Unsupervised Learning Methods
Problem Set II –
PCA and KPCA

Due: 24.05.2021

Guidelines

- Answer all questions (PDF + Jupyter notebook).
- You must type your solution manual (handwriting is not allowed).
- Submission in pairs (use the forum if needed).
- You **may** submit the entire solution in a single ipynb file (or in PDF + ipynb files).
- You **may** (and should) use the forum if you have any questions.
- Good luck!

1 PCA

1.1 Eigendecomposition

Trace

- Let $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ be a diagonalizable matrix, that is, $\boldsymbol{A} = \boldsymbol{V} \boldsymbol{\Lambda} \boldsymbol{V}^{-1}$ where $\boldsymbol{\Lambda}$ is a diagonal matrix.

1.1.1

Prove that:

$$\mathrm{Tr}\left\{\boldsymbol{A}\right\} = \sum_{i=1}^{d} \lambda_i\left(\boldsymbol{A}\right)$$

where $\lambda_i\left(\boldsymbol{A}\right) = \boldsymbol{\Lambda}\left[i,i\right]$ is the ith eigenvalue of $\boldsymbol{A}$.

– –

Similarity

- Two (square) matrices $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ and $\boldsymbol{B} \in \mathbb{R}^{d \times d}$ are called similar, namely, $\boldsymbol{A} \sim \boldsymbol{B}$, if exists an (invertible) matrix $\boldsymbol{P} \in \mathbb{R}^{d \times d}$ such that:

$$\boldsymbol{B} = \boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^{-1}$$

1.1.2

Prove that if $\boldsymbol{A}$ is diagonalizable and $\boldsymbol{A} \sim \boldsymbol{B}$, then, $\boldsymbol{A}$ and $\boldsymbol{B}$ share the same set of eigenvalues, namely:

$$\boldsymbol{A} \sim \boldsymbol{B} \implies \{\lambda_i(\boldsymbol{A})\}_{i=1}^{d} = \{\lambda_i(\boldsymbol{B})\}_{i=1}^{d}$$

SPD matrices

A symmetric matrix $\boldsymbol{A} = \boldsymbol{A}^T$ is an Symmetric Positive Definite (SPD), namely $\boldsymbol{A} \succ 0$ if either:

1. $\lambda_i(\boldsymbol{A}) > 0$ for all i.
2. $\boldsymbol{v}^T\boldsymbol{A}\boldsymbol{v} > 0$ for all $\boldsymbol{v} \neq \boldsymbol{0}$.

1.1.3

Prove that the two conditions are equivalent, that is:

$$\lambda_i(\boldsymbol{A}) > 0 \iff \boldsymbol{v}^T\boldsymbol{A}\boldsymbol{v} > 0 \quad \forall \boldsymbol{v} \neq \boldsymbol{0}$$

1.2 PCA

Full PCA

- Consider the data $\mathcal{X} = \left\{\boldsymbol{x}_i \in \mathbb{R}^D\right\}_{i=1}^{N}$ with mean $\boldsymbol{\mu}_x \in \mathbb{R}^D$ and covariance $\boldsymbol{\Sigma}_x \in \mathbb{R}^{D \times D}$.
- Let $\boldsymbol{\Sigma}_x = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$ be the eigendecomposition of $\boldsymbol{\Sigma}_x$.
- Let $\boldsymbol{z}_i = \boldsymbol{U}^T(\boldsymbol{x}_i - \boldsymbol{\mu}_x)$

1.2.1

Prove that:

1. The mean of $\mathcal{Z} = \{\boldsymbol{z}_i\}_{i=1}^{N}$ is zero, that is, $\boldsymbol{\mu}_z = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{z}_i = \boldsymbol{0}$.
2. The covariance of $\mathcal{Z}$ is diagonal, that is $\boldsymbol{\Sigma}_z$ is diagonal.
3. $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 = \|\boldsymbol{z}_i - \boldsymbol{z}_j\|_2$ for all i and j .

Geometric PCA

- Let $\boldsymbol{U}_d \in \mathbb{R}^{D \times d}$ be a full rank matrix (with $d \leq D$).

1.2.2

Show that exists an invertible matrix $\boldsymbol{M} \in \mathbb{R}^{d \times d}$ such that $\boldsymbol{O} = \boldsymbol{U}_d \boldsymbol{M} \in \mathbb{R}^{D \times d}$ is semi-orthogonal, that is:

$$\boldsymbol{O}^T \boldsymbol{O} = \boldsymbol{I}_d$$

- Consider the data $\boldsymbol{X} \in \mathbb{R}^{D \times N}$ with zero mean $\boldsymbol{X}\boldsymbol{1}_N = \boldsymbol{0} \in \mathbb{R}^D$ and covariance $\boldsymbol{\Sigma}_x = \frac{1}{N}\boldsymbol{X}\boldsymbol{X}^T \in \mathbb{R}^{D \times D}$.
- Consider the following optimization problems:

1. Reconstruction error minimization:
$$\begin{cases} \arg\min_{\boldsymbol{U}_d \in \mathbb{R}^{D \times d}} \left\| \boldsymbol{X} - \boldsymbol{U}_d \boldsymbol{U}_d^T \boldsymbol{X} \right\|_F^2 \\ \text{s.t. } \boldsymbol{U}_d^T \boldsymbol{U}_d = \boldsymbol{I}_d \end{cases}$$

2. Variance maximization:
$$\begin{cases} \arg\max_{\boldsymbol{U}_d \in \mathbb{R}^{D \times d}} \operatorname{Tr}\left\{ \boldsymbol{U}_d^T \boldsymbol{\Sigma}_x \boldsymbol{U}_d \right\} \\ \text{s.t. } \boldsymbol{U}_d^T \boldsymbol{U}_d = \boldsymbol{I}_d \end{cases}$$

1.2.3

Prove that both problems have the same optimal solution $\boldsymbol{U}_d^{\star}$.

PCA analysis

- Consider the data $\left\{ \boldsymbol{x}_i \in \mathbb{R}^D \right\}_{i=1}^N$ with mean $\boldsymbol{\mu}_x \in \mathbb{R}^D$ and covariance $\boldsymbol{\Sigma}_x \in \mathbb{R}^{D \times D}$.
- Let $\boldsymbol{U}_d \in \mathbb{R}^{D \times d}$ be a semi-orthogonal matrix, that is, $\boldsymbol{U}_d^T \boldsymbol{U}_d = \boldsymbol{I}_d$
- Let $\boldsymbol{z}_i = \boldsymbol{U}_d^T \left(\boldsymbol{x}_i - \boldsymbol{\mu}_x \right) \in \mathbb{R}^d$.
- Let $\hat{\boldsymbol{x}}_i = \boldsymbol{U}_d \boldsymbol{z}_i + \boldsymbol{\mu}_x \in \mathbb{R}^D$
- Let $\boldsymbol{\epsilon}_i = \boldsymbol{x}_i - \hat{\boldsymbol{x}}_i \in \mathbb{R}^D$

1.2.4

Prove that:

$$\operatorname{Tr}\left\{ \boldsymbol{\Sigma}_x \right\} = \operatorname{Tr}\left\{ \boldsymbol{\Sigma}_z \right\} + \operatorname{Tr}\left\{ \boldsymbol{\Sigma}_\epsilon \right\}$$

where:

- $\boldsymbol{\Sigma}_z \in \mathbb{R}^{d \times d}$ is the covariance of $\left\{ \boldsymbol{z}_i \right\}_{i=1}^N$.
- $\boldsymbol{\Sigma}_\epsilon \in \mathbb{R}^{D \times D}$ is the covariance of $\left\{ \boldsymbol{\epsilon}_i \right\}_{i=1}^N$.

1.2.5

- Let $\boldsymbol{U}_d \in \mathbb{R}^{D \times d}$ be the top d eigenvectors corresponding to the d largest eigenvalues of $\boldsymbol{\Sigma}_x$.
- Show that:

$$\operatorname{Tr}\left\{\boldsymbol{\Sigma}_\epsilon\right\} = \sum_{i=d+1}^{D} \lambda_i\left(\boldsymbol{\Sigma}_x\right)$$

  where we assume $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D$

– –

High-dimensional data PCA

- Consider the data $\boldsymbol{X} \in \mathbb{R}^{D \times N}$ where $D > N$.

1.2.6

- Provide a (tight) upper bound on the number of non-zero eigenvalues.
- Consequently, can you apply PCA to $\boldsymbol{X} \in \mathbb{R}^{D \times N}$ to obtain $\boldsymbol{Z} \in \mathbb{R}^{d \times N}$ with $d < D$ such that there is no loss of information?
  Explain your answer.

– –

Rank minimization

- Let $\boldsymbol{A} \in \mathbb{R}^{D \times N}$.
- Consider the following rank minimization problem:

$$\begin{cases} \min_{\boldsymbol{M} \in \mathbb{R}^{D \times N}} \|\boldsymbol{A} - \boldsymbol{M}\|_F^2 \\ \text{s.t. } \operatorname{rank}(\boldsymbol{M}) \leq d \end{cases}$$

1.2.7

- Solve the optimization problem.
- Write your final solution using the (truncated) matrices obtained by the SVD decomposition of $\boldsymbol{A}$, namely, $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$

Hints:

1. Any matrix $\boldsymbol{M} \in \mathbb{R}^{D \times N}$ with $\operatorname{rank}(\boldsymbol{M}) = d$ can be written as $\boldsymbol{M} = \boldsymbol{B}\boldsymbol{C}$
   where:
   (a) $\boldsymbol{B} \in \mathbb{R}^{D \times d}$
   (b) $\boldsymbol{C} \in \mathbb{R}^{d \times N}$

   use this result to formulate (and solve) an equivalent unconstrained problem.
2. There is a strong connection to PCA.

– –

1.3 Implementation and applications

Solve this section in the attached notebook.

---

2 KPCA

2.1

Centering matrix

- Let $\boldsymbol{J} = \boldsymbol{I} - \frac{1}{N}\mathbf{1}\mathbf{1}^T \in \mathbb{R}^{N \times N}$ be the centering matrix.

2.1.1

Prove that $\boldsymbol{J}$ is idempotent, that is, $\boldsymbol{J}^2 = \boldsymbol{J}$.
In words, after applying centering once, the second centering has no effect.

Kernel matrix

- Let $\boldsymbol{X} \in \mathbb{R}^{D \times N}$ and let:

$$\boldsymbol{\Sigma}_x := \boldsymbol{X}\boldsymbol{X}^T$$
$$\boldsymbol{K}_x := \boldsymbol{X}^T\boldsymbol{X}$$

Let $(\boldsymbol{u}_i, \lambda_i)$ be an eigen pair of $\boldsymbol{\Sigma}_x$ such that $\boldsymbol{\Sigma}_x \boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i$ with $\lambda_i > 0$.

2.1.2

1. Show that λ_i is an eigenvalue of $\boldsymbol{K}_x$ as well.
2. Find its corresponding eigenvector such that $\boldsymbol{K}_x \boldsymbol{w}_i = \lambda_i \boldsymbol{w}_i$.

Kernel functions

- Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and consider $\left\{\boldsymbol{x}_i \in \mathbb{R}^d\right\}_{i=1}^{N}$.

2.1.3

Show that if k can be written as an inner product, that is

$$k\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = \left\langle \phi\left(\boldsymbol{x}_i\right), \phi\left(\boldsymbol{x}_j\right)\right\rangle$$

for some ϕ, then, the matrix defined by:

$$\boldsymbol{K}_x\left[i, j\right] = k\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)$$

is an SPSD matrix, namely, $\boldsymbol{K}_x \succeq 0$.

- Let $\boldsymbol{A}$ be an SPD matrix, and let:

$$k\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = \boldsymbol{x}_i^T \boldsymbol{A} \boldsymbol{x}_j$$

2.1.4

Prove or disprove:
k is a kernel function.

- Let $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathbb{R}^d$ and consider:

$$k\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = \left(1 + \boldsymbol{x}_i^T \boldsymbol{x}_j\right)^2$$

Prove or disprove:
k is a kernel function.

- Consider $\left\{\boldsymbol{x}_i \in \mathbb{R}^d\right\}_{i=1}^{N}$, and consider the kernel:

$$k\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) := \left\langle \phi\left(\boldsymbol{x}_i\right), \phi\left(\boldsymbol{x}_j\right)\right\rangle$$

  for some ϕ.

- Let:

$$\tilde{k}\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) := \left\langle \phi\left(\boldsymbol{x}_i\right) - \boldsymbol{\mu}_\phi, \phi\left(\boldsymbol{x}_j\right) - \boldsymbol{\mu}_\phi\right\rangle$$

  be the centered version, where:

$$\boldsymbol{\mu}_\phi = \frac{1}{N} \sum_{i=1}^{N} \phi\left(\boldsymbol{x}_i\right)$$

2.1.5

Show that $\tilde{k}$ can be written using only k, and without using ϕ and $\boldsymbol{\mu}_\phi$ explicitly.

- Let $\boldsymbol{K}_x \in \mathbb{R}^{N\times N}$ be a kernel matrix, that is:

$$\boldsymbol{K}_x[i,j] = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

  for some kernel function k.

- Let $\widetilde{\boldsymbol{K}}_x$ be the centered version, that is:

$$\widetilde{\boldsymbol{K}}_x = \boldsymbol{J}\boldsymbol{K}_x\boldsymbol{J}$$

  where $\boldsymbol{J} = \boldsymbol{I} - \frac{1}{N}\mathbf{1}\mathbf{1}^T$.

2.1.6

Prove or disprove:
$\widetilde{\boldsymbol{K}}_x$ is an SPD matrix.

Out of sample extension

- Let $\boldsymbol{K}_x$ be the kernel matrix obtained from the training set $\mathcal{X} = \left\{\boldsymbol{x}_i \in \mathbb{R}^D\right\}_{i=1}^N$.

- Let $\boldsymbol{Z} \in \mathbb{R}^{d\times N}$ be the low-dimensional representation obtained by applying KPCA, that is:

$$\boldsymbol{Z} = \boldsymbol{\Sigma}_d \boldsymbol{V}_d^T$$

  where $\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}^T = \boldsymbol{J}\boldsymbol{K}_x\boldsymbol{J}$ is an eigendecomposition (see lecture notes).

- Let $\boldsymbol{X}^\star \in \mathbb{R}^{D\times N^\star}$ be a set of new unseen data-points.

2.1.7

Write an expression (in a matrix form) for $\boldsymbol{Z}^\star \in \mathbb{R}^{d\times N^\star}$, the KPCA out of sample extension applied to $\boldsymbol{X}^\star$.

- Let $\mathcal{X}^{\star} = \{\boldsymbol{x}_i^{\star}\}_{i=1}^{N^{\star}} \subseteq \mathcal{X}$ be a subset of the training set $\mathcal{X}$.
- Let $\boldsymbol{X}^{\star} \in \mathbb{R}^{D \times N^{\star}}$be the matrix from of $\mathcal{X}^{\star}$.
- Let $\boldsymbol{Z}^{\star} \in \mathbb{R}^{d \times N}$ be the low-dimensional representation obtained by the training encoding.

2.1.8

Prove that the out of sample encoding applied to $\boldsymbol{X}^{\star}$ coincide with the training encoding $\boldsymbol{Z}^{\star}$.

2.2 Implementation and applications

Solve this section in the attached notebook.