# Knowledge Discovery and Data Mining 1 (VO) (707.003)

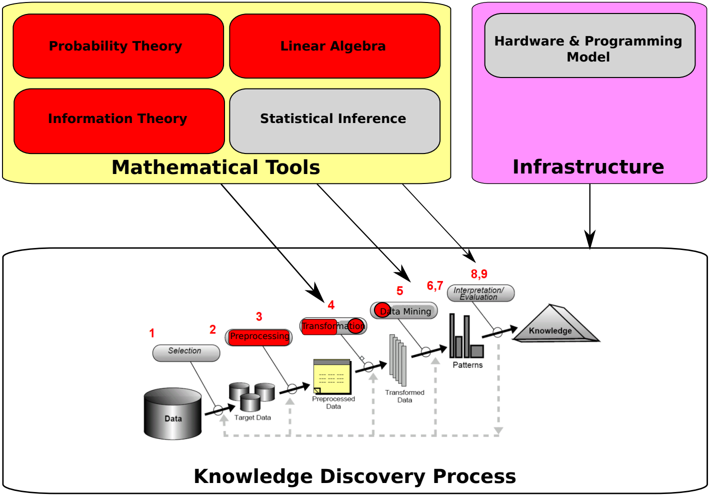
## Principal Component Analysis

Denis Helic

KTI, TU Graz

Nov 13, 2014

# Big picture: KDDM



Probability Theory

Linear Algebra

Information Theory

Statistical Inference

Mathematical Tools

Hardware & Programming Model

Infrastructure

Knowledge Discovery Process

## Outline

1. Introduction
2. Eigenvalues and Eigenvectors
3. Dimensionality Reduction
4. Principal-Component Analysis
5. Advanced: Power Method
6. Advanced: Maximizing the Variance in PCA

# Recap

## Recap

Review of data matrices

# Recap – Representing data

- **Given:** Preprocessed data objects as a set of features
- E.g. for text documents set of words, bigrams, n-grams, . . .
- **Given:** Feature statistics for each data object
- E.g. number of occurrences, magnitudes, ticks, . . .
- **Find:** Mathematical model for calculations
- E.g. similarity, distance, add, subtract, transform, . . .

# Recap – Representing data

- Now, an intuitive representation of the data is a matrix
- In a general case
- Columns correspond to features, i.e. dimensions or coordinates in an $m$-dimensional space
- Rows correspond to data objects, i.e. data points
- An element $d_{ij}$ in the $i$-th row and the $j$-th column is the $j$-th coordinate of the $i$-th data point
- The data is represented by a matrix $\mathbf{D} \in \mathbb{R}^{n \times m}$, where $n$ is the number of data points and $m$ the number of features

# Recap – The Document-Term Matrix

- For text
- Columns correspond to terms, words, and so on
- I.e. each word is a dimension or a coordinate in an *m*-dimensional space
- Rows correspond to documents
- An element $d_{ij}$ in the $i$-th row and the $j$-th column is the e.g. number of occurrences of the $j$-th word in the $i$-th document
- This matrix is called **document-term** matrix

# Recap – The Utility Matrix

- In a recommender system there are two classes of entities: users and items
- Users have preferences for certain items and we have to mine for these preferences
- The data is represented by a utility matrix $\mathbf{M} \in \mathbb{R}^{n \times m}$, where $n$ is the number of users and $m$ the number of items
- The matrix gives a value for each user-item pair what is known about the preference of that user for that item
- E.g. the values can come from an ordered set (1 to 5) and represent a rating that a user gave for an item

# Eigenvalues and eigenvectors

## Eigenvalues and eigenvectors

Given a square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, we say that $\lambda \in \mathbb{C}$ is an eigenvalue of $\mathbf{A}$ and $\mathbf{x} \in \mathbb{C}^n$ is the corresponding eigenvector if

$$\mathbf{A}\mathbf{x} = \lambda \mathbf{x}, \mathbf{x} \neq \mathbf{0}$$

# Interpretation of eigenvalues and eigenvectors

## Example

Suppose we have a Web-based question-answer system, where standard users (S) post questions about certain topics expert (E) and other users answer those questions. For example, Stackoverflow is an example of such a system. The questions are about programming, software development, and so on. We observe the following developments with our system:

- The system is very successful and we observe increasing numbers of both standard (S) as well as expert users (E). Each month we have 4x more (S) users and 2x more (E) users.
- The numbers of these two different types of users seem to be *decoupled* because we can not observe that they affect each other.

# Interpretation of eigenvalues and eigenvectors

## Example

This *decoupling* may arise in the following situation:

- The standard users (S) talk only to each other: "Hey, there is this great system where you get all the answers from experts very quickly".
- The expert users (E) also talk only to each other: "Hey, there is this great system where you can show what you know and become popular very quickly".
- (S) and (E) do not talk to each other and do not influence each other.

## Note

Such a system is completely hypothetical and is not realistic, but illustrates very well the educational point.

# Interpretation of eigenvalues and eigenvectors

## Question

How will our system develop? How many standard users (S) we will have in e.g. two, three, .. years. How many expert users (E), and hence how many users in total.

## Model

Let us denote the number of standard users (S) with $x$ and the number of expert users (E) with $y$, or in matrix notation:

$$\mathbf{u} = \begin{pmatrix} x \\ y \end{pmatrix}$$

# Interpretation of eigenvalues and eigenvectors

### Model

We know from the data that $x$ increases four times each month and $y$ increases two times each month:

$$\begin{aligned} x(t+1) &= 4x(t) \\ y(t+1) &= 2y(t), \end{aligned}$$

where $x(t)$ is the number of e.g. standard users in month $t$ and $x(t+1)$ is the number of standard users next month.

# Interpretation of eigenvalues and eigenvectors

- How does the system evolve?

$$\begin{aligned} x(1) &= 4 \cdot x(0) \\ y(1) &= 2 \cdot y(0) \end{aligned}$$

$$\begin{aligned} x(2) &= 4 \cdot x(1) = 4 \cdot 4 \cdot x(0) = 4^2 \cdot x(0) \\ y(2) &= 2 \cdot y(1) = 2 \cdot 2 \cdot y(0) = 2^2 \cdot y(0) \end{aligned}$$

# Interpretation of eigenvalues and eigenvectors

$$\begin{aligned} x(3) &= 4 \cdot x(2) = 4 \cdot 4^2 \cdot x(0) = 4^3 \cdot x(0) \\ y(3) &= 2 \cdot y(2) = 2 \cdot 2^2 \cdot y(0) = 2^3 \cdot y(0) \end{aligned}$$

$$\begin{aligned} x(t+1) &= 4 \cdot x(t) = 4 \cdot 4^t \cdot x(0) = 4^{t+1} \cdot x(0) \\ y(t+1) &= 2 \cdot y(t) = 2 \cdot 2^t \cdot y(0) = 2^{t+1} \cdot y(0) \end{aligned}$$

# Interpretation of eigenvalues and eigenvectors

- How does the system evolve?
- Suppose we have the following initial conditions: we have a single standard user (S) and a single expert user (E) in the beginning.
- How the numbers of standard and expert users compare after e.g. five years?

$$\begin{aligned} x(0) &= 1 \\ y(0) &= 1 \end{aligned}$$

$$\begin{aligned} x(60) &= 4^{60} \cdot x(0) = 4^{60} \\ y(60) &= 2^{60} \cdot y(0) = 2^{60} \end{aligned}$$

# Interpretation of eigenvalues and eigenvectors

- The ratio: $\frac{x}{y}$

$$\frac{x^{60}}{y^{60}} = \frac{4^{60}}{2^{60}} = 2^{60}$$

- $2^{60}$ is a huge number, i.e. standard users (S) completely dominate the expert users (E)
- The total number of users is approx. equal to the number of standard users (S)

# Interpretation of eigenvalues and eigenvectors

- But, how is this related to the eigenvalues and eigenvectors?

### Model

We can express the previous equations in the matrix form:

$$\mathbf{u}(t+1) = \mathbf{A}\mathbf{u}(t)$$

# Interpretation of eigenvalues and eigenvectors

$$\mathbf{A} = \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix}$$

$$\mathbf{u} = \begin{pmatrix} x \\ y \end{pmatrix}$$

# Interpretation of eigenvalues and eigenvectors

$$\mathbf{A} = \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix}$$

$$det(\lambda \mathbf{I} - \mathbf{A}) \quad = \quad det(\begin{pmatrix} \lambda - 4 & 0 \\ 0 & \lambda - 2 \end{pmatrix}) = (\lambda - 4)(\lambda - 2)$$

- Thus, $\lambda_1 = 4$, and $\lambda_2 = 2$ are eigenvalues of $\mathbf{A}$
- **The eigenvalues of a diagonal matrix are equal to its diagonal entries**
- We now solve $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ for each eigenvalue to find the corresponding eigenvectors

# Interpretation of eigenvalues and eigenvectors

- For $\lambda_1 = 4$

$$\begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = 4 \begin{pmatrix} x \\ y \end{pmatrix}$$

$$\begin{aligned} 4x &= 4x \\ y &= 4y \end{aligned}$$

- Thus, $y = 0$ and we might pick $x = 1$, i.e. $\mathbf{v}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$

# Interpretation of eigenvalues and eigenvectors

- For $\lambda_1 = 2$

$$\begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = 2 \begin{pmatrix} x \\ y \end{pmatrix}$$

$$\begin{aligned} 4x &= x \\ y &= y \end{aligned}$$

- Thus, $x = 0$ and we might pick $y = 1$, i.e. $\mathbf{v}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$
- **The eigenvectors of a diagonal matrix form a standard basis for a Euclidean space**

# Interpretation of eigenvalues and eigenvectors

- But, how is this related to the eigenvalues and eigenvectors?

### Model

We can express the previous equations in the matrix form:

$$\mathbf{u}(t+1) = \mathbf{A}\mathbf{u}(t)$$

# Interpretation of eigenvalues and eigenvectors

- How does the system evolve (in matrix form)?

$$\mathbf{u}(1) = \mathbf{A}\mathbf{u}(0)$$

$$\mathbf{u}(2) = \mathbf{A}\mathbf{u}(1) = \mathbf{A}\mathbf{A}\mathbf{u}(0) = \mathbf{A}^2\mathbf{u}(0)$$

$$\mathbf{u}(3) = \mathbf{A}\mathbf{u}(2) = \mathbf{A}\mathbf{A}^2\mathbf{u}(0) = \mathbf{A}^3\mathbf{u}(0)$$

$$\mathbf{u}(t+1) = \mathbf{A}\mathbf{u}(t) = \mathbf{A}\mathbf{A}^t\mathbf{u}(0) = \mathbf{A}^{t+1}\mathbf{u}(0)$$

# Interpretation of eigenvalues and eigenvectors

- Since the eigenvectors form a basis for the space
- We can write $\mathbf{u}(0)$ as a linear combination of the eigenvectors $\mathbf{v}_i$ of the matrix (for appropriate choice of constants $c_i$)
- $\mathbf{v}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$
- $\mathbf{v}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$

$$\mathbf{u}(0) = c_1\mathbf{v}_1 + c_2\mathbf{v}_1$$

# Interpretation of eigenvalues and eigenvectors

- We have the following initial conditions: we have a single standard user (S) and a single expert user (E) in the beginning.

$$\mathbf{u}(0) = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

$$\mathbf{u}(0) = \mathbf{v}_1 + \mathbf{v}_2$$

# Interpretation of eigenvalues and eigenvectors

- We know from before:

$$\mathbf{u}(t+1) = \mathbf{A}^{t+1}\mathbf{u}(0)$$

$$\mathbf{u}(t+1) = \mathbf{A}^{t+1}(\mathbf{v}_1 + \mathbf{v}_2) = \mathbf{A}^{t+1}\mathbf{v}_1 + \mathbf{A}^{t+1}\mathbf{v}_2$$

# Interpretation of eigenvalues and eigenvectors

- We have: $\mathbf{A}\mathbf{v}_1 = \lambda_1 \mathbf{v}_1$ and $\mathbf{A}\mathbf{v}_2 = \lambda_1 \mathbf{v}_2$
- By substituting:

$$\mathbf{u}(t+1) = \mathbf{A}^{t+1}\mathbf{v}_1 + \mathbf{A}^{t+1}\mathbf{v}_2 = \lambda_1^{t+1}\mathbf{v}_1 + \lambda_2^{t+1}\mathbf{v}_2$$

$$\mathbf{u}(t+1) = 4^{t+1}\begin{pmatrix}1\\0\end{pmatrix} + 2^{t+1}\begin{pmatrix}0\\1\end{pmatrix}$$

# Interpretation of eigenvalues and eigenvectors

- After five years:

$$\mathbf{u}(60) = 4^{60}\begin{pmatrix}1\\0\end{pmatrix} + 2^{60}\begin{pmatrix}0\\1\end{pmatrix}$$

- $4^{60}$ is a much larger number than $2^{60}$
- The first term dominates the second
- Standard users (S) completely dominate the expert users (E)

# Interpretation of eigenvalues and eigenvectors

- The system evolves along the directions of the matrix eigenvectors
- The eigenvalues determine the speed of the development
- Larger eigenvalues represent quicker development and dominant behavior
- In the long run the system develops in the direction of the eigenvector corresponding to the largest eigenvalue
- This is the leading eigenvector and leading eigenvalue

# Interpretation of eigenvalues and eigenvectors

- Another example: a *coupled* system (more realistic)
- The number of standard users (S) is influenced also by the number of expert users (E)
- I.e. a famous expert joins the system and as a consequence a lot of standard users joins as well
- Also, the number of expert users is influenced by the number of standard users
- Experts like to have an audience and are therefore attracted to the system

# Interpretation of eigenvalues and eigenvectors

## Model

Let say that we know from the data how $x$ and $y$ increase each month. For example:

$$\begin{aligned} x(t+1) &= 3x(t) + 6y(t) \\ y(t+1) &= x(t) + 4y(t), \end{aligned}$$

where $x(t)$ is the number of e.g. standard users in month $t$ and $x(t+1)$ is the number of standard users next month.

# Interpretation of eigenvalues and eigenvectors

- Again, in matrix form:

$$\mathbf{u}(t+1) = \mathbf{A}\mathbf{u}(t)$$

$$\mathbf{A} = \begin{pmatrix} 3 & 6 \\ 1 & 4 \end{pmatrix}$$

$$\mathbf{u} = \begin{pmatrix} x \\ y \end{pmatrix}$$

# Interpretation of eigenvalues and eigenvectors

$$\mathbf{A} = \begin{pmatrix} 3 & 6 \\ 1 & 4 \end{pmatrix}$$

$$\begin{aligned} det(\lambda \mathbf{I} - \mathbf{A}) &= det(\begin{pmatrix} \lambda - 3 & -6 \\ -1 & \lambda - 4 \end{pmatrix}) = (\lambda - 3)(\lambda - 4) - (-1)(-6) \\ &= \lambda^2 - 3\lambda - 4\lambda + 12 - 6 = \lambda^2 - 7\lambda + 6 \\ &= (\lambda - 6)(\lambda - 1) \end{aligned}$$

- Thus, $\lambda_1 = 6$, and $\lambda_2 = 1$ are eigenvalues of $\mathbf{A}$
- We now solve $\mathbf{Ax} = \lambda \mathbf{x}$ for each eigenvalue to find the corresponding eigenvectors

# Interpretation of eigenvalues and eigenvectors

- For $\lambda_1 = 6$

$$\begin{pmatrix} 3 & 6 \\ 1 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = 6 \begin{pmatrix} x \\ y \end{pmatrix}$$

$$\begin{aligned} 3x + 6y &= 6x \implies 3x = 6y \implies x = 2y \\ x + 4y &= 6y \implies x = 2y \end{aligned}$$

- Thus, we might pick $y = 1$ and then $x = 2$, i.e. $\mathbf{v}_1 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$

## Interpretation of eigenvalues and eigenvectors

- For $\lambda_1 = 1$

$$\begin{pmatrix} 3 & 6 \\ 1 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix}$$

$$\begin{aligned} 3x + 6y &= x \implies 2x = -6y \implies x = -3y \\ x + 4y &= y \implies x = -3y \end{aligned}$$

- Thus, we might pick $y = 1$ and then $x = -3$, i.e. $\mathbf{v}_2 = \begin{pmatrix} -3 \\ 1 \end{pmatrix}$

# Interpretation of eigenvalues and eigenvectors

- We know that he system evolves along the directions of the matrix eigenvectors
- The eigenvalues determine the speed of the development
- Larger eigenvalues represent quicker development and dominant behavior
- In the long run the system develops in the direction of the eigenvector corresponding to the largest eigenvalue
- This is the leading eigenvector and leading eigenvalue

# Interpretation of eigenvalues and eigenvectors

- We can write $\mathbf{u}(0)$ as a linear combination of the eigenvectors $\mathbf{v}_i$ of the matrix (for appropriate choice of constants $c_i$)
- $\mathbf{v}_1 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$
- $\mathbf{v}_2 = \begin{pmatrix} -3 \\ 1 \end{pmatrix}$

$$\mathbf{u}(0) = c_1\mathbf{v}_1 + c_2\mathbf{v}_1$$

# Interpretation of eigenvalues and eigenvectors

- We have the following initial conditions: we have a single standard user (S) and a single expert user (E) in the beginning.

$$\mathbf{u}(0) = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \frac{4}{5}\begin{pmatrix} 2 \\ 1 \end{pmatrix} + \frac{1}{5}\begin{pmatrix} -3 \\ 1 \end{pmatrix}$$

$$\mathbf{u}(0) = \frac{4}{5}\mathbf{v}_1 + \frac{1}{5}\mathbf{v}_2$$

# Interpretation of eigenvalues and eigenvectors

- As before we have:

$$\mathbf{u}(t+1) = \frac{4}{5}\mathbf{A}^{t+1}\mathbf{v}_1 + \frac{1}{5}\mathbf{A}^{t+1}\mathbf{v}_2 = \frac{4}{5}\lambda_1^{t+1}\mathbf{v}_1 + \frac{1}{5}\lambda_2^{t+1}\mathbf{v}_2$$

$$\mathbf{u}(t+1) = \frac{4}{5}6^{t+1}\begin{pmatrix}2\\1\end{pmatrix} + \frac{1}{5}\begin{pmatrix}-3\\1\end{pmatrix}$$

# Interpretation of eigenvalues and eigenvectors

- After five years:

$$\mathbf{u}(60) = \frac{4}{5}6^{60}\begin{pmatrix}2\\1\end{pmatrix} + \frac{1}{5}\begin{pmatrix}-3\\1\end{pmatrix}$$

- The first term dominates the second
- There are twice as many standard users (S) as expert users (E)

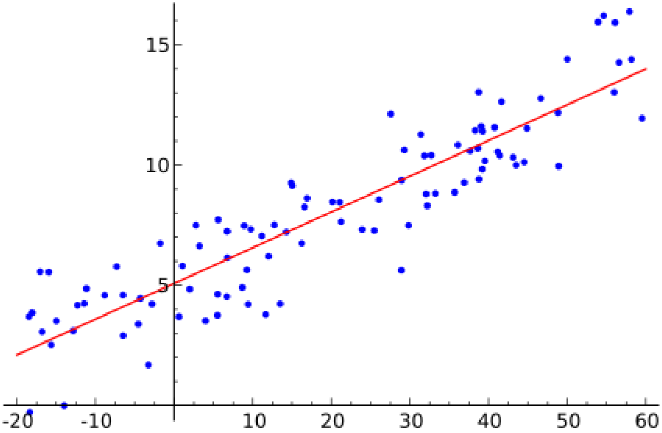
## The curse of dimensionality

- The data matrices are huge and have hundred thousands, even millions of rows and columns
- E.g. document-term matrix of Wikipedia: 10 millions rows and 100.000 columns
- Utility matrix at Amazon
- The number of customers times the number of articles
- The curse of dimensionality

# The curse of dimensionality

- In many cases we can summarize these matrices by finding narrower matrices that are in some sense close to the original
- These narrow matrices have small(er) numbers of rows and columns as compared to the original matrices
- We can use them more efficiently than the original matrices
- E.g. calculations are faster, we need less memory to store them, and so on.
- The process of finding those narrow matrices is called **dimensionality reduction**

# The curse of dimensionality



# Example

## Text documents

Suppose we have the following documents:

| DocID | Document |
|---:|---:|
| d1 | Iphone Iphone Iphone Apple |
| d2 | Google Google Google Apple |
| d3 | Apple Apple Apple Iphone Iphone Iphone Google |
| d4 | Google Google Google |
| d5 | Apple Iphone Apple |

# Representing data: Example

- Now, we take words as features
- We take word occurrences as the feature values
- Three features: *Apple*, *Iphone*, *Google*

| Doc \ Feature | *Apple* | *Iphone* | *Google* |
|---:|---:|---:|---:|
| d1 | 1 | 3 | 0 |
| d2 | 1 | 0 | 3 |
| d3 | 3 | 3 | 1 |
| d4 | 0 | 0 | 3 |
| d5 | 2 | 1 | 0 |

# Representing data: example

$$\mathbf{D} = \begin{pmatrix} 1 & 3 & 0 \\ 1 & 0 & 3 \\ 3 & 3 & 1 \\ 0 & 0 & 3 \\ 2 & 1 & 0 \end{pmatrix}$$

# Reducing dimensions

- Now, we have a matrix with three dimensions
- The question is how to reduce it to e.g. two or one dimension with a small loss of information
- One idea would be to find directions in which the data is aligned
- Depending on the "magnitude" (variation) of data along those lines we can keep the dominant directions
- This sounds very much like eigenvalues and eigenvectors

## Reducing dimensions

- Let start by investigating how data varies along the dimensions that we have
- E.g. how documents vary concerning the occurrences of “Apple”, “Iphone” and “Google”

$$var(\mathbf{f}) = \frac{1}{n}(\mathbf{f} - \overline{f}\mathbf{1})^T(\mathbf{f} - \overline{f}\mathbf{1}),$$

- where **f** is the column vector from the matrix **D**

# Reducing dimensions

- $\mathbf{f} - \bar{f}\mathbf{1}$ are column vectors moved to the center
- Lets call them $\mathbf{x}$
- Then, we have:

$$var(\mathbf{f}) = var(\mathbf{x}) = \frac{1}{n}\mathbf{x}^T\mathbf{x}$$

# Move the center

$$\mathbf{D} = \begin{pmatrix} 1 & 3 & 0 \\ 1 & 0 & 3 \\ 3 & 3 & 1 \\ 0 & 0 & 3 \\ 2 & 1 & 0 \end{pmatrix}$$

- $\overline{f_1} = \overline{f_2} = \overline{f_3} = 1.4$

# Move the center

$$\mathbf{X} = \begin{pmatrix} -0.4 & 1.6 & -1.4 \\ -0.4 & -1.4 & 1.6 \\ 1.6 & 1.6 & -0.4 \\ -1.4 & -1.4 & 1.6 \\ 0.6 & -0.4 & -1.4 \end{pmatrix}$$

# Variance along dimensions

$$var(\mathbf{x}_1) = 1.04$$

$$var(\mathbf{x}_2) = 1.84$$

$$var(\mathbf{x}_3) = 1.84$$

- The variance is smallest alongside the “Apple” dimension
- Why is that?

# Variance along dimensions

- Now a simple idea would be to throw away the “Apple” dimension
- It carries the smallest amount of information
- It appears in many documents
- Recollect also IDF

## Covariance between dimensions

- However, we can do even better
- Let us investigate how the data changes together alongside two dimensions
- How data co-varies
- That is the co-variance

$$cov(\mathbf{f}_i, \mathbf{f}_j) = cov(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{n}\mathbf{x}_i^T\mathbf{x}_j$$

# Covariance matrix

- We can even combine variance and covariance into the **covariance matrix** $\frac{1}{n}\mathbf{X}^T\mathbf{X}$
- It is an $m \times m$ matrix where each row and each column correspond to a feature, i.e. a dimension
- It connects features with each other

$$\frac{1}{n}\mathbf{X}^T\mathbf{X} = \begin{pmatrix} 1.04 & 0.84 & -0.76 \\ 0.84 & 1.84 & -1.36 \\ -0.76 & -1.36 & 1.84 \end{pmatrix}$$

# Covariance matrix

- It is a square symmetric matrix: eigenvalues are all real, eigenvectors are orthonormal
- The diagonal entries keep the variance along different dimensions
- The element in $i$-th row and $j$-th column keeps covariance between the dimensions $i$ and $j$

# Covariance matrix

- If this number is positive then the values either go up or go down together
- E.g. if a document mentions "Apple" many times then typically it also mentions "Iphone", and vice versa
- If the covariance is negative then whenever one value goes up the other goes down, or other way around
- E.g. if document mentions "Apple" then typically it does not mention "Google", and vice versa

# Eigenvalues and eigenvectors of the covariance matrix

- By analogy with our previous examples
- The eigenvectors of a system evolution matrix are directions along which the user numbers evolve
- The eigenvalues correspond to the speed of this development
- The eigenvectors of a covariance matrix give the directiosn along which the data varies
- The eigenvalues of a covariance matrix give the magnitude of the variance
- The leading eigenvalue and the leading eigevector represent the direction and the magnitude of the maximal variance

# PCA

- Principal-component analysis or PCA is a technique for transforming points from a high-dimensional space by finding the directions along which the points line up best
- The idea is to treat the data as a matrix moved to the center $\mathbf{X}$
- We then find the eigenvectors of the covariance matrix $\mathbf{X}^T\mathbf{X}$
- The matrix of these eigenvectors may be thought of as a rigid rotation in a high-dimensional space
- The axis corresponding to the principal eigenvector is the one with the maximal variance
- It carries most of the signal

# PCA

- The axis corresponding to the second eigenvector is the axis along which the variance of distances from the first axis is greatest and so on
- Thus, we can replace the original high-dimensional data by its projection onto the most important axes
- These axes are the ones corresponding to the largest eigenvalues
- Thus, the original data is approximated by data with fewer dimensions
- The new data summarizes the original data very good (according to a certain criteria)

# PCA: Example

$$\frac{1}{n}\mathbf{X}^T\mathbf{X} = \begin{pmatrix} 1.04 & 0.84 & -0.76 \\ 0.84 & 1.84 & -1.36 \\ -0.76 & -1.36 & 1.84 \end{pmatrix}$$

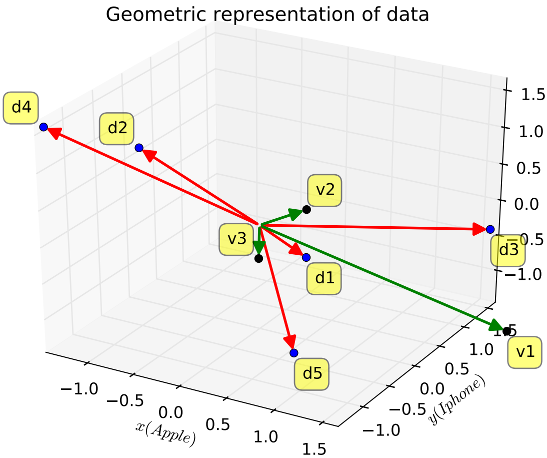
# PCA: Example

$$\lambda_1 = 3.68425,\ \mathbf{v}_1 = \begin{pmatrix} 0.39348 \\ 0.65496 \\ -0.64514 \end{pmatrix}$$

$$\lambda_2 = 0.58230,\ \mathbf{v}_2 = \begin{pmatrix} 0.818528 \\ 0.069895 \\ 0.570199 \end{pmatrix}$$

$$\lambda_3 = 0.45345,\ \mathbf{v}_3 = \begin{pmatrix} 0.41855 \\ -0.75242 \\ -0.50860 \end{pmatrix}$$

# PCA example


Geometric representation of data
d4
d2
v2
v3
d3
d1
d5
v1
x(Apple)
−1.0
−0.5
0.0
0.5
1.0
1.5
y(Iphone)
−1.0
−0.5
0.0
0.5
1.0
1.5
z(Google)
−1.0
−0.5
0.0
0.5
1.0
1.5

# PCA: example

- Now let us construct $\mathbf{E}$, which is the (orthogonal) matrix of eigenvectors for the matrix $\frac{1}{n}\mathbf{X}^T\mathbf{X}$

$$\mathbf{E} = \begin{pmatrix} 0.39348 & 0.818528 & 0.41855 \\ 0.65496 & 0.069895 & -0.75242 \\ -0.64514 & 0.570199 & -0.50860 \end{pmatrix}$$

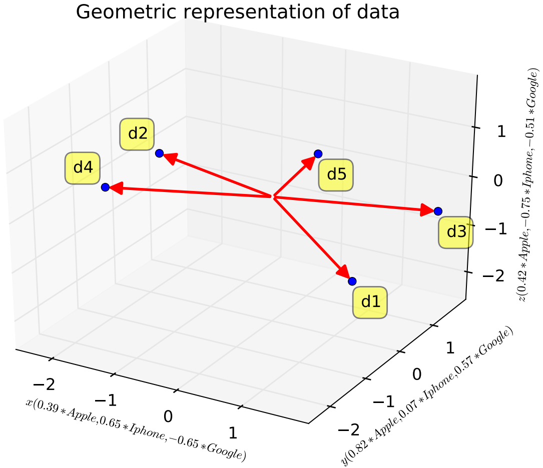
- Any orthogonal matrix represents a rotation of the axes of a Euclidean space
- Thus, multiplying the original matrix $\mathbf{X}$ with the $\mathbf{E}$ would project the points to the new axes

# PCA: example

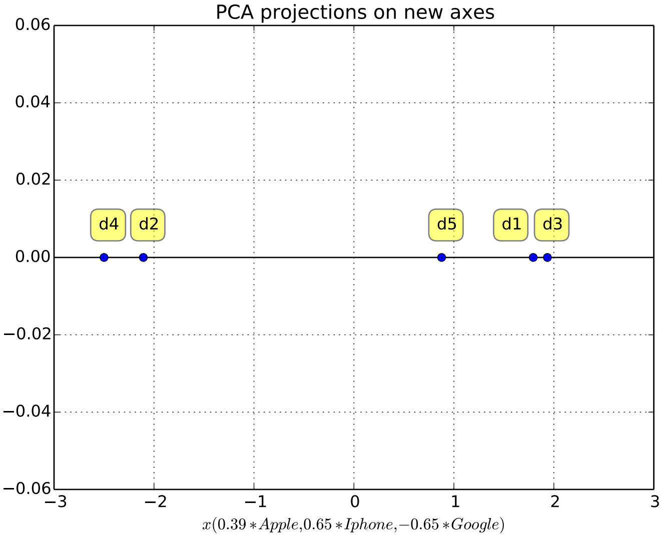
$$\mathbf{XE} = \begin{pmatrix} -0.4 & 1.6 & -1.4 \\ -0.4 & -1.4 & 1.6 \\ 1.6 & 1.6 & -0.4 \\ -1.4 & -1.4 & 1.6 \\ 0.6 & -0.4 & -1.4 \end{pmatrix} \begin{pmatrix} 0.39348 & 0.818528 & 0.41855 \\ 0.65496 & 0.069895 & -0.75242 \\ -0.64514 & 0.570199 & -0.50860 \end{pmatrix} =$$

$$\begin{pmatrix} 1.793731 & -1.013858 & -0.659259 \\ -2.106552 & 0.487054 & 0.072215 \\ 1.935562 & 1.193397 & -0.330762 \\ -2.500036 & -0.331474 & -0.346333 \\ 0.877295 & -0.335119 & 1.264139 \end{pmatrix}$$

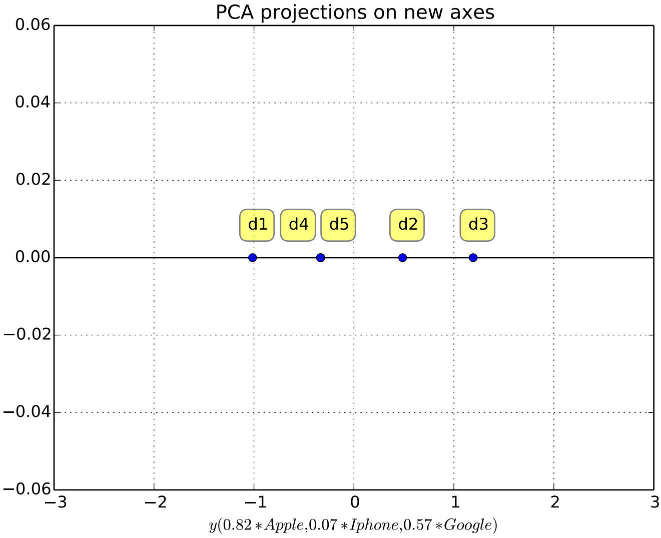
# PCA example


Geometric representation of data
d1
d2
d3
d4
d5
$x(0.39*Apple, 0.65*Iphone, -0.65*Google)$
$y(0.82*Apple, 0.07*Iphone, 0.57*Google)$
$z(0.42*Apple, -0.75*Iphone, -0.51*Google)$
−2
−1
0
1
−2
−1
0
1
1
0
−1
−2

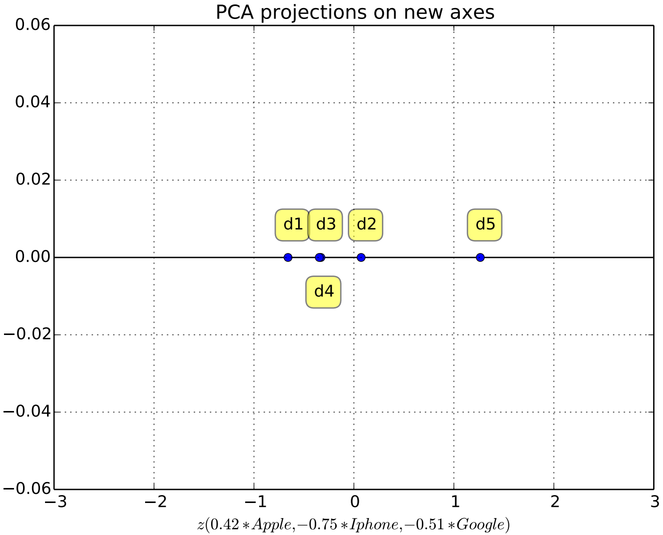
# PCA example


PCA projections on new axes
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
−3
−2
−1
0
1
2
3
d4
d2
d5
d1
d3
$x(0.39 * Apple, 0.65 * Iphone, -0.65 * Google)$

# PCA example


PCA projections on new axes
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
−3
−2
−1
0
1
2
3
d1
d4
d5
d2
d3
$y(0.82*Apple, 0.07*Iphone, 0.57*Google)$

# PCA example


PCA projections on new axes
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
−3
−2
−1
0
1
2
3
d1
d3
d2
d5
d4
$z(0.42 * Apple, −0.75 * Iphone, −0.51 * Google)$

# Variance along dimensions

$$var(\mathbf{X}\mathbf{v}_1) = 3.6843$$

$$var(\mathbf{X}\mathbf{v}_2) = 0.5823$$

$$var(\mathbf{X}\mathbf{v}_3) = 0.45345$$

- The variance along the axes decays with each new axes

# PCA: example

- From the example we also see the general principle
- The matrix **XE** keeps the transformed points
- Each column represent an axis in the new space
- The variance along the axes decays with each new axes, thus each new axis is less significant than the previous one
- Since the axes are orthogonal then the values along the axes are linearly uncorrelated
- We might drop less significant axes

# PCA: example

$$\mathbf{XE}_r = \begin{pmatrix} -0.4 & 1.6 & -1.4 \\ -0.4 & -1.4 & 1.6 \\ 1.6 & 1.6 & -0.4 \\ -1.4 & -1.4 & 1.6 \\ 0.6 & -0.4 & -1.4 \end{pmatrix} \begin{pmatrix} 0.39348 \\ 0.65496 \\ -0.64514 \end{pmatrix} =$$

$$\begin{pmatrix} 1.793731 \\ -2.106552 \\ 1.935562 \\ -2.500036 \\ 0.877295 \end{pmatrix}$$

# PCA: example

- Thus, we reduce dimensions
- PCA can be also understood as a data compression technique
- We remove (reduce) the values where the information content is small
- You can relate PCA with information theory
- It is possible to show that if the data is Gaussian then the PCA is also optimal from the information theoretic point of view, i.e. the most significant axes have the maximal information content

# PCA: Interpretation

- The first principal component: $(0.39 * \mathit{Apple}, 0.65 * \mathit{Iphone}, -0.65 * \mathit{Google})$
- This tells us that “Apple” and “Iphone” have a positive correlation
- They occur very often together in a document
- “Google” has a negative correlation with both “Apple” and “Iphone”
- When “Google” occurs in a document then “Apple” and “Iphone” typically do not occur
- This document collection is either about “Google” or about “Apple” / “Iphone”

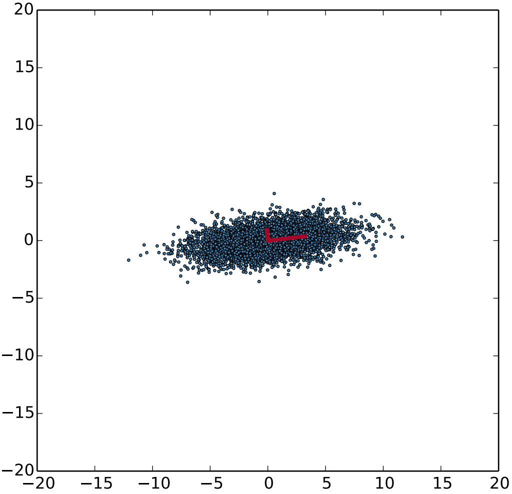
# PCA: Interpretation

- This a typical process when applying PCA
- You can interpret the first couple of principal components to learn something about the dataset
- Data mining
- However, be very careful: you can not generalize from this that all documents are about "Apple" or "Google"
- You might have document collection with topics such as technology and sports
- Then you can expect that "Apple" and "Google" positively correlate because they occur together in documents about the technology

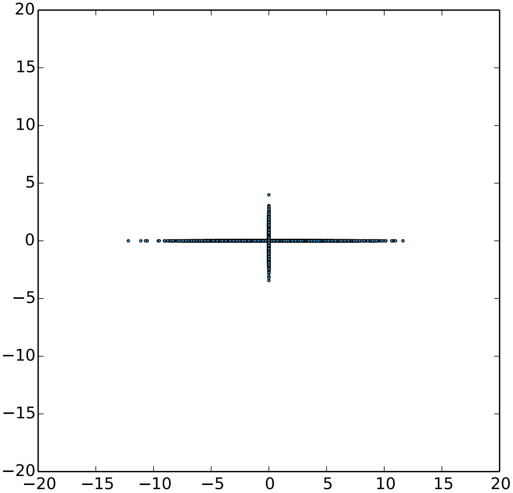
# PCA: Algorithm

- Organize data as an $n \times m$ matrix, with $n$ data points and $m$ features
- **Subtract the average for each feature** to obtain centered data matrix $\mathbf{X}$
- Calculate the covariance matrix $\frac{1}{n}\mathbf{X}^T\mathbf{X}$
- Calculate the eigenvalues and the eigenvectors of the covariance matrix
- Select the top $r$ eigenvectors
- Project the data to the new space spanned by those $r$ eigenvectors: $\mathbf{XE} \in \mathbb{R}^{n \times r}$, where $\mathbf{E} \in \mathbb{R}^{m \times r}$

# PCA example



# PCA example



# PCA example

- IPython Notebook examples
- http://kti.tugraz.at/staff/denis/courses/kddm1/pca.ipynb

### Command Line

```
ipython notebook –pylab=inline pca.ipynb
```

# PCA: Limitations

- PCA transforms the set of correlated observations into a set of linearly uncorrelated observations
- I.e. the goal of the analysis is to decorrelate the data
- In other words, the goal is to remove second-order dependencies in the data
- However, if the higher-order dependencies in the data exist removing only the second-order dependencies will not completely decorrelate the data
- First workaround: apply a nonlinear (kernel) transformation first
- Second workaround: require data to be statistically independent rather than linearly independent along the dimensions

# Advanced: Power Method

- Since the eigenvectors form a basis for the vector space (e.g. in case of a symmetric matrix)
- We can write $\mathbf{u}(0)$ as a linear combination of the eigenvectors $\mathbf{v}_i$ of the matrix (for appropriate choice of constants $c_i$)

$$\mathbf{u}(0) = \sum_i c_i \mathbf{v}_i$$

$$\mathbf{u}(t) = \mathbf{A}^t \sum_i c_i \mathbf{v}_i = \sum_i c_i \mathbf{A}^t \mathbf{v}_i = \sum_i c_i \lambda_i^t \mathbf{v}_i = \lambda_1^t \sum_i c_i \left[\frac{\lambda_i}{\lambda_1}\right]^t \mathbf{v}_i$$

# Advanced: Power Method

$$\mathbf{u}(t) = \lambda_1^t \sum_i c_i \left[\frac{\lambda_i}{\lambda_1}\right]^t \mathbf{v}_i$$

- $\lambda_i$ are eigenvalues, and $\lambda_1$ is the largest of themselves
- $\frac{\lambda_i}{\lambda_1} < 1$ for all $i > 1$
- When $t \to \infty$ $\frac{\lambda_i}{\lambda_1} \to 0$, for all $i > 1$
- When $t \to \infty$ $\mathbf{u}(t) \to c_1 \lambda_1^t \mathbf{v}_1$

# Advanced: Power Method

- In other words, the limiting behavior of the system is proportional to the leading eigenvector of the matrix
- The system evolves in the direction of the eigenvectors
- However, the leading eigenvector dominates all other eigenvectors
- The limiting behavior is therefore dominated by the leading eigenvalue and the leading eigenvector

# Advanced: Power Method

- Typically, we would calculate the leading eigenvector and leading eigenvalue iteratively
- A standard approach is the power method
- We make an initial guess about the eigenvector $\mathbf{x}^0$
- Then we iteratively calculate $\mathbf{x}^t$ (which converges to the leading eigenvector)

$$\mathbf{x}^t = \frac{\mathbf{A}\mathbf{x}^{(t-1)}}{||\mathbf{A}\mathbf{x}^{(t-1)}||_2}$$

# Advanced: Power Method

- In other words, the limiting vector is approximately equal the leading eigenvector of the matrix
- At the end of the iteration the leading (principal) eigenvalue can be calculated as:

$$\lambda_1 = \mathbf{x}^T \mathbf{A} \mathbf{x}$$

# Advanced: Power Method

- To find the second eigenpair we create a new matrix $\mathbf{A}^* = \mathbf{A} - \lambda_1 \mathbf{x}\mathbf{x}^T$
- We then again use the power iteration to calculate the leading eigenpair of $\mathbf{A}^*$
- This leading eigenpair corresponds to the second largest eigenpair of the original matrix $\mathbf{A}$
- Intuitively, we have eliminated the influence of a given eigenvector by setting its associated eigenvalue to zero

# Advanced: Power Method

- More formally, if $\mathbf{A}^* = \mathbf{A} - \lambda_1 \mathbf{x}\mathbf{x}^T$ where $\lambda_1$ is the leading eigenvalue of $\mathbf{A}$ and $\mathbf{x}$ is the leading eigenvector of $\mathbf{A}$ then
  1. $\mathbf{x}$ is also an eigenvector of $\mathbf{A}^*$ where the corresponding eigenvalue is 0.
  2. If $\mathbf{v}$ and $\lambda_v$ are eigenpair of $\mathbf{A}$ other then the principal eigenpair that they are also an eigenpair of $\mathbf{A}^*$

# Advanced: Power Method

Proof.

- We assume that **A** is a symmetric matrix

1. $\mathbf{A}^*\mathbf{x} = (\mathbf{A} - \lambda_1 \mathbf{x}\mathbf{x}^T)\mathbf{x} = \mathbf{A}\mathbf{x} - \lambda_1 \mathbf{x}\mathbf{x}^T\mathbf{x} = \mathbf{A}\mathbf{x} - \lambda_1 \mathbf{x} = \mathbf{0} = 0\mathbf{x}$
2. $\mathbf{A}^*\mathbf{v} = (\mathbf{A}^*)^T\mathbf{v} = (\mathbf{A} - \lambda_1 \mathbf{x}\mathbf{x}^T)^T\mathbf{v} = \mathbf{A}^T\mathbf{v} - \lambda_1 \mathbf{x}\mathbf{x}^T\mathbf{v} = \mathbf{A}^T\mathbf{v} = \mathbf{A}\mathbf{v} = \lambda_v \mathbf{v}$

□

# Advanced: Maximizing the variance

- We can specify an axis by a unit vector **w** lying on that axis
- A projection of another (centered) vector **x** onto the axis specified by **w** is given by the inner product of those two vectors:

$$\mathbf{x}^T\mathbf{w}$$

- Centered vector is a vector where the average has been subtracted
- If we combine all (centered) data vectors into a matrix **X** then the projection of the matrix onto the axis specified by **w** is given by:

$$\mathbf{X}\mathbf{w}$$

# Advanced: Maximizing the variance

- The variance of a single row from the matrix is given by:

$$(\mathbf{x}^T \mathbf{w})^2$$

- The variance of the complete projection is then given by:

$$\sigma^2 = \frac{1}{m} \sum_i (\mathbf{x}_i^T \mathbf{w})^2$$

# Advanced: Maximizing the variance

- In matrix form the variance is given by:

$$\sigma^2 = \frac{1}{m}(\mathbf{Xw})^T(\mathbf{Xw}) = \frac{1}{m}\mathbf{w}^T\mathbf{X}^T\mathbf{Xw} = \mathbf{w}^T\frac{\mathbf{X}^T\mathbf{X}}{m}\mathbf{w} = \mathbf{w}^T\mathbf{Vw}$$

- Now, we want to choose a unit vector **w** that maximizes $\sigma^2$
- It must be a unit vector, thus the constraint $\mathbf{w}^T\mathbf{w} = 1$ must be satisfied

# Advanced: Constrained optimization: Lagrange multipliers

- Original objective function that we want to minimize: $\mathbf{w}^T\mathbf{V}\mathbf{w}$
- This function is subject to constraint: constrained optimization
- Typically solved by the method of Lagrange multipliers

$$\begin{aligned} \text{Objective function:} \quad f(\mathbf{w}) &= \mathbf{w}^T\mathbf{V}\mathbf{w} \\ \text{Subject to:} \quad \mathbf{w}^T\mathbf{w} &= 1 \end{aligned}$$

# Advanced: Lagrange multipliers

- For each constraint we need one Lagrange multiplier, e.g. $\lambda$
- Lagrange formulation of the optimization problem will be a new objective function that is a function of **s** and $\lambda$

$$L(\mathbf{w}, \lambda) = \mathbf{w}^T\mathbf{V}\mathbf{w} - \lambda(\mathbf{w}^T\mathbf{w} - 1)$$

# Advanced: Constrained optimization

- To minimize $L$ we find $\mathbf{w}$ and $\lambda$ that make its gradient 0
- $\bigtriangledown L = 0$ :

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{w}} &= 0 \\ \frac{\partial L}{\partial \lambda} &= 0 \end{aligned}$$

# Advanced: Constrained optimization

- $\frac{\partial L}{\partial \lambda} = 0$ give back the constraint

$$\frac{\partial L}{\partial \mathbf{w}} = 2\mathbf{V}\mathbf{w} - 2\lambda\mathbf{w} = 0$$

$$\mathbf{V}\mathbf{w} = \lambda\mathbf{w}$$

# Advanced: Constrained optimization

- Thus, desired vector $\mathbf{w}$ is an eigenvector of the covariance matrix $\mathbf{V}$
- The maximizing vector will be the one associated with the largest eigenvalue $\lambda$
- $\mathbf{V}$ is a covariance matrix, thus it will be symmetric
- The eigenvectors are orthogonal and can be found by the power method
- They are called principal components