Multivariate Analysis
Clustering

Shaobo Jin

Department of Mathematics

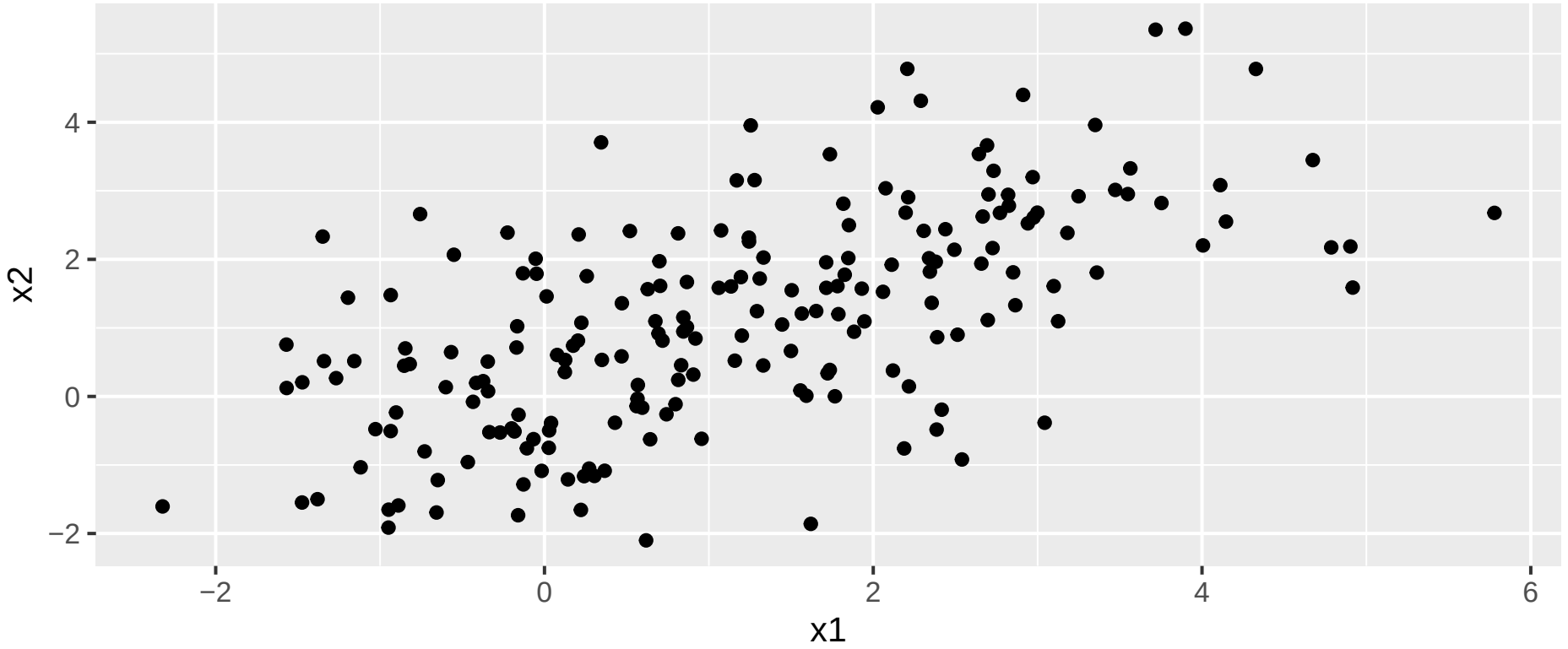
Intended Learning Outcome

Through this chapter, you should be able to

- apply hierarchical clustering methods,
- apply nonhierarchical clustering methods,
- apply EM algorithm for clustering.

Applications of Multivariate Analysis: Clustering

Classification pertains to a known number of groups, and the operational objective is to assign new observations to one of these groups. In clustering, no assumptions are made concerning the number of groups or the group structure.



Group Items or Variables

In cluster analysis, we can either group items/observations or variables.

- Suppose that we have a data set of different football players.
- For each player, we observe some variables (e.g., minutes played, number of passes, number of tackle, number of shoots, etc).
- We can group players (which players are similar) or variables (which measurements are similar).

Our focus in this course is to cluster items.

Distance for Items/Observations

Most efforts to produce a simple group structure from a complex data set require a measure of distance among items/observations. For continuous variables, we can compute

- Euclidean distance: $d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(\boldsymbol{x}-\boldsymbol{y})^T (\boldsymbol{x}-\boldsymbol{y})}$.
- Weighted Euclidean distance: $d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{(\boldsymbol{x}-\boldsymbol{y})^T \boldsymbol{A} (\boldsymbol{x}-\boldsymbol{y})}$.
- Minkowski distance: $d(\boldsymbol{x}, \boldsymbol{y}) = [\sum_{i=1}^{p} |x_i - y_i|^m]^{1/m}$. It becomes the Euclidean distance when $m = 2$.

Clustering Algorithms

It is not possible to examine all grouping possibilities. Hence, we want to find reasonable clusters without having to look at all configurations.

1. Agglomerative hierarchical methods start with each individual object as a cluster and combine the most similar objects.
2. Divisive hierarchical methods start with an initial single group that contains all objects and divide it into dissimilar subgroups.

Linkage Methods

Linkage methods are commonly used agglomerative hierarchical procedures, which are based on cluster distance. Let d_{ij} be the distance between objects (items or variables) i in cluster U and j in cluster V

1. Single linkage (nearest neighborhood): distance between their nearest members

$$\min_{i \in U, j \in V} d_{ij}.$$

2. Complete linkage: distance between their farthest members

$$\max_{i \in U, j \in V} d_{ij}.$$

3. Average linkage: average distance between pairs of members

$$\frac{\sum_i \sum_j d_{ik}}{N_U N_V},$$

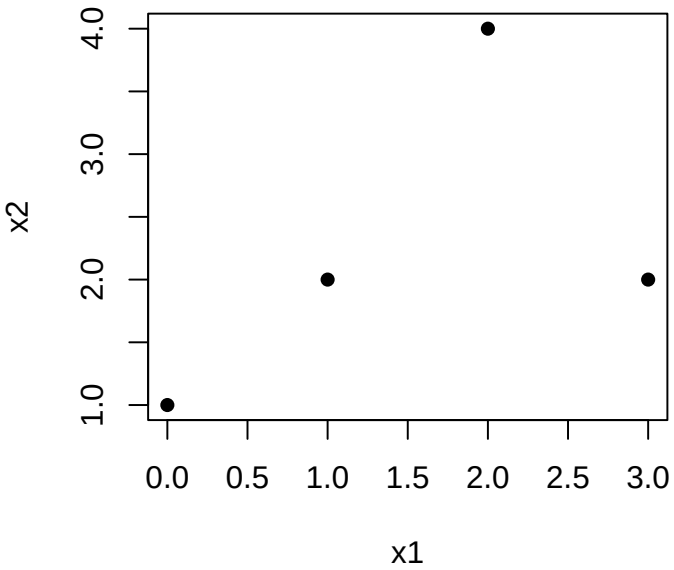
where N_U and N_V are the number of objects in clusters U and V.

Agglomerative Hierarchical Algorithm

Algorithm 1: Agglomerative Hierarchical Algorithm

1 Suppose that we have N objects (either items or variables) ;
2 Each object form its own cluster (N clusters) ;
3 **while** *total number of clusters is not* 1 **do**
4 Compute the matrix of distance or similarities for the current clusters. Each entry in the matrix is the distance between two clusters ;
5 Search the distance matrix for the nearest (most similar) pair of clusters (U, V) ;
6 Merge clusters U and V. Label the newly formed cluster (UV) ;
7 **end**

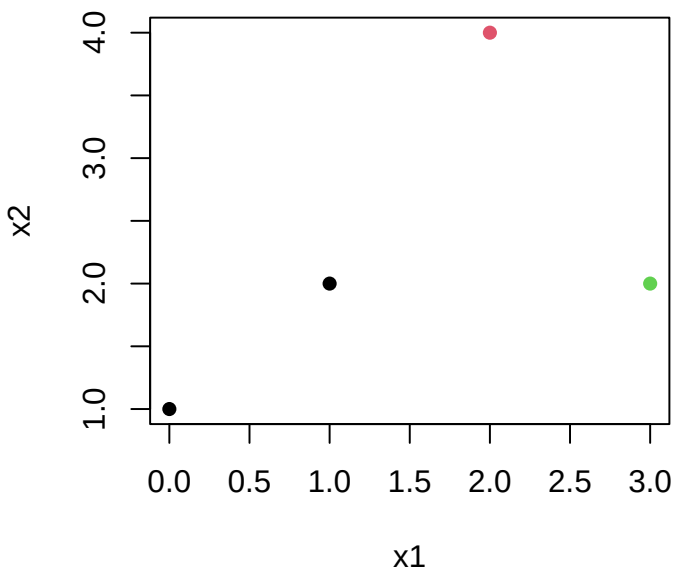
An Example



The distance matrix is

$$\boldsymbol{D} = \begin{bmatrix} 0 & \sqrt{2} & \sqrt{13} & \sqrt{10} \\ \sqrt{2} & 0 & \sqrt{5} & \sqrt{4} \\ \sqrt{13} & \sqrt{5} & 0 & \sqrt{5} \\ \sqrt{10} & \sqrt{4} & \sqrt{5} & 0 \end{bmatrix}.$$

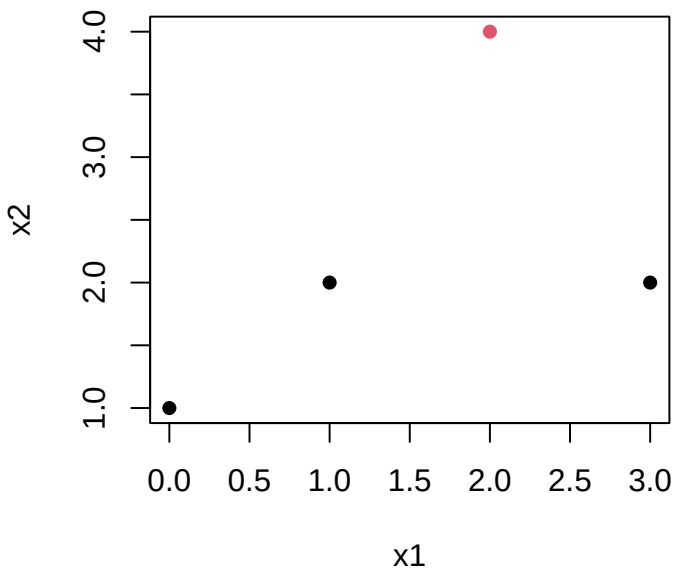
Example: Single Linkage



- The initial clusters are (1), (2), (3), and (4).
- The smallest distance is $d_{12} = \sqrt{2}$. Hence, the updated clusters are (12), (3), (4).
- The updated distance matrix is

$$\begin{bmatrix} 0 & \sqrt{5} & \sqrt{4} \\ \sqrt{5} & 0 & \sqrt{5} \\ \sqrt{4} & \sqrt{5} & 0 \end{bmatrix}.$$

Example: Single Linkage

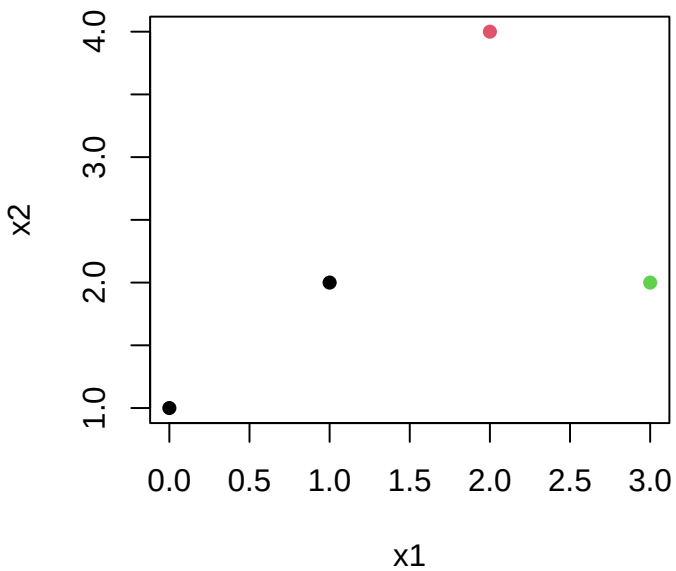


- The smallest distance is $d_{(12)4} = \sqrt{4}$. Hence, the updated clusters are (124), (3).
- The updated distance matrix is

$$\begin{bmatrix} 0 & \sqrt{5} \\ \sqrt{5} & 0 \end{bmatrix}.$$

The final cluster is (12345).

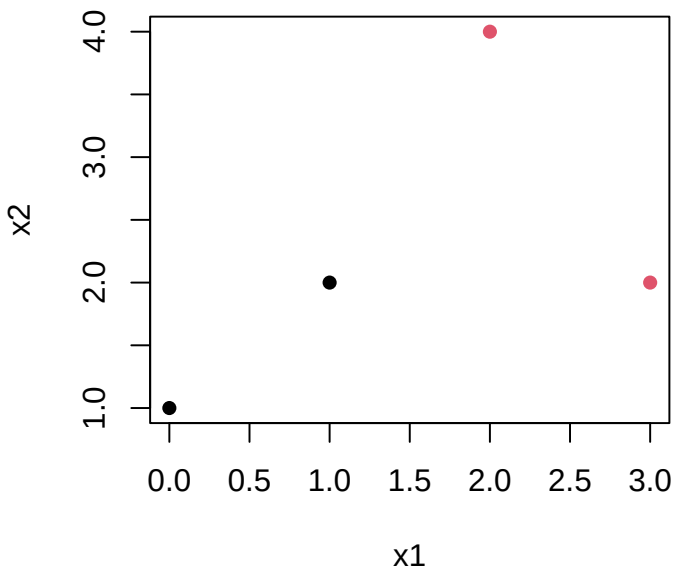
Example: Complete Linkage



- The initial clusters are (1), (2), (3), and (4).
- The smallest distance is $d_{12} = \sqrt{2}$. Hence, the updated clusters are (12), (3), (4).
- The updated distance matrix is

$$\begin{bmatrix} 0 & \sqrt{13} & \sqrt{10} \\ \sqrt{13} & 0 & \sqrt{5} \\ \sqrt{10} & \sqrt{5} & 0 \end{bmatrix}.$$

Example: Complete Linkage

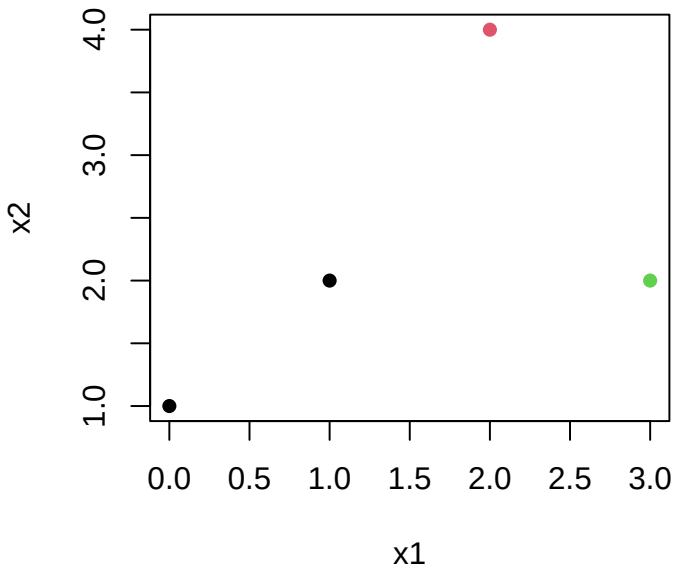


- The smallest distance is $d_{34} = \sqrt{5}$. Hence, the updated clusters are (12), (34).
- The updated distance matrix is

$$\begin{bmatrix} 0 & \sqrt{13} \\ \sqrt{13} & 0 \end{bmatrix}.$$

The final cluster is (12345).

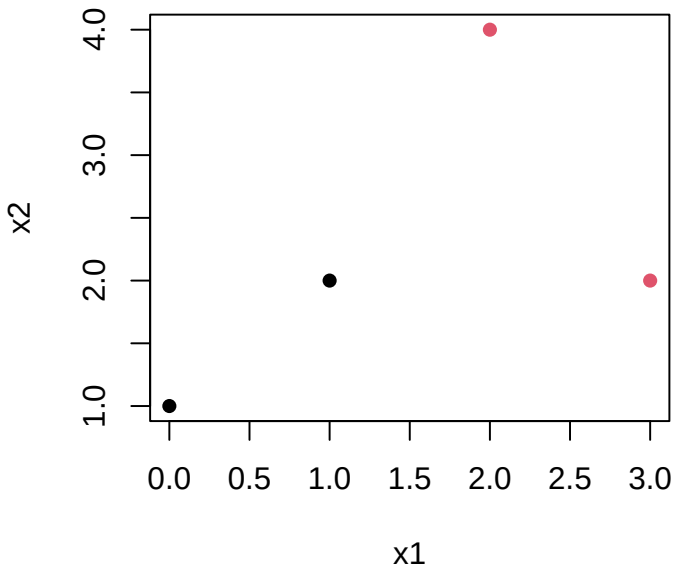
Example: Average Linkage



- The initial clusters are (1), (2), (3), and (4).
- The smallest distance is $d_{12} = \sqrt{2}$. Hence, the updated clusters are (12), (3), (4).
- The updated distance matrix is

$$\begin{bmatrix} 0 & \frac{\sqrt{13}+\sqrt{5}}{2} & \frac{\sqrt{10}+\sqrt{4}}{2} \\ \frac{\sqrt{13}+\sqrt{5}}{2} & 0 & \sqrt{5} \\ \frac{\sqrt{10}+\sqrt{4}}{2} & \sqrt{5} & 0 \end{bmatrix}.$$

Example: Average Linkage



- The smallest distance is $d_{34} = \sqrt{5}$. Hence, the updated clusters are (12), (34).
- The updated distance matrix is

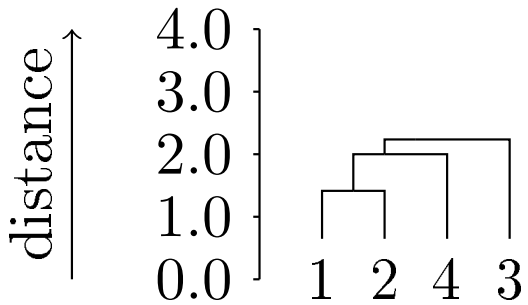
$$\begin{bmatrix} 0 & \frac{\sqrt{13}+\sqrt{5}+\sqrt{10}+\sqrt{4}}{4} \\ \frac{\sqrt{13}+\sqrt{5}+\sqrt{10}+\sqrt{4}}{4} & 0 \end{bmatrix}.$$

The final cluster is (12345).

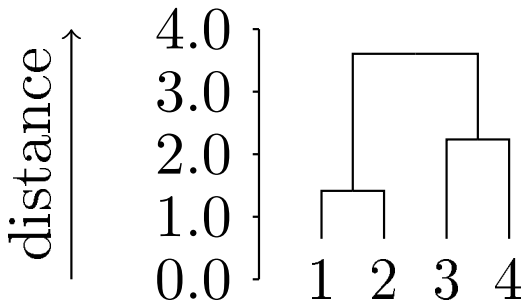
Dendrogram

The results of hierarchical clustering can be displayed as a dendrogram that illustrates the mergers or divisions.

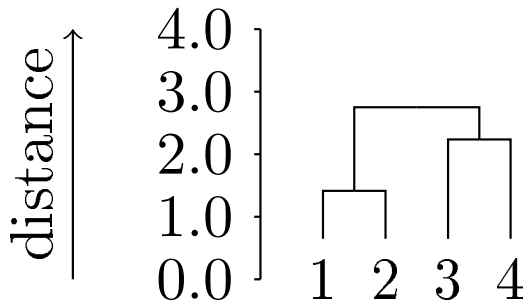
Single



Complete



Average



More Linkage Methods

- Let $\bar{\boldsymbol{X}}_U$ and $\bar{\boldsymbol{X}}_V$ be the group averages. Then, the cluster distance used by centroid linkage is

$$\sqrt{\left(\bar{\boldsymbol{X}}_U - \bar{\boldsymbol{X}}_V\right)^T \left(\bar{\boldsymbol{X}}_U - \bar{\boldsymbol{X}}_V\right)}.$$

- Minmax linkage is based on the cluster distance defined by

$$\min_{i \in U \cup V} \max_{j \in U \cup V} d_{ij},$$

  the smallest radius that encompassing all points in U and V.
- There are more linkage methods, each yield an agglomerative clustering method.

Ward's Hierarchical Clustering Method

Ward's hierarchical clustering method is based on the loss of information from joining two groups.

1. Start with each cluster consisting of a single item.
2. For a given cluster k, let

$$\mathrm{SS}_k = \sum_{j \in \text{cluster } k} \left(\boldsymbol{x}_j - \bar{\boldsymbol{x}}^{(k)}\right)^T \left(\boldsymbol{x}_j - \bar{\boldsymbol{x}}^{(k)}\right),$$

   where $\bar{\boldsymbol{x}}^{(k)}$ is the sample average of cluster k. Compute $\mathrm{SS} = \sum_{j=1}^{K} \mathrm{SS}_j$.
3. Consider the union of every possible pair of clusters. The two clusters whose combination results in the smallest increase in SS are joined.

Some Remarks

- Some clustering procedures may encounter inversion. An inversion occurs if an object joints an cluster at a small distance than that of a previous consolidation.
  - For example, C joints (AB) at distance 32. But D joints (ABC) at distance 30.
  - Inversions can occur when there is no clear cluster structure.
- For a given data set, it is a good idea to try several clustering methods and a couple different ways of assigning distances.
  - You can also add some small noises to the observed data and apply agglomerative hierarchical clustering procedure and check how stable they are.

Pros and Cons

- Single linkage cannot discern poorly separated clusters. But it can delineate nonellipsoidal clusters. It often creates a cluster with a long string shape, known as chaining.
- Complete linkage ensures that all items in a cluster are within some maximum distance of each other. But it may not merge close groups because of outliers in the group.
- Average linkage accounts for the distance between all pairs of items. But even a monotone transformation of distance may change the results.
- Centroid linkage is easy to understand, but can create inversion.
- Minimax linkage creates clusters whose centers are among the data points.
- Ward's method requires the Euclidean distance. It works well when the clusters of observations are expected to be roughly elliptically shaped.

K-Means Method

Nonhierarchical clustering methods are designed to group items, rather than variables into clusters. One commonly used approach is the $K-$means method.

Algorithm 2: K-Means Method

```
Specify a K;
Partition the items into K initial clusters ;
while convergence not met do
    for i from 1 to N do
        Assign item i to the cluster whose centroid (mean) is nearest;
        Recalculate the centroid for the cluster receiving the new
         item and for the cluster losing the item ;
    end
    Break the while-loop until no more reassignments take place ;
end
```

A Hand-Calculation Example

Suppose that we observe

Item	X_1	X_2
A	5	3
B	-1	1
C	1	-2
D	-3	-2

1. We want to divide them into $K = 2$ clusters.
2. (AB) and (CD). The centroids are $(2, 2)$ and $(-2, -2)$.
3. For each item, we compute its squared (Euclidean) distance from the group centroid. For item A,

	A does not move		A is moved	
Item	(AB)	(CD)	B	(ACD)
A	10	61	40	27.1

A is not moved to a new cluster, since it is closest to (AB).

A Hand-Calculation Example

4 For item B,

	B does not move		B is moved	
Item	(AB)	(CD)	A	(BCD)
B	10	9	40	4

B is closer to (BCD). Hence, we move B to the new cluster and obtain clusters A and (BCD).

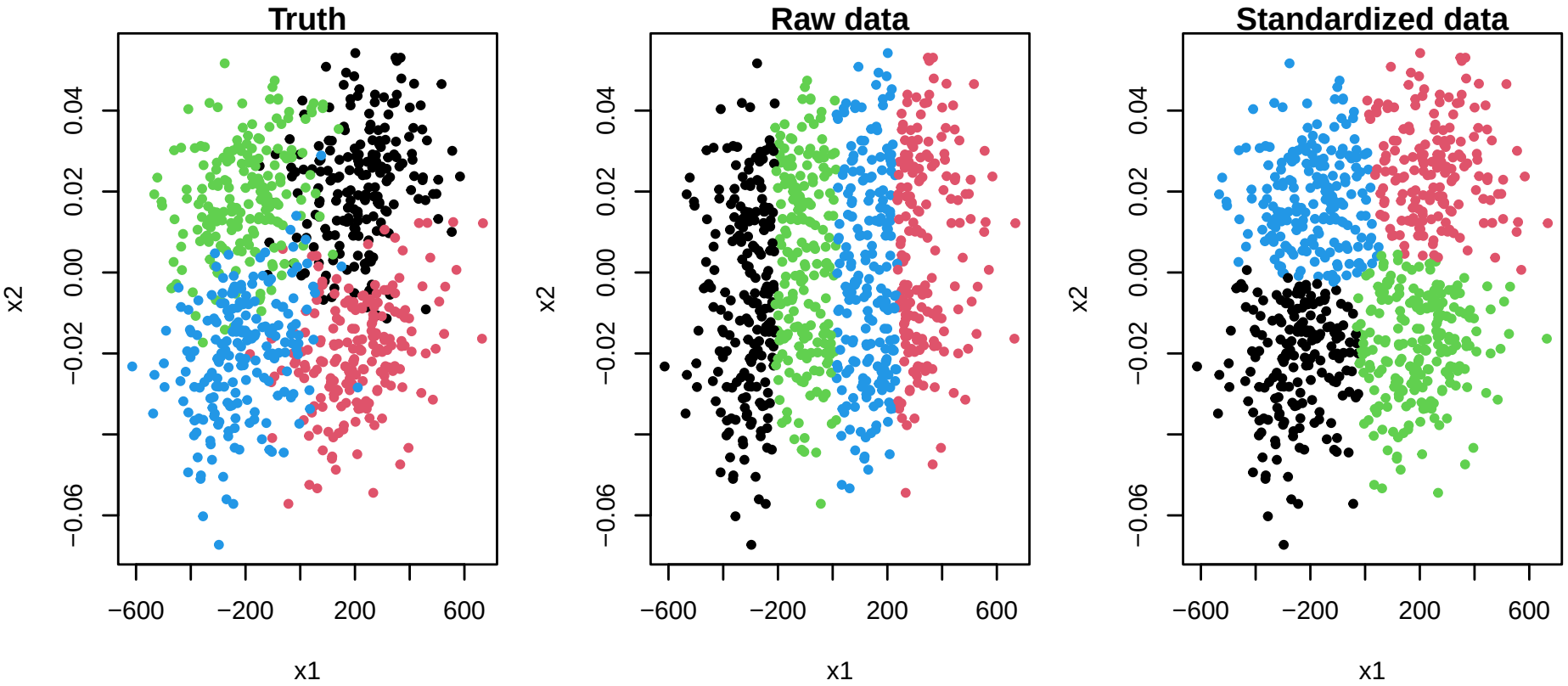
5 We proceed forward with C, D, A, B, C, D, A, ..., until no reassignment occurs. The final partition is A and (BCD).

Some Details

1. For a given K, you need to specify K initial clusters. You can
   - specify some initial partitions of items
   - or specify the centroid of each cluster.

   For example, you can have random partition, or pick the farthest items.
2. The final clusters are likely to depend on the initial partition. Hence, you can apply different random initial clusters and check whether the method is robust. You can also choose the best partition among them.
3. Scale of your data is also important.

Scale is Important



How to Choose Number of Clusters

In both hierarchical clustering and nonhierarchical clustering, we need to prespecify the number of clusters, which is subject to discussion. In practice, you can

1. use subject matter knowledge,
2. or, specify several different values and see which result makes more sense to you,
3. or, specify several different values and compute some measures.

But keep in mind, this is a super hard problem!

CH Index and Gap Statistic

For K clusters, we want the within sum of squares $W(K)$ to be small and the between sum of squares $B(K)$ to be large.

- We pick K that maximize the CH index

$$\mathrm{CH}(K) = \frac{B(K)/(K-1)}{W(K)/(n-K)}.$$

- Gap statistic: compare $W(K)$ to the within-cluster variation of uniformly distributed points.
- We pick K that maximizes the silhouette statistic:

$$\sum_{i=1}^{n} \frac{b_i - a_i}{\max\{a_i, b_i\}}$$

where a_i is the average distance between subject i and all other data points in the same cluster, and b_i is the average distance from subject i to the closet cluster that does not contain i.

Parametric Approach: Mixing Distribution

The above methods are mostly "model-free". It is also common to perform cluster analysis with some assumptions on our model. We assume that our data are generated in a hierarchical manner.

- Let Z be a multinomial random variable that indicates which group the item comes from:

$$P(Z = k) \quad = \quad p_k \geq 0.$$

- Within group k $(Z = k)$, our data follow some distribution with density function $f_k(\boldsymbol{x})$. That is,

$$\boldsymbol{X} \mid Z = k \quad \sim \quad f_k(\boldsymbol{x}; \boldsymbol{\theta}_k).$$

- Then, the observation vector for a single object is modeled as a finite mixing distribution

$$f(\boldsymbol{x}) \quad = \quad \sum_{k=1}^{K} f(\boldsymbol{x}, Z = k) = \sum_{k=1}^{K} p_k f_k(\boldsymbol{x}; \boldsymbol{\theta}_k).$$

Complete Likelihood and Observed Likelihood

If both $\boldsymbol{X}$ and Z were observed, then the likelihood function is

$$L_C = \prod_{j=1}^{N} f\left(\boldsymbol{x}_j, Z_j\right) = \prod_{j=1}^{N} \left[f\left(\boldsymbol{x}_j, Z_j = k\right)\right]^{I(Z_j=k)} = \prod_{j=1}^{N} \left[p_k f_k\left(\boldsymbol{x}; \boldsymbol{\theta}_k\right)\right]^{I(Z_j=k)}.$$

However, we only observe $\boldsymbol{X}$ in practice, the likelihood is

$$L = \prod_{j=1}^{N} f\left(\boldsymbol{x}_j\right) = \prod_{j=1}^{N} \left[\sum_{k=1}^{K} f\left(\boldsymbol{x}_j, Z_j = k\right)\right] = \prod_{j=1}^{N} \left[\sum_{k=1}^{K} p_k f_k\left(\boldsymbol{x}; \boldsymbol{\theta}_k\right)\right].$$

L_C is known as the complete likelihood and L is known as the observed likelihood. The maximum likelihood estimators of unknown parameters maximize the observed likelihood.

EM Algorithm

However, if we do not observe Z, directly maximizing the observed likelihood is computationally not easy. The Expectation-Maximization (EM) algorithm is often used instead, which is one of the most cited method in statistics.

Algorithm 3: EM Algorithm

Suppose that the parameter vector is $\boldsymbol{\theta}$. Obtain initial guess $\boldsymbol{\theta}^{(0)}$;

while *At step t* **do**

  E step: Find conditional expectation $\mathbb{E}\left(\log f\left(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\theta}\right) \mid \boldsymbol{X}; \hat{\boldsymbol{\theta}}^{(t)}\right)$ using $P\left(\boldsymbol{Z} \mid \boldsymbol{X}; \hat{\boldsymbol{\theta}}^{(t)}\right)$;

  M step: Maximize the conditional expectation with respect to $\boldsymbol{\theta}$ and obtain the updated value $\hat{\boldsymbol{\theta}}^{(t+1)}$;

end

E Step

The E step computes $\mathbb{E}\left(\log f\left(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\theta}\right) \mid \boldsymbol{X}; \hat{\boldsymbol{\theta}}^{(t)}, \hat{\boldsymbol{p}}^{(t)}\right)$ given by

$$Q\left(\boldsymbol{\theta}, \boldsymbol{p} \mid \hat{\boldsymbol{\theta}}^{(t)}, \hat{\boldsymbol{p}}^{(t)}\right) := \sum_{j=1}^{N} \sum_{k=1}^{K} \left[\log p_k + \log f\left(\boldsymbol{x}_j \mid Z_j = k; \boldsymbol{\theta}\right)\right] P_{(t)}\left(Z_j = k \mid \boldsymbol{x}_j\right),$$

where $f\left(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\theta}\right)$ is the complete likelihood L_C, and, according to the Bayes rule,

$$P_{(t)}\left(Z_j = k \mid \boldsymbol{x}_j\right) = \frac{P\left(\boldsymbol{x}_j \mid Z_j = k; \hat{\boldsymbol{\theta}}^{(t)}\right) \hat{p}_k^{(t)}}{\sum_{k=1}^{K} P\left(\boldsymbol{x}_j \mid Z_j = k; \hat{\boldsymbol{\theta}}^{(t)}\right) \hat{p}_k^{(t)}}.$$

M Step

The M step maximizes $\mathbb{E}\left(\log f\left(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\theta}\right) \mid \boldsymbol{X}; \hat{\boldsymbol{\theta}}^{(t)}, \hat{\boldsymbol{p}}^{(t)}\right)$. The gradient with respect to p_k is

$$\frac{\partial Q\left(\boldsymbol{\theta}, \boldsymbol{p} \mid \hat{\boldsymbol{\theta}}^{(t)}, \hat{\boldsymbol{p}}^{(t)}\right)}{\partial p_k} = \sum_{j=1}^{N}\left[\frac{P_{(t)}\left(Z_j = k \mid \boldsymbol{x}_j\right)}{p_k} - \frac{P_{(t)}\left(Z_j = K \mid \boldsymbol{x}_j\right)}{p_K}\right],$$

and the gradient with respect to $\boldsymbol{\theta}$ is

$$\frac{\partial Q\left(\boldsymbol{\theta}, \boldsymbol{p} \mid \hat{\boldsymbol{\theta}}^{(t)}, \hat{\boldsymbol{p}}^{(t)}\right)}{\partial \boldsymbol{\theta}} = \sum_{j=1}^{N}\sum_{k=1}^{K}\frac{\partial \log f\left(\boldsymbol{x}_j \mid Z_j = k; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} P_{(t)}\left(Z_j = k \mid \boldsymbol{x}_j\right).$$

The maximizer is the updated $\boldsymbol{\theta}$, denoted by $\hat{\boldsymbol{\theta}}^{(t+1)}$, either closed form expression or numerical methods.

EM Estimator

- It can be shown that the solution $\left(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{p}}\right)$ of the EM algorithm is equivalent to the MLE estimator of the observed likelihood.
- We can then compute the (posterior) probability of a subject belonging to each cluster using

$$P\left(Z_j = k \mid \boldsymbol{x}_j; \hat{\boldsymbol{\theta}}\right) = \frac{P\left(\boldsymbol{x}_j \mid Z_j = k; \hat{\boldsymbol{\theta}}\right) \hat{p}_k}{\sum_{k=1}^{K} P\left(\boldsymbol{x}_j \mid Z_j = k; \hat{\boldsymbol{\theta}}\right) \hat{p}_k}.$$

- We still need to determine K. One way is to use the information criterion

$$\begin{aligned} \text{AIC} &= -2\ell\left(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{p}}\right) + 2 \times \text{number of parameters}, \\ \text{BIC} &= -2\ell\left(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{p}}\right) + \log n \times \text{number of parameters}. \end{aligned}$$

Gaussian Mixture

The most common model is the Gaussian mixture where $f_k(\boldsymbol{x})$ is the density of $N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.

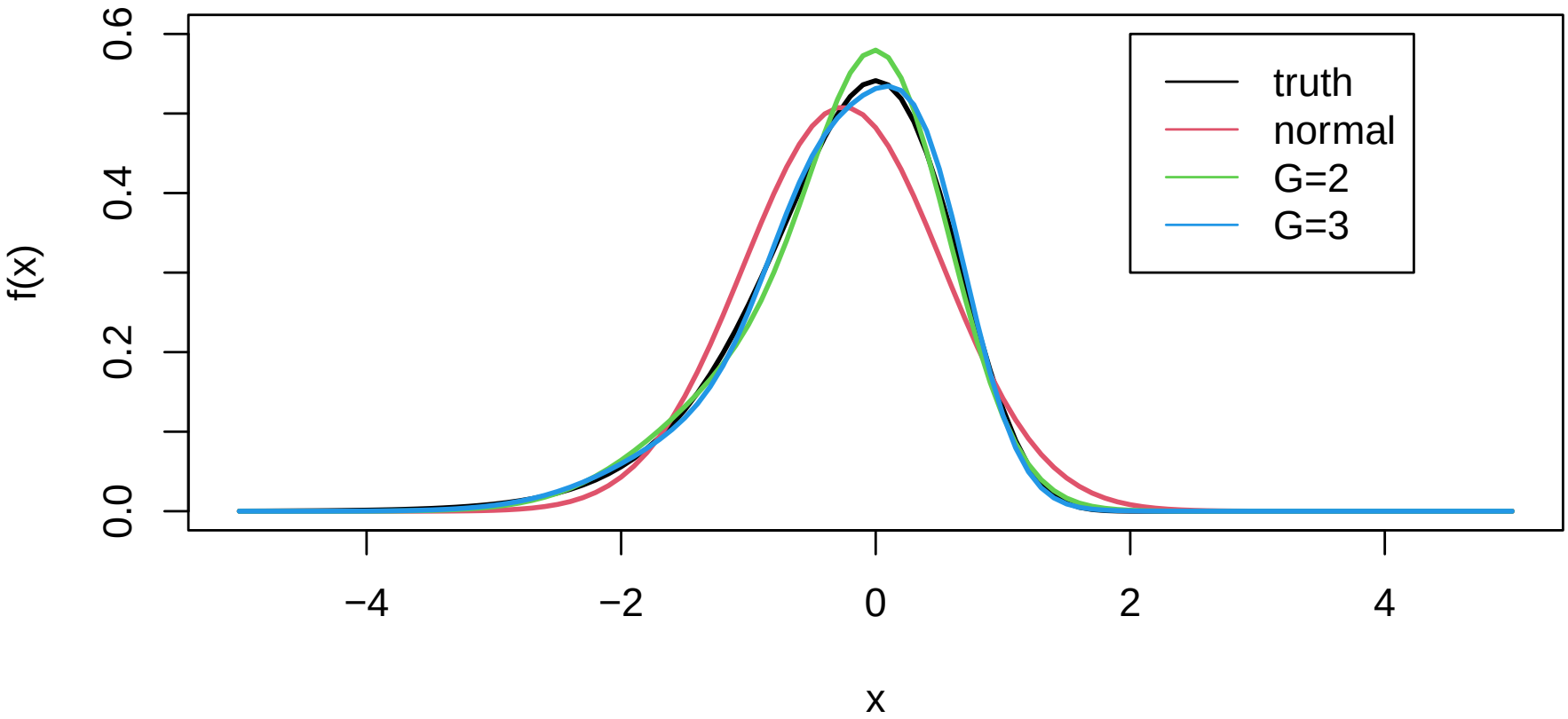
EM Algorithm For Gaussian Mixture

Suppose that we have observed a random sample $\boldsymbol{x}_1$, ..., $\boldsymbol{x}_n$, but we do not observe Z. We assume that

$$\begin{aligned} \boldsymbol{X} \mid Z = k &\sim N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \\ p_k &= P(Z = k). \end{aligned}$$

Find the expression of p_k, $\boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma}_k$.

Gaussian Mixture For Other Applications



Other Finite Mixture Applications

For many statistical models that require normal assumptions, violating normality can often cause severe consequences. However, we can assume that data are heterogeneity.

- We can assume that our data come from different unobserved populations. Each population follows a normal distribution. The distribution of the observed data is

$$f(\boldsymbol{x}) = \sum_{k=1}^{K} p_k f_k(\boldsymbol{x}; \boldsymbol{\theta}_k).$$

- EM algorithm is used to estimate the parameters and cluster observations into unknown clusters.
- For example, in factor analysis we can assume normality as

$$\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{L}\boldsymbol{F} + \boldsymbol{e} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

or a Gaussian mixture as in cluster k,

$$\boldsymbol{X}_k = \boldsymbol{\mu}_k + \boldsymbol{L}_k\boldsymbol{F}_k + \boldsymbol{e}_k \sim N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$