Optimization and Data Science

Lecture 17: Principal Component Analysis

Prof. Dr. Thomas Slawig

Kiel University - CAU Kiel
Dep. of Computer Science

Summer 2020

Contents

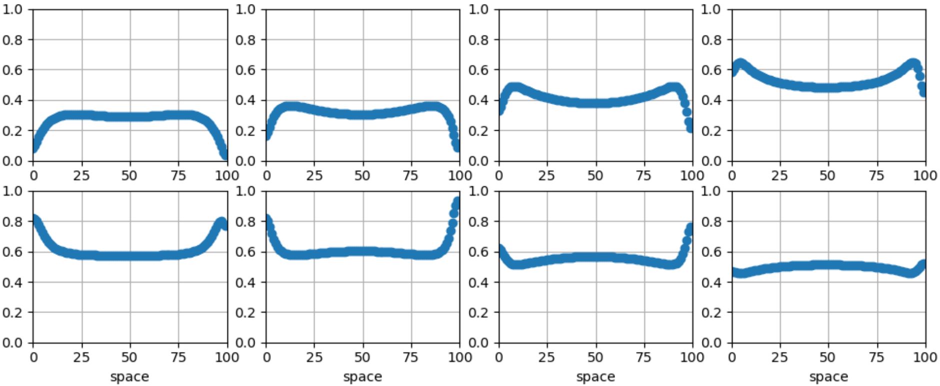
Contents

Principal Component Analysis

- What is it?
  Identification of main (principal) components of a data set, i.e., a matrix built of different data vectors ...
  ... coming from different experiments of from observations of a system at different times
  Other names: Proper orthogonal decomposition POD, Singular value decomposition SVD
- Why are we studying this?
  One important method of data analysis
- How does it work?
  Using either eigenvalue or singular value decomposition
- What if we can use it?
  Analysis of main components of data
  Statistical interpretation is possible when applied on covariance matrix
  Complexity reduction
  Uncertainty analysis

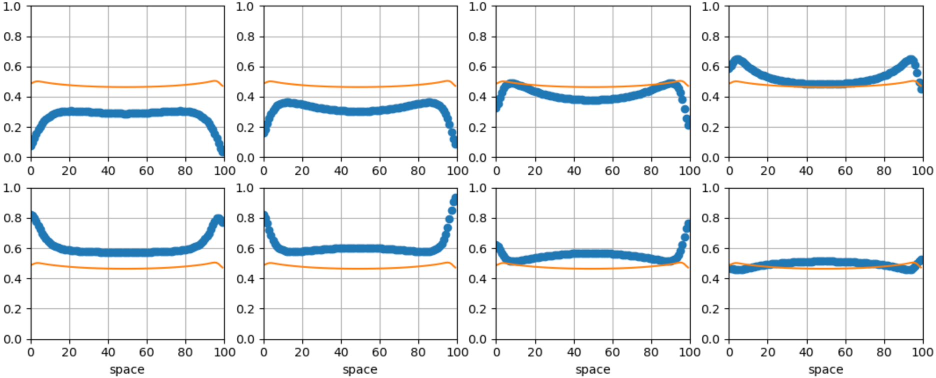
Basic idea

- Consider m sample data vectors where each $X_i = (X_{ij})_{j=1}^n$ is n-dimensional, ...
- ... e.g., m different measurements at n different spatial points.
- Example: $m = 8$ temporal snapshots, spatial points $n = 100$:



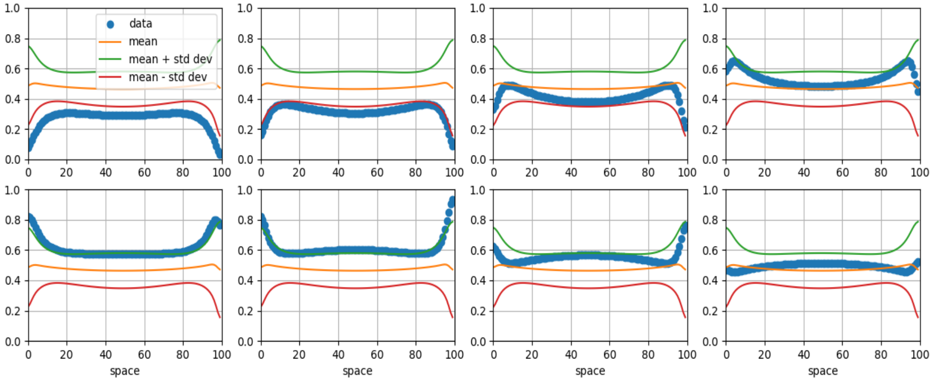
Component-wise mean (unbiased estimator for the expectation)

$$\bar{X} := \left(\bar{X}_j\right)_{j=1}^{n} := \left(\frac{1}{m}\sum_{i=1}^{m} X_{ij}\right)_{j=1}^{n}$$



(Unbiased) Estimator for component-wise variance

$$\frac{1}{m-1}\sum_{i=1}^{m}\left(X_{ij}-\bar{X}_j\right)^2, \quad j=1,\dots,n.$$



Estimator for the covariance

- Samples row-wise in a matrix

$$X := \begin{pmatrix} X_{11} & \cdots & X_{1n} \\ \vdots & & \vdots \\ X_{i1} & \cdots & X_{in} \\ \vdots & & \vdots \\ X_{m1} & \cdots & X_{mn} \end{pmatrix} = (X_{ij})_{i=1,\dots,m,j=1,\dots,n}$$

- Difference to the component-wise mean $\bar{X}_j$:

$$X - \bar{X} := \begin{pmatrix} X_{11} - \bar{X}_1 & \cdots & X_{1n} - \bar{X}_n \\ \vdots & & \vdots \\ X_{m1} - \bar{X}_1 & \cdots & X_{mn} - \bar{X}_n \end{pmatrix}$$

Sample covariance matrix

- The **sample covariance matrix**

$$\frac{1}{m-1}(X-\bar{X})^\top (X-\bar{X}), \quad \text{(if samples are row-wise)},$$

  is an unbiased estimator for the covariance of X.
- For a concrete realization $x \in \mathbb{R}^{m\times n}$ of X, the matrix

$$C := (x-\bar{x})^\top (x-\bar{x}) \in \mathbb{R}^{n\times n}$$

  is symmetric positive semi-definite.
- It is positive definite if $m \geq n$ and the data matrix $x \in \mathbb{R}^{m\times n}$ has full rank n, i.e., the data vectors are linear independent.

Contents

Recall: Eigenvalues and eigenvectors

Definition

- $\lambda \in \mathbb{C}$ is called **eigenvalue** of $A \in \mathbb{C}^{n \times n}$, if there exists $x \in \mathbb{C}^n \setminus \{0\}$ with $Ax = \lambda x$.
- x is called the corresponding **eigenvector**.

- Eigenvalues can be computed from

$$\begin{aligned} Ax = \lambda x, x \neq 0 \;&\Leftrightarrow\; (A - \lambda I)x = 0, x \neq 0 \\ &\Leftrightarrow\; (A - \lambda I) \text{ is singular} \\ &\Leftrightarrow\; \det(A - \lambda I) = 0. \end{aligned}$$

- A matrix in $\mathbb{C}^{n \times n}$ has n eigenvalues $\lambda_i \in \mathbb{C}, i = 1, \ldots, n$ (that do not have to be different).
- A *real* matrix in $\mathbb{R}^{n \times n}$ may have *complex* eigenvalues $\lambda_i \in \mathbb{C}, i = 1, \ldots, n$.
- The matrix we consider here is symmetric and positive (semi-) definite.

Properties of the eigenvectors of a symmetric matrix

Theorem

For a symmetric matrix $C \in \mathbb{R}^{n\times n}$, we have the following properties:

- *All eigenvalues $\lambda_i, i = 1, \dots, n$, are real.*
- *C can be diagonalized, i.e., there exists an invertible matrix $S \in \mathbb{R}^{n\times n}$ with*
$$C = SDS^{-1}, \quad D = diag((\lambda_i)_{i=1}^n) \in \mathbb{R}^{n\times n}.$$
The i-th column vector of S is an eigenvector to the eigenvalue λ_i.
- *The eigenvectors $s_i \in \mathbb{R}^n, i = 1, \dots, n$, form an orthonormal basis of $\mathbb{R}^n$, i.e.,*
$$s_i^\top s_j = \left\{ \begin{array}{ll} 1, & i = j \\ 0, & i \neq j \end{array} \right\}, i,j = 1, \dots, n. \tag{1}$$
- *As a consequence, every vector in $\mathbb{R}^n$ is a linear combination of the eigenvectors, i.e.,*
$$\forall v \in \mathbb{R}^n \, \exists c_i \in \mathbb{R}, i = 1, \dots, n : \quad v = \sum_{i=1}^n c_i s_i.$$

Properties of the estimator for the covariance matrix

For a realization $x \in \mathbb{R}^{m \times n}$ of X, the matrix

$$C := (x - \bar{x})^\top (x - \bar{x}) \in \mathbb{R}^{n \times n} \tag{2}$$

is symmetric positive (semi-)definite.

Theorem

Every symmetric positive (semi-)definite matrix $C \in \mathbb{R}^{n \times n}$ has only positive (non-negative) real eigenvalues

$$\lambda_1 \geq \ldots \geq \lambda_n > (\geq)\, 0.$$

Definition

The eigenvector to the i-the eigenvalue of the matrix C defined in (2) is called the i-th **principal component** of the data set x.

Properties of principal components

The above Theorem on page 12 has some useful consequences:

- Every ("new") data vector $w \in \mathbb{R}^n$ can be represented as

$$w = \sum_{i=1}^{n} c_i s_i,$$

- ..., i.e., it is a linear combination of the principal components.

$\rightsquigarrow$ We directly see which components are dominant and which are not.

- The same holds for its deviation from the mean $\bar{x}$:

$$w - \bar{x} = \sum_{i=1}^{n} \hat{c}_i s_i,$$

Properties of principal components

- Because of the orthonormality of the principal components/eigenvectors we have

$$w^\top s_k = \left(\sum_{i=1}^{n} c_i s_i \right)^\top s_k = \sum_{i=1}^{n} c_i \underbrace{s_i^\top s_k}_{=0 \text{ for } i \neq k} = c_k \underbrace{s_k^\top s_k}_{=1} \quad \text{for all } k = 1, \dots, n.$$

⇝ The coefficients are given by

$$c_k = w^\top s_k, \quad k = 1, \dots, n.$$

- Similarly for the deviation from the mean:

$$\hat{c}_k = (w - \bar{x})^\top s_k, \quad k = 1, \dots, n.$$

Contents

Recall: Singular value decomposition (SVD)

Theorem (Singular value decomposition)

Every matrix $A \in \mathbb{R}^{m\times n}$ *can be decomposed in the form*

$$A = U\Sigma V^\top$$

where

- $U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{n\times n}$ *are orthogonal matrices,*
- $\Sigma \in \mathbb{R}^{m\times n}$ *with* $\Sigma_{ij} = \left\{ \begin{array}{ll} \sigma_j \geq 0, & i = j, \\ 0, & i \neq j \end{array} \right\}, i = 1, \dots, m, j = 1, \dots, n.$
- *The* σ_j *are ordered by magnitude, i.e.,* $\sigma_j \geq \sigma_{j+1}$ *for all* j.

Singular value decomposition (SVD)

- For $m \geq n$:

$$\boxed{A} = \boxed{U} \underbrace{\begin{bmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_n \\ & & \\ 0 & \cdots & 0 \end{bmatrix}}_{=\Sigma} \boxed{V^\top}$$

- For $m < n$:

$$\boxed{A} = \boxed{U} \underbrace{\begin{bmatrix} \sigma_1 & & 0 & \cdots & 0 \\ & \ddots & & & \vdots \\ 0 & & \sigma_m & \cdots & 0 \end{bmatrix}}_{=\Sigma} \boxed{V^\top}$$

Eigenvalue vs. singular value decomposition

- Singular value decomposition of $B := x - \bar{x} \in \mathbb{R}^{m \times n}$:

$$B = U \Sigma V^\top$$

with orthogonal matrices $U \in \mathbb{R}^{m \times m}$, $V \in \mathbb{R}^{n \times n}$ gives

$$B^\top B = \left(U \Sigma V^\top\right)^\top U \Sigma V^\top = V \Sigma^\top \underbrace{U^\top U}_{=I} \Sigma V^\top = V \Sigma^\top \Sigma V^\top,$$

where for $m > n$:

$$\Sigma^\top \Sigma = \boxed{\begin{matrix} \sigma_1 & & 0 & \cdots & 0 \\ & \ddots & & & \vdots \\ 0 & & \sigma_n & \cdots & 0 \end{matrix}} \boxed{\begin{matrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_n \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{matrix}} = \operatorname{diag}\left(\sigma_1^2, \ldots, \sigma_n^2\right) \in \mathbb{R}^{n \times n}.$$

Eigenvalue vs. singular value decomposition

- For $m \leq n$ we get

$$\Sigma^\top \Sigma = \begin{pmatrix} \sigma_1 & & 0 \\ & \ddots & \\ 0 & & \sigma_m \\ \vdots & & \vdots \\ 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} \sigma_1 & & 0 & \cdots & 0 \\ & \ddots & & & \vdots \\ 0 & & \sigma_m & \cdots & 0 \end{pmatrix} = \operatorname{diag}\left(\sigma_1^2, \dots, \sigma_m^2, 0, \dots, 0\right) \in \mathbb{R}^{n \times n}.$$

Eigenvalue vs. singular value decomposition

- Singular value decomposition
$$B = U\Sigma V^\top \quad \text{with } B := x - \bar{x} \in \mathbb{R}^{m\times n}$$
gives with V orthogonal, i.e., $V^{-1} = V^\top$:
$$B^\top B = V \operatorname{diag}\big(\sigma_1^2, \ldots, \sigma_{\min(m,n)}^2, \underbrace{0, \ldots, 0}_{n-m \text{ times}}\big) V^\top.$$

- Eigenvalue decomposition gives
$$B^\top B = SDS^{-1}.$$

- Correspondence:
$$S = V, \quad D = \operatorname{diag}\left(\sigma_1^2, \ldots, \sigma_n^2\right), \sigma_i = 0 \text{ for } i > m.$$
and
$$\sigma_i \text{ singular value of } B \Leftrightarrow \lambda_i = \sigma_i^2 \text{ eigenvalue of } B^\top B.$$

What is important?

- Principal component analysis is a tool for data analysis.
- It is based on eigenvalue or singular value decomposition of a data matrix.
- Both methods are equivalent.
- The data matrix can be interpreted as (estimator of the) covariance matrix, if data are considered as realizations of random variables.
- The eigenvalues/singular values help to detect most important (i.e., principal) components of the data.
- This can be used to reduce the data size by omitting components corresponding to small eigenvalues.
- Data can be also clustered w.r.t. the magnitude of selected principal components.