Introduction to cluster analysis

L. Amendola

Comp. Stat. and Data Analysis

SS2025

What is cluster analysis?

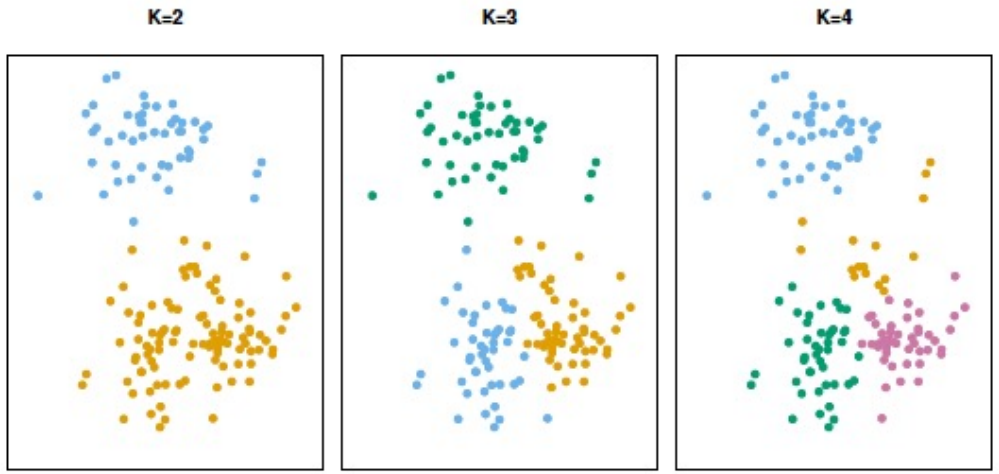
Finding subgroups or clusters in a dataset

Classify a number of objects (eg measurements) into groups in which an element of the group is *closer* to other elements of that group than to any other element

Is an example of *unsupervised learning*: we do not have a model to compare to

Difficult problem: there are K^n partitions of n points into K clusters !

Questions



Gareth James • Daniela Witten • Trevor Hastie • Robert Tibshirani
An Introduction to Statistical Learning
with Applications in R, Springer

How many groups are there?

How good the classification is?

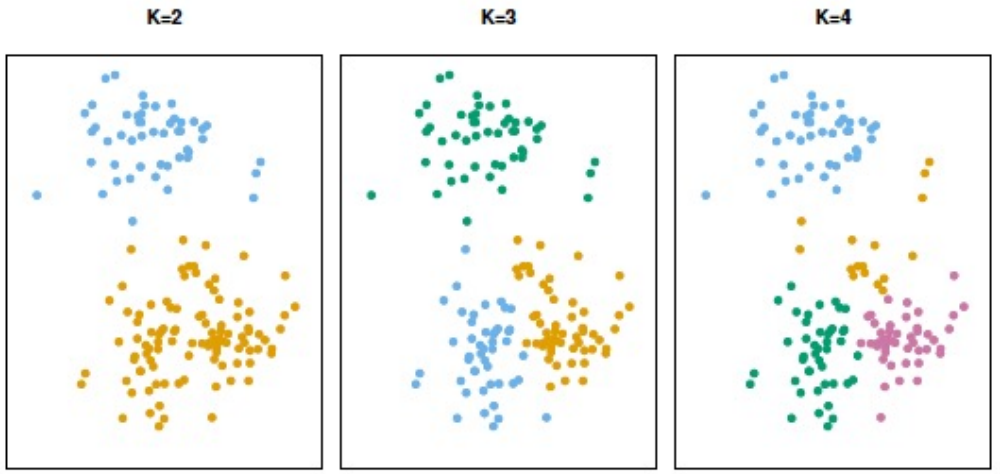
How to define “closer than”?

Examples

Science: we want to identify galaxies in a N-body simulations by collecting nearby “particles”

Sociology: we want to group people according to interests, work experience, training, etc

Marketing: we want to classify people in groups according to income, occupation, food habit, hobbies, etc

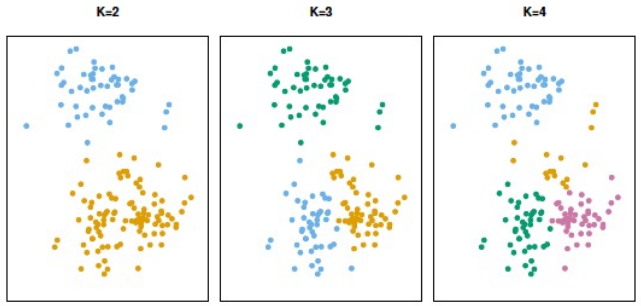


Gareth James • Daniela Witten • Trevor Hastie • Robert Tibshirani
An Introduction to Statistical Learning
with Applications in R, Springer

Here the two dimensions might be temperature versus pressure, household income versus education level, weight versus food habits, etc

Defining a Distance

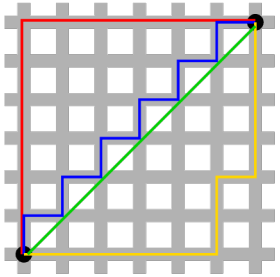
Gareth James • Daniela Witten • Trevor Hastie • Robert Tibshirani
An Introduction to Statistical Learning
with Applications in R, Srpinger



Here the two dimensions might be temperature versus pressure, household income versus education level, weight versus food habits, etc…but then, which is the “distance” between points?

Euclidean distance $$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{ip} - x_{jp})^2}$$

Manhattan distance



Wikimedia, public domain

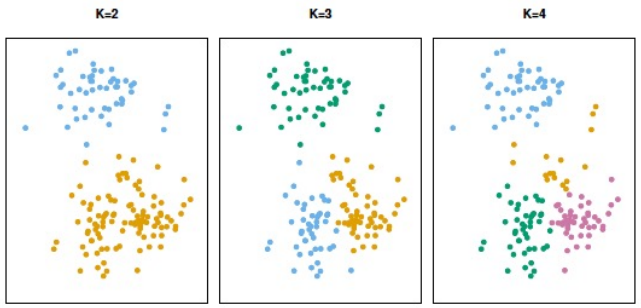
Distances

Names	Formula
Euclidean distance	$\|a-b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$
Squared Euclidean distance	$\|a-b\|_2^2 = \sum_i (a_i - b_i)^2$
Manhattan (or city block) distance	$\|a-b\|_1 = \sum_i \|a_i - b_i\|$
Maximum distance (or Chebyshev distance)	$\|a-b\|_\infty = \max_i \|a_i - b_i\|$
Mahalanobis distance	$\sqrt{(a-b)^\top S^{-1} (a-b)}$ where S is the Covariance matrix

Wikipedia

Defining a Distance between clusters

Gareth James • Daniela Witten • Trevor Hastie • Robert Tibshirani
An Introduction to Statistical Learning
with Applications in R, Srpinger

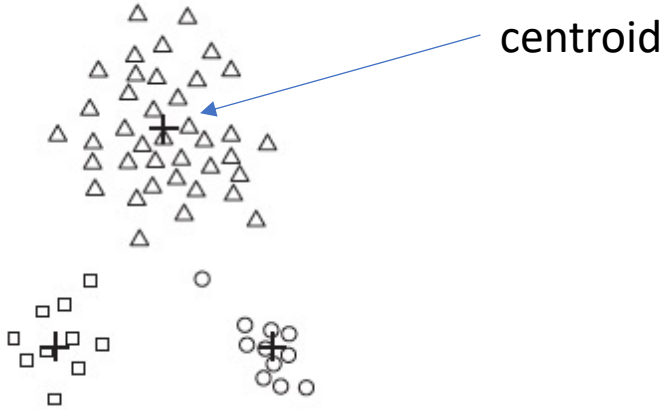


Measuring the distance between every pair might be too demanding.
We can define a distance between clusters, for instance centroids

For a Euclidean distance, the centroid is just the mean position of the cluster's members:

$$\mathbf{c}_i = \frac{1}{m_i} \sum_{\mathbf{x} \in C_i} \mathbf{x}$$

C_i = i-th cluster
c_i = i-th centroid
m_i = number of points in i-th cluster



Introduction to Data Mining
By Tan, Steinbach, Kumar

Algorithms

If you know the number of clusters and the data are roughly globular: K-means

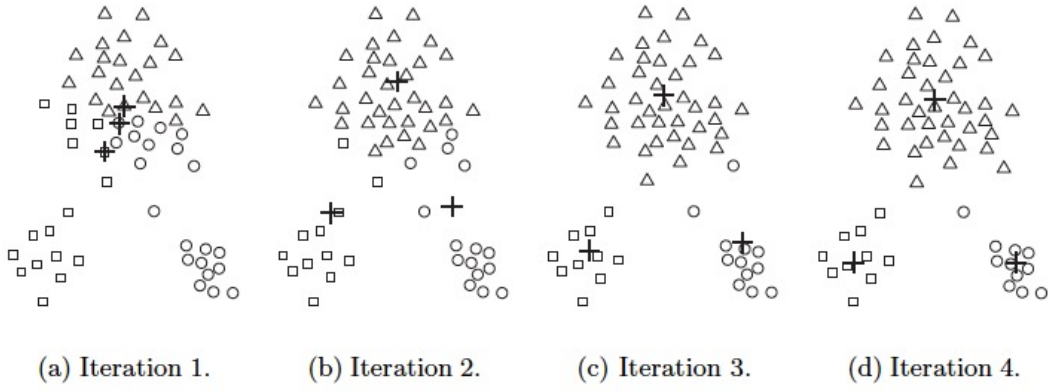
otherwise: Hierarchical Clustering

If you know the physics of clustering: Friends-of-friends algorithm

Many more (eg, density based) and many variants!

K-means

Introduction to Data Mining
By Tan, Steinbach, Kumar



Each point is a “measurement”

General algorithm:

- A) Choose K, the number of clusters
- B) Generate K random points (initial centroids)
- C) Assign each measurement to the closest centroid
- D) Determine new centroids
- E) repeat from C) until the centroids do not change anymore

K-means

General algorithm:

- A) Choose K, the number of clusters
- B) Generate K random points (initial centroids)
- C) Assign each measurement to the closest centroid
- D) Determine new centroids
- E) repeat from C) until the centroid do not change anymore

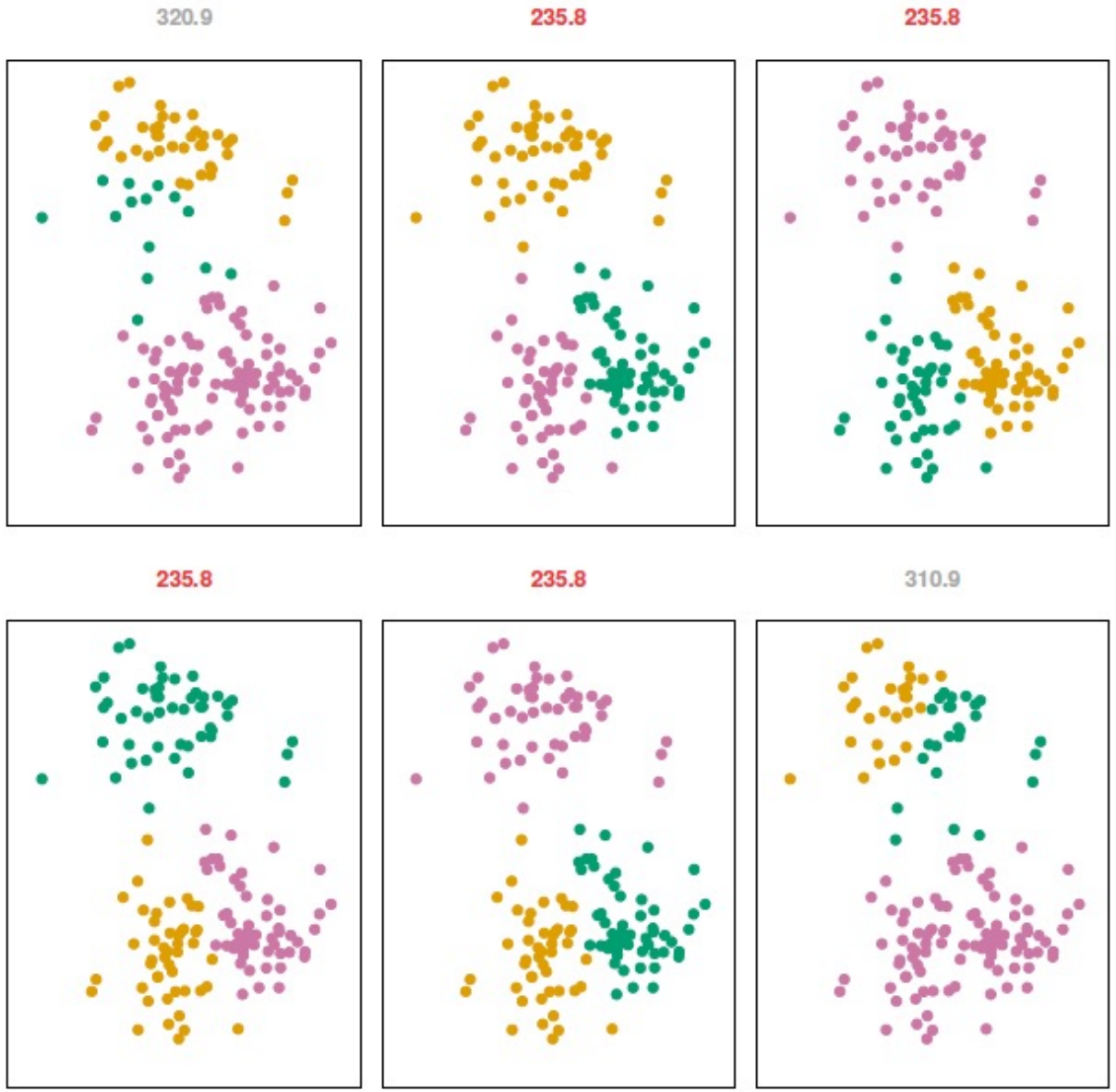
The K-means algorithm is guaranteed to find a *local minimum* of the objective function, in this case the total within-cluster squared distance sum

Objective function: Sum over clusters (average of within-clusters squared distances)

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}.$$

Gareth James • Daniela Witten • Trevor Hastie • Robert Tibshirani
An Introduction to Statistical Learning
with Applications in R, Springer

C_k = k-th cluster
$|C_k|$ = number of points in k-th cluster
p =dimensions
x_{ij} =i-th coordinate of j-th point

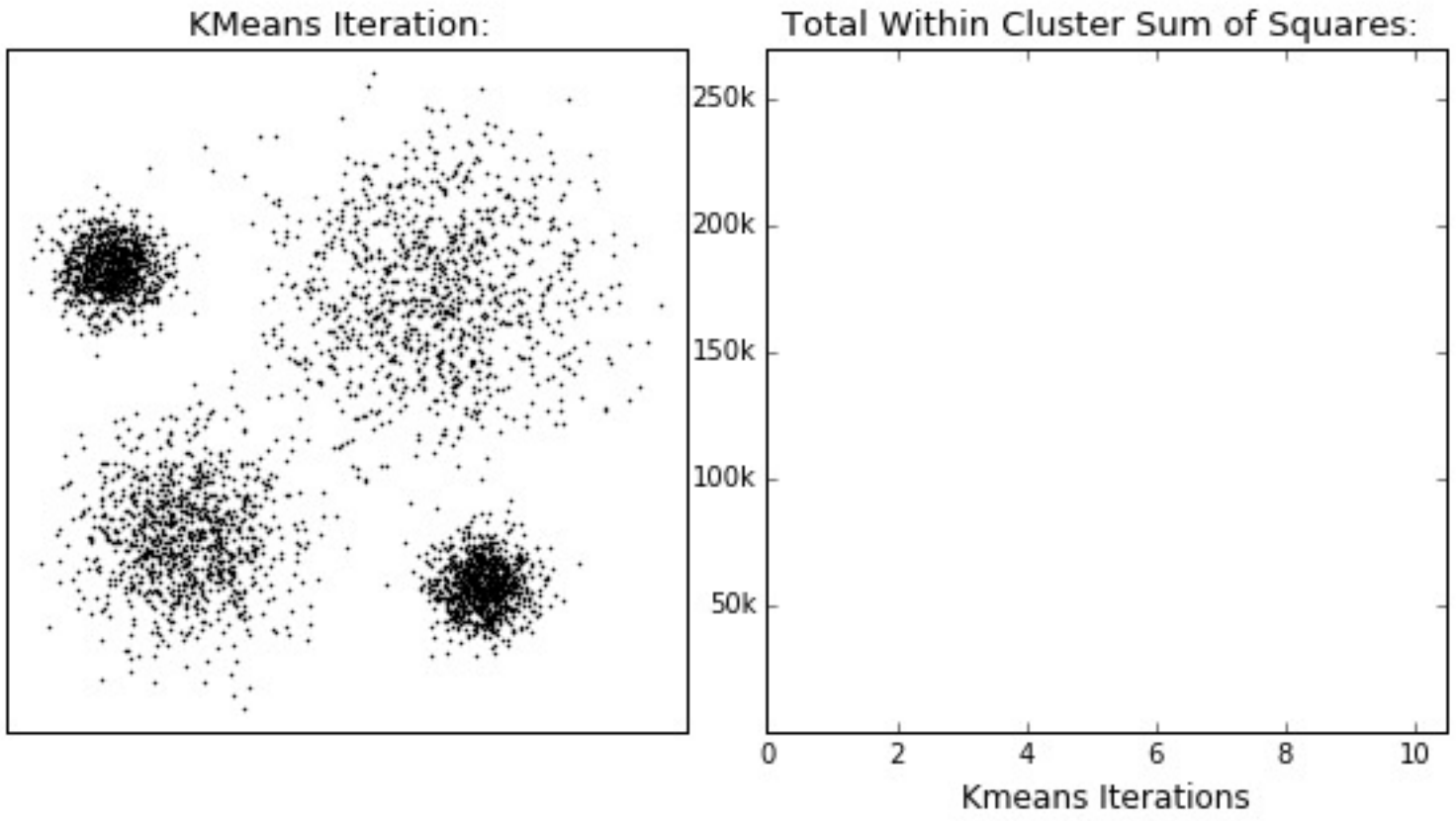


Several different local minima depending on the initial choice

Choose the one with the least value of the objective function

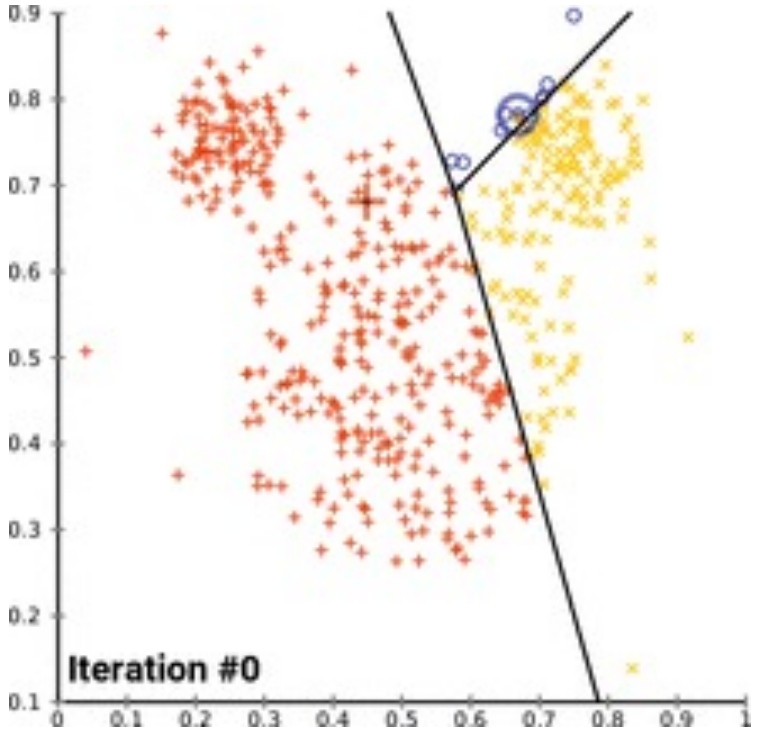
Gareth James • Daniela Witten •
Trevor Hastie • Robert Tibshirani
An Introduction to Statistical
Learning with Applications in R, Springer

K-means



David Sheehan
dashee87.github.io/data%20science/general/Clustering-with-Scikit-with-GIFs/

K-means

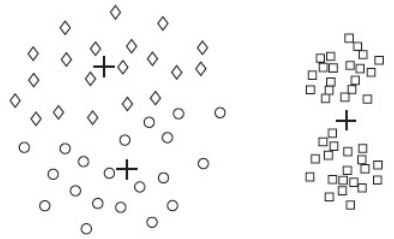


Chire, Wikimedia

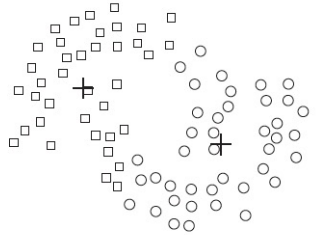
Pro’s and contra’s of K-means

Pros and contras

- Simple and efficient for small datasets, small k
- Does not handle well non-globular clusters
- Does not handle well clusters of very different sizes and densities
- Problems with outliers
- Complexity O(n^{kD+1}) for full convergence, O(nkDi) after i iterations

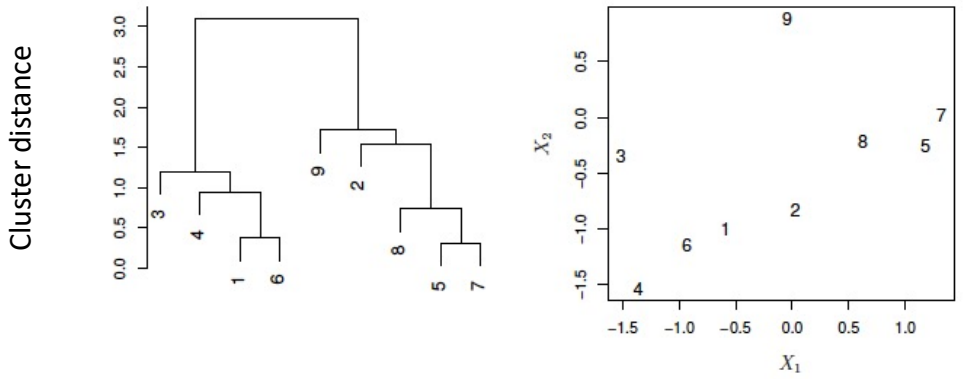


Unequal densities



Non-globular clusters

Hierarchical clustering



Dendrogram Data

Gareth James • Daniela Witten • Trevor Hastie • Robert Tibshirani
An Introduction to Statistical Learning with Applications in R, Springer

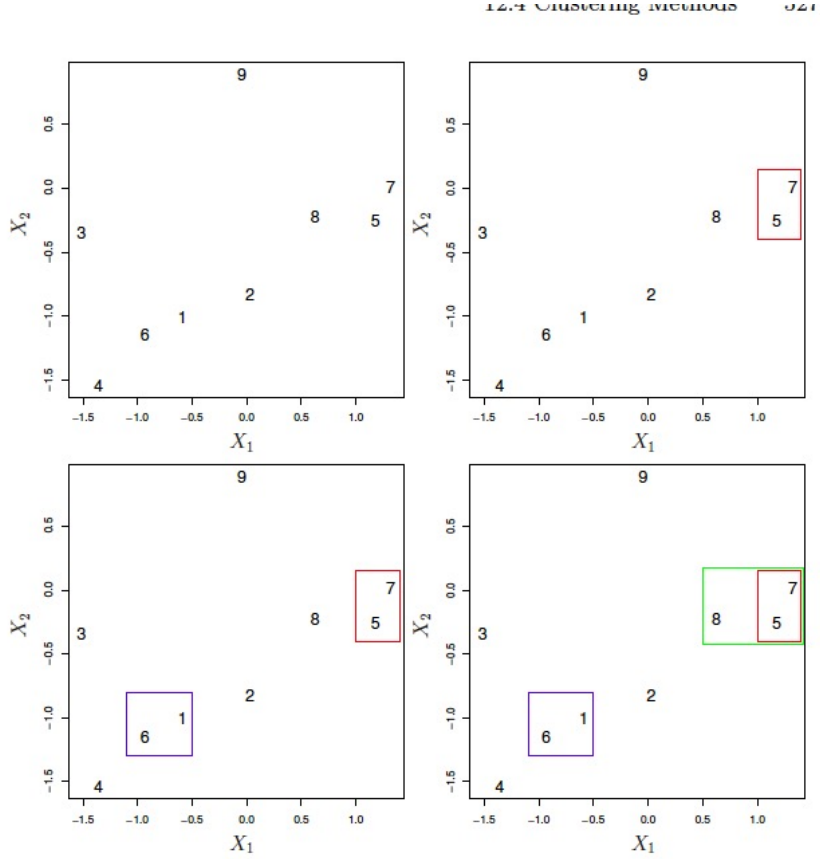
General algorithm (agglomerative HC):

- A) Initially, each of the *m* points is a cluster
- B) cluster the pairs that are closest (eg 1-6 and 5-7 in the figure above)
- C) Repeat B) until there is only one cluster
- D) build the dendrogram

Hierarchical clustering

General algorithm (agglomerative HC):

- A) Initially, each of the *m* points is a cluster
- B) cluster the pairs that are closest (eg 1-6 and 5-7 in the figure above)
- C) Repeat B) until there is only one cluster
- D) build the dendrogram



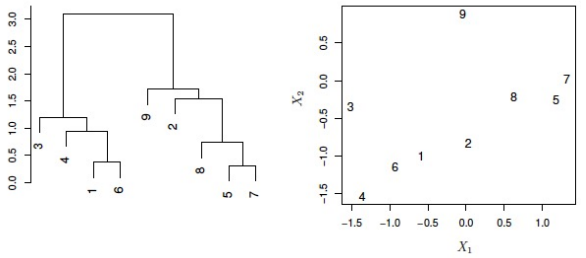
Gareth James • Daniela Witten • Trevor Hastie • Robert Tibshirani
An Introduction to Statistical Learning with Applications in R, Springer

Complexity: $m^2 \log m$

Hierarchical clustering

General algorithm:

- A) Initially, each point is a cluster
- B) cluster the pairs that are closest (eg 1-6 and 5-7 in the figure)
- C) Repeat B) until there is only one cluster
- D) build the dendrogram, and choose a cut



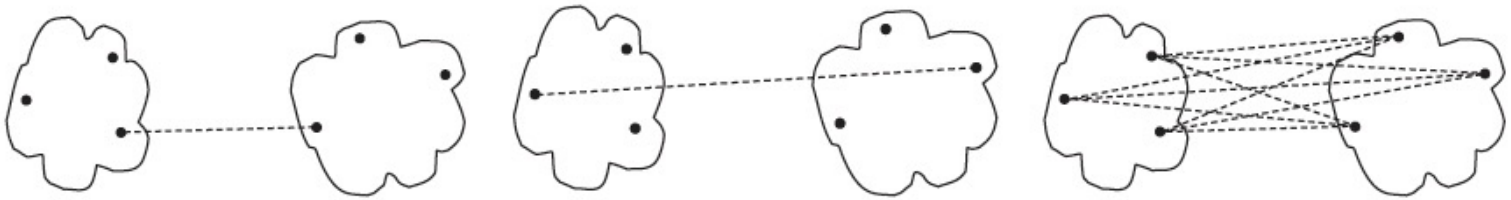
Dendrogram data

Gareth James • Daniela Witten • Trevor Hastie • Robert Tibshirani
An Introduction to Statistical Learning with Applications in R, Srpinger

Two issues:

- Linkage: how to determine “closest clusters”
- Cut: where to cut the dendrogram

Hierarchical clustering: Linkage

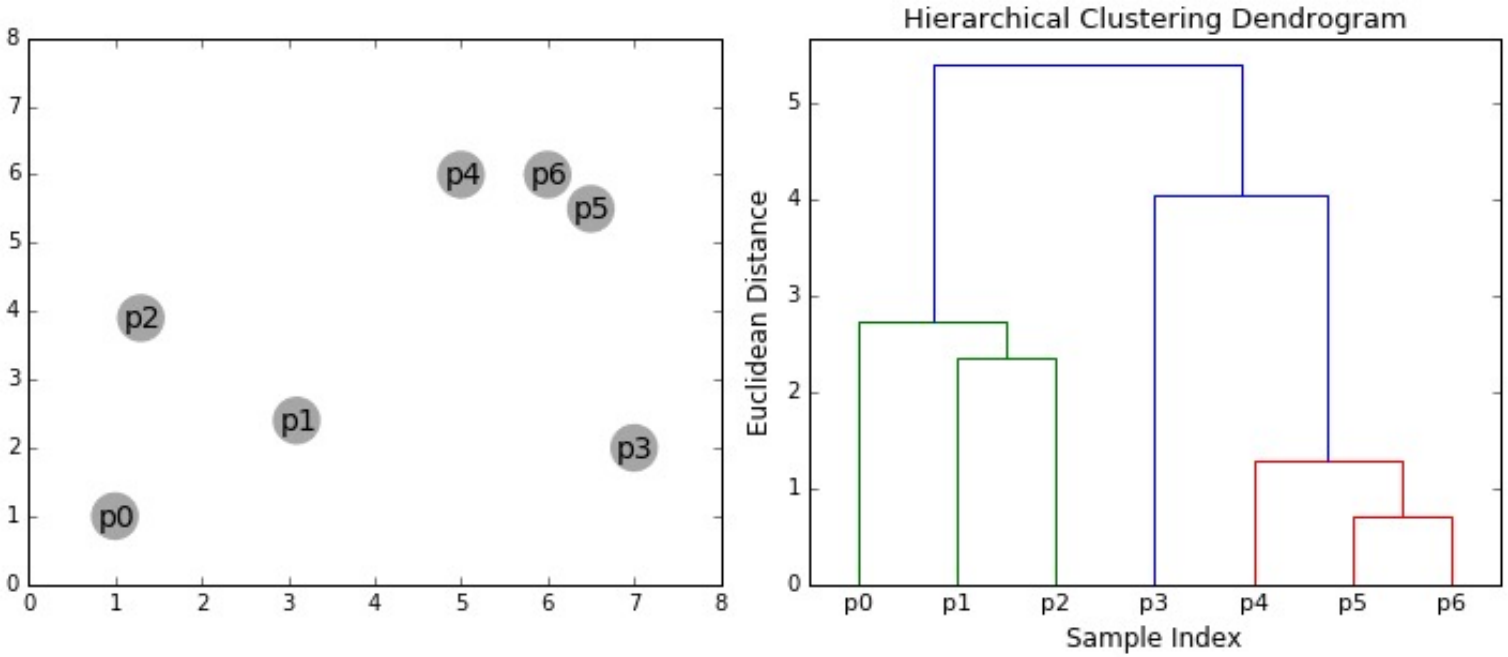


(a) MIN (single link.) (b) MAX (complete link.) (c) Group average.

Gareth James • Daniela Witten • Trevor Hastie • Robert Tibshirani
An Introduction to Statistical Learning with Applications in R, Springer

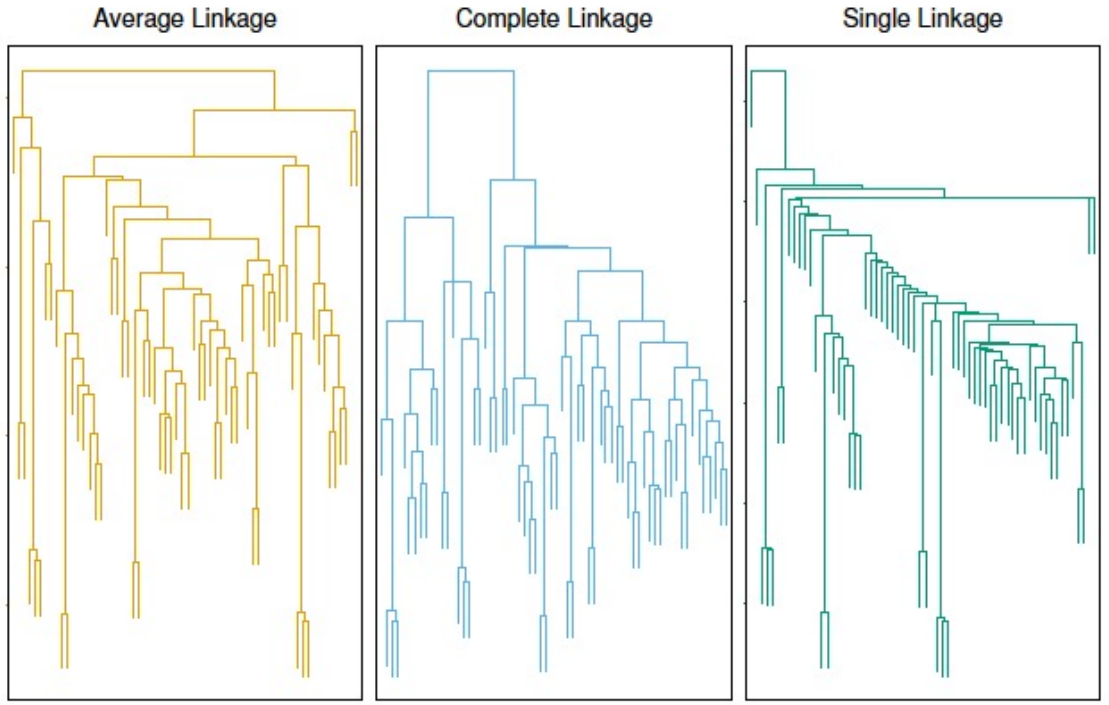
Instead of centroids, three alternative cluster distances are often used

Hierarchical clustering



David Sheehan
dashee87.github.io/data%20science/general/Clustering-with-Scikit-with-GIFs/

Dependence on Linkage

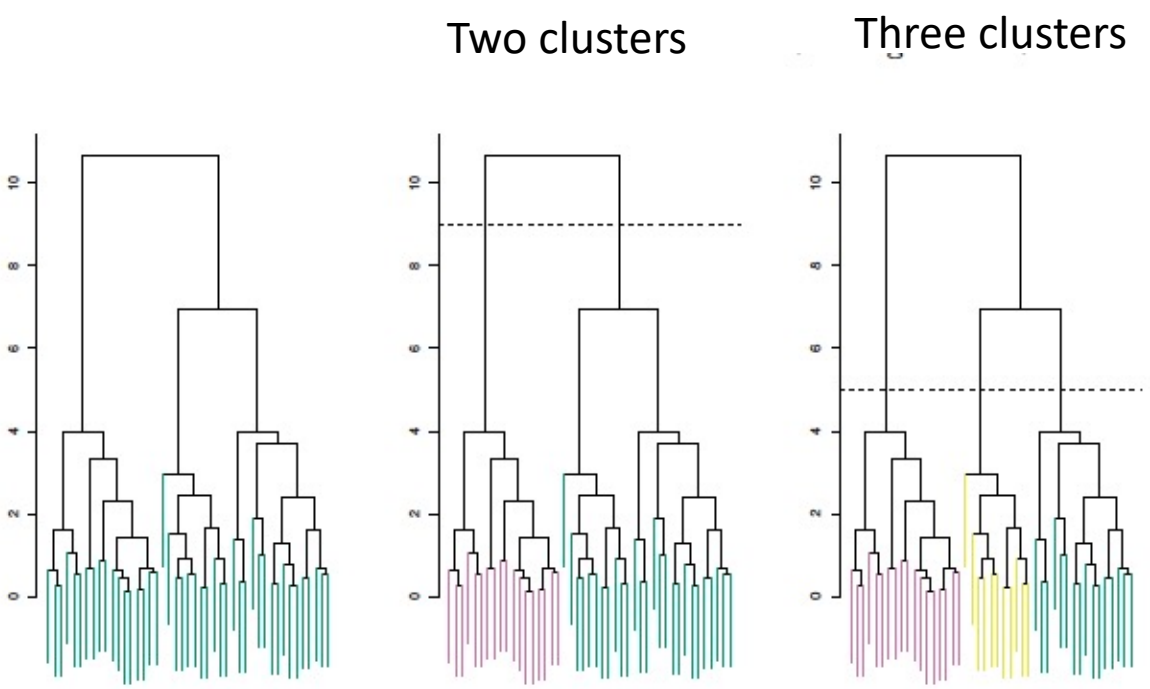


Gareth James • Daniela Witten • Trevor Hastie • Robert Tibshirani
An Introduction to Statistical Learning with Applications in R, Springer

(Less balanced)

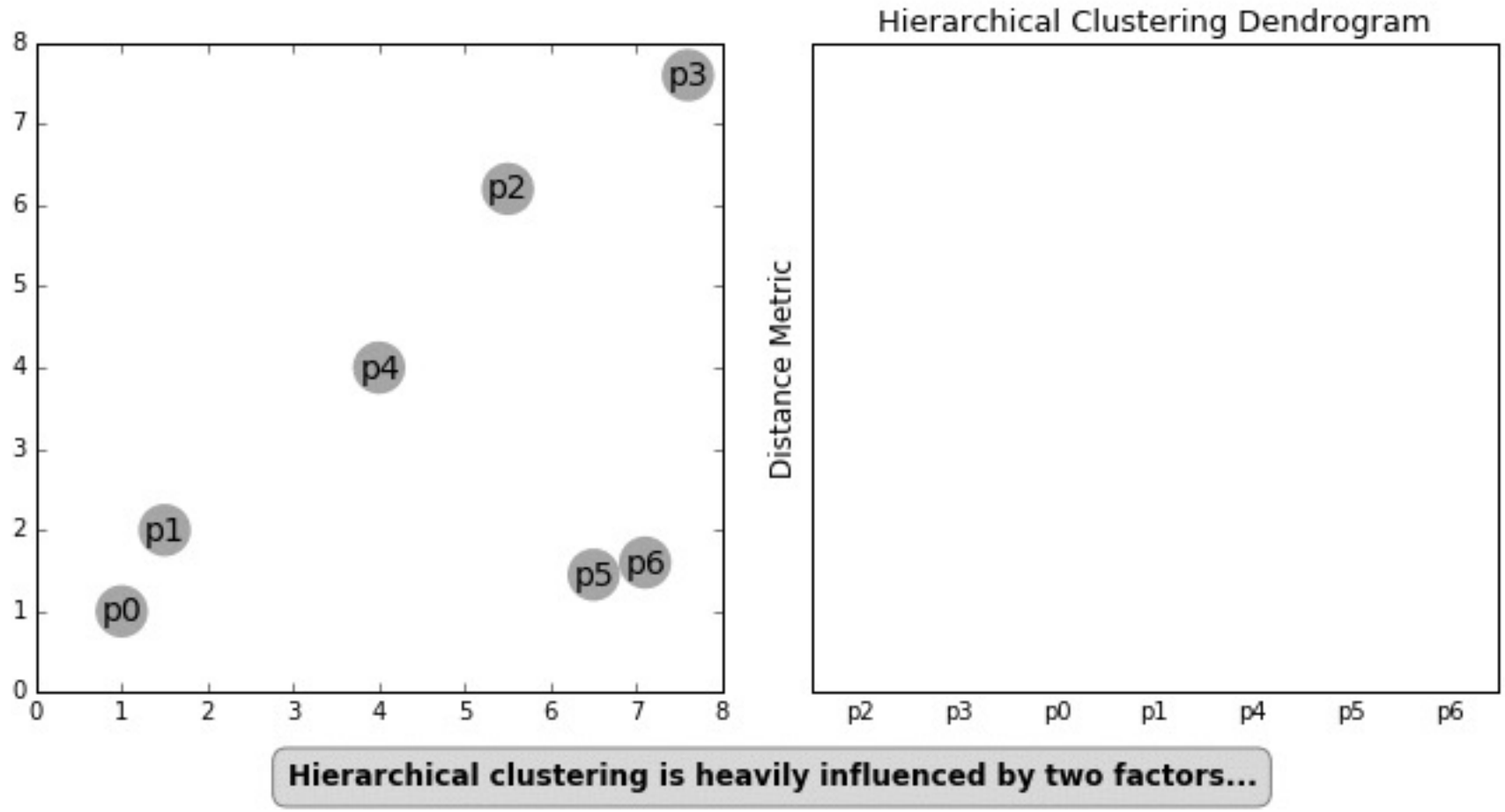
Dependence on Cut

More and more different data collected in the same clusters



Gareth James • Daniela Witten • Trevor Hastie • Robert Tibshirani
An Introduction to Statistical Learning with Applications in R, Springer

Hierarchical clustering: different distances and linkages



David Sheehan
dashee87.github.io/data%20science/general/Clustering-with-Scikit-with-GIFs/

Pro’s and contra’s of HC

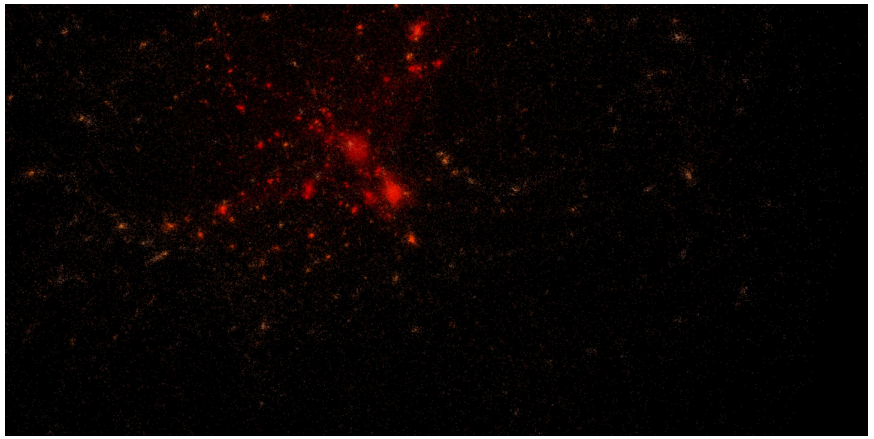
Pros and contras

- No need to pre-set number of clusters
- Simple, fast
- Many choices (linkage, cut, distance)
- Sensitive to outliers
- Complexity: $m^2 \log m$

Friend-of-Friends algorithm

If you know that points cluster due to some physical mechanism,
and that the clusters should have known properties as e.g. size or density,
then you can define a linking length, i.e. a distance below which points
should be in the same cluster

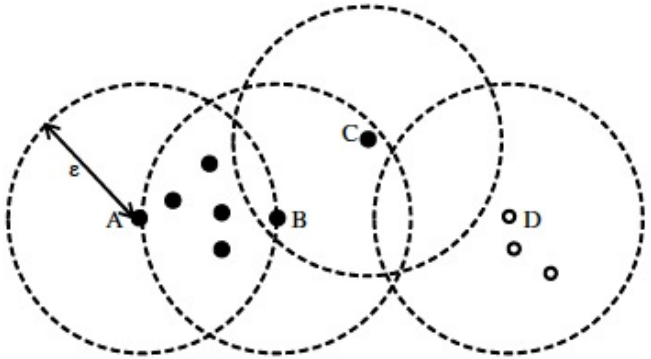
Method of choice to identify astrophysical objects (galaxies, clusters) in N-body simulations



Michael Umbricht Wikimedia

Friend-of-Friends algorithm

A, B,C are in one cluster
D is in another

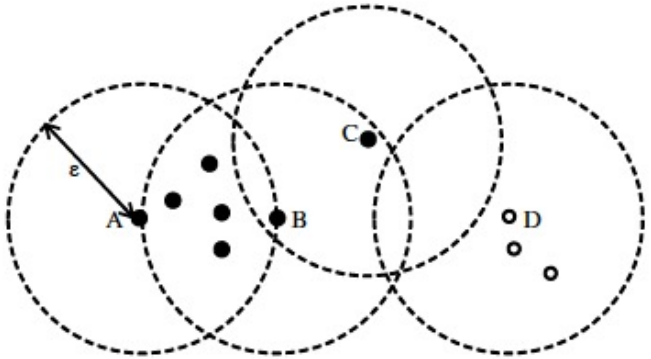


Kwon, YongChul & Nunley, Dylan & Gardner, Jeffrey & Balazinska, Magdalena & Howe, Bill & Loebman, Sarah. (2010).
Scalable Clustering Algorithm
for N-Body Simulations in a Shared-Nothing Cluster.

Two points are friends if their distance is smaller
than a pre-assigned *linking length*
Friends-of-friends belong to the same cluster

Friend-of-Friends algorithm

A, B,C are in one cluster
D is in another



Kwon, YongChul & Nunley, Dylan & Gardner, Jeffrey & Balazinska, Magdalena & Howe, Bill & Loebman, Sarah. (2010).
Scalable Clustering Algorithm
for N-Body Simulations in a Shared-Nothing Cluster.

In N=body simulations, the linking length is almost always chosen as 0.2 times the mean interparticle distance, because it produces virialized halos

References

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, *An Introduction to Statistical Learning with Applications in R*, Springer

Tan, Steinbach, Kumar, *Introduction to Data Mining*

Everitt, B.S., Landau, S. and Leese, M. (2001), *Cluster Analysis*, Fourth edition, Arnold.