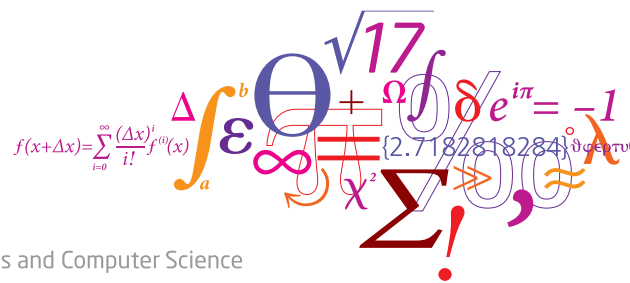
02450: Introduction to Machine Learning and Data Mining

Data, feature extraction and PCA

Georgios Arvanitidis

DTU Compute, Technical University of Denmark (DTU)



DTU Compute
Department of Applied Mathematics and Computer Science

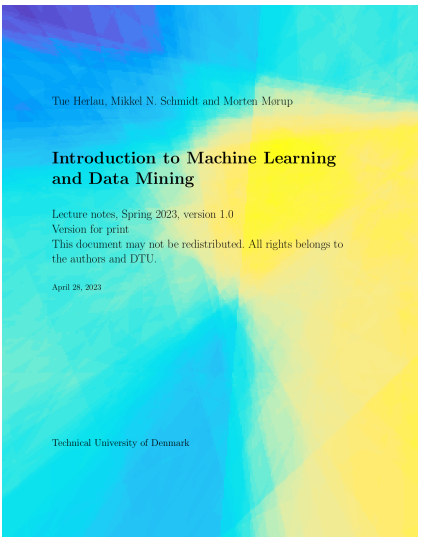
Today

Feedback Groups of the day:

Abrahim Deiaa El Din Abbas, Johanne Abildgaard, Julieta Aceves, Helle Achari, Magnus Møller Aggernæs, Subhayan Akhuli, Melis Cemre Akyol, Malek Al Abed, Mohammad Al-Ansari, Maximillian Al-Helo, Ismail Ali, Yusuf Mohamed Alin, Mads Albert Alkjærsig, Javier Alonso Fernandez, Rikke Alstrup, Muhammad Hussain Rashid Al-Takmaji, Mohamad Malaz Mohamed Alzarrad, Saeed Mohamud Amin, William Kirk Andersen, Mikkel Arn Andersen, Simon Rung Andersen, Oline Melinda Andersen, Jeppe Aarup Andersen, Mathias Vith Andersen, Giulia Andreatta, Sander Skjolden Andresen, ¿eljko Antunovi¿, Theo Rønne Appel, Pedro Aragon Fernandez, Ivan Antonino Arena, Alba Arias Martínez, Enerel Ariunbold, Amari Karakandi Arun, Matthew Hiroto Asano, Jacob Frederik Aslan-Lassen, Bergur Ástráðsson, Salomé Anaïs Aubri, Mike Auer, Eseme Ida Elena Ayiwe, Md Amin Azad, Gursharandeep Singh Badhesha, Haydar Hamid Abbas Bahr, Eline Agnes Jacoueline Balland, Volodymyr Baran, Samira Sanjay Barve, Laura Bauer, Quim Bech Vilaseca, Aslan Dalhoff Behbahani, Nikolaj Ivø Beier, Alex Belai, Magnus Johan Berg-Arnbak, Toms Rudolfs Berzins, Kaushik Amol Bhat, Mark Bidstrup, Kawa Shawki Bilal, Christian Lundgaard Bjerregaard, Magnus Bjørnskov, Emma Louise Blair, Louise Toft Blankensteiner

Reading material:

Chapter 2, Chapter 3



Lecture Schedule

1. Introduction
   29 August: C1

Data: Feature extraction, and visualization

2. **Data, feature extraction and PCA**
   5 September: C2, C3
3. Measures of similarity, summary statistics and probabilities
   12 September: C4, C5
4. Probability densities and data visualization
   19 September: C6, C7

Supervised learning: Classification and regression

5. Decision trees and linear regression
   26 September: C8, C9
6. Overfitting, cross-validation and Nearest Neighbor
   3 October: C10, C12 **(Project 1 due before 13:00)**
7. Performance evaluation, Bayes, and Naive Bayes
   10 October: C11, C13
8. Artificial Neural Networks and Bias/Variance
   24 October: C14, C15
9. AUC and ensemble methods
   31 October: C16, C17

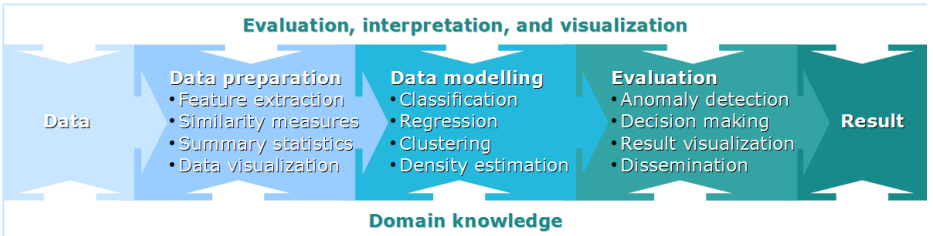
Unsupervised learning: Clustering and density estimation

10. K-means and hierarchical clustering
    7 November: C18
11. Mixture models and density estimation
    14 November: C19, C20 **(Project 2 due before 13:00)**
12. Association mining
    21 November: C21

Recap

13. Recap and discussion of the exam
    28 November: C1-C21

Online help: Discussion Forum (Piazza) on DTU Learn
Videos of lectures: `https://panopto.dtu.dk`
Streaming of lectures: Zoom (link on DTU Learn)



Learning Objectives

- Understand the types of data, their attributes and data issues
- Understand the bag of word representation
- Be able to apply principal component analysis for data visualization and feature extraction

What is data?

- Collection of data objects and their attributes
- An attribute is a property or characteristic of an object
  - Also known as variable, field, characteristic, or feature
- Collection of attributes describe an object
  - Also known as record, point, case, sample, entity, or instance

Attributes

Data objects

ID	Age	Gender	Name
1	31	F	Alex
2	24	M	Ben
3	52	F	Cindy
4	35	M	Dan
5	58	M	Eric
6	46	F	Fay
7	42	M	George

Discrete / continuous attributes

- **Discrete**
  - Finite (or countably infinite) set of values
  - Examples:
    - Zip codes
    - Counts
    - Set of words in a collection of documents
  - Often represented as integer variables
- **Continuous**
  - Has real numbers as attribute values
  - Examples:
    - Temperature
    - Height
    - Weight.
  - Often represented as floating point variables

Types of attributes

- **Nominal:** Objects belong to a category (Equal / Not equal)
  - ID numbers
  - Eye color
  - Zip codes
- **Ordinal:** Objects can be ranked (Greater than / Less than)
  - Taste of potato chips on a scale from 1-10
  - Grades
  - Height in {short, medium, tall}
- **Interval:** Distance between objects can be measured (Addition / Subtraction)
  - Calendar dates
  - Temperature in Fahrenheit and Celcius
- **Ratio:** Zero means absence of what is measured (Multiplication / Division)
  - Length
  - Time
  - Counts
  - Temperature in Kelvin

Discussion

- **Classify the following attributes**
  a) Military rank
  b) Angles measured in degrees
  c) A persons year of birth
  d) A persons age in years
  e) Coat check number
  f) Distance from center of campus
  g) Number of patients in a hospital

- **Discrete**
  - Finite (or countably infinite) set of values
- **Continuous**
  - Real number

- **Nominal** (Equal / Not equal)
  - Objects belong to a category
- **Ordinal** (Greater than / Less than)
  - Objects can be ranked
- **Interval** (Addition / Subtraction)
  - Distance between objects can be measured
- **Ratio** (Multiplication / Division)
  - Zero means absence of what is measured

Quiz 1: Attribute types (Spring 2012)

No.	Attribute description	Abbrev.
x_1	Type (0 = served cold, 1 = served hot)	TYPE
x_2	Calories per serving	CAL
x_3	Grams of protein	PROT
x_4	Grams of fat	FAT
x_5	Milligrams of sodium	SOD
x_6	Grams of dietary fiber	FIB
x_7	Grams of complex carbohydrates	CARB
x_8	Grams of sugars	SUG
x_9	Milligrams of potassium	POT
x_{10}	Vitamins and minerals in 0%, 25%, or 100% of FDA recommendations	VIT
x_{11}	Shelf position (1, 2, or 3, counting from the floor)	SHELF
x_{12}	Weight in ounces of one serving	WEIGHT
x_{13}	Number of cups in one serving	CUPS
x_{14}	Name of cereal brand	NAME
y	Average rating of the cereal (from 0 to 100)	RAT

Table 1: Attributes in a study of cereals (i.e. breakfast products, taken from http://lib.stat.cmu.edu/DASL/Datafiles/Cereals.html).

In a study of healthy breakfast habits 77 cereal brands were investigated. The attributes of the data are given in Table 1. There are a total of 14 attributes denoted x_1–x_{14} and one output variable y which defines the average rating of the cereal products by the consumers.

Which statement about the attributes in the data set is *incorrect*?

A. NAME is discrete and nominal.

B. PROT, FAT and SOD are all continuous and ratio.

C. TYPE and VIT are both discrete and ordinal.

D. An attribute that is ratio will also be interval.

E. Don't know.

Types of data sets

- **Record data**
  - Collection of data objects and their attributes
  - Representation: Table
- **Relational data**
  - Collection of data objects and their relation
  - Representation: Graph
- **Ordered data**
  - Ordered collection of data objects
  - Representation: Sequence

Record data example: Market basket data

- **Transaction data table**

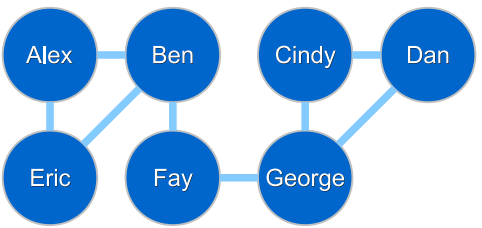
ID	Items
1	Bread, Soda, Milk
2	Beer, Bread
3	Beer, Soda, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Soda, Diaper, Milk

- **Matrix**

ID	Bread	Soda	Milk	Beer	Diaper
1	1	1	1	0	0
2	1	0	0	1	0
3	0	1	1	1	1
4	1	0	1	1	1
5	0	1	1	0	1

Relational data example: Who knows who?

- **Graph**

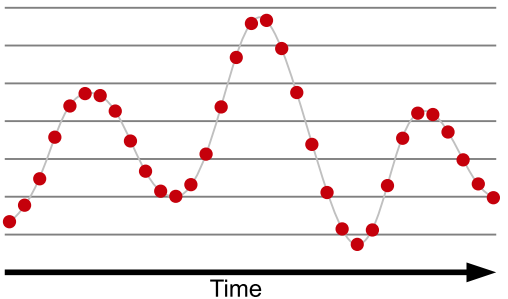


- **Matrix**

	A	B	C	D	E	F	G
A	0	1	0	0	1	0	0
B	1	0	0	0	1	1	0
C	0	0	0	1	0	0	1
D	0	0	1	0	0	0	1
E	1	1	0	0	0	0	0
F	0	1	0	0	0	0	1
G	0	0	1	1	0	1	0

Ordered data example: Time series

- **Sequence**



- **Matrix**

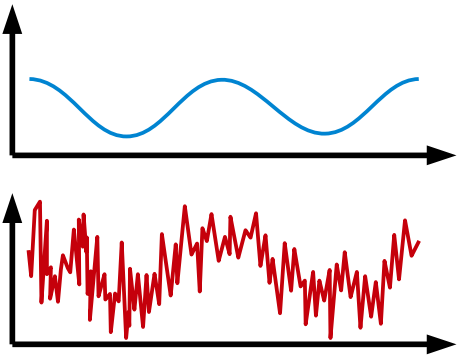
Time	Value
0	1.3
1	1.8
2	2.5
3	3.6
4	4.4
5	4.7
6	4.6
7	4.3
8	2.4
9	2.1
10	2.0
11	2.3
12	3.1

Data quality

- **Data is of high quality if they**
  - Are fit for their intended use
  - Correctly represent the phenomena they correspond to
- **Examples of quality problems**
  - Noise
  - Outliers
  - Missing values

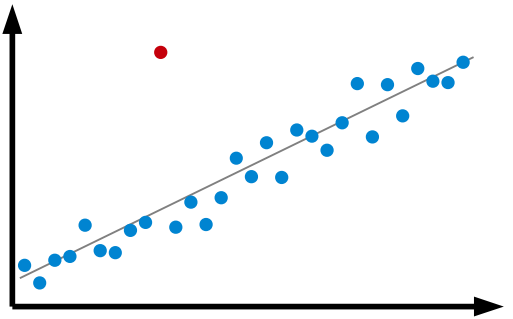
Noise

- **Definition**
  - Unwanted pertubation to a signal
  - Unwanted data
- **Reasons for noise**
  - Limits in measurement accuracy
  - Interference from other signals
  - Measurement of attributes not related to the data modeling task
- **Handling noise**
  - Exclude noisy attributes
  - Remove noise by filtering
  - Include a model of the noise



Outliers

- **Definition**
  - Data objects which are significantly different from most others
- **Reasons for outliers**
  - Measurement error
  - Natural property of data
- **Handling outliers**
  - Identify & exclude outliers
  - Model the outliers



Missing values

- **Definition**
  - No value is stored for an attribute in a data object
- **Reasons for missing values**
  - Information is not collected or measured
    - People decline to give their age
  - Attribute is not applicable
    - Annual income is not applicable to children
- **Handling missing values**
  - Eliminate data objects
  - Eliminate attributes
  - Estimate missing values (e.g. an average)
  - Ignore the missing value in analysis
  - Model the missing values

ID	Age	Gender	Name
1	31	F	Alex
2	(?)	M	Ben
3	52	F	Cindy
4	35	(?)	Dan
5	(?)	M	Eric
6	(?)	F	Fay
7	42	M	(?)

Discussion

- A group of people were asked to write how many children they have
  - Their response was this

- A research assistant typed the results into a table
  - His table looked like this

Children	3	1	0	2	7	5	15	0	1	3	-2	0	0	0	1

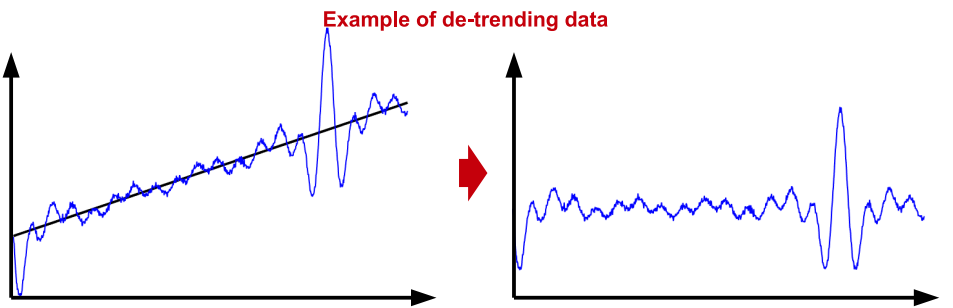
- Are there any data quality issues?
  - Noise?
  - Outliers?
  - Missing values?
- Why have these issues occured, and how should they be handled?

Dataset manipulations

- **Sampling**
  - Selecting a representative subset of data points
- **Feature subset selection**
  - Choose a subset of attributes
- **Feature extraction/transformation**
  - Create new features from existing attributes
  - Discretization and binarization
  - Apply a fixed transformation to an attribute
  - Aggregation several attributes into a single attribute
- **Dimensionality reduction**
  - Project data to a low-dimensional subspace

Feature processing

- Eliminating, suppressing, or attenuating certain aspects of the data
  - Noise removal in audio signals
  - Elimination of common words in text documents
  - Removal of background in images
  - Removal of examples which are corrupted
  - De-trending data (if it is not stationary)



Common feature transformations

ID	MPG	Cylinders	Horsepower	Weight	Year	Safety	Acceleration	Origin
1	18	8	150	3436	70	4	11	France
2	28	4	79	2625	82	4	18.6	USA
3	26	4	79	2255	76	3	17.7	USA
3	29	4	70	1937	76	1	14.2	Germany
4	NaN	8	175	3850	70	2	11	USA
5	24	4	90	2430	70	3	14.5	Germany
6	17.5	6	95	3193	76	4	17.8	USA
7	25	4	87	2672	70	-100	17.5	France
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
142	15	8	198	4341	70	2	10	USA

$$\boldsymbol{X} = \begin{bmatrix} 18 & 8 & 150 & 3436 & 70 & 4 & 11 & 3 \\ 28 & 4 & 79 & 2625 & 82 & 4 & 18.6 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 15 & 8 & 198 & 4341 & 70 & 2 & 10 & 1 \end{bmatrix}$$

Standardize:

$$\boldsymbol{X} = \begin{bmatrix} \cdots & (X_{1j} - \hat{\mu}_j)/\hat{\sigma}_j & \cdots \\ \cdots & (X_{2j} - \hat{\mu}_j)/\hat{\sigma}_j & \cdots \\ & \vdots & \\ \cdots & (X_{Nj} - \hat{\mu}_j)/\hat{\sigma}_j & \cdots \end{bmatrix}$$

$$\hat{\mu}_j = \frac{1}{N}\sum_{i=1}^{N} X_{ij}, \quad \hat{\sigma}_j = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_{ij} - \mu_j)^2}$$

Binarize/threshold:

$$\boldsymbol{X} = \begin{bmatrix} \cdots & 1_{[\theta,\infty[}(x_{1j}) & \cdots \\ \cdots & 1_{[\theta,\infty[}(x_{2j}) & \cdots \\ & \vdots & \\ \cdots & 1_{[\theta,\infty[}(x_{Nj}) & \cdots \end{bmatrix}$$

$$1_{[\theta,\infty[}(x) = 1 \text{ if } x \geq \theta \text{ otherwise } 0$$

One-out-of K encoding

One-out-of-K coding

$X =$

Age	Height	Weight	Nationality
-0.2248	-0.4762	-0.2097	'Sweden'
-0.5890	0.8620	0.6252	'Sweden'
-0.2938	-1.3617	0.1832	'Sweden'
-0.8479	0.4550	-1.0298	'Sweden'
-1.1201	-0.8487	0.9492	'Norway'
2.5260	-0.3349	0.3071	'Norway'
1.6555	0.5528	0.1352	'Norway'
0.3075	1.0391	0.5152	'Norway'
-1.2571	-1.1176	0.2614	'Norway'
-0.8655	1.2607	-0.9415	'Sweden'
-0.1765	0.6601	-0.1623	'Norway'
0.7914	-0.0679	-0.1461	'Denmark'
-1.3320	-0.1952	-0.5320	'Denmark'
-2.3299	-0.2176	1.6821	'Sweden'
-1.4491	-0.3031	-0.8757	'Sweden'
0.3335	0.0230	-0.4838	'Sweden'
0.3914	0.0513	-0.7120	'Denmark'
0.4517	0.8261	-1.1742	'Sweden'
-0.1303	1.5270	-0.1922	'Norway'
0.1837	0.4669	-0.2741	'Denmark'

Age	Height	Weight			
-0.2248	-0.4762	-0.2097	0	0	1
-0.5890	0.8620	0.6252	0	0	1
-0.2938	-1.3617	0.1832	0	0	1
-0.8479	0.4550	-1.0298	0	0	1
-1.1201	-0.8487	0.9492	0	1	0
2.5260	-0.3349	0.3071	0	1	0
1.6555	0.5528	0.1352	0	1	0
0.3075	1.0391	0.5152	0	1	0
-1.2571	-1.1176	0.2614	0	1	0
-0.8655	1.2607	-0.9415	0	0	1
-0.1765	0.6601	-0.1623	0	1	0
0.7914	-0.0679	-0.1461	1	0	0
-1.3320	-0.1952	-0.5320	1	0	0
-2.3299	-0.2176	1.6821	0	0	1
-1.4491	-0.3031	-0.8757	0	0	1
0.3335	0.0230	-0.4838	0	0	1
0.3914	0.0513	-0.7120	1	0	0
0.4517	0.8261	-1.1742	0	0	1
-0.1303	1.5270	-0.1922	0	1	0
0.1837	0.4669	-0.2741	1	0	0

Bag of words representation

- First three sentences on **wikipedia.org**
  - The bag-of-words model is a simplifying assumption used in natural language processing and information retrieval
  - In this model, a text (such as a sentence or a document) is represented as an unordered collection of words, disregarding grammar and even word order
  - The bag-of-words model is used in some methods of document classification

(Image source: https://pixabay.com/p-297223/)

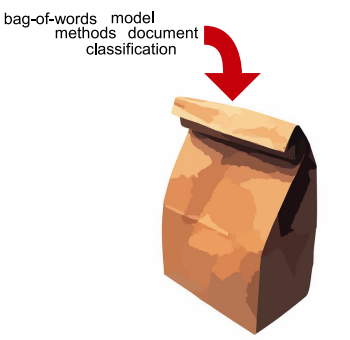
Bag of words representation

- First three sentences on **wikipedia.org**
  - The bag-of-words model is a simplifying assumption used in natural language processing and information retrieval
  - In this model, a text (such as a sentence or a document) is represented as an unordered collection of words, disregarding grammar and even word order
  - The bag-of-words model is used in some methods of document classification

- We will treat **this text** as a data set and create a bag-of-words model of it

Bag of words representation

- Elimination of common words (so-called stop words)
  - The bag-of-words model is a simplifying assumption used in natural language processing and information retrieval
  - In this model, a text (such as a sentence or a document) is represented as an unordered collection of words, disregarding grammar and even word order
  - The bag-of-words model is used in some methods of document classification



Bag of words representation

- Representation as matrix

Word	Sentence		
	1	2	3
bag-of-words	1		1
model	1	1	1
simplifying	1		
assumption	1		
natural	1		
language	1		
processing	1		
information	1		
retrieval	1		
text		1	
sentence		1	
document		1	1
represented		1	
unordered		1	
collection		1	
words		1	
disregarding		1	
grammar		1	
word		1	
order		1	
methods			1
classification			1

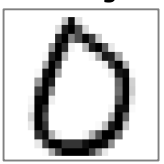
Bag of words representation

- Stemming

Word	Sentence		
	1	2	3
bag-of-word*	1		1
model*	1	1	1
simplif*	1		
assum*	1		
natural*	1		
languag*	1		
process*	1		
information*	1		
retriev*	1		
text*		1	
sentence*		1	
document*		1	1
represent*		1	
unorder*		1	
collect*		1	
word*		2	
disregard*		1	
grammar*		1	
order*		1	
method*			1
classif*			1

Image representation

- **Example: Handwritten digits**
- Preprocessing
  - Digitalization
  - Centering
  - Rotation
  - Scaling



28×28

$$M_0 = \begin{bmatrix} 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0.3 & 1 & 0.2 & 0 & \dots & 0 \\ \vdots & & & & & & & \vdots \\ 0 & \dots & 0 & 0 & 0 & 0 & \dots & 0 \end{bmatrix}$$

- Vectorization

1×784

$$x_0 = \begin{bmatrix} 0 & \cdots\cdots & 0 & 0.3 & 1 & 0.2 & 0 & \cdots\cdots & 0 \end{bmatrix}^\top$$

- Matrix representation of data set

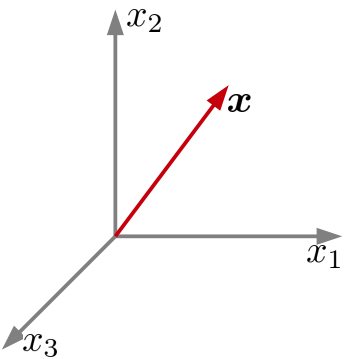
$$\boldsymbol{X} = \begin{bmatrix} \text{---} & \boldsymbol{x}_1 & \text{---} \\ \text{---} & \boldsymbol{x}_2 & \text{---} \\ & \vdots & \\ \text{---} & \boldsymbol{x}_N & \text{---} \end{bmatrix}$$

If each image is 28×28 pixels then $\boldsymbol{X}$ is a $N \times 784$ matrix.

Vector space representation

- All these data objects have a vector space representation

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_M \end{bmatrix}$$



Plan for the rest of today:

- Linear algebra recap (subspaces and projections)
- The goal of Principal Component Analysis (PCA)
- Derivation of PCA
- Singular Value Decomposition used to implement PCA
- Use of PCA for data visualization

Vectors and matrices

- Common matrix notation

$$\boldsymbol{A}, A, \overline{\overline{A}}$$

$$\boldsymbol{A} = \begin{bmatrix} a_{1,1} & \cdots & a_{1,M} \\ \vdots & & \vdots \\ a_{N,1} & \cdots & a_{N,M} \end{bmatrix} \in \mathbb{R}^{N \times M}$$

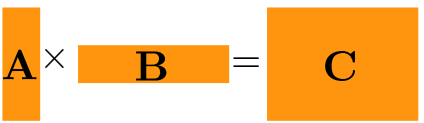
- Common vector notation

$$\boldsymbol{x}, x, \overline{x}, \vec{x}$$

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_M \end{bmatrix} \in \mathbb{R}^{M}$$

Matrix multiplication

- Two matrices can be multiplied $\boldsymbol{AB} = \boldsymbol{C}$
  - if the number of columns in the first equals the number of rows in the second



$$\underset{3\times 4 \text{ matrix}}{\begin{bmatrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 1 & 2 & 3 & 4 \end{bmatrix}} \underset{4\times 5 \text{ matrix}}{\begin{bmatrix} \cdot & \cdot & \cdot & a & \cdot \\ \cdot & \cdot & \cdot & b & \cdot \\ \cdot & \cdot & \cdot & c & \cdot \\ \cdot & \cdot & \cdot & d & \cdot \end{bmatrix}} = \underset{3\times 5 \text{ matrix}}{\begin{bmatrix} \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & x_{3,4} & \cdot \end{bmatrix}}$$

$$x_{3,4} = 1 \cdot a + 2 \cdot b + 3 \cdot c + 4 \cdot d$$

Example: $\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix} = \begin{bmatrix} 19 & 22 \\ 43 & 50 \end{bmatrix}$

Matrix transpose

- The transpose of a matrix

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 4 & 6 \\ 7 & 8 & 9 \end{bmatrix} \quad \boldsymbol{A}^\top = \begin{bmatrix} 1 & 3 & 7 \\ 2 & 4 & 8 \\ 3 & 6 & 9 \end{bmatrix}$$

- Transpose of a sum

$$(\boldsymbol{A} + \boldsymbol{B})^\top = \boldsymbol{A}^\top + \boldsymbol{B}^\top$$

- Transpose of a product

$$(\boldsymbol{AB})^\top = \boldsymbol{B}^\top \boldsymbol{A}^\top$$

$$(\boldsymbol{Ax})^\top \boldsymbol{y} = \boldsymbol{x}^\top \boldsymbol{A}^\top \boldsymbol{y} = \boldsymbol{x}^\top (\boldsymbol{A}^\top \boldsymbol{y})$$

The identity matrix

- Ones on the diagonal and zeros everywhere else

$$\boldsymbol{I} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \qquad \boldsymbol{I}^\top = \boldsymbol{I}$$

- Multiplying by the identity does not change anything

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

$$\boldsymbol{I}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \boldsymbol{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

$$\boldsymbol{I}_2\boldsymbol{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

- For a square matrix, the inverse satisfies

$$\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$$

Norms

- The (Euclidian) norm of a vector measures it's length (magnitude):

$$\|\boldsymbol{x}\|_2 = \sqrt{\sum_{n=1}^{N} |x_n|^2} = \sqrt{\boldsymbol{x}^\top \boldsymbol{x}}$$

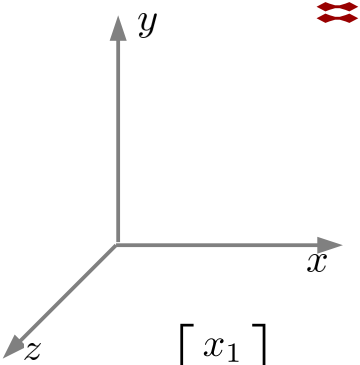
- The Frobenius norm of a matrix measures it's magnitude:

$$\|\boldsymbol{X}\|_F^2 = \sum_{i,j} x_{i,j}^2 = \mathrm{trace}(\boldsymbol{X}\boldsymbol{X}^T) = \mathrm{trace}(\boldsymbol{X}^T\boldsymbol{X})$$

Where trace takes the sum of the diagonal elements, i.e. $\mathrm{trace}(\boldsymbol{A}) = \sum_{i=1}^{N} a_{i,i}$

Vector spaces

- A M-dimensional vector space is just $\mathbb{R}^M$
- This is the set of all M-dimensional vectors
- A vector space is closed under **linear combinations**



$$a_1\boldsymbol{x}_1 + a_2\boldsymbol{x}_2 + \cdots + a_n\boldsymbol{x}_n$$

$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ Vectors

$\mathrm{a}_1, \ldots, a_n$ Numbers

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_M \end{bmatrix} \in \mathbb{R}^M$$

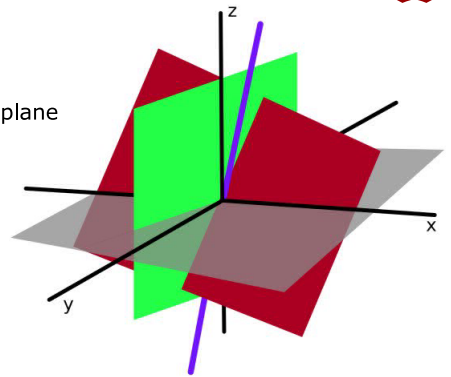
Subspaces

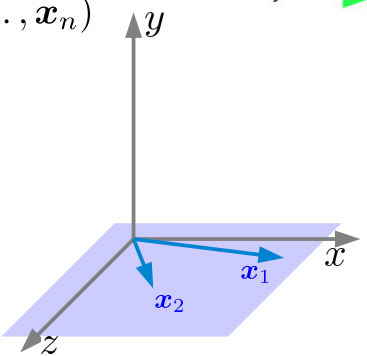
- A **subspace** generalizes the concept of a line/plane
- If we consider n vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ the **span** is then all linear combinations

$$\boldsymbol{z} = a_1\boldsymbol{x}_1 + a_2\boldsymbol{x}_2 + \cdots + a_n\boldsymbol{x}_n$$

and it is said to be a **subspace**

$$\mathrm{V} = \mathrm{span}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$$





Basis of a (sub)space

- Vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are said to be **linearly independent** if

$$\mathbf{0} = a_1\boldsymbol{x}_1 + a_2\boldsymbol{x}_2 + \cdots + a_n\boldsymbol{x}_n$$

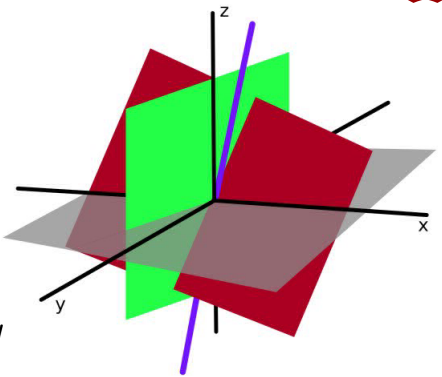
implies $a_1 = a_2 = \cdots = a_n = 0$

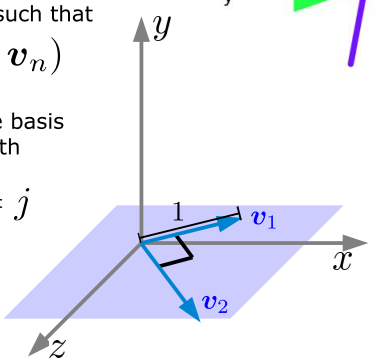
- A **basis** of a vector space V are n linearly independent vectors such that

$$\mathrm{V} = \mathrm{span}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)$$

- A basis is **orthonormal** if the basis is orthogonal and of unit length

$$\boldsymbol{v}_i^T \boldsymbol{v}_j = 0 \text{ for } i \neq j$$
$$\|\boldsymbol{v}_i\| = 1$$





Basis of a (sub)space

- A **basis** of a vector space V are n linearly independent vectors such that

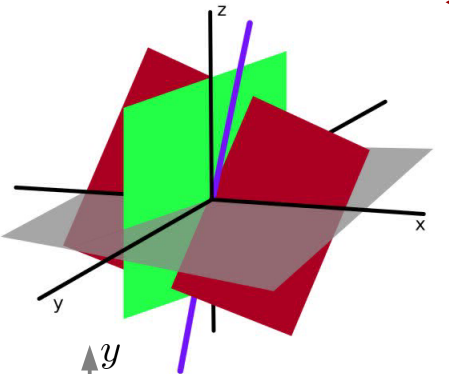
$$\mathrm{V} = \mathrm{span}(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)$$

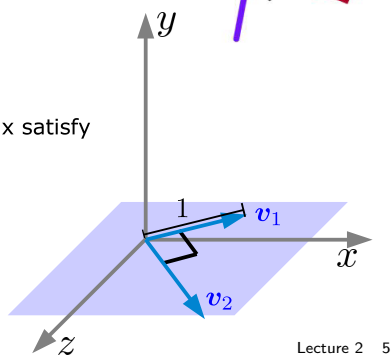
- We collect the basis into a matrix

$$\boldsymbol{V} = \begin{bmatrix} | & | & & | \\ \boldsymbol{v}_1 & \boldsymbol{v}_2 & \cdots & \boldsymbol{v}_N \\ | & | & & | \end{bmatrix}$$

- If the basis is orthonormal the matrix satisfy

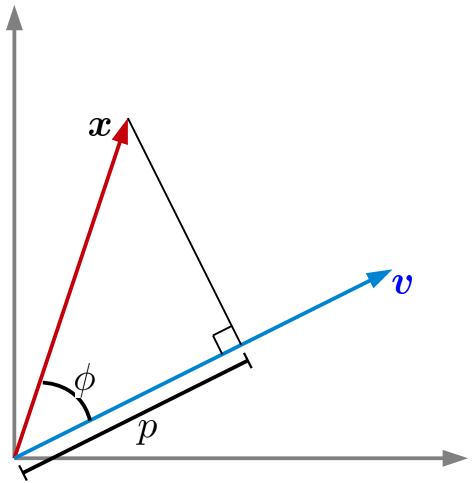
$$\boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{I}, \quad \boldsymbol{V}^\top = \boldsymbol{V}^{-1}$$





Projection

- Projection onto a vector



- Angle between vectors

$$\cos(\phi) = \frac{\boldsymbol{v}^\top \boldsymbol{x}}{\|\boldsymbol{x}\|_2 \|\boldsymbol{v}\|_2}$$

- Length of projection

$$p = \|\boldsymbol{x}\|_2 \cos(\phi) = \frac{\boldsymbol{v}^\top \boldsymbol{x}}{\|\boldsymbol{v}\|_2}$$

- Projection onto unit vector

$$p = \boldsymbol{v}^\top \boldsymbol{x}$$

Projection onto a subspace

- **Projection onto a subspace**
  - Subspace of dimension n defined by a orthonormal basis matrix $\boldsymbol{V}$
  - Projection of $\boldsymbol{x}$ (M dimensional) onto V given by

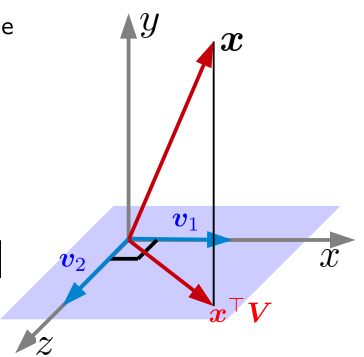
$$\boldsymbol{b}^T = \boldsymbol{x}^T \boldsymbol{V}$$

  - 'Reconstruction' can be found as: $\boldsymbol{x}' = \boldsymbol{V}\boldsymbol{b}$

Example: Projection of 3-D vector onto the (x,z) plane

$$\boldsymbol{V} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix} \qquad \boldsymbol{x} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

$$\boldsymbol{x}^\top V = \begin{bmatrix} x & y & z \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} x & z \end{bmatrix}$$



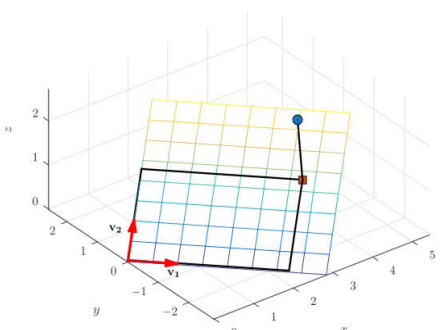
Projection onto a subspace

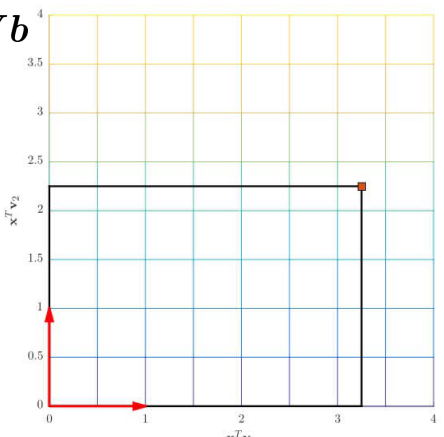
- **Projection onto a subspace**
  - Subspace of dimension n defined by a orthonormal basis matrix $\boldsymbol{V}$
  - Projection of $\boldsymbol{x}$ (M dimensional) onto V given by

$$\boldsymbol{b}^T = \boldsymbol{x}^T \boldsymbol{V}$$

  - 'Reconstruction' can be found as: $\boldsymbol{x}' = \boldsymbol{V}\boldsymbol{b}$

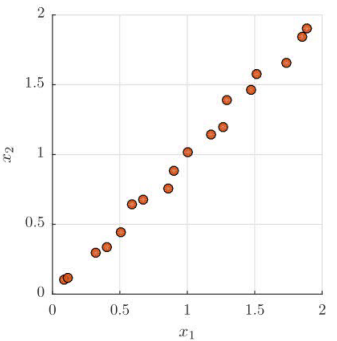
Example 2:





PCA for high-dimensional data

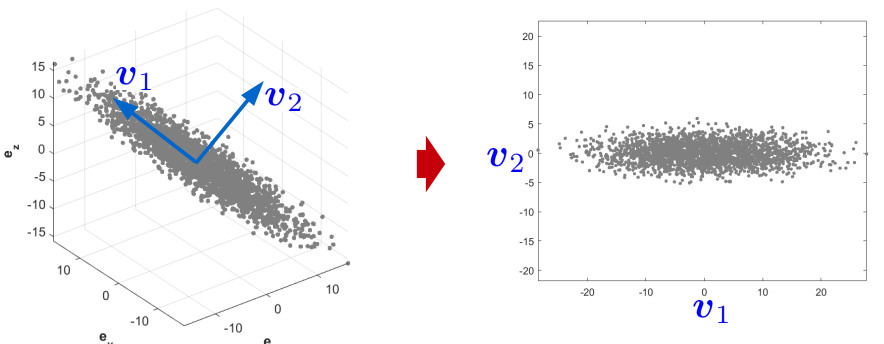
- Much data is high-dimensional
- We want to find a **lower**-dimensional representation of the **high**-dimensional data



(2 dimensional but really 1 dimensional)

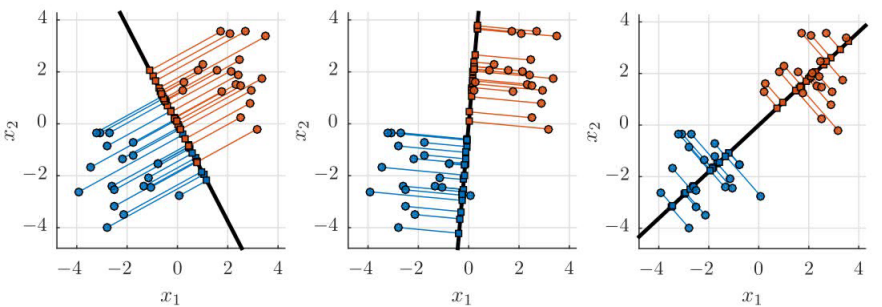
PCA for high-dimensional data

- Much data is high-dimensional
- We can **project high** dimensional data to a **lower** dimensional **subspace**
- But what is a good projection?



PCA for high-dimensional data

- Much data is high-dimensional
- We can project high dimensional data to a lower dimensional subspace
- But what is a good projection?
- Select projection that maximizes the variance of the projected data

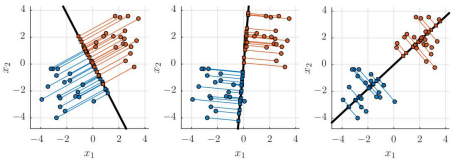


PCA derivation

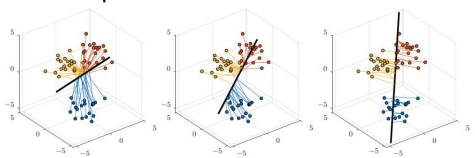
Projection of $\boldsymbol{x}_i$ onto unit vector $\boldsymbol{v}$: $b_i = \boldsymbol{x}_i^\top \boldsymbol{v}$

$$\begin{aligned}
\mathrm{Var}[b] &= \frac{1}{N-1}\sum_{i=1}^{N}\left[b_i - \frac{1}{N}\sum_{j=1}^{N} b_j\right]^2 \\
&= \frac{1}{N-1}\sum_{i=1}^{N}\left[\boldsymbol{x}_i^\top \boldsymbol{v} - \frac{1}{N}\sum_{j=1}^{N} \boldsymbol{x}_j^\top \boldsymbol{v}\right]^2 \\
&= \frac{1}{N-1}\sum_{i=1}^{N}\left[\left(\boldsymbol{x}_i^\top - \frac{1}{N}\sum_{j=1}^{N} \boldsymbol{x}_j^\top\right)\boldsymbol{v}\right]^2 \\
&= \frac{1}{N-1}\sum_{i=1}^{N}\left(\tilde{\boldsymbol{x}}_i^\top \boldsymbol{v}\right)^2 \qquad \boxed{\tilde{\boldsymbol{x}}_i = \boldsymbol{x}_i - \boldsymbol{m}} \\
&= \frac{1}{N-1}\sum_{i=1}^{N}\boldsymbol{v}^\top \tilde{\boldsymbol{x}}_i \tilde{\boldsymbol{x}}_i^\top \boldsymbol{v} = \frac{1}{N-1}\boldsymbol{v}^\top \tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} \boldsymbol{v}
\end{aligned}$$

2D example



3D example



$$\boldsymbol{A}\boldsymbol{v} = \lambda \boldsymbol{v}, \qquad \boldsymbol{A} \text{ is a } N \times N \text{ matrix}$$

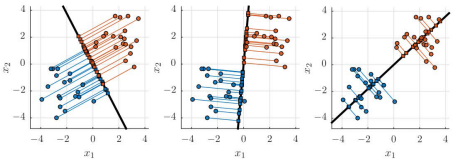
We say $\boldsymbol{v}$ is an eigenvector with eigenvalue λ

PCA derivation

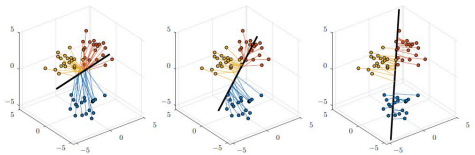
Projection of $\boldsymbol{x}_i$ onto unit vector $\boldsymbol{v}$: $b_i = \boldsymbol{x}_i^\top \boldsymbol{v}$

$$\begin{aligned}
\mathrm{Var}[b] &= \frac{1}{N-1}\sum_{i=1}^{N}\left[b_i - \frac{1}{N}\sum_{j=1}^{N} b_j\right]^2 \\
&= \frac{1}{N-1}\sum_{i=1}^{N}\left[\boldsymbol{x}_i^\top \boldsymbol{v} - \frac{1}{N}\sum_{j=1}^{N} \boldsymbol{x}_j^\top \boldsymbol{v}\right]^2 \\
&= \frac{1}{N-1}\sum_{i=1}^{N}\left[\left(\boldsymbol{x}_i^\top - \frac{1}{N}\sum_{j=1}^{N} \boldsymbol{x}_j^\top\right)\boldsymbol{v}\right]^2 \\
&= \frac{1}{N-1}\sum_{i=1}^{N}\left(\tilde{\boldsymbol{x}}_i^\top \boldsymbol{v}\right)^2 \qquad \tilde{\boldsymbol{x}}_i = \boldsymbol{x}_i - \boldsymbol{m} \\
&= \frac{1}{N-1}\sum_{i=1}^{N}\boldsymbol{v}^\top \tilde{\boldsymbol{x}}_i \tilde{\boldsymbol{x}}_i^\top \boldsymbol{v} = \frac{1}{N-1}\boldsymbol{v}^\top \tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}}\boldsymbol{v}
\end{aligned}$$

2D example



3D example



$\boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v}, \qquad \boldsymbol{A}$ is a $N \times N$ matrix

We say $\boldsymbol{v}$ is an eigenvector with eigenvalue λ

$$\arg\max_{\boldsymbol{v}} \mathrm{Var}[b] = \arg\max_{\boldsymbol{v}} \boldsymbol{v}^T \tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}}\boldsymbol{v} \quad \text{s.t.} \quad ||\boldsymbol{v}||^2 = \boldsymbol{v}^T\boldsymbol{v} = 1$$

$$\mathcal{L}(\boldsymbol{v}, \lambda) = \boldsymbol{v}^T \tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}}\boldsymbol{v} - \lambda(\boldsymbol{v}^T\boldsymbol{v} - 1), \qquad \frac{\partial \mathcal{L}}{\partial \boldsymbol{v}} = 2\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}}\boldsymbol{v} - 2\lambda\boldsymbol{v} = 0$$

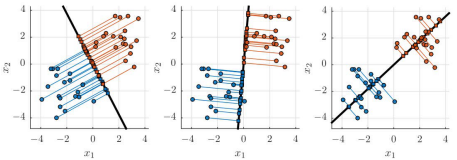
XTX = A or $\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}}\boldsymbol{v} = \lambda\boldsymbol{v}$

PCA derivation

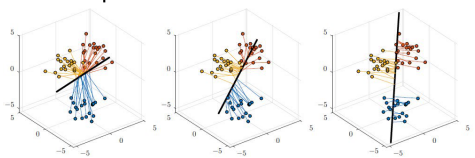
Projection of $\boldsymbol{x}_i$ onto unit vector $\boldsymbol{v}$: $b_i = \boldsymbol{x}_i^\top \boldsymbol{v}$

$$\begin{aligned}
\mathrm{Var}[b] &= \frac{1}{N-1}\sum_{i=1}^{N}\left[b_i - \frac{1}{N}\sum_{j=1}^{N} b_j\right]^2 \\
&= \frac{1}{N-1}\sum_{i=1}^{N}\left[\boldsymbol{x}_i^\top \boldsymbol{v} - \frac{1}{N}\sum_{j=1}^{N} \boldsymbol{x}_j^\top \boldsymbol{v}\right]^2 \\
&= \frac{1}{N-1}\sum_{i=1}^{N}\left[\left(\boldsymbol{x}_i^\top - \frac{1}{N}\sum_{j=1}^{N} \boldsymbol{x}_j^\top\right)\boldsymbol{v}\right]^2 \\
&= \frac{1}{N-1}\sum_{i=1}^{N}\left(\tilde{\boldsymbol{x}}_i^\top \boldsymbol{v}\right)^2 \qquad \boxed{\tilde{\boldsymbol{x}}_i = \boldsymbol{x}_i - \boldsymbol{m}} \\
&= \frac{1}{N-1}\sum_{i=1}^{N}\boldsymbol{v}^\top \tilde{\boldsymbol{x}}_i \tilde{\boldsymbol{x}}_i^\top \boldsymbol{v} = \frac{1}{N-1}\boldsymbol{v}^\top \tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}}\boldsymbol{v}
\end{aligned}$$

2D example



3D example



$\boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v}, \qquad \boldsymbol{A}$ is a $N \times N$ matrix

We say $\boldsymbol{v}$ is an eigenvector with eigenvalue λ

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{v}} = 2\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}\boldsymbol{v} - 2\lambda\boldsymbol{v} = 0 \quad \text{or} \quad \tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}\boldsymbol{v} = \lambda\boldsymbol{v}$$

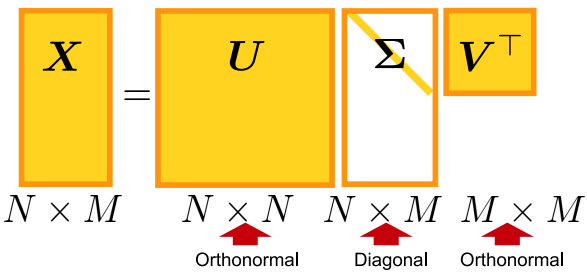
This means that $\mathrm{Var}[b] = \frac{1}{N-1}\boldsymbol{v}^T\lambda\boldsymbol{v} = \frac{1}{N-1}\lambda$

The Singular Value Decomposition (SVD)

(Eugino Beltrami & Camille Jordan, independently, 1873-1874)

Any N x M matrix can be decomposed as follows:

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$$



$\sigma_1, \ldots, \sigma_M$ is known as the singular values

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_M \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

$$\boldsymbol{U} = \begin{bmatrix} \boldsymbol{u}_1 & \cdots & \boldsymbol{u}_N \end{bmatrix} \quad \boldsymbol{V} = \begin{bmatrix} \boldsymbol{v}_1 & \cdots & \boldsymbol{v}_M \end{bmatrix}$$

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_M \geq 0$$

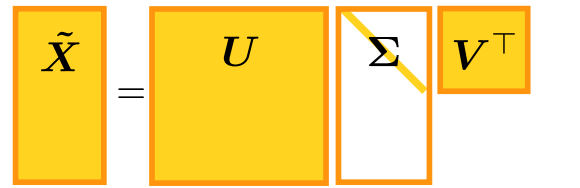
$$\text{if } i \neq j\text{: } \boldsymbol{\Sigma}_{i,j} = 0, \quad \boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}_{N\times N}, \quad \boldsymbol{V}^\top\boldsymbol{V} = \boldsymbol{I}_{M\times M}$$

The Singular Value Decomposition (SVD)

(Eugino Beltrami & Camille Jordan, independently, 1873-1874)

Any N x M matrix can be decomposed as follows:

$$\tilde{\boldsymbol{X}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$$



$$\begin{aligned}(\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}})\boldsymbol{v}_i &= (\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top)^\top \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top \boldsymbol{v}_i \\ &= (\boldsymbol{V}\boldsymbol{\Sigma}^\top \boldsymbol{U}^\top \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top)\boldsymbol{v}_i \\ &= \boldsymbol{V}\boldsymbol{\Sigma}^\top \boldsymbol{\Sigma}\boldsymbol{e}_i = \sigma_{ii}^2 \boldsymbol{v}_i\end{aligned}$$

$$\boldsymbol{A}\boldsymbol{v} = \lambda \boldsymbol{v}, \qquad \boldsymbol{A} \text{ is a } N \times N \text{ matrix}$$

We say $\boldsymbol{v}$ is an eigenvector with eigenvalue λ

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_M \geq 0$$

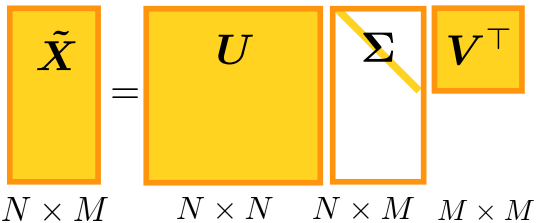
is known as the singular values

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_M \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

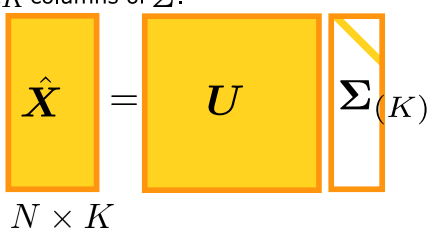
Principal component analysis (PCA)

(Karl Pearson, 1901)

1) Subtract the mean from each observation $\tilde{\boldsymbol{x}}_i = \boldsymbol{x}_i - \boldsymbol{m}$
2) Apply singular value decomposition (SVD) $\tilde{\boldsymbol{X}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$

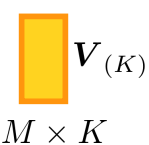


3) Select first K columns of $\boldsymbol{V}$ (the PCA projection operation) and first K columns of $\boldsymbol{\Sigma}$.



(PCA components or PCA projection of the data)

$$\boldsymbol{V}_{(K)} = \begin{bmatrix} \boldsymbol{v}_1 & \cdots & \boldsymbol{v}_K \end{bmatrix}$$



(PCA loadings)

Principal component analysis (PCA)

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$$

- Entries in the diagonal matrix $\boldsymbol{\Sigma}$ are called **singular values**
  - They are sorted (largest first)
  - Indicate how much variability is explained by the corresponding component
    - 1st component explains most of the variability
    - 2nd component explains most of the remaining variability
      Etc.

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_N \end{bmatrix}$$

越大越上面

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_N \geq 0$$

Explained Variance

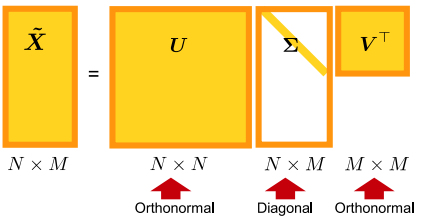
Recall that from SVD: $\tilde{\boldsymbol{X}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$

In the original space, the coordinates of $\tilde{\boldsymbol{X}}$ project onto the first K components are:

$$\boldsymbol{X}' = \underbrace{\boldsymbol{U}\boldsymbol{\Sigma}_{(K)}}_{\text{N * K}}\underbrace{\boldsymbol{V}^T_{(K)}}_{\text{K * D}}$$

We can measure how much variance is retained in the reconstruction $\boldsymbol{X}'$:

$$\text{Explained var.} = \frac{\|\boldsymbol{X}'\|_F^2}{\|\tilde{\boldsymbol{X}}\|_F^2}$$



$$\text{cov}(\tilde{\boldsymbol{X}}) = \tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}$$

$$\text{trace}(\boldsymbol{A}\boldsymbol{B}) = \text{trace}(\boldsymbol{B}\boldsymbol{A})$$

$$\|\tilde{\boldsymbol{X}}\|_F^2 = \text{trace}(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}) = \text{trace}(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T)$$

Explained Variance

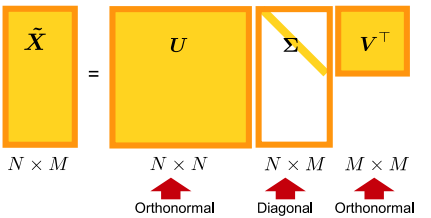
Recall that from SVD: $\tilde{\boldsymbol{X}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$

In the original space, the coordinates of $\tilde{\boldsymbol{X}}$ project onto the first K components are:

$$\boldsymbol{X}' = \boldsymbol{U}\boldsymbol{\Sigma}_{(K)}\boldsymbol{V}_{(K)}^T$$

We can measure how much variance is retained in the reconstruction $\boldsymbol{X}'$:

$$\text{Explained var.} = \frac{\|\boldsymbol{X}'\|_F^2}{\|\tilde{\boldsymbol{X}}\|_F^2} = \frac{\sum_{i=1}^K \sigma_i^2}{\sum_{i=1}^M \sigma_i^2}$$



$$\text{cov}(\tilde{\boldsymbol{X}}) = \tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}$$

$$\text{trace}(\boldsymbol{A}\boldsymbol{B}) = \text{trace}(\boldsymbol{B}\boldsymbol{A})$$

$$\begin{aligned}
\|\tilde{\boldsymbol{X}}\|_F^2 &= \text{trace}(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}) = \text{trace}(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T) \\
&= \text{trace}(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T)^T) \\
&= \text{trace}(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{\Sigma}^T\boldsymbol{U}^T) \\
&= \text{trace}(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\boldsymbol{U}^T) \\
&= \text{trace}(\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T) \\
&= \text{trace}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T) = \sum_i \sigma_i^2
\end{aligned}$$

Explained Variance

Recall that from SVD: $\tilde{\boldsymbol{X}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$

In the original space, the coordinates of $\tilde{\boldsymbol{X}}$ project onto the first K components are:

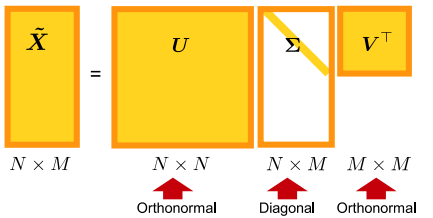
$$\boldsymbol{X}' = \boldsymbol{U}\boldsymbol{\Sigma}_{(K)}\boldsymbol{V}_{(K)}^T$$

We can measure how much variance is retained in the reconstruction $\boldsymbol{X}'$:

$$\text{Explained var.} = \frac{\|\boldsymbol{X}'\|_F^2}{\|\tilde{\boldsymbol{X}}\|_F^2} = \frac{\sum_{i=1}^K \sigma_i^2}{\sum_{i=1}^M \sigma_i^2}$$

Similarly, the fraction of explained variance for the i'th component is

$$\text{Explained var.} = \frac{\sigma_i^2}{\sum_{i=1}^M \sigma_i^2}$$



$$\text{cov}(\tilde{\boldsymbol{X}}) = \tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}$$

$$\text{trace}(\boldsymbol{A}\boldsymbol{B}) = \text{trace}(\boldsymbol{B}\boldsymbol{A})$$

$$\begin{aligned}
\|\tilde{\boldsymbol{X}}\|_F^2 &= \text{trace}(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}) = \text{trace}(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^T) \\
&= \text{trace}(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T)^T) \\
&= \text{trace}(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{\Sigma}^T\boldsymbol{U}^T) \\
&= \text{trace}(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\boldsymbol{U}^T) \\
&= \text{trace}(\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T) \\
&= \text{trace}(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T) = \sum_i \sigma_i^2
\end{aligned}$$

Quiz 2: PCA (Fall 2012)

No.	Attribute description	Abbrev.
x_1	Age (in years)	AGE
x_2	Gender (Female=0, Male=1)	GDR
x_3	Total Bilirubin	TB
x_4	Direct Bilirubin	DB
x_5	Alkaline Phosphotase	AP
x_6	Alamine Aminotransferase	AlA
x_7	Aspartate Aminotransferase	AsA
x_8	Total Protiens	TP
x_9	Albumin	AB
x_{10}	Albumin to Globulin ratio	A/G
y	0=No liver disease, 1=Liver disease	LD

Table 1: Attributes in a study on liver disease among Indians living in the north eastern part of Andhra Pradesh, India. (taken from http://archive.ics.uci.edu/ml/datasets/ILPD +%28Indian+Liver+Patient+Dataset%29). The data has 10 input attributes x_1–x_{10} and one output variable y which defines whether the subject considered has a liver disease ($y = 1$) or not ($y = 0$). x_3–x_9 are non-negative measurements giving the concentrations of various quantities measured in a blood test. x_{10} gives the ratio of Albumin to Globulin in the blood.

A PCA analysis is applied to the standardized data based on the attributes x_1–x_{10}. The squared Frobenius norm of the standardized data matrix $\boldsymbol{X}$ is given by $\|\boldsymbol{X}\|_F^2 = 5780.0$. The first four singular values are $\sigma_1 = 40.1$, $\sigma_2 = 34.2$, $\sigma_3 = 28.1$, and $\sigma_4 = 24.8$, Which of the following statements is *correct*?

A. The first PCA component accounts for more than 35 % of the variation.

B. The second PCA component accounts for more than 30 % of the variation.

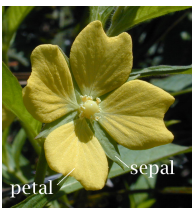
C. The first three PCA components account for less than 70 % of the variation in the data.

D. The fourth PCA component accounts for less than 10 % of the variation in the data.

E. Don't know.

use explained var to calculate

Fishers Iris Data

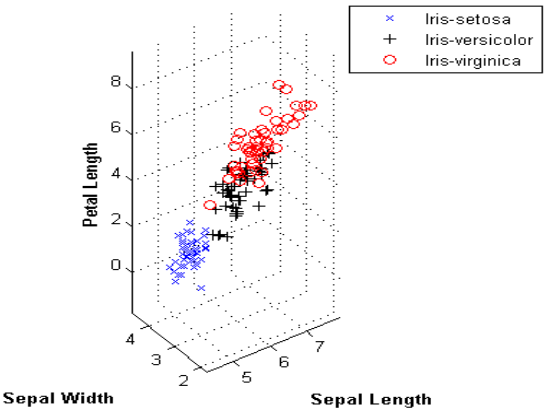


Three types of flowers:
Iris Setosa, Iris Versicolor, Iris Virginica

Flower ID	Attribute			
	Sepal Length	Sepal Width	Petal Length	Petal Width
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
.	.	.	.	.
.	.	.	.	.
150	5.9	3.0	5.1	1.8

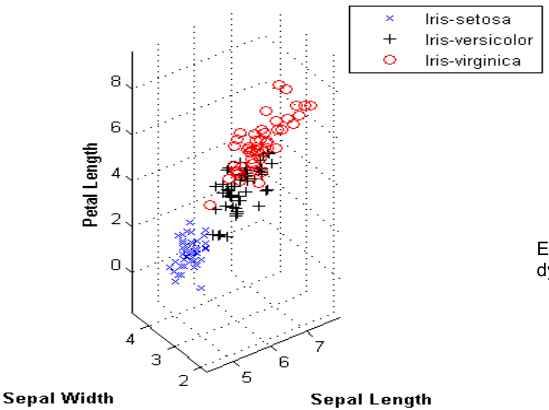
We will presently consider the first 3 attributes, i.e. Sepal length, Sepal Width and Petal Length.

3D scatter plot of Iris Data



What fraction of the total variation in the data will the first principal component account for?

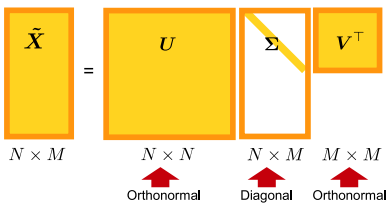
3D scatter plot of Iris Data



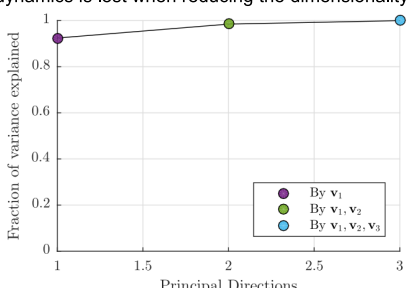
What fraction of the total variation in the data will the first principal component account for?

1) Subtract the mean
2) Apply singular value decomposition (SVD)

$$\tilde{\boldsymbol{X}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$$

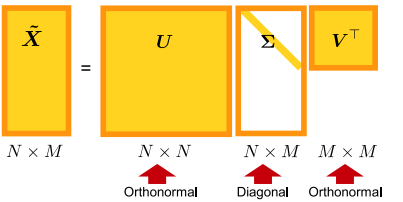


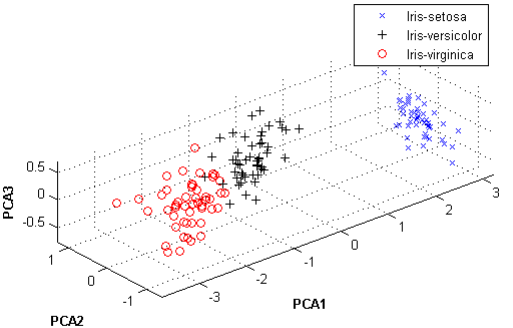
Evaluate the singular values to determine how much of the dynamics is lost when reducing the dimensionality



Visualization of the PCA projections of the data

$$\tilde{\boldsymbol{X}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$$

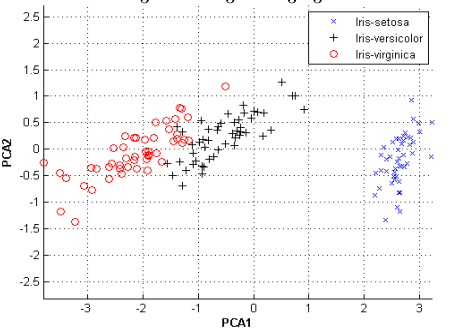




$$PCA1\colon\ \boldsymbol{b}_1 = \tilde{\boldsymbol{X}}\boldsymbol{v}_1 = \boldsymbol{u}_1\sigma_1$$

$$PCA2\colon\ \boldsymbol{b}_2 = \tilde{\boldsymbol{X}}\boldsymbol{v}_2 = \boldsymbol{u}_2\sigma_2$$

$$PCA3\colon\ \boldsymbol{b}_3 = \tilde{\boldsymbol{X}}\boldsymbol{v}_3 = \boldsymbol{u}_3\sigma_3$$



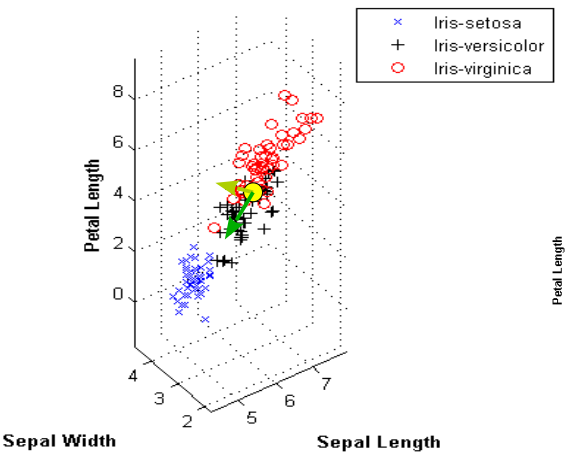
The principal directions V

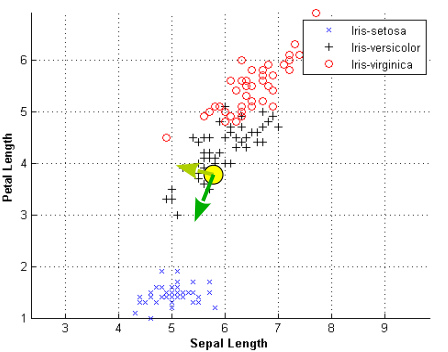
Sepal Length
Sepal Width
Petal Length

$$\boldsymbol{V} = \begin{bmatrix} -0.39 & -0.64 & -0.66 \\ 0.09 & -0.74 & 0.66 \\ -0.92 & 0.20 & 0.35 \end{bmatrix}$$

$$\mu = \begin{bmatrix} 5.8 \\ 3.1 \\ 3.8 \end{bmatrix}, \quad v_1 = \begin{bmatrix} -0.39 \\ 0.09 \\ -0.92 \end{bmatrix}, \quad v_2 = \begin{bmatrix} -0.64 \\ -0.74 \\ 0.20 \end{bmatrix}$$

Sepal Length
Sepal Width
Petal Length





Quiz 3: PCA Cont. (Fall 2012)

No.	Attribute description	Abbrev.
x_1	Age (in years)	AGE
x_2	Gender (Female=0, Male=1)	GDR
x_3	Total Bilirubin	TB
x_4	Direct Bilirubin	DB
x_5	Alkaline Phosphotase	AP
x_6	Alamine Aminotransferase	AlA
x_7	Aspartate Aminotransferase	AsA
x_8	Total Protiens	TP
x_9	Albumin	AB
x_{10}	Albumin to Globulin ratio	A/G
y	0=No liver disease, 1=Liver disease	LD

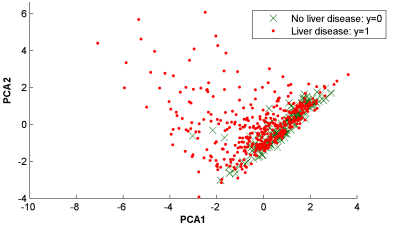


Figure 1: Principal component 1 (PCA1) plotted against principal component 2 (PCA2).

The first and second principal component directions of the liver-dataset are

$$\boldsymbol{v}_1 = \begin{bmatrix} -0.1404 \\ -0.1090 \\ -0.4115 \\ -0.4179 \\ -0.2468 \\ -0.2682 \\ -0.3009 \\ 0.2781 \\ 0.4375 \\ 0.3638 \end{bmatrix}, \quad \boldsymbol{v}_2 = \begin{bmatrix} -0.2859 \\ 0.0130 \\ 0.2510 \\ 0.2622 \\ 0.0525 \\ 0.4162 \\ 0.3927 \\ 0.4197 \\ 0.4323 \\ 0.3052 \end{bmatrix}.$$

In the figure, the data projected onto the first two principal components is plotted, and the colors indicate the presence of liver disease. Which of the following statements is *correct*?

A. Relatively high values of AGE, GDR, TB, DB, AP, AlA, and AsA and low values of TP, AB, and A/G will result in a positive projection onto the first principal component.

B. Relatively low values of the projection onto PCA1 and high values of the projection onto PCA2 indicates the subject does not have a liver disease.

C. PCA2 mainly discriminate between old subjects with low measurements of TB, DB, AlA, AsA, TP, AB, and A/G from young subjects with high values of TB, DB, AlA, AsA, TP, AB, and A/G.

D. The principal component directions are not guaranteed to be orthogonal to each other since the data has been standardized.

Visualization of hand written digits

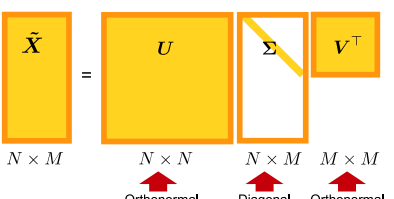
- Data matrix

$$\boldsymbol{X} = \begin{bmatrix} \text{---} & \boldsymbol{x}_1 & \text{---} \\ \text{---} & \boldsymbol{x}_2 & \text{---} \\ & \vdots & \\ \text{---} & \boldsymbol{x}_N & \text{---} \end{bmatrix}$$

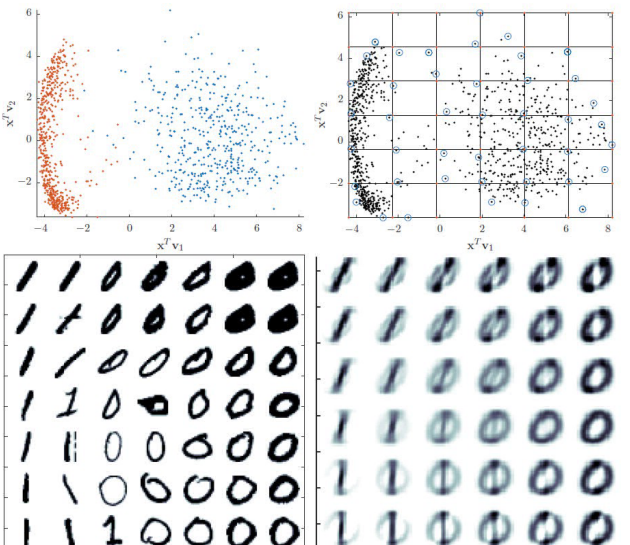
If each image is 28 x 28 pixels then X is a N x 784 matrix

- Principal component analysis

$$\tilde{\boldsymbol{X}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$$

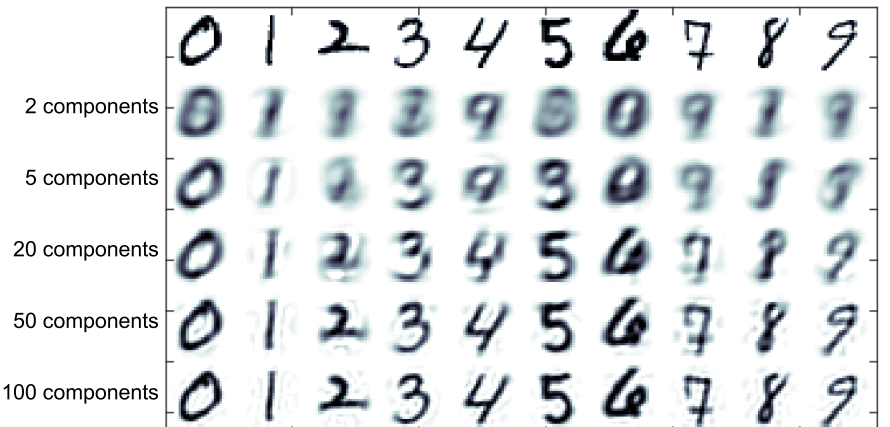


Visualization of hand written digits



PCA as compression

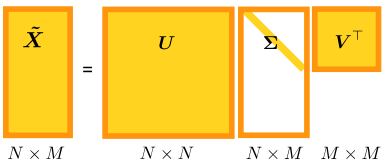
Only include a few components: $\hat{\boldsymbol{x}}_{\boldsymbol{i}} = \boldsymbol{V}\boldsymbol{b} + \boldsymbol{m}$ n=2,5,20,50,100



Data and domain driven feature extraction

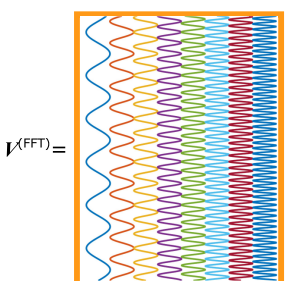
PCA is an example of a data driven approach for feature extraction

i.e., we define from data the features extracted in terms of the projections $\boldsymbol{V}^{(PCA)}$ that preserve most of the variance in the data



The fourier transform is an example of a domain driven approach for feature extraction

i.e., *in the analysis of sound good features are often to use spectral representations. These can be found by projecting the data using the so-called fourier transform matrix $\boldsymbol{V}^{(FFT)}$ where the components are defined as specific frequencies such that the projection of the data onto these frequencies defines the extend to which these frequencies are present in the data.*



Resources

http://www2.imm.dtu.dk Our online PCA demo which highlights key concepts of PCA such as the effect of normalization, variance explained, and much more (http://www2.imm.dtu.dk/courses/02450/DemoPCA.html)

https://arxiv.org A great and more in-depth tutorial on PCA (https://arxiv.org/abs/1404.1100)

https://www.3blue1brown.com An great, animated recap of linear algebra (https://www.3blue1brown.com/essence-of-linear-algebra-page/)