WeST People and Knowledge Networks | Institute for Web Science & Technologies

UNIVERSITÄT KOBLENZ · LANDAU

Machine Learning and Data Mining WS21/22

"3 Data Transformation"

Dr. Zeyd Boukhers

@ZBoukhers

Institute for Web Science and Technologies

University of Koblenz-Landau

November 10, 2021

From previous lecture

- How to define a MLDM task.
- How to design high quality features
- How to pre-process the data.
  - Remove outliers.
  - Scale features.
  - Measure the correlation between features.
  - Handle missing data.

Agenda

- Data dimension reduction

Data dimension reduction

Data dimension reduction

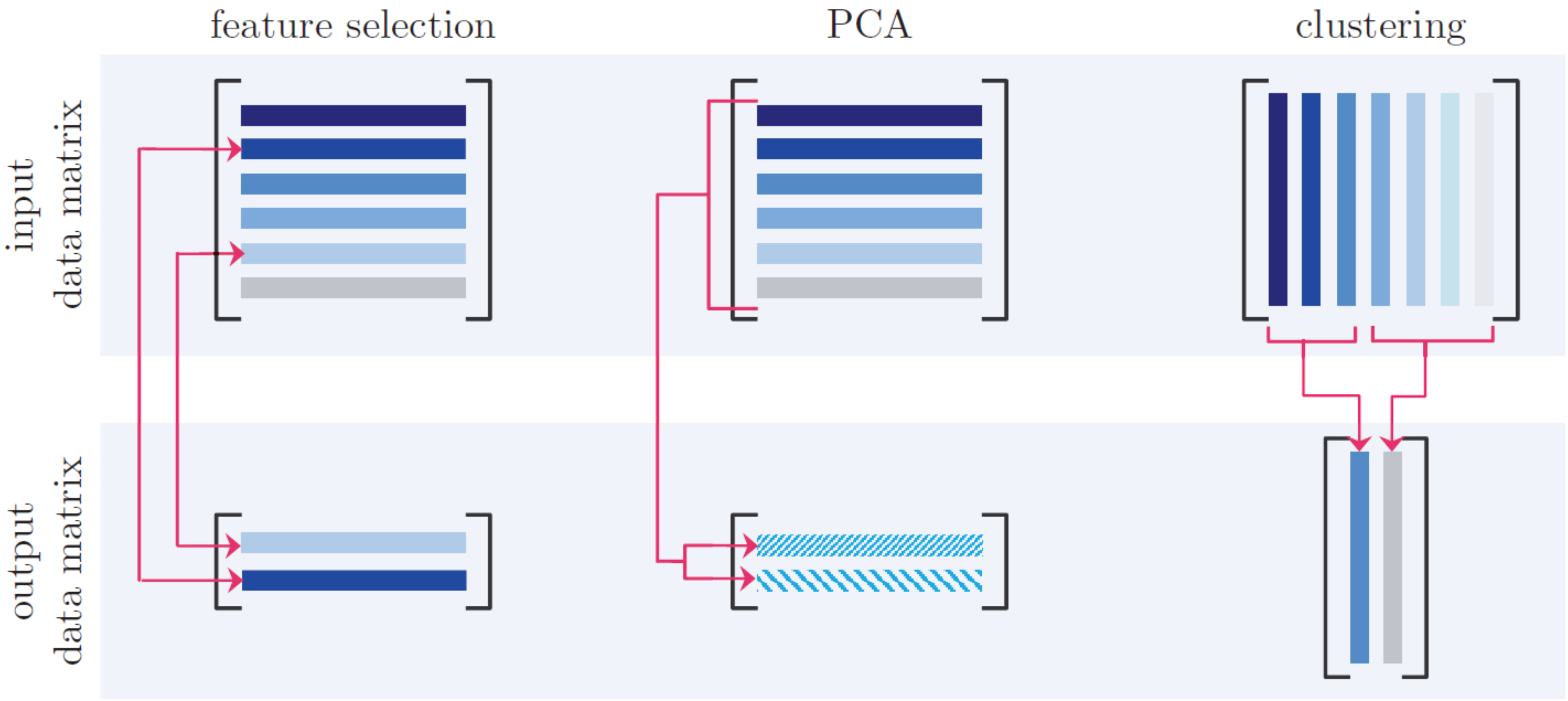


Image source: Watt, J., Borhani, R., & Katsaggelos, A. K. (2019). Machine learning refined.

PCA: Principal Components Analysis

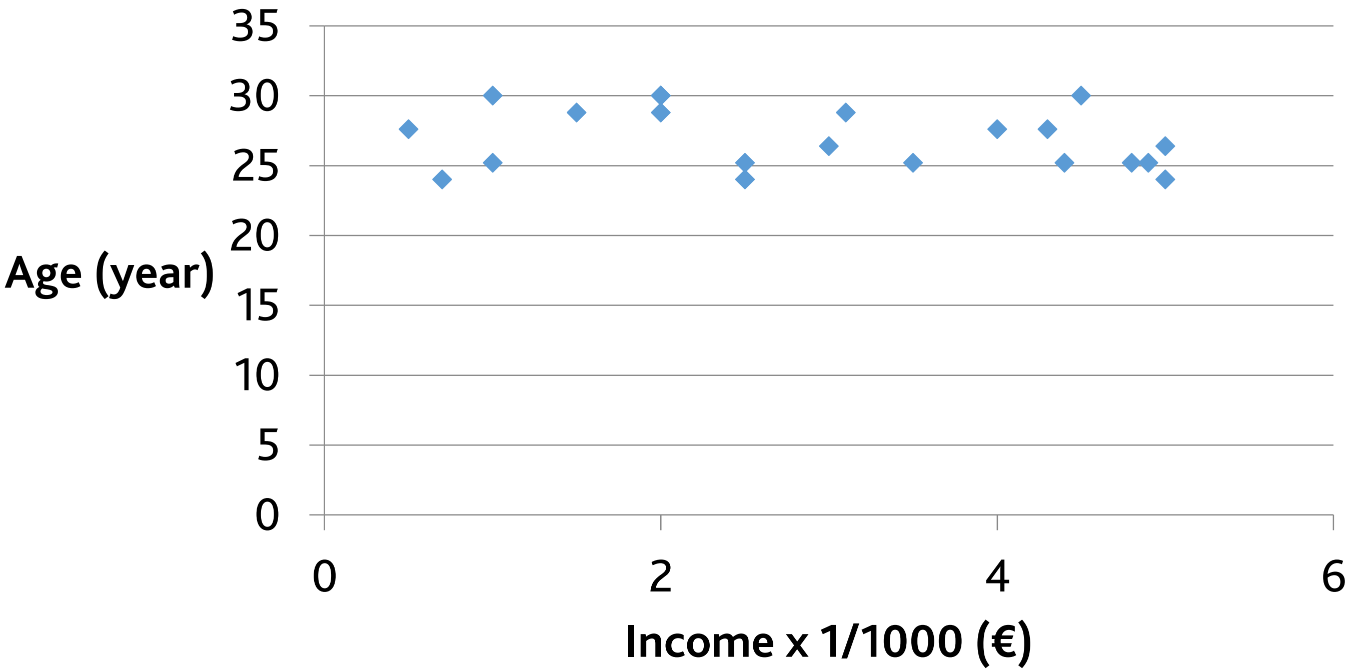
PCA: Principal Components Analysis

- What is PCA?
    - Linear transformation, invented in 1901 by Pearson and in 1933 by Hotteling.
    - Known also as « Principal Factor Analysis », referring to its first application in 1954 by Goodall.
    - Summarizes correlated data of l attributes by K uncorrelated axes (Principal Components).
    - The first component displays the maximum variance, the second displays the second maximum variance and so on.
- Why PCA?
    - Efficient reduction of the data.
    - Better visualization of the data.
    - Better classification.
    - Etc.

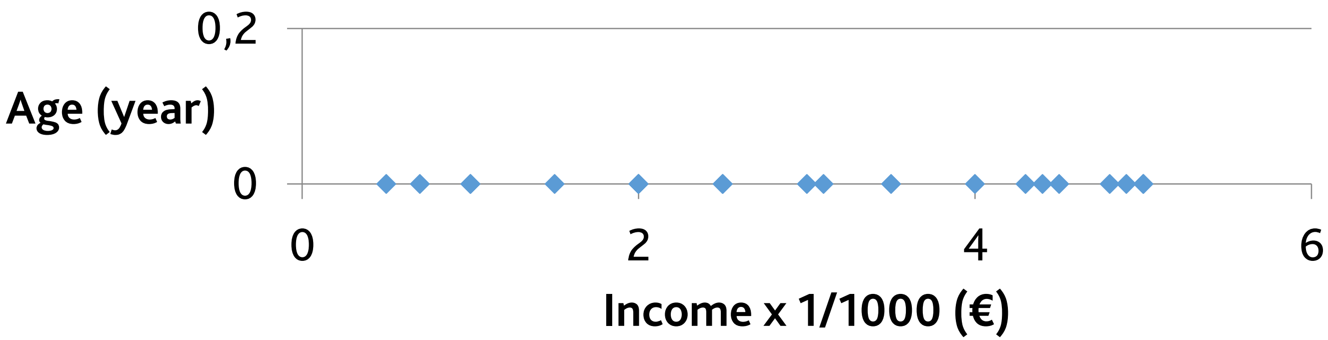
PCA: Principal Components Analysis

- Examples:
  - How to efficiently present a data X of size $l \times N$ by K principal components without information loss?
    - $X: \boldsymbol{x}_1, \dots, \boldsymbol{x}_N \to X': \boldsymbol{x}'_1, \dots, \boldsymbol{x}'_N: \ \mathbb{R}^l \to \mathbb{R}^K$
  - How to achieve the same or better accuracy with less dimensions?
    - Classification
    - Clustering
    - Any Machine learning / data mining task

- Overview:
  - Considering the following samples, each of which is presented by income and age.
  - Are both attributes important to understand the data?

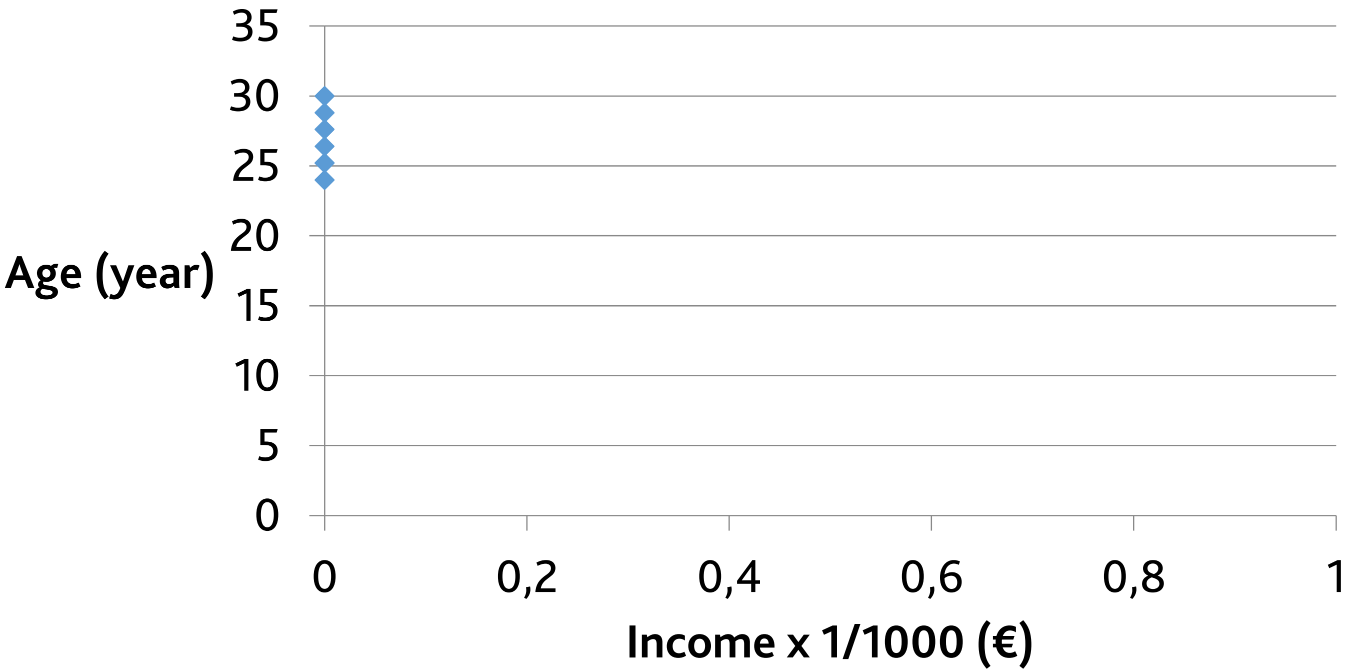


- Overview:
  - What if we simply ignore age, does this change anything in understanding the data?
  - Age does not give such valuable information which can help to understand and analyze the data.

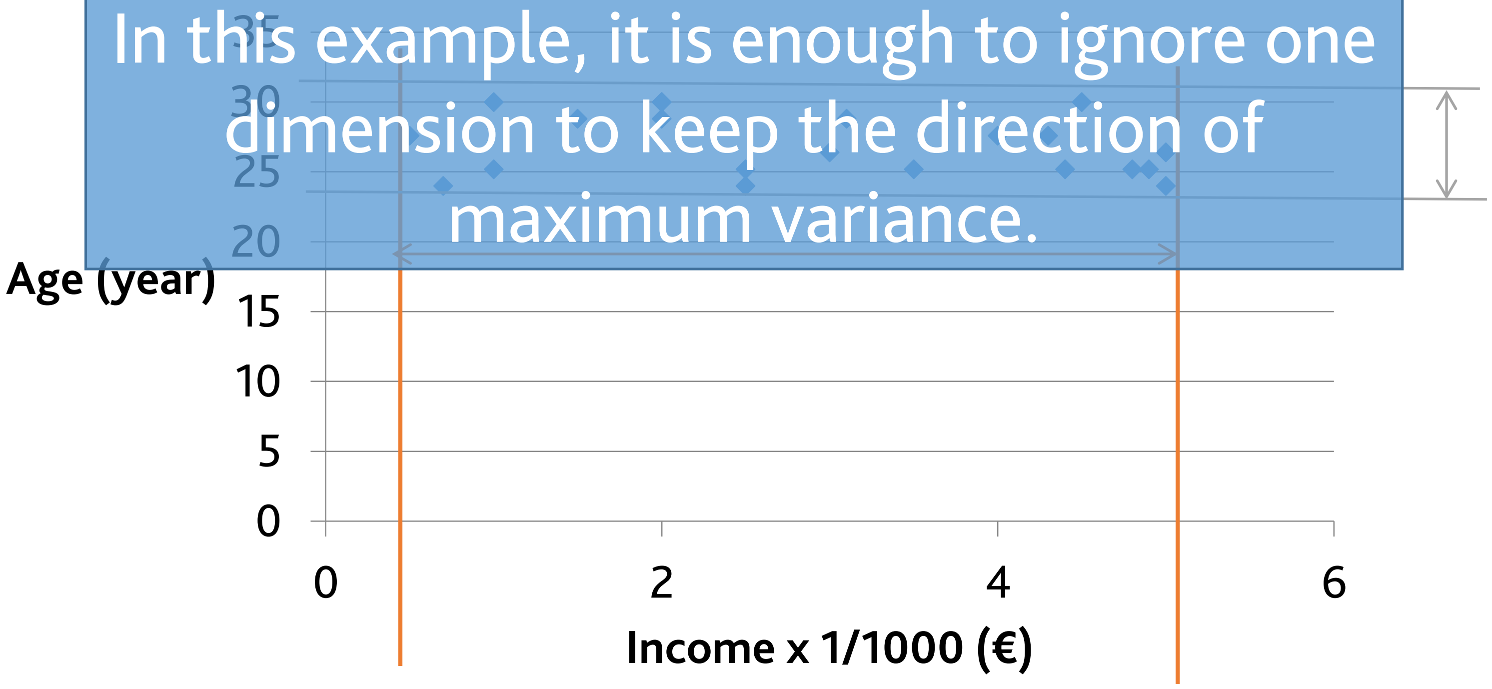


PCA: Principal Components Analysis

- Overview:
  - What if we simply ignore income, does this change anything in understanding the data?
  - Income was important to keep a high variance among samples.
  - But, what makes us consider Income much important feature than age by only visual assessments?

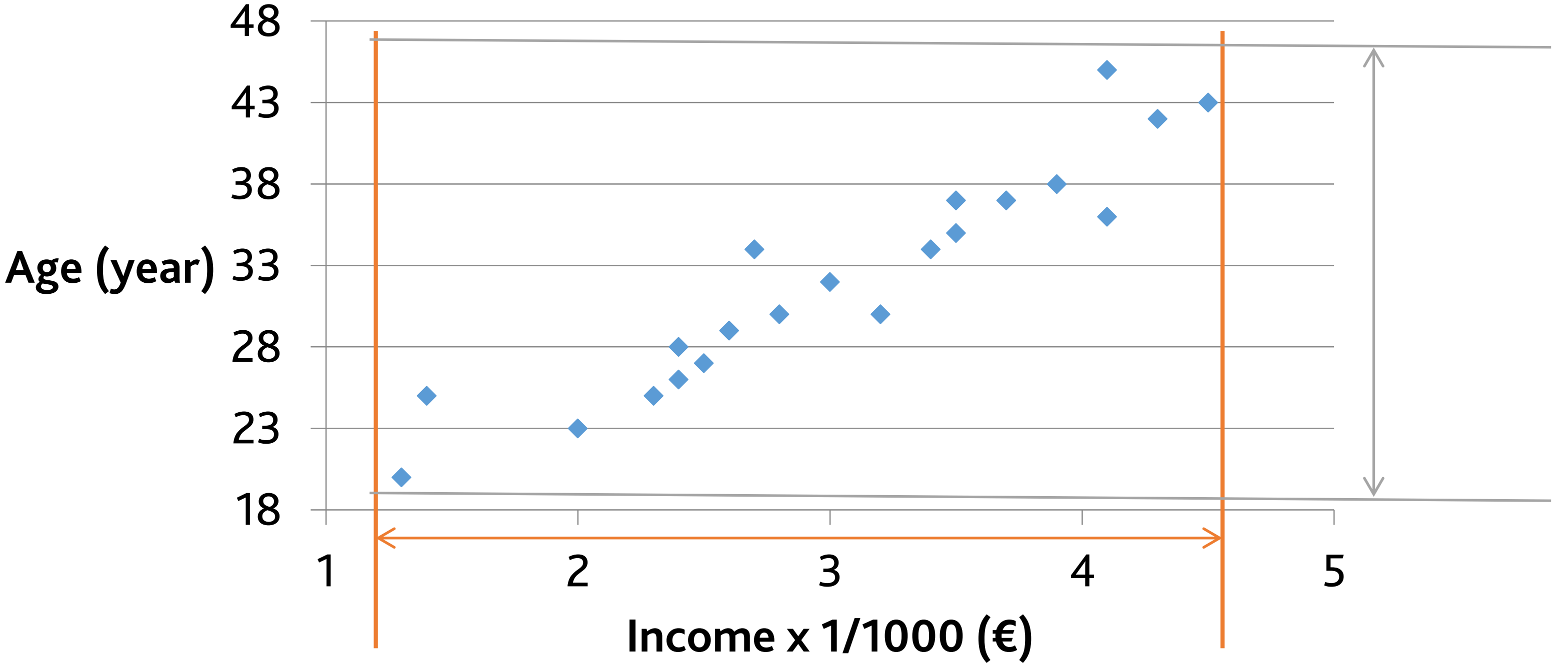


- Overview:
  - what makes us consider Income much important feature than age by only visual assessment?
  - **The answer is:** how the data is spread out or in another word the variance of the data.

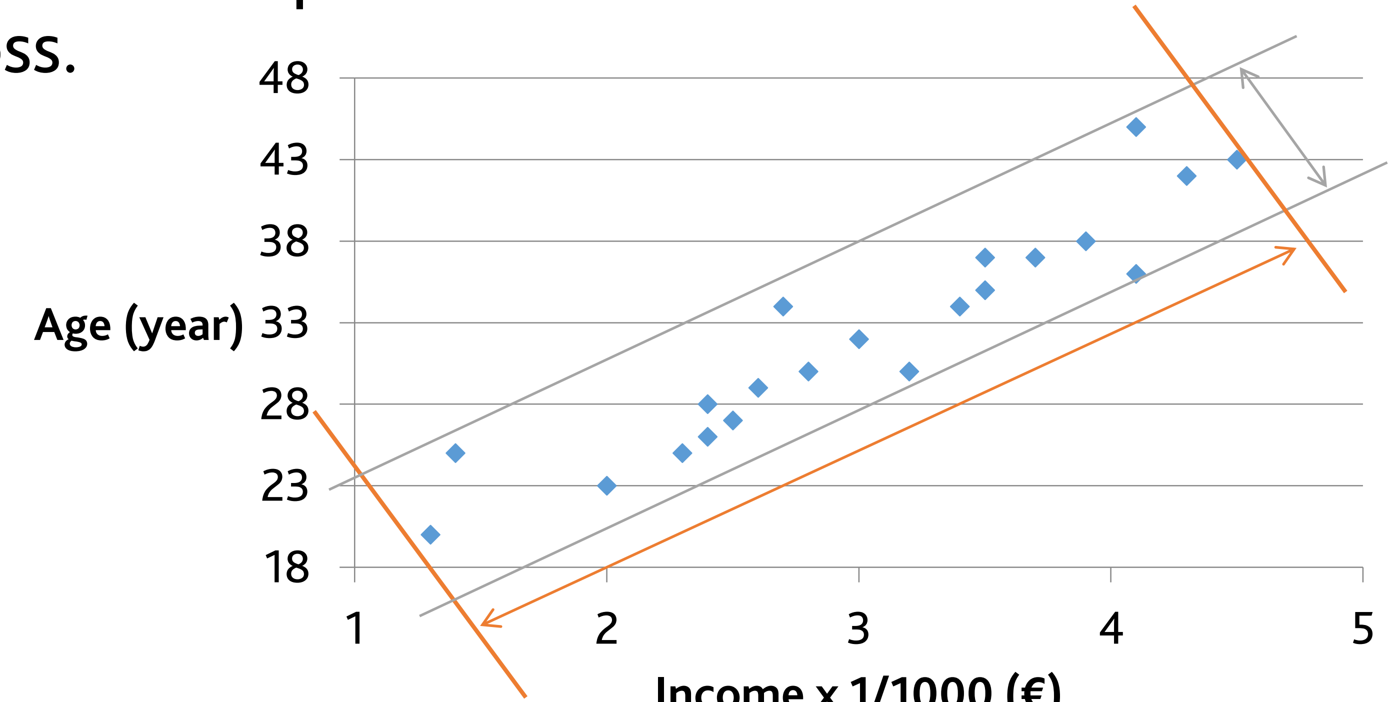


In this example, it is enough to ignore one dimension to keep the direction of maximum variance.

- Overview:
  - It is not possible to flat the data on one axis because the one-d variations along both dimensions are quite similar.
  - However, if we consider a rotated coordinate system (45°), we can make it different.

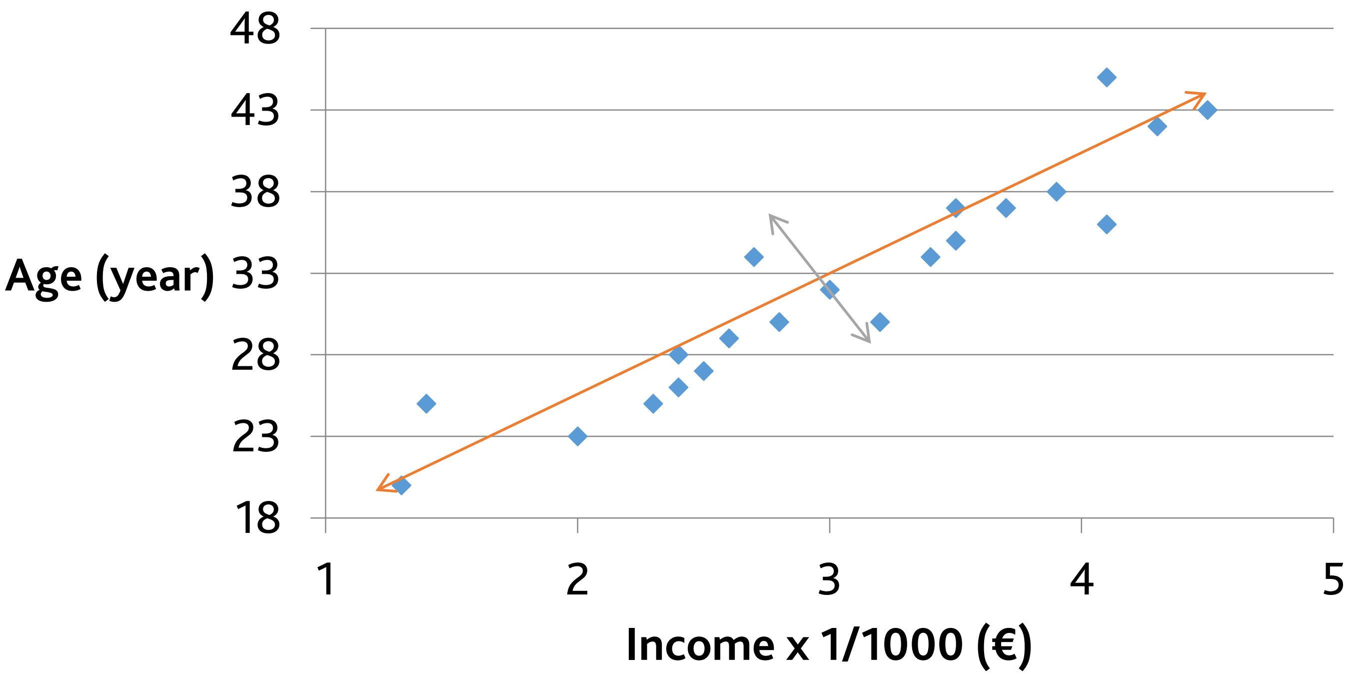


- Overview:
  - The new axes in the rotated coordinate system represent neither age nor income, but the combination of both.
  - Now, the data can be represented on one axis without too much information loss.



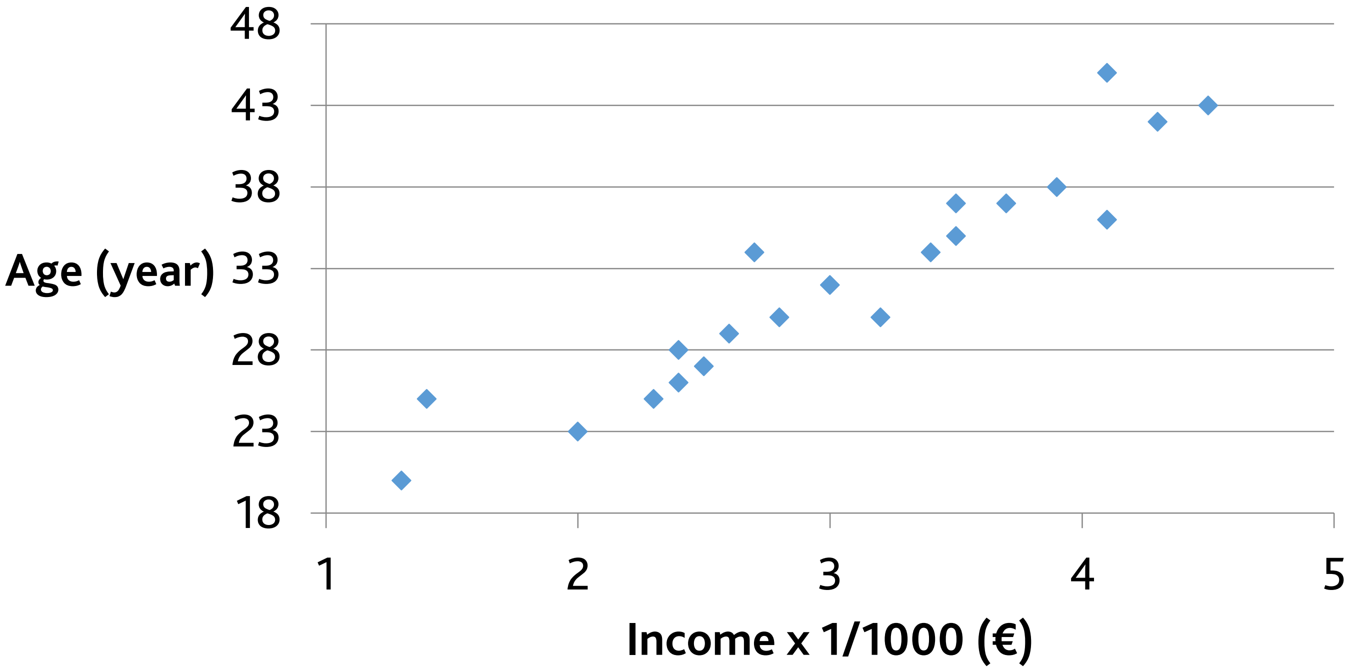
PCA: Principal Components Analysis

- The new axes that span the direction of the highest variances are called Principal components.
- Note that the first PC (Principal Component) describes the highest variance, the second PC describes the second highest variance and the k^{th} PC describes the k^{th}highest variance.



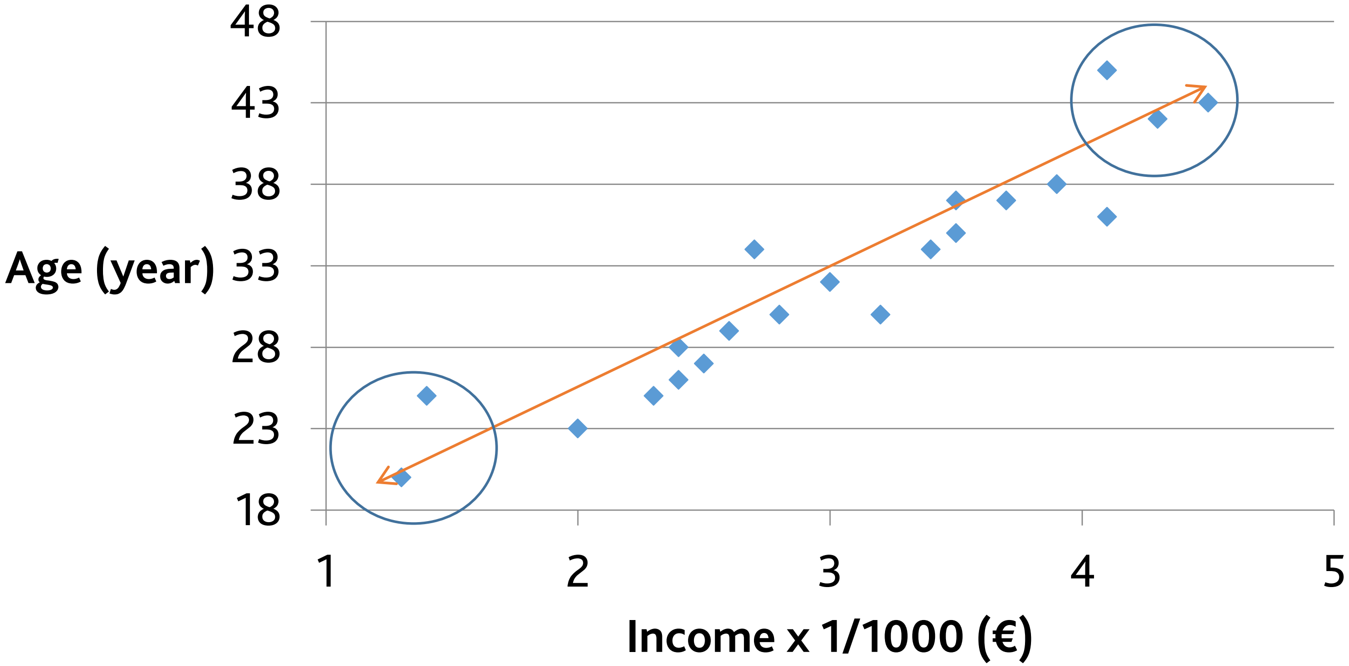
PCA: Principal Components Analysis

- In this example, which data points influence more the PCs in terms of length and direction?
- Considering only PC1:



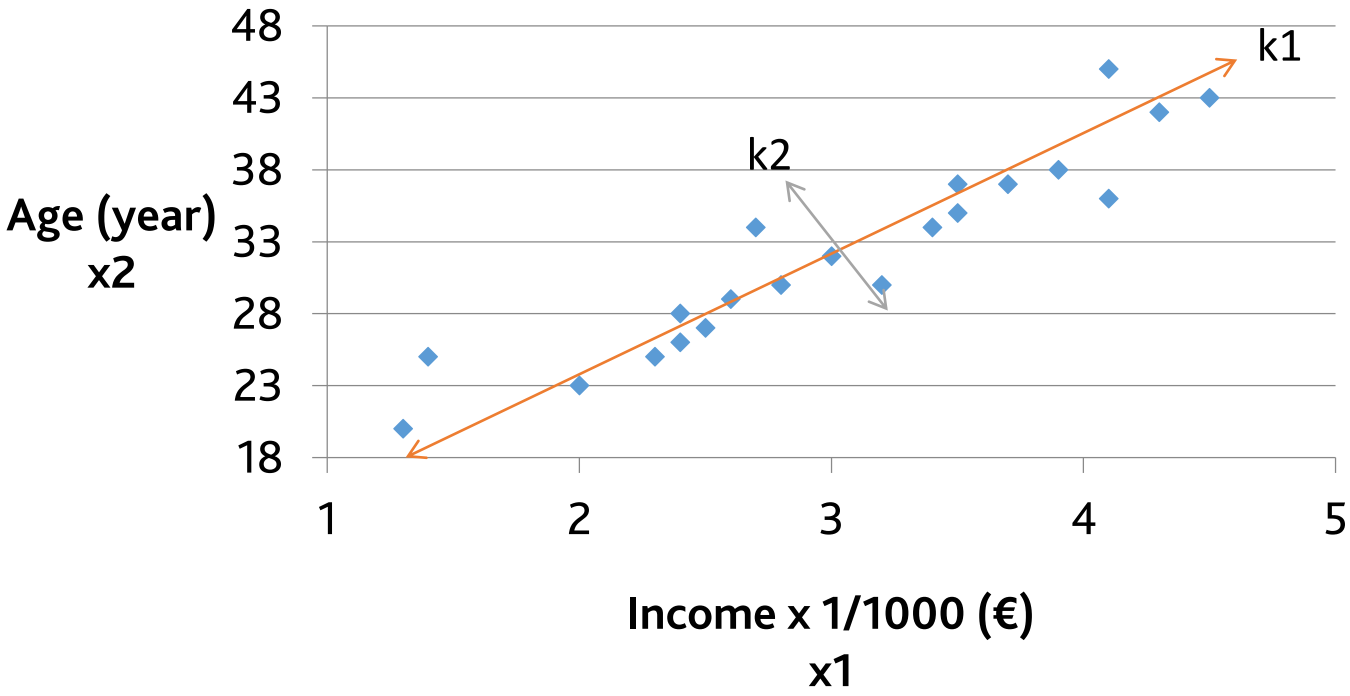
PCA: Principal Components Analysis

- In this example, which datapoints influence more the PCs in terms of length and direction?
- Considering only PC1:
- The ones at endpoints influence more the PC, since they maximize the variance more than other data points.



PCA: Principal Components Analysis

- Given N training samples: D: $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, where $\boldsymbol{x}_i$ and y_i denotes the feature vector and the label of the ith instance, respectively.
- **Goal:** Project X of dimension $l \times N$ onto X' having dimensionality $K \times N (K < l)$ while maximizing the variance of X' .



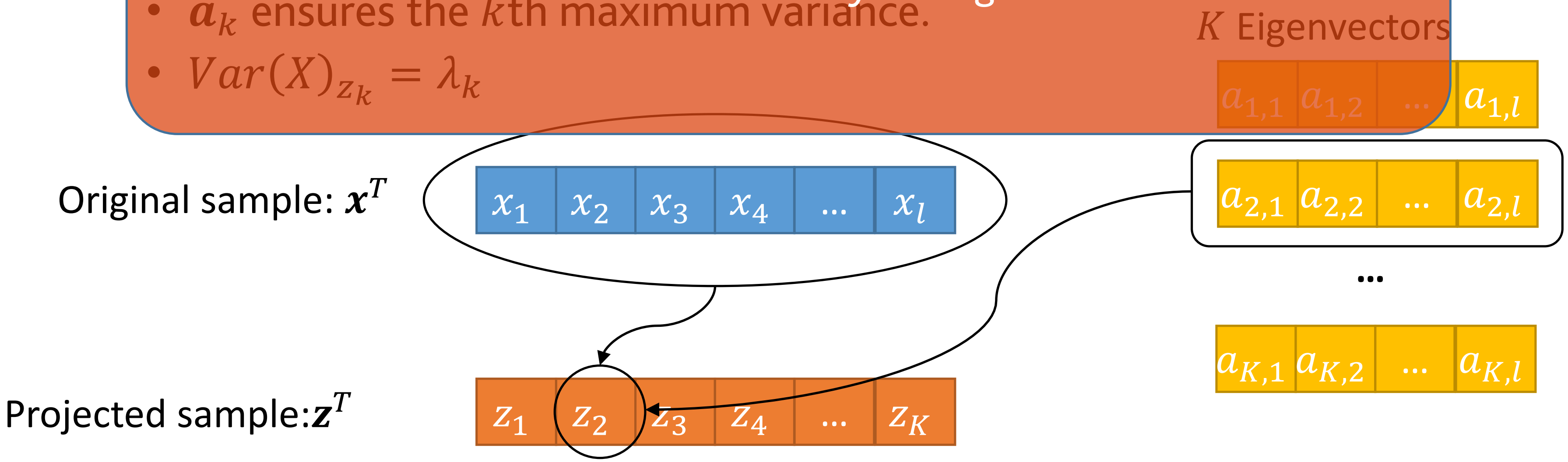
PCA: Principal Components Analysis

- Goal: find the K principal vectors such that:

$$z_k = \boldsymbol{a}_k^T \boldsymbol{x} = \sum_{j=1}^{l} a_{j,k} x_j$$

  - Where
    - z_k is the kth principal component.
    - $|\boldsymbol{a}_k| = l$, for $k = 1,2,\dots,K$
    - $\boldsymbol{a}_1$ ensures the maximum variance.
    - $\boldsymbol{a}_k$ ensures the kth maximum variance.
    - $Var(X)_{z_k} = \lambda_k$

The order is controlled by the eigenvalues



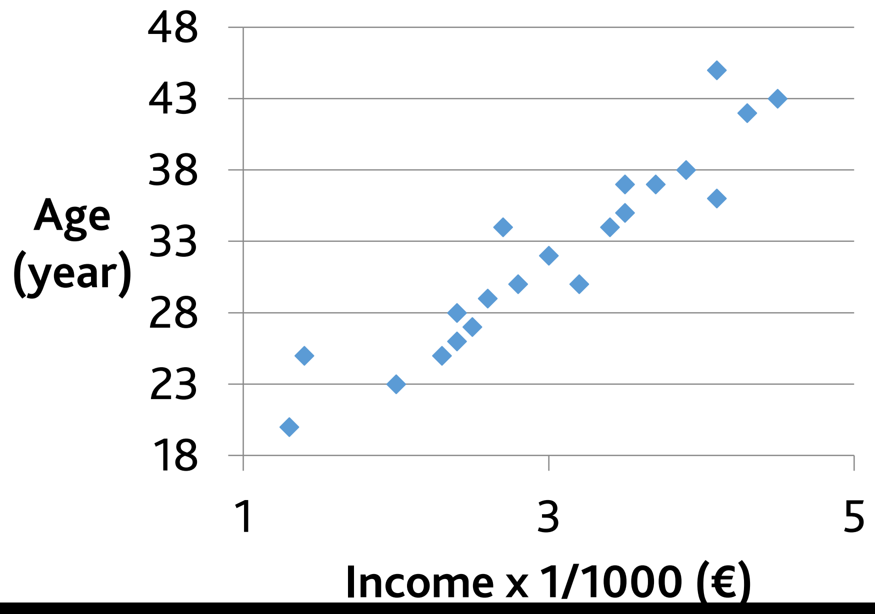
PCA: Principal Components Analysis

- Goal: find $\boldsymbol{a}_{k=1}$ that maximizes $Var(\boldsymbol{z}_{k=1})$
  - Until $k \rightarrow K$
- $\boldsymbol{a}_k$ may contain very big values
  - $\boldsymbol{a}_k^T \boldsymbol{a}_k = 1$ (unit vector)

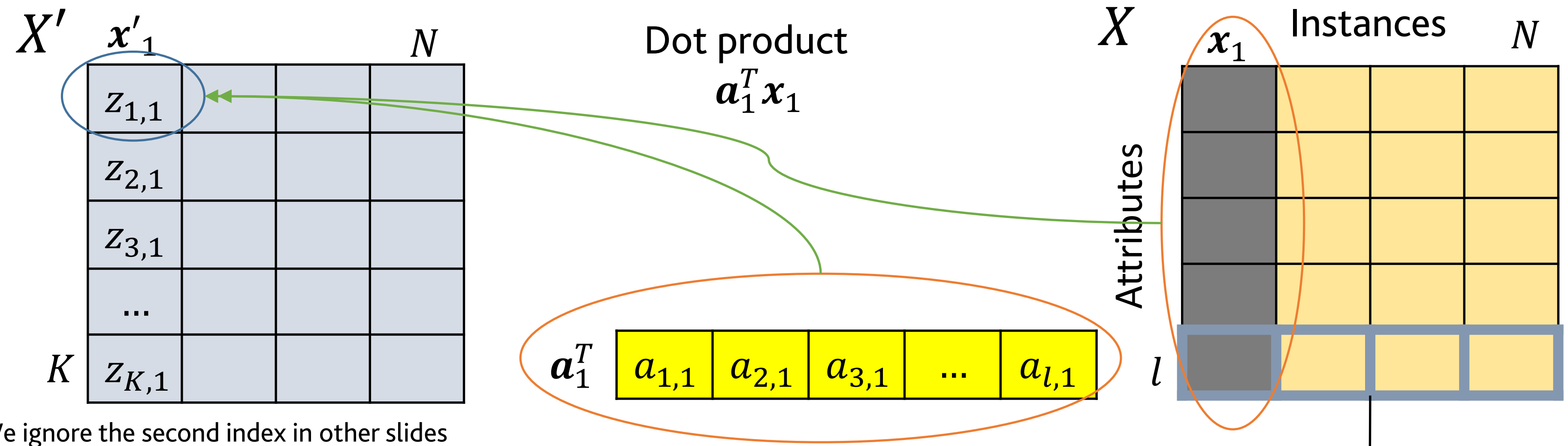
PCA: Principal Components Analysis

- $Var(income) = \frac{1}{N}\sum_{i=1}^{N}(x_{i,income} - \bar{\boldsymbol{x}}_{income})^2$
- Given that $z_k = \boldsymbol{a}_k^T \boldsymbol{x}$:
  - $Var(\boldsymbol{z}_k) = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{a}_k^T \boldsymbol{x}_i - \boldsymbol{a}_k^T \bar{X})^2$
    - $\bar{X}$ is the mean over all samples
  - $Var(\boldsymbol{z}_k) = \boldsymbol{a}_k^T \boldsymbol{S} \boldsymbol{a}_k$
    - $\boldsymbol{S}$ is the covariance matrix
  - $\boldsymbol{S} = Cov(X)$
  - $= \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i - \bar{X})(\boldsymbol{x}_i - \bar{X})^T$

See next « bonus » slide!



Bonus slide



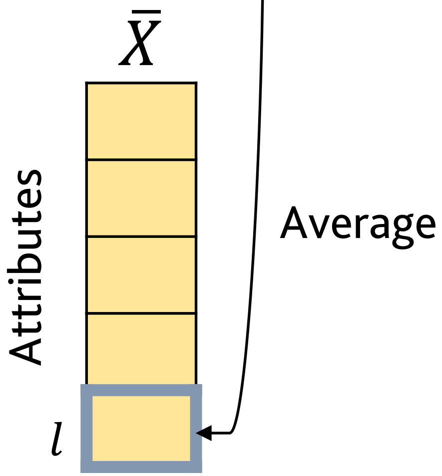
We ignore the second index in other slides for the sake of simplicity.

Why $\overline{\boldsymbol{a}_k^T X} = \boldsymbol{a}_k^T \bar{X}$?

$$\overline{\boldsymbol{a}_k^T X} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{a}_k^T \boldsymbol{x}_i$$

$$= \frac{\boldsymbol{a}_k^T}{N} \sum_{i=1}^{N} \boldsymbol{a}_k^T \boldsymbol{x}_i$$

$$= \boldsymbol{a}_k^T \bar{X}$$



PCA: Principal Components Analysis

- For the first PC, the task becomes now:
  - Maximizing $\boldsymbol{a}_1^T \boldsymbol{S} \boldsymbol{a}_1$ with the constraint $\boldsymbol{a}_k^T \boldsymbol{a}_k = 1$
    - Lagrange multiplier
- The maximization becomes unconstrained

$$\boldsymbol{a}_1^T \boldsymbol{S} \boldsymbol{a}_1 + \lambda_1 (1 - \boldsymbol{a}_1^T \boldsymbol{a}_1)$$

- By differentiating w.r.t $\boldsymbol{a}_1$

$$\frac{\partial}{\partial \boldsymbol{a}_1} \left(\boldsymbol{a}_1^T \boldsymbol{S} \boldsymbol{a}_1 - \lambda_1 (\boldsymbol{a}_1^T \boldsymbol{a}_1 - 1) \right) = 0$$

$$\Longrightarrow \boldsymbol{S} \boldsymbol{a}_1 - \lambda_1 \boldsymbol{a}_1 = 0$$

$$\Longrightarrow \boldsymbol{S} \boldsymbol{a}_1 = \lambda_1 \boldsymbol{a}_1$$

$$\Longrightarrow \boldsymbol{a}_1^T \boldsymbol{S} \boldsymbol{a}_1 = \lambda_1$$

Remember: $\boldsymbol{a}_1^T \boldsymbol{a}_1 = 1$

The variance will be a maximum when we set $\boldsymbol{a}_1$ equals to the eigenvector having the largest eigenvalue λ_1

- $\boldsymbol{a}_1$ is called the first principal component
- The second principal component is the one that maximises $\boldsymbol{a}_2^T \boldsymbol{S} \boldsymbol{a}_2$ subject to: $\boldsymbol{z}_2$ and $\boldsymbol{z}_1$ are uncorrelated.
  - $Cov(\boldsymbol{z}_1, \boldsymbol{z}_2) = 0$

$$\Longrightarrow \boldsymbol{a}_1^T \boldsymbol{S} \boldsymbol{a}_2 = 0$$
$$\Longrightarrow \boldsymbol{a}_2^T \boldsymbol{S} \boldsymbol{a}_1 = 0$$
$$\Longrightarrow \boldsymbol{a}_2^T \lambda_1 \boldsymbol{a}_1 = 0$$

  - Two constraints:
    - $\boldsymbol{a}_2^T \boldsymbol{a}_2 = 1$
    - $\boldsymbol{a}_2^T \lambda_1 \boldsymbol{a}_1 = 0$
    - Two Lagrange multipliers: λ_2 and ϕ

PCA: Principal Components Analysis

- The unconstrained maximization of $\boldsymbol{a}_2^T\boldsymbol{S}\boldsymbol{a}_2$ becomes

$$\boldsymbol{a}_2^T\boldsymbol{S}\boldsymbol{a}_2 + \lambda_2\left(1 - \boldsymbol{a}_2^T\boldsymbol{a}_2\right) + \phi(0 - \boldsymbol{a}_2^T\boldsymbol{a}_1)$$

- By differentiating w.r.t $\boldsymbol{a}_2$

$$\frac{\partial}{\partial \boldsymbol{a}_2}\left(\boldsymbol{a}_2^T\boldsymbol{S}\boldsymbol{a}_2 - \lambda_2(\boldsymbol{a}_2^T\boldsymbol{a}_2 - 1) - \phi\boldsymbol{a}_2^T\boldsymbol{a}_1\right) = 0$$
$$\Longrightarrow \boldsymbol{S}\boldsymbol{a}_2 - \lambda_2\boldsymbol{a}_2^T - \phi\boldsymbol{a}_1 = 0$$
$$\Longrightarrow \boldsymbol{a}_1^T\,\boldsymbol{S}\boldsymbol{a}_2 - \lambda_2\boldsymbol{a}_1^T\boldsymbol{a}_2^T - \phi\boldsymbol{a}_1^T\boldsymbol{a}_1 = 0$$
$$\Longrightarrow 0 - 0 - \phi = 0$$
$$\Longrightarrow \phi = 0$$
$$\Longrightarrow \boldsymbol{S}\boldsymbol{a}_2 - \lambda_2\boldsymbol{a}_2^T = 0$$
$$\Longrightarrow \boldsymbol{a}_2^T\,\boldsymbol{S}\boldsymbol{a}_2 = \lambda_2$$

- $\boldsymbol{a}_2$ is the eigenvector associated with the second largest eigenvalue λ_2 yielding the second PC
- This process can be repeated up to $K = l$ eigenvectors.

PCA: Principal Components Analysis

- Let A be an orthogonal $K \times l$ matrix consisting of K eigenvectors:
  - $\boldsymbol{z} = A^T \boldsymbol{x}$
  - Then: $Cov(Z) = \Lambda = A^T \boldsymbol{S} A$
  - For the data matrix X:
    - $Z = A^T X$

PCA: Principal Components Analysis

- How to compute the eigenvectors and eigenvalues of $\boldsymbol{S}$?
  - If $\boldsymbol{S}$ is a square matrix, a non zero vector $\boldsymbol{a}_k$ is an eigenvector of $\boldsymbol{S}$ such that: $\boldsymbol{S}\boldsymbol{a}_k = \lambda_k \boldsymbol{a}_k$, where λ_k is the corresponding eigenvalue. For example:
    - $\begin{pmatrix} 2 & 6 \\ 1 & 1 \end{pmatrix}\begin{pmatrix} 6 \\ 2 \end{pmatrix} = \begin{pmatrix} 24 \\ 8 \end{pmatrix} = 4\begin{pmatrix} 6 \\ 2 \end{pmatrix}$
  - Considering all eigenvectors:

$$\boldsymbol{S}A = \boldsymbol{\lambda} I A$$

Let $\Psi = \boldsymbol{\lambda} I$

$$\boldsymbol{S} = A\Psi A^T$$

Remember A is orthonormal $\Longrightarrow \boldsymbol{a}_k^T \boldsymbol{a}_k = 1 \Longrightarrow A^{-1} = A^T$

Eigendecomposition

See https://www.scss.tcd.ie/Rozenn.Dahyot/CS1BA1/SolutionEigen.pdf

PCA steps

- Data standardization: $X = \text{standardization}(X)$
- Covariance matrix calculation: $\boldsymbol{S} = Cov(X)$
- Finding eigenvectors and eigenvalues
- Eigenvectors ordering (descending) w.r.t eigenvalues
- Project X w.r.t the eigenvectors.

SVD: Singular Value Decomposition

SVD: Singular Value Decomposition

- Given that any $a \times b$ matrix X can be uniquely expressed as:

$$X_{a\times b} = U_{a\times a}\Sigma_{a\times b}V_{b\times b}^T$$

- Where:
  - U is a unity matrix
  - Σ is the diagonal matrix of singular values.
- Assuming that X is centred, i.e. zero mean:

$$Cov(X) = \frac{1}{N-1}X^TX$$

$$\frac{1}{N-1}V\Sigma U^TU\Sigma V^T$$

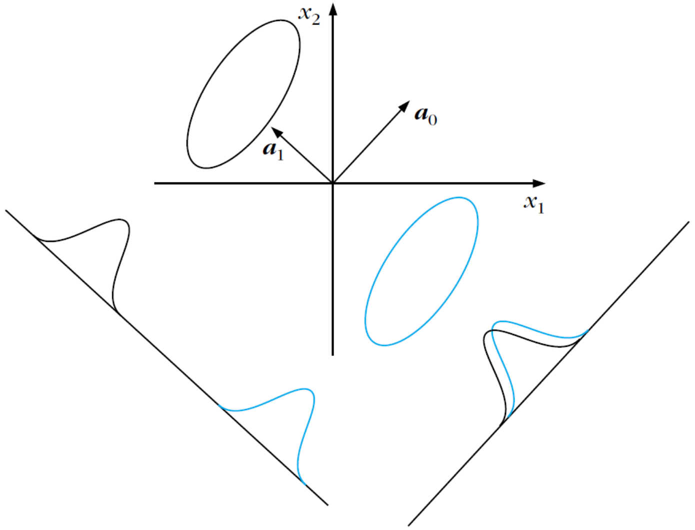
$$= V\frac{\Sigma^2}{N-1}V^T$$

SVD: Singular Value Decomposition

- Given that
  - $Cov(X) = \boldsymbol{S} = A\Psi A^T$
  - $Cov(X) = V\frac{\Sigma^2}{N-1}V^T$
- This means that V are principal vectors (eigenvectors) and the diagonal of $\frac{\Sigma^2}{N-1}$ are eigenvalues
- Very important: This is correct only when X is centred.

Q&A

- Obviously, when we apply PCA on the training data, we need to use the same eigenvectors to transform the unseen data.
- Is PCA always helpful?
  - No, the eigenvector with the largest eigenvalue might make the two classes coincide. The classes might be better separated by another eigenvector.



Source: Theodoridis, S., & Koutroumbas, K. "Pattern recognition. " Fourth Edition, 9781597492720, 2008

Summary

Summary

- Data dimension reduction
  - PCA
  - SVD

Thank you!

Zeyd Boukhers

E-mail: Boukhers@uni-koblenz.de
Phone: +49 (0) 261 287-2765
Web: Zeyd.Boukhers.com

University of Koblenz-Landau
Universitätsstr. 1
56070 Koblenz