f(\cdot)$ by maximizing $\mathrm{var}(f(X))$.
- We can understand KPCA in a different way.
- Feature mapping: $X \to \phi(X) \in \mathcal{H}$
- In linear PCA, we find the variance matrix Σ then take eigen decomposition to extract the directions.
- Can we do that for $\phi(X)$ in $\mathcal{H}$?

Understand Kernel PCA (Advanced)

Define the covariance operator $\Gamma : \mathcal{H} \to \mathcal{H}$

$$\Gamma = E[\{\phi(X) - E(\phi(X))\} \otimes \{\phi(X) - E(\phi(X))\}]$$

Note: $\phi(X)(\cdot) = \kappa(X, \cdot)$

$$\Gamma_n = n^{-1} \sum_{i=1}^{n} [\{\kappa(X_i, \cdot) - E_n(\kappa(X, \cdot))\} \otimes \{\kappa(X_i, \cdot) - E_n(\kappa(X, \cdot))\}]$$

For $f \in \mathcal{H}_n$, the reproducing property tells us that

$$f(X_i) = \langle f, \kappa(X_i, \cdot) \rangle_{\mathcal{H}_n}, \text{ thus, } \quad E_n(f(X)) = \langle f, E_n(\kappa(X, \cdot)) \rangle_{\mathcal{H}_n}$$

Using that, we have

$$\mathrm{var}_n(f(X)) = \langle f, \Gamma_n f \rangle_{\mathcal{H}_n}$$

Understand Kernel PCA (Advanced)

Define the covariance operator $\Gamma : \mathcal{H} \to \mathcal{H}$

$$\Gamma = E[\{\phi(X) - E(\phi(X))\} \otimes \{\phi(X) - E(\phi(X))\}]$$

Note: $\phi(X)(\cdot) = \kappa(X, \cdot)$

$$\Gamma_n = n^{-1} \sum_{i=1}^{n} [\{\kappa(X_i, \cdot) - E_n(\kappa(X, \cdot))\} \otimes \{\kappa(X_i, \cdot) - E_n(\kappa(X, \cdot))\}]$$

Using that, it can be easily shown that (why?)

$$\mathrm{var}_n(f(X)) = \langle f, \Gamma_n f \rangle_{\mathcal{H}_n}$$

Thus, the maximizing function f is the eigenfunction (with the largest eigenvalue) of Γ_n.

Kernel Trick: Summary

- Feature mapping $X_i \to \phi(X_i) \in \mathcal{H}$
- Conduct linear methods over $\mathcal{H}$
- Represented estimator becomes a linear combination of $\kappa(\cdot, x)$: nonlinear estimator in terms of x.

Role of Kernel & Tuning Parameters

Properties of Gaussian RBF

- When we consider $\mathrm{var}(f(X))$ over $f \in \mathcal{H}$, we would like $\mathcal{H}$ to be rich enough to cover all possible functions $f : \mathcal{X} \to \mathbb{R}$.
- If we take Gaussian RBF as the reproducing kernel $\kappa(\cdot, \cdot)$ for the RKHS $\mathcal{H}$, i.e.,

$$\mathcal{H} = \overline{\mathsf{span}\{\kappa(x, \cdot) : x \in \mathcal{X}\}},$$

with $\kappa(x, y) = \exp(-\gamma \|x - y\|^2)$, then $\mathcal{H}$ is a dense subset of $L_2(\mathcal{X})$.

$$\kappa(x, \cdot) \in \mathcal{H}, \text{ if } g = \kappa(x, \cdot)$$
$$g(y) = \kappa(x, y) \in \mathbb{R}$$
$$K_x = \kappa(x, \cdot), \quad K_y = \kappa(y, \cdot)$$

Tuning parameter of Gaussian RBF

- Gaussian radial basis function (Gaussian RBF)

$$\kappa(x, y) = \exp(-\gamma \|x - y\|_{\mathcal{X}}^2), \quad \gamma > 0$$

γ affects the results greatly. What is the role of γ?

Role of γ in the Gaussian RBF

- Gaussian radial basis function (Gaussian RBF)

$$\kappa(x, y) = \exp(-\gamma \|x - y\|_{\mathcal{X}}^2), \quad \gamma > 0$$

γ affects the results greatly. What is the role of γ?

- Remark: we are working with $\phi(X)(\cdot) = \kappa(\cdot, X)$.
- Instead of using the geometry in $\|x - y\|_{\mathcal{X}}$ in $x, y \in \mathcal{X}$
- we use the geometry in $\mathcal{H}$, $\|\kappa(\cdot, x) - \kappa(\cdot, y)\|_{\mathcal{H}}$.

Role of γ in the Gaussian RBF

For $X_a, X_b \in \mathcal{H}$, the squared distance between two elements in $\mathcal{H}$ is

$$\begin{aligned}\|\phi(X_a) - \phi(X_b)\|_{\mathcal{H}} &= \|\kappa(\cdot, X_a) - \kappa(\cdot, X_b)\|_{\mathcal{H}}^2 \\ &= \kappa(X_a, X_a) + \kappa(X_b, X_b) - 2\kappa(X_a, X_b) \\ &= 2\{1 - \kappa(X_a, X_b)\} \\ &= 2\{1 - \exp(-\gamma\|X_a - X_b\|_{\mathcal{X}}^2)\}.\end{aligned}$$

By Taylor approximation of $\exp(\cdot)$,

$$\begin{aligned}\|\kappa(\cdot, X_a) - \kappa(\cdot, X_b)\|_{\mathcal{H}}^2 &= 2\{1 - 1 + \gamma\|X_a - X_b\|_{\mathcal{X}}^2 + O(\gamma^2)\} \\ &= 2\gamma\|X_a - X_b\|_{\mathcal{X}}^2 + O(\gamma^2).\end{aligned}$$

Consequently, when γ is small, the KPCA can be approximated by the linear PCA, which is based on the distance $\|X_a - X_b\|_{\mathcal{X}}$.

Role of γ in the Gaussian RBF

- Gaussian radial basis function (Gaussian RBF)

$$\kappa(x, y) = \exp(-\gamma \|x - y\|_{\mathcal{X}}^2), \quad \gamma > 0$$

γ affects the results greatly. What is the role of γ?

- γ controls the nonlinearity of kernel trick.
- large γ: more nonlinear
- small γ: close to the linear method

Choice of γ

Initial choice

In practice, we find the γ via a grid search over $\rho \in [10^{-8}, 10^{2}]$ defined by

$$\gamma = \rho/(2\sigma^2), \quad \sigma^2 = \binom{n}{2}^{-1} \sum_{1 \le a < b \le n} \|X_a - X_b\|^2.$$

Next, which criteria? It's unsupervised.

Choice of γ

Initial choice

In practice, we find the γ via a grid search over $\rho \in [10^{-8}, 10^2]$ defined by

$$\gamma = \rho/(2\sigma^2), \quad \sigma^2 = \binom{n}{2}^{-1} \sum_{1 \le a < b \le n} \|X_a - X_b\|^2.$$

- We are seeking ρ that maximizes

$$\mathrm{var}_n(f(X)) = \langle f, \Gamma f \rangle = v^\top K^{1/2} Q K^{1/2} v$$

- The maximum value is the eigenvalue of $K^{1/2} Q K^{1/2}$.
- But it is not fair if we compare it directly since $\mathcal{H}_n$ has different inner product structure depending on γ.

Summary

- Choice of kernel function: a rich class of nonlinear functions - Gaussian RBF
- Tuning parameter selection also relates to how many PCs we consider
- See Song and Li (2021). Nonlinear and additive principal component analysis for functional data, *JMVA*

Nonlinear FPCA

RKHS over a function space

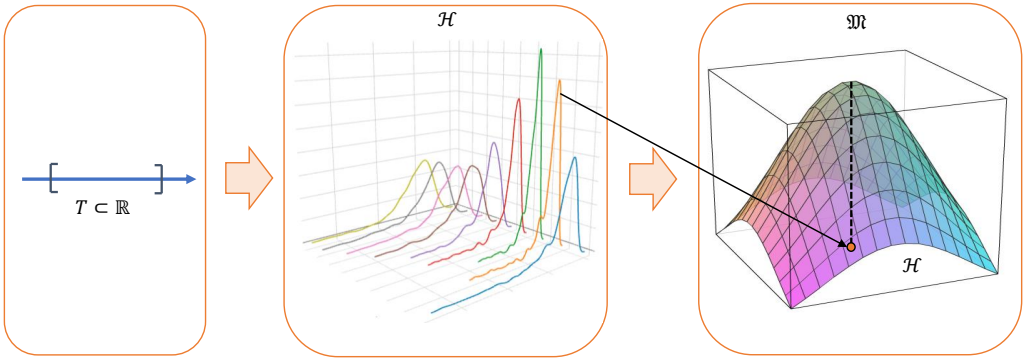
- Recall many of the kernel functions in RKHS use the inner product of the space.
- We also did not specify the dimension of the original space.
- We can build RKHS over any Hilbert space.

Song and Li (2021). Nonlinear and additive principal component analysis for functional data, *JMVA*

Nested reproducing kernel Hilbert space

Two levels of functions

$$f(X_i) = \sum_{\phi \in \mathcal{H}} \kappa_X(\phi, X_i)$$



1st level function space
$\mathcal{H}$: a separable Hilbert space

2nd level function space
$\mathfrak{M}$: a RKHS over $\mathcal{H}$

Mean and Covariance

Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space. Let X be a **random element** in $\mathcal{H}$, i.e.,

$$X : \Omega \to \mathcal{H} = \{f : T \subseteq \mathbb{R} \to \mathbb{R}p\},$$

where $\mathcal{H}$ is a separable Hilbert space of vector-valued functions defined on an interval $T \subseteq \mathbb{R}$.

Then $\kappa(\cdot, X)$ is a **random element** in $\mathfrak{M}$. Consider the linear functional defined on $\mathfrak{M}$, for any $f \in \mathfrak{M}$,

$$f \in \mathfrak{M} \mapsto E[f(X)].$$

Let $\mu_X \in \mathfrak{M}$ be the Riesz representation, which is the unique element that satisfies

$$\langle \mu_X, f \rangle_{\mathfrak{M}} = E[f(X)] = E[\langle f, \kappa(\cdot, X) \rangle_{\mathfrak{M}}], \quad \text{for all } f \in \mathfrak{M}.$$

We call μ_X be the mean of $\kappa(\cdot, X)$.

Mean and Covariance

Let $\mathfrak{M}^0$ be the completion of the space spanned by

$$\{\kappa(\cdot, x) - \mu_X : x \in \mathcal{H}\}.$$

For two members $f, g \in \mathfrak{M}^0$, the bilinear form

$$(f, g) \in \mathfrak{M} \times \mathfrak{M} \mapsto \mathrm{cov}[f(X), g(X)],$$

gives us a unique linear operator, say Σ, in $\mathcal{B}(\mathfrak{M}, \mathfrak{M})$ such that

$$\langle f, \Sigma g \rangle_{\mathfrak{M}} = \mathrm{cov}[f(X), g(X)].$$

We call Σ the covariance operator.

Remark 1. Note that it is **not** the covariance of X. It is the covariance operator of a collection of functions of X.

Remark 2. Under a bounded assumption, it can be shown that $\overline{\mathrm{ran}}(\Sigma) = \mathfrak{M}^0$. Since $\ker(\Sigma)$ is not our interest, we will consider only for the members in $\ker(\Sigma)^\perp = \mathfrak{M}^0$.

Nonlinear PCA for functional data

Definition. The population-level additive functional principal components are defined through the following iterative maximization: at step k, $\phi^{(k)}$ is obtained by

$$\begin{aligned} &\text{maximizing} \quad \operatorname{var}[\phi(X)] \\ &\text{subject to} \quad \phi \in \mathfrak{M},\ \langle \phi, \phi^{(1)} \rangle_{\mathfrak{M}} = \cdots = \langle \phi, \phi^{(k-1)} \rangle_{\mathfrak{M}} = 0,\ \|\phi\|_{\mathfrak{M}} = 1. \end{aligned}$$

The random variable $\phi^{(k)}(X)$ is called the kth additive functional principal component of X.

Nonlinear PCA for functional data

With our results and modification, the problem is equivalent to find the eigen functions of an self-adjoint operator.

Definition. The population-level functional additive principal components are defined through the following iterative maximization: at step k, ϕ_k is obtained by

$$\begin{aligned} &\text{maximizing} \quad \langle \phi, \Sigma\phi \rangle_{\mathfrak{M}} \\ &\text{subject to} \quad \phi \in \mathfrak{M}^0,\ \langle \phi, \phi^{(1)} \rangle_{\mathfrak{M}} = \cdots \langle \phi, \phi^{(k)} \rangle_{\mathfrak{M}} = 0,\ \|\phi\|_{\mathfrak{M}} = 1. \end{aligned}$$

The random variable $\phi^{(k)}(X)$ is called the k-th additive functional principal component of X.

Operators in Second-level space

Goal: Find the matrix representation of Σ.

Idea: Use MME.

$$\begin{aligned}\langle f, \hat{\Sigma} g\rangle &= [f] K_X [\hat{\Sigma}] [g] \\ \langle f, \hat{\Sigma} g\rangle &= \mathrm{cov}_n(f(X), g(X) \\ &= n^{-1} \textstyle\sum_{i=1}^{n} [f(X_i) - E_n f(X)][g(X_i) - E_n g(X)]\end{aligned}$$

Results: Solving the above two equations, we get $[\hat{\Sigma}] = n^{-1} Q K_X Q$, where $Q = I - 11^{\mathsf{T}}$. Let $G_X = Q K_X Q$.

Sample level Estimation - nonlinear FAPCA

The optimization problem of nonlinear PCA at the sample level is

$$\begin{aligned} &\text{maximize} \quad \langle \phi, \hat{\Sigma}\phi \rangle_{\mathfrak{M}_X} = [\phi]^\mathsf{T} G_X^2 [\phi] \\ &\text{subject to} \quad [\phi]^\mathsf{T} G_X [\phi] = 1, [\phi]^\mathsf{T} G_X [\phi_1] = \ldots = [\phi]^\mathsf{T} G_X [\phi_{k-1}] = 0. \end{aligned}$$

Let $v = G_X^{1/2}[\phi]$. Then this generalized eigenvalue problem can be rephrased as a normal eigenvalue problem:

$$\begin{aligned} &\text{maximize} \quad v^\mathsf{T} G_X^{-1/2} G_X^2 G_X^{-1/2} v = v^\mathsf{T} G_X v \\ &\text{subject to} \quad v^\mathsf{T} v = 1, v^\mathsf{T} v_1 = \ldots = v^\mathsf{T} v_{k-1} = 0. \end{aligned}$$

Thus, k-th PC, $\phi^{(k)} \in \mathfrak{M}$ is

$$\phi^{(k)}(x) = v_k^\mathsf{T} G_X^{\dagger 1/2} c_X(x),$$

where v_k is the k-th eigenvector of G_X and
$c_X(x) = (\kappa_X(x, X_1) - \mu_X(x), \ldots, \kappa_X(x, X_n) - \mu_X(x))^\mathsf{T}$.

Simulation Study

Model I

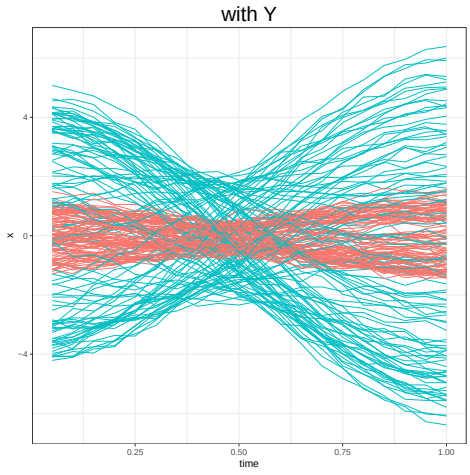
- with $n = 100$, generate $Y_i \sim Ber(0.5)$, $i = 1, \ldots, n$.
- Conditioning on Y, generate $X(t)$ on 20 equally spaced time points over $[0, 1]$

$$X_i(t)|Y_i = 0 \sim Z_{1i}\cos(\theta_{1i})\cos(\pi t) + Z_{1i}\sin(\theta_{1i})\sin(t) + \epsilon_i(t)$$
$$X_i(t)|Y_i = 1 \sim Z_{2i}\cos(\theta_{2i})\cos(\pi t) + Z_{2i}\sin(\theta_{2i})\sin(t) + \epsilon_i(t)$$

where $Z_{1i} \sim N(1, 0.2^2)$, $\theta_{1i} \sim U(0, 2\pi)$, $Z_{2i} \sim N(4, 0.5^2)$, $\theta_{2i} \sim U(0, 2\pi)$, and $\epsilon_i(t_j) \sim N(0, 0.1^2)$.

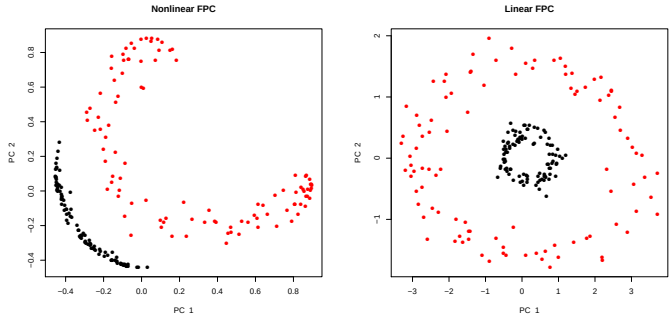
Simulation Study

Model I



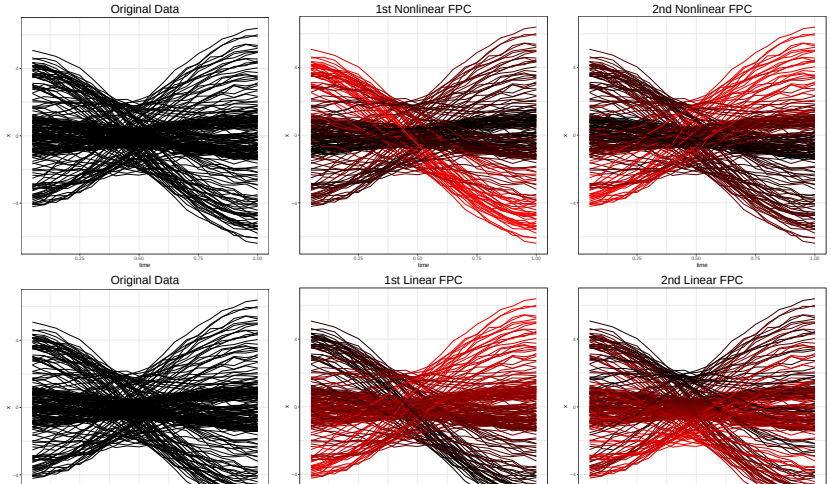
Simulation Study - Visualization

- If we chose the first two nonlinear PCs, each curve $X_i(t)$ is reduced to a point in a 2D plot with coordinates $(\phi_1(X_i), \phi_2(X_i))$.



Simulation Study - Visualization

- We want to see how the PCs are related with the original curves.
- i-th PC score $\phi_i(X)$ is a real value. Scale it to Red color.



Simulation Study - Visualization

- We want to see how the multiple PCs are related with the original curves.
- Scale them to RGB Color. R=$\phi_1(X)$, G=$\phi_2(X)$, B=$\phi_3(X)$.