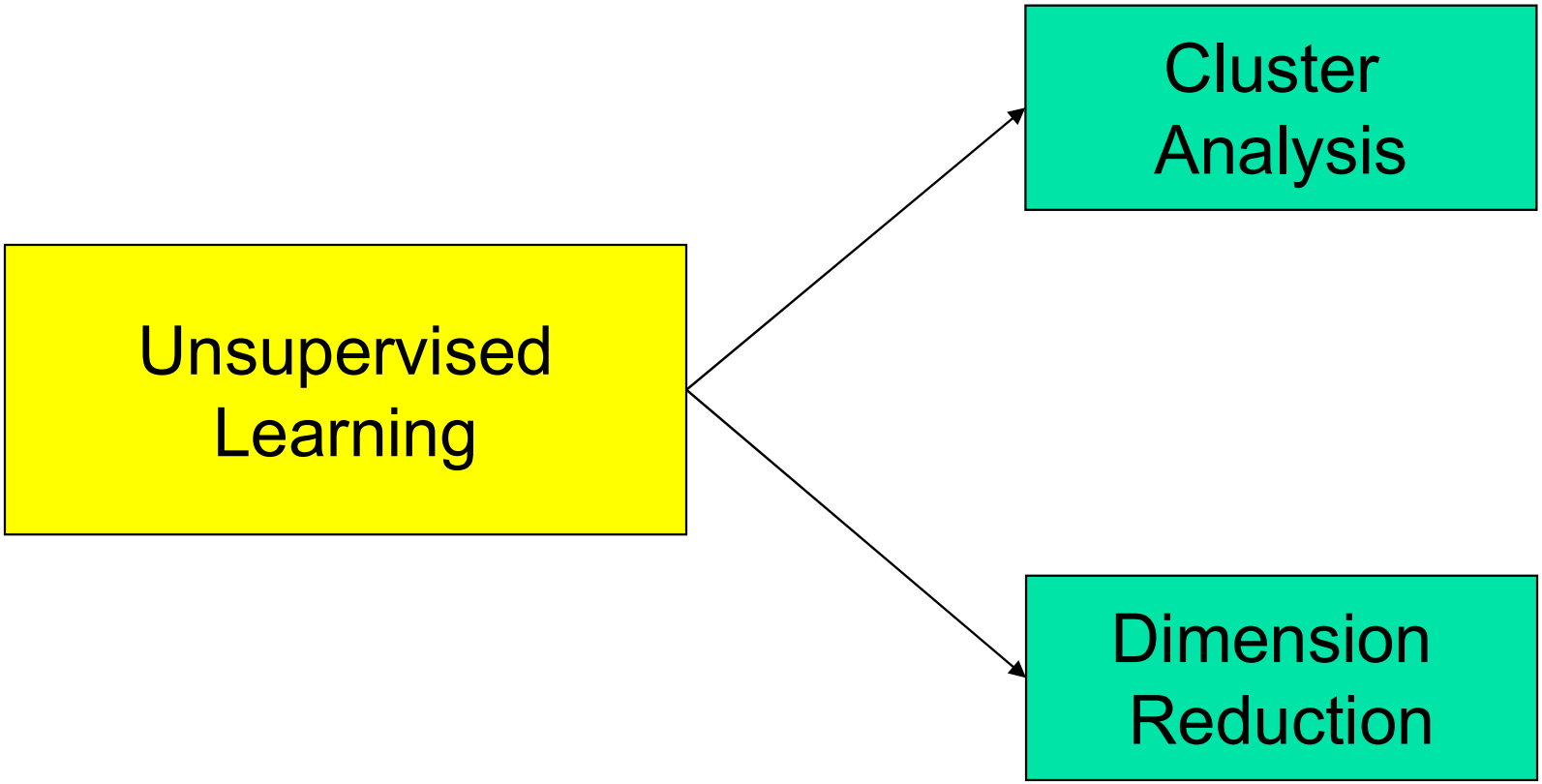
Session 10
Clustering Methods

PSYC 560

Heungsun Hwang

Unsupervised Learning

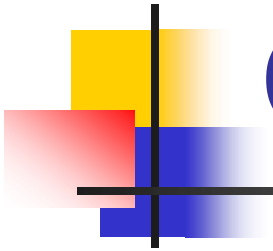


Unsupervised Learning Problems - Clustering

sample

Cluster/group

https://www.guru99.com/unsupervised-machine-learning.html#2

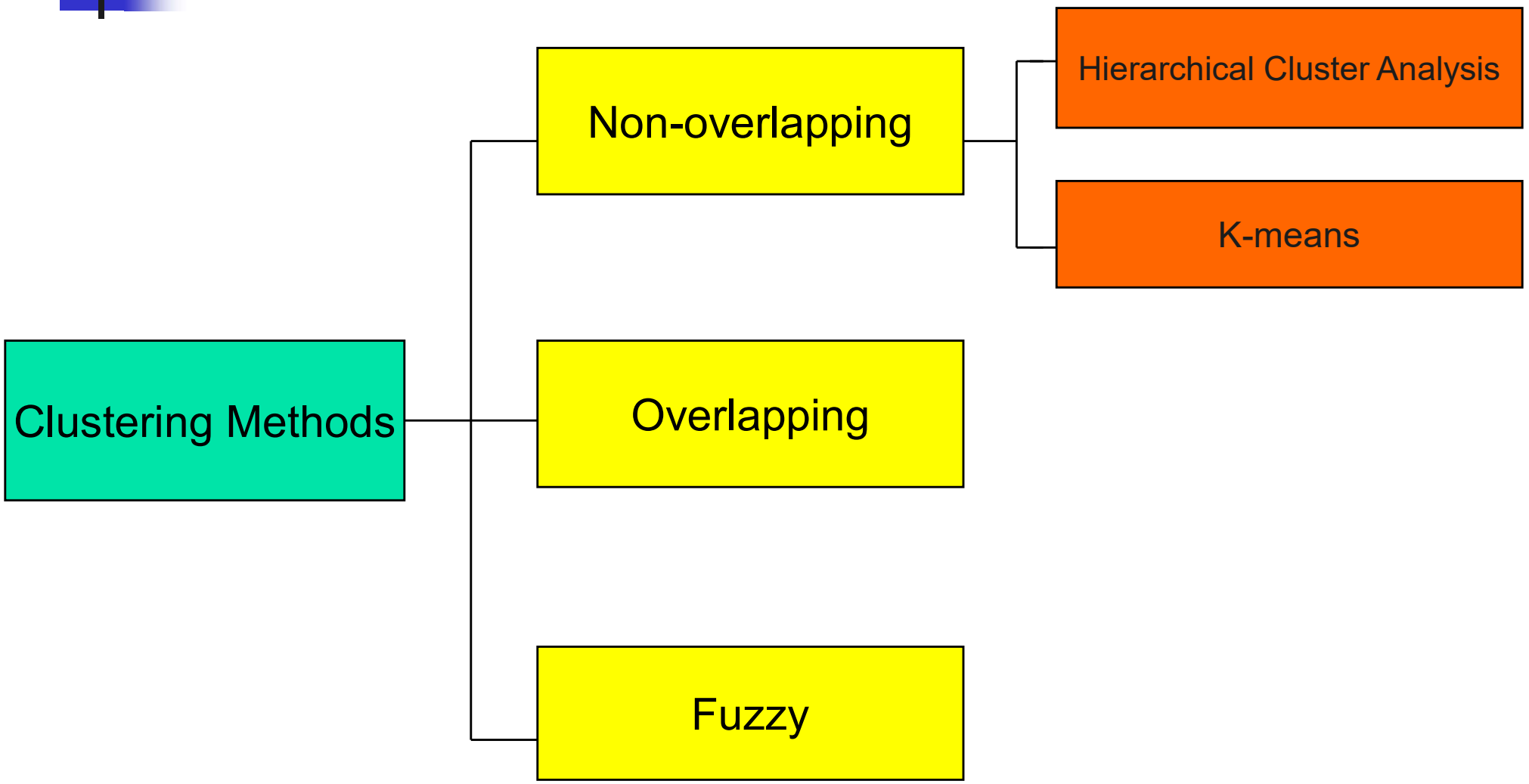


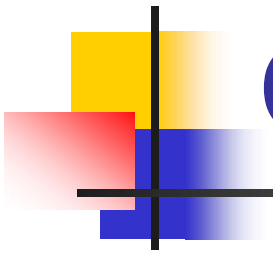
Cluster Analysis

- Cluster analysis involves dividing a set of observations (cases/variables) into smaller groups so that the observations within the same group are relatively homogeneous.

- Cluster analysis is a type of unsupervised learning in which we have only a set of variables measured on N observations.
  - We are not interested in prediction. Rather the goal is to discover unknown subgroups in data.

Classification of Clustering Methods



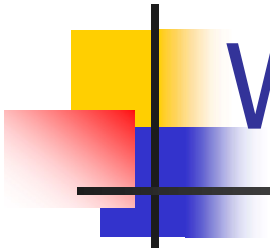


Cluster Analysis

- Generally, cluster analysis is based on two ingredients:
  - Input data
    - Original observations
    - Distance measure: Quantification of proximity or similarity of observations
  - Cluster algorithm: A procedure to group observations
    - Aim: high intra-cluster similarity while low inter-cluster similarity

Some Distance Measures

- Euclidean distance: $$d(x,y)=\sqrt{\sum_{i=1}^{N}(x_i-y_i)^2}$$
- City-block distance: $$d(x,y)=\sum_{i=1}^{N}\left|x_i-y_i\right|$$
- Correlation distance: $$d(x,y)=1-\frac{\sum_{i=1}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}(x_i-\bar{x})^2\sum_{i=1}(y_i-\bar{y})^2}}$$



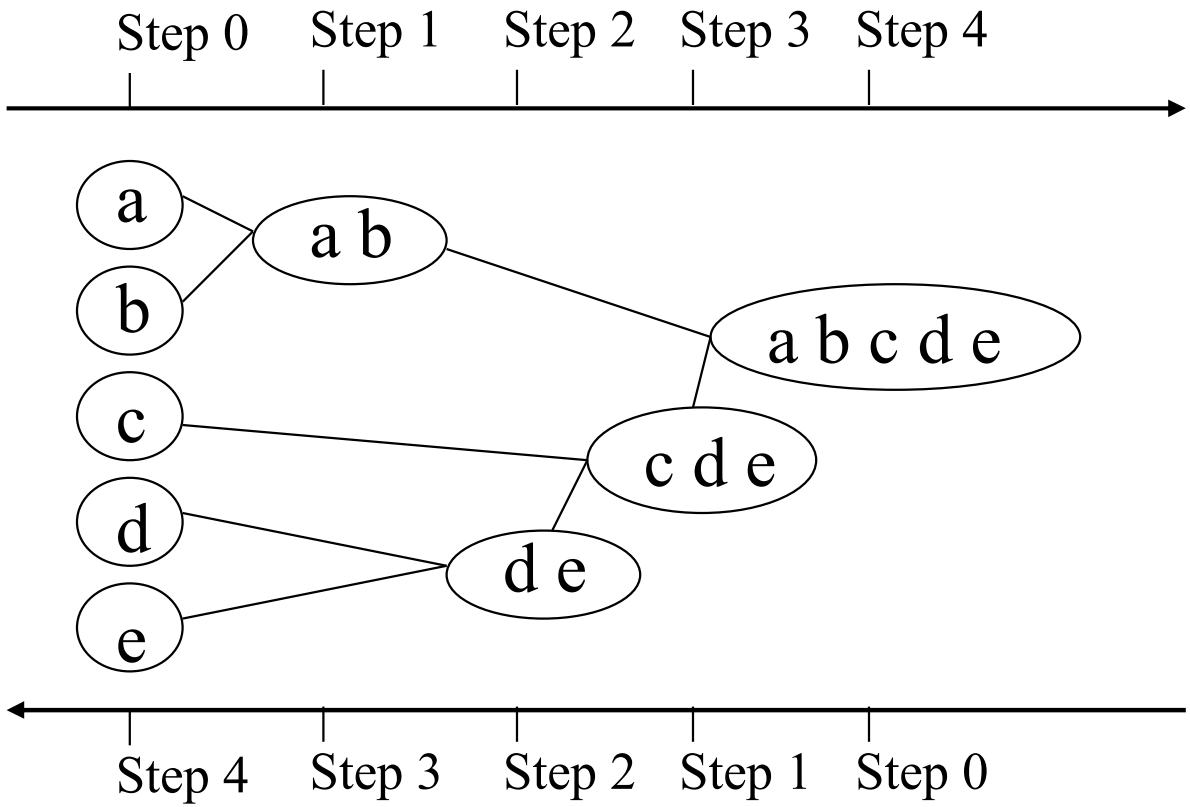
Which distance measure to use?

- The choice of distance measure should be based on the application area.
  - Euclidean and city-block distances both measure absolute differences between observations. The city-block distance is more robust against outliers.
  - If standardized observations are used, Euclidean and correlation distances become equivalent.

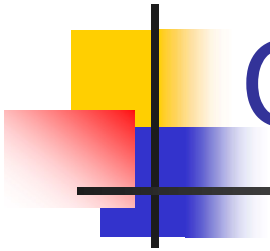
Hierarchical Cluster Analysis (HCA)

- Represents clustering which results in a hierarchical tree structure of observations, called a dendrogram.
- Input data = Distance measure of observations
  - Continuous, counts, or binary data
- Cluster algorithms
  - Agglomerative ('build-up')
    - Single linkage (Nearest neighbour)
    - Complete linkage (Furthest neighbour)
    - Average linkage (Between-groups linkage)
    - Centroid/Median linkage
    - Ward's minimum variance linkage (Ward's method)
  - Divisive ('split-down')

Hierarchical Cluster Analysis



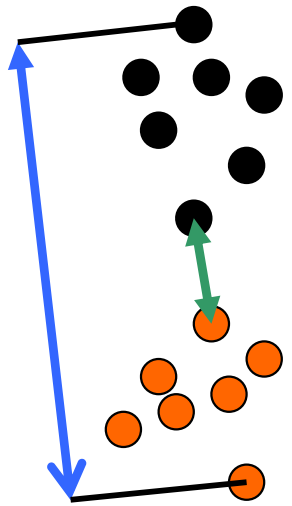
Agglomerative Hierarchical Clustering



- Bottom-up algorithm
  - Top-down (divisive) algorithm is less common.
  - Start with individual observations as clusters.
  - In each iteration, merge two clusters with a minimal distance from each other - until ending up with a single cluster comprising all observations.
  - But what is the distance between two clusters?

Distances between clusters used for hierarchical clustering

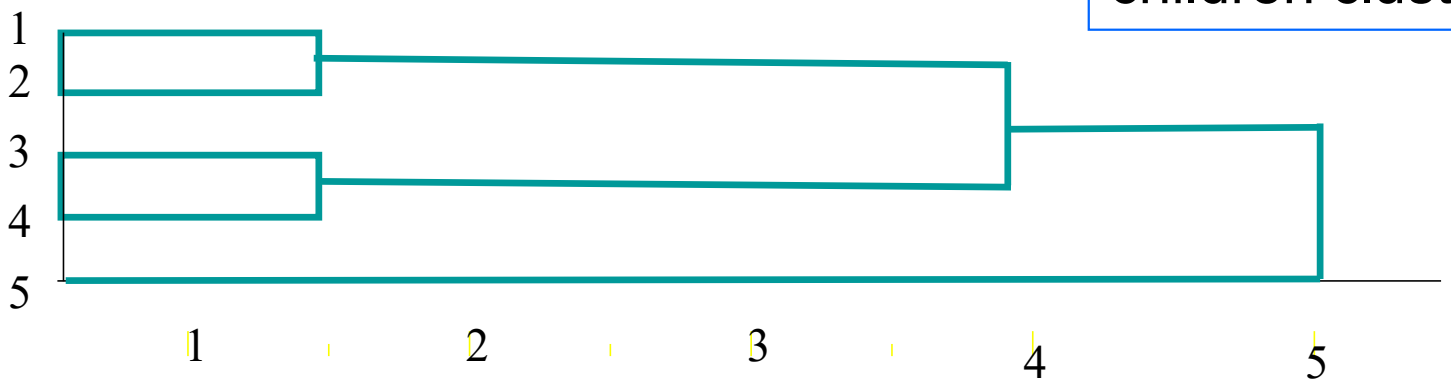
- Complete linkage: largest distance
- Average linkage: average of the distances between all pairs of data points
- Single linkage: smallest distance
- Centroid linkage: distance between cluster means
- Ward's method: seek to join two clusters whose merger leads to the minimum within-group variance



Single Linkage Cluster Example

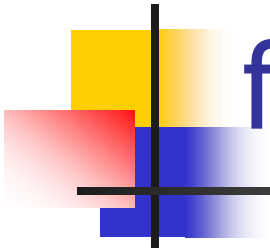
Distance Matrix

	obs 1	obs 2	obs 3	obs 4	obs 5
obs 1	0				
obs 2	1.49	0			
obs 3	3.42	2.29	0		
obs 4	1.81	1.99	1.48	0	
obs 5	5.05	4.82	4.94	4.83	0

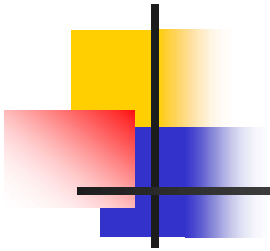


The height of a node in the dendrogram represents the distance of the two children clusters

Distances between clusters used for hierarchical clustering



- Which algorithm?
  - No clear answer…
  - Related to which distance measure is chosen
    - (Squared) Euclidean distances should be used with the Ward and centroid methods.



Hierarchical Cluster Analysis

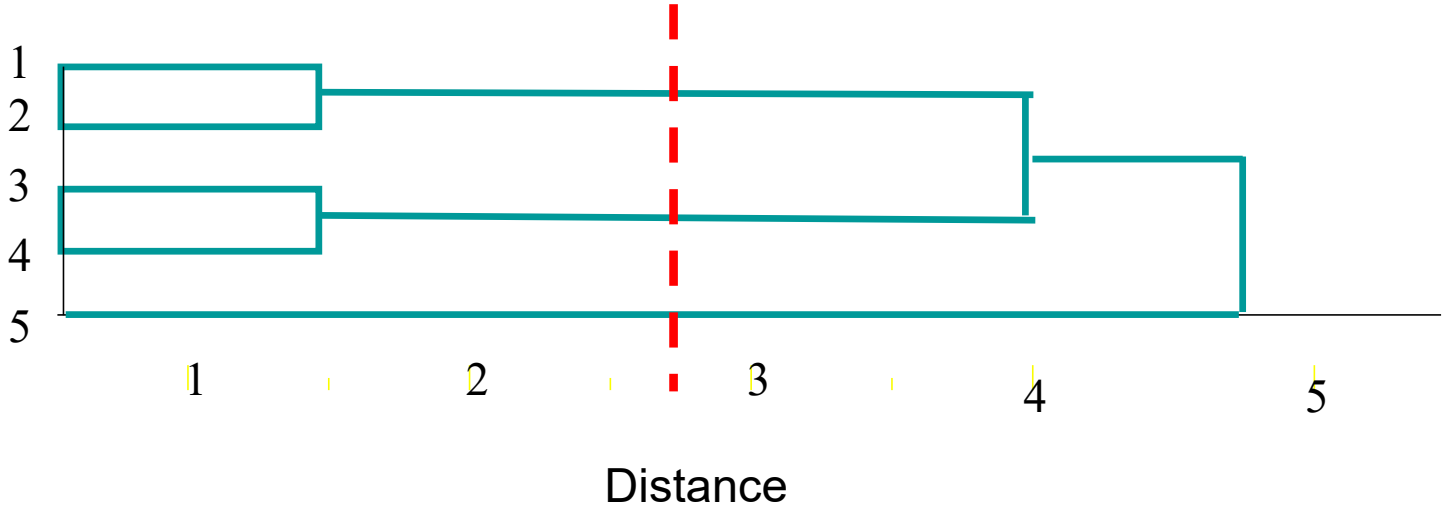
- How many clusters?
  - No definite answer…
  - Looking at the dendrogram, look for a relatively wide range of distances over which the number of clusters in the solution does not change.
    - Often, very subjective.
  - Non-statistical criteria

Single Linkage Cluster Example

Distance Matrix

	obs 1	obs 2	obs 3	obs 4	obs 5
obs 1	0				
obs 2	1.49	0			
obs 3	3.42	2.29	0		
obs 4	1.81	1.99	1.48	0	
obs 5	5.05	4.82	4.94	4.83	0

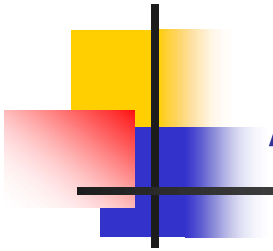
The height of a node in the dendrogram represents the distance of the two children clusters



Hierarchical Cluster Analysis

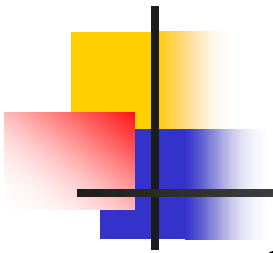
- Advantages:
  - Can be used for continuous, counts, or binary data.
  - No need to specify the number of clusters in advance.
  - Nested structure of clusters can be represented.

- Disadvantages:
  - Not efficient for large data (N < 200)



K-means

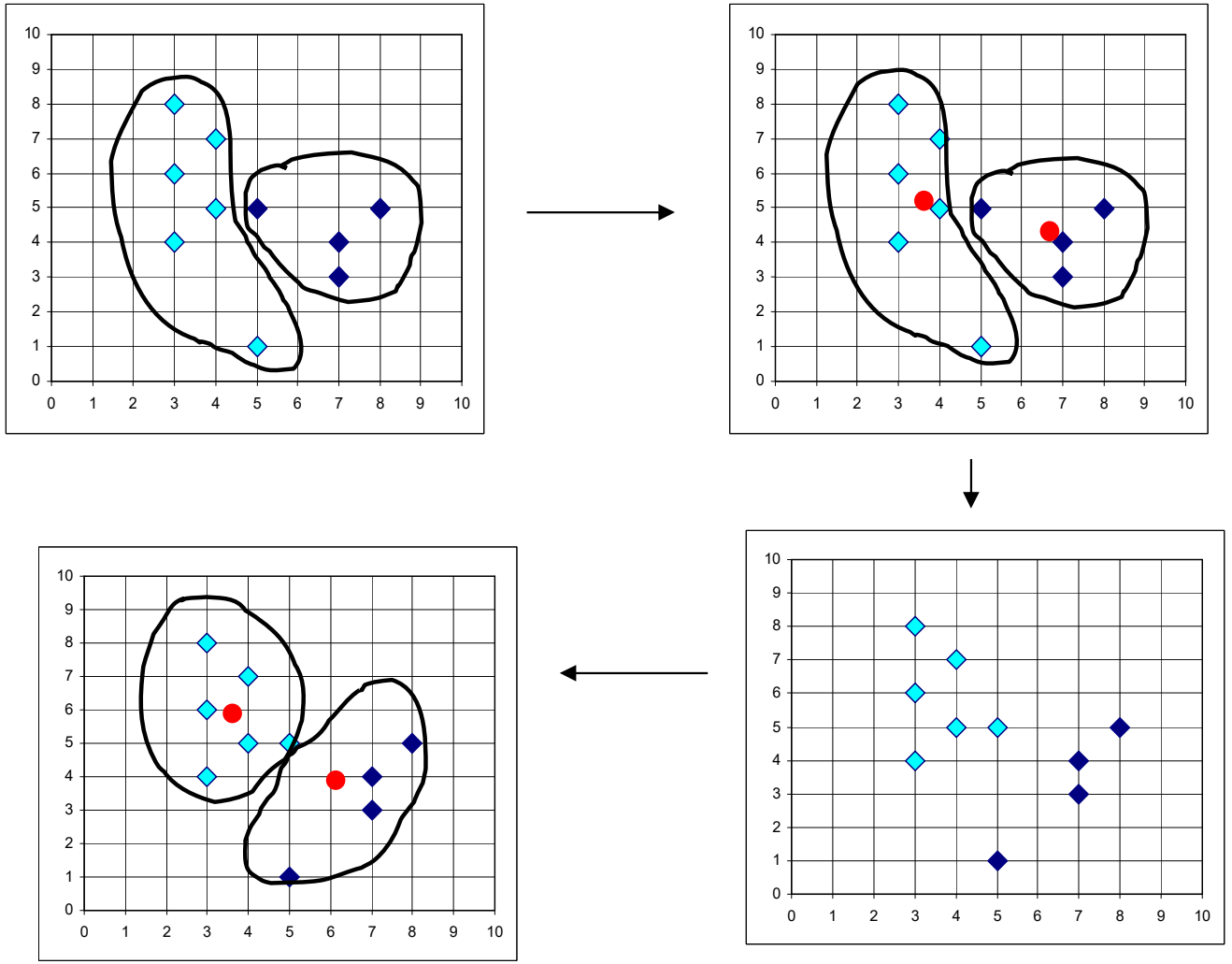
- Partition observations into a set of *K* clusters.
  - Each cluster is represented by the mean (centroid) of the cluster.
- Input data: *N* original observations
- The *K*-means algorithm (MacQueen, 1967)
  - Minimizes the sum of squared within-cluster distances



The *K*-means Algorithm

- Given *K*, the algorithm repeats the following steps:
  - Step 0: Initially, partition observations into *K* clusters.
  - Step 1: Compute the centroids of the clusters of the current partition.
  - Step 2: Assign each observation to the cluster with the nearest centroid.
  - Go back to Step 1, stop when no more changes in the centroids.

The *K*-means Algorithm

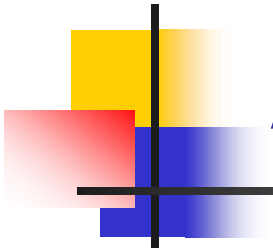


K-means

- How many clusters?
  - No definite answer.
  - Pseudo-F statistic (Calinski & Harabasz, 1974)

$$\text{pseudo - F} = \frac{SS(B)/K-1}{SS(W)/N-K}$$

    - A large value of the pseudo-*F* usually indicates a better clustering solution.
    - Apply *K*-means to the same data with varying numbers of clusters. Choose the number of clusters, which involves the largest pseudo-F statistic value.
  - Non-statistical criteria



K-means

- Advantages:
  - Relatively efficient for large data
- Disadvantages :
  - Applicable only when mean is defined (continuous data)
  - Need to specify *K*, the number of clusters, in advance
  - Difficult to handle outliers

Cluster Profiling

- Describe resultant clusters.

- Lilien and Rangaswamy (2003):
  - Bases — characteristics that tell us why clusters differ (e.g., needs, benefits, preferences).
  - Descriptors — characteristics that help us find and reach clusters.

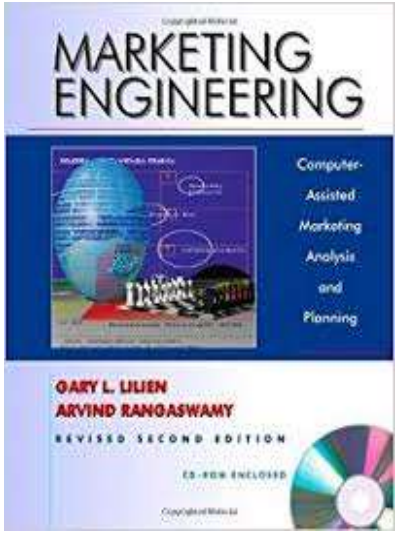
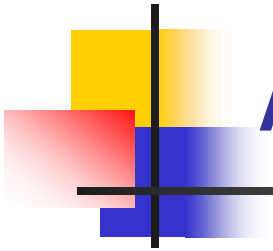


Table 7-2 Benefit Segmentation of the Toothpaste Market

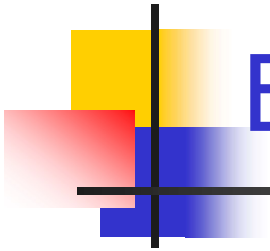
Benefit Segments	Demographics	Behaviour	Psychographics	Favoured Brands
Economy (low price)	Men	Heavy users	High autonomy, value oriented	Brands on sale
Medicinal (decay prevention)	Large families	Heavy users	Hypochondriac, conservative	Crest
Cosmetic (bright teeth)	Teens, young adults	Smokers	High sociability, active	Aqua-Fresh, Ultra Brite
Taste (good tasting)	Children	Spearmint lovers	High self-involvement, hedonistic	Colgate, Aim

Source: Adapted from Russell J. Haley, "Benefit Segmentation: A Decision-Oriented Research Tool," *Journal of Marketing,* July 1968, pp. 30–35. Also see Haley, "Benefit Segmentation: Backwards and Forwards," *Journal of Advertising Research*, February–March, 1, 1984, pp. 19–25; and Haley, "Benefit Segmentation—20 Years Later," *Journal of Consumer Marketing,* 1984: 5–14.

Combined Use of Cluster Analysis with Other Techniques



- Dimension reduction → cluster analysis
- Cluster analysis → other techniques
  - Clusterwise regression



Example: Antisocial Behaviour Data

- Part of the National Longitudinal Survey of Youth (NLSY) reported in Curran (1998).
  - Antisocial behaviour of 221 children measured in 1986
  - Three predictors:
    - Gender (female = 0 and male = 1)
    - Cognitive stimulation for children at home
    - Emotional support for children at home

Example: Antisocial Behaviour Data

- Step 1: *K*-means was applied to antisocial behaviour.
  - K = 3

- Step 2: A linear regression analysis was conducted for each cluster.
  - DV = antisocial behaviour
  - Predictors = Gender, cognitive stimulation, emotional support

Example: Antisocial Behaviour Data

- *K*-means (*K* = 3)

Report

anti1

Cluster Number of Case	Mean	N	Std. Deviation
1	.45	130	.500
2	2.59	80	.706
3	5.82	11	.874
Total	1.49	221	1.539

Example: Antisocial Behaviour Data

- Cluster 1:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.731	.211		3.469	.001
	gender	.076	.087	.076	.872	.385
	cogstm	-.045	.020	-.226	-2.284	.024
	emotsup	.010	.023	.046	.461	.645

a. Dependent Variable: anti1

Example: Antisocial Behaviour Data

- Cluster 2:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2.577	.417		6.183	.000
	gender	.407	.160	.283	2.546	.013
	cogstm	.012	.033	.040	.367	.715
	emotsup	-.038	.034	-.123	-1.101	.274

a. Dependent Variable: anti1

Example: Antisocial Behaviour Data

- Cluster 3:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4.300	.731		5.885	.001
	gender	-.425	.305	-.246	-1.396	.205
	cogstm	.380	.073	.913	5.232	.001
	emotsup	-.227	.065	-.547	-3.472	.010

a. Dependent Variable: anti1

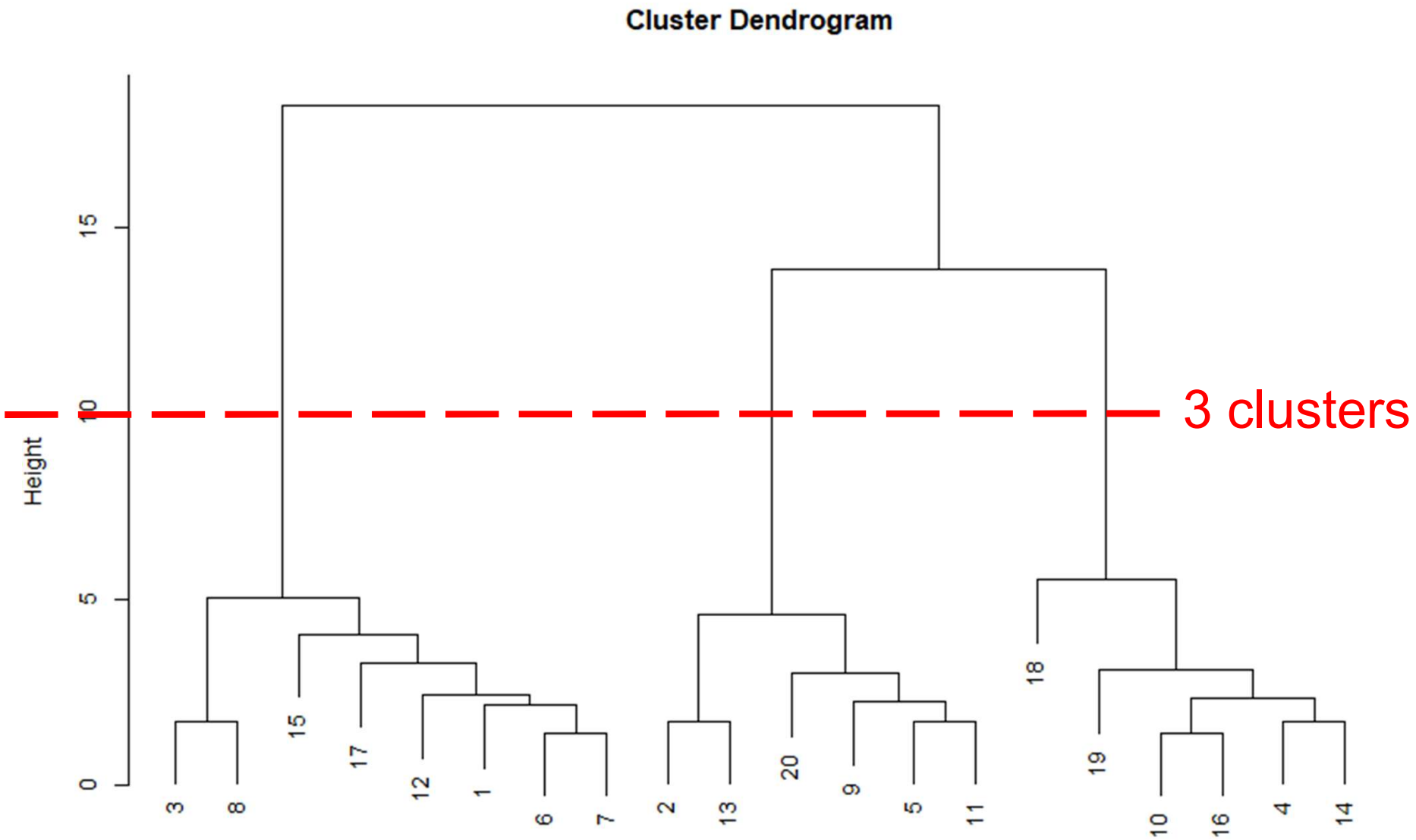
Example: HCA & K-means

- The shopping attitude data (Malhotra, 2004; shoppingattitude.csv). Six variables were measured on a 7-point scale (1= disagree, 7 = agree)
  - V1: Shopping is fun.
  - V2: Shopping is bad for your budget.
  - V3: I combine shopping with eating out.
  - V4: I try to get the best buys when shopping.
  - V5: I don't care about shopping.
  - V6: You can save a lot of money by comparing prices.
- N = 20

Example: HCA & K-means

- Hierarchical cluster analysis
  - Ward's method with squared Euclidean distances
- *K*-means (K = 3)

Example: HCA


Cluster Dendrogram
Height
0
5
10
15
3
8
15
17
12
1
6
7
2
13
20
9
5
11
18
19
10
16
4
14
3 clusters

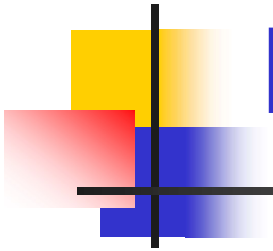
Example: K-means

```
K-means clustering with 3 clusters of sizes 6, 8, 6

Cluster means:
      v1        v2        v3     v4     v5     v6
1 1.666667 3.000000 1.833333 3.500 5.500 3.333333
2 5.750000 3.625000 6.000000 3.125 1.875 3.875000
3 3.500000 5.833333 3.333333 6.000 3.500 6.000000

Clustering vector:
 [1] 2 1 2 3 1 2 2 2 1 3 1 2 1 3 2 3 2 3 3 1

Within cluster sum of squares by cluster:
[1] 20.50000 34.00000 25.16667
 (between_SS / total_SS =  75.8 %)
```

Example: Cluster Profiling

- The shopping attitude data (Malhotra, 2004; shoppingattitude.csv) additionally include two demographic variables.
  - Gender (0: male, 1: female)
  - SES: Socioeconomic status measured on a 5-point Likert scale (1: very low - 5: very high)

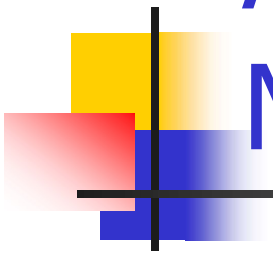
	Group 1	Group 2	Group 3	Entire
gender	0.500000	0.25	0.8333333	0.50
SES	1.833333	3.00	4.3333333	3.05



Finite Mixture Models

- In general, conventional clustering methods have limitations.
  - Choice from a large variety of similarity measures and clustering algorithms is mostly arbitrary.
  - No clear heuristics for the selection of the number of clusters.
  - Statistical properties of clustering solutions are largely unknown.
- Alternative?
  - Finite mixture models

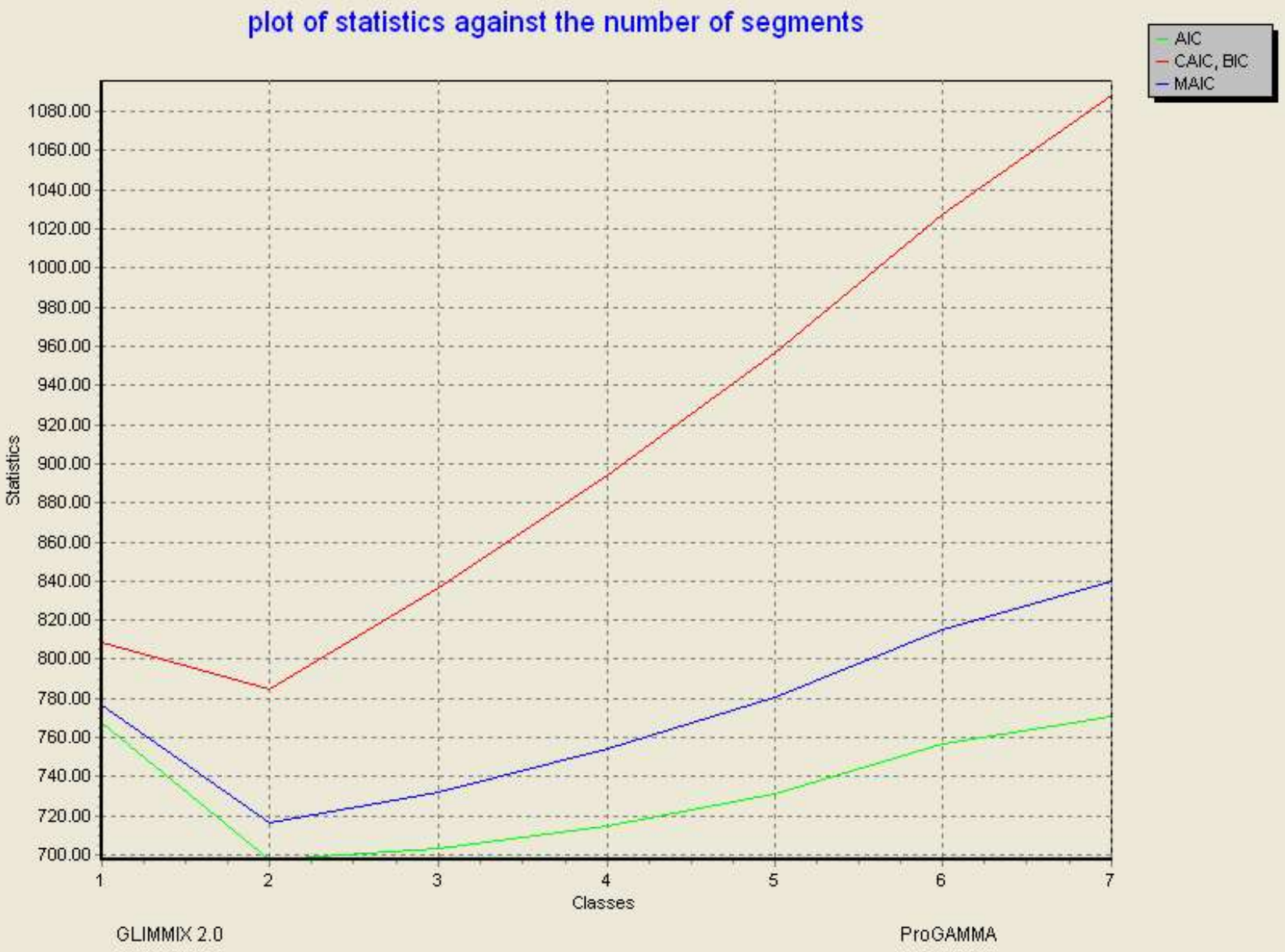
Advantages of Finite Mixture Models



- Statistical heuristics to select the number of clusters (AIC, BIC, CAIC..)
- Probabilistic assignment of observations into clusters
- Parameters can be estimated even if there are missing data (all the available data will be used)

Step 5 : View

Output item: plot of statistics against the number of segments | Output file...



Notes...

Editor

Copy...

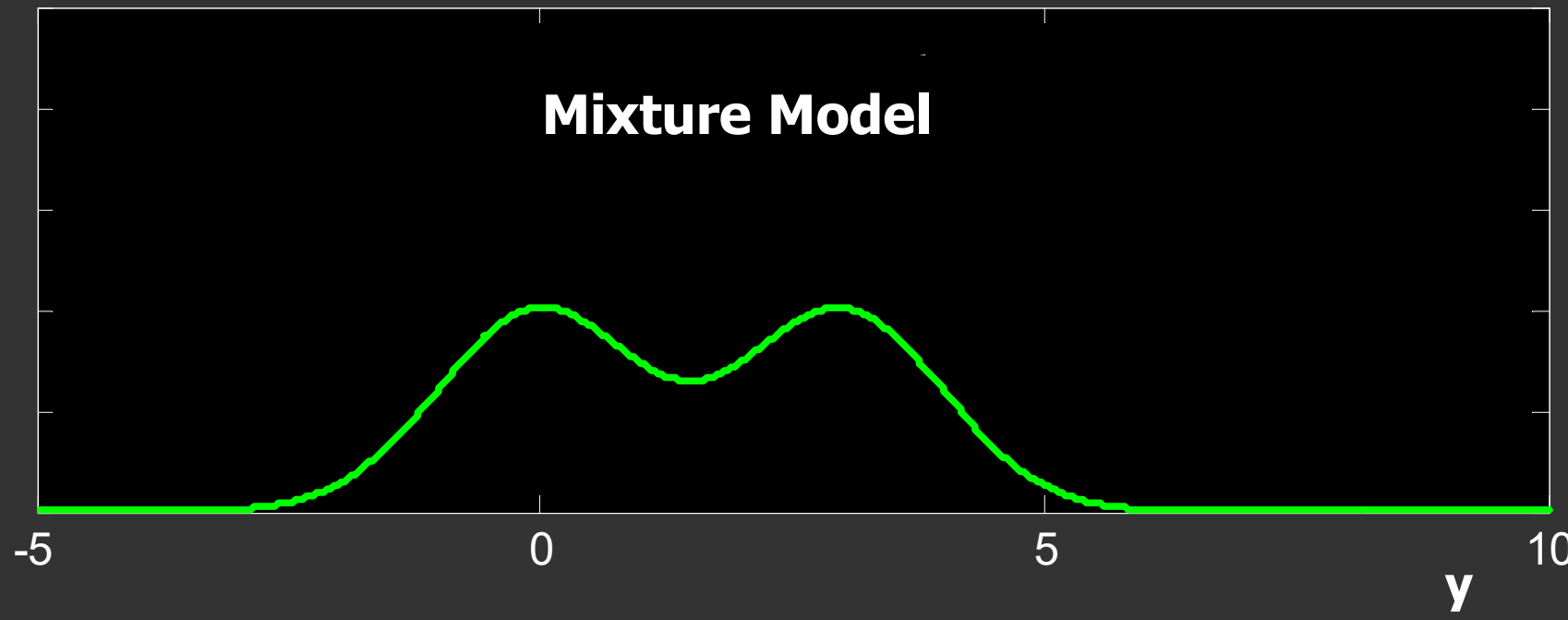
Print...

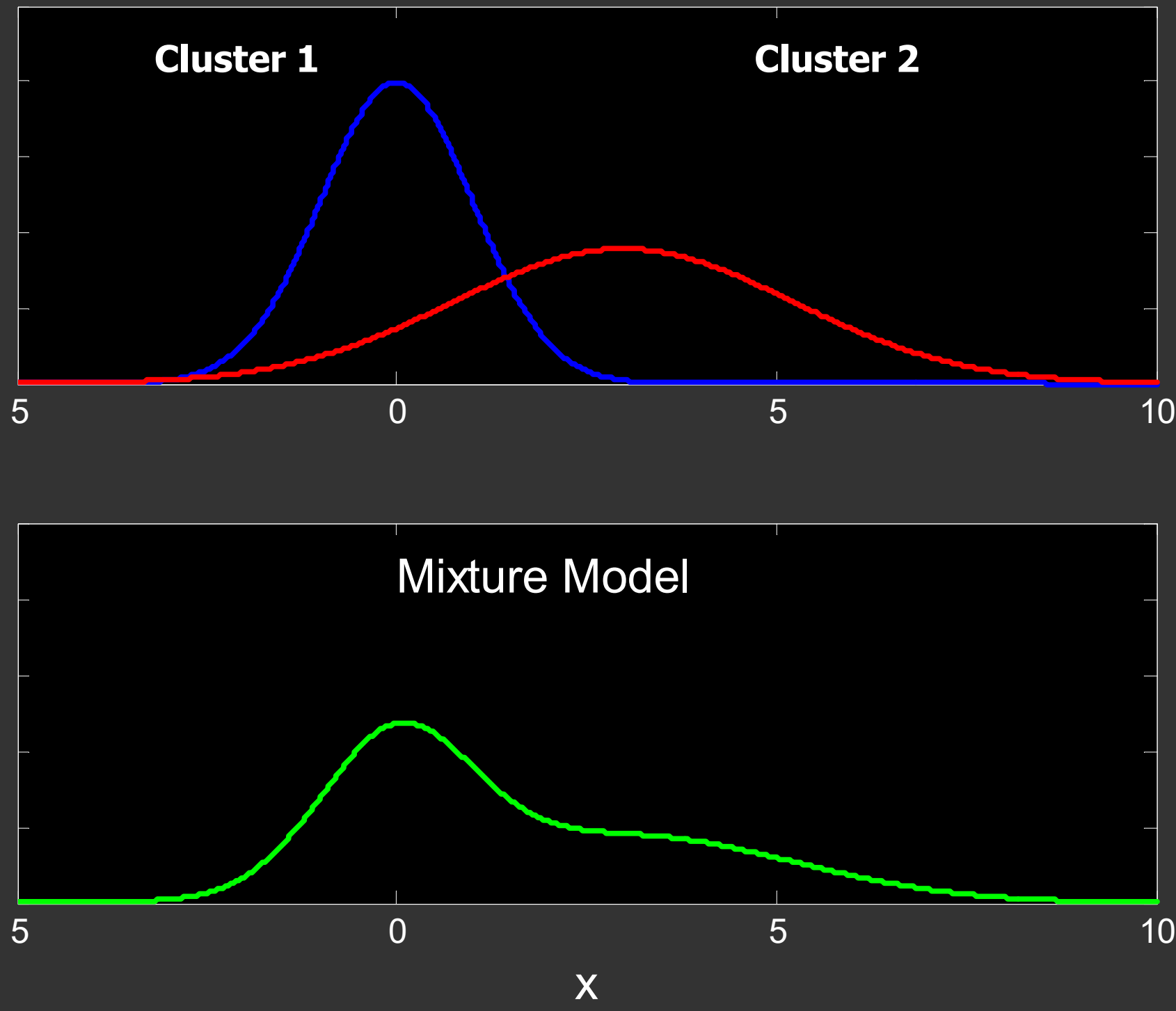
Save...

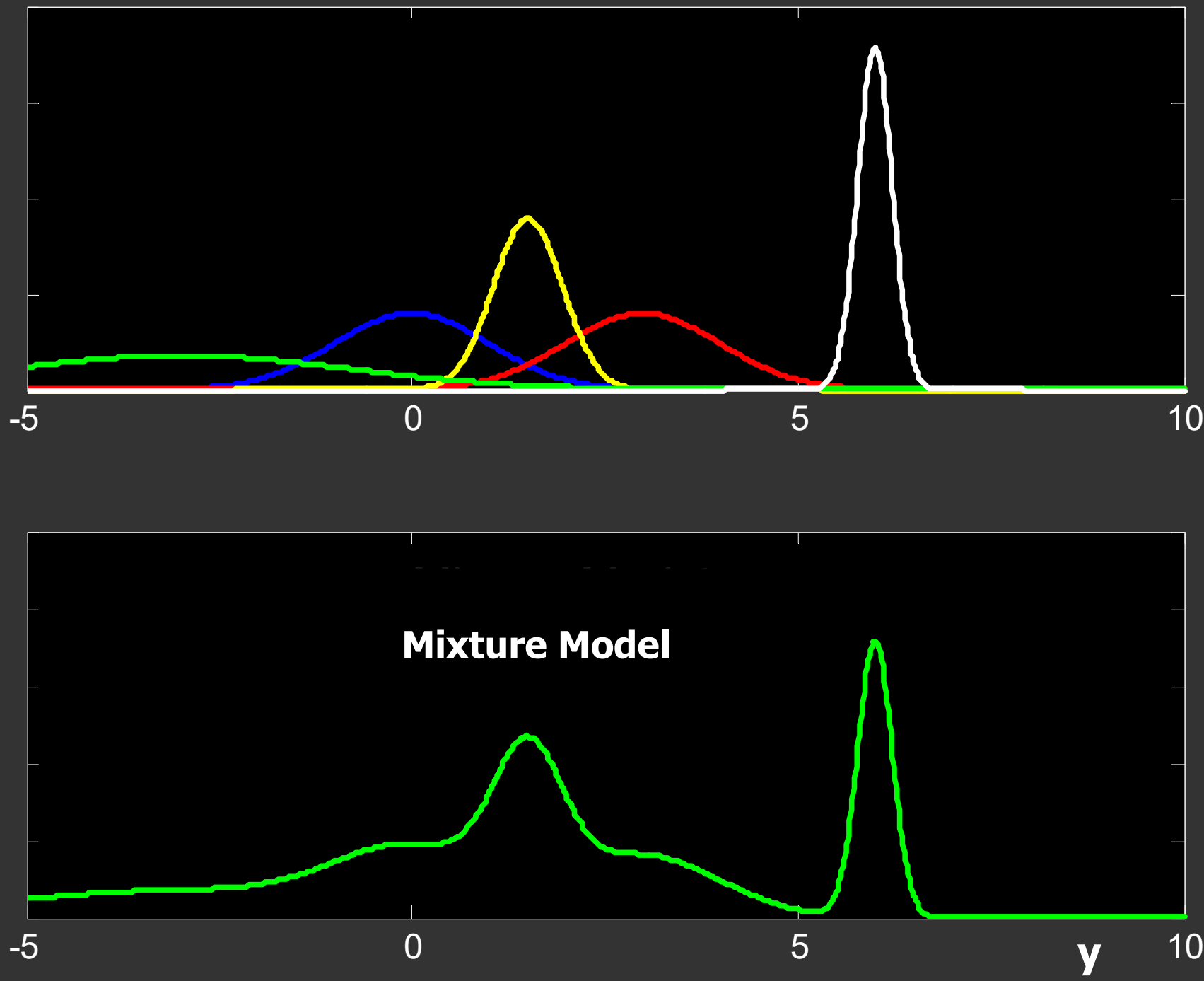
Done

Help

start
EN
Google
1:42 PM


Mixture Model
-5
0
5
10
y


Cluster 1
Cluster 2
5
0
5
10
Mixture Model
5
0
5
10
x


-5
0
5
10
Mixture Model
-5
0
5
y
10

Finite Mixture Models

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(\mathbf{x} \mid c_k)\,\alpha(c_k)$$

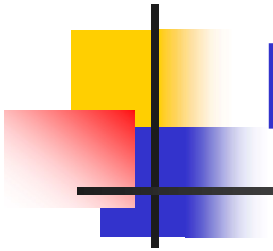
Mixture distribution

Cluster Distribution (e.g.,normal)

Cluster size

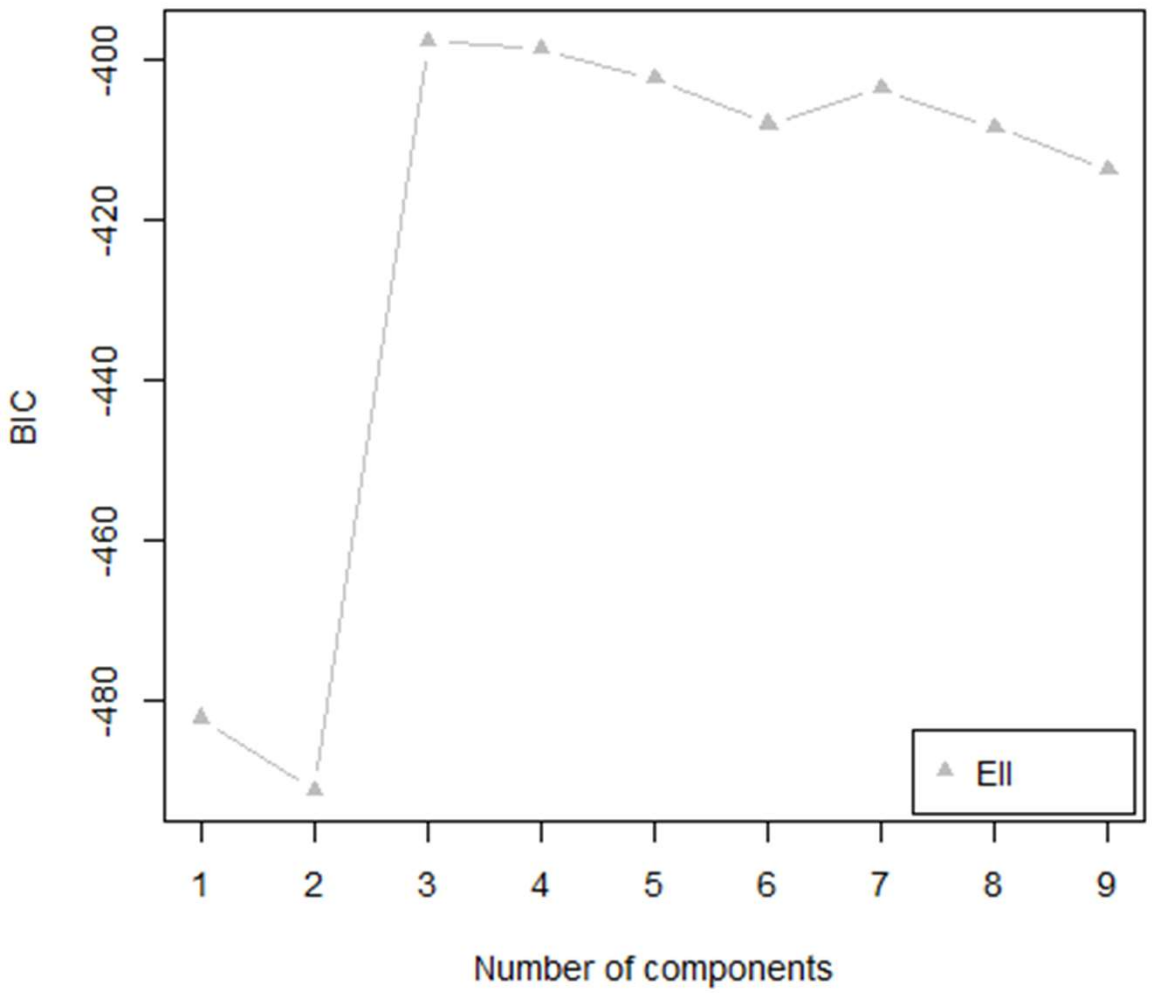
- If normal, it is often called (finite) Gaussian Mixture Models.

Disadvantages of Finite Mixture Models



- Require the assumption of knowing the correct distribution underlying the data.
- Require a large sample size.
- Computationally expensive and often slow.

Example: Finite Mixture Models




```
Best BIC values:
             EII,3       EII,4       EII,5
BIC      -397.8547 -398.889297 -402.594367
BIC diff    0.0000   -1.034601   -4.739672
```

Example: Finite Mixture Models

```
----------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
----------------------------------------------------

Mclust EII (spherical, equal volume) model with 3 components:

 log-likelihood  n df       BIC       ICL
      -167.4722 20 21 -397.8547 -397.8548

Clustering table:
1 2 3
8 6 6

Mixing probabilities:
   1         2         3
0.3999998 0.3000016 0.2999986

Means:
    [,1]     [,2]     [,3]
v1 5.750 1.666674 3.500004
v2 3.625 3.000021 5.833326
v3 6.000 1.833334 3.333341
v4 3.125 3.500013 5.999999
v5 1.875 5.499992 3.499996
v6 3.875 3.333332 6.000015
```

Example: Finite Mixture Models

<Probability that each observation belongs to the kth class>

	[,1]	[,2]	[,3]
[1,]	1.000000e+00	1.600322e-19	2.316242e-14
[2,]	2.204144e-16	1.000000e+00	1.194875e-08
[3,]	1.000000e+00	8.920410e-21	3.679966e-16
[4,]	3.751593e-07	1.901374e-11	9.999996e-01
[5,]	3.288983e-18	1.000000e+00	1.906140e-13
[6,]	1.000000e+00	8.848786e-15	3.650339e-10
[7,]	1.000000e+00	1.063858e-11	3.886805e-10
[8,]	1.000000e+00	2.894194e-24	4.530284e-15
[9,]	5.285443e-13	1.000000e+00	1.603604e-09
[10,]	7.224285e-12	9.682389e-08	9.999999e-01
[11,]	1.944188e-16	1.000000e+00	3.015997e-12
[12,]	9.999961e-01	1.650848e-11	3.947593e-06
[13,]	1.149260e-14	9.999999e-01	1.470960e-07
[14,]	1.739315e-11	1.612675e-13	1.000000e+00
[15,]	1.000000e+00	1.003257e-14	3.083577e-09
[16,]	1.633499e-11	1.820905e-10	1.000000e+00
[17,]	1.000000e+00	3.913198e-15	1.614388e-10
[18,]	2.485759e-12	3.116513e-05	9.999688e-01
[19,]	5.511315e-13	8.644719e-17	1.000000e+00
[20,]	4.145413e-19	1.000000e+00	8.974977e-14