Prelims Statistics and Data Analysis: Lectures 11-16

Lecturer: Dino Sejdinovic
Course designed by: Jonathan Marchini

Department of Statistics
University of Oxford

Course websites:
`https://goo.gl/PTz2pF`, `https://courses.maths.ox.ac.uk/node/37534`

TT 2019

Statistical learning

So far, this course introduced parameter estimation in statistical models: maximum likelihood, confidence intervals, and linear regression. The rest of the course is an introduction to **statistical learning** framework and **unsupervised learning** in particular.

Statistical learning refers to a vast set of tools for understanding (typically large quantities of) data, and is closely related to **Machine Learning**, **Data Science** and **Artifical Intelligence**.

Examples of recent advances in AI which make use of **machine learning** models:
learning game strategies from sensory input, computer vision, machine translation, AlphaGO

Statistical learning

Massive amounts of data are being collected in many different fields.

Financial institutions, businesses, governments, hospitals, and universities are all interested in utilizing and making sense of data they collect.

Statistical learning

Massive amounts of data are being collected in many different fields.

Financial institutions, businesses, governments, hospitals, and universities are all interested in utilizing and making sense of data they collect.

The majority of mathematics students will go on to work in careers that involve carrying out or interpreting analysis of data.

Statistical learning

Massive amounts of data are being collected in many different fields.

Financial institutions, businesses, governments, hospitals, and universities are all interested in utilizing and making sense of data they collect.

The majority of mathematics students will go on to work in careers that involve carrying out or interpreting analysis of data.

This course leads onto several more advanced courses offered by the Department of Statistics, including *Part B Statistical Machine Learning* and *Part C Advanced Topics in Statistical Machine Learning*.

Supervised vs unsupervised learning

$$Y = \alpha + \sum_{i=1}^{p} \beta_i X_i + \epsilon, \qquad \epsilon \sim N(0, \sigma^2)$$

Supervised vs unsupervised learning

$$Y = \alpha + \sum_{i=1}^{p} \beta_i X_i + \epsilon, \qquad \epsilon \sim N(0, \sigma^2)$$

Linear regression is an example of **supervised learning.**

i.e. we build a model to predict a response variable Y using a set of p variables (or features) $X_1, \ldots, X_p$.

Typically we will have data on n observations.

Supervised vs unsupervised learning

In **unsupervised learning** we just have observations on the a set of variables $X_1, \ldots, X_p$, measured on n observations.

Supervised vs unsupervised learning

In **unsupervised learning** we just have observations on the a set of variables $X_1, \ldots, X_p$, measured on n observations.

Interest lies in looking for patterns and structure in the data, which is often large and high-dimensional.

Supervised vs unsupervised learning

In **unsupervised learning** we just have observations on the a set of variables $X_1, \ldots, X_p$, measured on *n* observations.

Interest lies in looking for patterns and structure in the data, which is often large and high-dimensional.

Relevant questions include

1. Can we find a way to visualize the data that is informative?

Supervised vs unsupervised learning

In **unsupervised learning** we just have observations on the a set of variables $X_1, \ldots, X_p$, measured on n observations.

Interest lies in looking for patterns and structure in the data, which is often large and high-dimensional.

Relevant questions include

1. Can we find a way to visualize the data that is informative?
2. Can we compress the dataset without losing any relevant information?

Supervised vs unsupervised learning

In **unsupervised learning** we just have observations on the a set of variables $X_1, \ldots, X_p$, measured on n observations.

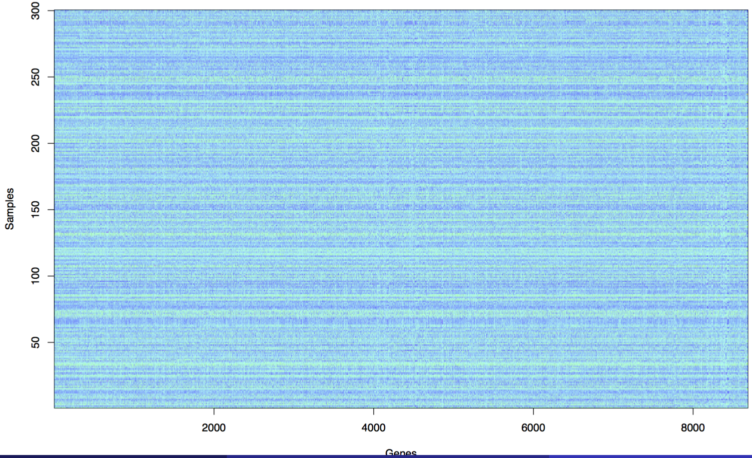
Interest lies in looking for patterns and structure in the data, which is often large and high-dimensional.

Relevant questions include

1. Can we find a way to visualize the data that is informative?
2. Can we compress the dataset without losing any relevant information?
3. Can we find separate subgroups (or clusters) of observations that describe the structure of the dataset?

Motivating example 1

300 cells each with measurements of activity of 8,686 genes
(click on image for 3D PCA projection)



Motivating example 2

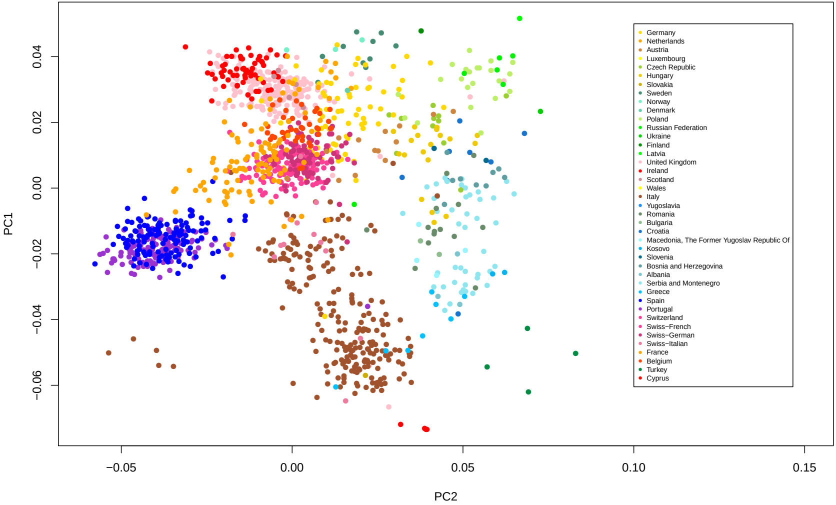
3,000 individuals from different European countries, each with measurements at $\sim$500,000 genes.

From the paper by Novembre et al. (2008) Nature 456:98-101

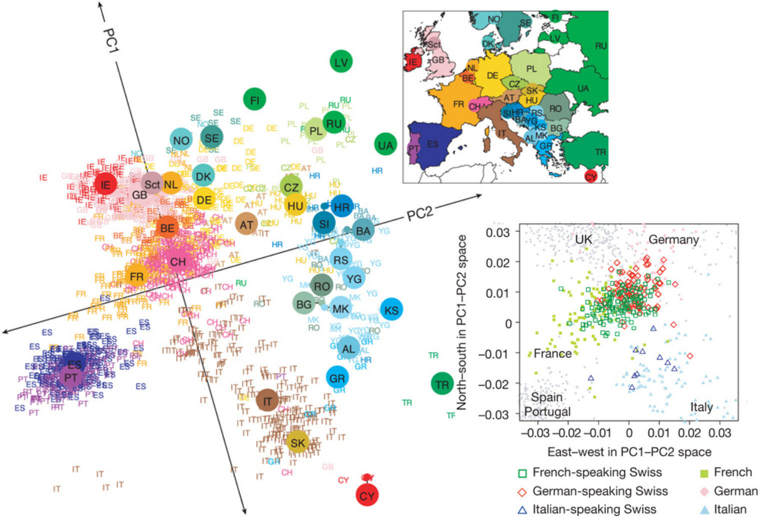
Scientific question

“not clear to what extent populations within continental regions exist as discrete genetic clusters versus as a genetic continuum, nor how precisely one can assign an individual to a geographic location on the basis of their genetic information alone."

Motivating example 2



Motivating example 2



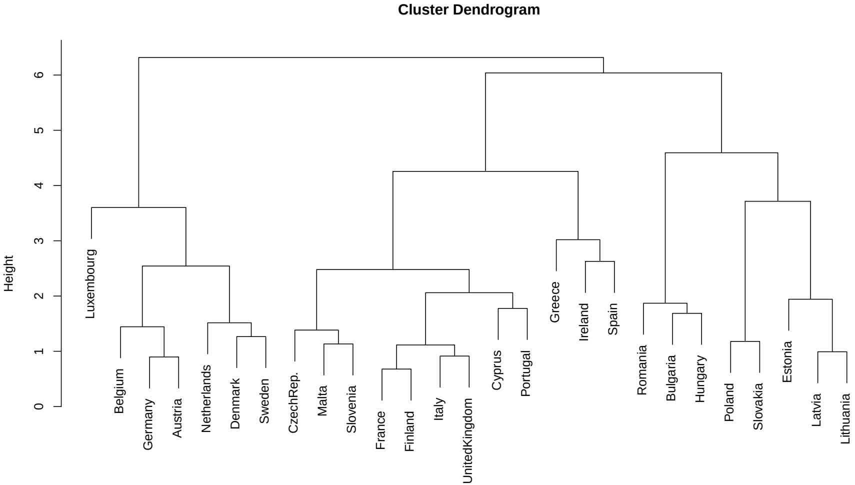
Genes mirror geography within Europe, Nature 2008

Motivating example 3

Economic indicators for 27 EU countries (data from 2012)

Country	CPI	UNE	INP	BOP	PRC	UN%
Belgium	116.03	4.77	125.59	908.60	6716.50	-1.60
Bulgaria	141.20	7.31	102.39	27.80	1094.70	3.50
CzechRep.	116.20	4.88	119.01	-277.90	2616.40	-0.60
Denmark	114.20	6.03	88.20	1156.40	7992.40	0.50
Germany	111.60	4.63	111.30	499.40	6774.60	-1.30
Estonia	135.08	9.71	111.50	153.40	2194.10	-7.70
Ireland	106.80	10.20	111.20	-166.50	6525.10	2.00
Greece	122.83	11.30	78.22	-764.10	5620.10	6.40
Spain	116.97	15.79	83.44	-280.80	4955.80	0.70
France	111.55	6.77	92.60	-337.10	6828.50	-0.90
Italy	115.00	5.05	87.80	-366.20	5996.60	-0.50
Cyprus	116.44	5.14	86.91	-1090.60	5310.30	-0.40
Latvia	144.47	12.11	110.39	42.30	1968.30	-3.60
Lithuania	135.08	11.47	114.50	-77.40	2130.60	-4.30
Luxembourg	118.19	3.14	85.51	2016.50	10051.60	-3.00
Hungary	134.66	6.77	115.10	156.20	1954.80	-0.10
Malta	117.65	4.15	101.65	359.40	3378.30	-0.60
Netherlands	111.17	3.23	103.80	1156.60	6046.00	-0.40
Austria	114.10	2.99	116.80	87.80	7045.50	-1.50
Poland	119.90	6.28	146.70	-74.80	2124.20	-1.00
Portugal	113.06	9.68	89.30	-613.40	4073.60	0.80
Romania	142.34	4.76	131.80	-128.70	1302.20	3.20
Slovenia	118.33	5.56	105.40	39.40	3528.30	1.80
Slovakia	117.17	9.19	156.30	16.00	2515.30	-2.10
Finland	114.60	5.92	101.00	-503.70	7198.80	-1.30
Sweden	112.71	6.10	100.50	1079.10	7476.70	-2.30
UnitedKingdom	120.90	6.11	90.36	-24.30	6843.90	-0.80

Motivating example 3



Data visualization

Campbell (1974) studied rock crabs of the genus leptograpsus. One species, L. variegatus, had been split into two new species according to their colour: orange and blue. Preserved specimens lose their colour, so it was hoped that morphological differences would enable museum material to be classified. Data are available on 50 specimens of each sex of each species.

Each specimen has measurements on:

- the width of the frontal lobe (FL),
- the rear width (RW),
- the length along the carapace midline (CL),
- the maximum width (CW) of the carapace,
- the body depth (BD) in mm.

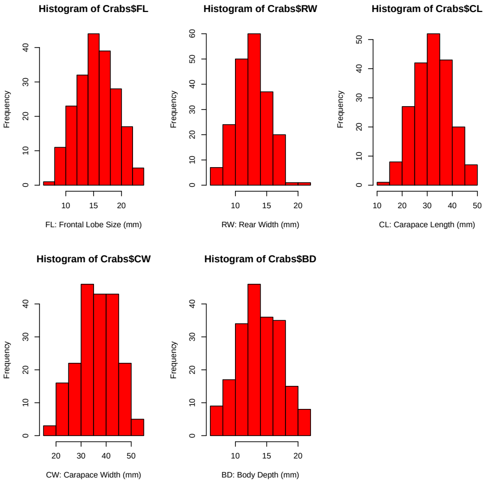
So the data matrix **X** has dimensions 200 × 5.

Crabs Data

```
      FL   RW   CL   CW   BD
1    8.1  6.7 16.1 19.0  7.0
2    8.8  7.7 18.1 20.8  7.4
3    9.2  7.8 19.0 22.4  7.7
4    9.6  7.9 20.1 23.1  8.2
5    9.8  8.0 20.3 23.0  8.2
6   10.8  9.0 23.0 26.5  9.8
7   11.1  9.9 23.8 27.1  9.8
8   11.6  9.1 24.5 28.4 10.4
9   11.8  9.6 24.2 27.8  9.7
10  11.8 10.5 25.2 29.3 10.3
11  12.2 10.8 27.3 31.6 10.9
12  12.3 11.0 26.8 31.5 11.4
13  12.6 10.0 27.7 31.7 11.4
14  12.8 10.2 27.2 31.8 10.9
15  12.8 10.9 27.4 31.5 11.0
16  12.9 11.0 26.8 30.9 11.4
17  13.1 10.6 28.2 32.3 11.0
18  13.1 10.9 28.3 32.4 11.2
19  13.3 11.1 27.8 32.3 11.3
20  13.9 11.1 29.2 33.3 12.1
```

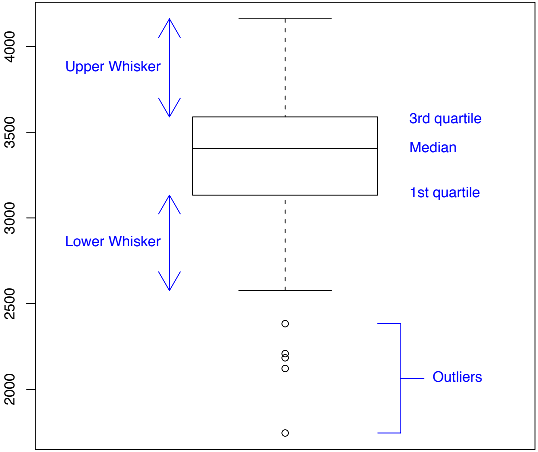
Histograms

A histogram is one of the simplest ways of visualizing the data from a single variable.



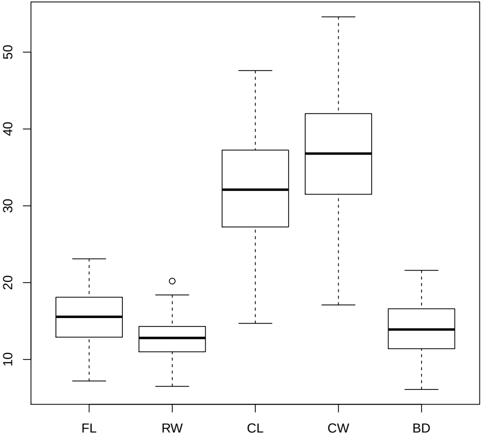
Boxplots

A Box Plot (sometimes called a Box-and-Whisker Plot) is a relatively sophisticated plot that summarises the distribution of a given variable.



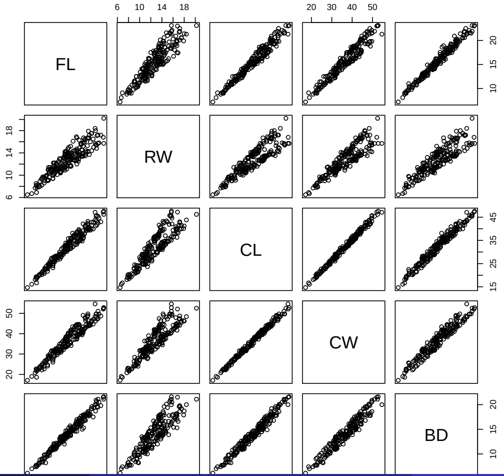
Boxplots

Boxplots of the crabs dataset



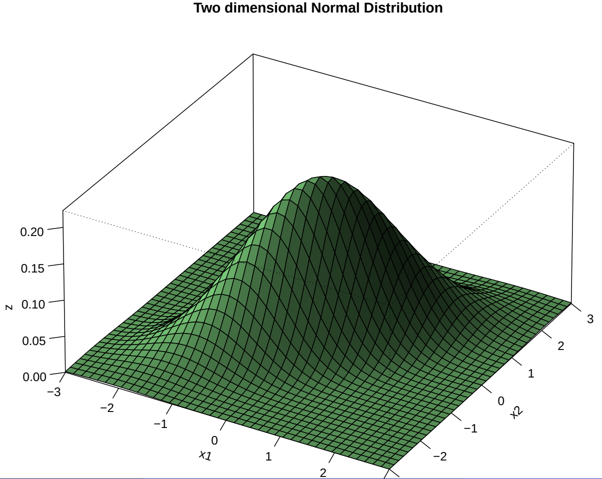
Pairs plots

Plotting pairs of variables together in a scatter plot can be helpful to see how variables co-vary.



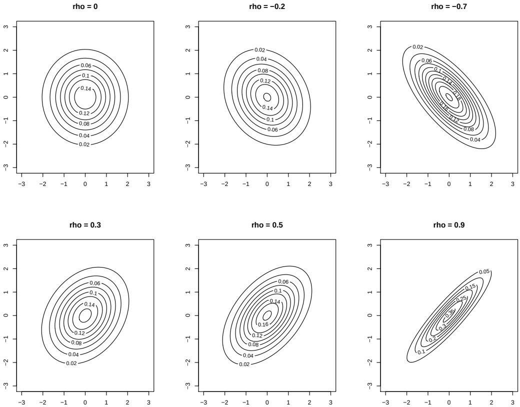
Multivariate Normal Density

$X \sim N_2(\mu, \boldsymbol{\Sigma})$ with $\mu = (0, 0)^T$ and $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}$



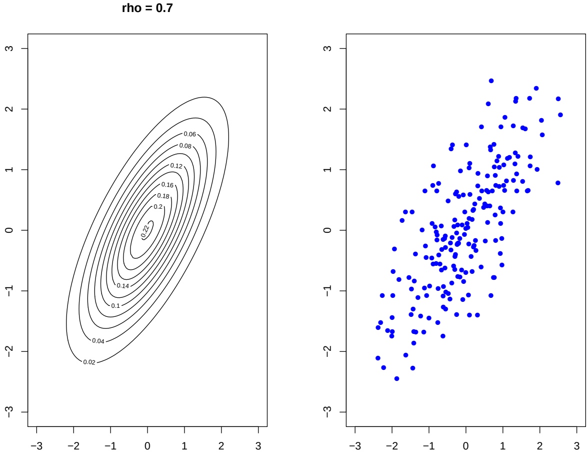
Multivariate Normal Density

$X \sim N_2(\mu, \boldsymbol{\Sigma})$ with $\mu = (0,0)^T$ and $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$



Multivariate Normal Density

Density (left) and Simulated Data from an MVN (right)



Sample covariance matrix

On the Crabs data the sample covariance matrix is

$$\mathbf{S} = \begin{array}{c|ccccc} & FL & RW & CL & CW & BD \\ \hline FL & 12.21 & 8.15 & 24.35 & 26.55 & 11.82 \\ RW & 8.15 & 6.62 & 16.35 & 18.23 & 7.83 \\ CL & 24.35 & 16.35 & 50.67 & 55.76 & 23.97 \\ CW & 26.55 & 18.23 & 55.76 & 61.96 & 26.09 \\ BD & 11.82 & 7.83 & 23.97 & 26.09 & 11.72 \end{array} .$$

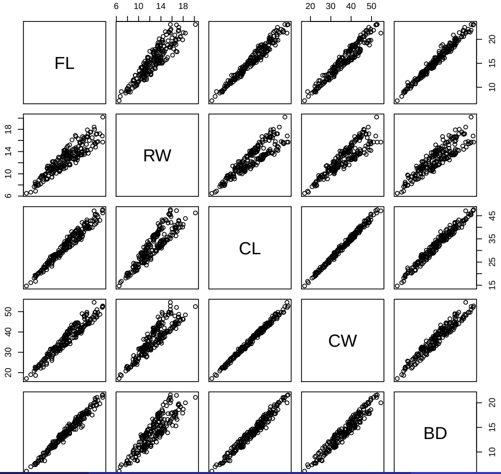
Sample correlation matrix

On the Crabs data the sample correlation matrix is

$$\mathbf{R} = \begin{array}{c|ccccc} & FL & RW & CL & CW & BD \\ \hline FL & 1.00 & 0.91 & 0.98 & 0.96 & 0.99 \\ RW & 0.91 & 1.00 & 0.89 & 0.90 & 0.89 \\ CL & 0.98 & 0.89 & 1.00 & 1.00 & 0.98 \\ CW & 0.96 & 0.90 & 1.00 & 1.00 & 0.97 \\ BD & 0.99 & 0.89 & 0.98 & 0.97 & 1.00 \end{array} .$$

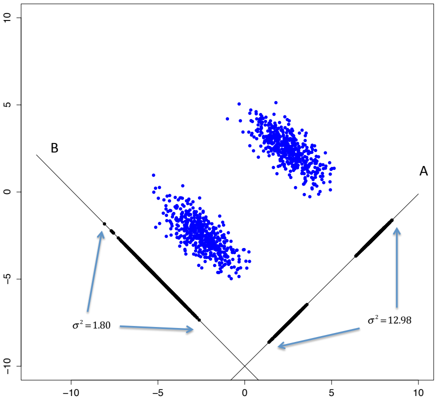
Pairs plots

Plotting pairs of variables together in a scatter plot can be helpful to see how variables co-vary.

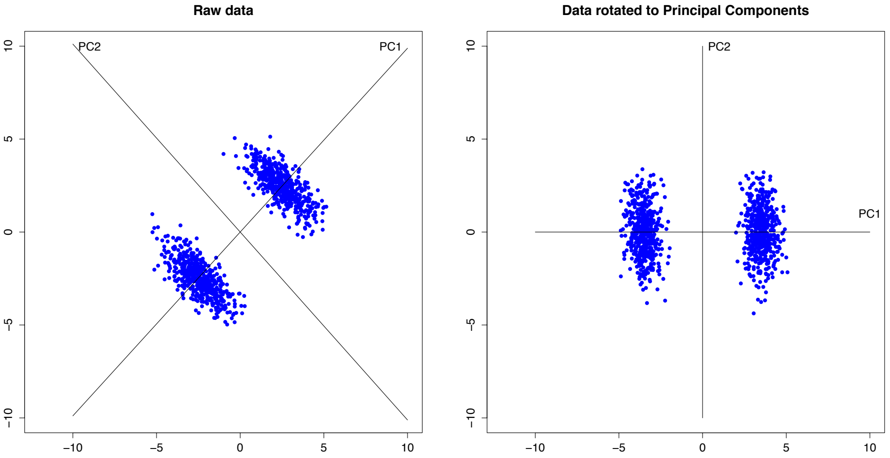


PCA

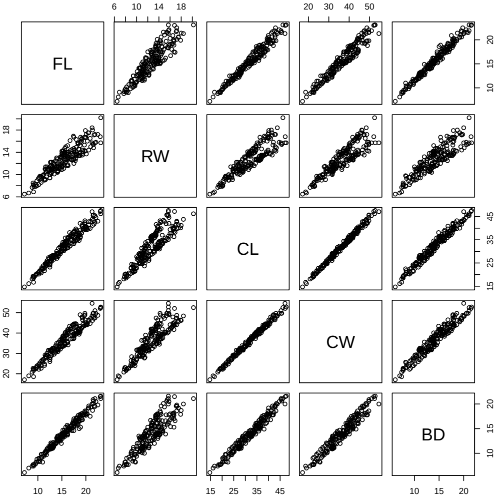
Projections that maximize variance can find useful structure in datasets. Projecting onto A separates clusters and has higher variance that projecting onto B.



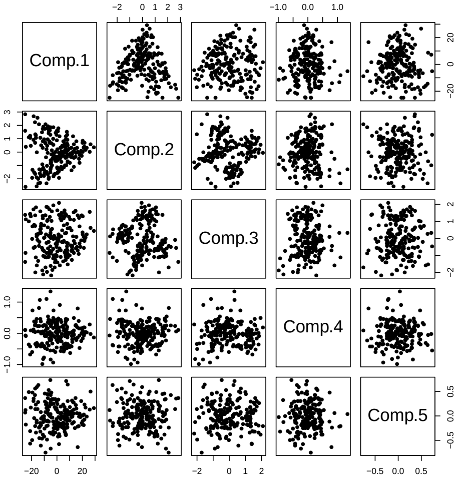
PCA


Raw data
PC2
PC1
Data rotated to Principal Components
PC2
PC1
10
5
0
−5
−10
−10
−5
0
5
10

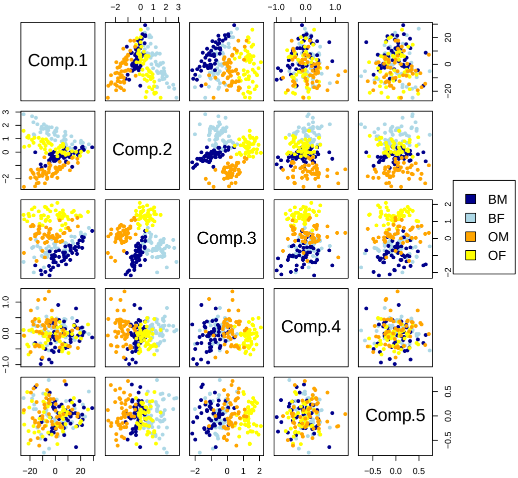
Pairs plots of Crabs dataset



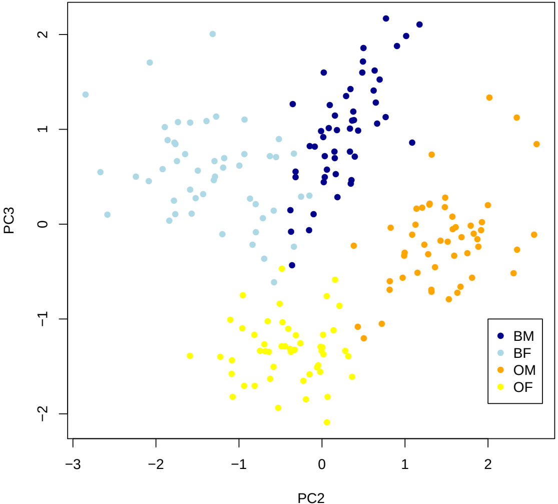
Pairs plot of PCA of Crabs dataset



Pairs plot of PCA of Crabs dataset



PC2 vs PC3 for the Crabs dataset



Loadings for the Crabs dataset

$$\mathbf{V} = \begin{array}{c|ccccc} & PC1 & PC2 & PC3 & PC4 & PC5 \\ \hline FL & \mathbf{0.28} & \mathbf{0.32} & \mathbf{-0.50} & 0.73 & 0.12 \\ RW & \mathbf{0.19} & \mathbf{0.86} & \mathbf{0.41} & -0.14 & -0.14 \\ CL & \mathbf{0.59} & \mathbf{-0.19} & \mathbf{-0.17} & -0.14 & -0.74 \\ CW & \mathbf{0.66} & \mathbf{-0.28} & \mathbf{0.49} & 0.12 & 0.47 \\ BD & \mathbf{0.28} & \mathbf{0.15} & \mathbf{-0.54} & -0.63 & 0.43 \end{array}$$

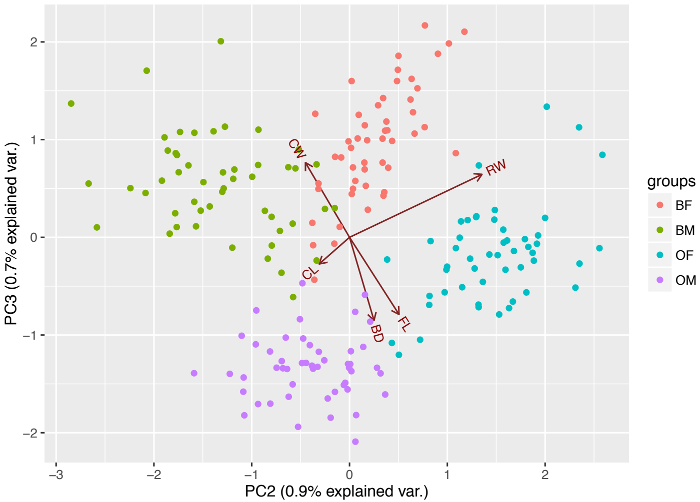
So for example, this means that the first, second and third PCs are

$$Z_1 = \mathbf{0.28}FL + \mathbf{0.19}RW + \mathbf{0.59}CL + \mathbf{0.66}CW + \mathbf{0.28}BD$$

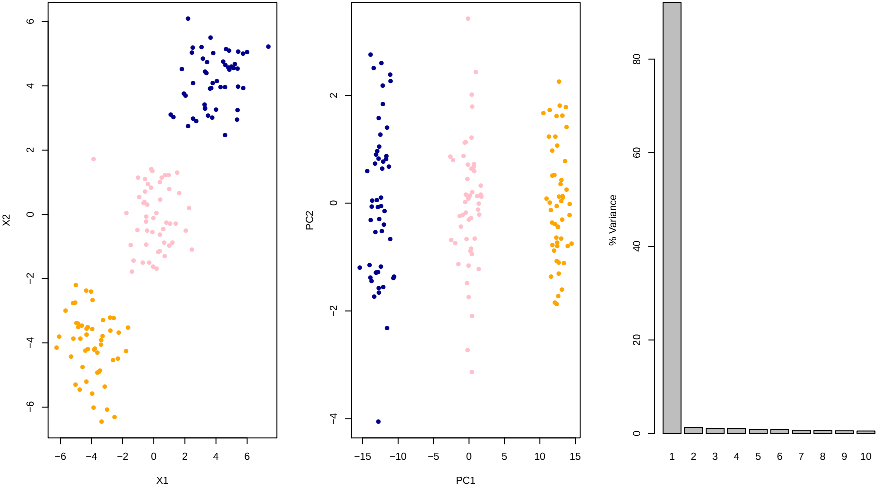
$$Z_2 = \mathbf{0.32}FL + \mathbf{0.86}RW\mathbf{-0.19}CL\mathbf{-0.28}CW + \mathbf{0.15}BD$$

$$Z_3 = \mathbf{-0.50}FL + \mathbf{0.41}RW\mathbf{-0.17}CL + \mathbf{0.49}CW\mathbf{-0.54}BD$$

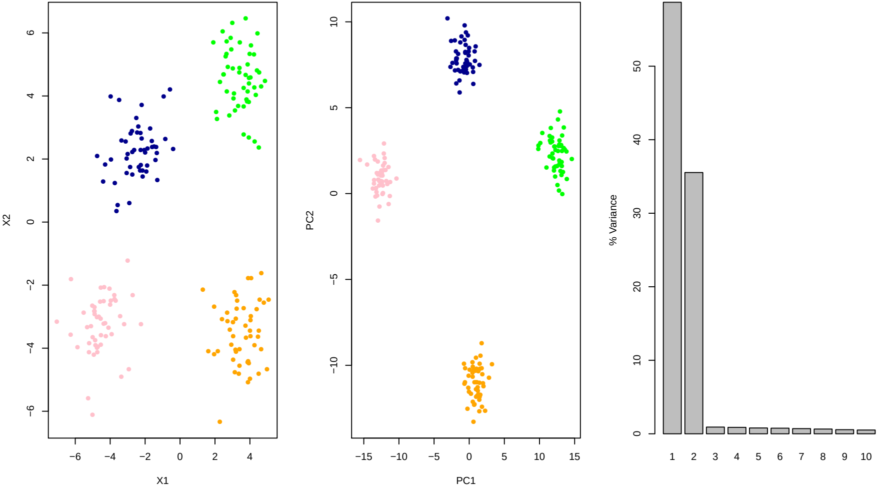
BiPlot of PCs 2 and 3 for the Crabs dataset.



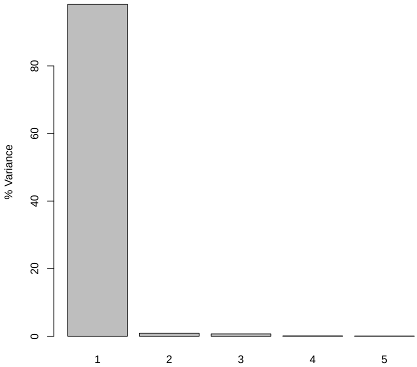
Scree plot example 1



Scree plot example 2



Scree plot for Crabs dataset



EU indicators dataset

Economic indicators for 27 EU countries

Country	CPI	UNE	INP	BOP	PRC	UN%
Belgium	116.03	4.77	125.59	908.60	6716.50	-1.60
Bulgaria	141.20	7.31	102.39	27.80	1094.70	3.50
CzechRep.	116.20	4.88	119.01	-277.90	2616.40	-0.60
Denmark	114.20	6.03	88.20	1156.40	7992.40	0.50
Germany	111.60	4.63	111.30	499.40	6774.60	-1.30
Estonia	135.08	9.71	111.50	153.40	2194.10	-7.70
Ireland	106.80	10.20	111.20	-166.50	6525.10	2.00
Greece	122.83	11.30	78.22	-764.10	5620.10	6.40
Spain	116.97	15.79	83.44	-280.80	4955.80	0.70
France	111.55	6.77	92.60	-337.10	6828.50	-0.90
Italy	115.00	5.05	87.80	-366.20	5996.60	-0.50
Cyprus	116.44	5.14	86.91	-1090.60	5310.30	-0.40
Latvia	144.47	12.11	110.39	42.30	1968.30	-3.60
Lithuania	135.08	11.47	114.50	-77.40	2130.60	-4.30
Luxembourg	118.19	3.14	85.51	2016.50	10051.60	-3.00
Hungary	134.66	6.77	115.10	156.20	1954.80	-0.10
Malta	117.65	4.15	101.65	359.40	3378.30	-0.60
Netherlands	111.17	3.23	103.80	1156.60	6046.00	-0.40
Austria	114.10	2.99	116.80	87.80	7045.50	-1.50
Poland	119.90	6.28	146.70	-74.80	2124.20	-1.00
Portugal	113.06	9.68	89.30	-613.40	4073.60	0.80
Romania	142.34	4.76	131.80	-128.70	1302.20	3.20
Slovenia	118.33	5.56	105.40	39.40	3528.30	1.80
Slovakia	117.17	9.19	156.30	16.00	2515.30	-2.10
Finland	114.60	5.92	101.00	-503.70	7198.80	-1.30
Sweden	112.71	6.10	100.50	1079.10	7476.70	-2.30
UnitedKingdom	120.90	6.11	90.36	-24.30	6843.90	-0.80
Variance	111.66	9.95	357.27	450057.15	5992520.48	7.12

PCA on covariance vs correlation matrix

When using the covariance matrix $\boldsymbol{S}$ the loadings of the 1st and 2nd PCs are

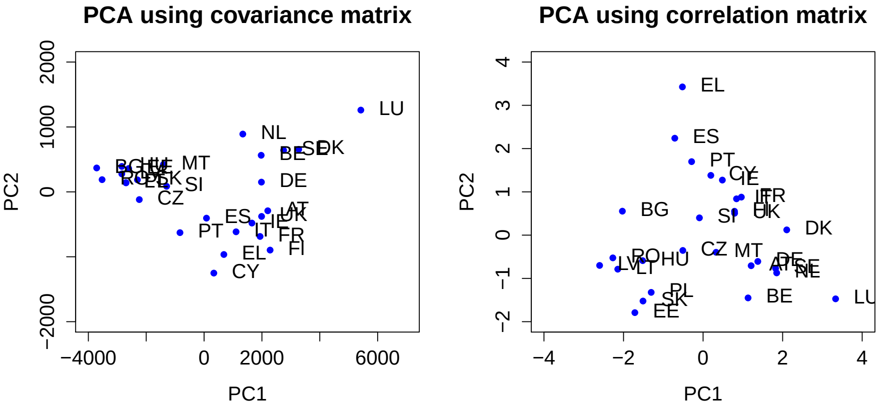
$$\begin{aligned} Z_1 &= -0.003CPI - 0.0004UNE - 0.0039INP + 0.121BOP + 0.993PRC - 0. \\ Z_2 &= 0.004CPI - 0.001UNE + 0.009INP + 0.992BOP - 0.121PRC - 0.0014 \end{aligned}$$

so it is the variables BOP and PRC that are dominating these PCs.
When using the correlation matrix $\boldsymbol{R}$ the loadings of the 1st and 2nd PCs are

$$\begin{aligned} Z_1 &= -0.51CPI - 0.37UNE - 0.29INP + 0.36BOP - 0.62PRC - 0.02UN\% \\ Z_2 &= -0.17CPI + 0.34UNE - 0.53INP - 0.49BOP + 0.12PRC + 0.56UN\% \end{aligned}$$

and the weightings for the variables are quite different.

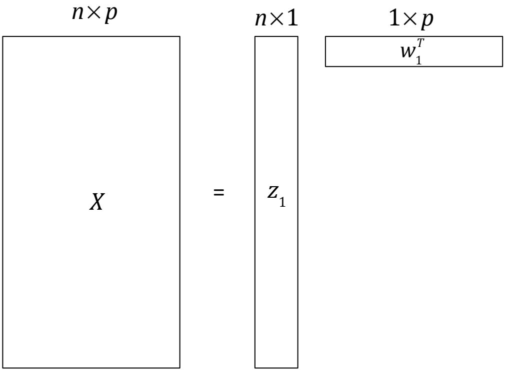
PCA for EU indicators dataset



PCA using covariance matrix

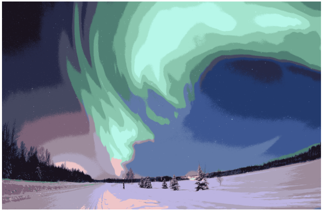
PCA using correlation matrix

Rank-1 approximation to the data matrix



Clustering

- Many datasets consist of multiple heterogeneous subsets.
- **Cluster analysis**: Given an unlabelled data, want algorithms that automatically group the datapoints into coherent subsets/clusters. Examples:
  - market segmentation of shoppers based on browsing and purchase histories
  - different types of cancer based on the gene expression measurements
  - discovering communities in social networks
  - image segmentation



The aim of clustering

- Clustering aims to group similar items together *and* to place separate dissimilar items into different groups
- Two objectives can contradict each other (similarity is not a transitive relation, while being in the same cluster is an equivalence relation)
- Notion of similarity/dissimilarity between data items is central: many ways to define and the choice will depend on the dataset being analyzed and dictated by domain specific knowledge
- *Partition-based* clustering: one divides n data items into K clusters $C_1, \ldots, C_K$ where for all $k, k' \in \{1, \ldots, K\}$,

$$C_k \subset \{1, \ldots, n\}, \quad C_k \cap C_{k'} = \emptyset \ \ \forall k \neq k', \quad \bigcup_{k=1}^{K} C_k = \{1, \ldots, n\}.$$

Within-cluster deviance

Goal: divide data items into a *pre-assigned number K of clusters* $C_1, \ldots, C_K$ where for all $k, k' \in \{1, \ldots, K\}$,

$$C_k \subset \{1, \ldots, n\}\,, \quad C_k \cap C_{k'} = \emptyset \;\; \forall k \neq k', \quad \bigcup_{k=1}^{K} C_k = \{1, \ldots, n\}\,.$$

Define $W(C_k)$ to be a measure of how different the observations are within cluster k, the most common choice is to use squared distances:

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \|x_i - x_{i'}\|_2^2 = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

Problem sheet:

$$\frac{1}{|C_k|} \sum_{i,i' \in C_k} \|x_i - x_{i'}\|_2^2 = 2 \sum_{i \in C_k} \|x_i - \mu_k\|_2^2, \tag{1}$$

where $\mu_k = \frac{1}{|C_k|} \sum_{i \in C_k} x_i$.

Within-cluster deviance

Each cluster is represented using a *prototype* or *cluster centroid* μ_k.
Within-cluster deviance:

$$W(C_k, \mu_k) = \sum_{i \in C_k} \|x_i - \mu_k\|_2^2 = \sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - \mu_{kj})^2$$

The overall quality of the clustering is given by the total within-cluster deviance:

$$W = \sum_{k=1}^{K} W(C_k, \mu_k) = \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - \mu_{kj})^2$$

K-means

$$W = \sum_{k=1}^{K} \sum_{i \in C_k} \|x_i - \mu_k\|_2^2 = \sum_{i=1}^{n} \|x_i - \mu_{c_i}\|_2^2$$

where $c_i = k$ if and only if $i \in C_k$.

- Given partition $\{C_k\}$, we can find the optimal prototypes easily by differentiating W with respect to μ_k:

$$\frac{\partial W}{\partial \mu_k} = 2 \sum_{i \in C_k} (x_i - \mu_k) = 0 \qquad \Rightarrow \mu_k = \frac{1}{|C_k|} \sum_{i \in C_k} x_i$$

- Given prototypes, we can easily find the optimal partition by assigning each data point to the closest cluster prototype:

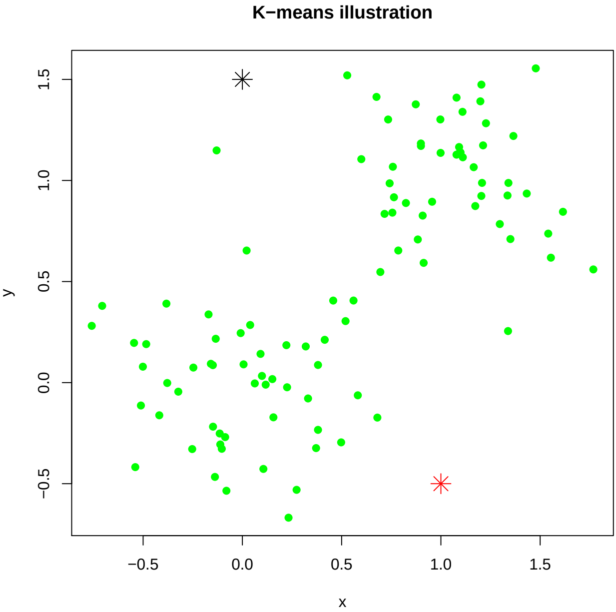
$$c_i = \text{argmin}_k \|x_i - \mu_k\|_2^2$$

But joint minimization over both is computationally difficult.

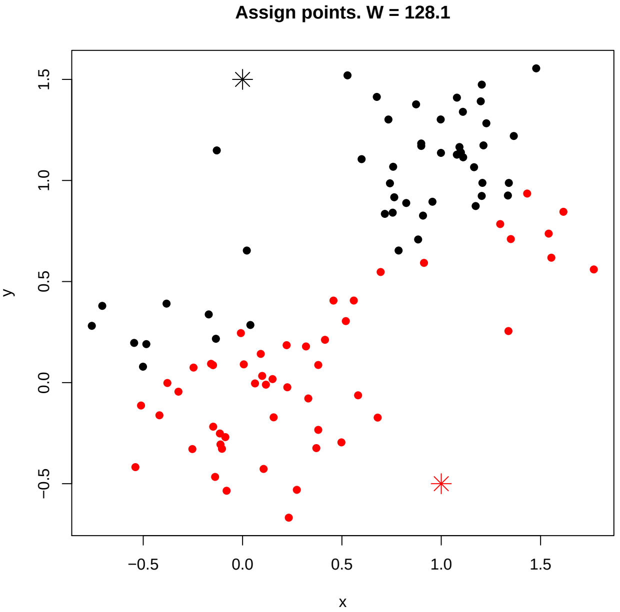
K-means

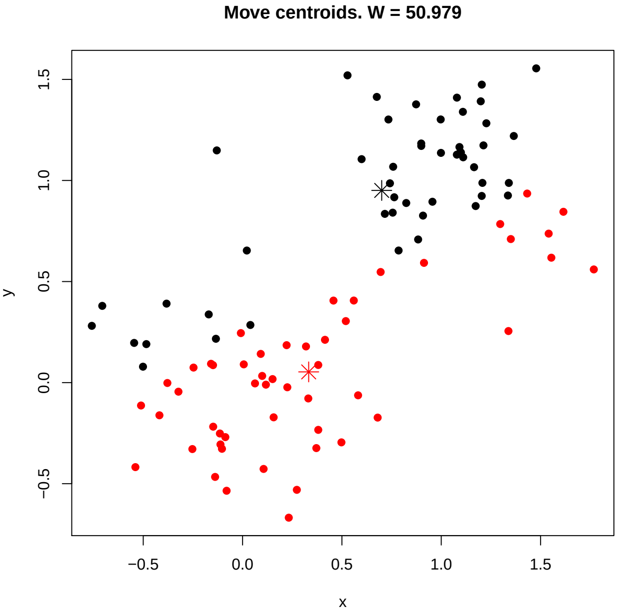
The K-means algorithm returns a *local optimum* of the objective function W, using iterative and alternating minimization.

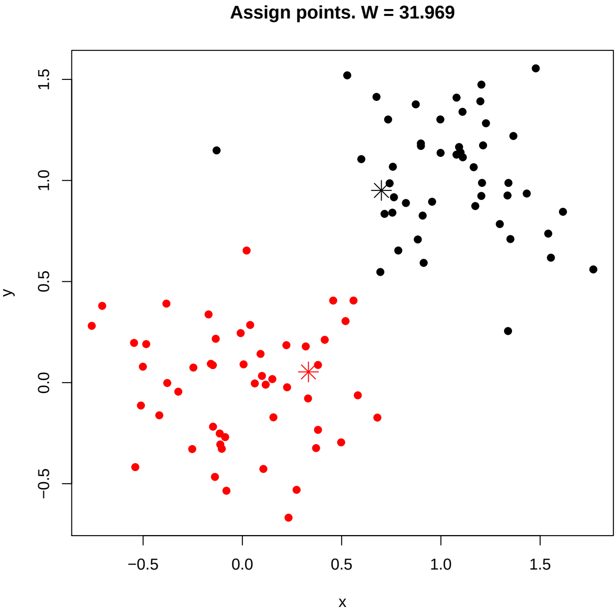
1. Randomly initialize K cluster centroids $\mu_1, \ldots, \mu_K$.
2. *Cluster assignment:* For each $i = 1, \ldots, n$, assign each x_i to the cluster with the nearest centroid,
$$c_i := \text{argmin}_k \|x_i - \mu_k\|_2^2$$
Set $C_k := \{i : c_i = k\}$ for each k.
3. *Move centroids:* Set $\mu_1, \ldots, \mu_K$ to the averages of the new clusters:
$$\mu_k := \frac{1}{|C_k|} \sum_{i \in C_k} x_i$$
4. Repeat steps 2-3 until convergence.
5. Return the partition $\{C_1, \ldots, C_K\}$ and means $\mu_1, \ldots, \mu_K$.

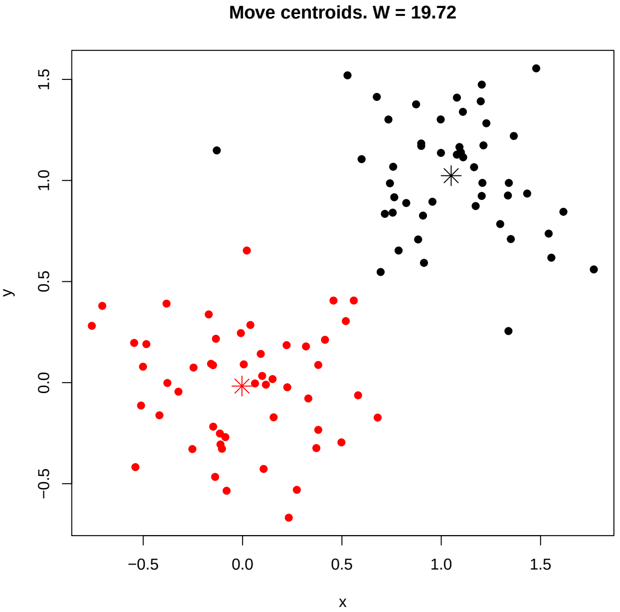


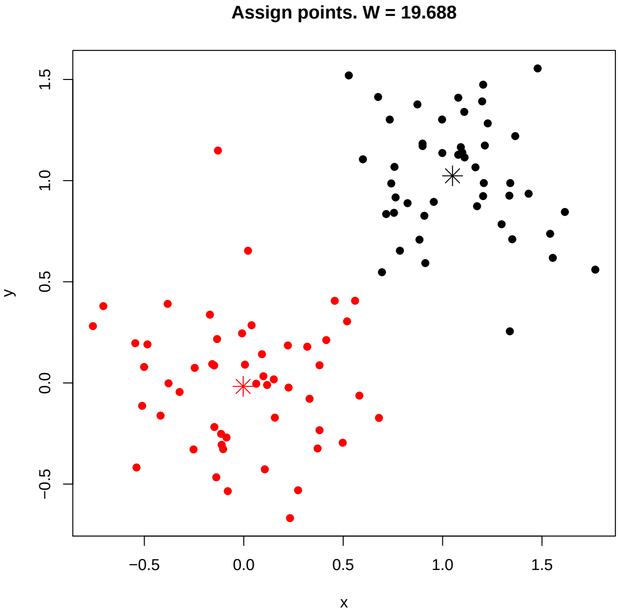
K−means illustration
y
−0.5
0.0
0.5
1.0
1.5
x
−0.5
0.0
0.5
1.0
1.5

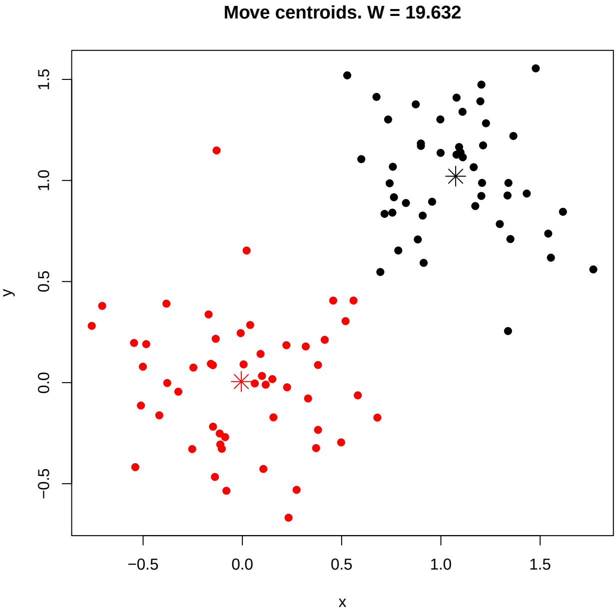

Assign points. W = 128.1
y
1.5
1.0
0.5
0.0
−0.5
x
−0.5
0.0
0.5
1.0
1.5


Move centroids. W = 50.979
y
x
−0.5
0.0
0.5
1.0
1.5
−0.5
0.0
0.5
1.0
1.5


Assign points. W = 31.969
y
1.5
1.0
0.5
0.0
−0.5
−0.5
0.0
0.5
1.0
1.5
x

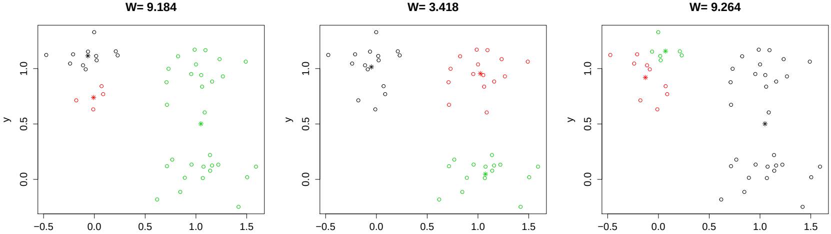

Move centroids. W = 19.72
y
x
1.5
1.0
0.5
0.0
−0.5
−0.5
0.0
0.5
1.0
1.5


Assign points. W = 19.688
y
x
−0.5
0.0
0.5
1.0
1.5
−0.5
0.0
0.5
1.0
1.5

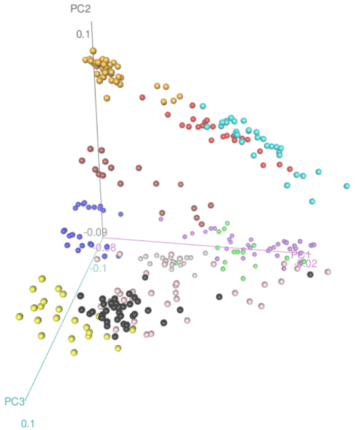

Move centroids. W = 19.632
y
1.5
1.0
0.5
0.0
−0.5
−0.5
0.0
0.5
1.0
1.5
x

K-means

- *The algorithm stops in a finite number of iterations.* Between steps 2 and 3, *W* either stays constant or it decreases, this implies that we never revisit the same partition. As there are only finitely many partitions, the number of iterations cannot exceed this.
- *The K-means algorithm need not converge to global optimum.* K-means can get stuck at suboptimal configurations and the result depends on the starting configuration. Typically perform a number of runs from different initial values, and pick the end result with minimum *W*.



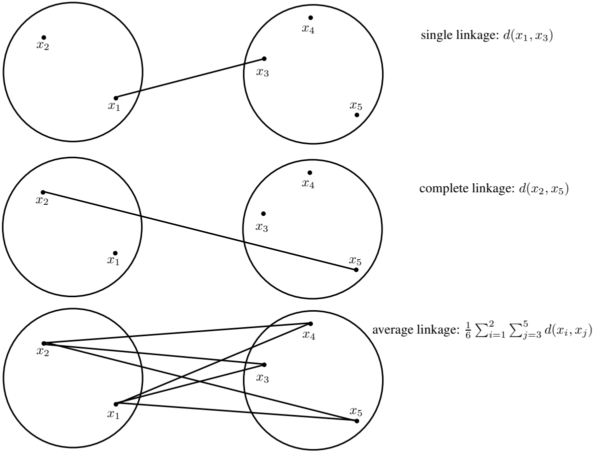
k-means clustering - single cell dataset



http://www.stats.ox.ac.uk/~sejdinov/teaching/movie.gif

Agglomerative Clustering

Iteratively join pairs of observations together to form clusters.
To join clusters C_i and C_j into larger clusters, we need a way to measure the dissimilarity $D(C_i, C_j)$ between them.



single linkage: $d(x_1, x_3)$

complete linkage: $d(x_2, x_5)$

average linkage: $\frac{1}{6}\sum_{i=1}^{2}\sum_{j=3}^{5} d(x_i, x_j)$

Measuring Dissimilarity Between Clusters

To join clusters C_i and C_j into super-clusters, we need a way to measure the dissimilarity $D(C_i, C_j)$ between them.

(a) *Single Linkage*: elongated, loosely connected clusters

$$D(C_i, C_j) = \min_{x,y} \left(d(x, y) | x \in C_i, y \in C_j \right)$$

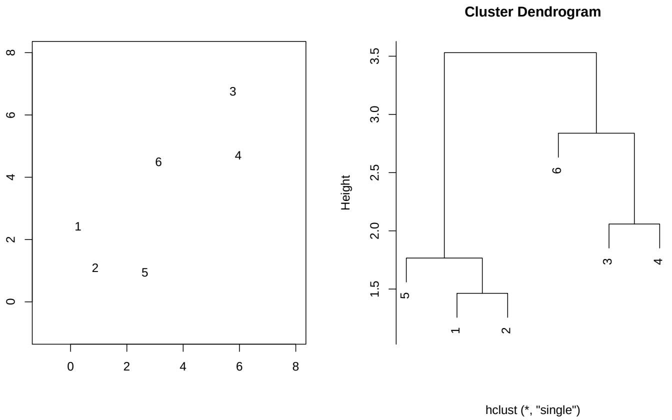
(b) *Complete Linkage*: compact clusters, relatively similar objects can remain separated at high levels

$$D(C_i, C_j) = \max_{x,y} \left(d(x, y) | x \in C_i, y \in C_j \right)$$

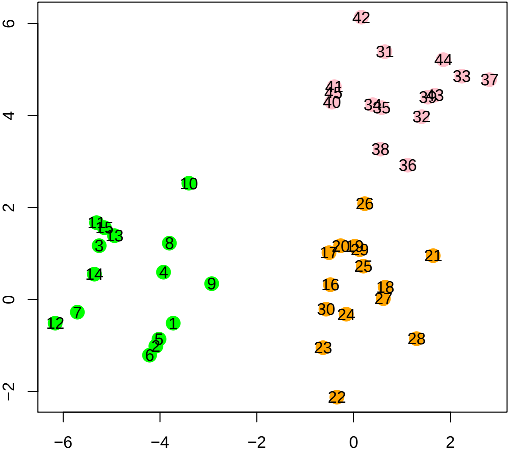
(c) *Average Linkage*: tries to balance the two above, but affected by the scale of dissimilarities

$$D(C_i, C_j) = \mathrm{avg}_{x,y} \left(d(x, y) | x \in C_i, y \in C_j \right)$$

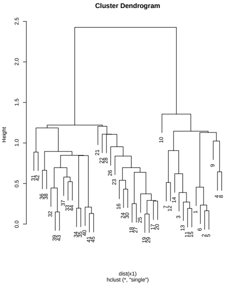
hierarchical clustering - example 1

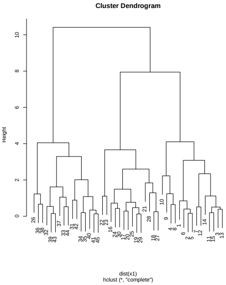


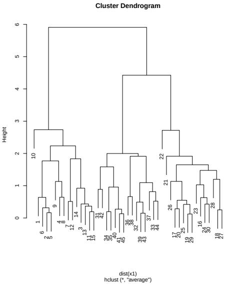
hierarchical clustering - example 2



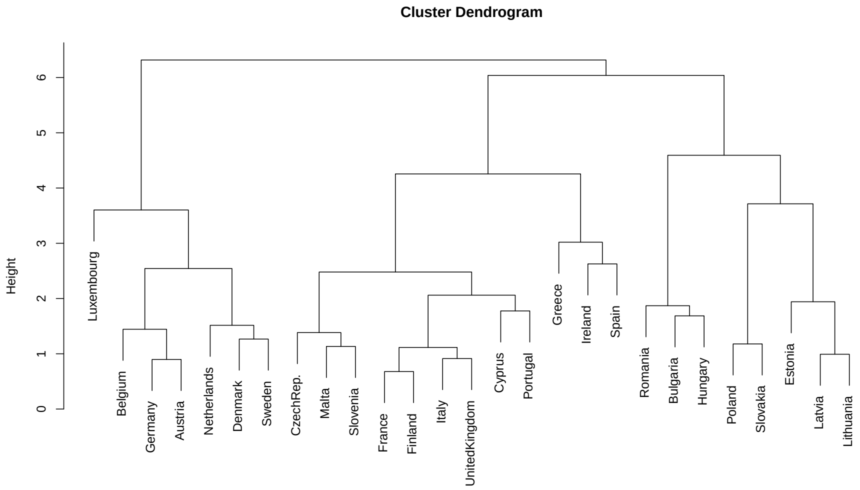
hierarchical clustering - example 2







hierarchical clustering - EU indicators



hierarchical clustering - extracting clusters