Hierarchical Clustering

Wolfgang Karl Härdle
Elizaveta Zinovyeva

Ladislaus von Bortkiewicz Professor of Statistics
Humboldt-Universität zu Berlin
BRC Blockchain Research Center
lvb.wiwi.hu-berlin.de
Charles University, WISE XMU, NCTU 玉山學者

Marketing

- Customer Segmentation
  - Help marketers to id customer groups and subsequently develop different targeting programs

Image Segmentation

- Drone id's landscape

Original

2 Clusters

4 Clusters

8 Clusters

Text Profiles

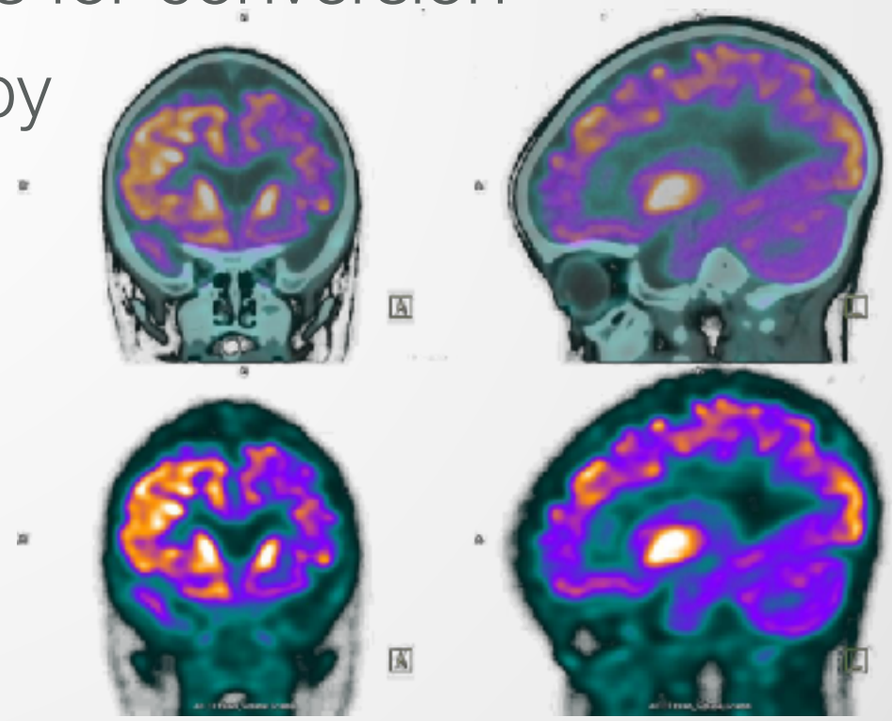
- Information Retrieval
  - Document Clustering
  - Youtube Ostap's Tolkien Clustering

- Groups obtained through similarity on
  - semantic similarity
  - style
  - length

Medicine

Cluster Analysis in medical imaging

- Differentiate between various types of tissue and blood in a 3D image
- Allows for accurate measurement of the rate of radioactive tracer which is delivered to the designation of interest
- Divide a fluence map into distinct regions for conversion into fields in MLC-based radiation therapy



Source: medicalnewstoday.com

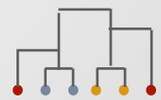
Biology

Cluster Analysis in bird habitat preferences

- Cluster movement into segments of motion
- Map the spatial distribution of behaviours
- When and for how long birds from different colonies engaged in each activity, during different stages of the breeding season

Source: deviantart.com

Source: mirrors.aggregate.org

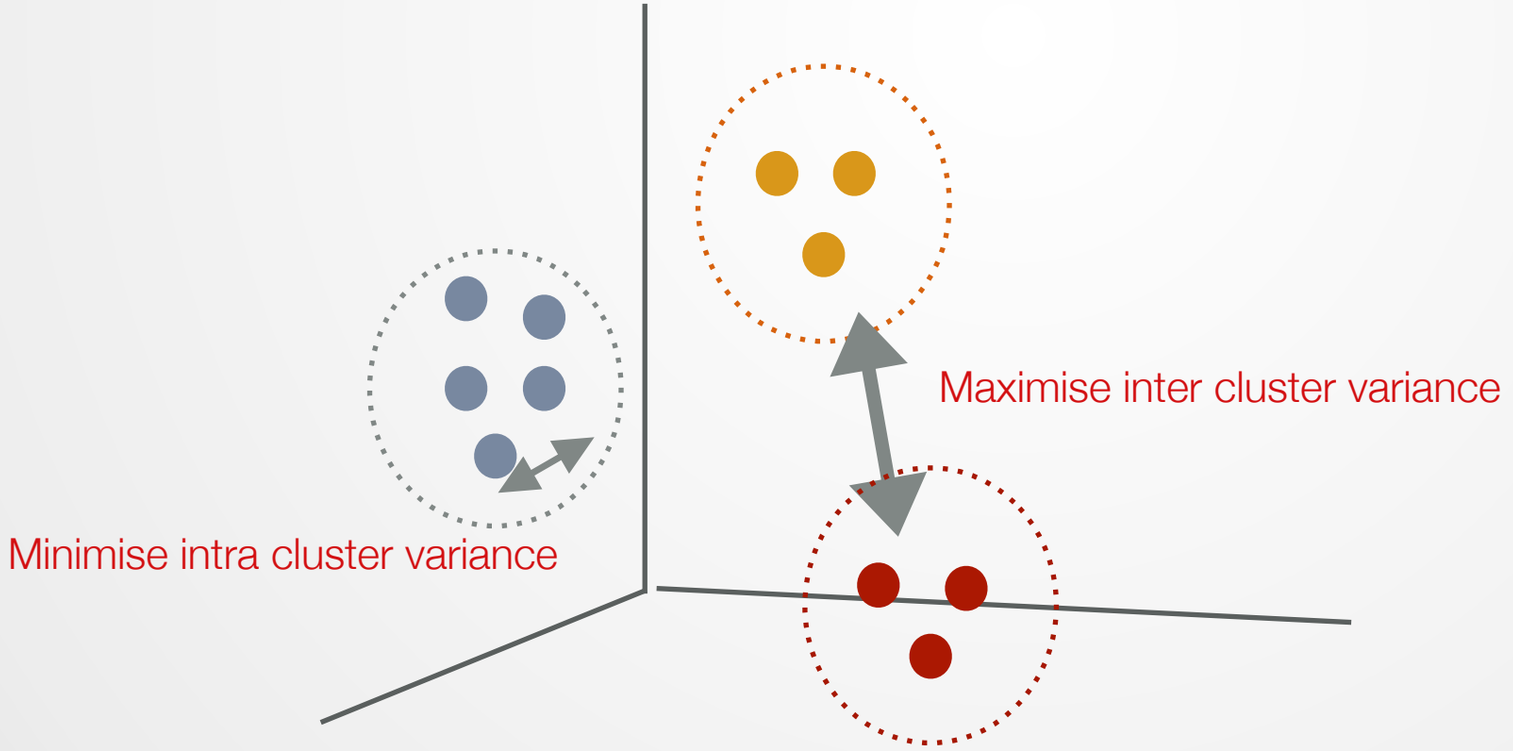


Outline

1. Motivation ✓
2. Cluster Analysis
3. Distance and Similarity
4. Cluster Algorithms
5. Real Data Application
6. Cluster VizTech

Search homogenous groups

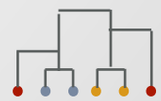
- objects within one cluster are more similar to each other than to objects from other groups
- minimise intra cluster variance
- maximise inter cluster variance



Cluster Analysis in 2 Steps

Cluster analysis can be divided into two fundamental steps.

1. Choice of a proximity measure
   - Find a way how to define which elements are close to each other
2. Choice of a group-building algorithm
   - Find a way how to assign objects to the clusters on the basis of the chosen proximity measure



Types of Clustering

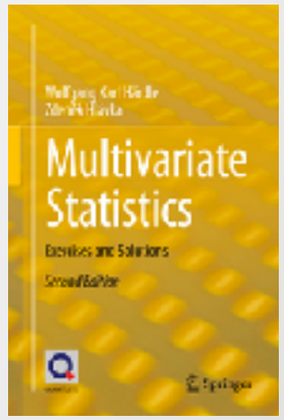
- Hierarchical vs. partitioning
  - In hierarchical clustering assignment to clusters is not changed
  - For partitioning techniques the assignment may change
- Exclusive vs. Overlapping vs. Fuzzy
  - non-exclusive clusterings: points belong to multiple clusters
  - fuzzy clustering: a point belongs to cluster with some $p \in (0,1)$
- Complete vs. Partial
  - Clustering parts of the data

US Health Data

Data set consists of 56 measurements of 22 variables. It states for one year (2005) the reported number of deaths in the 56 states and associated regions of the U.S. classified according to 15 categories:

X_1:	State (Area)	X_{12}:	Nephritis
X_2:	Total number of deaths (All)	X_{13}:	Accidents
X_3:	Human immunodeficiency virus (HIV)	X_{14}:	Vehicle Accidents
X_4:	Malignan neoplasms (Malignant)	X_{15}:	Suicide
X_5:	Diabetes	X_{16}:	Assault
X_6:	Alzheimer	X_{17}:	Firearms
X_7:	Heart	X_{18}:	Population
X_8:	Cerebrovascular (TIA)	X_{19}:	Area km^2
X_9:	Influenza	X_{20}:	Region
X_{10}:	Respiratory Diseases	X_{21}:	Division
X_{11}:	Liver	X_{22}:	State abbreviation (ANSI)

From Härdle and Hlávka (SMS), 2015 Springer



U.S. Health Data

Division (X21)		Region (X20)	
New England	1	Northeast	1
Mid-Atlantic	2	Midwest	2
E N Central	3	South	3
W N Central	4	West	4
C Atlantic	5		
E S Central	6		
W S Central	7		
Mountain	8		
Pacific	9		

U.S. Health Data

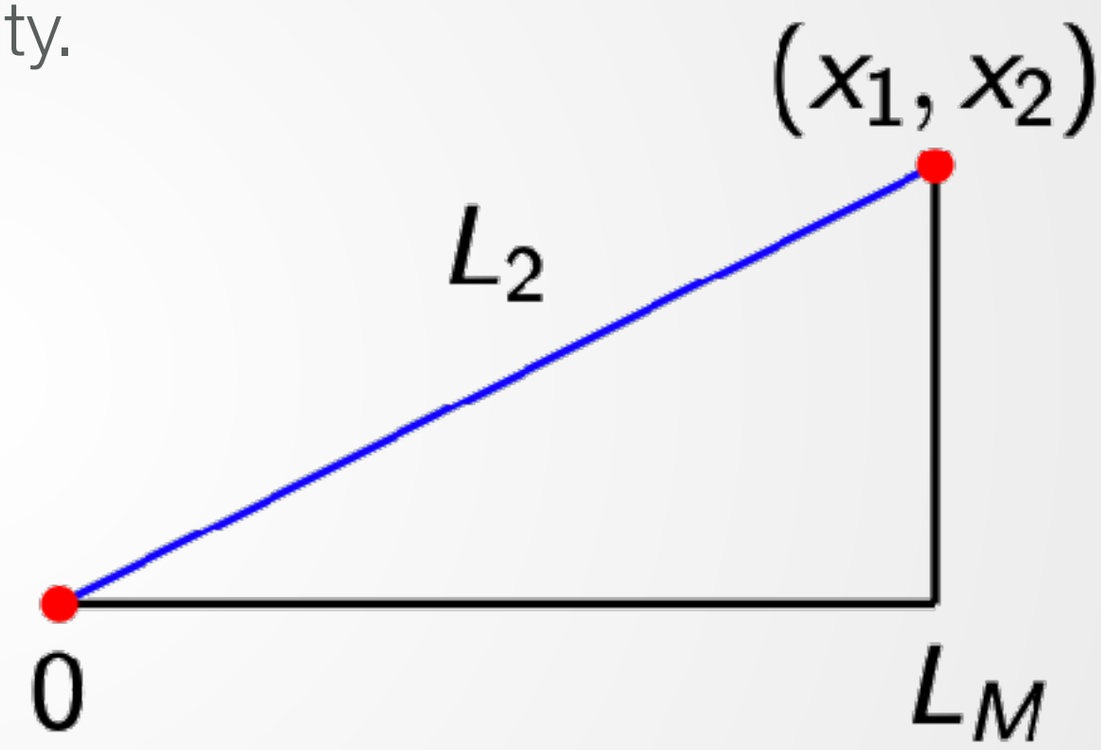
		Maine	New York	New Hampshire
X1	All	12868	152427	10194
X2	HIV	11	1644	13
X3	Malignant	3218	35556	2549
X4	Diabetes	385	4051	310
X5	Alzheimer	476	2065	376
X6	Heart	2941	51985	2530
X7	TIA	693	6622	497
X8	Influenza	352	5521	273
X9	Respiratory Deseases	830	6818	630
X10	Liver	116	1224	114
X11	Nephritis	250	2360	173
X12	Accidents	579	4645	477
X13	Vehicle Accidents	192	1530	162
X14	Suicide	175	1189	162
X15	Assault	22	901	19
X16	Firearms	109	1019	88
X17	Population	1321505	19254630	1309940
X18	Area.Km2	79883	122057	23187
X19	Region	Northeast	Northeast	Northeast
X20	Division	New England	Mid Atlantic	New England
X21	ANSI	ME	NY	NH

Clustering

- Cluster analysis is a set of tools and methods for building groups (clusters) from multivariate data objects
- Two fundamental steps:
  - Choice of a proximity measure
  - Choice of a group-building
- Maximize simultaneously:
  - Homogeneity within groups
  - Heterogeneity between groups

Proximity Measure

- D measures similarity or dissimilarity.
- For d_{ij} distance values: dissimilarity.
- For d_{ij} proximity measures: similarity.



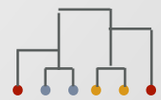
$$d_{ij} = \|x_i - x_j\|_2, \quad x_i, x_j \in \mathbb{R}^p$$

Proximity, Distance and Similarity

- For d_{ij} a distance: a proximity measure

$$d_{ij}' = \max_{ij}\{d_{ij}\} - d_{ij}$$

- Nominal values (like 0/1 variables) yield typically proximity values.
- Metric values lead (in general) to distance matrices.



Distance Measure for Continuous Variables

- A great variety of distance measures can be generated by L_r-norms, $r \geq 1$,

$$d_{ij} = \|x_i - x_j\|_r = \{ \sum_{k=1}^{p} |x_{ik} - x_{jk}|^r \}^{\frac{1}{r}} \qquad (1)$$

- x_{ik} denotes the value of the *k*-th variable on object *i*.
  - *1-norm distance* $\sum_{i=1}^{n} |x_i - y_i|$
  - *2-norm distance* $\left(\sum_{i=1}^{n} |x_i - y_i|^2 \right)^{1/2}$

Example

$x_1 = (0, 0)$, $x_2 = (1, 0)$ and $x_3 = (5, 5)$

distance matrix for the L_1-norm:

$$D_1 = \begin{pmatrix} 0 & 1 & 10 \\ 1 & 0 & 9 \\ 10 & 9 & 0 \end{pmatrix},$$

Squared L_2-(Euclidean) norm:

$$D_2 = \begin{pmatrix} 0 & 1 & 50 \\ 1 & 0 & 41 \\ 50 & 41 & 0 \end{pmatrix},$$

The third observation receives much more weight in the L_2-norm than in the L_1-norm.

French Railway Metric

Let (X, d) be a metric space (France), and fix $x_h \epsilon \chi$ (Paris). Set $r = 1$ in (1) and define a new metric d_h on χ by letting

$$d_h(x_i, x_j) \stackrel{\text{def}}{=} \begin{cases} 0 & x_i = x_j \\ d_{ih} + d_{hj} & \text{otherwise} \end{cases}$$

for $x_i, x_j \epsilon \chi$. Then (χ, d_h) is again a metric space.

Example: French Railway Metric

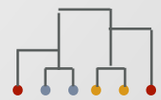
Source: Runde, V. (2005), A Taste of Topology. Ch. 2.

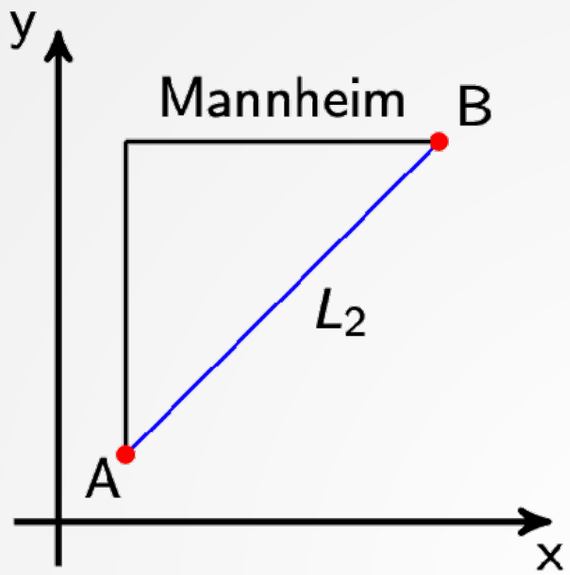
Example: Mannheim Metric

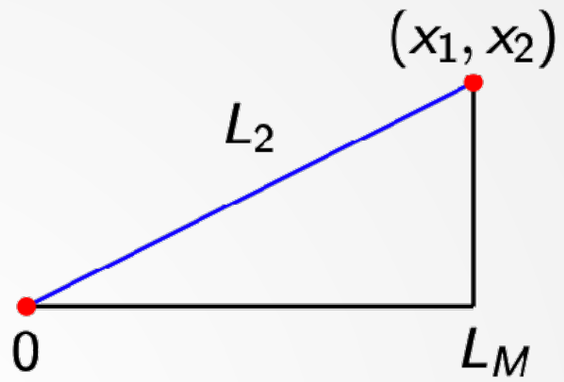
Let (χ, d) be a metric space (Mannheim). Set $r = 1$ in (1) and define a new metric d_m on χ by letting

$$d_m(x_i, x_j) \stackrel{\text{def}}{=} d_{ij}$$

for $x_i, x_j \in \chi$. Then (χ, d_m) is again a metric space.

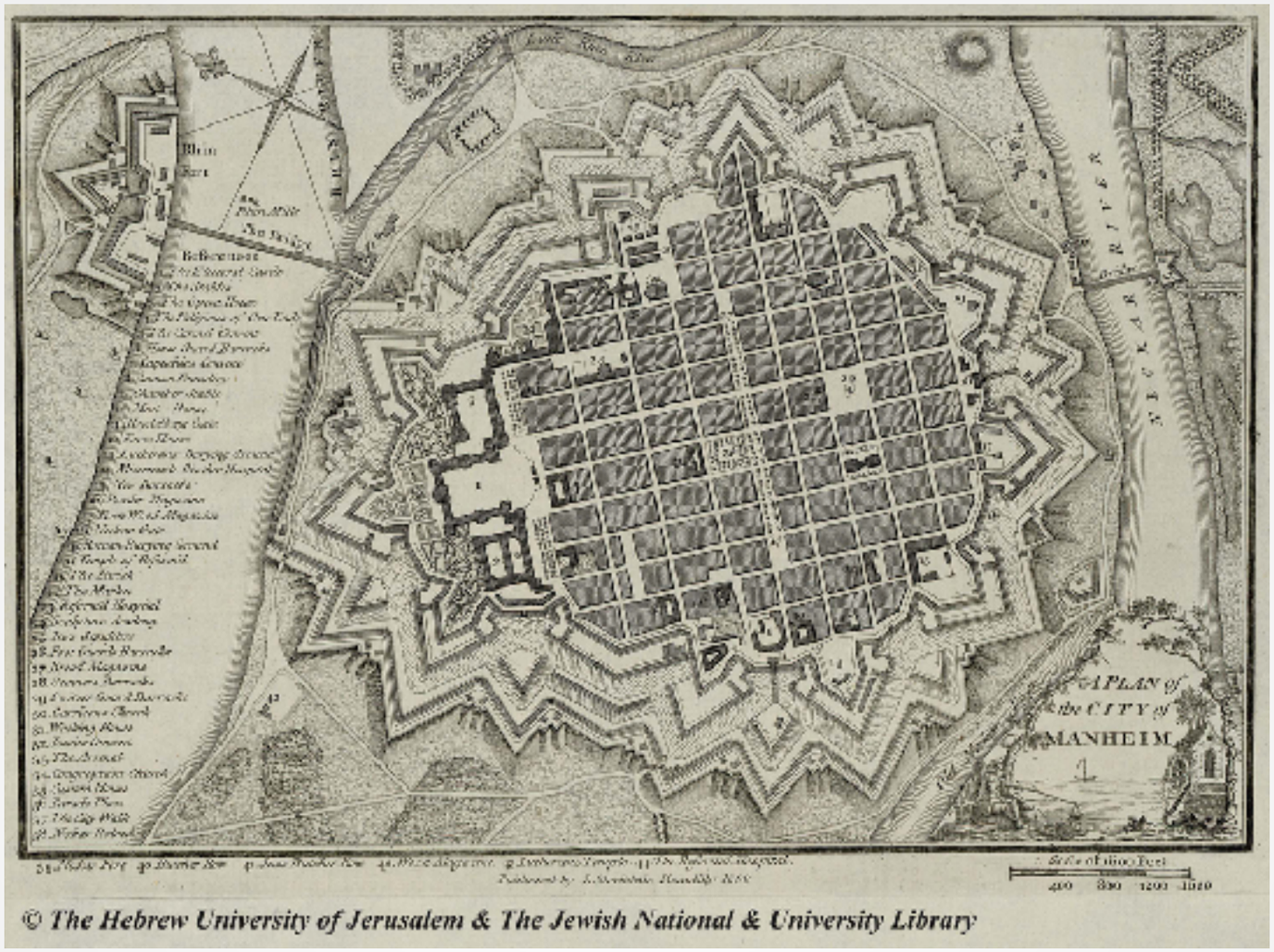






$$L_2 = x_1^2 + x_2^2$$

$$L_M = L_{Mannheim} = |x_1| + |x_2| = L_1 \text{ metric}$$



Mannheim around 1800

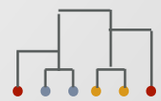
Source: http://historic-cities.huji.ac.i

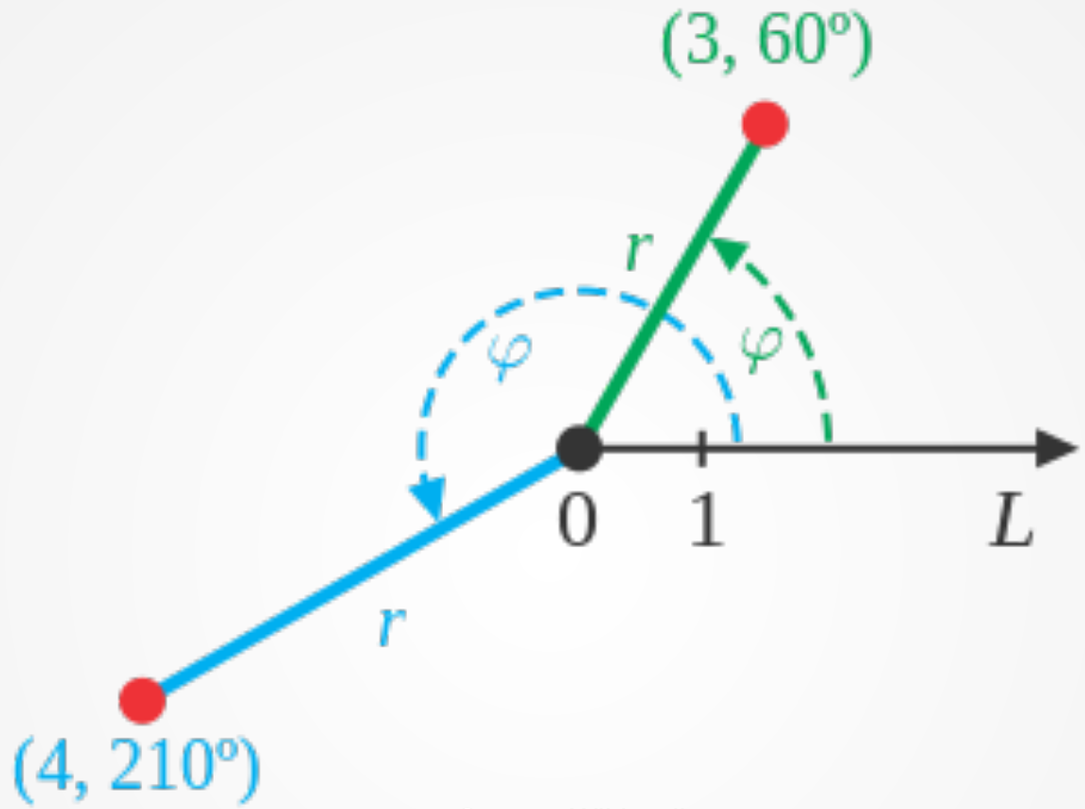
Example: Karlsruhe Metric

Let (χ, d) be a metric space (Karlsruhe). Set $r = 1$ in (1) and represent x_i, x_j in polar coordinates: $x_i = (d_{ih}, \varphi_{ih})$, $x_j = (d_{jh}, \varphi_{jh})$, where d_{ij}, d_{jh} are the distances from a fixed point x_h and φ_{ih}, φ_{jh} the angles from a fixed direction. Define:

$$d_k(x_i, x_j) \stackrel{\text{def}}{=} \begin{cases} \min(d_{ih}, d_{jh}) \cdot \delta(x_i, x_j) + |d_{ih} - d_{jh}| & 0 \leq \delta(x_i, x_j) \leq 2 \\ d_{ij} + d_{hj} & \delta(x_i, x_j) > 2 \end{cases}$$

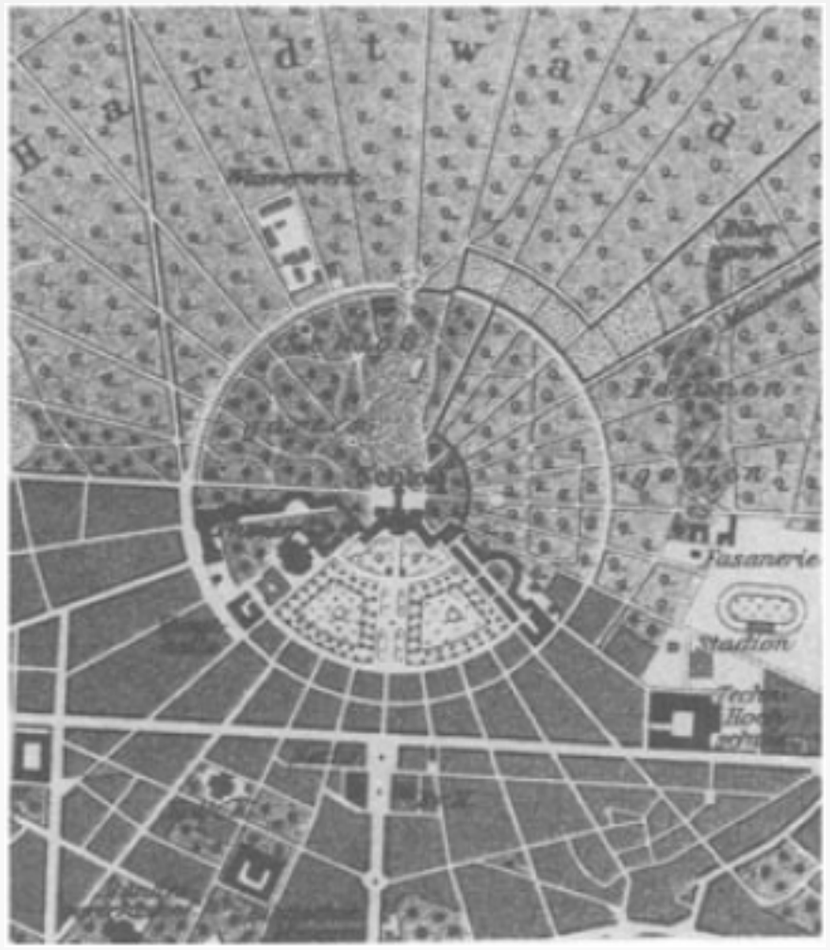
for $\delta(x_i, x_j) = \min(|\varphi_{ih} - \varphi_{jh}|, 2\pi - |\varphi_{ih} - \varphi_{jh}|)$. Then (χ, d_k) is again a metric space.





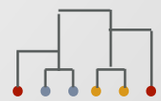
Source: Wikipedia

Polar Coordinates



Karlsruhe around 1800

Source: Proceedings of the 14th International Workshop on Graph-Theoretic Concepts in Computer Science, Voronoi Diagrams in the Moscow Metric, Rolf Klein (1989), Springer-Verlag



Distance Measure for Continuous Variables

- Example: US health 2005 data: ME, NH, NY

  - Euclidian distance $D = \begin{pmatrix} 0.0 & 850.1 & 59775.4 \\ & 0.0 & 60532.1 \\ & & 0.0 \end{pmatrix}$

  - Mannheim distance $D = \begin{pmatrix} 0.0 & 1811.0 & 108574.0 \\ & 0.0 & 110381.0 \\ & & 0.0 \end{pmatrix}$

  - Maximum distance $D = \begin{pmatrix} 0.0 & 669.0 & 49044.0 \\ & 0.0 & 49455.0 \\ & & 0.0 \end{pmatrix}$

SMSdisthealth05

Distance Measure for Continuous Variables

- An underlying hypothesis for application of L_r distances is that the variables are measured in the same scale.
- If this is not the case, use a more general metric ($A > 0$):

$$d_{ij}^2 = \|x_i - x_j\|_A = (x_i - x_j)^\top A (x_i - x_j) \qquad (2)$$

- Mahalanobis Distance:

$$d_{ij}^2 = \|x_i - x_j\|_A = (x_i - x_j)^\top S^{-1} (x_i - x_j)$$

- Ideally $A = \Sigma^{-1}$ with estimated cov-matrix $A = S^{-1}$

Mahalanobis Distance

Take $A = \Sigma^{-1}$ or S^{-1} in (2):

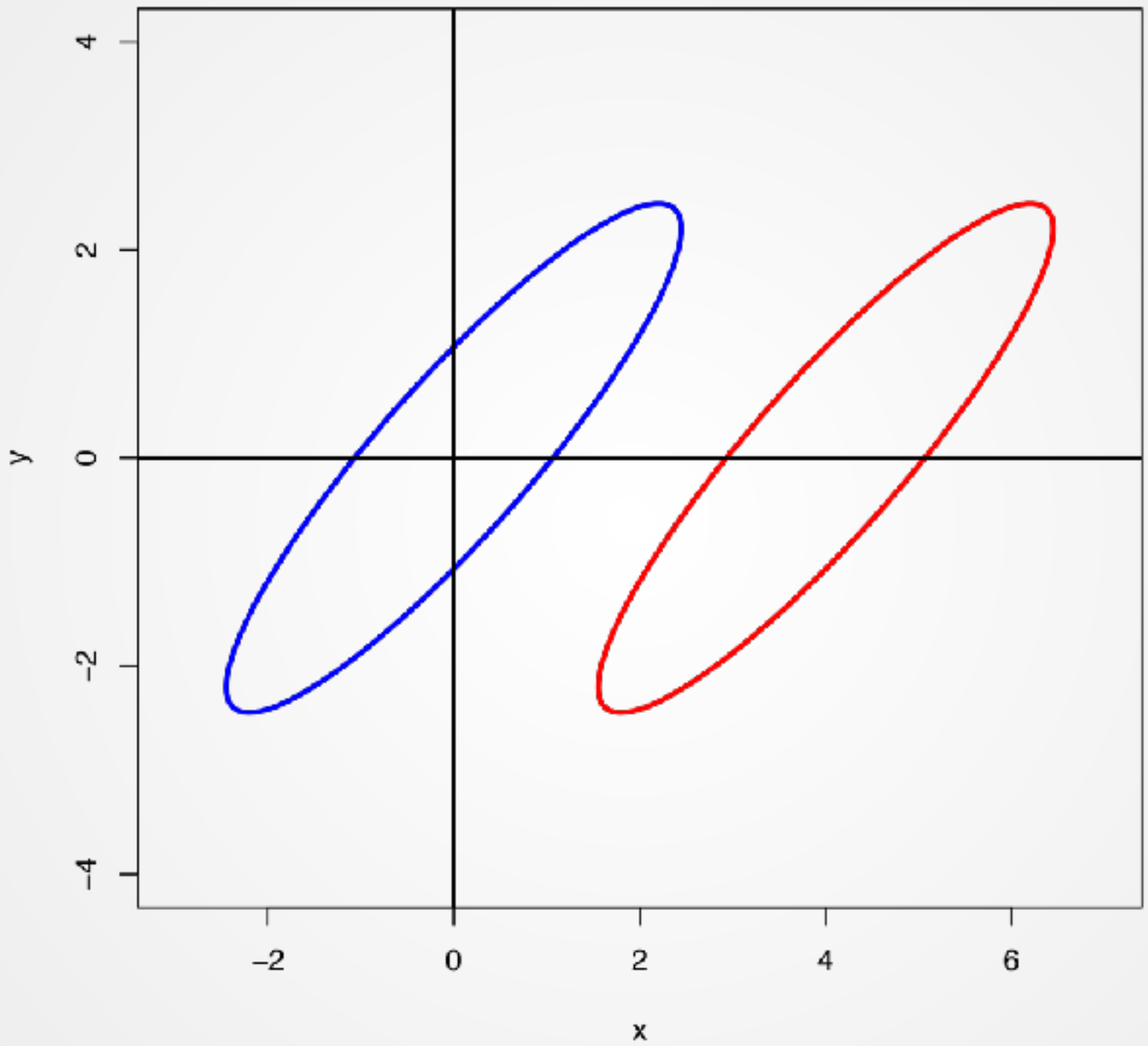
$$d_{ij}^2 = \|x_i - x_j\|_{S^{-1}} = (x_i - x_j)^\top S^{-1}(x_i - x_j)$$

Example

$$X \sim N_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \Sigma\right), \quad Y \sim N_2\left(\begin{pmatrix}4\\0\end{pmatrix}, \Sigma\right)$$

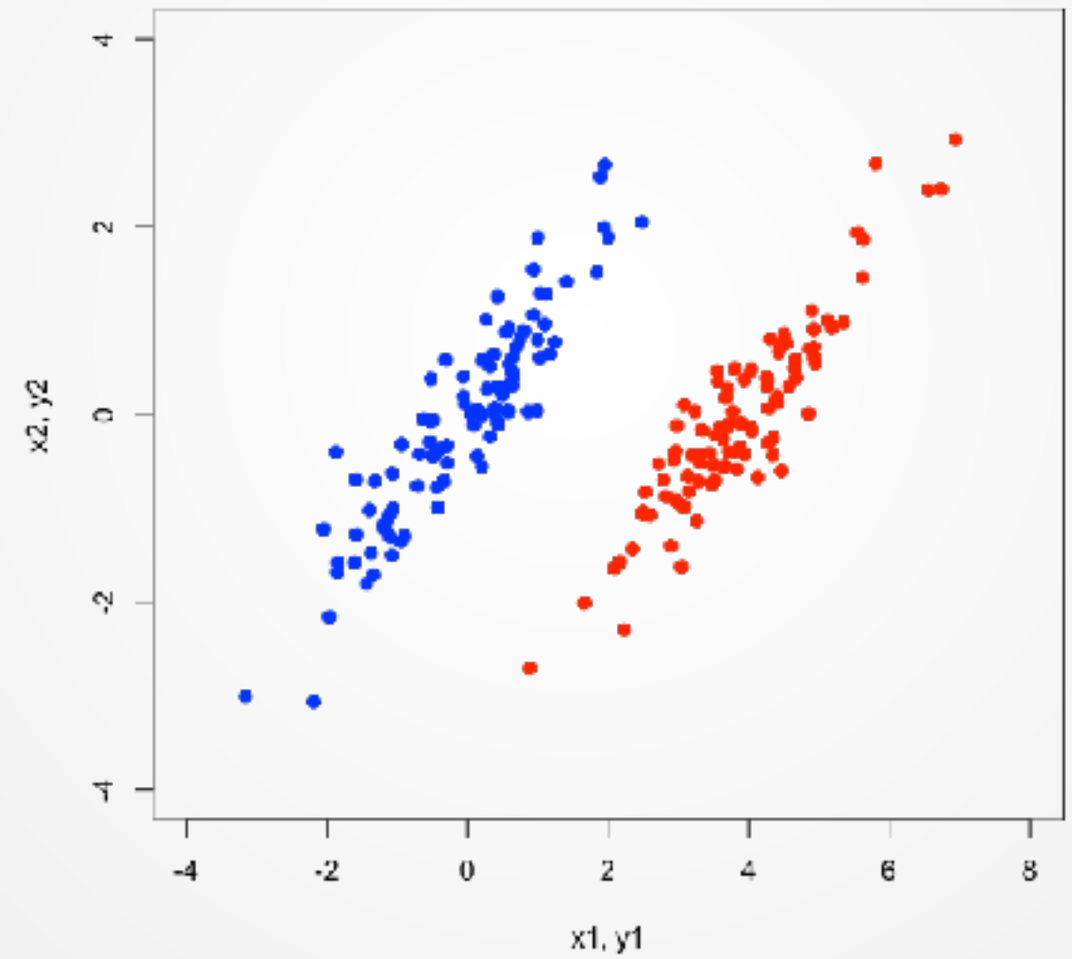
$$\Sigma = \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}, \quad \rho = 0.9$$

Example



Example

$$n = 100, \quad \mu_x = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \mu_y = \begin{pmatrix} 4 \\ 0 \end{pmatrix}, \quad \rho = 0.9$$



SMSmdmv

Calculate the D with $A = I_{100}^{-1}, A = S_{100}^{-1}$

Example: French Food expenditures

The data set consists of the average expenditures on food for several different types of families in France (manual workers = MA, employees = EM, managers = CA) with different numbers of children (2, 3, 4 or 5 family members)

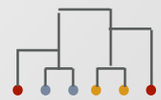
Source: cuisinealafrancaise.com

SMSdecofood

Example: French Food expenditures

Euclidean proximity (L_2-norm):

$$D = 10^4 \begin{pmatrix}
0.00 & 5.82 & 58.19 & 3.54 & 5.15 & 151.44 & 16.91 & 36.15 & 147.99 & 51.84 & 102.56 & 271.83 \\
 & 0.00 & 41.73 & 4.53 & 2.93 & 120.59 & 13.52 & 25.39 & 116.31 & 43.68 & 76.81 & 226.87 \\
 & & 0.00 & 44.14 & 40.10 & 24.12 & 29.95 & 8.17 & 25.57 & 20.81 & 20.30 & 88.62 \\
 & & & 0.00 & 0.76 & 127.85 & 5.62 & 21.70 & 124.98 & 31.21 & 72.97 & 231.57 \\
 & & & & 0.00 & 121.05 & 5.70 & 19.85 & 118.77 & 30.82 & 67.39 & 220.72 \\
 & & & & & 0.00 & 96.57 & 48.16 & 1.80 & 60.52 & 28.90 & 29.56 \\
 & & & & & & 0.00 & 9.20 & 94.87 & 11.07 & 42.12 & 179.84 \\
 & & & & & & & 0.00 & 46.95 & 6.17 & 18.76 & 113.03 \\
 & & & & & & & & 0.00 & 61.08 & 29.62 & 31.86 \\
 & & & & & & & & & 0.00 & 15.83 & 116.11 \\
 & & & & & & & & & & 0.00 & 53.77 \\
 & & & & & & & & & & & 0.00
\end{pmatrix}$$



Example: Using A = diag($s_{X_1X_1}^{-1}, \ldots, s_{X_7X_7}^{-1}$)

➢ Mahalanobis proximity D (L_2-norm)

$$D = 10^4 \begin{pmatrix} 0.00 & 6.85 & 10.04 & 1.68 & 2.66 & 24.90 & 8.28 & 8.56 & 24.61 & 21.55 & 30.68 & 57.48 \\ & 0.00 & 13.11 & 6.59 & 3.75 & 20.12 & 13.13 & 12.38 & 15.88 & 31.52 & 25.65 & 46.64 \\ & & 0.00 & 8.03 & 7.27 & 4.99 & 9.27 & 3.88 & 7.46 & 14.92 & 15.08 & 26.89 \\ & & & 0.00 & 0.64 & 20.06 & 2.76 & 3.82 & 19.63 & 12.81 & 19.28 & 45.01 \\ & & & & 0.00 & 17.00 & 3.54 & 3.81 & 15.76 & 14.98 & 16.89 & 39.87 \\ & & & & & 0.00 & 17.51 & 9.79 & 1.58 & 21.32 & 11.36 & 13.40 \\ & & & & & & 0.00 & 1.80 & 17.92 & 4.39 & 9.93 & 33.61 \\ & & & & & & & 0.00 & 10.50 & 5.70 & 7.97 & 24.41 \\ & & & & & & & & 0.00 & 24.75 & 11.02 & 13.07 \\ & & & & & & & & & 0.00 & 9.13 & 29.78 \\ & & & & & & & & & & 0.00 & 9.39 \\ & & & & & & & & & & & 0.00 \end{pmatrix}$$

Contingency Tables

If χ is a contingency table:

row i characterized by the conditional frequency $\frac{x_{ij}}{x_{i\bullet}}$

column j characterized by the conditional frequency $\frac{x_{ij}}{x_{\bullet j}}$

where $x_{i\bullet} = \sum_{j=1}^{p} x_{ij}$, $x_{\bullet j} = \sum_{i=1}^{n} x_{ij}$, $x_{\bullet\bullet} = \sum_{j=1}^{p}\sum_{i=1}^{n} x_{ij}$.

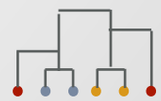
Contingency Tables

Distance between rows i_1 and i_2:

$$d^2(i_1, i_2) = \sum_{j=1}^{p} \frac{1}{(\frac{x_{\bullet j}}{x_{\bullet\bullet}})} (\frac{x_{i_1 j}}{x_{i_1 \bullet}} - \frac{x_{i_2 j}}{x_{i_2 \bullet}})^2$$

Similarly for columns:

$$d^2(j_1, j_2) = \sum_{i=1}^{n} \frac{1}{(\frac{x_{i\bullet}}{x_{\bullet\bullet}})} (\frac{x_{ij_1}}{x_{\bullet j_1}} - \frac{x_{ij_2}}{x_{\bullet j_2}})^2$$



Example: Cluster Analysis for US Crime data

- This is a data set consisting of 50 measurements of 7 variables. It states for one year (1985) the reported number of crimes in the 50states of the United States classified according to 7 categories (X_3-X_9):

variable	description
X_1:	land area (land)
X_2:	population 1985 (popu 1985)
X_3:	murder (murd)
X_4:	rape
X_5:	robbery (robb)
X_6:	assault (assa)
X_7:	burglary (burg)
X_8:	larceny (larc)
X_9:	auto theft (auto)
X_{10}:	U.S. states region number (reg)
X_{11}:	U.S. states division number (div)

Example

- Used are just the 7 crime variables. The matrix is interpreted as a contingency table. Derived is the distance matrix D.

$$D = \begin{pmatrix} 0.000 & 0.004 & 0.002 & 0.230 & 0.142 & 0.034 & \cdots & 0.009 \\ & 0.000 & 0.011 & 0.172 & 0.098 & 0.014 & \cdots & 0.001 \\ & & 0.000 & 0.272 & 0.176 & 0.051 & \cdots & 0.019 \\ & & & 0.000 & 0.010 & 0.087 & \cdots & 0.146 \\ & & & & 0.000 & 0.037 & \cdots & 0.079 \\ & & & & & 0.000 & \cdots & 0.008 \\ & & & & & & \ddots & \vdots \\ & & & & & & & 0.000 \end{pmatrix}$$

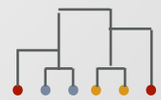
MVAclususcrime

Q-Correlation Distance

Apart from the L_r-distance measures, the Q-correlation coefficient can be used

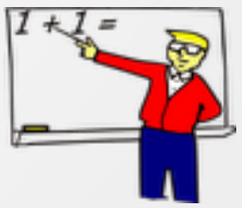
$$d_{ij} = \frac{\sum_{k=1}^{p} (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\{ \sum_{k=1}^{p} (x_{ik} - \bar{x}_i)^2 \sum_{k=1}^{p} (x_{jk} - \bar{x}_j)^2 \}^{1/2}}$$

where $\bar{x}_i$ denotes the mean of $x_{i1}, x_{i2}, \ldots, x_{ip}$



Summary: Proximity between Objects (continuous data)

- Proximity between data points is measured by a distance or similarity matrix D with components d_{ij}
  - Each d_{ij} gives the similarity coefficient or distance between two points x_i, x_j.
- Variety of similarity (distance) measures for continuous data (e.g., L_r-distances).
- The nature of the data may impose a particular metric A for defining the distance (standardization, χ^2-metric etc.).



Proximity Measure for Binary Variables

- Basic information on similarity between binary objects:

$$(x_i, x_j),\ \ x_i^\top = (x_{i1}, \ldots, x_{ip}),\ x_j^\top = (x_{ji}, \ldots, x_{jp}),\ x_{ik}, x_{jk} \in \{0,1\}$$

$$a_1 = \sum_{k=1}^{p} \mathbf{I}(x_{ik} = x_{jk} = 1),$$

$$a_2 = \sum_{k=1}^{p} \mathbf{I}(x_{ik} = 0, x_{jk} = 1),$$

$$a_3 = \sum_{k=1}^{p} \mathbf{I}(x_{ik} = 1, x_{jk} = 0),$$

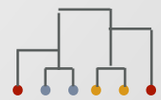
$$a_4 = \sum_{k=1}^{p} \mathbf{I}(x_{ik} = x_{jk} = 0)\,.$$

Proximity Measure for Binary Variables

- The following proximity measures are used in practice:

$$d_{ij} = \frac{a_1 + \delta a_4}{a_1 + \delta a_4 + \lambda(a_2 + a_3)}$$

Name	δ	λ	Definition
Jaccard	0	1	$\frac{a_1}{a_1 + a_2 + a_3}$
Tanimoto	1	2	$\frac{a_1 + a_4}{a_1 + 2(a_2 + a_3) + a_4}$
Simple Matching (M)	1	1	$\frac{a_1 + a_4}{p}$
Russel and Rao (RR)	-	-	$\frac{a_1}{p}$
Dice	0	0.5	$\frac{2a_1}{2a_1 + (a_2 + a_3)}$
Kulczynski	-	-	$\frac{a_1}{a_2 + a_3}$



Proximity Measure for Binary Variables

- Example: Let us consider a binary data set computed from the car data set

$$y_{ik} = \begin{cases} 1 & \text{if } x_{ik} > \overline{x_k}, \\ 0 & \text{if } else, \end{cases}$$

- for $i = 1,\ldots,n;\ k = 1,\ldots,p$

- Consider data points 17-19: Renault 19, Rover, and Toyota Corolla.
- This leads to (3 x 3) matrices.

Source: honestjohn.co.uk

Source: wikipedia.org

Source: classicsworld.co.uk

SMScarsim

Proximity Measure for Binary Variables

- Jaccard measure

$$D = \begin{pmatrix} 1.000 & 0.000 & 0.400 \\ & 1.000 & 0.167 \\ & & 1.000 \end{pmatrix}$$

- Simple matching

$$D = \begin{pmatrix} 1.000 & 0.000 & 0.625 \\ & 1.000 & 0.375 \\ & & 1.000 \end{pmatrix}$$

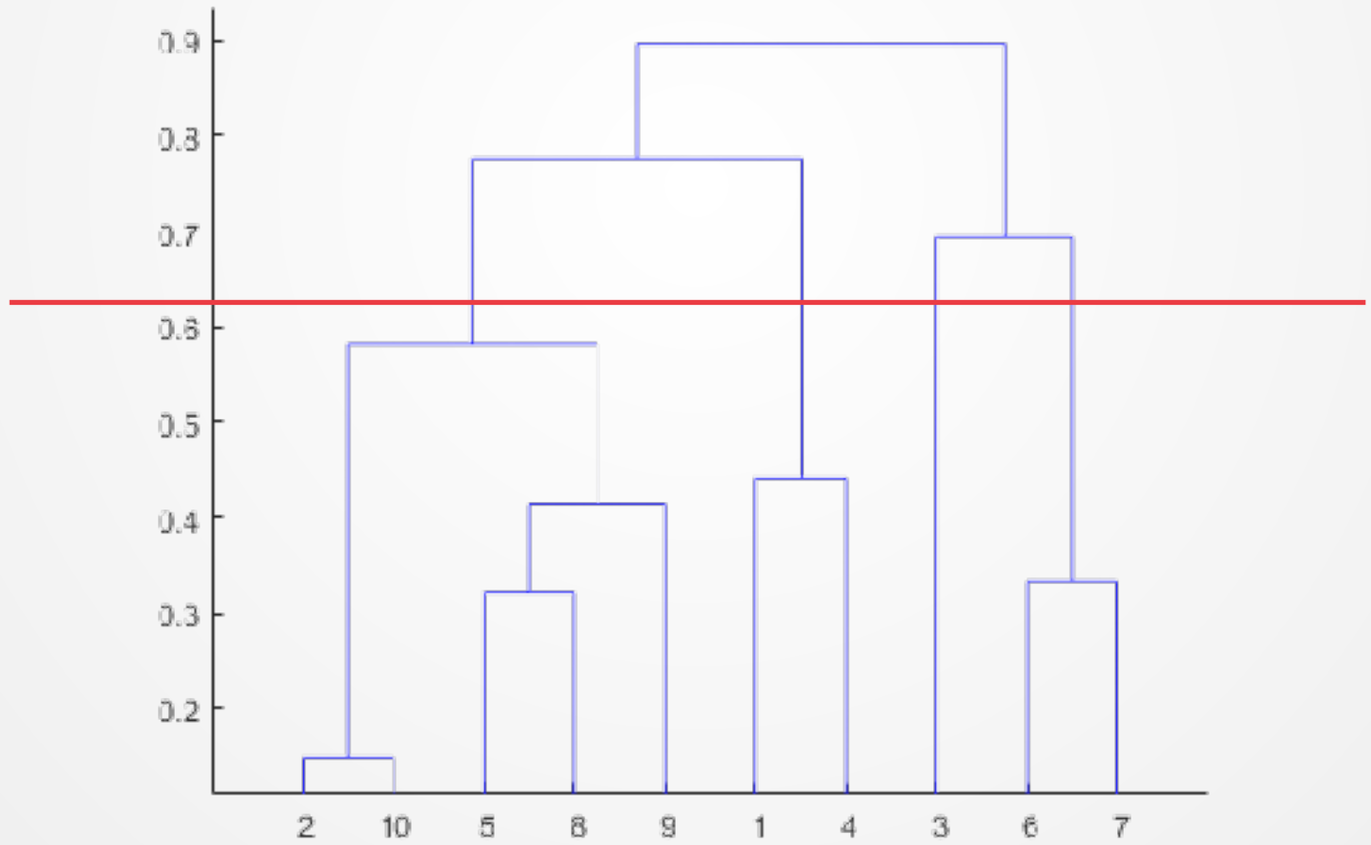
- Tanimoto measure

$$D = \begin{pmatrix} 1.000 & 0.000 & 0.455 \\ & 1.000 & 0.231 \\ & & 1.000 \end{pmatrix}$$

SMScarsim

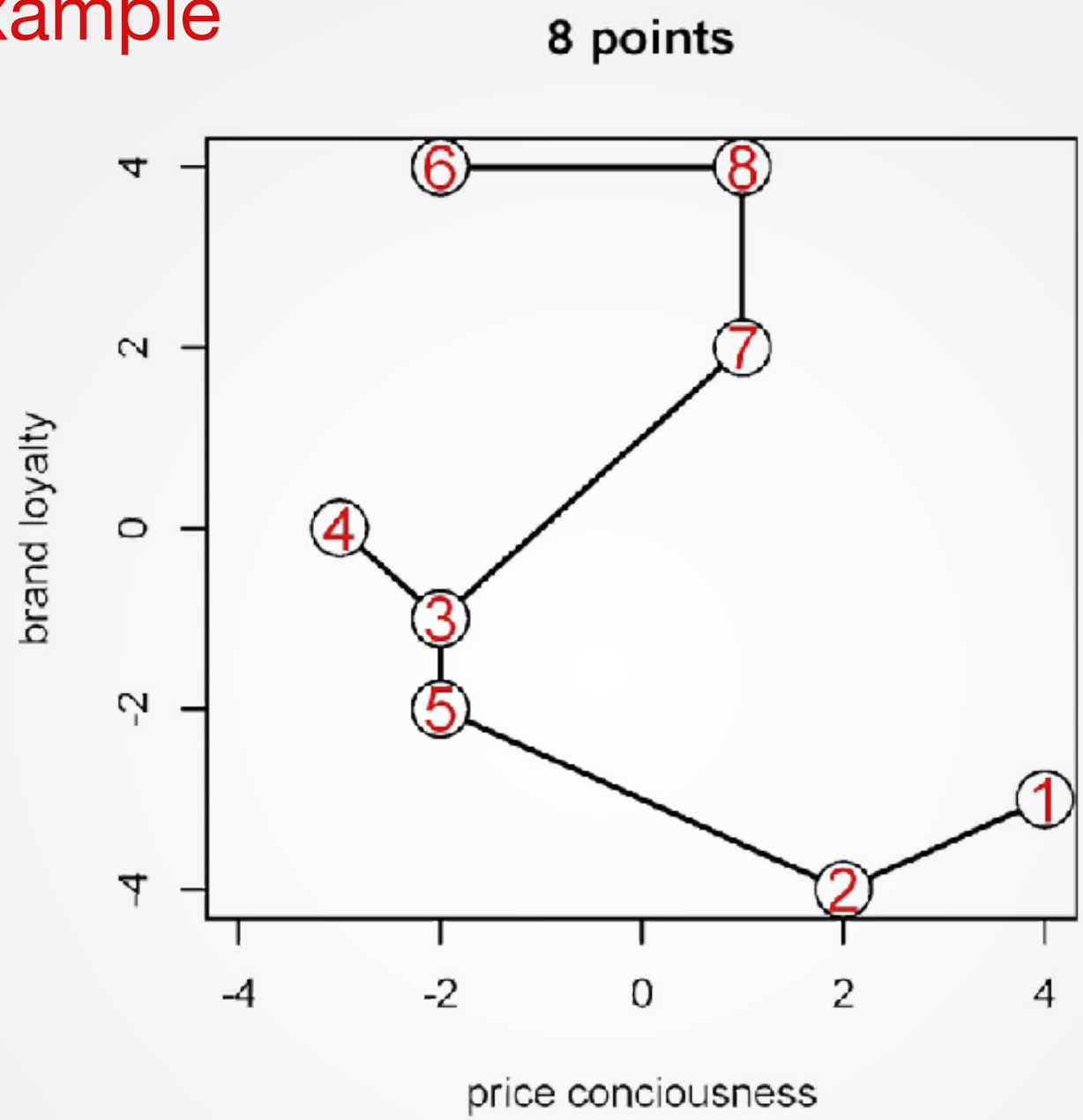
Dendrogram

- graphical representation of the sequence of clustering
- displays observations and sequence of clusters
- Horizontal axis: obs index, Vertical axis: cluster distance



Source: mathworks.com

A simple example



The 8 points example MVAclus8p

Example The distance matrix D (L_2 distances) is

$$D = \begin{pmatrix} 0 & 10 & 53 & 73 & 50 & 98 & 41 & 65 \\ & 0 & 25 & 41 & 20 & 80 & 37 & 65 \\ & & 0 & 2 & 1 & 25 & 18 & 34 \\ & & & 05 & 17 & 20 & 32 & \\ & & & & 0 & 36 & 25 & 45 \\ & & & & & 0 & 13 & 9 \\ & & & & & & 0 & 4 \\ & & & & & & & 0 \end{pmatrix}$$

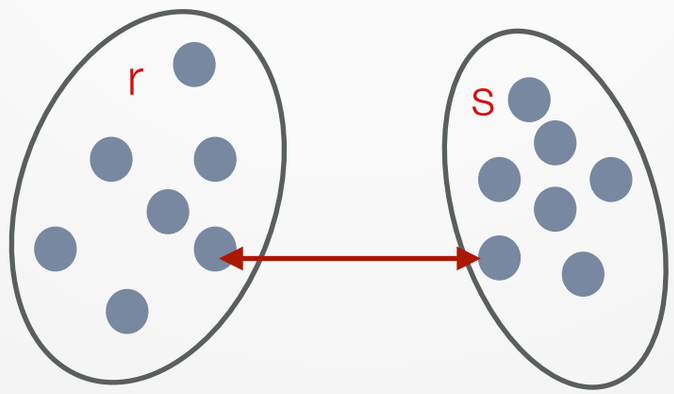
The 8 points example

Single Linkage Algorithm

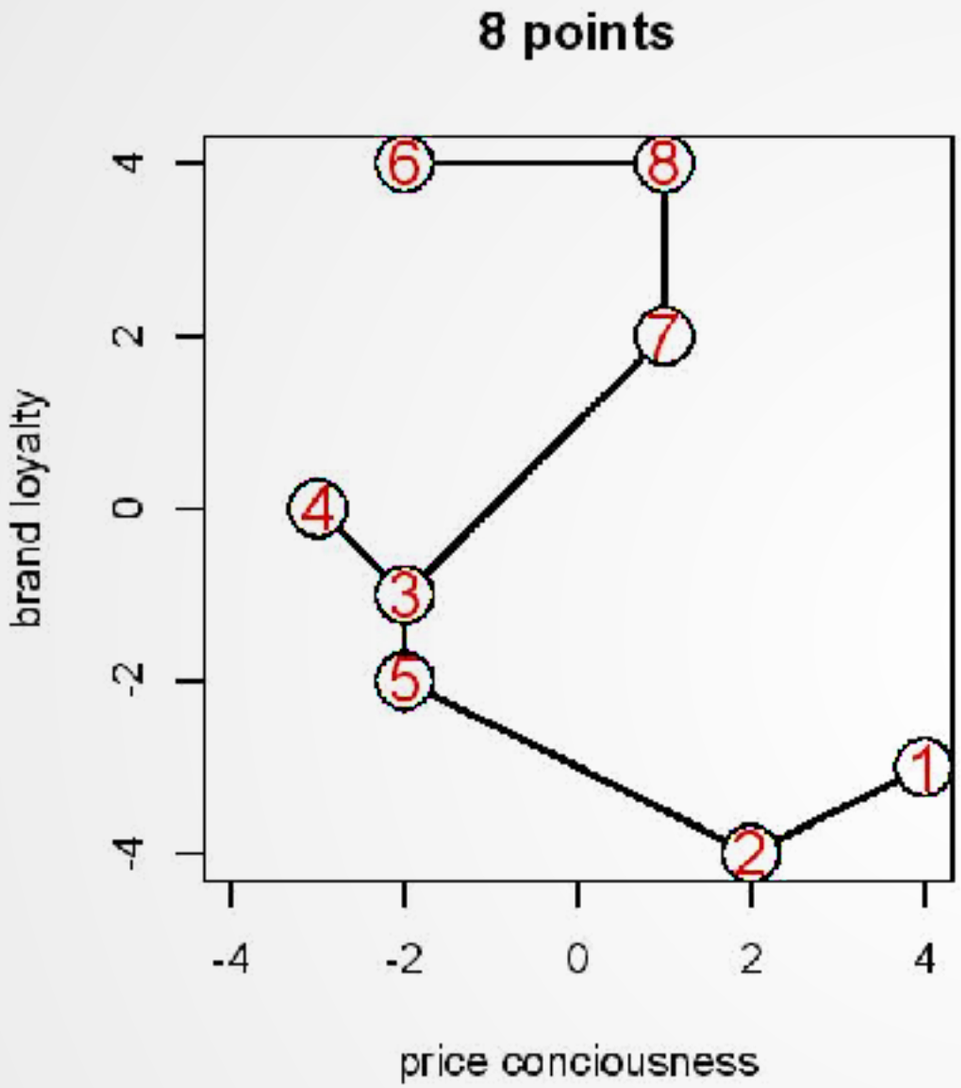
- distance between two clusters r and s : the smallest value of the individual distances.

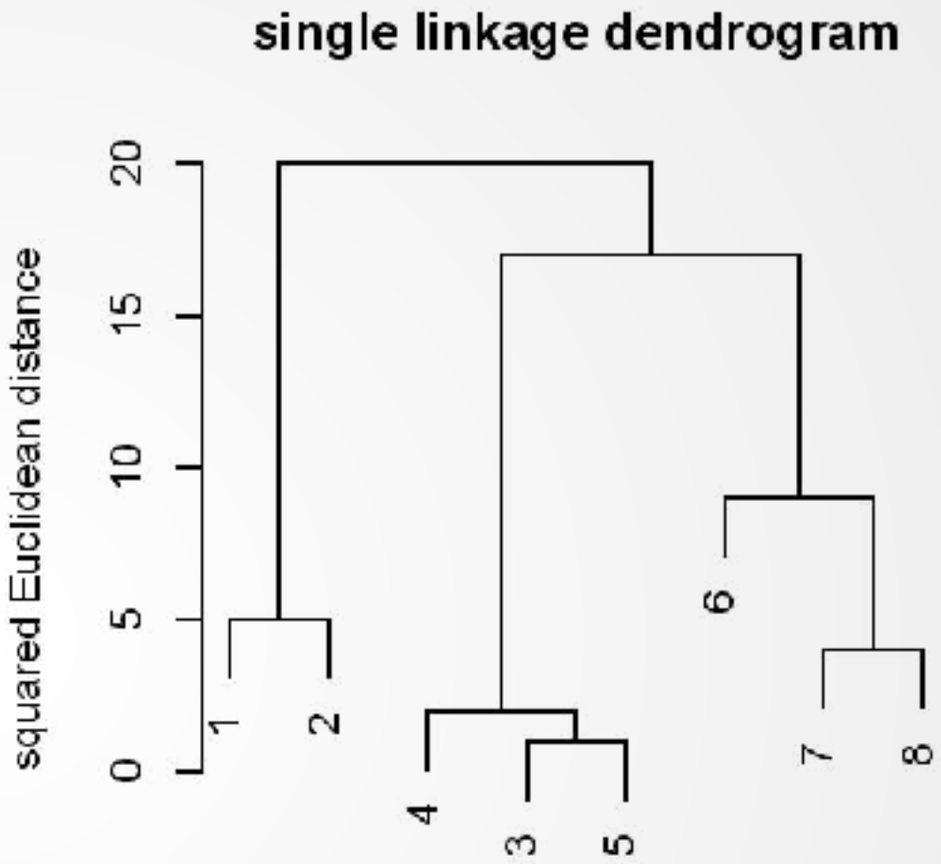
$$L(r,s) = \min\{D(x_{ri}, x_{sj})\}$$

- Also called the Nearest Neighbor algorithm.
- Single linkage Algo tends to build large groups.



A simple example





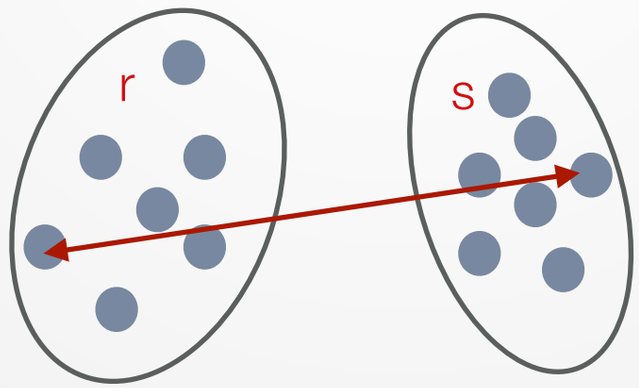
Single linkage algorithm on squared Euclidean distance for 8 point example with dendrogram. SMSclus8pd

Complete Linkage Algorithm

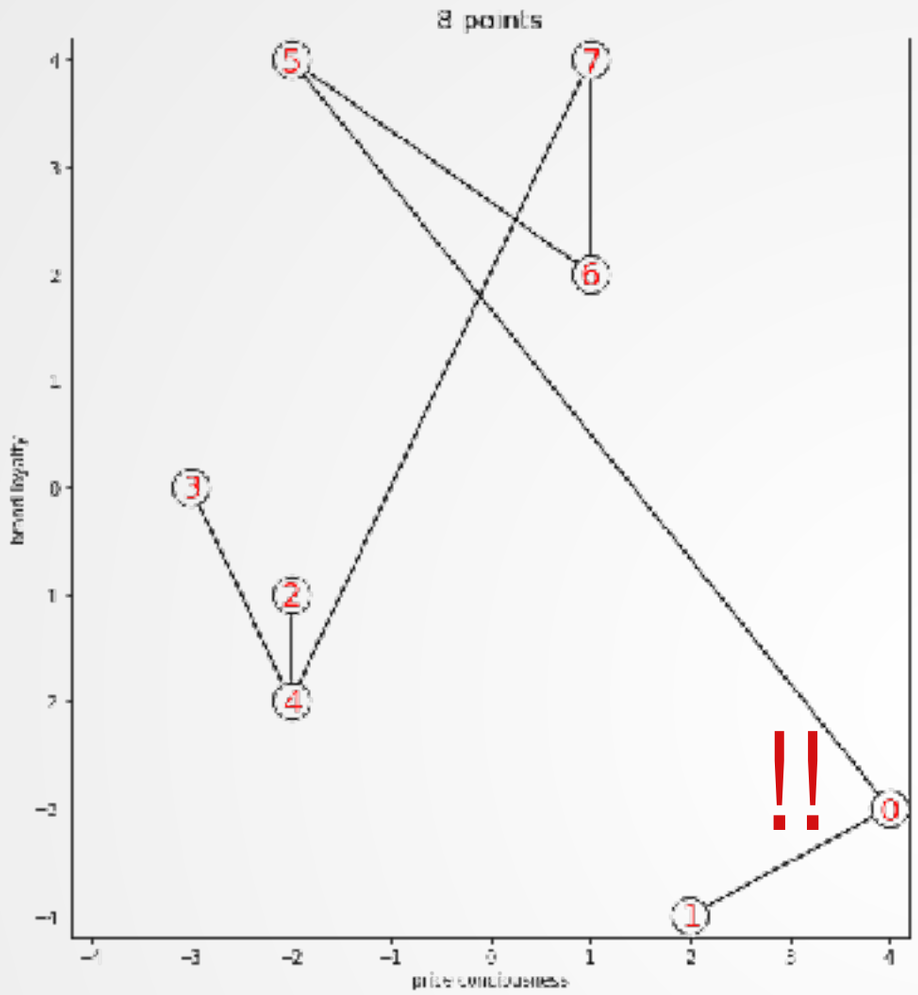
- Considers the largest (individual) distance

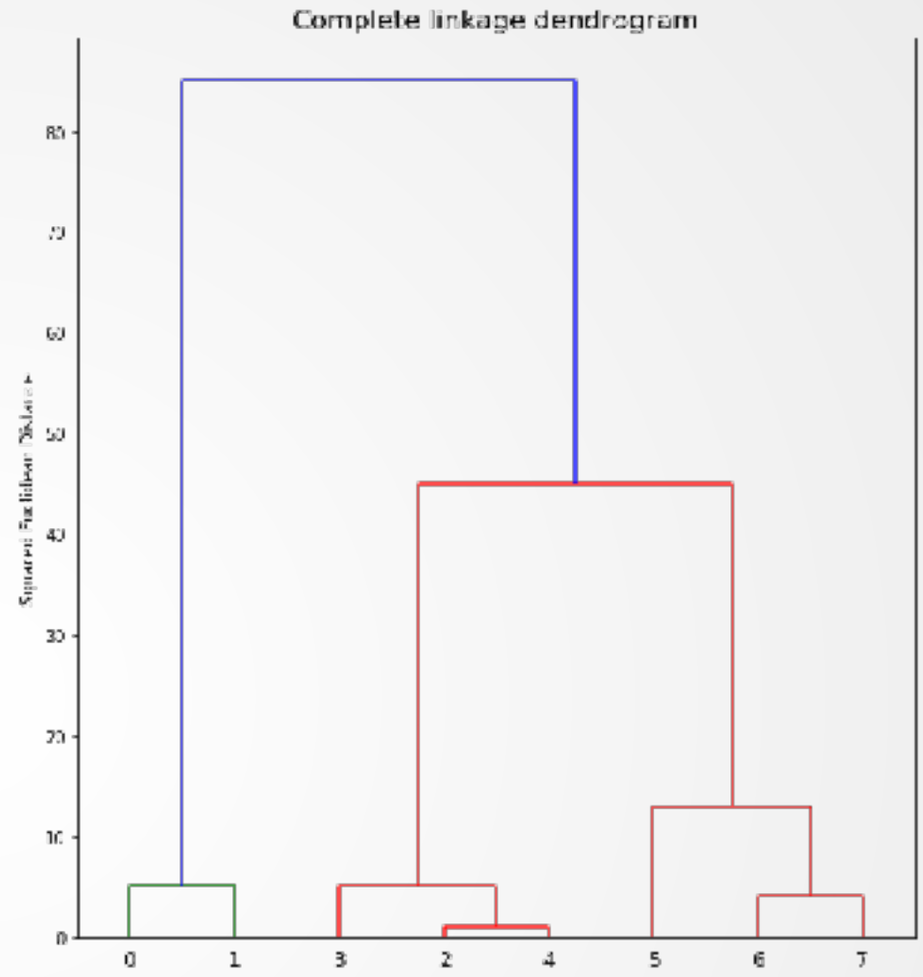
$$L(r,s) = \max\{D(x_{ri}, x_{sj})\}$$

- Also called Farthest Neighbor algorithm.
- Will cluster groups where all the points are proximate, since it compares the largest distances.



A simple example



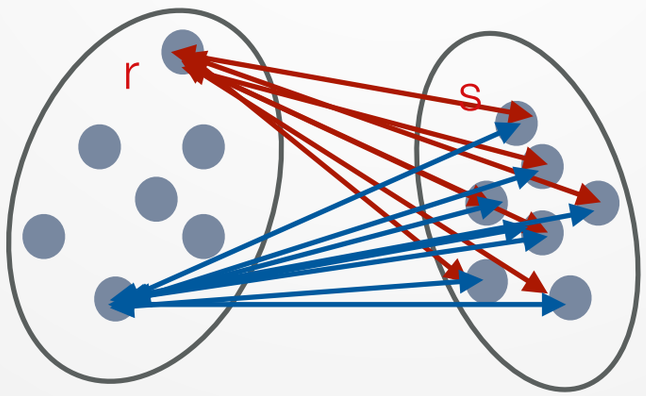


Complete linkage algorithm on squared Euclidean distance for 8 point example with dendrogram. SMSclus8pd

Average Linkage Algorithm

A compromise between nearest and farthest neighbor distance.
Average all mean distances:

$$L(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(x_{ri}, x_{sj})$$



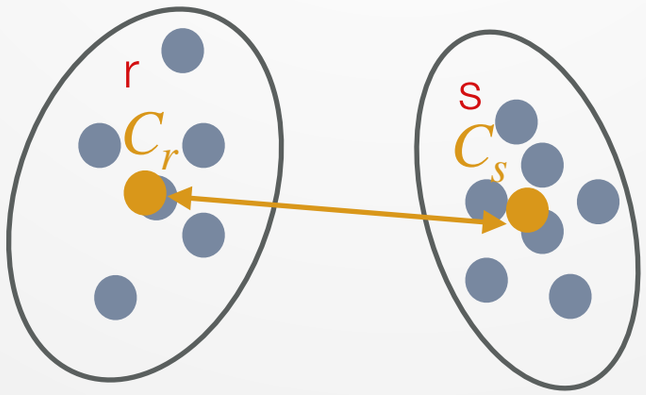
Centroid Algorithm

Employs a natural geometrical distance between clusters *r* and *s*: the weighted center of gravity of *r* and *s*.

C_r, C_s: centroids of clusters r and s

$$L(r, s) = D(C_r, C_s)$$

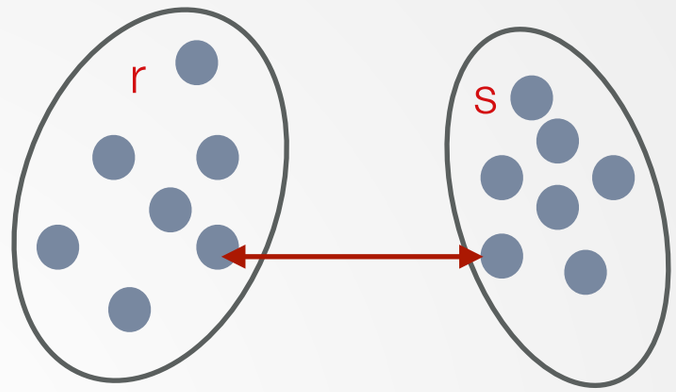
$$C_r = \frac{1}{n_r} \sum_{i=1}^{n_r} x_{ri}, \; C_s = \frac{1}{n_s} \sum_{i=1}^{n_s} x_{si}$$



Types of Linkage

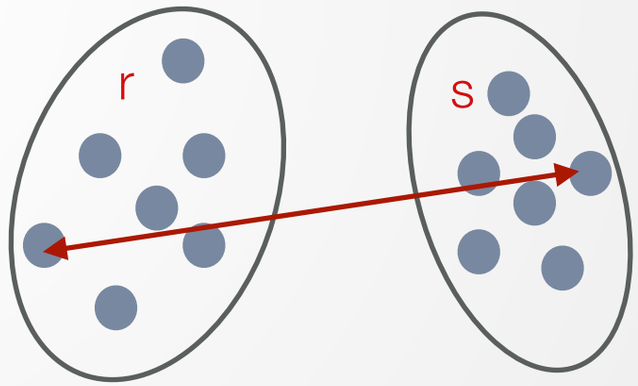
- Single Linkage

$$L(r,s) = \min\{D(x_{ri}, x_{sj})\}$$



- Complete Linkage

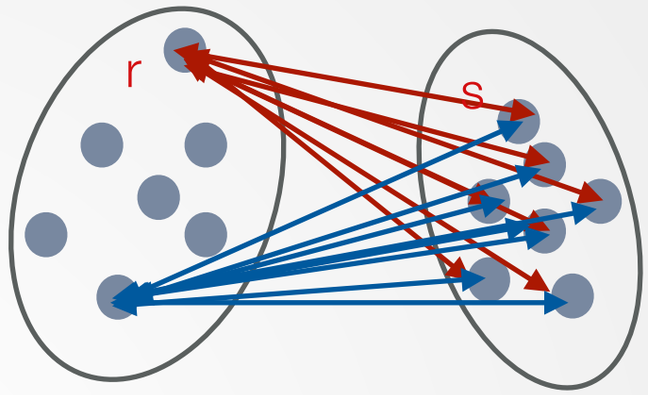
$$L(r,s) = \max\{D(x_{ri}, x_{sj})\}$$



Types of Linkage

- Average Linkage

$$L(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(x_{ri}, x_{sj})$$

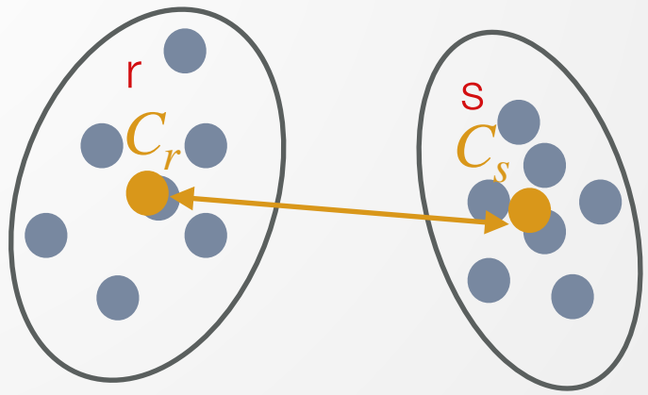


- Centroid

$$L(r,s) = D(C_r, C_s)$$

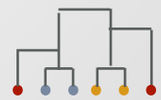
C_r, C_s: centroids of clusters r and s

$$C_r = \frac{1}{n_r} \sum_{i=1}^{n_r} x_{ri},\ C_s = \frac{1}{n_s} \sum_{i=1}^{n_s} x_{si}$$



Ward clustering algorithm

- Ward and linkage have different unification procedures
- Linkage: join groups with smallest distance. Ward join groups that do not increase too much heterogeneity
- Ward procedure: make sure that the variation inside groups is not increased too drastically



Ward clustering algorithm

Measure (Inertia) of heterogeneity for group R:

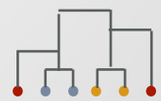
$$I_R = \frac{1}{n_R} \sum_{i=1}^{n_R} d^2(x_i, \bar{x}_R)$$

$\bar{x}_R$ center of gravity (mean)

I_R is a measure of the group dispersion around its center of gravity

When two objects or groups P and Q will be joined, the new group $P + Q$ will have a larger inertia I_{P+Q}

$$I_P + I_Q \leq I_{P+Q}$$



Ward clustering algorithm

Corresponding increase of inertia is given by

$$\Delta(P,Q) = \frac{n_P n_Q}{n_P + n_Q} d^2(P,Q)$$

Ward algorithm idea: “Join groups that give the smallest increase in $\Delta(P,Q)$“

Unification of P and Q: Ward algorithm is related to the centroid algorithm - with ‘inertial‘ distance Δ rather than the ‘geometric‘ distance d^2

Hierarchical Clustering

- Agglomerative
  - top-down approach
  - each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
- Divisive
  - bottom-up approach
  - all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

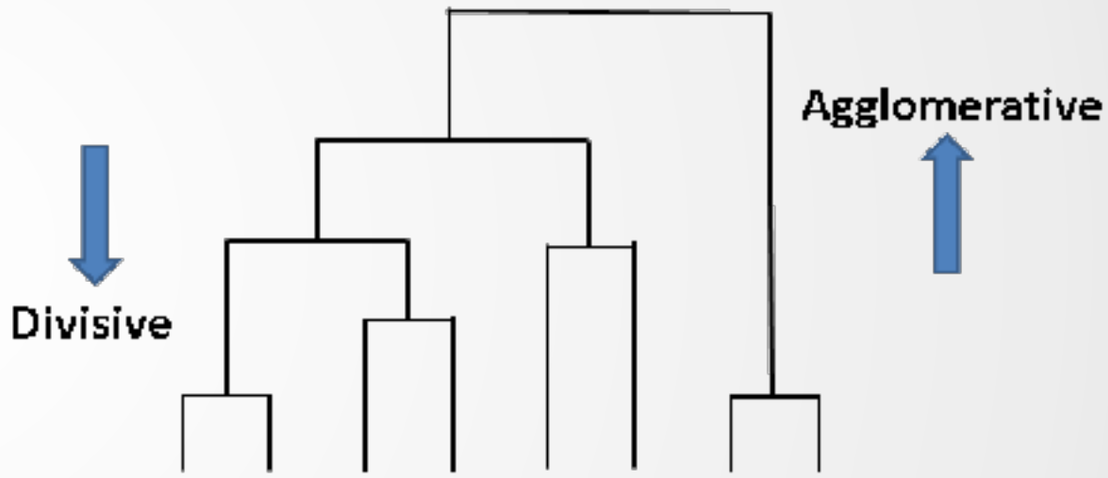


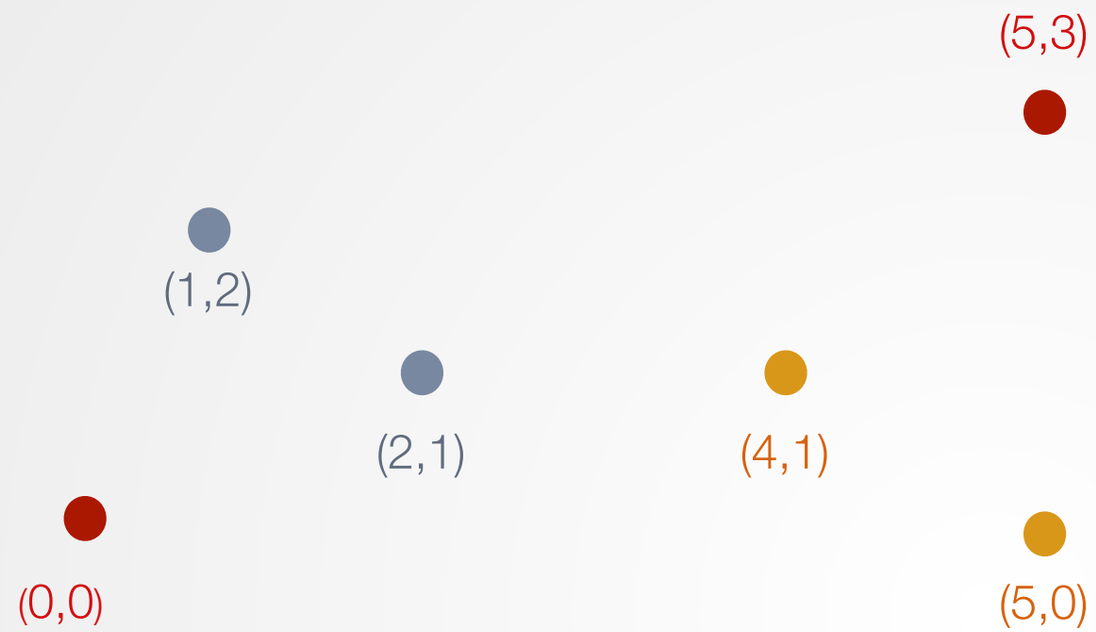
Image Source: towardsdatascience.com

Agglomerative Algorithm

1. Construct the finest partition, i.e. each observation is one cluster
2. Compute the distance metric D according to chosen linkage criterion.
3. Find two clusters with the closest distance.
4. Unite the two clusters into one cluster.
5. Compute the distance between the new groups and obtain a reduced distance metric D.

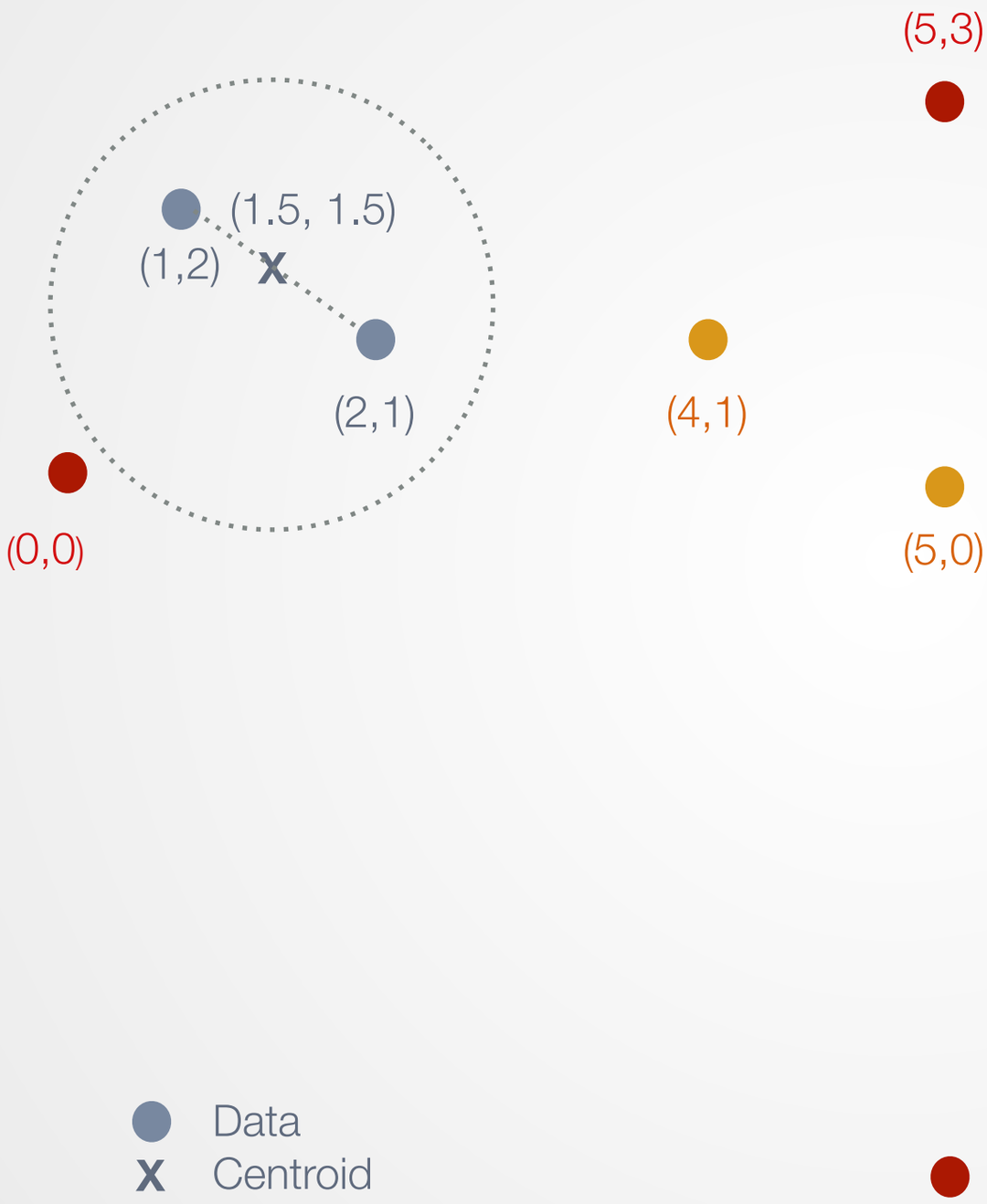
Repeat until all clusters are agglomerated.

Simple Example

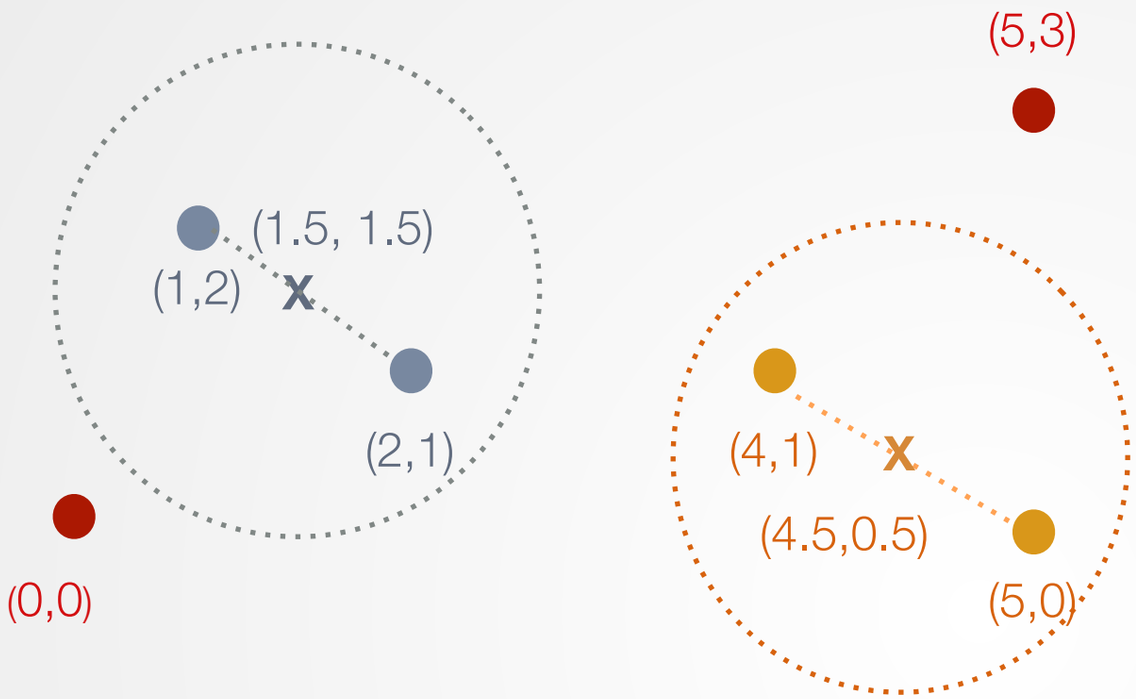


Data
X Centroid

Simple Example

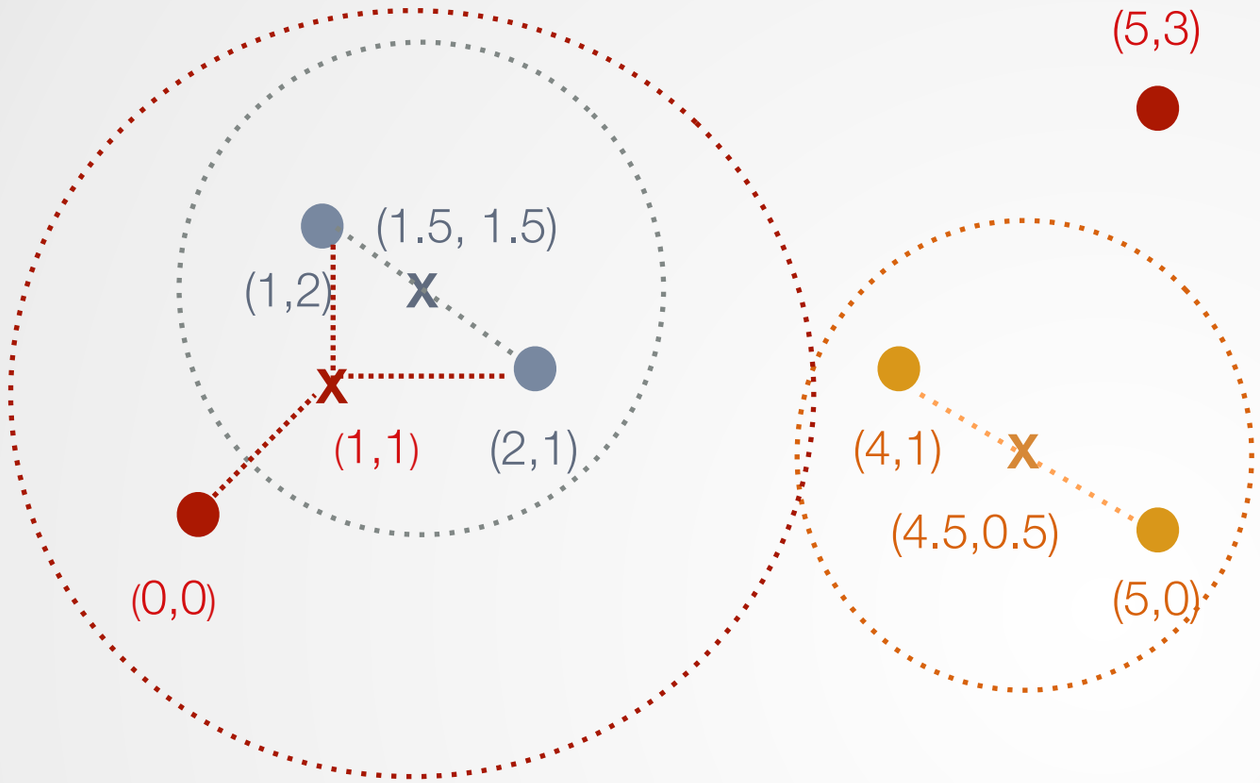


Simple Example

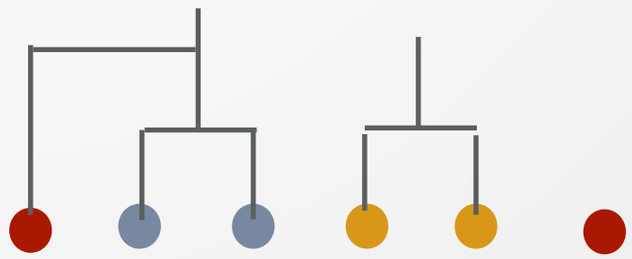


Data
x Centroid

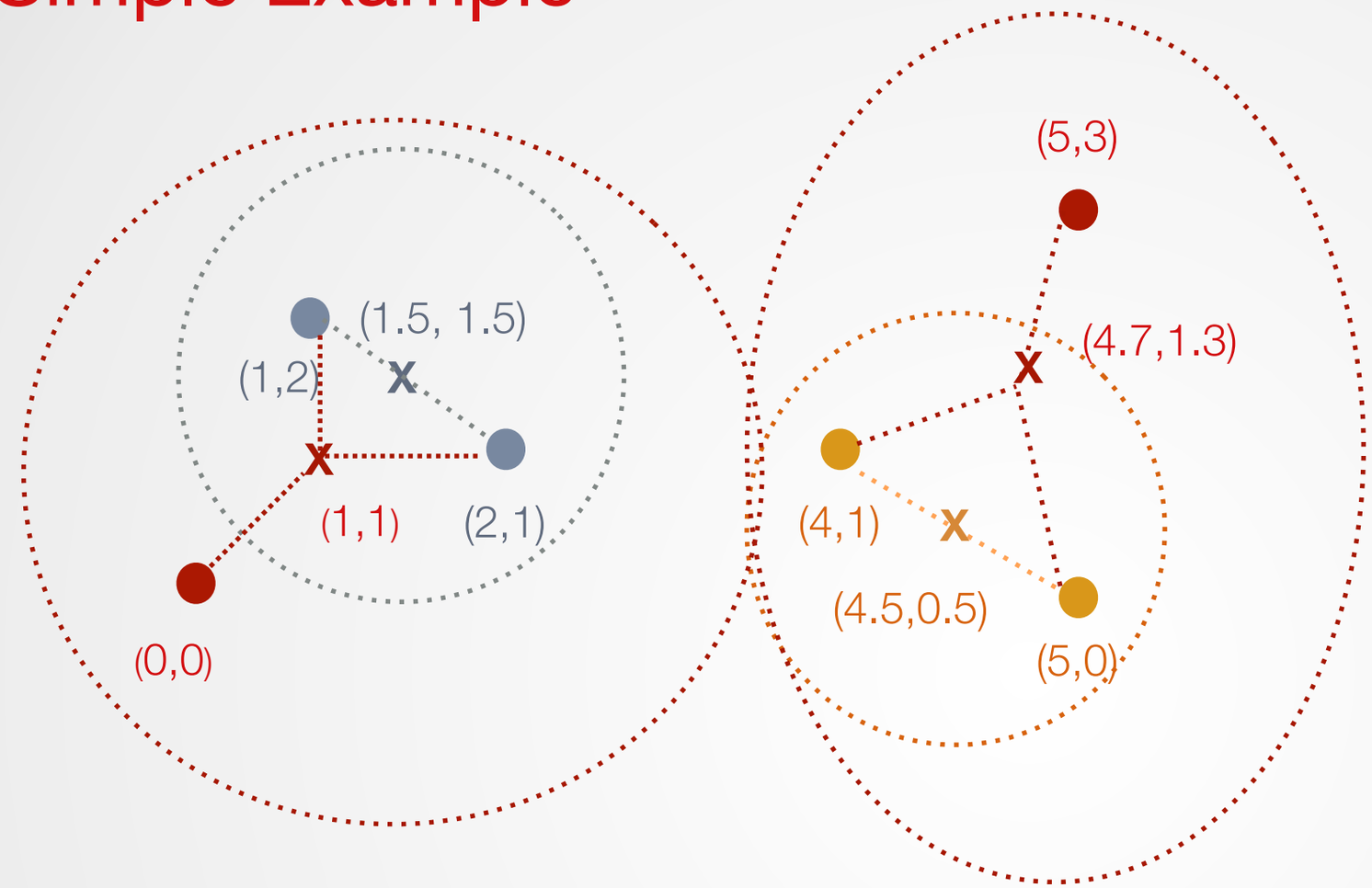
Simple Example



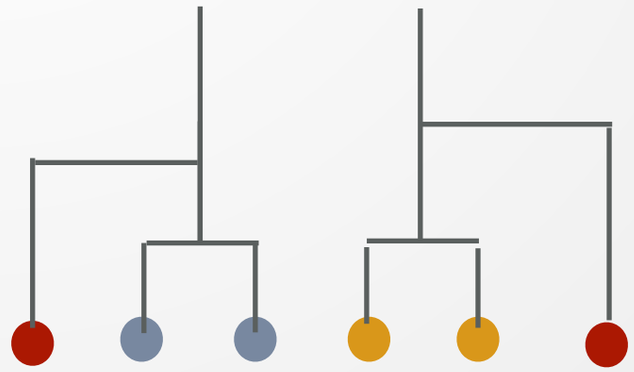
Data
x Centroid



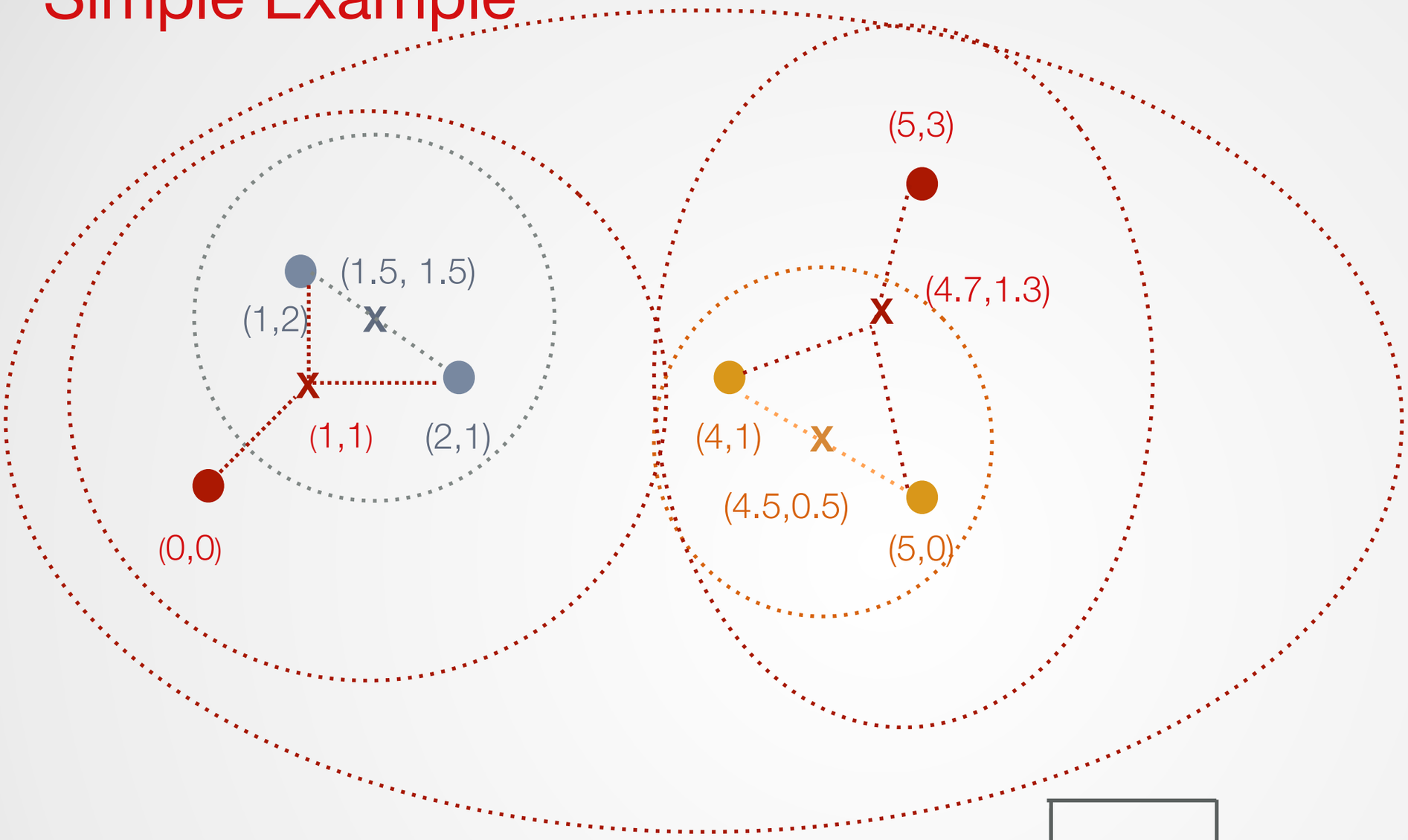
Simple Example



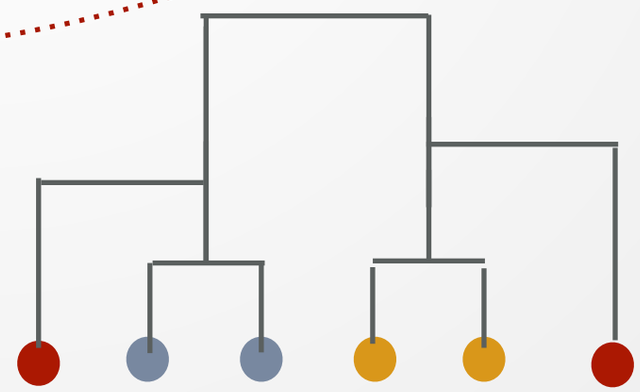
Data
Centroid



Simple Example



Data
Centroid



Swiss banknotes forgery

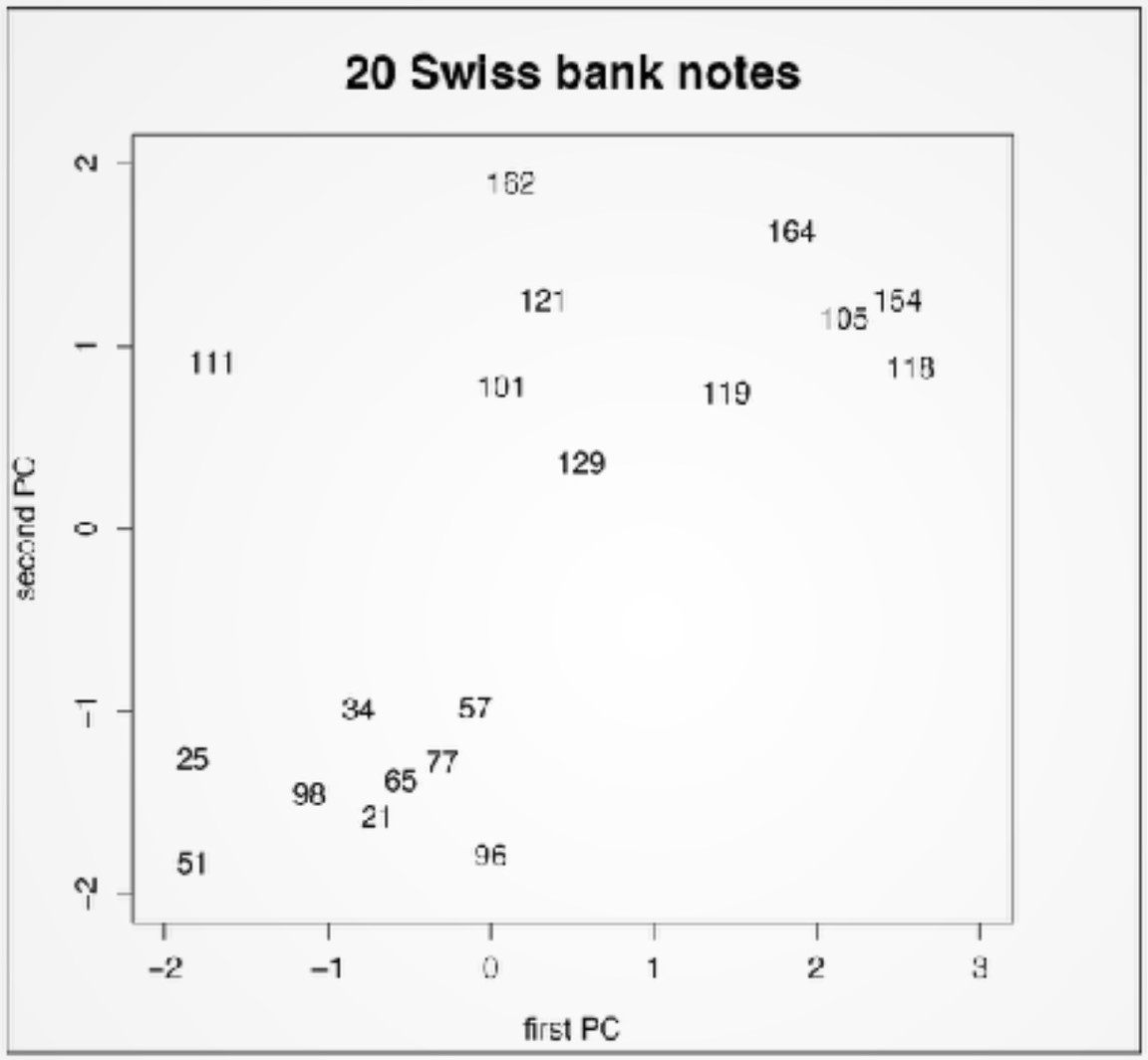
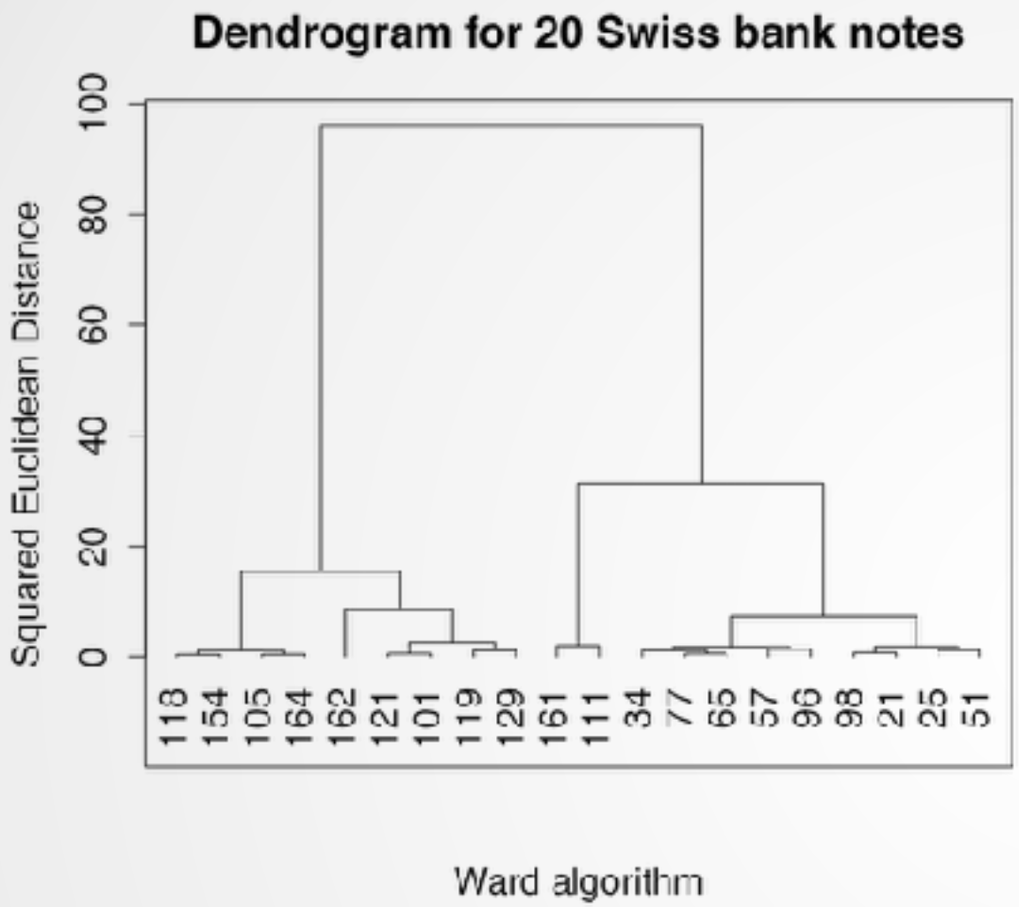
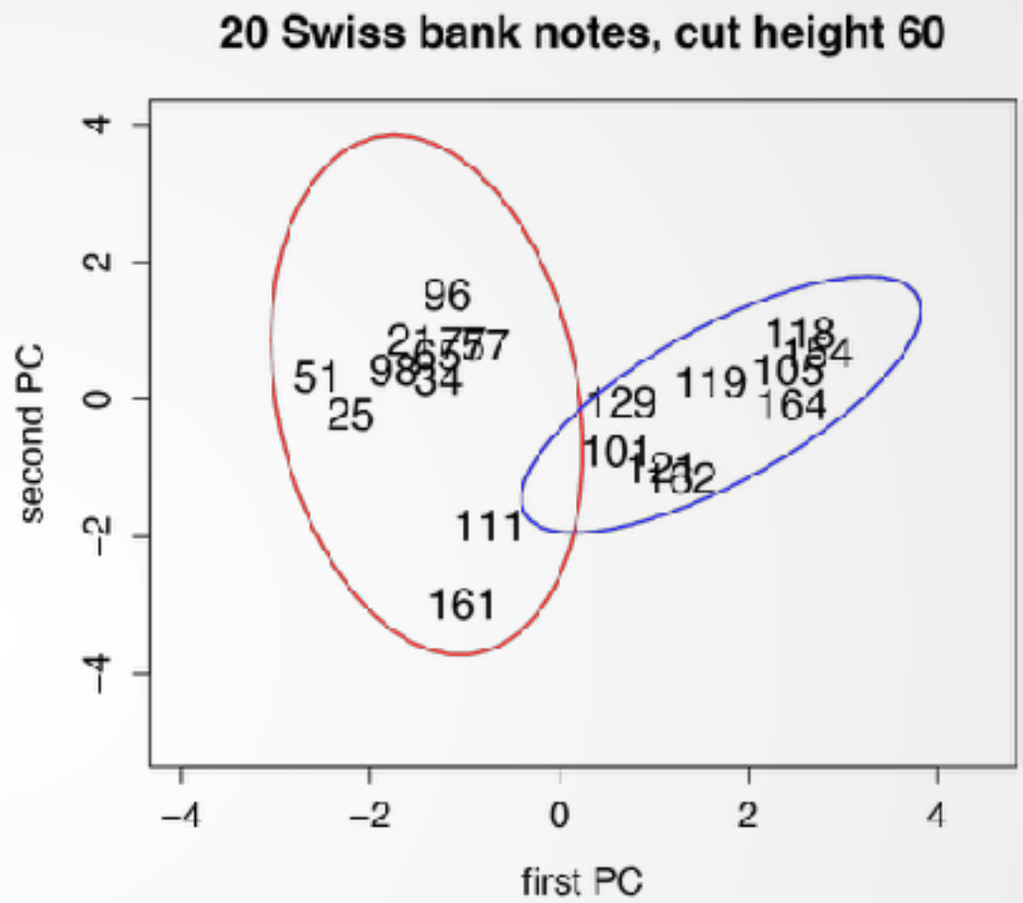


Figure: PCA for 20 randomly chosen bank notes MVAclusbank

Swiss banknotes forgery



(a) The dendrogram for the 20 bank notes, Ward algorithm



(b) PCA for 20 randomly chosen bank notes

MVAclusbank

French Food

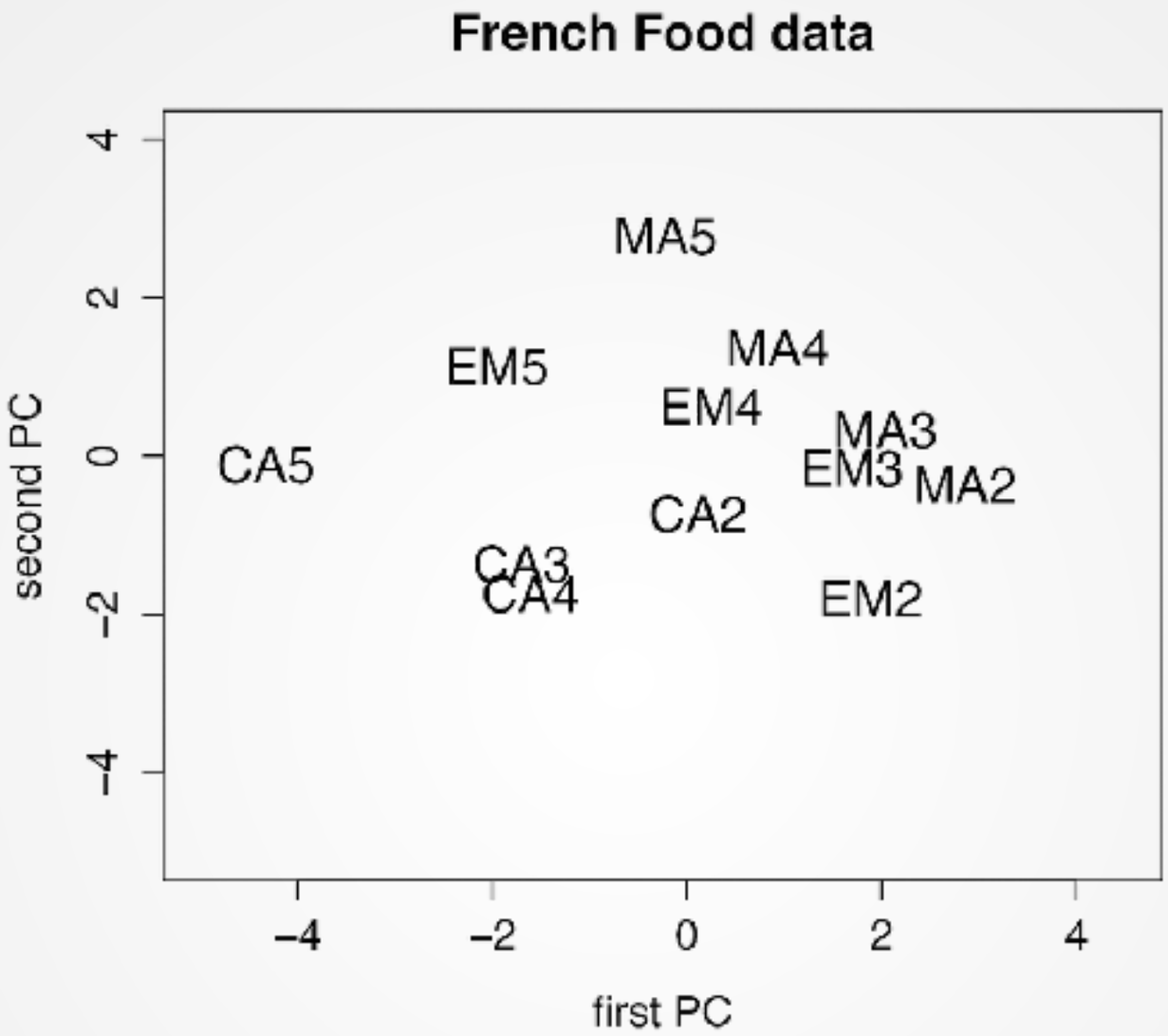


Figure: PCA for the standardised French food expenditures

MVAclusfood

Simple Example

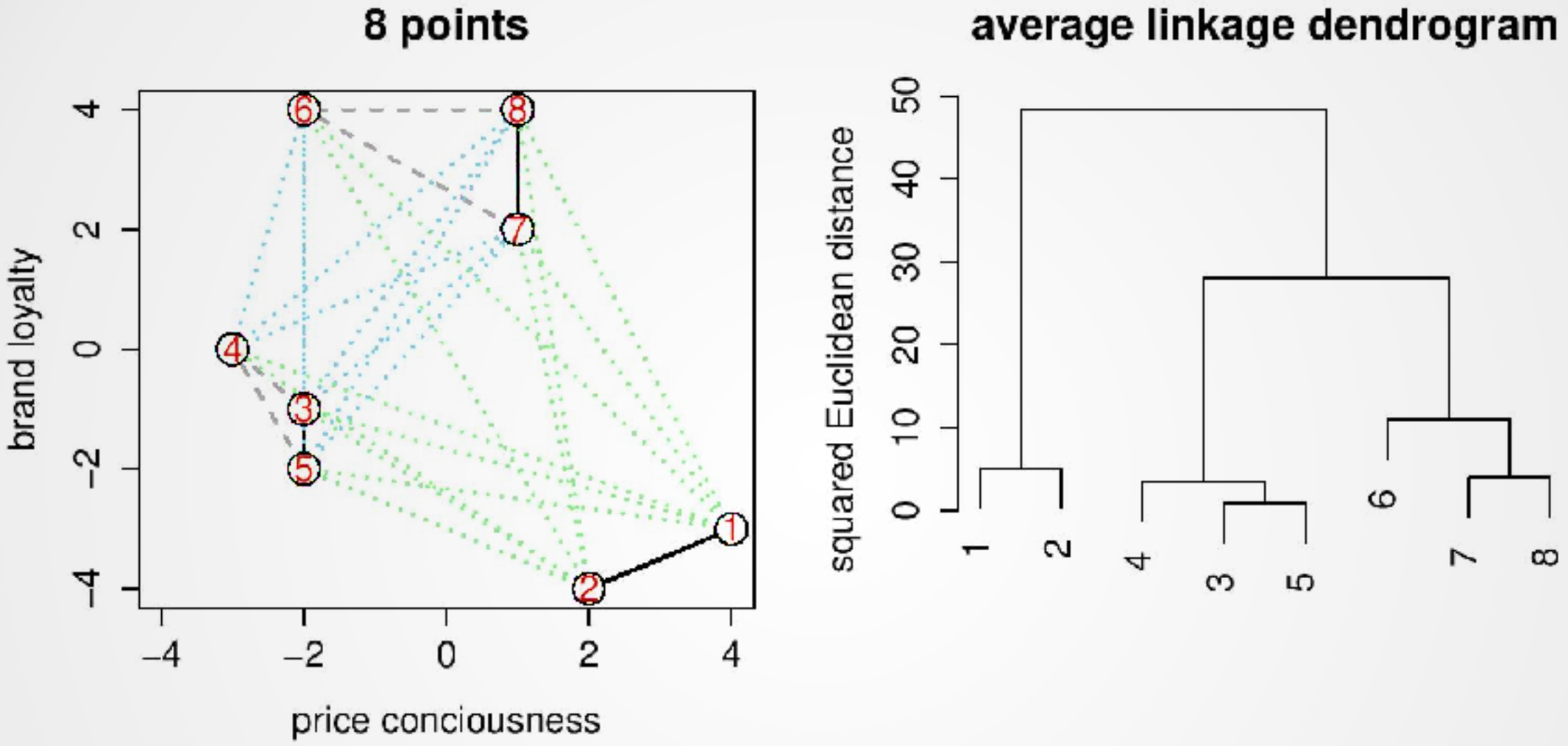
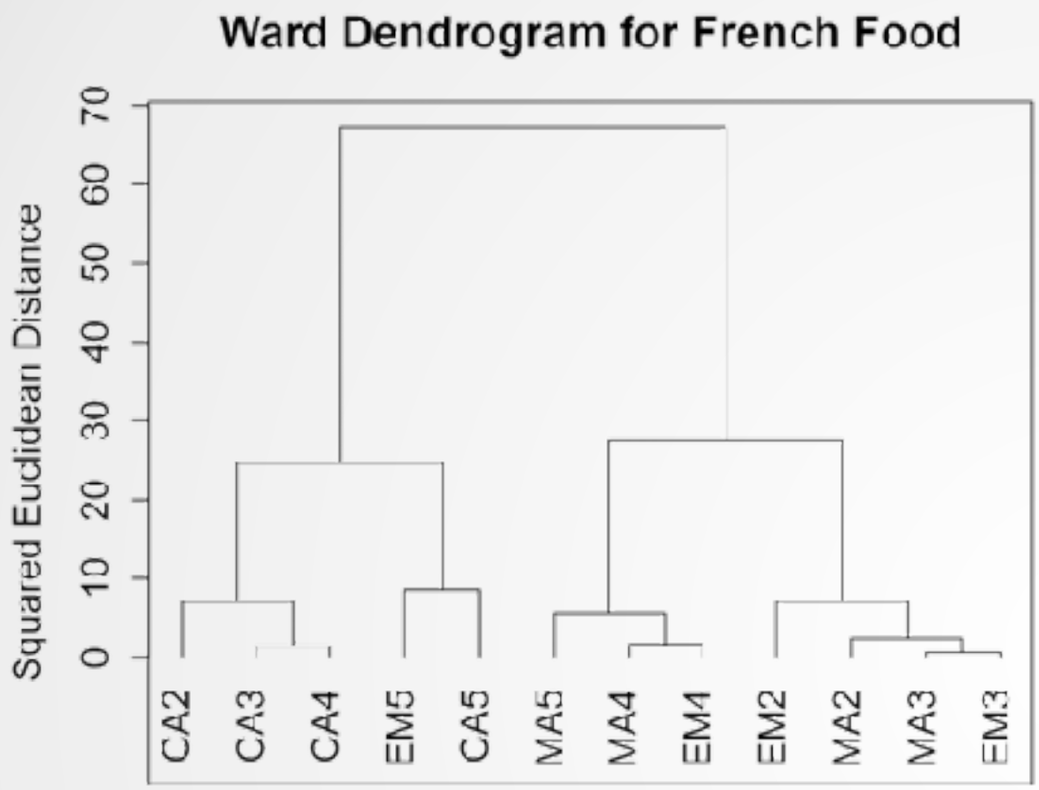
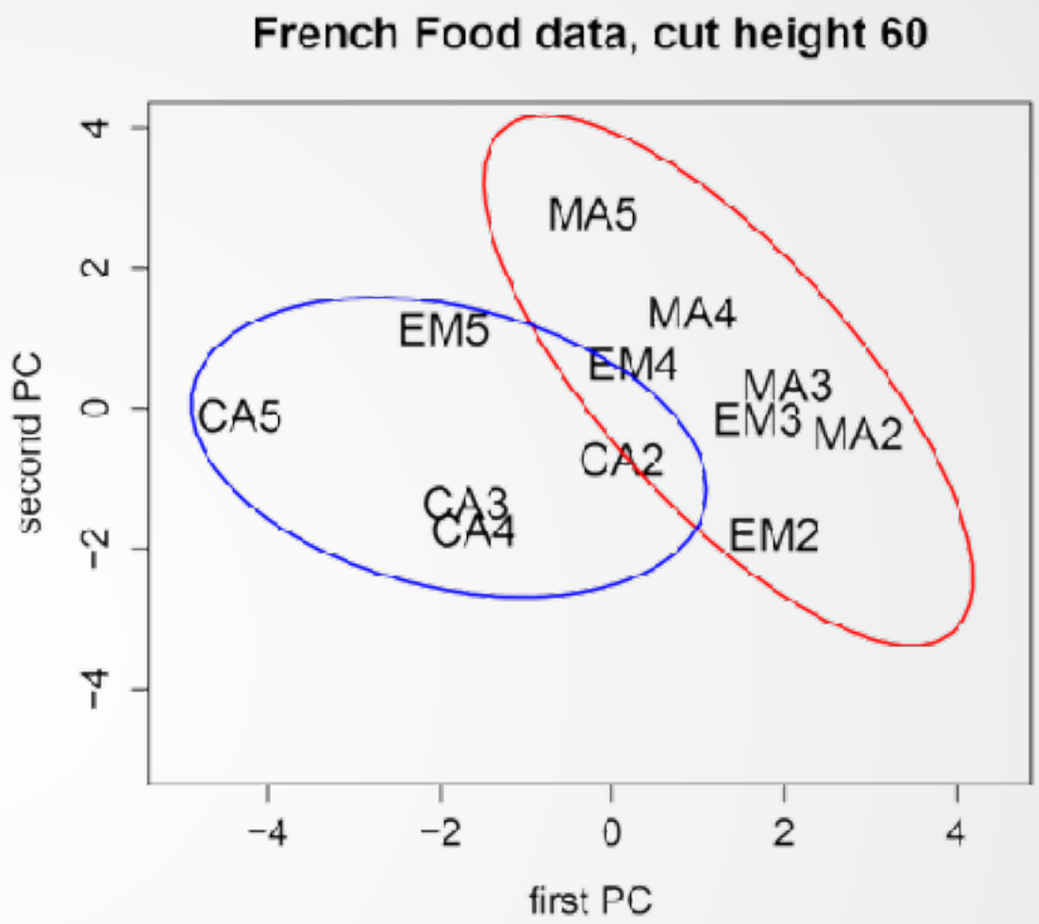


Figure: Average linkage algorithm on squared Euclidean distance for 8 point example with dendrogram. SMSclus8pa

French Food



(a) The dendrogram for the standardized French food expenditures, Wald algorithm



(b) PCA for the standardized French food expenditures

Figure 14: MVAclusfood

French Food

- Cluster 1: managers with 2-5 children, employees with 5 children
- Cluster 2: employees with 2-4 children, manual workers with 2-5 children

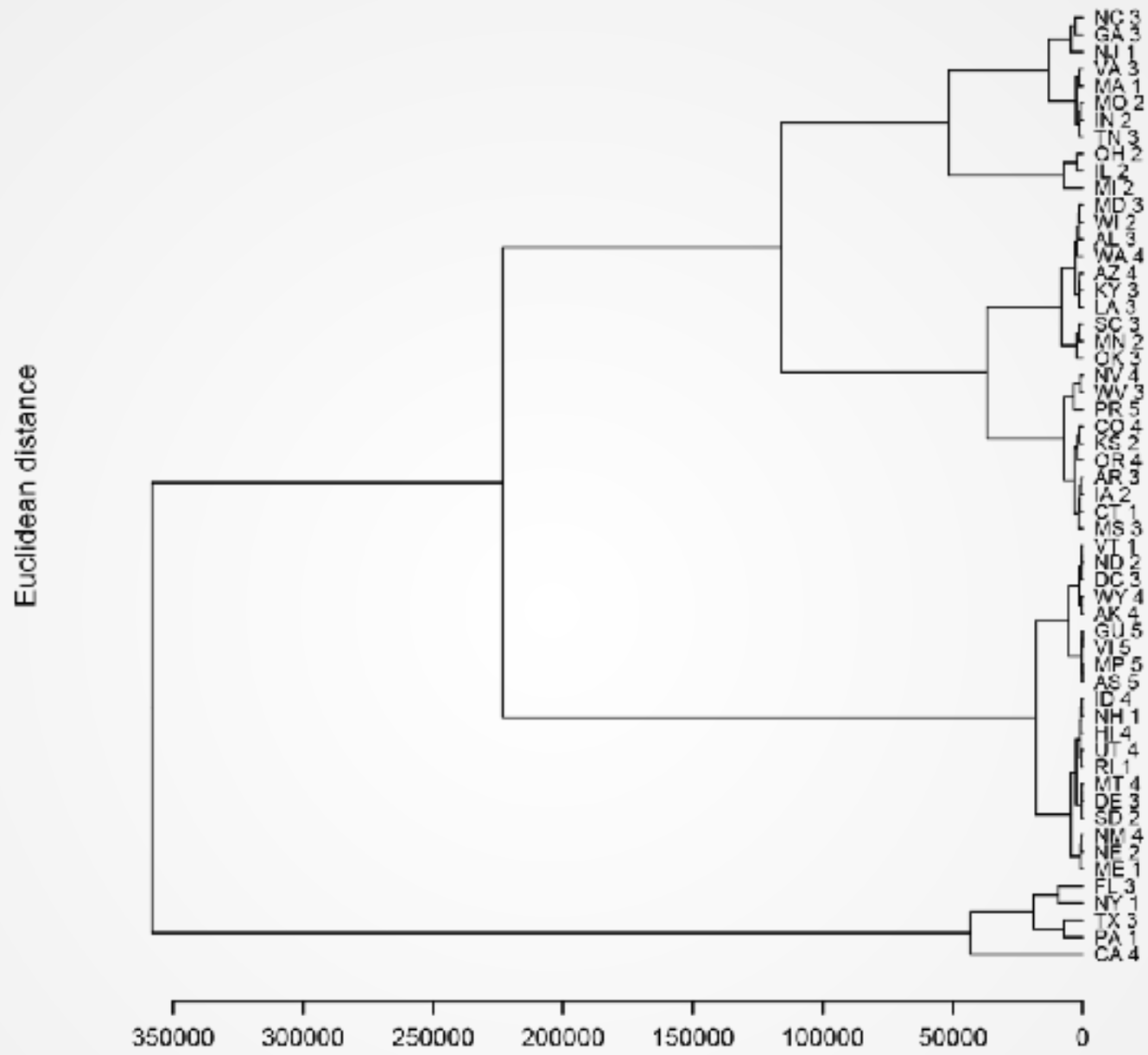
- Employment status more important than number of children
  - All managers in cluster 1, regardless of children
  - Only employees with 5 children also in cluster 1
  - All other groups in cluster 2, regardless of children

Example: US health data 2005

- *Perform the cluster analysis of the U.S. health 2005 data set.*
  Interest in the numbers of deaths related to diseases. Use Euclidean distance with Ward clustering.

Clusters	States										
1	IL	IN	MI	MO	OH	MA	NJ	GA	NC	TN	VA
2	IA	KS	MN	WI	CT	PR	AL	AR	KY	LA	MD
	MS	OK	SC	WV	AV	CO	NV	OR	WA		
3	NE	ND	SD	ME	NH	RI	VT	VI	GU	AS	MP
	DE	DC	AK	HI	ID	MT	NM	UT	WY		
4	NY	PA	FL	TX	CA						

US Health



Cluster analysis of U.S. health data set using Ward algorithm and Euclidean distance.

SMSclushealth05

US Health

	HIV	Malignant	Diabetes	Alzheimer	Heart
1	315,73	16597.09	2203.45	2104.36	19044.73
2	157.30	7727.10	1147.10	1150.15	8723.75
3	22.90	1626.70	237.65	221.05	1759.85
4	1198.60	38957.40	5219.80	4487.40	47907.80
	TIA	Influenza	Respiratory Diseases	Liver	Nephritis
1	4261.00	1837.18	3827.82	721.73	1500.82
2	2066.75	883.25	1918.45	372.50	657.75
3	422.10	182.40	405.50	84.85	105.70
4	9716.80	4519.60	8725.00	2140.40	2619.00

Table 3: The averages of the U.S. health data set within the 4 clusters.

SMSclushealth05

US Health

- Cluster 1: IL, IN, MI, MO, OH (Midwest), MA, NJ (Northeast), GA, NC, TN, VA (South)
- Medium to large states: population > 60 mio.
- Medium numbers of HIV, diabetes and hearth related deaths.
- High numbers of cancer (Maliagnacy) related deaths.
- Cluster 4: NY, PA, FL, TX, CA (regional inhomogeneous)
- Large states: population > 12 mio.
- Highest numbers of HIV, cancer, diabetes and heart related deaths.

US Health

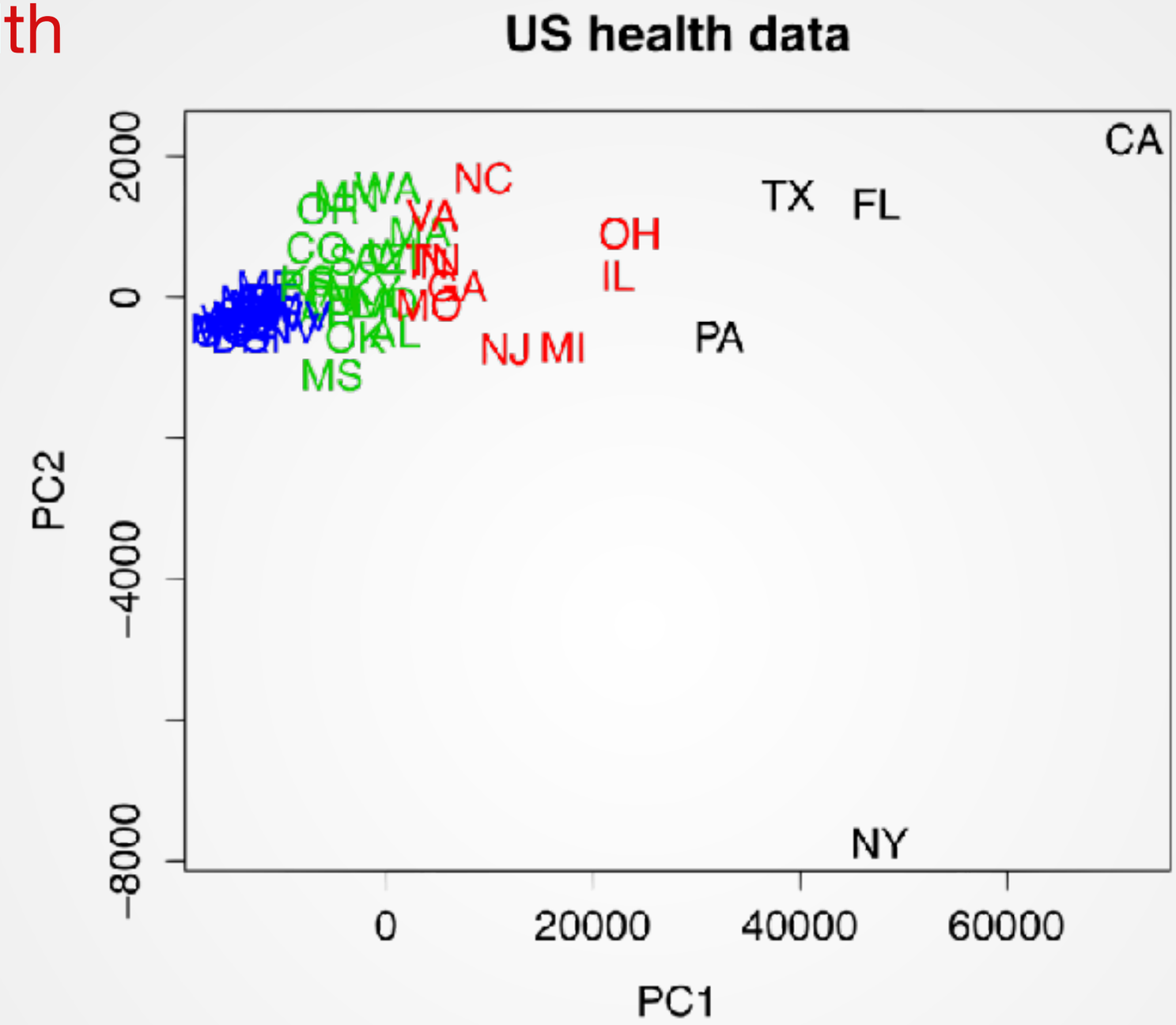


Figure: Plot of the first two principal components of the U.S. health 2005 data.

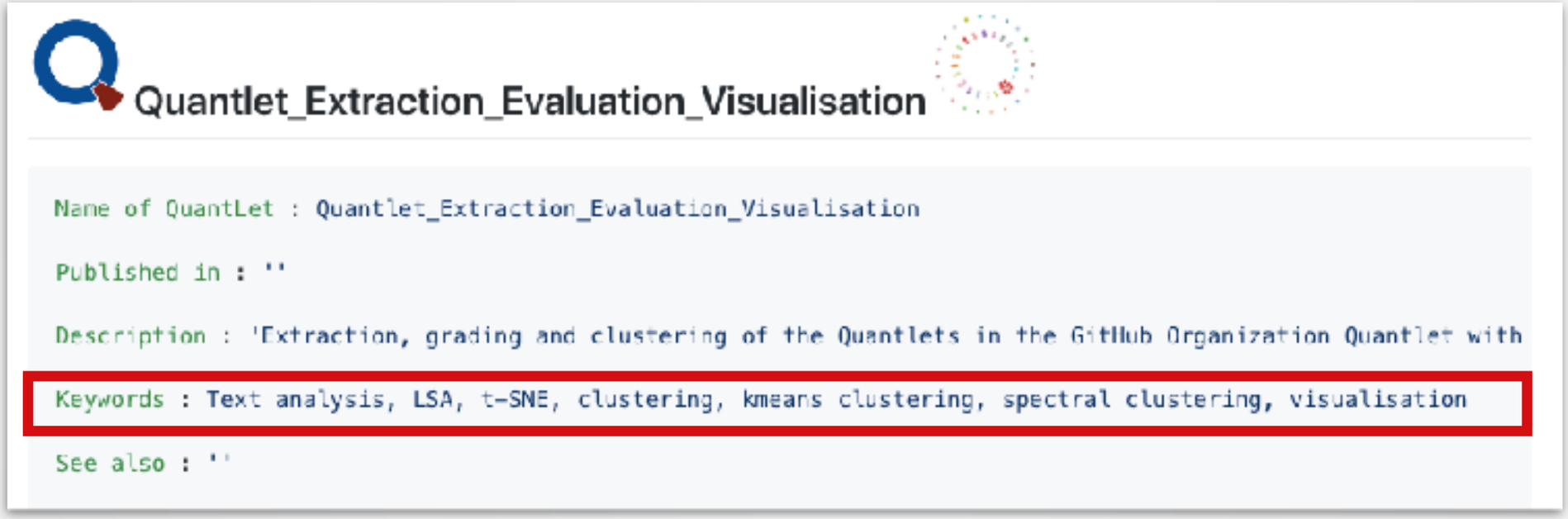
SMSclushealth5

Quantlets!

- Hierarchical clustering of Quantlets using keywords

Quantlet Website

Quantlet Metainfo Example



Quantlet_Extraction_Evaluation_Visualisation

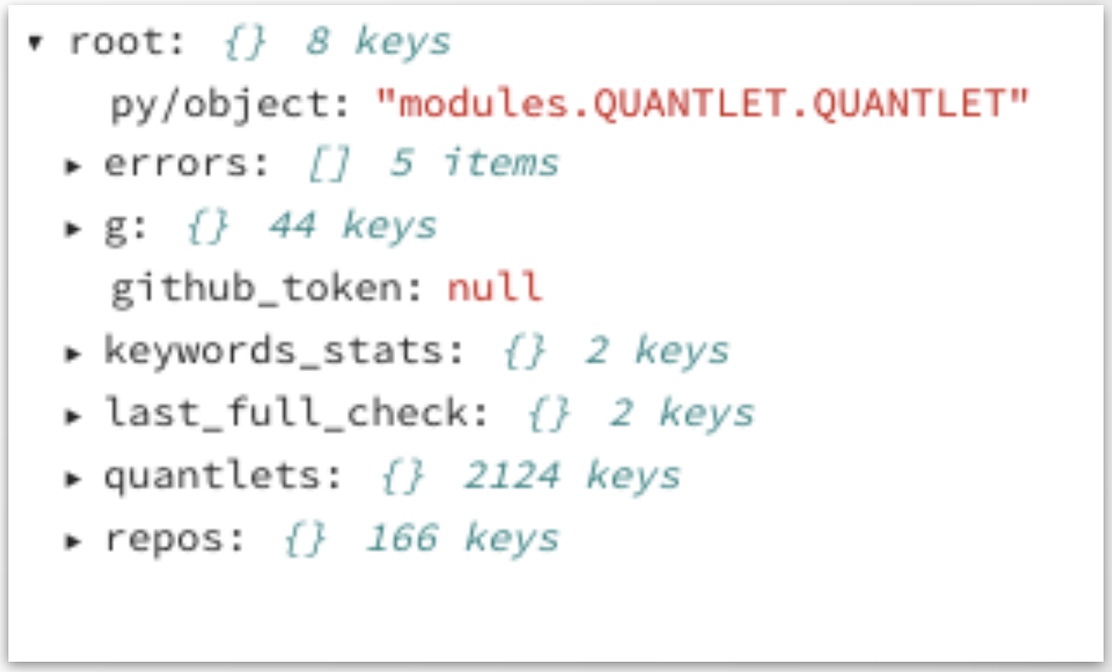
```
Name of QuantLet : Quantlet_Extraction_Evaluation_Visualisation

Published in : ''

Description : 'Extraction, grading and clustering of the Quantlets in the GitHub Organization Quantlet with

Keywords : Text analysis, LSA, t-SNE, clustering, kmeans clustering, spectral clustering, visualisation

See also : ''
```

Quantlet_Extraction_Evaluation_Visualisation

Data


```
"BCS/BCS_Binpdf/MetaInfo.txt": {"py/object": "modules.METAFILE.METAFILE",
"wagneran7@gmail.com", "name": "Palindroma"}, "comment_count": 0, "commit
"Palindroma"}, "message": "MDI\n\nChapter 3 complete", "tree": {"sha": "3
"https://api.github.com/repos/QuantLet/BCS/git/trees/3dc1ce0c7513a666fea2
"https://api.github.com/repos/QuantLet/BCS/git/commits/e5dc9c20a5ca99d6c3
"signature": null, "verified": false}}, "sha": "e5dc9c20a5ca99d6c303bb2ee
09T16:04:15Z", "email": "admin@quantnet.technology", "name": "Lukas Borke
"admin@quantnet.technology", "name": "Lukas Borke"}, "message": "auto cre
"https://api.github.com/repos/QuantLet/BCS/git/trees/a1b05133c04b288863c2
"https://api.github.com/repos/QuantLet/BCS/git/commits/701a6d4bcd2368579
"signature": null, "verified": false}}, "sha": "701a6d4bcd23685790e2b9468
{"comment": "", "description_length": 43, "description_length_wo_stopword
"submitted_year": "2018"}, "is_debuggable": true, "keyword_list": ["appro
"visualization"], "last_modified": "Mon, 28 May 2018 16:09:32 GMT", "meta
distribution is plotted for sample sizes 5, 10 and 100 and probabilities
values and show when it can likely be approximated by the normal distribu
representation, pdf, visualization", "name of quantlet": "BCS_Binpdf", "c
```

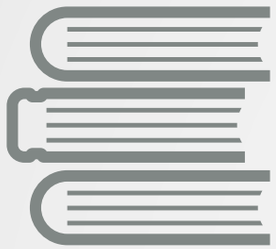

Original Json File



Data Tree Structure

Data Source

Preprocessing example



Example sentences:
„the cat sat on a mat“
„book from the shelf“

Dictionary of Words

1. the
2. cat
3. sat
4. on
5. a
6. mat
7. book
8. from
9. the
10. shelf

Remove stop words

cat sat mat

Tokenize to a Tensor

the cat sat on a mat
book from the shelf

cat	sat	mat	book	shelf
1	1	1	0	0
0	0	0	1	1

Modelling

- Distance Matrix of a subsample of 10 quantlets (Hamming Distance)
  - Given two vectors $u, v \in F^n$ we define the Hamming distance $d_{Hamming}(u, v)$ between u and v to be the number of places where u and v differ, divided by the dimension n of the vectors
- Example:

$$u_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, v_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, d_{Hamming}(u_1, v_1) = \frac{1}{3}$$

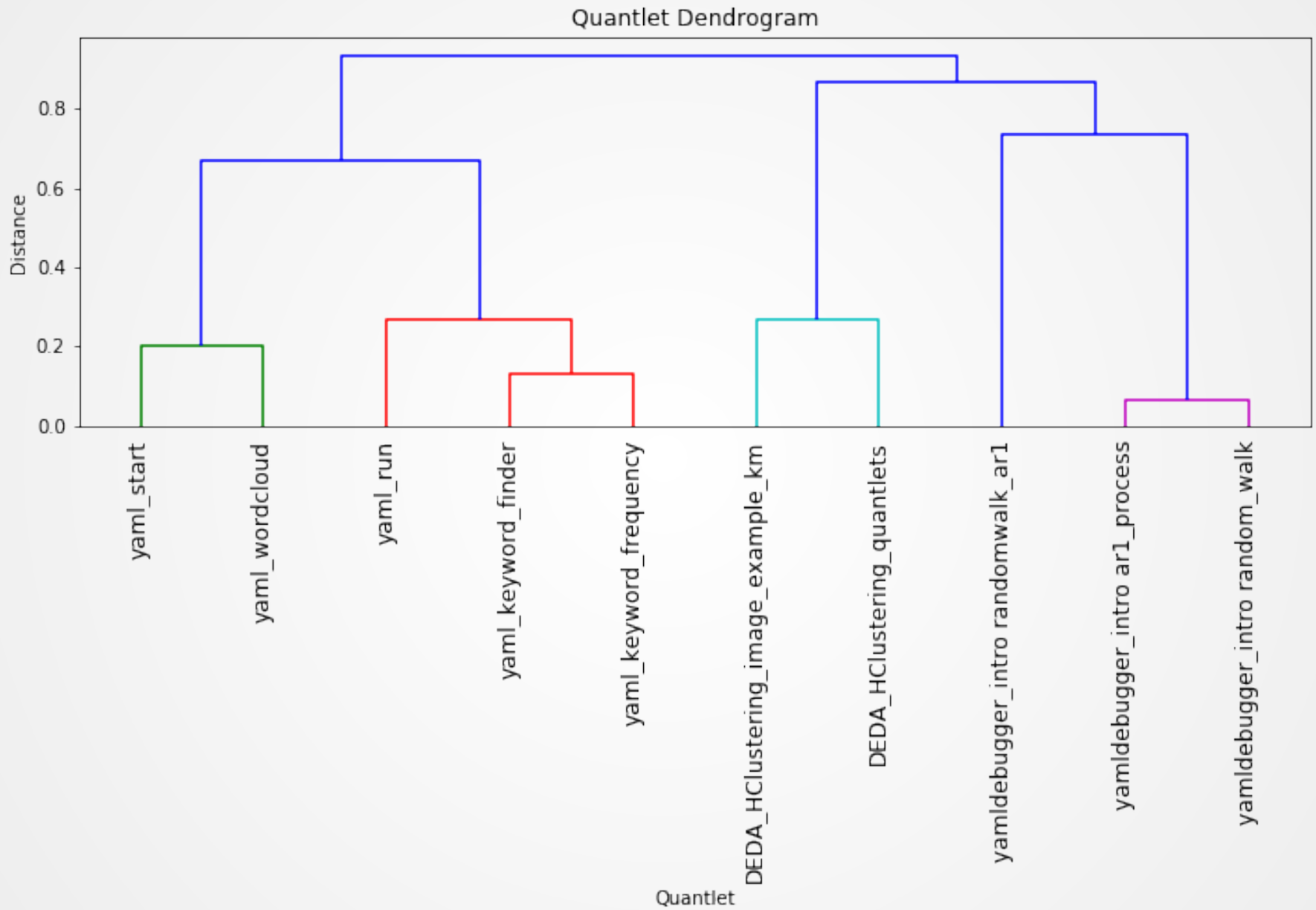
$$u_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, v_2 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, d_{Hamming}(u_1, v_1) = \frac{2}{3}$$

Modelling

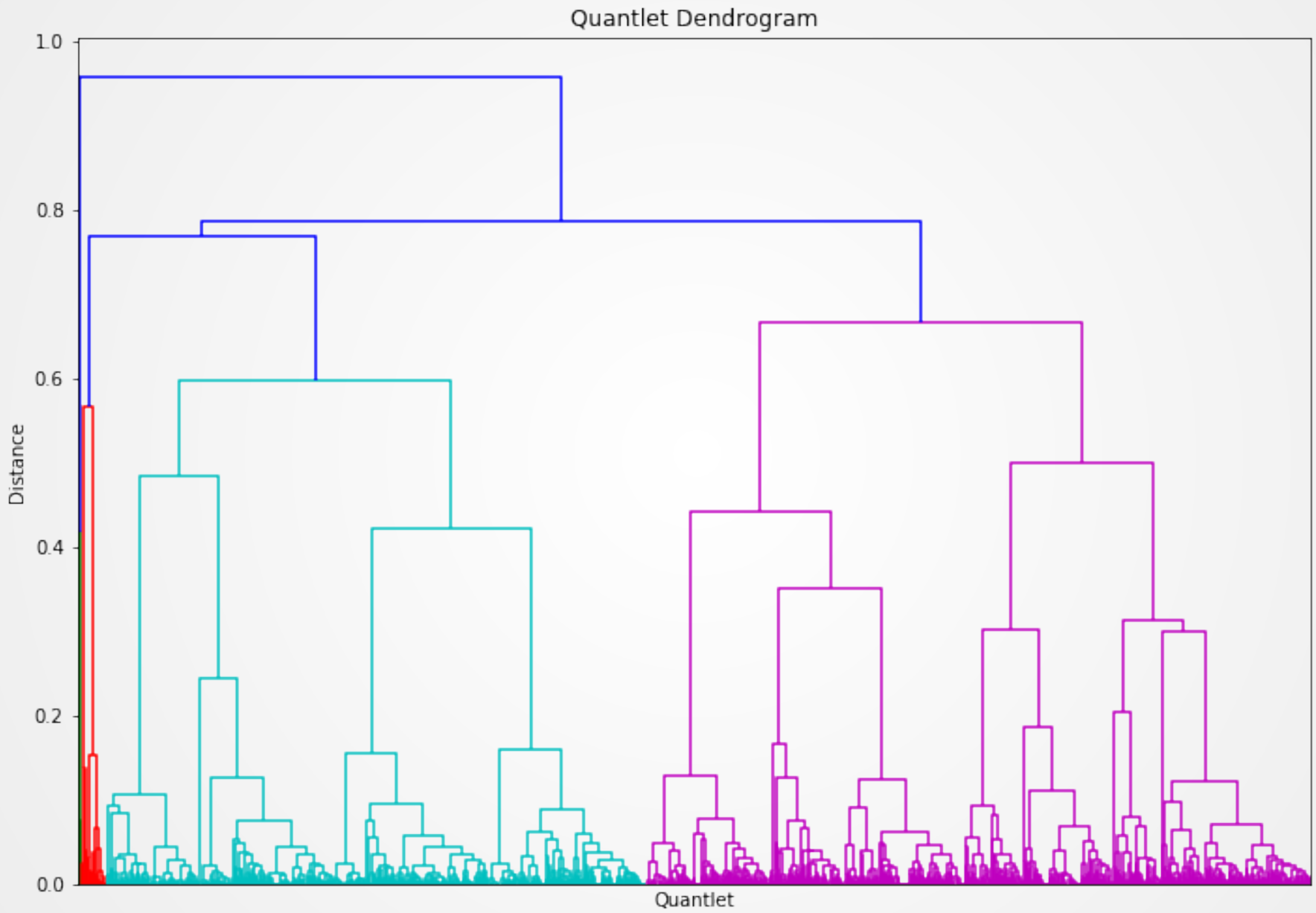
- Quantlet Hamming Distances:

$$D = \begin{pmatrix} 0. & 0.13 & 0.27 & 1. & 1. & 1. & 0.87 & 1. & 1. & 1. \\ 0.13 & 0. & 0.33 & 1. & 1. & 1. & 0.93 & 1. & 1. & 1. \\ 0.27 & 0.33 & 0. & 1. & 1. & 1. & 0.67 & 0.8 & 1. & 1. \\ 1. & 1. & 1. & 0. & 0.07 & 0.73 & 0.93 & 1. & 0.87 & 0.87 \\ 1. & 1. & 1. & 0.07 & 0. & 0.8 & 0.93 & 1. & 0.87 & 0.87 \\ 1. & 1. & 1. & 0.73 & 0.8 & 0. & 1. & 1. & 1. & 1. \\ 0.87 & 0.93 & 0.67 & 0.93 & 0.93 & 1. & 0. & 0.2 & 0.93 & 0.93 \\ 1. & 1. & 0.8 & 1. & 1. & 1. & 0.2 & 0. & 1. & 1. \\ 1. & 1. & 1. & 0.87 & 0.87 & 1. & 0.93 & 1. & 0. & 0.27 \\ 1. & 1. & 1. & 0.87 & 0.87 & 1. & 0.93 & 1. & 0.27 & 0. \end{pmatrix}$$

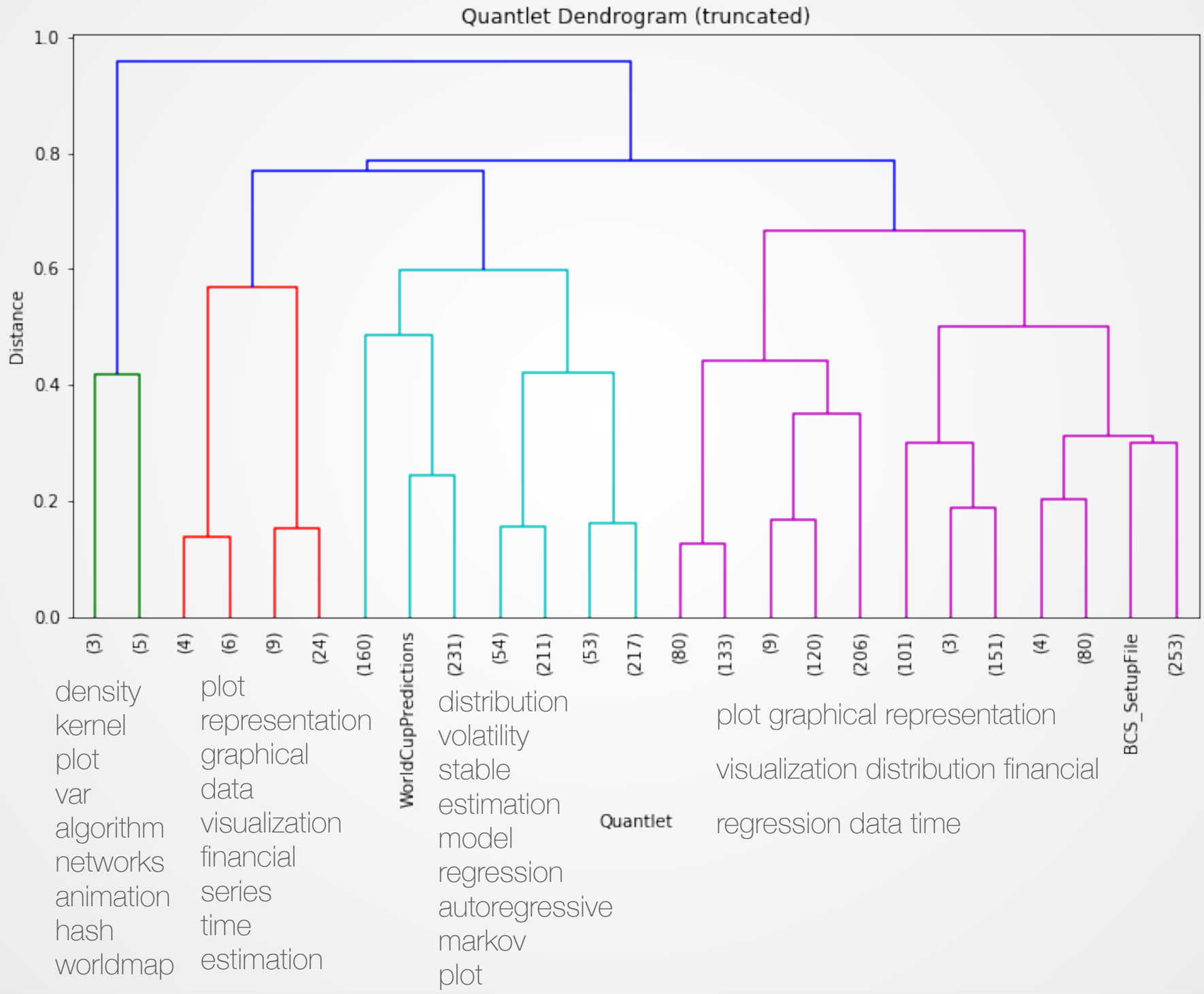
Modelling (cont.)



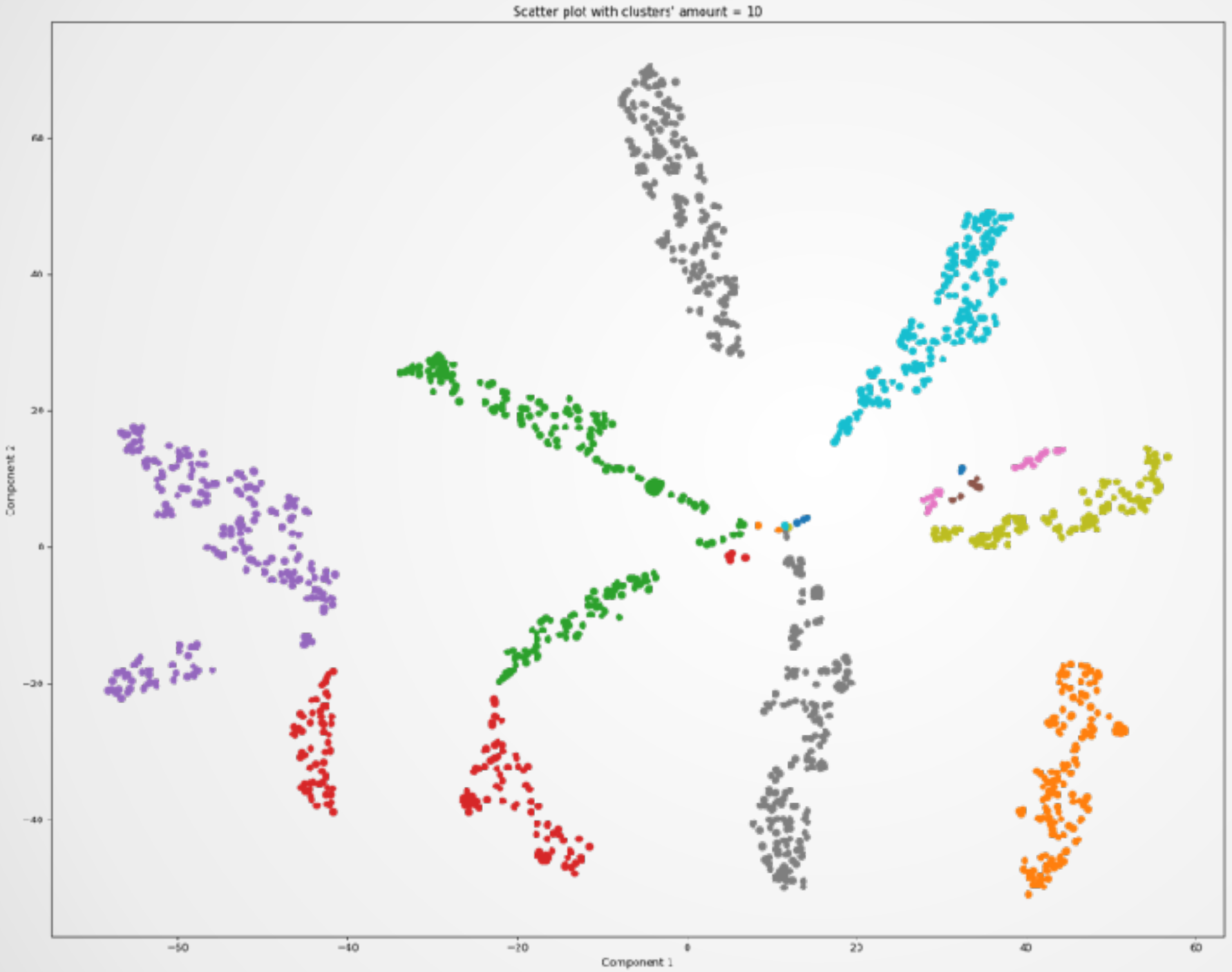
Modelling



Modelling (cont.)

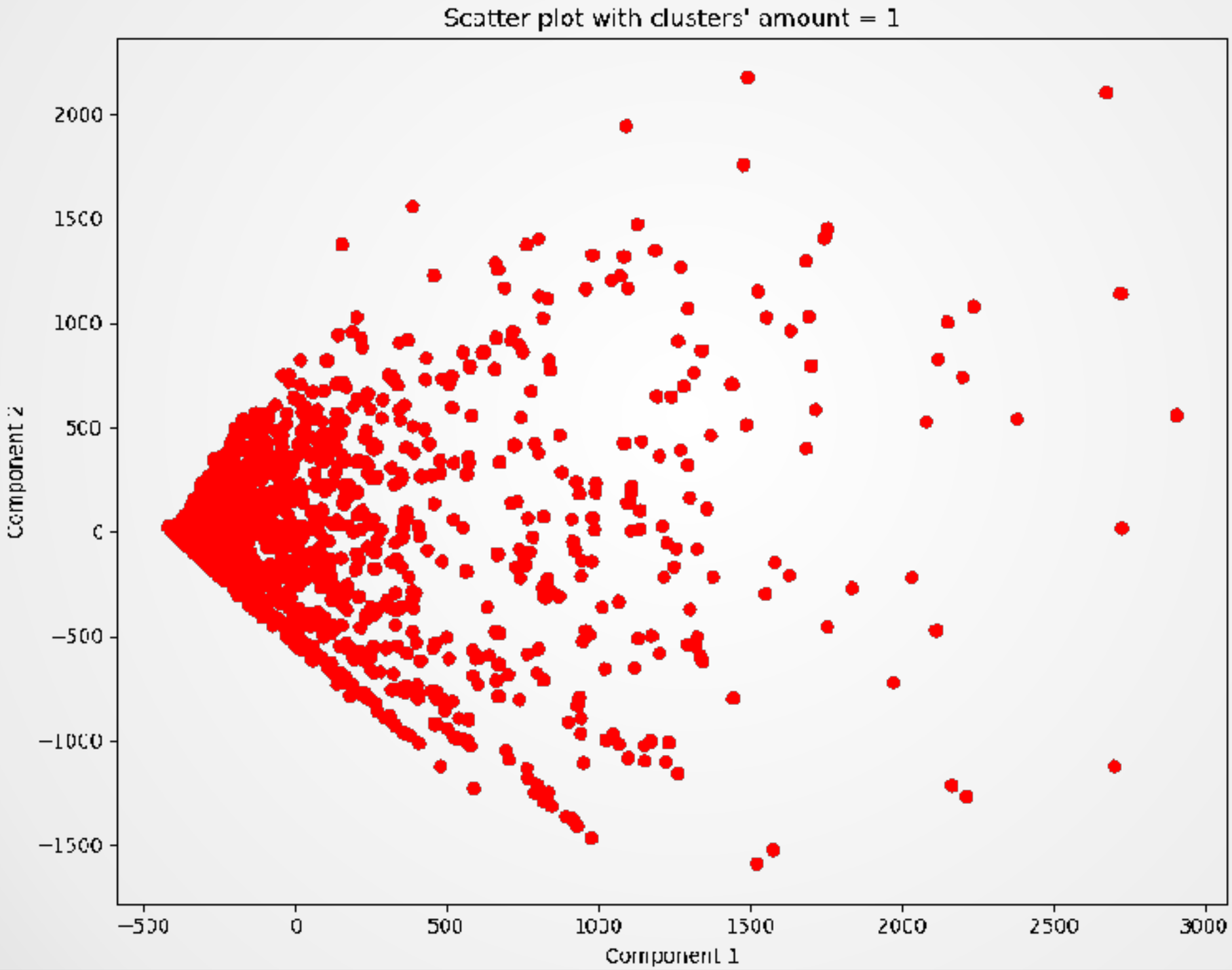


Visualisation *t*-SNE

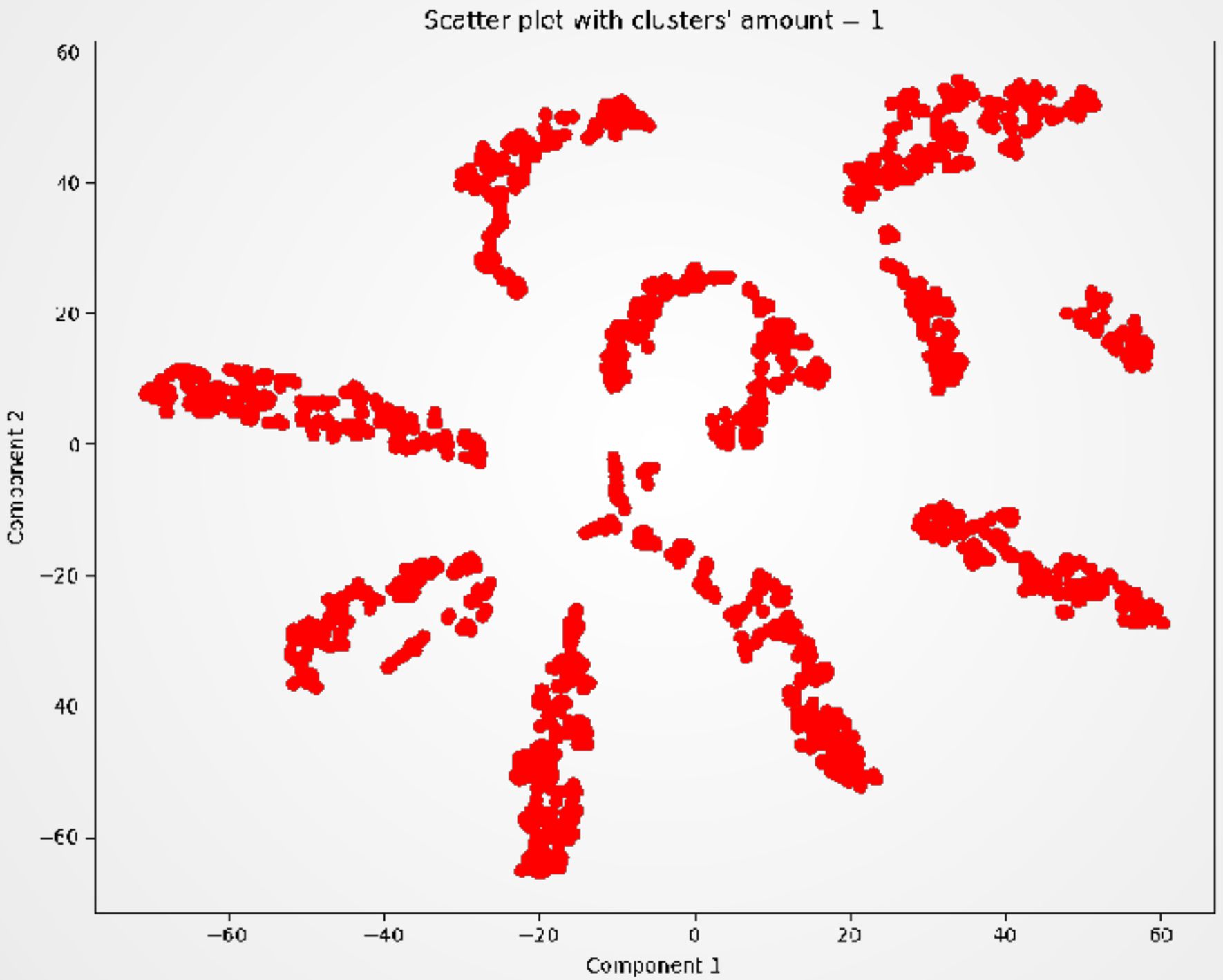


- density kernel plot animation world map
- plot financial time series distribution
- plot financial time simulation regression
- plot financial time series estimation
- data financial time series estimation
- plot clustering pareto qq capm
- volatility stable estimation autoregressive markov
- financial distribution data representation plot
- plot estimation regression density quantile
- plot simulation copula regression time

Dynamics of different amount of clusters (PCA)



Dynamics of different amount of clusters (t-SNE)

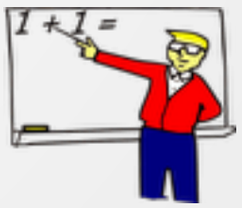


Discussion

- Useful if the underlying application has a taxonomy.
- Agglomerative hierarchical clustering algorithms are expensive in terms of their computational and storage requirements.
- Merges are final and cannot be undone at a later time, preventing global optimisation and causing trouble for noisy, high-dimensional data.

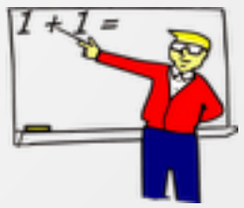
Summary: Cluster Algorithms

- The class of clustering algorithms can be divided into two types: hierarchical and partitioning algorithms.
- Hierarchical algorithms start with the finest (coarsest) possible partition and put groups together (split groups) from step to step.
- Partitioning algorithms start from a preliminary clustering and exchange group elements until a certain score is reached.
- The agglomerative procedure depends on the definition of the distance between two clusters. Often used distances are single linkage, complete linkage, Ward distance.



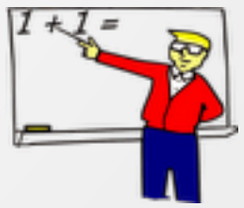
Summary: Cluster Algorithms

- The process of the unification of clusters can be graphically represented by a dendrogram.
- Hierarchical agglomerative techniques are frequently used in practice. They start from the finest possible structure (all data points form clusters), compute the distance matrix for all these clusters and join the clusters with the smallest distance. This step is replied until all points are united in one cluster.



Summary:

- After reading this chapter you should understand:
- Basic concepts of Cluster Analysis
- How the clustering algorithms work
- Different types of clustering algorithms
- The numerical aspects of Clustering
- Important: The distinction between proximity and dissimilarity



Links

- https://www-users.cs.umn.edu/~kumar001/dmbook/ch8.pdf
- https://www.youtube.com/watch?v=rg2cjfMsCk4
- https://www3.nd.edu/~rjohns15/cse40647.sp14/www/content/lectures/13%20-%20Hierarchical%20Clustering.pdf
- t-SNE: https://github.com/Divyagash/t-SNE/blob/master/tSNE_Presentation.pdf
- http://www.cs.toronto.edu/~hinton/absps/tsne.pdf

Hierarchical Clustering

Wolfgang Karl Härdle
Elizaveta Zinovyeva

Ladislaus von Bortkiewicz Professor of Statistics
Humboldt-Universität zu Berlin
BRC Blockchain Research Center
lvb.wiwi.hu-berlin.de
Charles University, WISE XMU, NCTU 玉山學者