Machine Learning and Intelligent Systems

Principal Components Analysis

Maria A. Zuluaga

Jan 26, 2023

EURECOM - Data Science Department

Table of contents

Preliminaries

Unsupervised Learning: Latent Variables

- **Latent Variable:** Variables that can only be inferred indirectly through a model from other observable variables that can be directly observed or measured.

Unsupervised Learning: Latent Variables

- **Latent Variable:** Variables that can only be inferred indirectly through a model from other observable variables that can be directly observed or measured.
- Clustering techniques allowed us to find a *discrete* latent variable z.
- How? We wanted to cluster our data into different groups, where z told us what group a data point $\mathbf{x}$ should go in.

Unsupervised Learning: Latent Variables

- **Latent Variable:** Variables that can only be inferred indirectly through a model from other observable variables that can be directly observed or measured.
- Clustering techniques allowed us to find a *discrete* latent variable z.
- How? We wanted to cluster our data into different groups, where z told us what group a data point $\mathbf{x}$ should go in.

In this slide deck:

- We switch to a *continuous* latent variable z

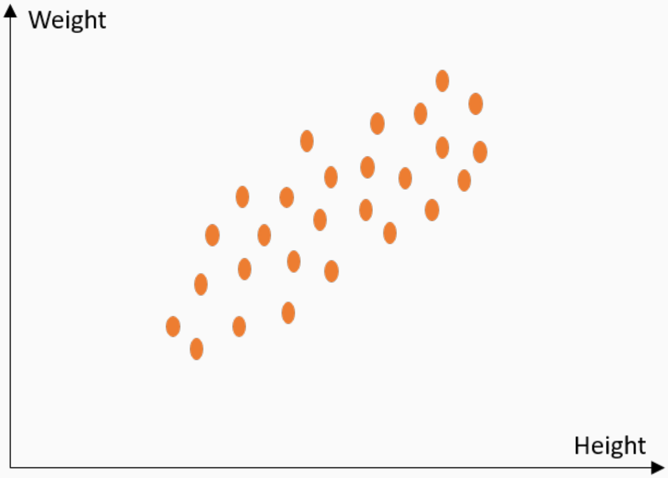
Unsupervised Learning: Latent Variables

- **Latent Variable:** Variables that can only be inferred indirectly through a model from other observable variables that can be directly observed or measured.
- Clustering techniques allowed us to find a *discrete* latent variable z.
- How? We wanted to cluster our data into different groups, where z told us what group a data point $\mathbf{x}$ should go in.

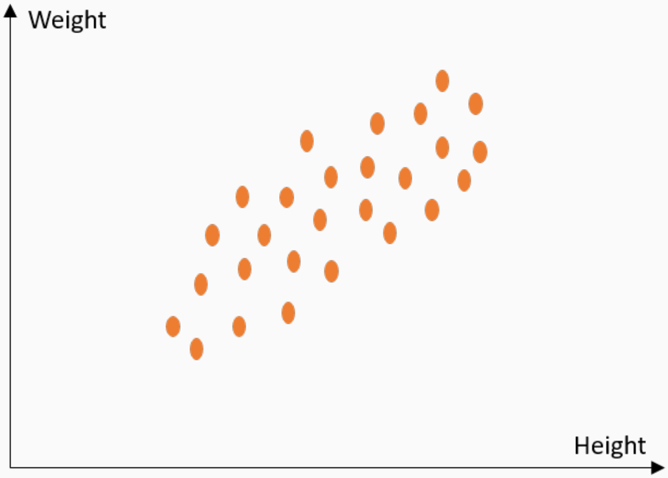
In this slide deck:

- We switch to a *continuous* latent variable z
- **Global idea:** Instead of grouping things into k discrete clusters, we try to summarize the data into k continuous dimensions

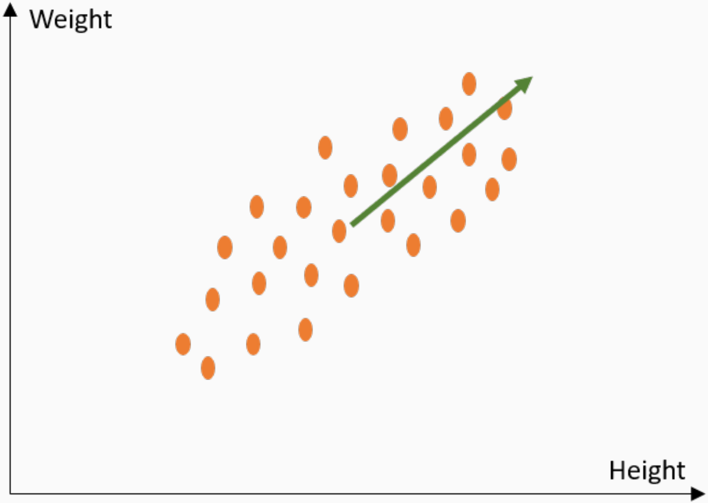
Intuition: An example



Intuition: An example



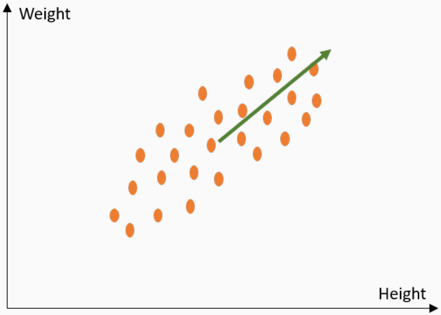
Can we find a vector that approximates this 2D space?



Yes. How can we identify this hidden axis?

Principal Component Analysis

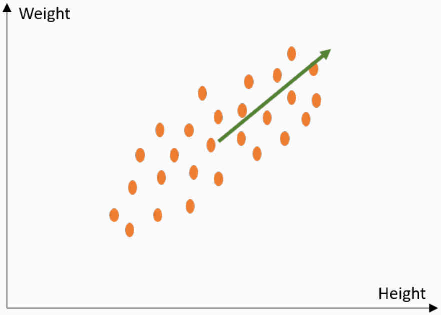
Intuition: First principal component



The green line can be more formally denoted as the **first principal component** (PC1).

PC1 is the line in the D-dimensional variable space ($D = 2$) that best approximates the data in the least squares sense.

Intuition: First principal component

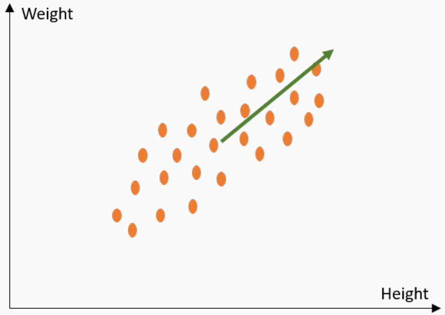


The green line can be more formally denoted as the **first principal component** (PC1).

PC1 is the line in the D-dimensional variable space ($D = 2$) that best approximates the data in the least squares sense.

Question: What does it mean to best approximate the data in the least squares sense?

Intuition: First principal component

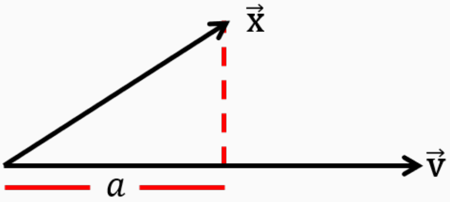


The green line can be more formally denoted as the **first principal component** (PC1).

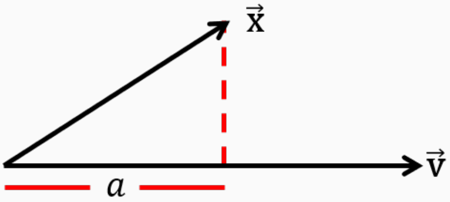
PC1 is the line in the D-dimensional variable space ($D = 2$) that best approximates the data in the least squares sense.

Question: What does it mean to best approximate the data in the least squares sense?

Each observation may now be *projected* onto this line in order to get a coordinate value along the PC-line



Let us recap the concept of projection already seen in the course (perceptron, SVM)



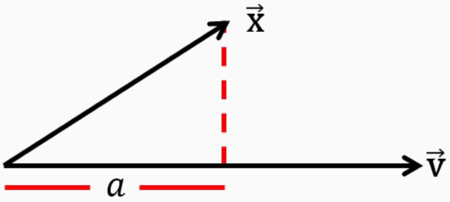
Let us recap the concept of projection already seen in the course (perceptron, SVM)

Length of projection of $\vec{x}$ onto $\vec{v}$:

$$a = \vec{\mathbf{v}}^T \vec{\mathbf{x}}$$

if $\|\mathbf{x}\|_2 = 1$ and $\|\mathbf{v}\|_2 = 1$

Linear Algebra: Projections



Let us recap the concept of projection already seen in the course (perceptron, SVM)

Length of projection of $\vec{x}$ onto $\vec{v}$:

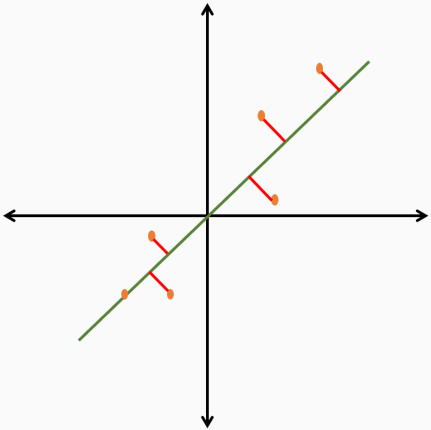
$$a = \vec{\mathbf{v}}^T \vec{\mathbf{x}}$$

if $\|\mathbf{x}\|_2 = 1$ and $\|\mathbf{v}\|_2 = 1$

Vector representing that projection:

$$a\vec{\mathbf{v}} = (\vec{\mathbf{v}}^T \vec{\mathbf{x}})\vec{\mathbf{v}}$$

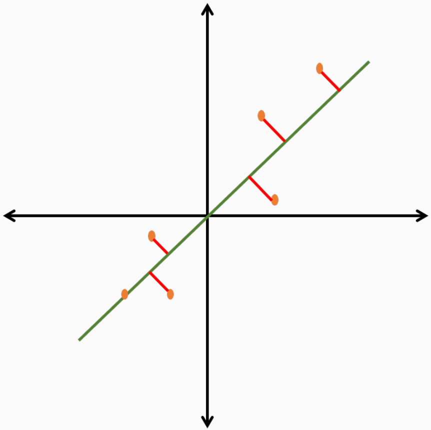
Minimizing the Reconstruction Error



Let us denote $\hat{\mathbf{v}}$ the vector that we want to find, i.e. PC1.

Approximating the data in the least squares sense, accounts to minimizing the difference between a given $\mathbf{x}$ and its approximation (projection).

Minimizing the Reconstruction Error



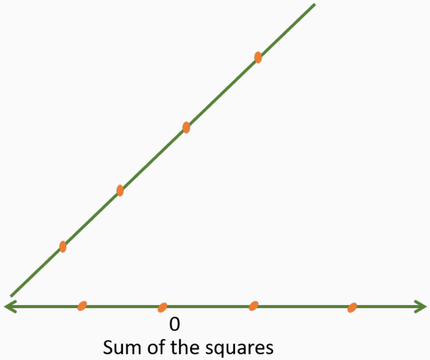
Let us denote $\hat{\mathbf{v}}$ the vector that we want to find, i.e. PC1.

Approximating the data in the least squares sense, accounts to minimizing the difference between a given $\mathbf{x}$ and its approximation (projection).

$$\begin{aligned}\hat{\mathbf{v}} &= \arg\min_{\vec{\mathbf{v}}} \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{x}_i - (\text{projection of } \mathbf{x}_i)\|_2^2 \\ &= \arg\min_{\vec{\mathbf{v}}} \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{x}_i - (\vec{\mathbf{v}}^T\mathbf{x}_i)\vec{\mathbf{v}}\|_2^2\end{aligned}$$

with $\|\mathbf{v}\|_2 = 1$

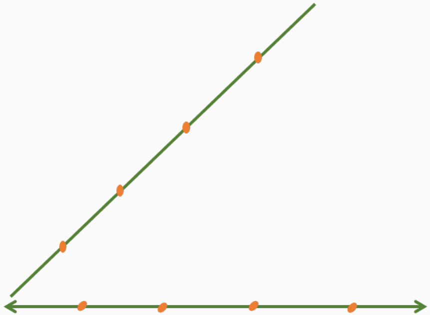
An alternative view: Maximizing the Variance



We can alternatively think about the problem as variance preservation

We want the vectors that capture the most variance in the data $\mathbf{x}$.

An alternative view: Maximizing the Variance



We can alternatively think about the problem as variance preservation

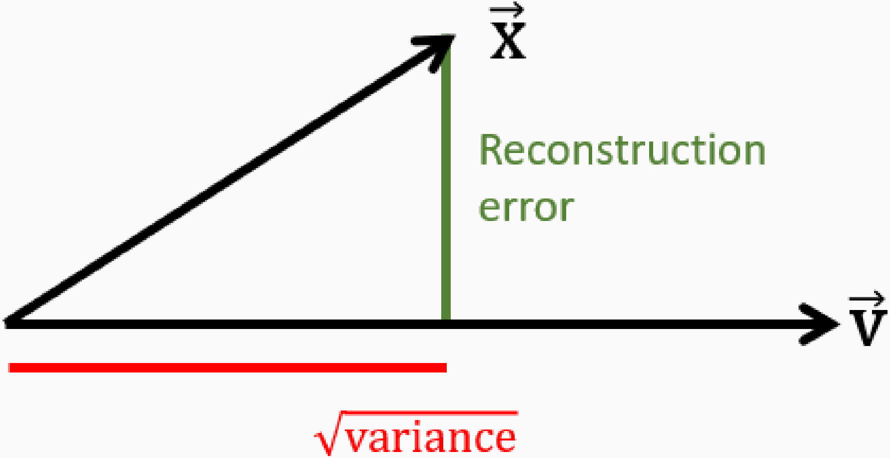
We want the vectors that capture the most variance in the data $\mathbf{x}$.

$$\hat{\mathbf{v}} = \arg\max_{\vec{\mathbf{v}}} \frac{1}{N} \sum_{i=1}^{N} (\text{projection length of } \mathbf{x}_i))^2$$

$$= \arg\max_{\vec{\mathbf{v}}} \frac{1}{N} \sum_{i=1}^{N} (\vec{\mathbf{v}}^T \mathbf{x}_i)^2$$

with $\|\mathbf{v}\|_2 = 1$

Equivalence: Reconstruction error vs. Variance



By Pythagoras theorem, minimizing the green is equivalent to maximizing the red

Principal Components Analysis

Choosing a subspace to maximize the projected variance, or minimize the reconstruction error, is called **principal component analysis** (PCA)

Principal Components Analysis

Choosing a subspace to maximize the projected variance, or minimize the reconstruction error, is called **principal component analysis** (PCA)

Theorem:
The vector that **maximizes the variance** is the **eigenvector** of $\boldsymbol{\Sigma}$, the sample covariance matrix, with **largest eigenvalue**

Linear Algebra: Quick refresher

Reminder 1:
$\vec{\mathbf{v}}$ is an **eigenvector** of $\boldsymbol{\Sigma}$ if $\boldsymbol{\Sigma}\vec{\mathbf{v}} = \lambda\vec{\mathbf{v}}$ for some **eigenvalue** $\lambda \in \mathbb{R}$.

Reminder 2:
The **eigenvectors** of a symmetric matrix are orthogonal to each other

Reminder 3:
Covariance matrices ($\boldsymbol{\Sigma}$) are symmetric and positive semidefinite.

Principal Component Analysis

The optimal PCA subspace is spanned by the top $K \ll D$ eigenvectors of $\mathbf{\Sigma}$.

More precisely, choose the first K of any orthonormal eigenbasis for $\mathbf{\Sigma}$.

Projection:

$$U_i = \begin{bmatrix} \vec{\mathbf{v}}_1^T \mathbf{x}_i \\ \vec{\mathbf{v}}_2^T \mathbf{x}_i \\ \cdots \\ \vec{\mathbf{v}}_K^T \mathbf{x}_i \end{bmatrix}$$

$\vec{\mathbf{v}}_1$ is the eigenvector of $\mathbf{\Sigma}$ with largest eigenvalue
$\vec{\mathbf{v}}_2$ is the eigenvector of $\mathbf{\Sigma}$ with the 2nd largest eigenvalue
$\vec{\mathbf{v}}_K$ is the eigenvector of $\mathbf{\Sigma}$ with Kth largest eigenvalue

Collectively, we obtain a set of vectors $\vec{\mathbf{v}}_1, \ldots, \vec{\mathbf{v}}_K$ that minimize the reconstruction error and that are orthogonal to each other

Computing the PCA eigenvectors and eigenvalues

- Covariance method
    - Simplest way of doing PCA
    - Based on eigenvectors interpretation
    - It can be very slow for high dimensions

- Singular Value Decomposition (SVD)
    - Faster for high dimensions
    - Numerically stable
    - Truncated SVD allows to compute only top PCs, making it even faster

Generally, SVD is the preferred method

Covariance method: HOWTO

1. Center the data, i.e. $\mathbf{x}_i - \boldsymbol{\mu} \quad \forall i$, where $\boldsymbol{\mu}$ denotes the mean
2. It is also a good idea to have unit variance along each feature dimension

Covariance method: HOWTO

1. Center the data, i.e. $\mathbf{x}_i - \boldsymbol{\mu} \quad \forall i$, where $\boldsymbol{\mu}$ denotes the mean
2. It is also a good idea to have unit variance along each feature dimension
3. Collect all the data into an $N \times D$ matrix $\mathbf{X}$

Covariance method: HOWTO

1. Center the data, i.e. $\mathbf{x}_i - \boldsymbol{\mu} \quad \forall i$, where $\boldsymbol{\mu}$ denotes the mean
2. It is also a good idea to have unit variance along each feature dimension
3. Collect all the data into an $N \times D$ matrix $\mathbf{X}$
4. Estimate the covariance matrix

$$\boldsymbol{\Sigma} = \frac{1}{N-1}\mathbf{X}^T\mathbf{X}$$

Covariance method: HOWTO

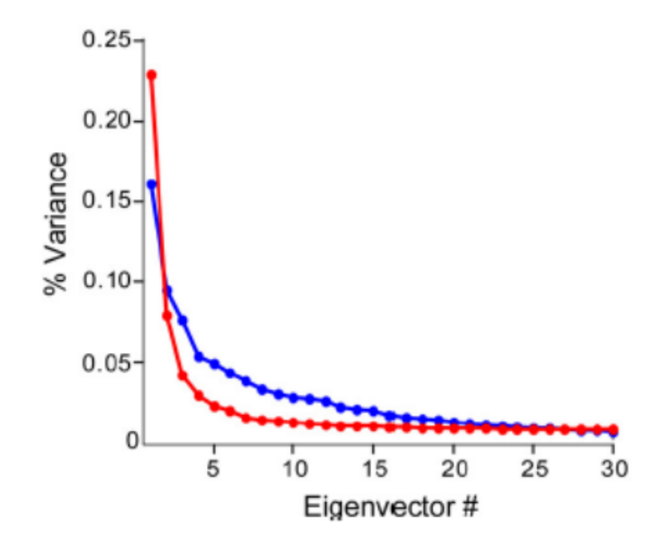
1. Center the data, i.e. $\mathbf{x}_i - \boldsymbol{\mu} \quad \forall i$, where $\boldsymbol{\mu}$ denotes the mean
2. It is also a good idea to have unit variance along each feature dimension
3. Collect all the data into an $N \times D$ matrix $\mathbf{X}$
4. Estimate the covariance matrix
$$\mathbf{\Sigma} = \frac{1}{N-1}\mathbf{X}^T\mathbf{X}$$
5. Compute the eigendecomposition $\mathbf{\Sigma} = \mathbf{V}\Lambda\mathbf{V}^T$

Covariance method: HOWTO

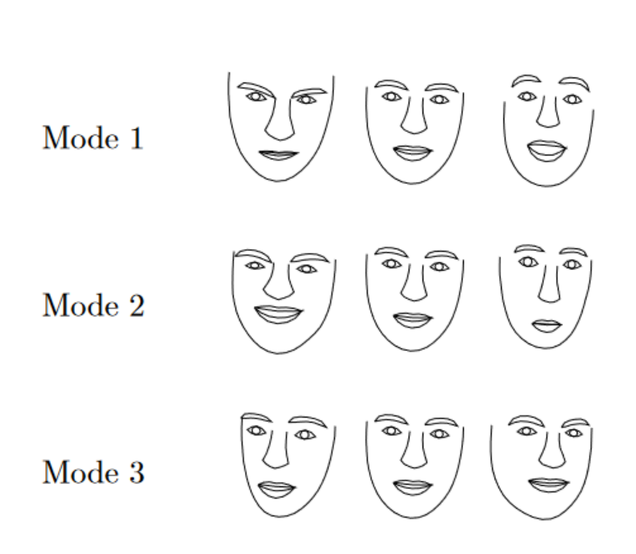
1. Center the data, i.e. $\mathbf{x}_i - \boldsymbol{\mu} \quad \forall i$, where $\boldsymbol{\mu}$ denotes the mean
2. It is also a good idea to have unit variance along each feature dimension
3. Collect all the data into an $N \times D$ matrix $\mathbf{X}$
4. Estimate the covariance matrix
$$\boldsymbol{\Sigma} = \frac{1}{N-1}\mathbf{X}^T\mathbf{X}$$
5. Compute the eigendecomposition $\boldsymbol{\Sigma} = \mathbf{V}\Lambda\mathbf{V}^T$

- $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_D)$ where $(\lambda_1 \geq \cdots \geq \lambda_D)$ are the eigenvalues of $\boldsymbol{\Sigma}$.
- $\mathbf{V}$ is orthogonal and its k^{th} column is the k^{th} eigenvector of $\boldsymbol{\Sigma}$

Variance vs. PCs



Application Example: Facial recognition



Source: T. Cootes, An Introduction to Active Shape Models

Wrap-up

Wrap-up

- We introduced principal component analysis, a technique for dimensionality reduction
- We reviewed the intuition behind the concept under two perspectives: minimizing the reconstruction error and maximizing the variance
- We covered the covariance method to obtain the principal components

Key Concepts

- Eigenvalues
- Eigenvectors
- Covariance matrix
- Singular value decomposition

References

Further Reading and Useful Material

Source	Notes
The Elements of Statistical Learning	Section 14.5