High-dimensional data analysis

Lecture 3

Singular Value Decomposition and Principal Component Analysis

Innopolis University

Fall 2018

Similar matrices

- Matrices $A, B \in \mathbb{R}^{d \times d}$ are called <u>similar</u> if $\exists C \in \mathbb{R}^{d \times d}$ such that
  - C is invertible
  - $B = C^{-1}AC$

- Similar matrices have same:
  - Rank
  - Determinant
  - Trace
  - Eigenvalues
  - Frobenius norm ($\|A\|_F = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2\right)^{1/2} = \sqrt{\operatorname{tr} A^{\mathrm{T}} A}$)

Spectral Decomposition

- Let $A \in \mathbb{R}^{d \times d}$, symmetric and positive-definite, then
$$A = \Gamma \Lambda \Gamma^T,$$
  where
  - $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \dots, \lambda_d)$
    - $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d > 0$
  - $\Gamma = [\boldsymbol{\eta}_1 \quad \boldsymbol{\eta}_2 \quad \cdots \quad \boldsymbol{\eta}_d]$
    - $\forall i = 1 \dots d$: $A\boldsymbol{\eta}_i = \lambda_i \boldsymbol{\eta}_i$, $\|\boldsymbol{\eta}_i\| = 1$

- Property:
  - Γ is orthogonal ($\Gamma^T \Gamma = \Gamma \Gamma^T = \boldsymbol{I}$)

Spectral Decomposition

- Let $A \in \mathbb{R}^{d \times d}$, $r \stackrel{\text{def}}{=} \operatorname{rank} A < d$, then
$$A = \Gamma_r \Lambda_r \Gamma_r^T,$$
  where
  - $\Lambda_r = \operatorname{diag}(\lambda_1, \dots, \lambda_r) \in \mathbb{R}^{r \times r}$
    - $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0, \lambda_{r+1} = \cdots = \lambda_d = 0$
  - $\Gamma_r = [\boldsymbol{\eta}_1 \quad \boldsymbol{\eta}_2 \quad \cdots \quad \boldsymbol{\eta}_r] \in \mathbb{R}^{d \times r}$
    - $\forall i = 1 \dots d$: $A\boldsymbol{\eta}_i = \lambda_i \boldsymbol{\eta}_i$, $\|\boldsymbol{\eta}_i\| = 1$

- Property:
  - Γ is r-orthogonal ($\Gamma_r^T \Gamma_r = \boldsymbol{I}$, $\Gamma_r \Gamma_r^T \neq \boldsymbol{I}$)

Spectral Decomposition

- Let $A \in \mathbb{R}^{d\times d}$, rank $A = d$, $A = \Gamma\Lambda\Gamma^T$, then
  - $A = \sum_{i=1}^{d} \lambda_i \boldsymbol{\eta}_i \boldsymbol{\eta}_i^T$
  - $\exists\Theta$: $\mathrm{A} = \Theta^T\Theta$ (e.g., $\Theta = \Lambda^{1/2}\Gamma$)
  - $\forall q \in \mathbb{Q}$: $A^q = \Gamma\Lambda^q\Gamma^T$

Singular Value Decomposition

- Let $A \in \mathbb{R}^{d \times n}$, rank $A = r \leq d$,
$$A = UDV^T$$
where
  - $D = \text{diag}(d_1, d_2, \dots, d_r) \in \mathbb{R}^{r \times r}$
    - $d_1 \geq d_2 \geq \cdots \geq d_r > 0$
  - $U = [\boldsymbol{u}_1 \quad \boldsymbol{u}_2 \quad \cdots \quad \boldsymbol{u}_r] \in \mathbb{R}^{d \times r}$
  - $V = [\boldsymbol{v}_1 \quad \boldsymbol{v}_2 \quad \cdots \quad \boldsymbol{v}_r] \in \mathbb{R}^{n \times r}$
    - $\forall i = 1 \dots r$: $A^T \boldsymbol{u}_i = d_i \boldsymbol{v}_i, A\boldsymbol{v}_i = d_i \boldsymbol{u}_i$
- Property:
  - $V^T V = U^T U = \boldsymbol{I}$

Decompositions for the Sample Case

- Let $\mathbb{X} \in \mathbb{R}^{d\times n}$, $\mathbb{X} \sim \mathrm{Sam}(\bar{\boldsymbol{X}}, S)$, $\mathbb{X}_0 \stackrel{\text{def}}{=} \mathbb{X} - \bar{\boldsymbol{X}}$,
  - $\mathbb{X}_0 = UDV^T$,
  - $S = \Gamma\Lambda\Gamma^T$,
- then:
  - $U = \Gamma$
  - $D^2 = (n-1)\Lambda$

Principal Component Analysis

- Goals of Multivariate data analysis:
  - **<u>understand the structure in the data;</u>**
  - **<u>summarize data in simpler ways;</u>**
  - find the relationship between parts of the data;
  - make decisions based on the data.

Principal Component Analysis

- Instead of looking at individual features, let’s look at their combinations
- How to choose these combinations?
  - Linear combinations,
  - that best explain the variability of data
- How many combinations to choose?
  - More combinations: higher accuracy
  - Less combinations: faster computations, better interpretability

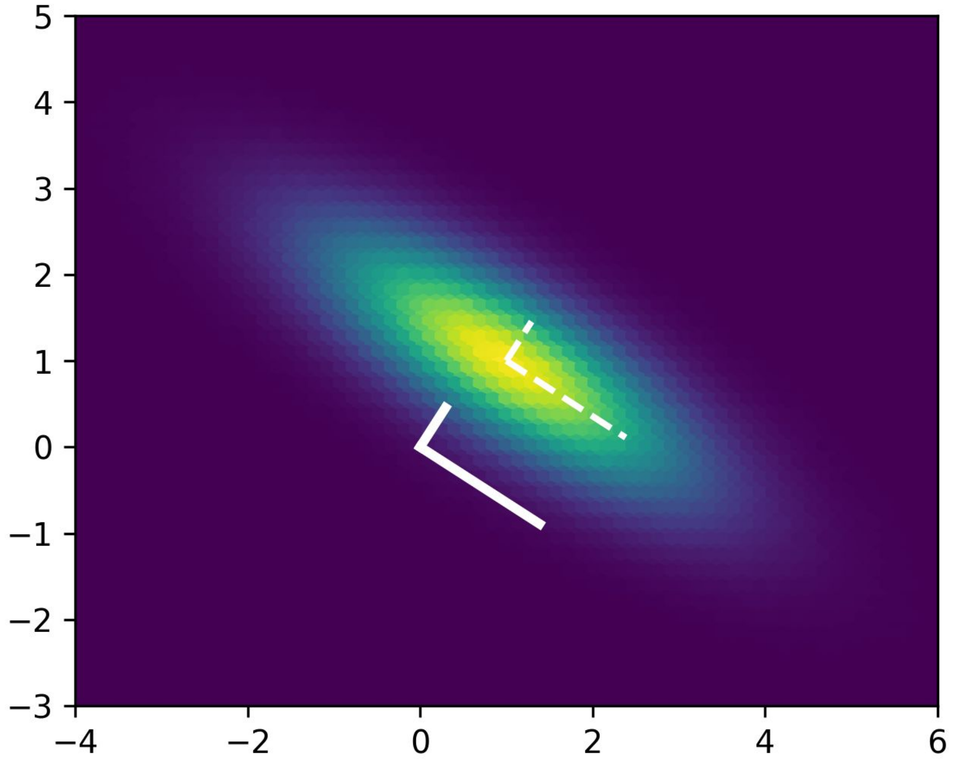
PCA: Population case

- $\boldsymbol{X} \sim (\boldsymbol{\mu}, \Sigma)$, $\Sigma = \Gamma\Lambda\Gamma^T$, $r = \operatorname{rank}\Sigma$, then for $k = 1 \ldots r$:
  - $W_k = \boldsymbol{\eta}_k^T(\boldsymbol{X} - \boldsymbol{\mu})$ – k-th principal component score
  - $\boldsymbol{W}^{(k)} = [W_1, \ldots, W_k]^T = \Gamma_k^T(\boldsymbol{X} - \boldsymbol{\mu})$ – k-th principal component vector
  - $\boldsymbol{P}_k = \boldsymbol{\eta}_k\boldsymbol{\eta}_k^T(\boldsymbol{X} - \boldsymbol{\mu}) = W_k\boldsymbol{\eta}_{\boldsymbol{k}}$ – k-th principal component projection

- First PC – largest eigenvalue – direction in which data has largest variance

PCA: Population case

- Example:
  - 2D Multivariate normal distribution
    - $\mu = [1 \;\; 1]^T$
    - $\Sigma = \begin{bmatrix} 2 & -1.1 \\ -1.1 & 1 \end{bmatrix}$
    - $\Lambda \approx \mathrm{diag}(2.7, 0.3)$
    - $\Gamma \approx \begin{bmatrix} 0.84 & 0.54 \\ -0.54 & 0.84 \end{bmatrix}$
  - Solid lines – principal components
  - Dashed lines – PCs, transposed by μ



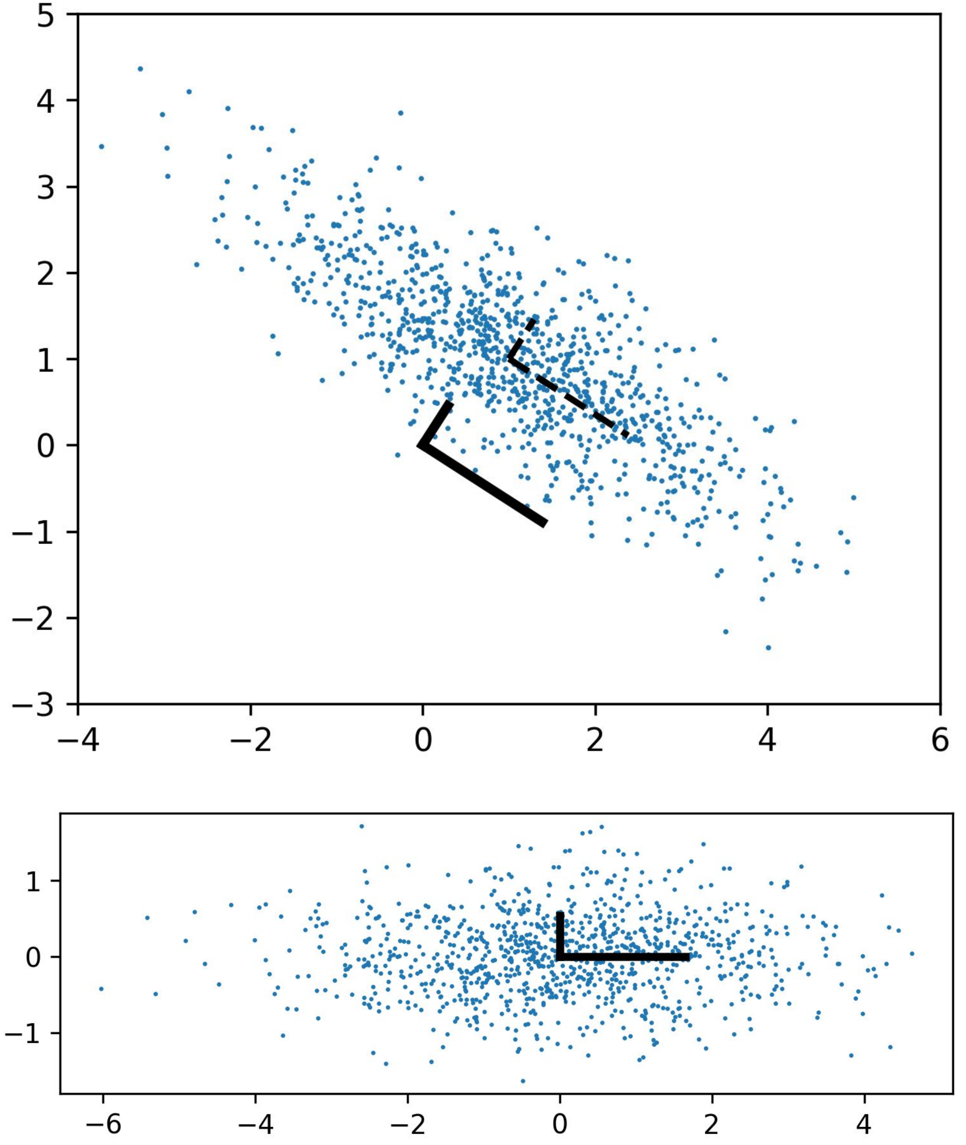
PCA: Sample case

- $\mathbb{X} \in \mathbb{R}^{d\times n}, \mathbb{X} \sim \mathrm{Sam}(\overline{\boldsymbol{X}}, S), S = \Gamma\Lambda\Gamma^T, r = \mathrm{rank}\, S$, then for $k = 1 \ldots r$:
  - $\boldsymbol{W}_{\blacksquare k} = \boldsymbol{\eta}_k^T(\mathbb{X} - \overline{\boldsymbol{X}}) \in \mathbb{R}^{1\times n}$ – k-th principal component score
  - $\mathbb{W}^{(k)} = [\boldsymbol{W}_{\blacksquare 1}, \ldots, \boldsymbol{W}_{\blacksquare k}]^T = \Gamma_k^T(\mathbb{X} - \overline{\boldsymbol{X}}) \in \mathbb{R}^{k\times n}$– k-th principal component data
  - $\mathbb{P}_{\blacksquare k} = \boldsymbol{\eta}_k\boldsymbol{\eta}_k^T(\mathbb{X} - \overline{\boldsymbol{X}}) = \boldsymbol{\eta}_{\boldsymbol{k}}\boldsymbol{W}_{\blacksquare k} \in \mathbb{R}^{d\times n}$– k-th principal component projection

- First PC – largest eigenvalue – direction in which data has largest variance
  - $\mathrm{Var}\left(\boldsymbol{\eta}_k^T\mathbb{X}\right) = \boldsymbol{\eta}_k^T S \boldsymbol{\eta}_k = \boldsymbol{\eta}_k^T\Gamma\Lambda\Gamma^T\boldsymbol{\eta}_k = \lambda_k$

PCA: Sample case

- Example:
  - 2D data
    - $\mathbb{X} \sim \mathrm{Sam}\left(\begin{bmatrix}1\\1\end{bmatrix}, \begin{bmatrix}2 & -1.1\\-1.1 & 1\end{bmatrix}\right)$
  - Solid lines – principal components
  - Dashed lines – PCs, transposed by μ
  - Lower figure: $\mathbb{W}^{(2)}$ (2nd PC data)



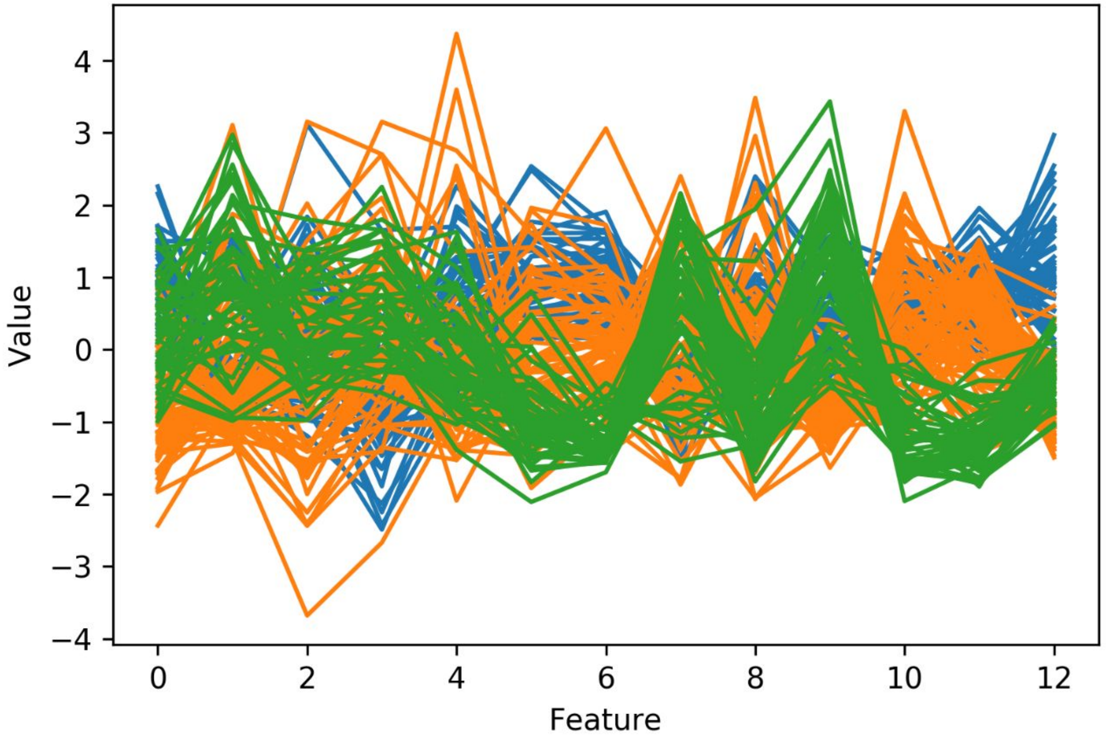
PCA: Dimensionality reduction

- We look for the subspace of dimension k onto which we can project our d-dimensional data
  - And we want to keep as much information as possible

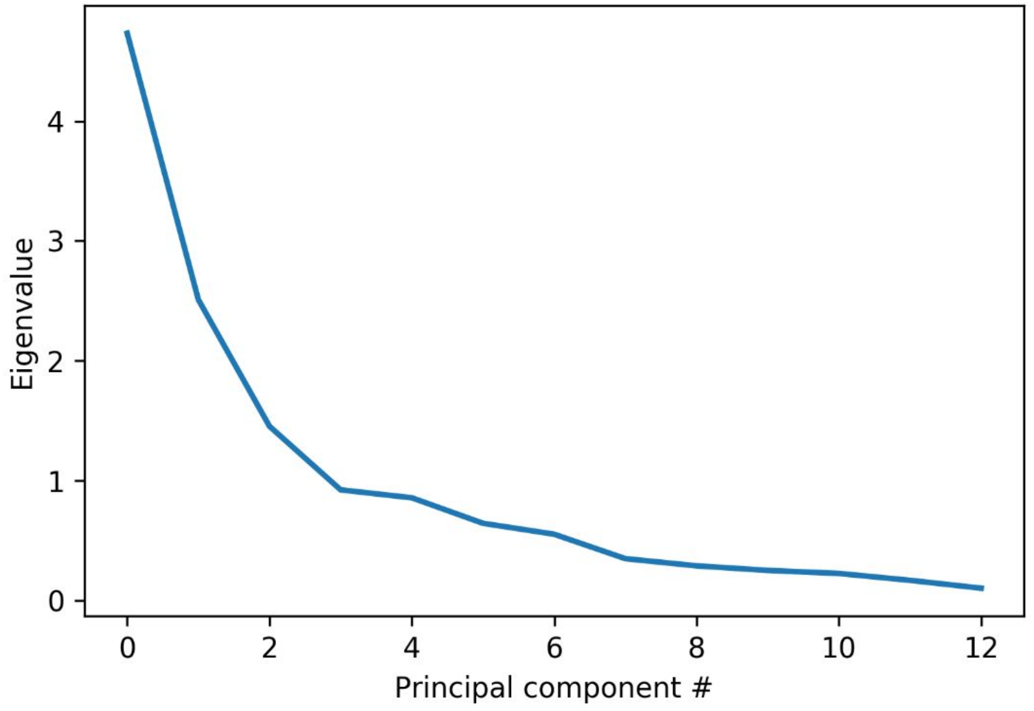
- PCA "preserves" covariance: $\mathrm{Var}\left(\boldsymbol{\eta}_k^T \mathbb{X}\right) = \lambda_k$

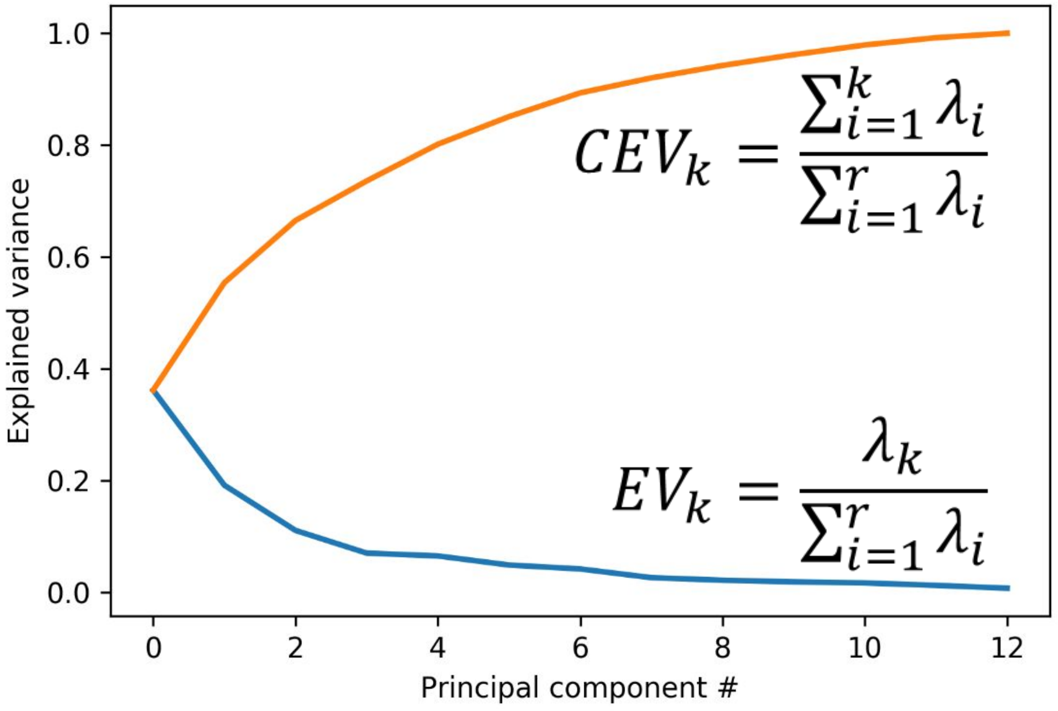
PCA: Dimensionality reduction

- UCI ML Wine Data Set:
  - 13 features (x-axis)
  - 178 samples (lines)
  - Values scaled to $\mathrm{Sam}(0, 1)$ (y-axis)
  - 3 classes (color)
- Principal components often correspond to underlying structure of the data
- As with most linear methods, data must be scaled for practical applications



PCA: Dimensionality reduction





PCA: Dimensionality reduction