Cluster Analysis

What do you think about “Cluster”?

]

Outline

- Introduction
- Conceptual support of Cluster Analysis
- How does Cluster Analysis work?
- Measuring Similarity
- Forming clusters:
  - Hierarchical Clustering Methods
  - Non Hierarchical Clustering Method
- Measuring Heterogeneity (number of cluster)

Introduction

- Searching the data for a structure of "natural" groupings is an important exploratory technique.
- Groupings can provide an informal means for assessing:
  - dimensionality,
  - identifying outliers,
  - and suggesting interesting hypotheses concerning relationships.

- Cluster analysis is a group of multivariate techniques whose primary purpose is to group objects (e.g. respondents, products or other entities) based on characteristics they possess.
- No assumptions are made concerning the number of groups or the group structure.
- Grouping is done on the basis of similarities or distances (dissimilarities) .

- Cluster analysis differs from factor analysis:
  - Cluster analysis groups objects, whereas factor analysis groups variables
  - Factor analysis makes the grouping based on pattern of variation (correlation) in the data, whereas cluster analysis makes grouping on the basis of distance (proximity)

- Example:

Marketing – In the area of marketing, we use clustering to explore and select customers that are potential buyers of the product. This differentiates the most likeable customers from the ones who possess the least tendency to purchase the product. After the clusters have been developed, *businesses can keep a track of their customers and make necessary decisions to retain them in that cluster.*

Retail – Retail industries make use of clustering to group customers based on their preferences, style, choice of wear as well as store preferences. This allows them to manage their stores in a much more efficient manner.

Medical Science – Medicine and health industries make use of clustering algorithms to *facilitate efficient diagnosis and treatment of their patients as well as the discovery of new medicines.* Based on the age, group, genetic coding of the patients, these organizations are better capable to understand diagnosis through robust clustering.

Sociology – Clustering is used in Data Mining operations to *divide people based on their demographics, lifestyle, socioeconomic status,* etc. This can help the law enforcement agencies to group potential criminals and even identify them with an efficient implementation of the clustering algorithm.

The Objectives of Cluster Analysis

1. Taxonomy description
    - The most traditional use of cluster analysis has been for exploratory purposes and formation of a taxonomy (an empirically based classification of objects)
    - It can also generate hypotheses related to the structure of the objects.
    - It also can be used for confirmatory purposes
2. Data simplification
    - Cluster analysis also develop a simplified perspective by grouping observation for further analysis
    - Whereas factor analysis attempt to provide dimensions/structure to variables, cluster analysis performs the same tasks for observations
3. Relationship identification
    - With the cluster defined, the researcher has a means revealing relationship among the observations that typically is not possible with the individual observations.

Conceptual support of Cluster Analysis

The most common criticism must be addressed by conceptual support:

- Cluster analysis is descriptive, atheoretical, and non inferential,
  - It is only an exploratory technique
  - Nothing guarantee unique solutions
- Cluster analysis is always create clusters, regardless of the actual existence of any structure in the data.
  - When using cluster analysis, the researcher is making an assumption of some structure among the objects.
  - The researcher should always remember that just because clusters can be found, does not validate their existence.
- The cluster solution is not generalizable because is totally dependent upon the variables used as the basis for the similarity measures

How does Cluster Analysis work?

The primary objective of cluster analysis is to define the structure of the data by placing the most similar observations into groups.

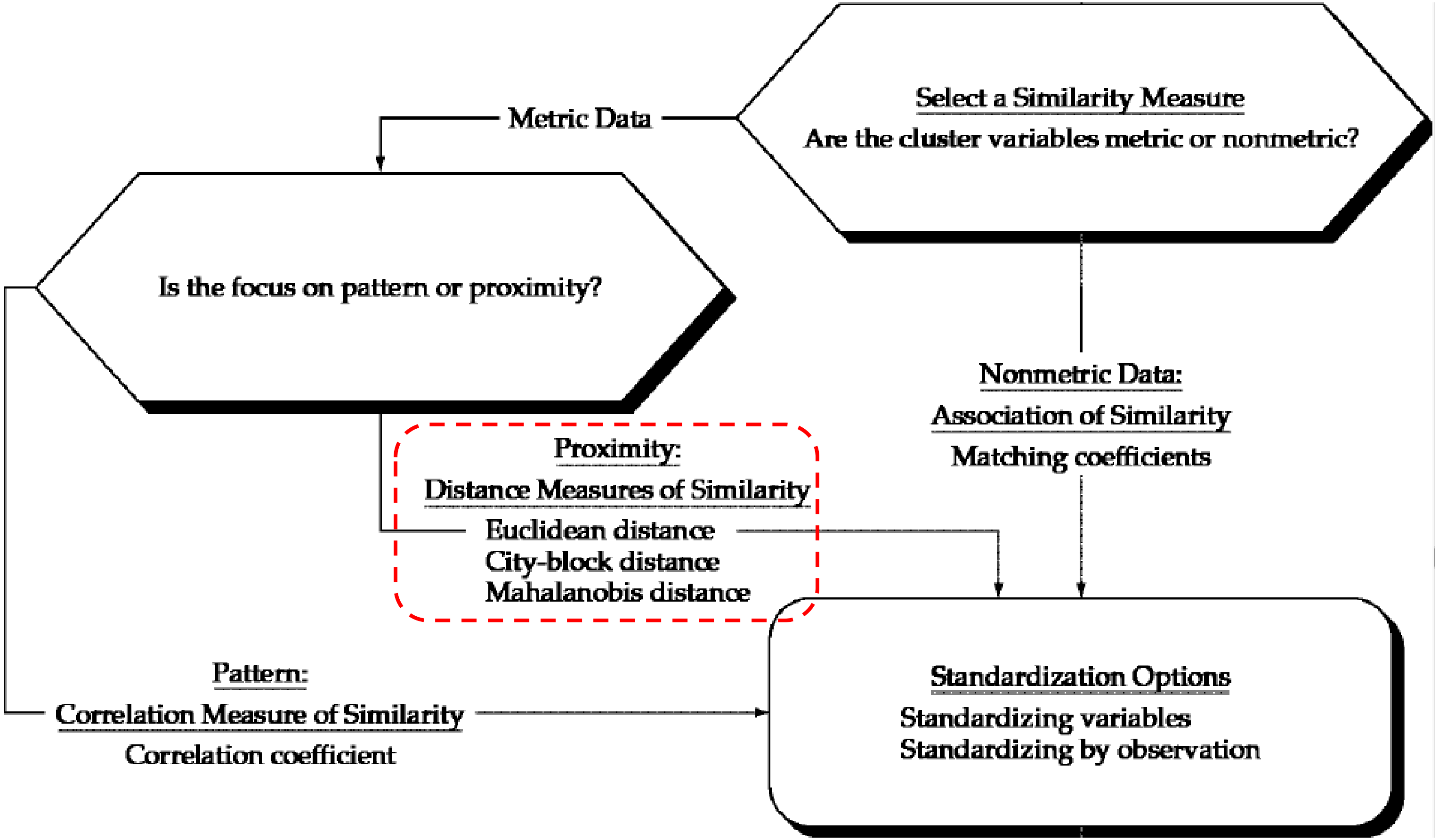
To accomplish the tasks, we must address three basic questions:

1. How do we **measure similarity**?
2. How do we **form clusters**?
   - Hierarchical method
   - Non hierarchical method
3. How many **groups** do we form?
   - Measuring heterogeneity

Cluster analysis decision process:

- Partitioning the data set to form clusters and selecting to cluster solution
- Interpreting the cluster to understand the characteristics of each cluster and develop a name/label that appropriately defines its nature.
- Validating the result of the final cluster solution (i.e. , determining the stability)

Measuring Similarity



Select a Similarity Measure

Are the cluster variables metric or nonmetric?

Metric Data

Is the focus on pattern or proximity?

Proximity:

Distance Measures of Similarity

Euclidean distance

City-block distance

Mahalanobis distance

Nonmetric Data:

Association of Similarity

Matching coefficients

Pattern:

Correlation Measure of Similarity

Correlation coefficient

Standardization Options

Standardizing variables

Standardizing by observation

Distance measures:

Euclidian distance (It is often preferred for clustering).

The Euclidian distance between two p-dimensional observations (items), $\mathbf{x}' = [x_1, x_2, \dots, x_p]$ and $\boldsymbol{y}' = [y_1, y_2, \dots, y_p]$ is

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_p - y_p)^2}$$

$$= \sqrt{(\mathbf{x} - \mathbf{y})'(\mathbf{x} - \mathbf{y})}$$

Squared (absolute) Euclidian distance

$$d^2(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})'(\mathbf{x} - \mathbf{y})$$

It is recommended for centroid or Ward's method of clustering

Mahalanobis distance (statistical distance)

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})'\boldsymbol{S}^{-1}(\mathbf{x} - \mathbf{y})}$$

When the variables are correlated, the Mahalanobis distance is likely the most appropriate

Minkowski metric

$$d(\mathbf{x}, \mathbf{y}) = \left[\sum_{i=1}^{p} |x_i - y_i|^m\right]^{1/m}$$

For $m = 1$, it measures the **city-block** distance

For $m = 2$, it become the Euclidian distance

Two additional popular measures of "distance" or dissimilarity are given by the **Canberra metric** and the **Czekanowski coefficient**. Both of these measures are defined for nonnegative variables only

Canberra metric

$$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{p} \frac{|x_i - y_i|}{(x_i + y_i)}$$

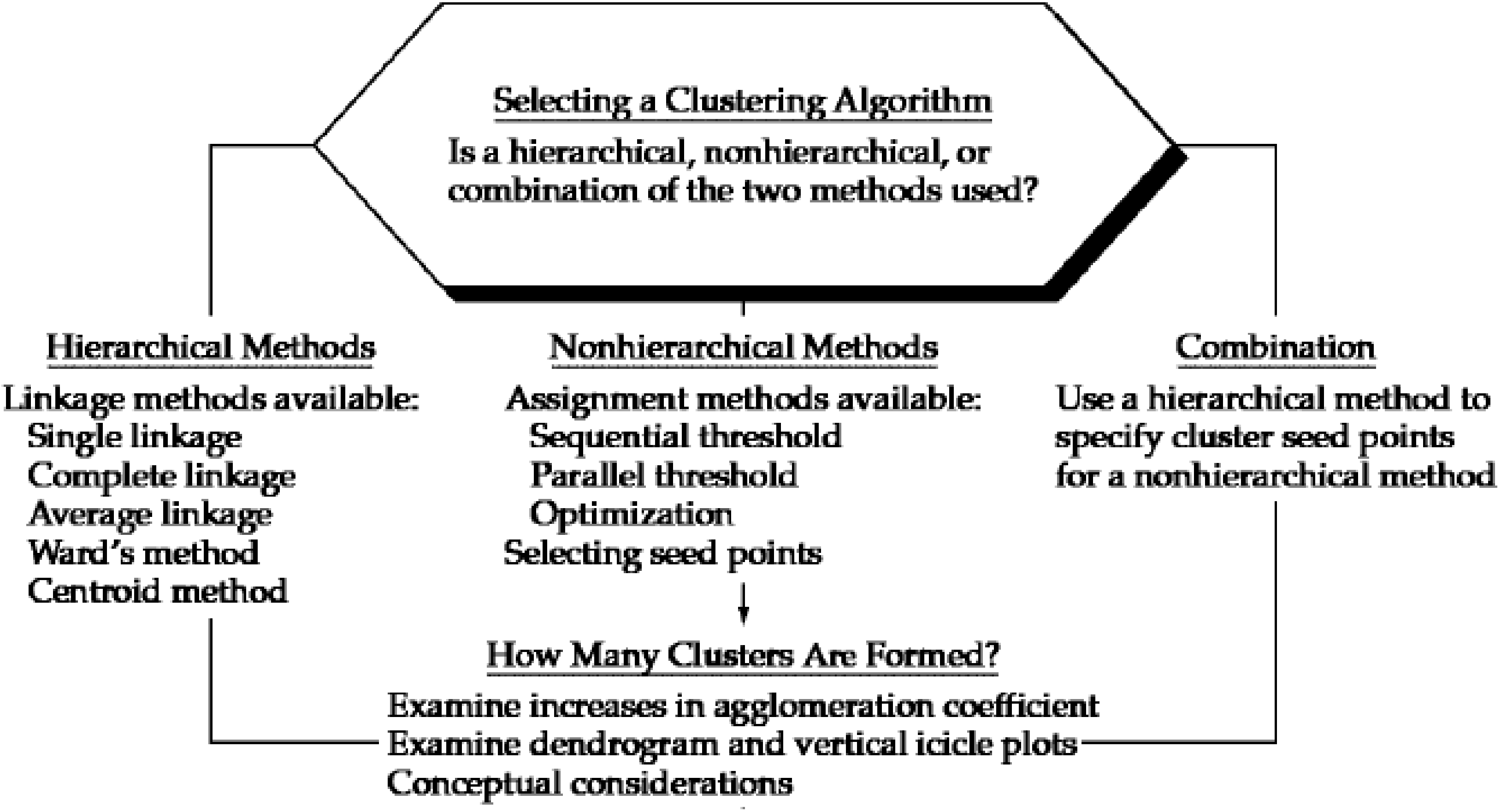
Czekanowski coefficient

$$d(\mathbf{x}, \mathbf{y}) = 1 - \frac{2 \sum_{i=1}^{p} \min(x_i, y_i)}{\sum_{i=1}^{p} (x_i + y_i)}$$

Greater distance means observations are less similar

“The researcher is encouraged to explore alternative cluster solution obtained when using different distance measures in an effort to best represent the underlying data patterns.”

Forming Cluster



Selecting a Clustering Algorithm

Is a hierarchical, nonhierarchical, or combination of the two methods used?

Hierarchical Methods

Linkage methods available:
- Single linkage
- Complete linkage
- Average linkage
- Ward's method
- Centroid method

Nonhierarchical Methods

Assignment methods available:
- Sequential threshold
- Parallel threshold
- Optimization

Selecting seed points

Combination

Use a hierarchical method to specify cluster seed points for a nonhierarchical method

How Many Clusters Are Formed?

Examine increases in agglomeration coefficient
Examine dendrogram and vertical icicle plots
Conceptual considerations

Three types of clustering methods (Beverrit & Hothorn, 2011):

- Agglomerative hierarchical techniques,
- k-means clustering, and
- model-based clustering.

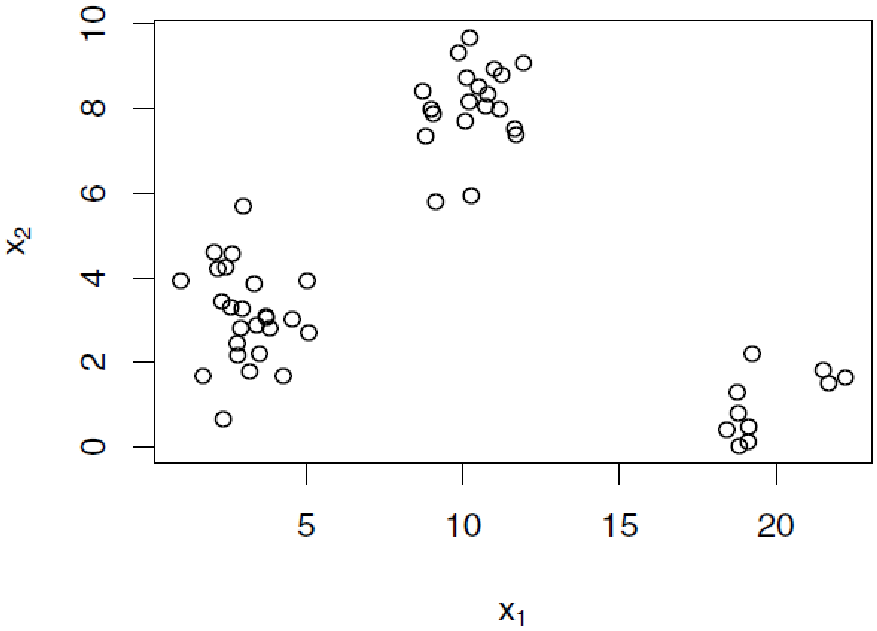
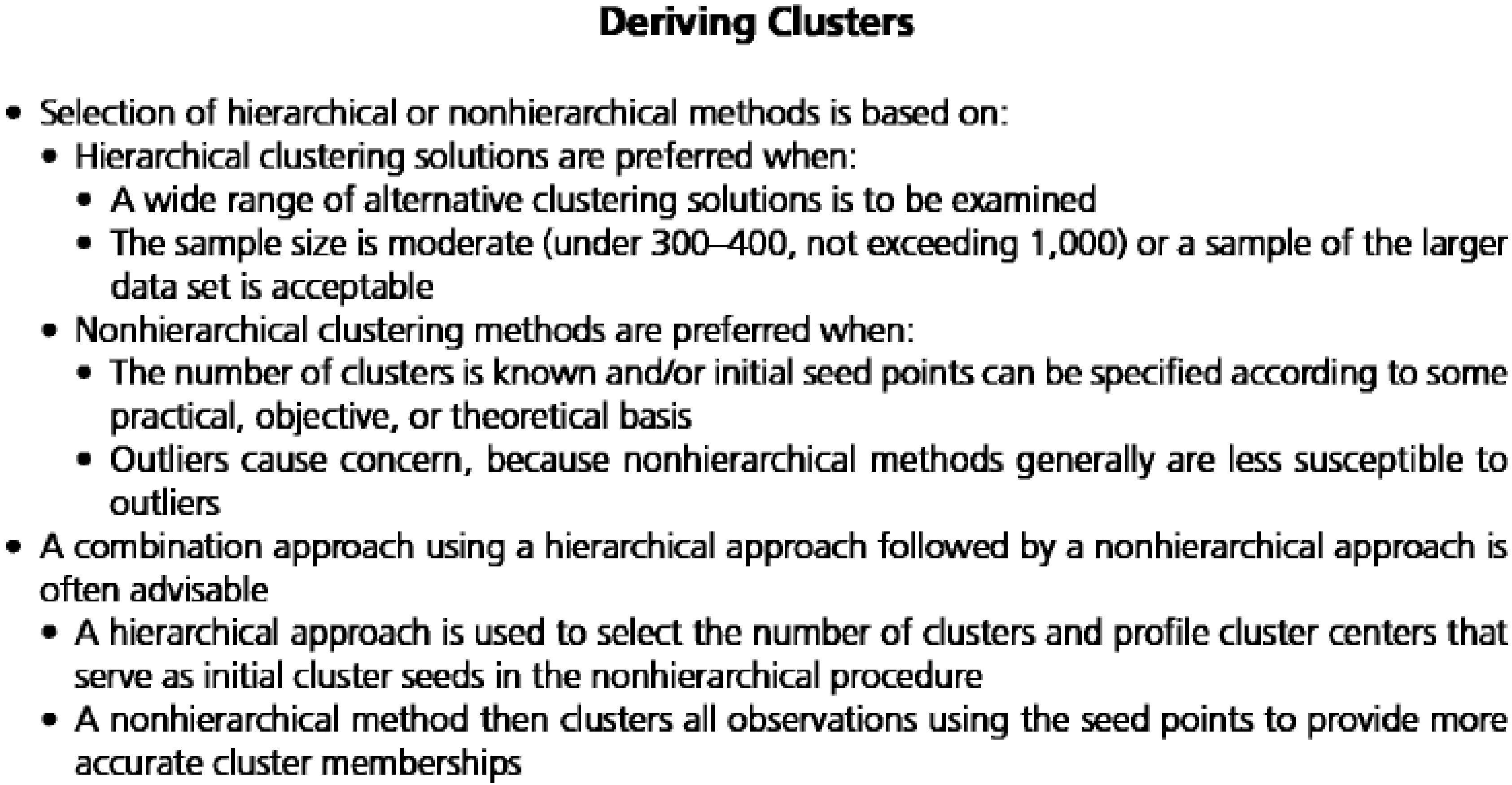


Fig. 6.1. Bivariate data showing the presence of three clusters.



Deriving Clusters

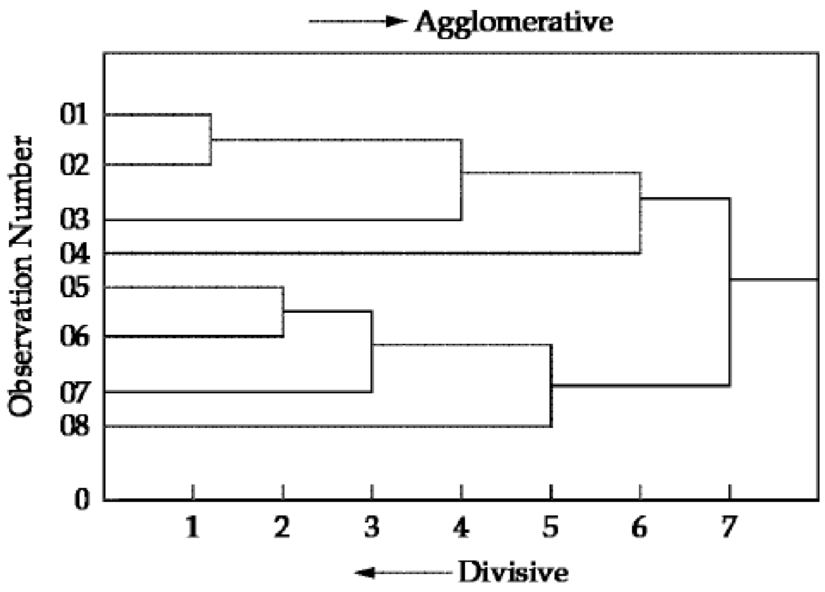
- Selection of hierarchical or nonhierarchical methods is based on:
  - Hierarchical clustering solutions are preferred when:
    - A wide range of alternative clustering solutions is to be examined
    - The sample size is moderate (under 300–400, not exceeding 1,000) or a sample of the larger data set is acceptable
  - Nonhierarchical clustering methods are preferred when:
    - The number of clusters is known and/or initial seed points can be specified according to some practical, objective, or theoretical basis
    - Outliers cause concern, because nonhierarchical methods generally are less susceptible to outliers
- A combination approach using a hierarchical approach followed by a nonhierarchical approach is often advisable
  - A hierarchical approach is used to select the number of clusters and profile cluster centers that serve as initial cluster seeds in the nonhierarchical procedure
  - A nonhierarchical method then clusters all observations using the seed points to provide more accurate cluster memberships

Hierarchical Clustering Method

The two basic types of hierarchical clustering procedure are:

- Agglomerative (linkage methods)
  Start with the individual objects (N clusters). The most similar objects are first grouped, and these initial groups are merged according to their similarities. Eventually, as the similarity decreases, all subgroups are fused into a single cluster.
- Divisive (work in the opposite direction).
  An initial single group of objects is divided into two subgroups such that the objects in one subgroup are "far from" the objects in the other. These subgroups are then further divided into dissimilar subgroups; the process continues until there are as many subgroups as objects-that is, until each object forms a group

The results of both agglomerative and divisive methods may be displayed in the form of a two-dimensional diagram known as a dendrogram

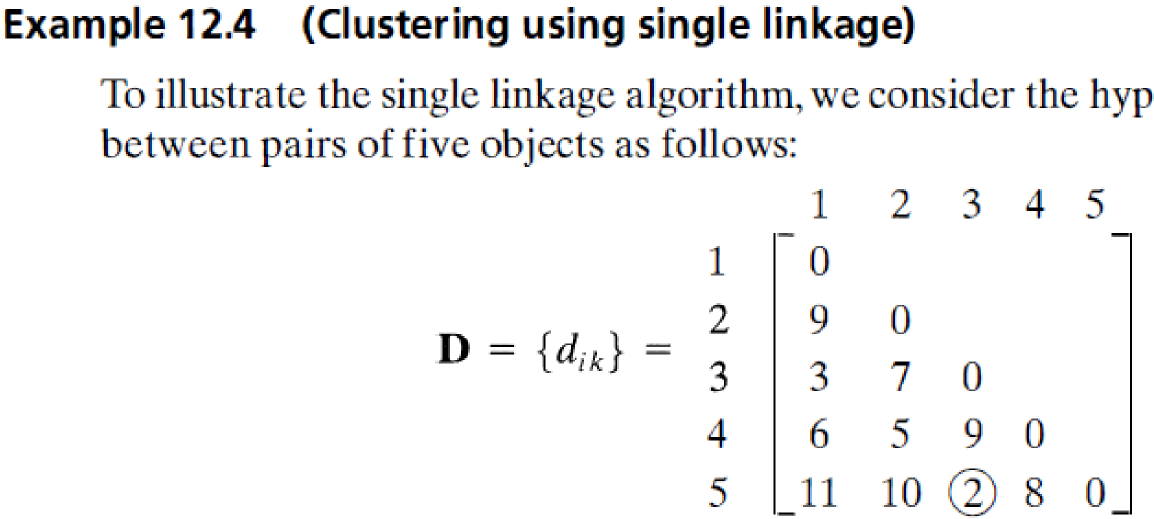


The following are the steps in the agglomerative hierarchical clustering algorithm for grouping N objects (items or variables) :

1. Start with N clusters, each containing a single entity and an $N \times N$ symmetric matrix of distances (or similarities) $\mathbf{D} = \{ d_{ik} \}$.
2. Search the distance matrix for the nearest (most similar) pair of clusters. Let the distance between "most similar" clusters U and V be d_{UV}.
3. Merge clusters U and V. Label the newly formed cluster (UV) . Update the entries in the distance matrix by (a) deleting the rows and columns corresponding to clusters U and V and (b) adding a row and column giving the distances between cluster (UV) and the remaining clusters.
4. Repeat Steps 2 and 3 a total of $N - 1$ times. (All objects will be in a single cluster after the algorithm terminates.) Record the identity of clusters that are merged and the levels (distances or similarities) at which the mergers take place.

Linkage Methods:

- single linkage (minimum distance or nearest neighbor),
- complete linkage (maximum distance or farthest neighbor), and
- average linkage (average distance)
- Centroid Method (the similarity between two clusters is the distance between cluster centroid). $d(\mathbf{x},\mathbf{y}) = \sqrt{(\bar{\mathbf{x}} - \bar{\mathbf{y}})^2}$
- Ward’s method (the similarity between two clusters is not a single measure similarity, but rather the sum of squares within clusters summed over all variables).



Example 12.4 (Clustering using single linkage)

To illustrate the single linkage algorithm, we consider the hypothetical distances between pairs of five objects as follows:

$$\mathbf{D} = \{d_{ik}\} = \begin{array}{cc} & \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} & \begin{bmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & \textcircled{2} & 8 & 0 \end{bmatrix} \end{array}$$

$$d_{(35)1} = \min\{d_{31}, d_{51}\} = \min\{3, 11\} = 3$$
$$d_{(35)2} = \min\{d_{32}, d_{52}\} = \min\{7, 10\} = 7$$
$$d_{(35)4} = \min\{d_{34}, d_{54}\} = \min\{9, 8\} = 8$$

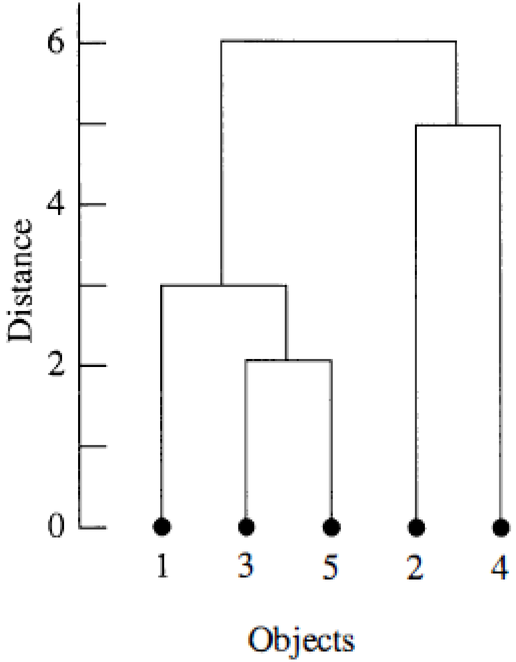
$$d_{(135)2} = \min\{d_{(35)2}, d_{12}\} = \min\{7, 9\} = 7$$
$$d_{(135)4} = \min\{d_{(35)4}, d_{14}\} = \min\{8, 6\} = 6$$

$$\begin{array}{c|ccc} & (135) & 2 & 4 \\ \hline (135) & 0 & & \\ 2 & 7 & 0 & \\ 4 & 6 & \textcircled{5} & 0 \end{array}$$

$$d_{(135)(24)} = \min\{d_{(135)2}, d_{(135)4}\} = \min\{7, 6\} = 6$$

$$\begin{array}{c|cc} & (135) & (24) \\ \hline (135) & 0 & \\ (24) & \textcircled{6} & 0 \end{array}$$

$$\begin{array}{c|cccc} & (35) & 1 & 2 & 4 \\ \hline (35) & 0 & & & \\ 1 & \textcircled{3} & 0 & & \\ 2 & 7 & 9 & 0 & \\ 4 & 8 & 6 & 5 & 0 \end{array}$$



Nonhierarchical Clustering Method

- The number of clusters, K, may either be specified in advance or determined as part of the clustering procedure.
- Nonhierarchical methods can be applied to much larger data sets than can hierarchical techniques.
- Nonhierarchical methods start from either (1) an initial partition of items into groups or (2) an initial set of seed points, which will form the nuclei of clusters
- One of the more popular nonhierarchical procedures, the K-means method.

K-Means Method

1. Partition the items into K initial clusters.
2. Proceed through the list of items, assigning an item to the cluster whose centroid (mean) is nearest. (Distance is usually computed using Euclidean distance with either standardized or unstandardized observations.) Recalculate the centroid for the cluster receiving the new item and for the cluster losing the item.
3. Repeat Step 2 until no more reassignments take place.

Example 12.12 (Clustering using the *K*-means method)

Suppose we measure two variables X_1 and X_2 for each of four items A, B, C, and D. The data are given in the following table:

Item	Observations	
	x_1	x_2
A	5	3
B	−1	1
C	1	−2
D	−3	−2

The objective is to divide these items into $K = 2$ clusters such that the items within a cluster are closer to one another than they are to the items in different clusters. To implement the $K = 2$-means method, we *arbitrarily* partition the items into two clusters, such as (AB) and (CD), and compute the coordinates $(\bar{x}_1, \bar{x}_2)$ of the cluster centroid (mean). Thus, at Step 1, we have

Step 1, we have

Cluster	Coordinates of centroid	
	$\bar{x}_1$	$\bar{x}_2$
(AB)	$\frac{5+(-1)}{2}=2$	$\frac{3+1}{2}=2$
(CD)	$\frac{1+(-3)}{2}=-1$	$\frac{-2+(-2)}{2}=-2$

Step 2, compute Euclidian distance

$$d^2(A,(AB)) = (5-2)^2 + (3-2)^2 = 10$$

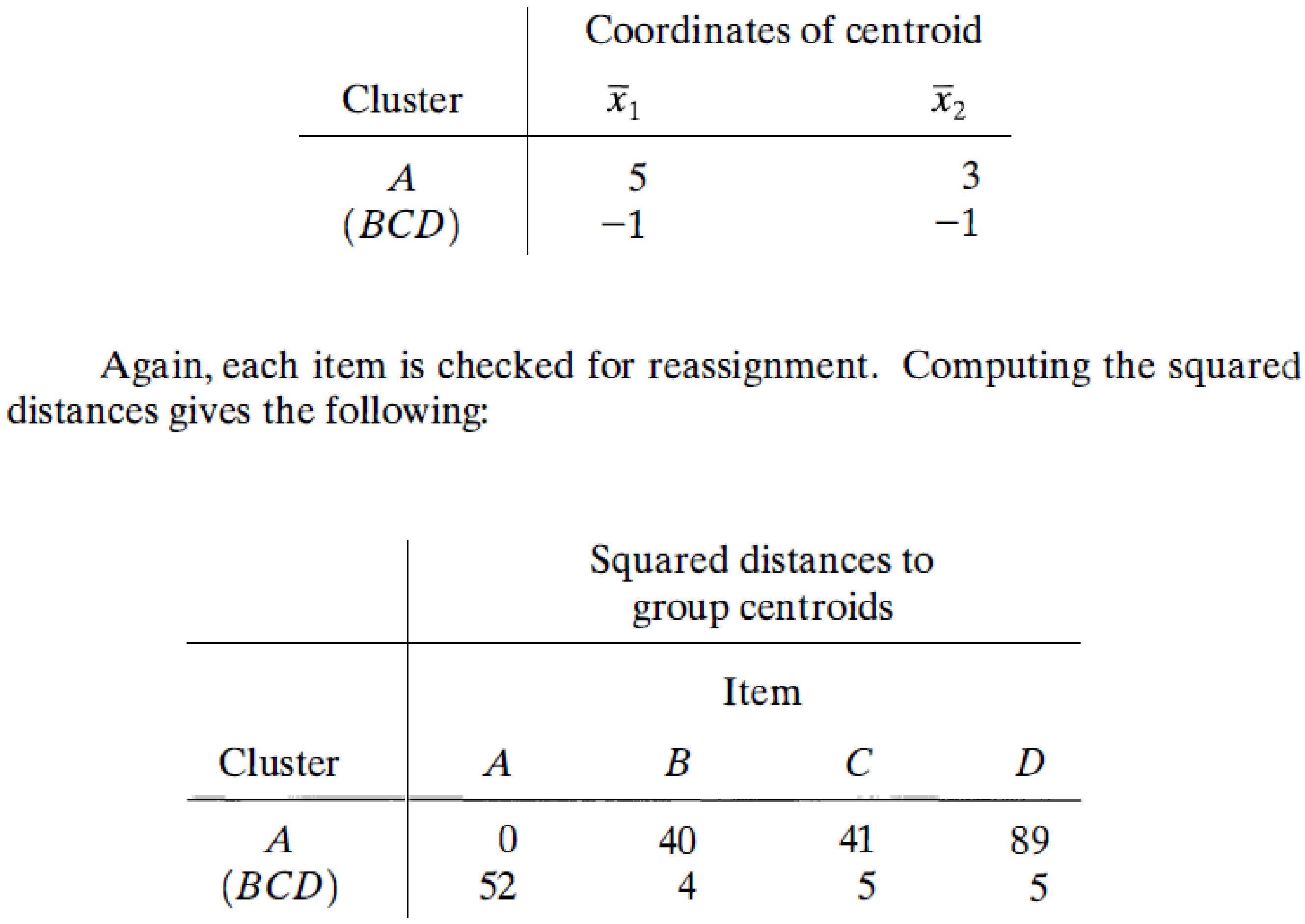
$$d^2(A,(CD)) = (5+1)^2 + (3+2)^2 = 61$$

Since A is closer to cluster (AB) than to cluster (CD), it is not reassigned. Continuing, we get

$$d^2(B,(AB)) = (-1-2)^2 + (1-2)^2 = 10$$

$$d^2(B,(CD)) = (-1+1)^2 + (1+2)^2 = 9$$

and, consequently, B is reassigned to cluster (CD) , giving cluster (BCD) and the following updated coordinates of the centroid:

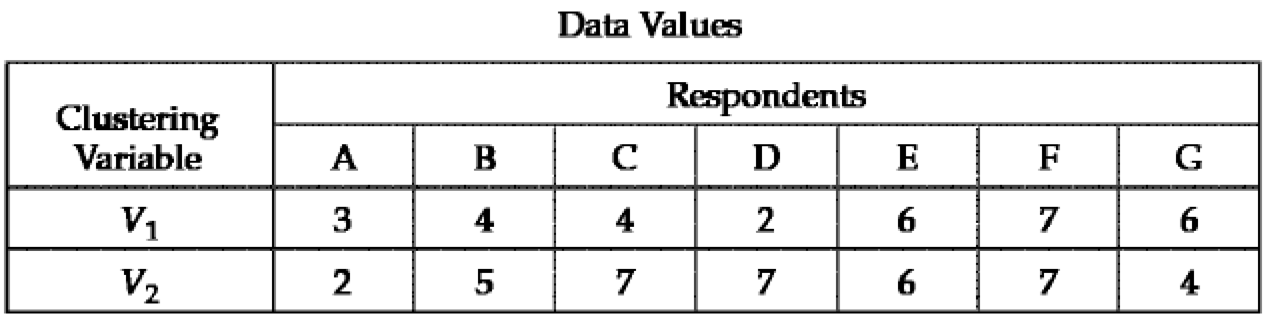


Cluster	Coordinates of centroid	
	$\bar{x}_1$	$\bar{x}_2$
A	5	3
(BCD)	−1	−1

Again, each item is checked for reassignment. Computing the squared distances gives the following:

	Squared distances to group centroids			
	Item			
Cluster	A	B	C	D
A	0	40	41	89
(BCD)	52	4	5	5

Illustration (simple example)



Data Values

Clustering Variable	Respondents						
	A	B	C	D	E	F	G
V_1	3	4	4	2	6	7	6
V_2	2	5	7	7	6	7	4

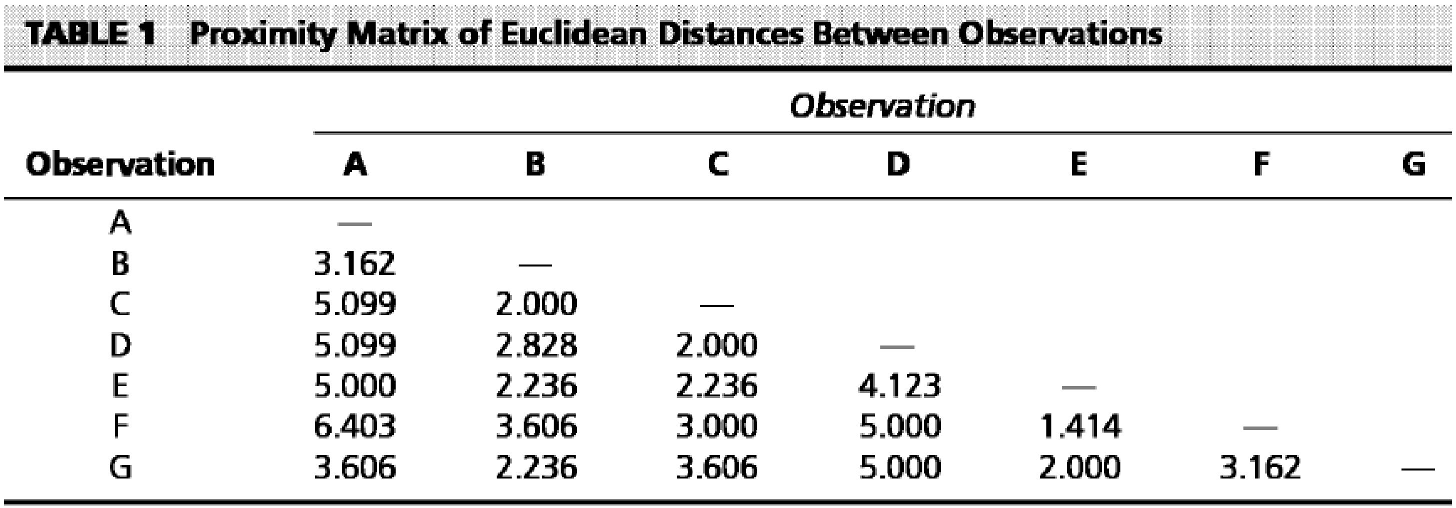


TABLE 1 Proximity Matrix of Euclidean Distances Between Observations

	Observation						
Observation	**A**	**B**	**C**	**D**	**E**	**F**	**G**
A	—						
B	3.162	—					
C	5.099	2.000	—				
D	5.099	2.828	2.000	—			
E	5.000	2.236	2.236	4.123	—		
F	6.403	3.606	3.000	5.000	1.414	—	
G	3.606	2.236	3.606	5.000	2.000	3.162	—

Forming Cluster

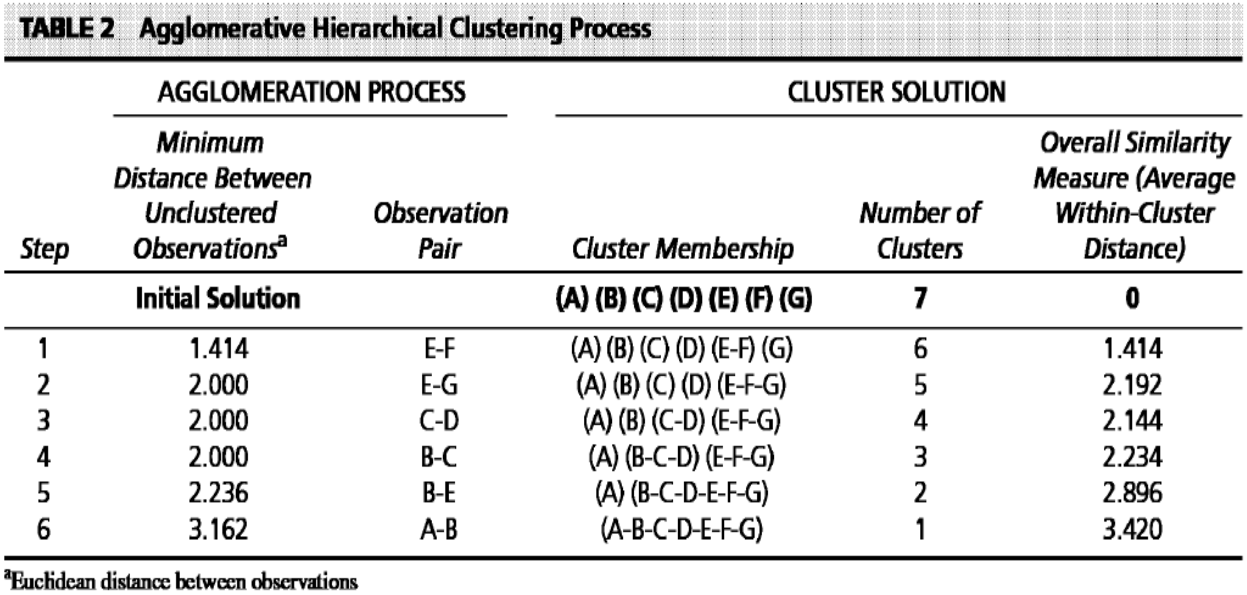
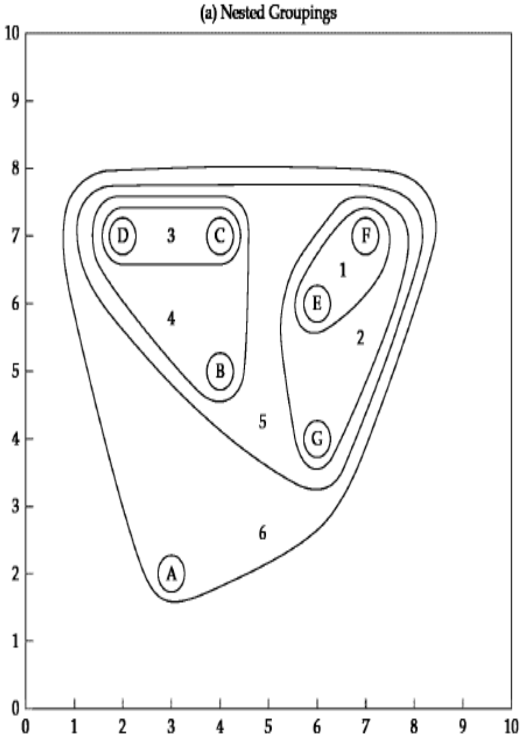


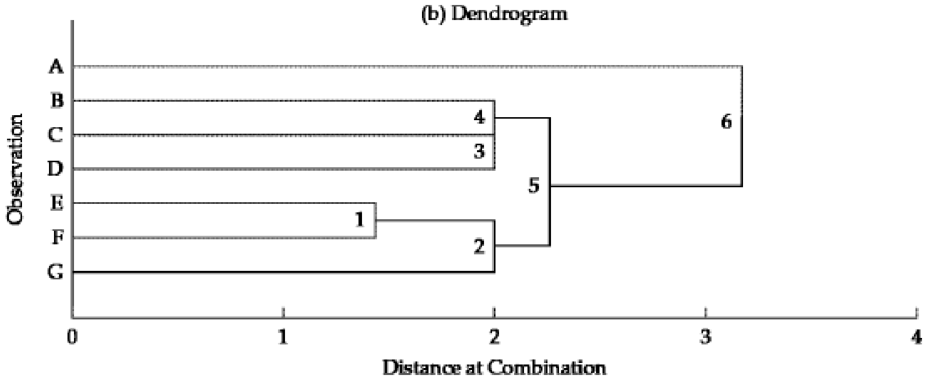
TABLE 2 Agglomerative Hierarchical Clustering Process

	AGGLOMERATION PROCESS		CLUSTER SOLUTION		
Step	*Minimum Distance Between Unclustered Observations*[a]	*Observation Pair*	*Cluster Membership*	*Number of Clusters*	*Overall Similarity Measure (Average Within-Cluster Distance)*
	Initial Solution		**(A) (B) (C) (D) (E) (F) (G)**	**7**	**0**
1	1.414	E-F	(A) (B) (C) (D) (E-F) (G)	6	1.414
2	2.000	E-G	(A) (B) (C) (D) (E-F-G)	5	2.192
3	2.000	C-D	(A) (B) (C-D) (E-F-G)	4	2.144
4	2.000	B-C	(A) (B-C-D) (E-F-G)	3	2.234
5	2.236	B-E	(A) (B-C-D-E-F-G)	2	2.896
6	3.162	A-B	(A-B-C-D-E-F-G)	1	3.420

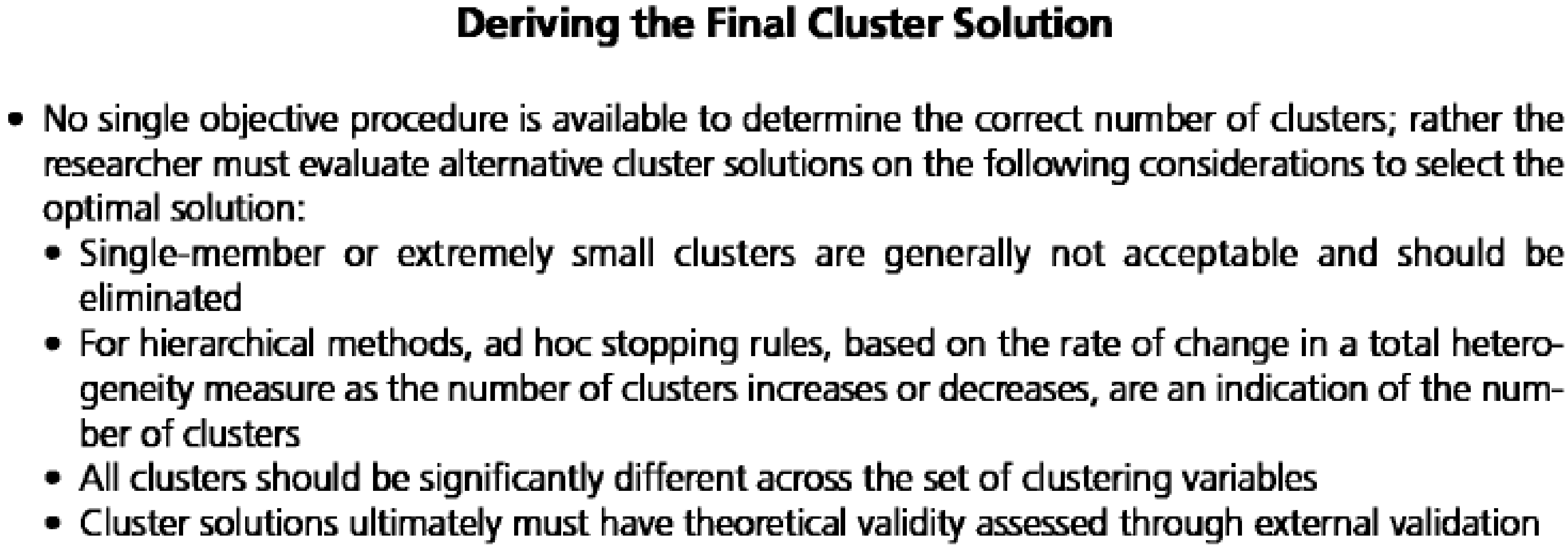
[a] Euclidean distance between observations



(a) Nested Groupings



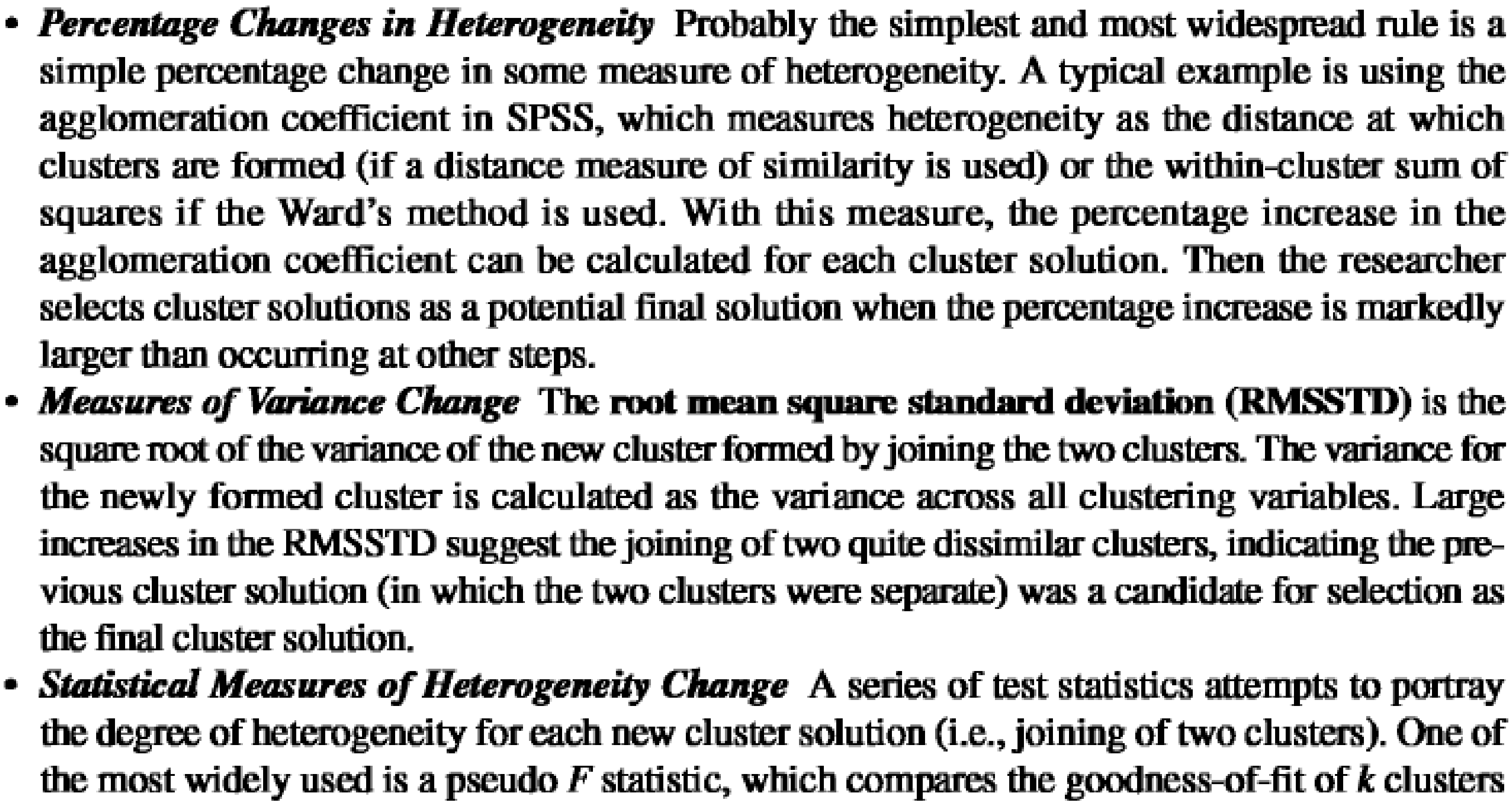
(b) Dendrogram



Deriving the Final Cluster Solution

- No single objective procedure is available to determine the correct number of clusters; rather the researcher must evaluate alternative cluster solutions on the following considerations to select the optimal solution:
  - Single-member or extremely small clusters are generally not acceptable and should be eliminated
  - For hierarchical methods, ad hoc stopping rules, based on the rate of change in a total heterogeneity measure as the number of clusters increases or decreases, are an indication of the number of clusters
  - All clusters should be significantly different across the set of clustering variables
  - Cluster solutions ultimately must have theoretical validity assessed through external validation

Measuring Heterogeneity



- ***Percentage Changes in Heterogeneity*** Probably the simplest and most widespread rule is a simple percentage change in some measure of heterogeneity. A typical example is using the agglomeration coefficient in SPSS, which measures heterogeneity as the distance at which clusters are formed (if a distance measure of similarity is used) or the within-cluster sum of squares if the Ward's method is used. With this measure, the percentage increase in the agglomeration coefficient can be calculated for each cluster solution. Then the researcher selects cluster solutions as a potential final solution when the percentage increase is markedly larger than occurring at other steps.
- ***Measures of Variance Change*** The **root mean square standard deviation (RMSSTD)** is the square root of the variance of the new cluster formed by joining the two clusters. The variance for the newly formed cluster is calculated as the variance across all clustering variables. Large increases in the RMSSTD suggest the joining of two quite dissimilar clusters, indicating the previous cluster solution (in which the two clusters were separate) was a candidate for selection as the final cluster solution.
- ***Statistical Measures of Heterogeneity Change*** A series of test statistics attempts to portray the degree of heterogeneity for each new cluster solution (i.e., joining of two clusters). One of the most widely used is a pseudo *F* statistic, which compares the goodness-of-fit of *k* clusters

Evaluating cluster size (Application on SPSS)

How many cluster should we have? The basic rationale is that when large increases in heterogeneity occur in moving from one stage to the next, the researcher selects the prior cluster solution

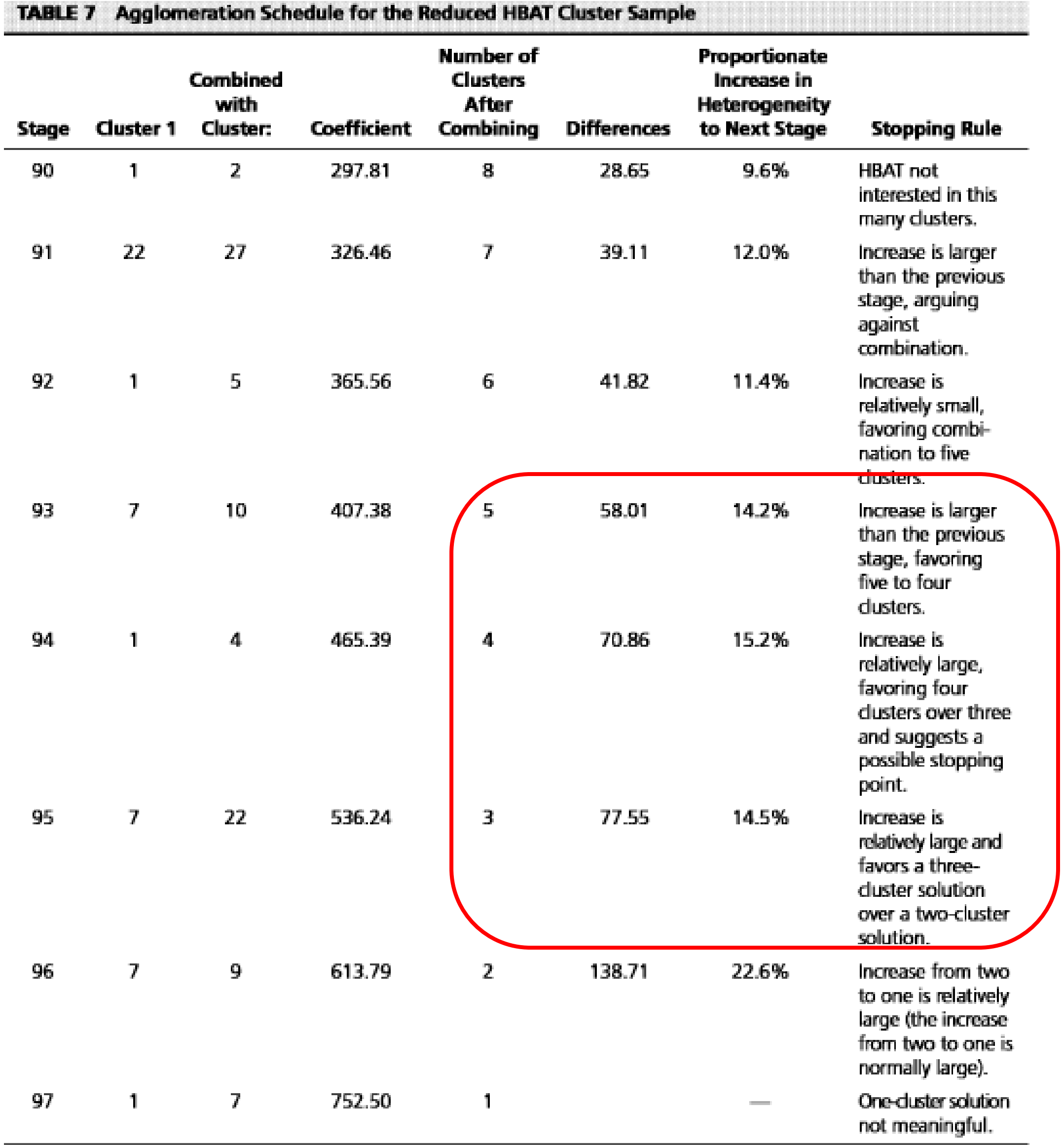
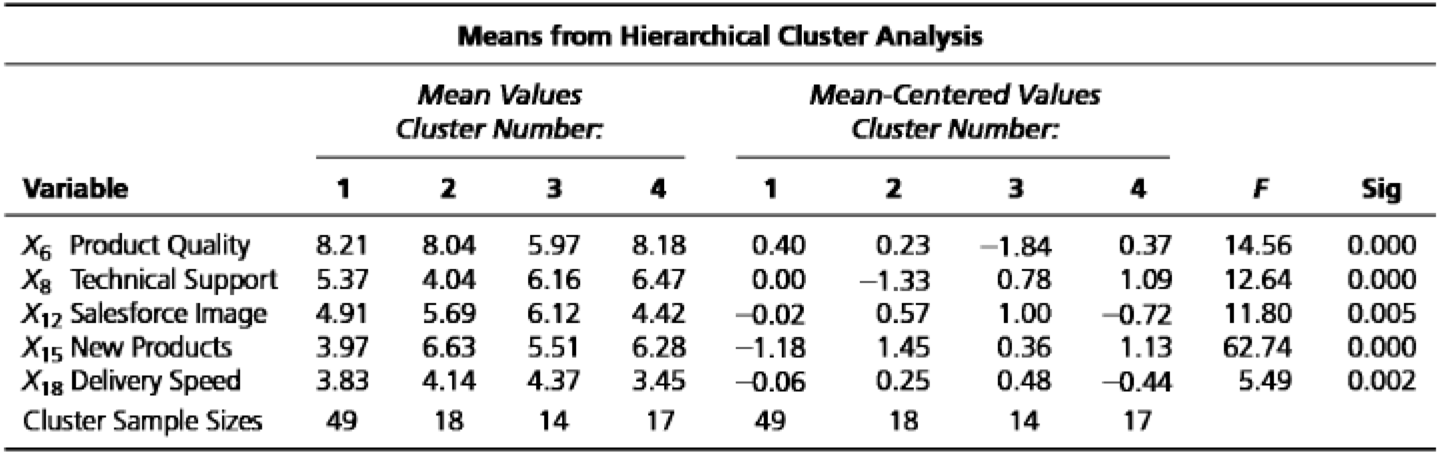


TABLE 7 Agglomeration Schedule for the Reduced HBAT Cluster Sample

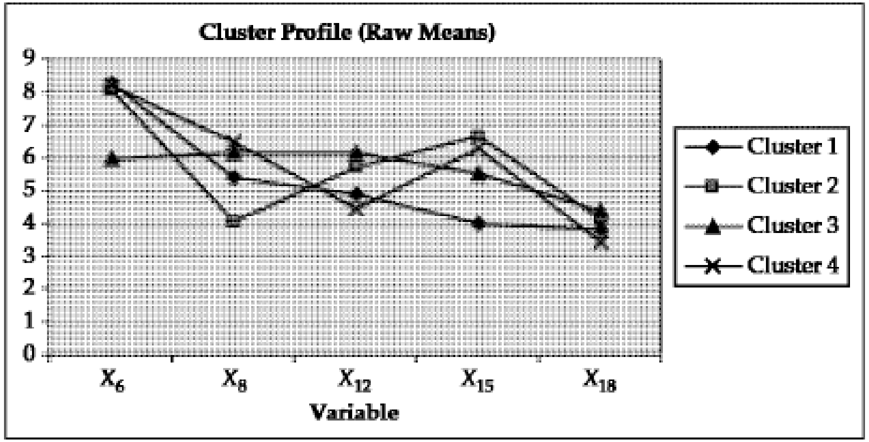
Stage	Cluster 1	Combined with Cluster:	Coefficient	Number of Clusters After Combining	Differences	Proportionate Increase in Heterogeneity to Next Stage	Stopping Rule
90	1	2	297.81	8	28.65	9.6%	HBAT not interested in this many clusters.
91	22	27	326.46	7	39.11	12.0%	Increase is larger than the previous stage, arguing against combination.
92	1	5	365.56	6	41.82	11.4%	Increase is relatively small, favoring combination to five clusters.
93	7	10	407.38	5	58.01	14.2%	Increase is larger than the previous stage, favoring five to four clusters.
94	1	4	465.39	4	70.86	15.2%	Increase is relatively large, favoring four clusters over three and suggests a possible stopping point.
95	7	22	536.24	3	77.55	14.5%	Increase is relatively large and favors a three-cluster solution over a two-cluster solution.
96	7	9	613.79	2	138.71	22.6%	Increase from two to one is relatively large (the increase from two to one is normally large).
97	1	7	752.50	1		—	One-cluster solution not meaningful.

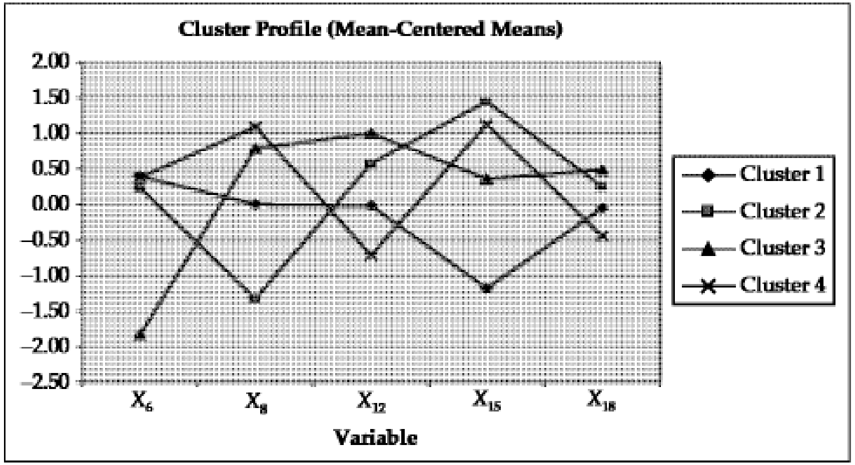
Four-cluster solution



Means from Hierarchical Cluster Analysis

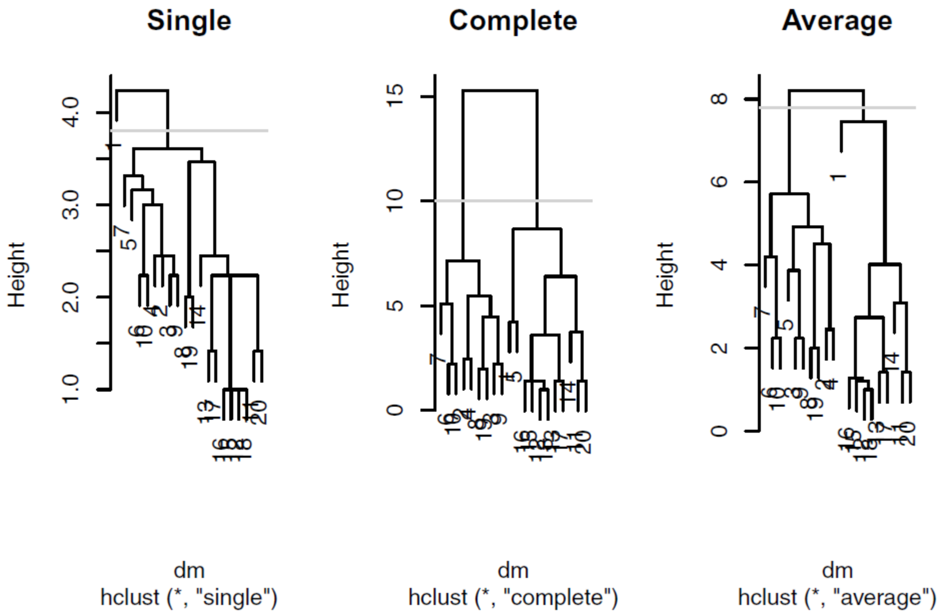
Variable	*Mean Values Cluster Number:* 1	2	3	4	*Mean-Centered Values Cluster Number:* 1	2	3	4	*F*	Sig
X_6 Product Quality	8.21	8.04	5.97	8.18	0.40	0.23	−1.84	0.37	14.56	0.000
X_8 Technical Support	5.37	4.04	6.16	6.47	0.00	−1.33	0.78	1.09	12.64	0.000
X_{12} Salesforce Image	4.91	5.69	6.12	4.42	−0.02	0.57	1.00	−0.72	11.80	0.005
X_{15} New Products	3.97	6.63	5.51	6.28	−1.18	1.45	0.36	1.13	62.74	0.000
X_{18} Delivery Speed	3.83	4.14	4.37	3.45	−0.06	0.25	0.48	−0.44	5.49	0.002
Cluster Sample Sizes	49	18	14	17	49	18	14	17		



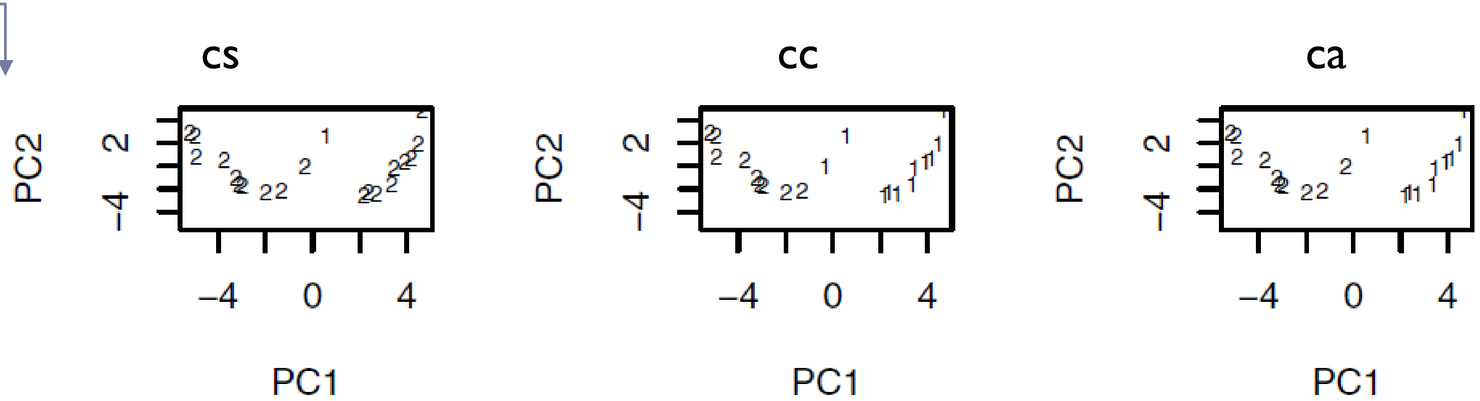


Hierarchical clustering in R

```
dm <- dist(measure[, c("chest", "waist", "hips")])
plot(cs <- hclust(dm, method = "single"))
plot(cc <- hclust(dm, method = "complete"))
plot(ca <- hclust(dm, method = "average"))
```



```
body_pc <- princomp(dm, cor = TRUE)
xlim <- range(body_pc$scores[,1])
plot(body_pc$scores[,1:2], type = "n", xlim = xlim, ylim = xlim)
lab <- cutree(cs, h = 3.8) # for single linkage
text(body_pc$scores[,1:2], labels = lab, cex = 0.6)
```



cs

cc

ca

single linkage solutions often contain long “straggly" clusters that do not give a useful description of the data (Everitt & Hothorn, 2011)

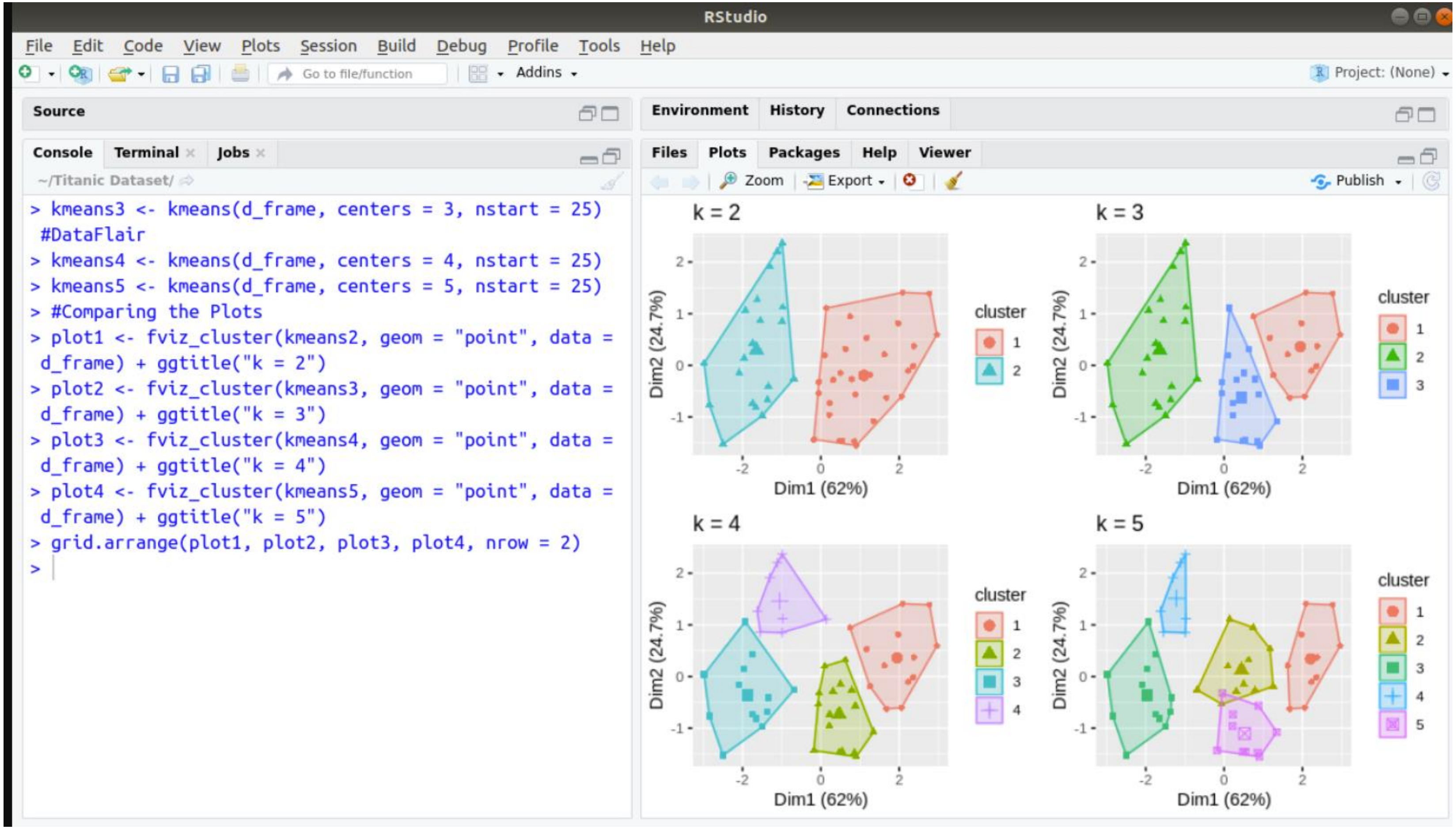
K-Means Clustering Method

```
library(tidyverse)
library(cluster)
library(factoextra)
library(gridExtra)
data('USArrests')
d_frame <- USArrests
d_frame <- na.omit(d_frame) #Removing the missing values
d_frame <- scale(d_frame)   # standardizing
head(d_frame) # show the data

kmeans2 <- kmeans(d_frame, centers = 2, nstart = 25)
str(kmeans2) # to see the variables that are used
fviz_cluster(kmeans2, data = d_frame) #make cluster plot
```

```r
kmeans3 <- kmeans(d_frame, centers = 3, nstart = 25) #DataFlair
kmeans4 <- kmeans(d_frame, centers = 4, nstart = 25)
kmeans5 <- kmeans(d_frame, centers = 5, nstart = 25)

#Comparing the Plots
plot1 <- fviz_cluster(kmeans2, geom = "point", data = d_frame) +
ggtitle("k = 2")
plot2 <- fviz_cluster(kmeans3, geom = "point", data = d_frame) +
ggtitle("k = 3")
plot3 <- fviz_cluster(kmeans4, geom = "point", data = d_frame) +
ggtitle("k = 4")
plot4 <- fviz_cluster(kmeans5, geom = "point", data = d_frame) +
ggtitle("k = 5")
grid.arrange(plot1, plot2, plot3, plot4, nrow = 2)
```



```
> kmeans3 <- kmeans(d_frame, centers = 3, nstart = 25)
 #DataFlair
> kmeans4 <- kmeans(d_frame, centers = 4, nstart = 25)
> kmeans5 <- kmeans(d_frame, centers = 5, nstart = 25)
> #Comparing the Plots
> plot1 <- fviz_cluster(kmeans2, geom = "point", data =
 d_frame) + ggtitle("k = 2")
> plot2 <- fviz_cluster(kmeans3, geom = "point", data =
 d_frame) + ggtitle("k = 3")
> plot3 <- fviz_cluster(kmeans4, geom = "point", data =
 d_frame) + ggtitle("k = 4")
> plot4 <- fviz_cluster(kmeans5, geom = "point", data =
 d_frame) + ggtitle("k = 5")
> grid.arrange(plot1, plot2, plot3, plot4, nrow = 2)
>
```

Model based clustering

Finite mixture Gaussian model:

- Assumption: $\boldsymbol{X}$ ~ multivariate normal

```
library(mclust)
mc <- Mclust(X)
```

best model: ellipsoidal, equal shape with 3 components

Other Examples

- SPSS: https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/cluster-analysis/
- R: https://www.statmethods.net/advstats/cluster.html

Exercises

Do exercises at Johnson & Wichern (2002):
12.8; 12.10; 12.12 & 12.16