CARLETON UNIVERSITY

SCHOOL OF
MATHEMATICS AND STATISTICS

HONOURS PROJECT

TITLE: Selected Methods of Cluster Analysis

AUTHOR: Qiang Xu

SUPERVISOR: Dr. Natalia Stepanova

DATE: May 3rd, 2021

Contents

Abstract

The goal of this honours project is to explore the basics of cluster analysis from Applied Multivariate Statistical Analysis (6th Edition) by Richard A. Johnson and Dean W. Wichern [7], illustrate the use of various cluster procedures by solving selected exercises from Chapter 12, and hone the skills of writing technical papers using LaTeX 2_ε.

1 Introduction

> An intelligent being cannot treat every object it sees as a unique entity unlike anything else in the universe. It has to put objects in categories so that it may apply its hard-won knowledge about similar objects encountered in the past, to the object at hand.
>
> Steven Pinker, How the Mind Works [10]

One of basic abilities of human being is grouping of similar objects to produce classification; the idea of sorting similar things into categories is essential to human being's mind. For example, early man must have been able to realize that many individual objects had certain properties such as being edible, or poisonous, etc. Classification is also fundamental to most branches of scientific pursues. For instance, the classification of the elements in the periodic table produced by D. I. Mendeleyev in the 1860s has helped to understand the structure of the atom profoundly. A classification represents a convenient method for organizing a large data set so that it can be understood more easily and information can be retrieved more efficiently. If the data can be summarized by a small number of groups of objects, then the group labels may provide a very concise description of patterns of similarities and differences in the data. Because of the growing number of large databases, the need to summarize data sets in this way and extract useful information or knowledge is increasingly important.

1.1 Definitions

Cluster, group, and class have been used in the statistical literature interchangeably without formal definition. Many, for example, Gordon [3], attempt to define what is a cluster in terms of homogeneity (internal cohesion) and external isolation (separation); but others, for example, Bonner [1], suggest that the criterion for evaluating the meaning of cluster is the value judgment of the researcher, that is, if the term of cluster produces an answer of value to the researcher, that is all that matters. This may explain why so many attempts to differentiate the concepts of homogeneity and separation mathematically proliferate numerous and diverse criteria, sometimes leading to confusion in practice. The attempt to define the term of cluster mathematically precise is beyond the scope of the project, so we will leave it as it be.

Cluster analysis concerns about searching for patterns in a data set by grouping the multivariate observations into clusters in order to find an optimal grouping by which

the observations or objects within each cluster are similar but the clusters are dissimilar to each other. The goal of cluster analysis is to find the "natural groupings" in the data set that make sense to the researcher.

Cluster analysis is also referred to as classification, pattern recognition (specifically, unsupervised learning), and numerical taxonomy. Classification and clustering are often used interchangeably in the literature, but we wish to differentiate them for the project. In classification, we assign new observations to one of several groupings, the number of which is predefined. In cluster analysis, neither the number of groups nor the groups themselves are unknown in advance.

The basic data for most applications of cluster analysis is the usual $n \times p$ multivariate data matrix containing the variable values describing each object to be clustered. The data matrix can be defined as

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}_1^\top \\ \mathbf{y}_2^\top \\ \vdots \\ \mathbf{y}_n^\top \end{pmatrix} = \left(\mathbf{y}_{(1)}, \mathbf{y}_{(2)}, \cdots \mathbf{y}_{(p)}\right), \tag{1}$$

where $\mathbf{y}_i^\top$ is a row observation vector and $\mathbf{y}_{(j)}$ is a column corresponding to a variable. We usually wish to group $n\mathbf{y}_i^\top$'s rows into g clusters, but may also cluster the columns $\mathbf{y}_{(j)}$, $j = 1, 2, \ldots, p$.

Many techniques employed in cluster analysis begin with similarities between all pairs of observations. It means that the data points required to be measured similarly or at least from which similarities can be computed. In practice, there exist time constraints that make it almost impossible to determine the best grouping of similar objects from a list of all possible structure. Thus, we should settle for algorithms searching for good, but not necessarily the best, groupings.

1.2 Examples of the use of cluster analysis

Cluster analysis appears in many disciplines such as biology, botany, medicine, psychology, geography, marketing, image processing, psychiatry, archaeology, etc. We briefly discuss several applications of cluster analysis.

1.2.1 Astronomy

Large multivariate astronomical data frequently contain relatively distinct groups of objects which can be distinguished from each other. Astronomers want to know how

many groups of stars there exist on the some sorts of statistical criterion. Cluster analysis can be used to classify astronomical objects and let astronomers find unusual objects. For example, discoveries of high-redshift quasars, type 2 quasars, and brown dwarfs. The Hertzsprung-Russell plot (see Figure 1: Hertzsprung-Russell-Diagram, Montmerle [9]) of temperature vs. luminosity helps to classify stars into dwarf stars, giant stars, and main sequence stars that develops theories of stellar evolution and forecasts the life expectancy of the Sun.

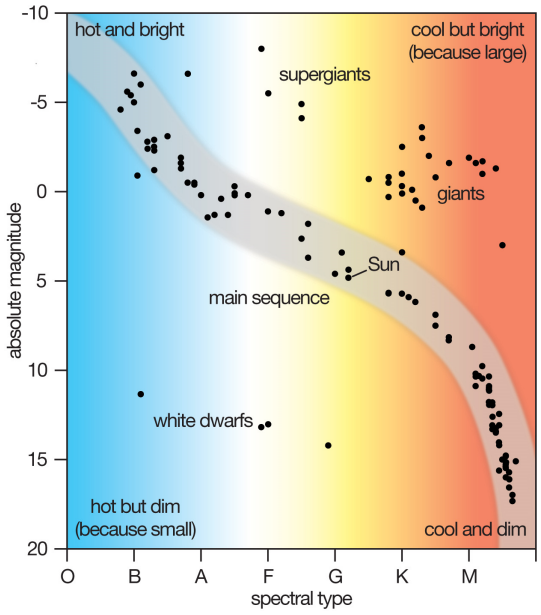


Figure 1: Hertzsprung-Russell Diagram

1.2.2 Archaeology

In archaeology, the classification of artifacts help to find their different uses, the periods they were used and who used by. And the study of fossilized material can help to understand how prehistoric societies lived. For example, Hodson et al. [6] applied single linkage and average linkage clustering to brooches from the Iron Age and found classifications of archaeological significance; Hodson [5] used a *k-means* clustering technique to classify hand axes found in the British Isles, and variables used to describe the axes included length, breadth and pointedness at the tip, which resulted in two clusters: thin, small axes and thick, large axes used for different purposes.

1.2.3 Market research

The basic strategies of marketing is to divide customers into homogeneous groups. For example, Green et al. [4] employed cluster analysis to classify the cities into a small numbers of groups on the basis of 14 variables: city size, newspaper circulation, per capita income, etc. Cities within a group were expected to be very similar to each other, so choosing one city from each group was a means of sampling the test markets. Also Chakrapani [2] described that some survey suggests that buying a sports car is not solely based on one's means or one's age but a lifestyle decision, and so car manufacturer employs cluster analysis to identify people with a lifestyle most associated with buying sports cars, to propose a focused marketing campaign.

2 Similarity Measures

Any attempt to identify clusters of observations from a complex data set requires knowledge on how 'close' items are to each other, i.e., a measure of "closeness" or "similarity." The nature of the variables, such as discrete, continuous, binary, scales of measurement, such as nominal, ordinal, interval, ratio, and subject matter knowledge have been suggested as considerations. Clustering items (units or cases) employs the measure of proximity by some sort of distance; grouping variables utilizes the measure of proximity by correlation coefficients or like measures of association.

2.1 Distances and Similarity Coefficients for Pairs of Items

The Euclidean (straight-line) distance between two p-dimensional observations (items) $\mathbf{x}^\top = (x_1, x_2, \ldots, x_p)$ and $\mathbf{y}^\top = (y_1, y_2, \ldots, y_p)$ is given as

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

The statistical distance between the same observations is of the form

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^\top \mathbf{A} (\mathbf{x} - \mathbf{y})}, \quad (3)$$

where $\mathbf{A} = \mathbf{S}^{-1}$ and $\mathbf{S}$ the matrix of sample variances and covariances. However, clustering analysis is about to find the distinct groups in the data set, so these sample quantities cannot be computed. Thus, we would prefer Euclidean distance for clustering.

The third distance measure is the Minkowski metric

$$d(\mathbf{x},\mathbf{y}) = \left[\sum_{i=1}^{p} |x_i - y_i|^m\right]^{1/m}. \tag{4}$$

Two additional popular measures of "distance" are Canberra metric and the Czekanowski coefficient which are both defined for nonnegative variables. We have

Canberra metric:

$$d(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{p} \frac{|x_i - y_i|}{(x_i + y_i)}. \tag{5}$$

Czekanowski coefficient:

$$d(\mathbf{x},\mathbf{y}) = 1 - \frac{2\sum_{i=1}^{p} \min(x_i, y_i)}{\sum_{i=1}^{p} (x_i + y_i)}. \tag{6}$$

Manhattan distance:

$$d(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{p} |x_i - y_i|. \tag{7}$$

It is always advisable to use "true" distances for clustering objects. However, most clustering algorithms do accept subjectively assigned distance numbers that may not satisfy the triangle inequality.

In situation where items cannot be accessed by meaningful p-dimensional measurements, we would compare them on the basis of the presence or absence of certain characteristics. The presence or absence of a characteristic can be represented mathematically by a *binary variable*, which assumes the value of 1 if the characteristic is present and the value of 0 if absent. For example, for $p = 5$ binary variables, the "scores" of two items i and k might be arranged as in Table 1.

Table 1: Example

	Variables				
	1	2	3	4	5
Item i	1	0	0	1	1
Item k	1	1	0	1	0

In this table, we have two $1-1$ matches, one $1-0$ mismatch, one $0-1$ mismatch, and one $0-0$ matches.

Let x_{ij} be the score (1 or 0) of the jth binary variable on the ith item and x_{kj} be

the score of the jth binary variable on the kth item, $j = 1, 2, \ldots, p$. Therefore, if we let

$$(x_{ij} - x_{kj})^2 = \begin{cases} 0 & \text{if} \quad x_{ij} = x_{kj} = 1 \quad \text{or} \quad x_{ij} = x_{kj} = 0, \\ 1 & \text{if} \quad x_{ij} \neq x_{kj}, \end{cases} \tag{8}$$

then the squared distance, $\sum_{j=1}^{p} (x_{ij} - x_{kj})^2$, counts the number of mismatches. A large distance corresponds to many mismatches, or dissimilar items. The above example gives us the following distance:

$$\sum_{j=1}^{5} (x_{ij} - x_{kj})^2 = (1-1)^2 + (0-1)^2 + (0-0)^2 + (1-1)^2 + (1-0)^2 = 2.$$

The result of 2 indicates that there are two mismatches in the example.

A distance based on (7) weights the $1-1$ and $0-0$ matches equally. In some cases, a $1-1$ match indicates a stronger similarity than a $0-0$ match. For instance, in grouping people, two persons both read Chinese is stronger support to similarity than the absence of this ability. To differential treatment of the $1-1$ matches and the $0-0$ matches, mathematicians have suggested several schemes for defining similarity coefficients.

First, we need to arrange the frequencies of matches and mismatches for items i and k in the form of a contingency table (see Table 2). Table 3 lists common similarity coefficients defined in terms of the frequencies in the Table 2. Coefficients 1, 2, and 3 in Table 3 are related monotonically. For instance, coefficient 1 is calculated for two contingency tables, Table I and Table II; then if $(a_{\mathbf{I}} + d_{\mathbf{I}})/p \geqslant (a_{\mathbf{II}} + d_{\mathbf{II}})/p$, we also have

$$2(a_{\mathbf{I}} + d_{\mathbf{I}}) / [2(a_{\mathbf{I}} + d_{\mathbf{I}}) + b_{\mathbf{I}} + c_{\mathbf{I}}] \geqslant 2(a_{\mathbf{II}} + d_{\mathbf{II}}) / [2(a_{\mathbf{II}} + d_{\mathbf{II}}) + b_{\mathbf{II}} + c_{\mathbf{II}}],$$

and coefficient 3 will be at least as large for Table I as it is for Table II. And coefficient 5, 6, and 7 also retain their relative orders.

Table 2: Contingency Table

		Item k		
		1	0	Totals
Item i	1	a	b	$a+b$
	0	c	d	$c+d$
Totals		$a+c$	$b+d$	$p=a+b+c+d$

Monotonicity (or maintaining relative orders) is important because certain clustering

algorithms are not affected if the definition of similarity is changed in a fashion that retains the relative orderings of similarities. For instance, the single linkage and complete linkage hierarchical procedures discussed below are not affected. Hence, any choice of the coefficients 1, 2, and 3 in Table 3 will produce the same groupings. Similarly, any choice of the coefficients 5, 6, and 7 will produce identical clusters.

Table 3: Similarity Coefficients for Clustering Items*

Coefficient	Rationale
1. $\frac{a+d}{p}$	Equal weights for $1-1$ matches and $0-0$ matches.
2. $\frac{2(a+d)}{2(a+d)+b+c}$	Double weight for $1-1$matches and $0-0$ matches.
3. $\frac{a+d}{a+d+2(b+c)}$	Double weight for unmatched pairs.
4. $\frac{a}{p}$	No $0-0$ matches in numerator.
5. $\frac{a}{a+b+c}$	No $0-0$ matches in numerator or denominator. (The $0-0$ matches are treated as irrelevant.)
6. $\frac{2a}{2a+b+c}$	No $0-0$ matches in numerator or denominator. Double weight for $1-1$ matches.
7. $\frac{a}{a+2(b+c)}$	No $0-0$ matches in numerator or denominator. Double weight for unmatched pairs.
8. $\frac{a}{b+c}$	Ratio of matches to mismatches with $0-0$ matches excluded.

*[p binary variables; see Table 2.]

2.2 Examples

Example 1 (Example 12.1 from Johnson and Wichern [7]: Calculating the values of a similarity coefficient). Suppose five individuals posses the following characteristics:

Table 4: Characteristics of Five Individuals

	Height	Weight	Eye color	Hair color	Handedness	Gender
Individual 1	68 in	140 lb	green	blond	right	female
Individual 2	73 in	185 lb	brown	brown	right	male
Individual 3	67 in	165 lb	blue	blond	right	male
Individual 4	64 in	120 lb	brown	brown	right	female
Individual 5	76 in	210 lb	brown	brown	left	male

Define six binary variables $X_1, X_2, X_3, X_4, X_5, X_6$ as follows:

$$X_1 = \begin{cases} 1 & \text{height} \geqslant 72 \text{ in.} \\ 0 & \text{height} < 72 \text{ in.} \end{cases} \qquad X_4 = \begin{cases} 1 & \text{blond hair} \\ 0 & \text{not blond hair} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{weight} \geqslant 150 \text{ lb} \\ 0 & \text{weight} < 150 \text{ lb} \end{cases} \qquad X_5 = \begin{cases} 1 & \text{right handed} \\ 0 & \text{left handed} \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{brown eyes} \\ 0 & \text{otherwise} \end{cases} \qquad X_6 = \begin{cases} 1 & \text{female} \\ 0 & \text{male} \end{cases}$$

The scores for individuals 1 and 2 on the $p = 6$ binary variables are as in Table 5,

Table 5: Binary Variables

		X_1	X_2	X_3	X_4	X_5	X_6
Individual	1	0	0	0	1	1	1
	2	1	1	1	0	1	0

and the number of matches and mismatches are indicated in the two-way array in Table 6.

Table 6: Two-way Array of Individual 1 and 2

	Individual 2		
Individual 1	1	0	Total
1	1	2	3
0	3	0	3
Total	4	2	6

Using similarity coefficient 1, which gives equal weight to matches, we have

$$\frac{a+d}{p} = \frac{1+0}{6} = \frac{1}{6}.$$

Under equal weight scheme, $1-1$ match occurs once every six times.

Example 2 (Exercise 12.1 from Johnson and Wichern [7]). Certain characteristics associated with a few recent U.S. presidents are listed in Table 7.

(a) Introducing appropriate binary variables, calculate similarity coefficient 1 in Table 7 for pairs of presidents.

(b) Proceeding as in part (a), calculate similarity coefficients 2 and 3 in Table 3. Verify the monotonicity relation of coefficient 1, 2, and 3 by displaying the order of the 15 similarities for each coefficient.

Table 7: Information Concerning Six Presidents of USA

President	Birthplace (region of United States	Elected first term?	Party	Prior U.S. congressional experience?	Served as vice president?
1. R. Reagan	Midwest	Yes	Republican	No	No
2. J. Carter	South	Yes	Democrat	No	No
3. G. Ford	Midwest	No	Republican	Yes	Yes
4. R. Nixon	West	Yes	Republican	Yes	Yes
5. L. Johnson	South	No	Democrat	Yes	Yes
6. J. Kennedy	East	Yes	Democrat	Yes	No

Solution: (a) Introduce five binary variables X_1, X_2, X_3, X_4 and X_5 as follows:

$$X_1 = \begin{cases} 1 & \text{South} \\ 0 & \text{Non-south} \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{Elected first term} \\ 0 & \text{Not Elected first term} \end{cases}$$

$$X_3 = \begin{cases} 1 & \text{Republican party} \\ 0 & \text{Democrat party} \end{cases}$$

$$X_4 = \begin{cases} 1 & \text{Prior U.S. congressional experience} \\ 0 & \text{Not Prior U.S. congressional experience} \end{cases}$$

$$X_5 = \begin{cases} 1 & \text{Served as vice president} \\ 0 & \text{Not Served as vice president} \end{cases}$$

For example, the score for Presidents 1 and 2 on the $p = 5$ binary variables is as in the following table:

Table 8: Binary Variables of President 1 and 2

	X_1	X_2	X_3	X_4	X_5
President 1	0	1	1	0	0
President 2	1	1	0	0	0

The contingency table of matches and mismatches is shown in Table 9.

Table 9: Two-way Array of Matches and Mismatches

		President 2		
		1	0	Total
President 1	1	1	1	2
	0	1	2	3
	Total	2	3	5

Coefficient 1 from Table 3 of the pair is calculated as

$$\frac{a+d}{p} = \frac{1+2}{5} = \frac{3}{5}.$$

Similarly, we can calculate coefficient 1 for the pairs of presidents and display them in a 6×6 symmetric matrix as follows:

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{pmatrix} 1 & \frac{6}{8} & \frac{4}{7} & \frac{6}{8} & 0 & \frac{6}{8} \\ \frac{6}{8} & 1 & 0 & \frac{2}{6} & \frac{4}{7} & \frac{6}{8} \\ \frac{4}{7} & 0 & 1 & \frac{8}{9} & \frac{6}{8} & \frac{4}{7} \\ \frac{6}{8} & \frac{2}{6} & \frac{8}{9} & 1 & \frac{4}{7} & \frac{6}{8} \\ 0 & \frac{4}{7} & \frac{6}{8} & \frac{4}{7} & 1 & \frac{4}{7} \\ \frac{6}{8} & \frac{6}{8} & \frac{4}{7} & \frac{6}{8} & \frac{4}{7} & 1 \end{pmatrix} \end{array}.$$

(b) The scores for Presidents 1 and 2 on the $p = 5$ binary variables are shown in Table 10.

Table 10: Binary Variables of Presidents 1 and 2

		X_1	X_2	X_3	X_4	X_5
President	1	0	1	1	0	0
President	2	1	1	0	0	0

The contingency table of matches and mismatches is as displayed in Table 11.

Table 11: Two-way Arrays of President 1 and 2

		President 2		
		1	0	Total
President 1	1	1	1	2
	0	1	2	3
	Total	2	3	5

We can calculate the similarity coefficient 2 defined in Table 3 as follows:

$$\frac{2(a+d)}{2(a+d)+b+c} = \frac{2(1+2)}{2(1+2)+1+1} = \frac{6}{8},$$

and for coefficient 3 we have

$$\frac{a+d}{a+d+2(b+c)} = \frac{1+2}{1+2+2(1+1)} = \frac{3}{7}.$$

Similarly, we can calculate coefficient 2 for the pairs of presidents and display them in a 6×6 symmetric matrix:

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{pmatrix} 1 & \frac{6}{8} & \frac{4}{7} & \frac{6}{8} & 0 & \frac{6}{8} \\ \frac{6}{8} & 1 & 0 & \frac{2}{6} & \frac{4}{7} & \frac{6}{8} \\ \frac{4}{7} & 0 & 1 & \frac{8}{9} & \frac{6}{8} & \frac{4}{7} \\ \frac{6}{8} & \frac{2}{6} & \frac{8}{9} & 1 & \frac{4}{7} & \frac{6}{8} \\ 0 & \frac{4}{7} & \frac{6}{8} & \frac{4}{7} & 1 & \frac{4}{7} \\ \frac{6}{8} & \frac{6}{8} & \frac{4}{7} & \frac{6}{8} & \frac{4}{7} & 1 \end{pmatrix} \end{array}.$$

Continuing with the similar computation, we can produce Table 12. The result reassures that similarity coefficients 1, 2, and 3 are monotonically related because they also retain their relative orders indicated by their ranking in the table. President R. Reagan and President L. Johnson have least similarities, and so does President J. Carter and President G. Ford; on the other hand, President G. Ford and President R. Nixon have most similarities.

Table 12: The Results of the Example 2

	Similarity Coefficient			**Ranking**		
Pair	1	2	3	1	2	3
1-2	0.6	0.75	0.429	4.5	4.5	4.5
1-3	0.4	0.571	0.25	10	10	10
1-4	0.6	0.75	0.429	4.5	4.5	4.5
1-5	0	0	0	14.5	14.5	14.5
1-6	0.6	0.75	0.429	4.5	4.5	4.5
2-3	0	0	0	14.5	14.5	14.5
2-4	0.2	0.333	0.111	13	13	13
2-5	0.4	0.571	0.25	10	10	10
2-6	0.6	0.75	0.429	4.5	4.5	4.5
3-4	0.8	0.889	0.667	1	1	1
3-5	0.6	0.75	0.429	4.5	4.5	4.5
3-6	0.4	0.571	0.25	10	10	10
4-5	0.4	0.571	0.25	10	10	10
4-6	0.6	0.75	0.429	4.5	4.5	4.5
5-6	0.4	0.571	0.25	10	10	10

Example 3 (Exercise 12.3 from Johnson and Wichern [7]). Show that the sample correlation coefficient given by formula (9) can be written as

$$r = \frac{ad - bc}{[(a+b)(a+c)(b+d)(c+d)]^{1/2}}$$

for two $0-1$ binary variables with the following frequencies:

Table 13: Frequencies of Variable 1 and 2

		Variable 2	
		0	1
Variable 1	0	a	b
	1	c	d

Solution: First, we form the table of binary variables as follows:

Table 14: Binary Values of Variables X and Y

		y		
		1	0	Total
x	1	a	b	$a+b$
	0	c	d	$c+d$
	Total	$a+c$	$b+d$	$p=a+b+c+d$

We have

$$\bar{x} = \frac{(a+b)}{p},$$

$$\bar{y} = \frac{(a+c)}{p}.$$

The simple correlation coefficient formula is

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}, \tag{9}$$

where the term $\sum_{i=1}^{n} (x_i - \bar{x})^2$ can be calculated as follows:

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = (a+b)\left(1 - \frac{(a+b)}{p}\right)^2 + (c+d)\left(0 - \frac{(a+b)}{p}\right)^2 = \frac{(c+d)(a+b)}{p}.$$

Similarly, we can write

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = (a+c)\left(1 - \frac{(a+c)}{p}\right)^2 + (b+d)\left(0 - \frac{(a+c)}{p}\right)^2 = \frac{(a+c)(b+d)}{p}.$$

Next, we have

$$\begin{aligned}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right) &= a-\frac{(a+c)(a+b)}{p}-\frac{(a+b)(a+c)}{p}-p\frac{(a+b)(a+c)}{p^2}\\ &= \frac{a(a+b+c+d)-(a+c)(a+b)}{p}\\ &= \frac{a^2+ab+ac+ad-a^2-ab-ac-ba}{p}\\ &= \frac{ad-bc}{p}.\end{aligned}$$

Now, we insert the above expressions into the simple correlation coefficient formula (9) and obtain

$$\begin{aligned}r &= \frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}}\\ &= \frac{\left(\frac{ad-bc}{p}\right)}{\left[\frac{(c+d)(a+b)}{p}\times\frac{(a+c)(b+d)}{p}\right]^{\frac{1}{2}}} = \frac{ad-bc}{\left[(a+b)(c+d)(a+c)(b+d)\right]^{\frac{1}{2}}}.\end{aligned}$$

Thus, the relation of interest is verified. Note that variables with large negative correlations are regarded as very dissimilar and variables with large positive correlations are regarded as very similar.

3 Hierarchical Clustering Methods

Hierarchical clustering is conducted by either a series of successive mergers or a series of successive divisions. *Agglomerative hierarchical methods* begin with the individual objects; so there are as many clusters as objects initially. They first group the most similar objects, and then these groups are merged according to their similarities. Eventually, all subgroups are merged into a single cluster. *Divisive hierarchical methods* initially divide all objects into two subgroups in which the objects in one subgroup are "far from" the objects in the other group. The subgroups are further divided into dissimilar subgroups until each object forms a group itself. The results of both methods could be depicted in a *dendrogram* that illustrates the mergers or divisions made at successive levels.

Our focus is agglomerative hierarchical procedures, and particularly, *linkage methods* which are suitable for clustering items as well as variables. In turn, *single linkage* (min-

imum distance or nearest neighbor), *complete linkage* (maximum distance or farthest neighbor), and *average linkage* (average distance) shall be discussed.

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

1. Begin with N clusters, each of which contains a single entity and an $N \times N$ symmetric matrix of distances (or similarities) $\mathbf{D} = \{d_{ik}\}$.

2. Find the distance matrix for the nearest pair of clusters. Suppose the distance between the nearest clusters U and V is d_{UV}.

3. Merge clusters U and V, and label the new cluster (UV). Recalculate the entries in the distance matrix by deleting the rows and columns corresponding to clusters U and V and adding a row and column giving the distances between cluster (UV) and the remaining clusters.

4. Follow Steps 2 and 3 a total of $N - 1$ times until all objects will be in a single cluster. Make a record of the identity of clusters that are merged and the distance at which the mergers take place.

3.1 Linkage Methods

3.1.1 Single Linkage

Distances or similarities between pairs of objects can be inputs to a single linkage algorithm. Clusters are merged from the individual entities by combining *nearest neighbors*, which mean the smallest distance or largest similarity.

First, we should identify the smallest distance in $\mathbf{D} = \{d_{ik}\}$ and merge the corresponding objects, say, U and V, to find the cluster (UV). For Step 3 in the above algorithm, the distance between (UV) and any other cluster W is calculated by

$$d_{(UV)W} = \min \{d_{UW}, d_{VW}\}. \tag{10}$$

The quantities d_{UW} and d_{VW} are the distances between the nearest neighbours of clusters U and W and clusters V and W, respectively.

In a typical application of hierarchical clustering, the intermediate results, where the objects are sorted into a moderate number of clusters, are of great interest. Because single linkage joins clusters by the shortest link between them, the method cannot discern poorly separated clusters. Also, the method is one of a few clustering methods that can delineate nonellipsoidal clusters. The tendency of single linkage to pick out

long stringlike clusters is known as *chaining* that can be misleading if items at opposite ends of the chain are quite dissimilar.

The clusters formed by the method will be unchanged by any initial assignment of distance (similarity) that gives the same relative orderings. In particular, any of a set of similarity coefficients from Table 3 that are monotonic to one another will produce the same clustering in the end.

3.1.2 Complete Linkage

Complete linkage clustering is conducted in the almost same way as single linkage clustering, with one important exception: at each stage, the distance between clusters is determined by the distance between the elements, one from each cluster, that are most distant. So, complete linkage guarantees that all items in a cluster are within some maximum distance (or minimum similarity) of each other.

The general agglomerative algorithm starts by finding the minimum entry in $\mathbf{D} = \{d_{ik}\}$ and merging the corresponding objects, such as U and V, to get cluster (UV). For Step 3 of the above algorithm, the distance between (UV) and any other cluster W is computed by

$$d_{(UV)W} = \max\{d_{UW}, d_{VW}\}. \tag{11}$$

Here d_{UW} and d_{VW} are the distances between the most distant members of clusters U and W and cluster V and W, respectively.

Similar to the single linkage method, a new assignment of distances (similarities) that have the same relative orderings as the initial distances will not change the configuration of the complete linkage clusters.

3.1.3 Average Linkage

Average linkage treats the distance between two clusters as the average distance between all pairs of items where one member of a pair belongs to each cluster. The method begins by searching the distance matrix $\mathbf{D} = \{d_{ik}\}$ to find the nearest (most similar) objects, for instance, U and V. These objects are merged to form the cluster (UV). For Step 3 of the above algorithm, the distance between (UV) and the other cluster W is determined by

$$d_{(UV)W} = \frac{\sum_i \sum_k d_{ik}}{N_{(UV)} N_W}, \tag{12}$$

where d_{ik} is the distance between object i in the cluster (UV) and object k in the cluster W, and $N_{(UV)}$ and N_W are the number of items in cluster (UV) and W, respectively.

For average linkage clustering, changes in the assignment of distances (similarities) can affect the arrangement of the final cluster configuration, but preserve relative orderings.

3.2 Examples

Example 4 (Exercise 12.5 from Johnson and Wichern [7]). Consider the matrix of distances

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{pmatrix} 0 & 1 & 11 & 5 \\ 1 & 0 & 2 & 3 \\ 11 & 2 & 0 & 4 \\ 5 & 3 & 4 & 0 \end{pmatrix} \end{array}$$

Cluster the four items using each of the following procedures.

(a) Single linkage hierarchical procedure.

(b) Complete linkage hierarchical procedure.

(c) Average linkage hierarchical procedure.

Draw the dendrograms and compare the results in (a), (b), and (c).

Solution: (a) We treat each subject in the above matrix as cluster, apply clustering by using single linkage, and merge the two closest items. Since

$$\min_{i,k}(d_{ik}) = d_{21} = 1,$$

we merge objects 2 and 1 to form a cluster (12). To move on to the next level of clustering, calculate the distances between the cluster (12) and the remaining objects 3 and 4. The nearest neighbour distances can be calculated as below:

$$d_{(12)3} = \min\{d_{13}, d_{23}\} \ = \min\{11, 2\} = 2,$$

and also

$$d_{(12)4} = \min\{d_{14}, d_{24}\} = \min\{5, 3\} = 3.$$

We now delete the rows and columns of the above matrix corresponding to objects 1 and 2, and add cluster (12) to obtain new distance matrix as below:

$$\begin{array}{c} \\ (12) \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccc} (12) & 3 & 4 \end{array} \\ \begin{pmatrix} 0 & 2 & 3 \\ 2 & 0 & 4 \\ 3 & 4 & 0 \end{pmatrix} \end{array}.$$

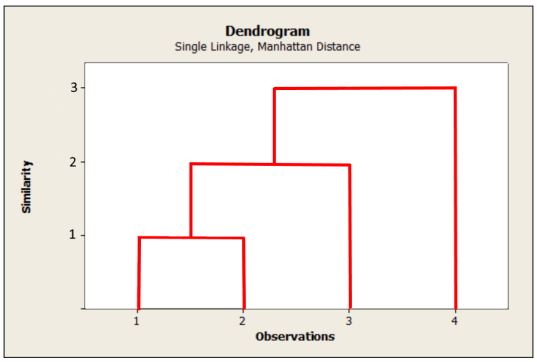
The smallest distance between pairs of clusters is now $d_{(12)3} = 2$. Hence, we merge cluster 3 with cluster (12) to get a new cluster (312). Next, we have

$$d_{(312)4} = \min\left\{d_{(12)4}, d_{(34)}\right\} = \min\{3, 4\} = 3,$$

and hence, the final matrix is as shown below:

$$\begin{array}{c} \\ (312) \\ 4 \end{array} \begin{array}{c} \begin{array}{cc} (312) & 4 \end{array} \\ \begin{pmatrix} 0 & 3 \\ 3 & 0 \end{pmatrix} \end{array}.$$

Finally, clusters (312) and 4 are merged together to form a single cluster (1234) when the nearest neighbour distance reaches 3. The dendrogram picturing the hierarchical clustering using single linkage algorithm just concluded is shown below. The groupings and the distance levels at which they occur are illustrated by the diagram.



(b) We treat each subject in the above matrix as cluster. To apply clustering by using complete linkage, merge the two closest items. Since

$$\max_{i,k}(d_{ik}) = d_{21} = 2,$$

we merge objects 2 and 1 to form a cluster (12). To move on to the next level of clustering, calculate the distance between the cluster (12) and the remaining objects 3 and 4.

The nearest neighbour distances can be calculated as follows:

$$d_{(12)3} = \max\{d_{13}, d_{23}\} = \max\{11, 2\} = 11,$$

and

$$d_{(12)4} = \max\{d_{14}, d_{24}\} = \max\{5, 3\} = 5.$$

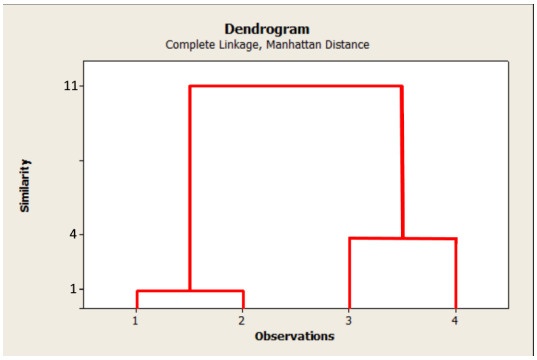
Delete the rows and columns of the above matrix corresponding to object 1 and 2, and add cluster (12) to obtain new distance matrix as below:

$$\begin{array}{c} \\ (12) \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccc} (12) & 3 & 4 \end{array} \\ \begin{pmatrix} 0 & 11 & 5 \\ 11 & 0 & 4 \\ 5 & 4 & 0 \end{pmatrix} \end{array}.$$

At the final step, clusters (312) and 4 are merged together to form a single cluster (1234) when the nearest neighbour distances reaches 11. The final matrix is as follows:

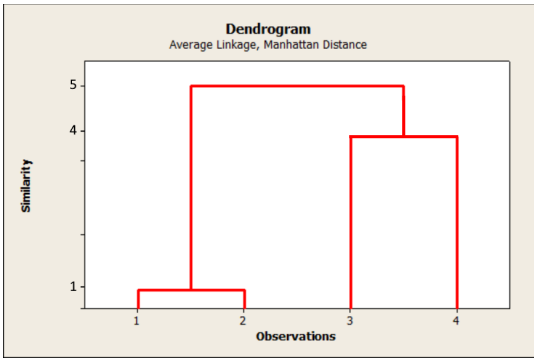
$$\begin{array}{c} \\ (312) \\ 4 \end{array} \begin{array}{c} \begin{array}{cc} (312) & 4 \end{array} \\ \begin{pmatrix} 0 & 11 \\ 11 & 0 \end{pmatrix} \end{array}.$$

The dendrogram picturing the hierarchical clustering using complete linkage algorithm just concluded is shown below. The groupings and the distance levels at which they occur are illustrated by the diagram.



(c) Due to intensive computation of average linkage procedures, we use Minitab to cluster the five items instead. The dendrogram picturing the hierarchical clustering

using average linkage algorithm just concluded is shown below. The groupings and the distance levels at which they occur are illustrated by the diagram.



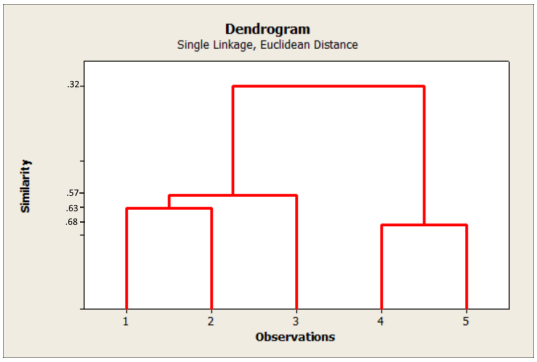
Example 5 (Exercise 12.7 from Johnson and Wichern [7]). Sample correlations for five stocks, rounded to two decimal places, are as given in Table 15.

Table 15: Sample correlations

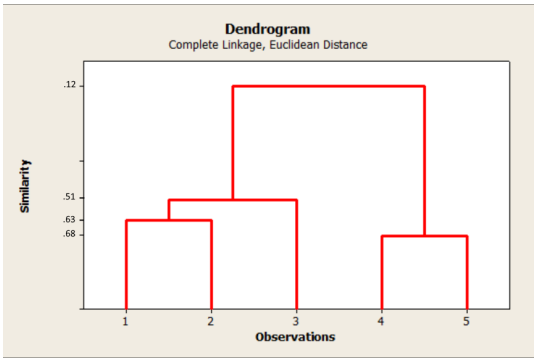
	JP Morgan	Citibank	Wells Fargo	Royal DutchShell	Exxon Mobil
JP Morgan	1				
Citibank	.63	1			
Wells Fargo	.51	.57	1		
Royal DutchShell	.12	.32	.18	1	
Exxon Mobil	.16	.21	.15	.68	1

Treating the sample correlations as similarity measures, cluster the stocks using the single linkage and complete linkage hierarchical procedures. Draw the dendrograms and compare the results.

Solution: (a) We use Minitab to cluster the five stocks using single linkage procedure.



(b) We use Minitab to cluster the five stocks using complete linkage procedure.



The above two methods produce almost the same clustering configuration in their final stage.

3.3 Ward's Hierarchical Clustering Method

J. H. Ward Jr. [11] proposed a hierarchical clustering method based on minimizing the "loss of information" from joining two groups. This method regards an increase in an error sum of squares, ESS, as loss of information. First, for a given cluster k, let ESS_k be the sum of the squared deviations of every item in the cluster from the cluster mean (centroid). If there are currently K clusters, define ESS as the sum of the ESS_ks or $\text{ESS} = \text{ESS}_1 + \text{ESS}_2 + \ldots + \text{ESS}_k$. At each step, the union of every possible pair of clusters is considered, and the two clusters whose combination results in the smallest increase in ESS (minimum loss of information) are joined. Initially, each cluster consists of a single item and, if there are N items, $\text{ESS}_k = 0, k = 1, 2, \ldots, N$, so $\text{ESS} = 0$. At the other extreme, when all the clusters are combined in a single group of N items, the

value of ESS is given by

$$\text{ESS} = \sum_{j=1}^{N} (\mathbf{x}_j - \bar{\mathbf{x}})^\top (\mathbf{x}_j - \bar{\mathbf{x}}), \tag{13}$$

where $\mathbf{x}_j$ is the multivariate measurement associated with the jth item and $\bar{\mathbf{x}}$ is the mean of all the items. The results of Ward's method can be displayed as a dendrogram in which the vertical axis gives the values of ESS where the mergers occur.

Ward's method is based on the assumption that the clusters of multivariate observations are expected to be roughly elliptically shaped. It is a hierarchical precursor to nonhierarchical clustering methods that optimize some criterion for dividing data into a *given* number of elliptical groups.

3.4 Final Comments on Hierarchical Procedures

Hierarchical procedures do not take sources of error and variation into account which means that a clustering method will be sensitive to outliers, or "noise points." Also, there is no provision for a reallocation of objects that may have been "incorrectly" grouped at an early stage; therefore, the final configuration of clusters should be carefully examined to see whether it is sensible or not. For a particular problem, it is a good idea to try several clustering methods and, within a given method, a couple of different ways of assigning distances. If the outcomes from the several methods are (roughly) consistent with one another, perhaps a case for "natural" groupings can be advanced.

The *stability* of a hierarchical solution can be checked by applying the clustering algorithm before and after *small* errors (perturbations) have been added to the data units. If the groups are fairly well distinguished, the clusterings before and after perturbation should agree.

Common values in the similarity or distance matrix can produce multiple solutions to a hierarchical clustering problem. This is not an inherent problem of any method; rather, multiple solutions occur for certain kinds of data. Some data sets and hierarchical clustering methods can produce *inversions* that occur when an object joins an existing cluster at a smaller distance (greater similarity) than that of a previous mergers. Inversions can occur when there is no clear cluster structure and are generally associated with two hierarchical clustering algorithms known as the centroid method and the median method.

4 Nonhierarchical Clustering Methods

Nonhierachical clustering methods aim to group *items* instead of *variables* into K clusters, where K can be either determined beforehand or during the clustering procedure. These methods can be utilized to much larger data sets than do hierarchical methods because a matrix of distances (similarities) does not have to be determined, and none of the basic data have to be stored during the computation. These methods start from the nuclei of clusters which can be formed from either an initial partition of items into groups or an initial set of seed points. Starting configurations should be chosen without overt biases, i.e., randomly select seed points from the items or randomly partition the items into initial groups. We will discuss the K-means method, one of the most popular nonhierarchical procedures.

4.1 K-means Method

J. B. MacQueen [8] gives the term K-means to one of his algorithms assigning each item to the cluster having the nearest centroid (mean). The algorithm is composed of three steps below:

1. Divide the items into K initial clusters.

2. Assign each item in the list to the cluster whose centroid (mean) is nearest, by the distance computed using Euclidean distance with either standardized or unstandardized observations. After assignment, recalculate the centroid for the cluster gaining the new item and for the cluster losing the item.

3. Repeat Step 2 until there is no more reassignments.

We could assign K initial centroids (seed points) and carry on Step 2 instead of partitioning all items into K preliminary groups in Step 1. The final assignment of all items will be influenced by the initial partition or the selection of centroids.

4.2 Example

Example 5 (Exercise 12.11 from Johnson and Wichern [7]). Suppose we measure two variables X_1 and X_2 for four items A, B, C and D. The data are collected in Table 16.

Table 16: Measurements of two variables

	Observations	
Item	x_1	x_2
A	5	4
B	1	-2
C	-1	1
D	3	1

Use the K-means clustering technique to divide the items into $K = 2$ clusters. Start with the initial groups (AB) and (CD).

Solution: We use the K-means clustering technique to divide the items into $K = 2$ clusters. The initial groups (AB) and (CD) are given as below:

	Coordinates of the centroid	
cluster	$\overline{x}_1$	$\overline{x}_2$
(AB)	$\frac{5+1}{2} = 3$	$\frac{4-2}{2} = 1$
(CD)	$\frac{-1+3}{2} = 1$	$\frac{1+1}{2} = 1$

Similarly, calculate the remaining clusters that are given as below:

	Coordinates of the centroid	
Cluster	$\overline{x}_1$	$\overline{x}_2$
(AC)	2	2.5
(AD)	4	2.5
(BD)	2	-0.5
(BC)	0	-0.5

The squared distance of the centroids from final cluster (AD) and four items A, B, C and D are computed as below:

$$\begin{aligned} d^2\left(A,(AD)\right) &= \left(x_1-\overline{x}_1\right)^2+\left(x_2-\overline{x}_2\right)^2 \\ &= (5-4)^2+(4-2.5)^2 \\ &= 1+2.25 \\ &= 3.25, \end{aligned}$$

$$
\begin{aligned}
d^2\left(B,(AD)\right) &= \left(x_1-\overline{x}_1\right)^2+\left(x_2-\overline{x}_2\right)^2 \\
&= (1-4)^2+(-2-2.5)^2 \\
&= 9+20.25 \\
&= 29.25,
\end{aligned}
$$

$$
\begin{aligned}
d^2\left(C,(AD)\right) &= \left(x_1-\overline{x}_1\right)^2+\left(x_2-\overline{x}_2\right)^2 \\
&= (-1-4)^2+(1-2.5)^2 \\
&= 25+2.25 \\
&= 27.25,
\end{aligned}
$$

$$
\begin{aligned}
d^2\left(D,(AD)\right) &= \left(x_1-\overline{x}_1\right)^2+\left(x_2-\overline{x}_2\right)^2 \\
&= (3-4)^2+(1-2.5)^2 \\
&= 1+2.25 \\
&= 3.25\,.
\end{aligned}
$$

Similarly, we calculate the squared distances of the centroids from the final cluster (BC) and four items A, B, C and D. The calculated distances are as follows:

$$
\begin{aligned}
d^2\left(A,(BC)\right) &= \left(x_1-\overline{x}_1\right)^2+\left(x_2-\overline{x}_2\right)^2 \\
&= (5-0)^2+(4+0.5)^2 \\
&= 25+20.25 \\
&= 45.25,
\end{aligned}
$$

$$
\begin{aligned}
d^2\left(B,(BC)\right) &= \left(x_1-\overline{x}_1\right)^2+\left(x_2-\overline{x}_2\right)^2 \\
&= (1-0)^2+(-2+0.5)^2 \\
&= 1+2.25 \\
&= 3.25,
\end{aligned}
$$

$$
\begin{aligned}
d^2\left(C,(BC)\right) &= \left(x_1-\overline{x}_1\right)^2+\left(x_2-\overline{x}_2\right)^2 \\
&= (-1+0)^2+(1+0.5)^2 \\
&= 1+2.25 \\
&= 3.25,
\end{aligned}
$$

$$
\begin{aligned}
d^2\left(D,(BC)\right) &= \left(x_1-\overline{x}_1\right)^2+\left(x_2-\overline{x}_2\right)^2 \\
&= (3-0)^2+(1+0.5)^2 \\
&= 9+2.25 \\
&= 11.25\,.
\end{aligned}
$$

For the final clusters, we have

	Squared distance to group centroids			
Cluster	A	B	C	D
(AD)	3.25	29.25	27.25	3.25
(BC)	45.25	3.25	3.25	11.25

Therefore, the final partition (AD) and (BC) is given as below:

	$\overline{x}_1$	$\overline{x}_2$
(AD)	4	2.5
(BC)	0	-0.5

The coordinates of Cluster (AD)'s centroid is $(4, 2.5)$; the ones of Cluster (BC)'s centroid is $(0, -0.5)$.

4.3 Final Comments–Nonhierarchical Procedures

To check the stability of the clustering procedure, we should run the algorithm with a different initial partition or centroids. Once clusters emerges, the list of items should be rearranged so that items in the first cluster appear first, those in the second cluster appear next, and so on. A table of the cluster centroids and within-cluster variances also help us to find the differences between clusters. The importance of individual variables in clustering should be evaluated from a multivariate perspective. Descriptive statistics can be helpful in assessing the importance of individual variables and the "success" of the clustering algorithm.

Some strong arguments for not fixing the number of clusters beforehand are:

1. Two or more seed points lying within a single cluster causes their resulting clusters to be poorly differentiated.

2. An outlier might cause at least one cluster in which items disperse widely.

3. A sample may not contain items from the rarest group in a population. Forcing the sample data into K groups would produce nonsensical clusters.

It is advisable to rerun the algorithm for several choices to the number of centroids.

5 Clustering Based on Statistical Models

The clustering methods discussed earlier, for instance, single linkage, complete linkage, average linkage, Ward's method and K-means clustering, are intuitively reasonable procedures. However, the introduction of statistical models has made major advances in clustering techniques because statistical models explain how the collection of $(p \times 1)$ measurements $\mathbf{x}_j$, from the N objects, was generated. We will discuss the most common model where cluster k has expected proportion p_k of the objects and the corresponding measurements are generated by a probability density function $f_k(\mathbf{x})$. If there are K clusters, the *mixing distribution* is used to model the observation vector for a single object:

$$f_{Mix}(\mathbf{x}) = \sum_{k=1}^{K} p_k f_k(\mathbf{x}), \qquad \mathbf{x} \in \mathbb{R}^p,$$

where each $p_k \geq 0$ and $\sum_{k=1}^{K} p_k = 1$. Because the observation comes from the component distribution $f_k(\mathbf{x})$ with probability p_k, the distribution $f_{Mix}(\mathbf{x})$ is a mixture of the K distributions $f_1(\mathbf{x})$, $f_2(\mathbf{x})$, $\ldots$, $f_K(\mathbf{x})$. Therefore, the collection of N observation vectors generated from this distribution should be a mixture of observations from the component distributions.

A mixture of multivariate normal distributions in which the kth component $f_k(\mathbf{x})$ is the $N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ density function is the most common mixture model. The normal mixture model for a single observation $\mathbf{x}$ is

$$f_{Mix}(\mathbf{x} \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K) = \sum_{k=1}^{K} \frac{p_k}{(2\pi)^{p/2} \mid \boldsymbol{\Sigma}_k \mid^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k)\right). \tag{14}$$

Clusters obtained from this model should be ellipsoidal in shape that is concentrated heavily near the center. Inferences are based on the likelihood for N objects and a fixed

number of clusters K, which is given by

$$\begin{aligned} L\left(p_1, ..., p_K, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, ..., \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_K\right) &= \prod_{j=1}^{N} f_{Mix}\left(\mathbf{x}_j \left|\boldsymbol{\mu}_1\right., \boldsymbol{\Sigma}_1, \cdots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K\right) \\ &= \prod_{j=1}^{N} \left(\sum_{k=1}^{K} P \cdot \exp\left(Q\right)\right), \end{aligned} \tag{15}$$

where the proportions $p_1, \ldots, p_k$, the mean vectors $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$, and the covariance matrices $\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_k$ are unknown, and $P = \frac{p_k}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_k|^{1/2}}$ and $Q = -\frac{1}{2}\left(\mathbf{x} - \boldsymbol{\mu}_k\right)^{\top} \boldsymbol{\Sigma}_k^{-1}\left(\mathbf{x} - \boldsymbol{\mu}_k\right)$. The measurements for different objects are regarded as independent and identically distributed observations from the mixture distribution $f_{Mix}\left(\mathbf{x}\right)$.

The likelihood-based procedure under the normal mixture model with all $\boldsymbol{\Sigma}_k$ being the same multiple of the identity matrix, $\eta\mathbf{I}$, under which certain conclusions can be drawn based on a heuristic clustering method, is approximately analogues to K-means clustering and Ward's method since these methods use the distance computed using the Euclidean distance with either standardized or unstandardized observations. It is a major advance that under the sequence of mixture models (12) for different K, the problems of choosing the number of clusters and an appropriate clustering method become the one of selecting an appropriate statistical model.

The maximum likelihood estimators $\hat{p}_1, \ldots, \hat{p}_K, \hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\Sigma}}_1, \ldots, \hat{\boldsymbol{\mu}}_K, \hat{\boldsymbol{\Sigma}}_K$ for a fixed number of clusters K must be obtained numerically using special software. The resulting estimator

$$L_{\max} := L\left(\hat{p}_1, \ldots, \hat{p}_K, \hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\Sigma}}_1, \ldots, \hat{\boldsymbol{\mu}}_K, \hat{\boldsymbol{\Sigma}}_K\right)$$

provides the basis for model selection. To compare models with different numbers of parameters, a penalty is subtracted from twice the maximum of the log-likelihood to give

$$-2\ln L_{\max} - \mathit{Penalty},$$

where the penalty depends on the number of parameters estimated and the number of observations N. Because the probabilities p_k sum up to 1, there are $K-1$ probabilities that must be estimated, $K \times p$ means and $K \times p\left(p+1\right)/2$ variances and covariances. For the Akaike information criterion (AIC), the penalty is $2N \times$ (number of parameters) so

$$\mathrm{AIC} = 2\ln L_{\max} - 2N\left(K\frac{1}{2}\left(p+1\right)\left(p+2\right) - 1\right). \tag{16}$$

The Bayesian information criterion (BIC) uses the logarithm of the number of pa-

rameters in the penalty function:

$$\text{BIC} = 2\ln L_{\max} - 2\ln(N)\left(K\frac{1}{2}(p+1)(p+2) - 1\right). \tag{17}$$

Under either AIC or BIC, the better model is indicated by the higher score using AIC or BIC formula.

To avoid occasional difficulty with too many parameters in the mixture model, we assume simple structures for the $\mathbf{\Sigma}_k$. Namely, we allow the covariance structures as listed in Table 17.

Table 17: BIC for Different Scenarios

Assumed form for $\mathbf{\Sigma}_k$	Total number of parameters	BIC
$\mathbf{\Sigma}_k = \eta\,\mathbf{I}$	$K(p+1)$	$\ln L_{\max} - 2\ln N K(p+1)$
$\mathbf{\Sigma}_k = \eta_k\,\mathbf{I}$	$K(p+2)-1$	$\ln L_{\max} - 2\ln N\,(K(p+2-1))$
$\mathbf{\Sigma}_k = \eta_k\,\text{Diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$	$K(p+2)+p-1$	$\ln L_{\max} - 2\ln N\,(K(p+2)+p-1)$

The package *MCLUST* available in the R software library, provides the combination of hierarchical clustering, the EM algorithm and the BIC criterion to develop an appropriate model for clustering.

Example 6 (Example 12.13: A model based clustering of the Iris data Johnson and Wichern [7]). For the Iris data in Table 11.5 (see Appendix I), we use MCLUST (version 5.4.7) which provides the EM algorithm to determine the model based on the best BIC values, using different structures of the covariance matrix. As a multivariate data set, the Iris data consists of 50 samples from each of three species of Iris: *Iris setosa*, *Iris virginica* and *Iris versicolor*, totaling 150 samples; the length and the width of the sepals and petals were measured from each sample. Thus, the Iris data set has five variables: sepal length, sepal width, petal length, petal width and species; it has three clusters: Iris setosa, Iris virginica and Iris versicolor. A matrix plot of the clusters for pairs of variables, among sepal length, sepal width, petal length and petal width, is as shown on Figure 7. The scatter plots of petal length and petal width suggest certain positive correlations between the two variables.

The best BIC value is -561.7285 (VEV, 2 clusters) shown on Figure 8, where VEV is an identifier for covariance parameter of $G + (d-1) + G\left[d(d-1)/2\right]$, with G denoting the number of mixture components and d denoting the dimension of the data. The EII,

VEE, etc. are the identifiers for covariance parameters in MCLUST package. Figure 8 shows that the "best" model can be estimated by fitting the model with VEV covariance parameter with 2 clusters.

The plot, located in the second row and the first column, and boxed in red lines, on Figure 9, demonstrates the best model based on the above computation.

This clustering algorithm suggests that the Iris data set can be divided into two, rather obvious, clusters: one of the clusters is Iris setosa, and the other contains both Iris virginica and Iris vesrsicolor.

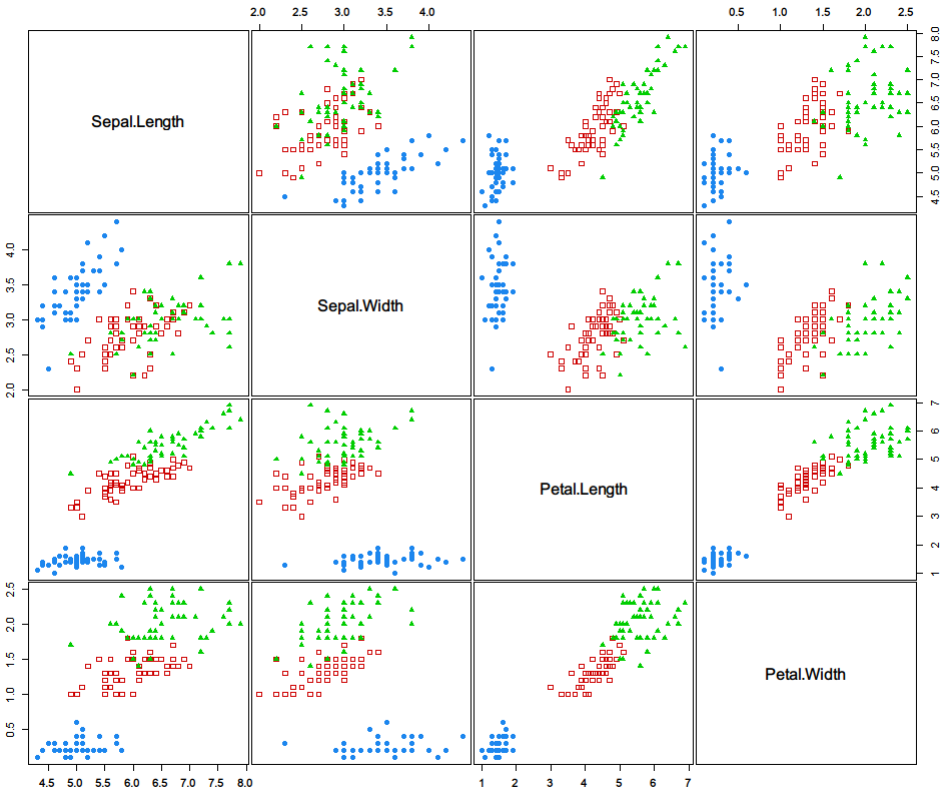


Figure 7: A scatter plot matrix for pairs of variables, where blue = setosa, red = versicolor, green = virginica

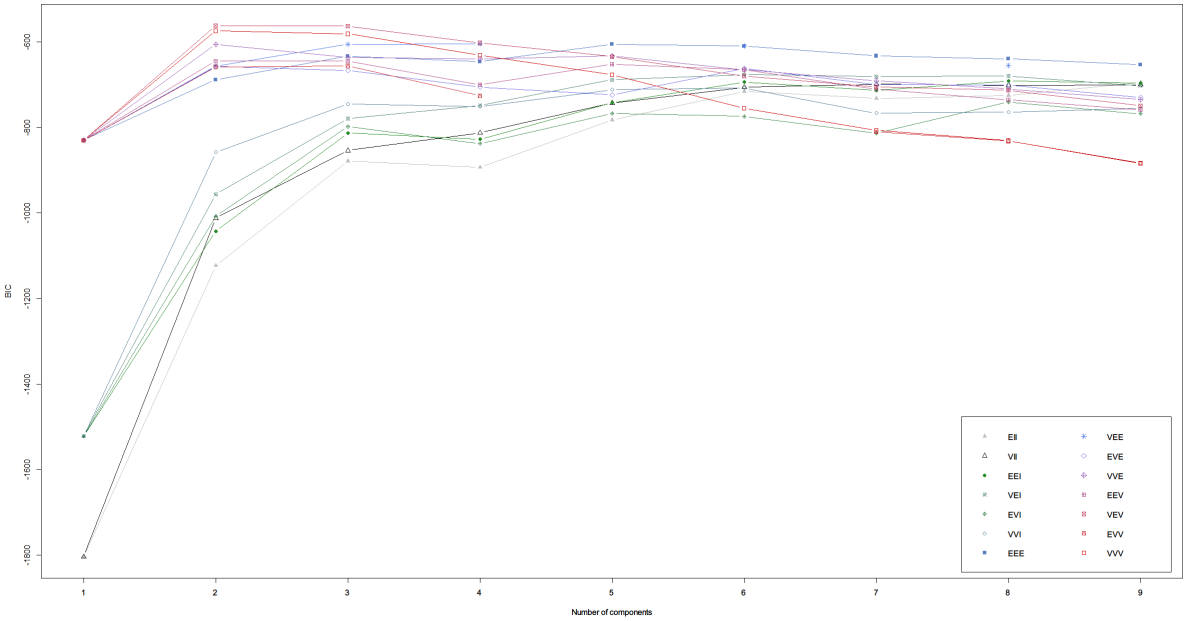


Figure 8: the BIC plot with different covariance parameter models, the "best" model is the model with VEV, 2 clusters.

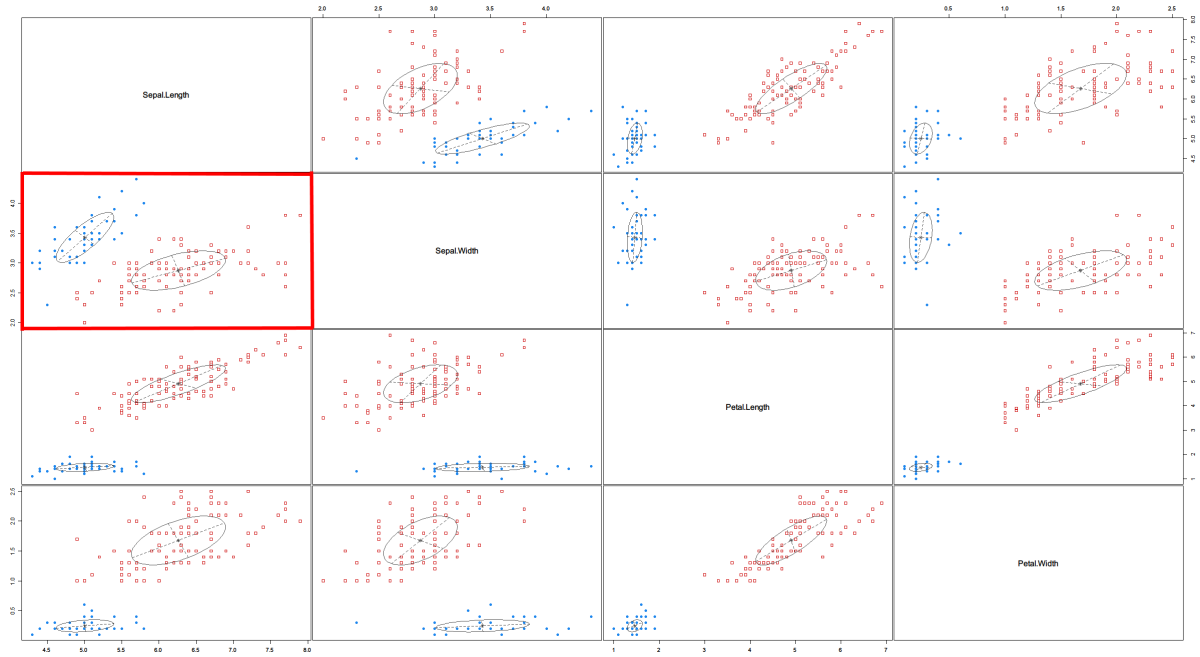


Figure 9: Matrix plot of different models

6 Correspondence Analysis

A graphical procedure that represents associations in a table of frequencies is called correspondence analysis. This project focuses on two-way table of frequencies or *contingency table*. Suppose the contingency table has I rows and J columns, the plot is produced by correspondence analysis containing two sets of points: a set of I points corresponding to the rows and a set of J points corresponding to the columns. Associations between the points is reflected by the positions of the points on the plot. Row points being close together show rows having similar profiles (conditional distributions) across the columns; column points being close together show columns having similar profiles (conditional distributions) down the rows; and row points being close to column points demonstrates combinations occurring more frequently than would be expected from an independence mode in which the row categories are unrelated to the column categories. The usual output from a correspondence analysis provides the "best" two-dimensional representation of the data, the coordinates of the plotted points, and the *inertia* of the amount of information retained in each dimension.

6.1 Algebraic Development of Correspondence Analysis

Let $\mathbf{X}$, with elements x_{ij}, be an $I \times J$ two-way table of unscaled frequencies. We take $I > J$ and assume that $\mathbf{X}$ is of full column rank J in the table of which the rows and columns correspond to different categories of two different characteristics. If n is the total of the frequencies in the data matrix $\mathbf{X}$, we should first construct a matrix of proportions $\mathbf{P} = \{p_{ij}\}$ by dividing each element of $\mathbf{X}$ by n. Thus, we obtain

$$p_{ij} = \frac{x_{ij}}{n}, \qquad i = 1, 2, \ldots, I, \quad j = 1, 2, \ldots, J, \qquad \text{or} \qquad \underset{(I\times J)}{\mathbf{P}} = \frac{1}{n}\underset{(I\times J)}{\mathbf{X}}. \tag{18}$$

The matrix $\mathbf{P}$ is called the *correspondence matrix*.

We define the vector of row sums $\mathbf{r} = (r_1, r_2, \ldots, r_{\mathrm{J}})^\top$ as follows:

$$r_i = \sum_{j=1}^{J} p_{ij} = \sum_{j=1}^{J} \frac{x_{ij}}{n}, \qquad i = 1, 2, \ldots, I, \qquad \text{or} \qquad \underset{(I\times 1)}{\mathbf{r}} = \underset{(I\times J)}{\mathbf{P}}\,\underset{(J\times 1)}{\mathbf{1}_J}$$

and the vector of column sum $\mathbf{c} = (c_1, c_2, \ldots, c_{\mathrm{I}})^\top$ as below:

$$c_j = \sum_{i=1}^{I} p_{ij} = \sum_{i=1}^{I} \frac{x_{ij}}{n}, \qquad j = 1, 2, \ldots, J, \qquad \text{or} \qquad \underset{(J\times 1)}{\mathbf{c}} = \underset{(J\times 1)}{\mathbf{P}^\top}\,\underset{(I\times 1)}{\mathbf{1}_I}$$

where $\mathbf{1}_J$ is a $J \times 1$ and $\mathbf{1}_I$ is a $I \times 1$ vector of $1's$. We define the diagonal matrix $\mathbf{D}_r$ with the elements of $\mathbf{r}$ on the main diagonal:

$$\mathbf{D}_r = \text{diag}\left(r_1, r_2, \ldots, r_I\right),$$

and the diagonal matrix $\mathbf{D}_c$ with the elements of $\mathbf{c}$ on the main diagonal:

$$\mathbf{D}_c = \text{diag}\left(c_1, c_2, \ldots, c_J\right).$$

Also, for scaling purposes, we define the square root matrices:

$$\mathbf{D}_r^{1/2} = \text{diag}\left(\sqrt{r_1}, \sqrt{r_2}, \ldots, \sqrt{r_I}\right), \tag{19}$$

$$\mathbf{D}_r^{-1/2} = \text{diag}\left(\frac{1}{\sqrt{r_1}}, \frac{1}{\sqrt{r_2}}, \ldots, \frac{1}{\sqrt{r_I}}\right),$$

$$\mathbf{D}_c^{1/2} = \text{diag}\left(\sqrt{c_1}, \sqrt{c_2}, \ldots, \sqrt{c_J}\right), \tag{20}$$

$$\mathbf{D}_c^{-1/2} = \text{diag}\left(\frac{1}{\sqrt{c_1}}, \frac{1}{\sqrt{c_2}}, \ldots, \frac{1}{\sqrt{c_J}}\right).$$

Correspondence analysis can be formulated in the form of the weighted least squares problem to select $\hat{\mathbf{P}} = \{\hat{p}_{ij}\}$, a matrix of specified reduced rank, to minimize

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(p_{ij} - \hat{p}_{ij}\right)^2}{r_i c_j} = \text{tr}\left[\left(\mathbf{D}_r^{-1/2}\left(\mathbf{P} - \hat{\mathbf{P}}\right)\mathbf{D}_c^{-1/2}\right)\left(\mathbf{D}_r^{-1/2}\left(\mathbf{P} - \hat{\mathbf{P}}\right)\mathbf{D}_c^{-1/2}\right)^\top\right], \tag{21}$$

because $\frac{(p_{ij} - \hat{p}_{ij})}{\sqrt{r_i c_j}}$ is the (i,j)element of $\mathbf{D}_r^{-1/2}\left(\mathbf{P} - \hat{\mathbf{P}}\right)\mathbf{D}_c^{-1/2}$.

Before introducing Result 12.1, we will review the relevant Result 2A.15 and 2A.16 from Johnson and Wichern [7]. These are standard results of Linear Algebra.

Result 2A.15. *Singular-Value Decomposition.* Let $\mathbf{A}$ be an $m \times k$ matrix of real numbers. Then there exist an $m \times m$ orthogonal matrix $\mathbf{U}$ and a $k \times k$ orthogonal matrix $\mathbf{V}$ such that

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top,$$

where the $m \times k$ matrix $\mathbf{\Lambda}$ has (i,i) entry $\lambda_i \geq 0$ for $i = 1, 2, \ldots, \min\left(m, k\right)$ and the other entries are zero. The positive constants λ_i are called the *singular values* of $\mathbf{A}$.

The singular-value decomposition can also be expressed as a matrix expansion that depends on the rank r of $\mathbf{A}$. There exist r positive constants $\lambda_1, \lambda_2, \ldots, \lambda_r$, r orthogonal

$m \times 1$ unit vectors $\mathbf{u}_1, \mathbf{u}_2 \ldots, \mathbf{u}_r$ and r orthogonal $k \times 1$ unit vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r$, such that

$$\mathbf{A} = \sum_{i=1}^{r} \lambda_i \mathbf{u}_i \mathbf{v}_i^\top = \mathbf{U}_r \mathbf{\Lambda}_r \mathbf{V}_r^\top, \tag{22}$$

where $\mathbf{U}_r = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r)$, $\mathbf{V}_r = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r)$, and $\mathbf{\Lambda}_r$ is an $r \times r$ diagonal matrix with diagonal entries λ_i. Here $\mathbf{A}\mathbf{A}^\top$ has eigenvalue-eigenvector pairs $(\lambda_i^2, \mathbf{u}_i)$, so

$$\mathbf{A}\mathbf{A}^\top \mathbf{u}_i = \lambda_i^2 \mathbf{u}_i$$

with $\lambda_1^2, \lambda_2^2, \ldots, \lambda_r^2 > 0 = \lambda_{r+1}^2, \lambda_{r+2}^2, \ldots, \lambda_m^2$ (for $m > k$). Then, in view of (21), $\mathbf{v}_i = \lambda_i^{-1} \mathbf{A}^\top \mathbf{u}_i$. Alternatively, the $\mathbf{v}_i$ are the eigenvectors of $\mathbf{A}^\top \mathbf{A}$ with the same nonzero eigenvalues λ_i^2.

The matrix expansion for the singular-value decomposition written in terms of the full dimensional matrices $\mathbf{U}$,$\mathbf{V}$,$\mathbf{\Lambda}$ is

$$\underset{(m\times k)}{\mathbf{A}} = \underset{(m\times m)}{\mathbf{U}} \underset{(m\times k)}{\Lambda} \underset{(k\times k)}{\mathbf{V}^\top},$$

where $\mathbf{U}$ has m orthogonal eigenvectors of $\mathbf{A}\mathbf{A}^\top$ as its columns, $\mathbf{V}$ has k orthogonal eigenvectors of $\mathbf{A}^\top \mathbf{A}$ as its columns, and $\mathbf{\Lambda}$ is specified in Result 2A.15.

Result 2A.16. *Error of Approximation.* Let $\mathbf{A}$ be an $m \times k$ matrix of real numbers with $m \geq k$ and singular value decomposition $\mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top$. Let $s < k = \text{rank}\,(\mathbf{A})$. Then

$$\mathbf{B} = \sum_{i=1}^{s} \lambda_i \mathbf{u}_i \mathbf{v}_i^\top$$

is the rank s least squares approximation to $\mathbf{A}$. It minimizes

$$\text{tr}\left[(\mathbf{A} - \mathbf{B})(\mathbf{A} - \mathbf{B})^\top\right]$$

over all $m \times k$ matrices $\mathbf{B}$ having rank no greater than s. The minimum value, or error of approximation, is $\sum_{i=s+1}^{k} \lambda_i^2$.

Proof. To establish this result, we use $\mathbf{U}\mathbf{U}^\top = \mathbf{I}_m$ and $\mathbf{V}\mathbf{V}^\top = \mathbf{I}_k$ to write the sum of

squares as

$$\begin{aligned}
\operatorname{tr}\left[(\mathbf{A}-\mathbf{B})(\mathbf{A}-\mathbf{B})^\top\right] &= \operatorname{tr}\left[\mathbf{U}\mathbf{U}^\top(\mathbf{A}-\mathbf{B})\mathbf{V}\mathbf{V}^\top(\mathbf{A}-\mathbf{B})^\top\right] \\
&= \operatorname{tr}\left[\mathbf{U}^\top(\mathbf{A}-\mathbf{B})\mathbf{V}\mathbf{V}^\top(\mathbf{A}-\mathbf{B})^\top\mathbf{U}\right] \\
&= \operatorname{tr}\left[(\mathbf{\Lambda}-\mathbf{C})(\mathbf{\Lambda}-\mathbf{C})^\top\right] \\
&= \sum_{i=1}^{m}\sum_{j=1}^{k}(\lambda_{ij}-c_{ij})^2 \\
&= \sum_{i=1}^{m}(\lambda_i-c_{ij})^2 + \sum\sum_{i\neq j} c_{ij}^2,
\end{aligned}$$

where $\mathbf{C} = \mathbf{U}^\top\mathbf{B}\mathbf{V} = (c_{ij})$. Clearly, the minimum occurs when $c_{ij} = 0$ for $i \neq j$ and $c_{ii} = \lambda_i$ for the s largest singular values; the other $c_{ii} = 0$. That is, $\mathbf{U}\mathbf{B}\mathbf{V}^\top = \Lambda_s$ or $\mathbf{B} = \sum_{i=1}^{s}\lambda_i\mathbf{u}_i\mathbf{v}_i^\top$.

□

Result 12.1. The term $\mathbf{r}\mathbf{c}^\top$ is common (the meaning of "common" is explained right below the proof of this result) to the approximation $\hat{\mathbf{P}}$ whatever the $I \times J$ correspondence matrix $\mathbf{P}$ in (18).

The reduced rank s approximation to $\mathbf{P}$ is a minimizer of the sum of squares (21). It is given by

$$\mathbf{P} \doteq \sum_{k=1}^{s}\tilde{\lambda}_k\left(\mathbf{D}_r^{1/2}\tilde{\mathbf{u}}_k\right)\left(\mathbf{D}_c^{1/2}\tilde{\mathbf{v}}_k\right) = \mathbf{r}\mathbf{c}^\top + \sum_{k=2}^{s}\tilde{\lambda}_k\left(\mathbf{D}_r^{1/2}\tilde{\mathbf{u}}_k\right)\left(\mathbf{D}_c^{1/2}\tilde{\mathbf{v}}_k\right)^\top,$$

where the $\tilde{\lambda}_k$ are the singular values and the $I \times 1$ vectors $\tilde{\mathbf{u}}_k$ and the $J \times 1$ vectors $\tilde{\mathbf{v}}_k$ are the corresponding singular vectors of the $I \times J$ matrix $\mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1/2}$. In view of Result 2A.16, the minimum value of (21) is $\sum_{k=s+1}^{J}\tilde{\lambda}_k^2$.

The reduced rank $K > 1$ approximation to $\mathbf{P} - \mathbf{r}\mathbf{c}^\top$ is

$$\mathbf{P} - \mathbf{r}\mathbf{c}^\top \doteq \sum_{k=1}^{K}\lambda_k\left(\mathbf{D}_r^{1/2}\mathbf{u}_k\right)\left(\mathbf{D}_c^{1/2}\mathbf{v}_k\right)^\top,$$

where the λ_k are the singular values and the $I \times 1$ vectors $\mathbf{u}_k$ and the $J \times 1$ vectors $\mathbf{v}_k$ are the corresponding singular vectors of the $I \times J$ matrix $\mathbf{D}_r^{-1/2}\left(\mathbf{P} - \mathbf{r}\mathbf{c}^\top\right)\mathbf{D}_c^{-1/2}$. And $\lambda_k = \tilde{\lambda}_{k+1}$, $\mathbf{u}_k = \tilde{\mathbf{u}}_{k+1}$, and $\mathbf{v}_k = \tilde{\mathbf{v}}_{k+1}$ for $k = 1, \ldots, J-1$.

Proof. First we consider a scaled version $\mathbf{B} = \mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1/2}$ of the correspondence matrix $\mathbf{P}$. According to Result 2A.16, the best row rank $= s$ approximation $\hat{\mathbf{B}}$ to $\mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1/2}$ is given by the first s terms in the singular-value decomposition

$$\mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1/2} = \sum_{k=1}^{J} \tilde{\lambda}_k \tilde{\mathbf{u}}_k \tilde{\mathbf{v}}_k^\top,$$

where

$$\mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1/2}\tilde{\mathbf{v}}_k = \tilde{\lambda}_k \tilde{\mathbf{u}}_k \qquad \tilde{\mathbf{u}}_k^\top \mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1/2} = \tilde{\lambda}_k \tilde{\mathbf{v}}_k^\top, \tag{23}$$

and

$$\left| \left(\mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1/2}\right)\left(\mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1/2}\right)^\top - \tilde{\lambda}_k^2 \mathbf{I} \right| = 0 \qquad \text{for} \qquad k = 1, \ldots, J.$$

The approximation to $\mathbf{P}$ is therefore given by

$$\hat{\mathbf{P}} = \mathbf{D}_r^{1/2}\hat{\mathbf{B}}\mathbf{D}_c^{1/2} \doteq \sum_{k=1}^{s} \tilde{\lambda}_k \left(\mathbf{D}_r^{1/2}\tilde{\mathbf{u}}_k\right)\left(\mathbf{D}_c^{1/2}\tilde{\mathbf{v}}_k\right)^\top$$

and, by Result 2.16, the error of approximation is $\sum_{k=s+1}^{J} \tilde{\lambda}_k^2$.

□

Whatever the correspondence matrix $\mathbf{P}$, the term $\mathbf{r}\mathbf{c}^\top$always provides the best rank one approximation; thus, it is common to the approximation $\hat{\mathbf{P}}$. This corresponds to the assumption of Independence of the rows and columns of $\mathbf{P}$. Let $\hat{\mathbf{u}}_1 = \mathbf{D}_r^{1/2}\mathbf{1}_I$ and $\tilde{\mathbf{v}}_1 = \mathbf{D}_c^{1/2}\mathbf{1}_J$, where $\mathbf{1}_I$ is $I \times 1$ and $J \times 1$ vector of $1's$. We can show that (22) holds for these choices. Indeed, we have

$$\begin{aligned}
\tilde{\mathbf{u}}_1^\top \left(\mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1/2}\right) &= \left(\mathbf{D}_r^{1/2}\mathbf{1}_I\right)^\top \left(\mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1/2}\right) \\
&= \mathbf{1}_I^\top \mathbf{P}\mathbf{D}_c^{-1/2} = \mathbf{c}^\top \mathbf{D}_c^{-1/2} \\
&= (\sqrt{c_1}, \ldots, \sqrt{c_J}) = \left(\mathbf{D}_c^{1/2}\mathbf{1}_J\right)^\top = \tilde{\mathbf{v}}_1^\top,
\end{aligned}$$

and

$$\begin{aligned}
\left(\mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1/2}\right)\tilde{\mathbf{v}}_1 &= \left(\mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1/2}\right)\left(\mathbf{D}_c^{1/2}\mathbf{1}_J\right) \\
&= \mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{1}_J = \mathbf{D}_r^{-1/2}\mathbf{r} \\
&= (\sqrt{r_1}, \ldots, \sqrt{r_I})^\top = \mathbf{D}_r^{1/2}\mathbf{1}_I = \tilde{\mathbf{u}}_1.
\end{aligned}$$

That is,

$$(\tilde{\mathbf{u}}_1, \tilde{\mathbf{v}}_1) = \left(\mathbf{D}_r^{1/2}\mathbf{1}_I, \mathbf{D}_c^{1/2}\mathbf{1}_J\right)$$

are singular vectors associated with singular value $\tilde{\lambda}_1 = 1$. For any correspondence matrix $\mathbf{P}$, the common term in every expansion is

$$\mathbf{D}_r^{1/2}\mathbf{u}_1\mathbf{v}_1^\top\mathbf{D}_c^{1/2} = \mathbf{D}_r\mathbf{1}_I\mathbf{1}_J^\top\mathbf{D}_c = \mathbf{r}\mathbf{c}^\top.$$

Therefore, we have established the first approximation and (22) can always be expressed as

$$\mathbf{P} = \mathbf{r}\mathbf{c}^\top + \sum_{k=2}^{J} \tilde{\lambda}_k \left(\mathbf{D}_r^{1/2}\tilde{\mathbf{u}}_k\right)\left(\mathbf{D}_c^{1/2}\tilde{\mathbf{v}}_k\right)^\top.$$

Because of the common term, the problem can be rephrased in terms of $\mathbf{P} - \mathbf{r}\mathbf{c}^\top$ and its scaled version $\mathbf{D}_r^{-1/2}\left(\mathbf{P} - \mathbf{r}\mathbf{c}'\right)\mathbf{D}_c^{-1/2}$. By the orthogonality of the singular vectors of $\mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1/2}$, we have $\tilde{\mathbf{u}}_k^\top\left(\mathbf{D}_r^{1/2}\mathbf{1}_I\right) = 0$ and $\tilde{\mathbf{v}}_k^\top\left(\mathbf{D}_\mathbf{c}^{1/2}\mathbf{1}_J\right) = 0$ for $k > 1$, so

$$\mathbf{D}_r^{-1/2}\left(\mathbf{P} - \mathbf{r}\mathbf{c}^\top\right)\mathbf{D}_c^{-1/2} = \sum_{k=2}^{J} \tilde{\lambda}_k\tilde{\mathbf{u}}_k\tilde{\mathbf{v}}_k^\top$$

is the singular-value decomposition of $\mathbf{D}_r^{-1/2}\left(\mathbf{P} - \mathbf{r}\mathbf{c}^\top\right)\mathbf{D}_c^{-1/2}$ in terms of the singular values and vectors obtained from $\mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1/2}$. Converting the singular values and vectors λ_k, $\mathbf{u}_k$, and $\mathbf{v}_k$ from $\mathbf{D}_r^{-1/2}\left(\mathbf{P} - \mathbf{r}\mathbf{c}^\top\right)\mathbf{D}_c^{-1/2}$ only amounts to changing k to $k-1$ so $\lambda_k = \lambda_{k+1}$, $\mathbf{u}_k = \tilde{\mathbf{u}}_{k+1}$, and $\mathbf{v}_k = \tilde{\mathbf{v}}_{k+1}$ for $k = 1,\ldots, J-1$.

In terms of the singular value decomposition for $\mathbf{D}_r^{-1/2}\left(\mathbf{P} - \mathbf{r}\mathbf{c}^\top\right)\mathbf{D}_c^{-1/2}$, the expansion for $\mathbf{P} - \mathbf{r}\mathbf{c}^\top$ takes the form

$$\mathbf{P} - \mathbf{r}\mathbf{c}^\top = \sum_{k=1}^{J-1} \lambda_k \left(\mathbf{D}_r^{1/2}\mathbf{u}_k\right)\left(\mathbf{D}_c^{1/2}\mathbf{v}_k\right)^\top.$$

The best rank K approximation to $\mathbf{D}_r^{-1/2}\left(\mathbf{P} - \mathbf{r}\mathbf{c}^\top\right)\mathbf{D}_c^{-1/2}$ is given by $\sum_{k=1}^{K}\lambda_k\mathbf{u}_k\mathbf{v}_k^\top$. Then the best approximation to $\mathbf{P} - \mathbf{r}\mathbf{c}^\top$ is

$$\mathbf{P} - \mathbf{r}\mathbf{c}^\top \doteq \sum_{k=1}^{K} \lambda_k \left(\mathbf{D}_r^{1/2}\mathbf{u}_k\right)\left(\mathbf{D}_c^{1/2}\mathbf{v}_k\right)^\top. \tag{24}$$

Remark. Note that the vectors $\mathbf{D}_r^{1/2}\mathbf{u}_k$ and $\mathbf{D}_c^{1/2}\mathbf{v}_k$ in the expansion (24) of $\mathbf{P} - \mathbf{r}\mathbf{c}^\top$

need not have length one but satisfy the scaling

$$\left(\mathbf{D}_r^{1/2}\mathbf{u}_k\right)^\top \mathbf{D}_r^{-1}\left(\mathbf{D}_r^{1/2}\mathbf{u}_k\right) = \mathbf{u}_k^\top \mathbf{u}_k = 1,$$
$$\left(\mathbf{D}_c^{1/2}\mathbf{v}_k\right)^\top \mathbf{D}_c^{-1}\left(\mathbf{D}_c^{1/2}\mathbf{v}_k\right) = \mathbf{v}_k^\top \mathbf{v}_k = 1.$$

Because of this scaling, the expansion in Result 12.1 is called a *generalized singular-value decomposition*.

The preceding approach is called the *matrix approximation method* and the approach to follow the *profile approximation method*. We will also discuss the profile approximation method using the row profiles that are the rows of the matrix $\mathbf{D}_r^{-1}\mathbf{P}$. Contingency analysis can also be defined as the approximation of the row profiles by points in a low-dimensional space. The row profiles are the row proportions that are calculated from the counts in the contingency table; the value of each cell in the row profiles is the count of the cell divided by the sum of the counts for the entire row; the row profiles for each row sum to approximately 1 (or 100%). Suppose the row profiles are approximated by the matrix $\mathbf{P}^* = \left(p_{ij}^*\right)$. Using the square-root matrices $\mathbf{D}_r^{1/2}$ and $\mathbf{D}_c^{1/2}$ defined in (19) and (20), we have

$$\left(\mathbf{D}_r^{-1}\mathbf{P} - \mathbf{P}^*\right)\mathbf{D}_c^{-1/2} = \mathbf{D}_r^{-1/2}\left(\mathbf{D}_r^{-1/2}\mathbf{P} - \mathbf{D}_r^{1/2}\mathbf{P}^*\right)\mathbf{D}_c^{-1/2}$$

and, with $p_{ij}^* = \frac{\hat{p}_{ij}}{r_i}$ the least squares criterion (20) can be written as

$$\begin{aligned}\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(p_{ij}-\hat{p}_{ij}\right)^2}{r_i c_j} &= \sum_{i=1}^{I} r_i \sum_{j=1}^{J}\frac{\left(p_{ij}/r_i - p_{ij}^*\right)^2}{c_j} \\ &= \text{tr}\left[\left[\left(\mathbf{Q}\right)\mathbf{D}_c^{-1/2}\right]\mathbf{D}_r^{1/2}\left[\left(\mathbf{Q}\right)\mathbf{D}_r^{-1/2}\right]^\top\right],\end{aligned}$$

where $\mathbf{Q} = \mathbf{D}_r^{-1/2}\mathbf{P} - \mathbf{D}_r^{1/2}\mathbf{P}^*$.

The proof of Result 12.1 deals with the minimization of the last expression for the trace, and $\mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1/2}$ has the singular-value decomposition

$$\mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1/2} = \sum_{k=1}^{J}\tilde{\lambda}_k\tilde{\mathbf{u}}_k\tilde{\mathbf{v}}_k^\top. \tag{25}$$

The best rank K approximation can be obtained by the first K terms of this expansion. Since $\mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1/2}$ can be approximated by $\mathbf{D}_r^{1/2}\mathbf{P}^*\mathbf{D}_c^{1/2}$, we can left multiply each side of (25) by $\mathbf{D}_r^{-1/2}$ and right multiply each side of (24) by $\mathbf{D}_c^{1/2}$ to obtain the

generalized singular-value decomposition

$$\mathbf{D}_r^{-1}\mathbf{P} = \sum_{k=1}^{J} \tilde{\lambda}_k \mathbf{D}_r^{-1/2}\tilde{\mathbf{u}}_k \left(\mathbf{D}_c^{1/2}\tilde{\mathbf{v}}_k\right)^\top,$$

where $(\tilde{\mathbf{u}}_1, \tilde{\mathbf{v}}_1) = \left(\mathbf{D}_r^{1/2}\mathbf{1}_I, \mathbf{D}_c^{1/2}\mathbf{1}_J\right)$, are singular vectors associated with singular value $\lambda_1 = 1$, where $\mathbf{1}_I$ is a $I \times 1$ and $\mathbf{1}_J$ is a $J \times 1$ vector of $1's$. Because $\mathbf{D}_r^{-1/2}\left(\mathbf{D}_r^{1/2}\mathbf{1}_I\right) = \mathbf{1}_I$ and $\left(\mathbf{D}_c^{1/2}\mathbf{1}_J\right)^\top \mathbf{D}_c^{1/2} = \mathbf{c}^\top$, the leading term in the above decomposition is $\mathbf{1}_I\mathbf{c}^\top$. Therefore, in terms of the singular values and vectors from $\mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{D}_c^{-1/2}$, the reduced rank $K < J$ approximation of the row profiles of $\mathbf{D}_r^{-1}\mathbf{P}$ is

$$\mathbf{P}^* \doteq \mathbf{1}_I\mathbf{c}^\top + \sum_{k=2}^{K} \tilde{\lambda}_k \mathbf{D}_r^{-1/2}\tilde{\mathbf{u}}_k \left(\mathbf{D}_c^{1/2}\tilde{\mathbf{v}}_k\right)^\top.$$

The above formula provides one of the mathematical foundations for statistical softwares to carry out the algorithm of correspondence analysis. In practice, the correspondence analysis can be conveniently implemented by using statistical software (see Example 7 below).

6.2 Inertia

Total inertia, defined as the weighted sum of squares below, is a measure of the variation in the count data:

$$\text{tr}\left[\mathbf{D}_r^{-1/2}\left(\mathbf{P} - \mathbf{r}\mathbf{c}^\top\right)\mathbf{D}_c^{-1/2}\left(\mathbf{D}_r^{-1/2}\left(\mathbf{P} - \mathbf{r}\mathbf{c}^\top\right)\mathbf{D}_c^{-1/2}\right)^\top\right] = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(p_{ij} - r_i c_j\right)^2}{r_i c_j} = \sum_{k=1}^{J-1}\lambda_k^2,$$

where λ_k are the singular values obtained from the single-value decomposition of $\mathbf{D}_r^{-1/2}\left(\mathbf{P} - \mathbf{r}\mathbf{c}^\top\right)\mathbf{D}_c^{-1/2}$ (see the proof of Result 12.1).

The inertia associated with the best reduced rank $K < J$ approximation to the centered matrix $\mathbf{P} - \mathbf{r}\mathbf{c}^\top$ (the K-dimensional solution) has inertia $\sum_{k=1}^{K}\lambda_k^2$. The residual inertia (variation) not accounted for by the rank K solution is equal to the sum of squares of the remaining singular values: $\lambda_{K+1}^2 + \lambda_{K+2}^2 + \ldots + \lambda_{J-1}^2$.

The total inertia is related to the chi-square measure of association in a two-way contingency table, $\chi^2 = \sum_{i,j}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$. Here $O_{ij} = x_{ij}$ is the observed frequency and E_{ij} is the expected frequency for the (i,j)th cell. In the correspondence analysis, it is

assumed that the row variable is independent of the column variable, $E_{ij} = nr_ic_j$, and

$$\text{Total inertia} = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(p_{ij} - r_ic_j)^2}{r_ic_j} = \frac{\chi^2}{n}.$$

The total inertia of all the rows in a contingency table is equal to the χ^2 statistic divided by n which is also known as Pearson's *mean-square contingency* used to assess the quality of its graphical representation in correspondence analysis.

6.3 Interpretation of Correspondence Analysis in Two Dimensions

Geometrically, we interpret a large value for the proportion $\frac{(\lambda_1^2+\lambda_2^2)}{\sum_{k=1}^{J-1}\lambda_k^2}$ as the associations in the centered data well represented by points in a plane, and this best approximating plane accounting for nearly all the variation in the data beyond that accounted for by the rank 1 solution (independence model). Algebraically, the approximation can be written as

$$\mathbf{P} - \mathbf{r}\mathbf{c}^\top \doteq \lambda_1\mathbf{u}_1\mathbf{v}_1^\top + \lambda_2\mathbf{u}_2\mathbf{v}_2^\top.$$

6.4 Example

Example 7 (Exercise 12.20 Johnson and Wichern [7]). A sample of $n = 1660$ people is cross-classified according to mental health status and socioeconomics status in Table 18. Perform a correspondence analysis of these data. Interpret the results. Can the associations in the data be well represented in one dimension?

Table 18

	A	B	C	D	E
Well	121	57	72	36	21
Mild Symptoms	188	105	141	97	71
Moderate Symptoms	112	65	77	54	54
Impaired	86	60	94	78	71

Solution: To perform correspondence analysis, we use the SAS software. The obtained results are shown on Figure 10.

Inertia and Chi-Square Decomposition

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent
0.16132	0.02602	43.1992	94.75	94.75
0.03709	0.00138	2.2833	5.01	99.76
0.00820	0.00007	0.1115	0.24	100.00
	0.02747	45.5941	100.00	

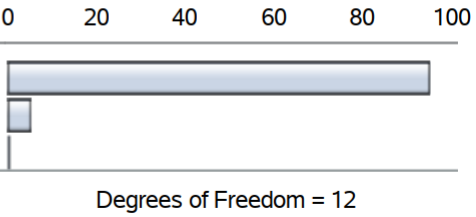


Figure 10: Inertia and Chi-Square Decomposition

According to the above analysis, the singular values are 0.16132, 0.03709, and 0.00820; the principal inertia are 0.02602, 0.00138, and 0.00007. The row coordinates and column coordinates are shown on Figure 11:

Row Coordinates

	Dim1	Dim2
Well	0.2597	-0.0133
MildSymp	0.0295	-0.0226
Moderate	-0.0142	0.0700
Impaired	-0.2373	-0.0197

Column Coordinates

	Dim1	Dim2
A	0.1829	0.0155
B	0.0590	0.0224
C	-0.0089	-0.0423
D	-0.1654	-0.0433
E	-0.2877	0.0619

Figure 11: Row Coordinates and Column Coordinates

A correspondence analysis plot of the mental health-socioeconomic data is shown on Figure 12 below. Hence, the lowest economic class (A and B in the plot) is located between moderate and impaired; the lower class (C and D) is closest to impaired.

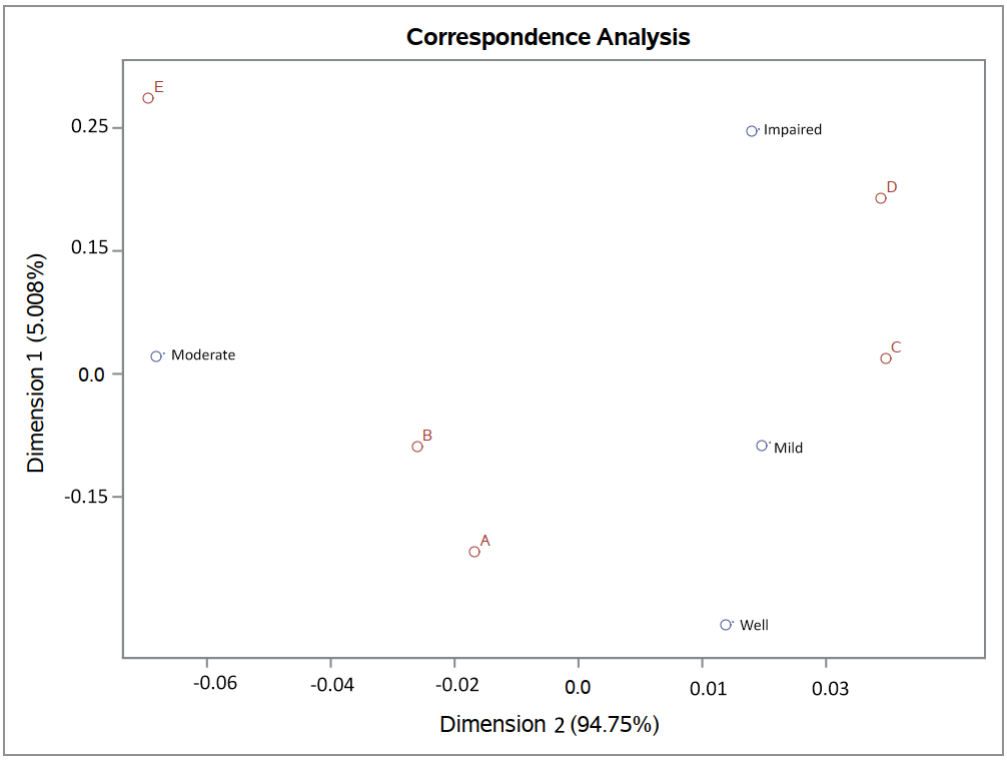


Figure 12: A Correspondence Analysis Plot of the Mental Health-socioeconomic Data

Correspondence analysis designed to represent associations in a low-dimensional space is primarily a graphical technique. But it is also relevant to principal component analysis and canonical correlation analysis.

7 Conclusion

Cluster analysis combines numerical methods of classification; it appears in many disciplines such as biology, botany, medicine, psychology, geography, marketing, image processing, psychiatry, archaeology, etc. Some examples of applications of cluster analysis appeared in the Introduction of the project. Cluster analysis concerns about searching for patterns in a data set by grouping the multivariate observations into clusters in order to find an optimal grouping by which the observations or objects within each cluster are similar but the clusters are dissimilar to each other. The goal of cluster analysis is to find the "natural groupings" in the data set that make sense to the researcher. Clustering items (units or cases) employs the measure of proximity by some sort of distance; grouping variables utilizes the measure of proximity by correlation coefficients or like measures of association.

The project has discussed some statistical theory behind several clustering methods related to hierarchical classification, nonhierachical partition, and graphical representation.

First, hierarchical clustering is conducted by either a series of successive mergers or a series of successive divisions. *Agglomerative hierarchical methods* begin with the individual objects; so there are as many clusters as objects initially. They first group the most similar objects, and then these groups are merged according to their similarities. Eventually, all subgroups are merged into a single cluster. *Divisive hierarchical methods* initially divide all objects into two subgroups in which the objects in one subgroup are "far from" the objects in the other group. The subgroups are further divided into dissimilar subgroups until each object forms a group itself. The results of both methods could be depicted in a *dendrogram* that illustrates the mergers or divisions made at successive levels. Our focus was on agglomerative hierarchical procedures, and particularly, *linkage methods* which are suitable for clustering items as well as variables. In turn, *single linkage* (minimum distance or nearest neighbor), *complete linkage* (maximum distance or farthest neighbor), and *average linkage* (average distance) were discussed.

Second, nonhierachical clustering methods aim to group *items* instead of *variables* into K clusters, the number K of which can be either determined beforehand or during the clustering procedure. Theses methods can be utilized to much larger data sets than do hierarchical methods because a matrix of distances (similarities) does not have to be determined, and none of the basic data have to be stored during the computation. These methods start from the nuclei of clusters which can be formed from either an initial partition of items into groups or an initial set of seed points. Starting configurations should be chosen without overt biases, i.e., randomly select seed points from the items or randomly partition the items into initial groups. We discuss the K-means method, one of the most popular nonhierarchical procedures.

Third, statistical model-based classification has made major advances in clustering techniques because statistical models explain how the collection of $(p \times 1)$ measurements $\mathbf{x}_j$, from the N objects, was generated. We discussed the most common model where cluster k has expected proportion p_k of the objects and the corresponding measurements are generated by a probability density function $f_k(\mathbf{x})$. If there are K clusters, the *mixing distribution* is used to model the observation vector for a single object:

$$f_{Mix}(\mathbf{x}) = \sum_{k=1}^{K} p_k \, f_k(\mathbf{x}), \qquad \mathbf{x} \in \mathbb{R}^p,$$

where each $p_k \geq 0$ and $\sum_{k=1}^{K} p_k = 1$. Because the observation comes from the com-

ponent distribution $f_k(\mathbf{x})$ with probability p_k, the distribution $f_{Mix}(\mathbf{x})$ is a mixture of the K distributions $f_1(\mathbf{x})$, $f_2(\mathbf{x})$, ... , $f_K(\mathbf{x})$. Therefore, the collection of N observation vectors generated from this distribution should be a mixture of observations from the component distributions. Statistical model-based classification involves both hierarchical or nonhierarchinal procedures.

Finally, a graphical procedure that represents associations in a table of frequencies is called correspondence analysis. This project has focused on two-way table of frequencies or *contingency table*. Suppose the contingency table has I rows and J columns, the plot is produced by correspondence analysis containing two sets of points: a set of I points corresponding to the rows and a set of J points corresponding to the columns. Associations between the points is reflected by the positions of the points on the plot. Row points being close together show rows having similar profiles (conditional distributions) across the columns; column points being close together show columns having similar profiles (conditional distributions) down the rows; and row points being close to column points demonstrates combinations occurring more frequently than would be expected from an independence mode in which the row categories are unrelated to the column categories. The usual output from a correspondence analysis provides the "best" two-dimensional representation of the data, the coordinates of the plotted points, and the *inertia* of the amount of information retained in each dimension.

Due to the scope of the project, it does not confront the difficult problem of cluster analysis: cluster validation or avoiding artefactual solutions to cluster analysis. The preceding discussion of clustering methods reveals that no one clustering method can be judged to be the "best" in all circumstances. However, it might be advisable to apply a number of clustering methods to the data set before reaching any conclusion. If all produce very similar solutions, it is justifiably concluded that the results are worthy of further investigation; otherwise, it might be evidence against any clustering solution. In this project, we have not only discussed in detail some common clustering techniques, but also illustrated their use by numerical examples with real-life data sets.

References

[1] R. E. Bonner. "On Some Clustering Techniques". In: *IBM Journal of Research and Development* 8.1 (Jan. 1964), pp. 22–32.

[2] C. Chakrapani. *Statistics in Market Research*. London: Arnold, 2004.

[3] A. D. Gordon. *Classification*. 2nd edition. Boca Raton, FL: Chapman and Hall/CRC, 1999.

[4] P. E. Green, R. E. Frank, and P. J. Robinson. "Cluster Analysis in Test Market Selection". In: *Management Science* 13.8 (1967), pp. 387–400.

[5] F. R. Hodson. *Numerical typology and prehistoric archaeology. In: Mathematics in the Archaeological and Historical Sciences*. Ed. by P. A. Tautu F. R. Hodson D. G. Kendall. Edinburgh: Edinburgh University Press, 1971.

[6] F. R. Hodson, P. H. A. Sneath, and J. E. Doran. "Some Experiments in the Numerical Analysis of Archaeological Data". In: *Biometrika* 53.3–4 (1966), pp. 311–324.

[7] R. A. Johnson and D. W. Wichern. *Applied Multivariate Statistical Analysis*. 6th edition. Upper Saddle River, NJ: Pearson, Mar. 2007.

[8] J. MacQueen. "Some methods for classification and analysis of multivariate observations". In: *5th Berkeley Symposium on Mathematical Statistics and Probability*. 1967, pp. 281–297.

[9] T. Montmerle. "Hertzsprung Russell Diagram". In: Berlin, Heidelberg: Springer, 2011, pp. 749–754.

[10] S. Pinker. *How the Mind Works*. New York, NY: W. W. Norton & Company, 2009.

[11] J. H. Ward. "Hierarchical Grouping to Optimize an Objective Function". In: *Journal of the American Statistical Association.* 58.301 (1963), pp. 236–244.

8 Appendix I

Table 11.5: the Iris data

5.1, 3.5, 1.4, 0.2, setosa
4.9, 3.0, 1.4, 0.2, setosa
4.7, 3.2, 1.3, 0.2, setosa
4.6, 3.1, 1.5, 0.2, setosa
5.0, 3.6, 1.4, 0.2, setosa
5.4, 3.9, 1.7, 0.4, setosa
4.6, 3.4, 1.4, 0.3, setosa
5.0, 3.4, 1.5, 0.2, setosa
4.4, 2.9, 1.4, 0.2, setosa
4.9, 3.1, 1.5, 0.1, setosa
5.4, 3.7, 1.5, 0.2, setosa
4.8, 3.4, 1.6, 0.2, setosa
4.8, 3.0, 1.4, 0.1, setosa
4.3, 3.0, 1.1, 0.1, setosa
5.8, 4.0, 1.2, 0.2, setosa
5.7, 4.4, 1.5, 0.4, setosa
5.4, 3.9, 1.3, 0.4, setosa
5.1, 3.5, 1.4, 0.3, setosa
5.7, 3.8, 1.7, 0.3, setosa
5.1, 3.8, 1.5, 0.3, setosa
5.4, 3.4, 1.7, 0.2, setosa
5.1, 3.7, 1.5, 0.4, setosa
4.6, 3.6, 1.0, 0.2, setosa
5.1, 3.3, 1.7, 0.5, setosa
4.8, 3.4, 1.9, 0.2, setosa
5.0, 3.0, 1.6, 0.2, setosa
5.0, 3.4, 1.6, 0.4, setosa
5.2, 3.5, 1.5, 0.2, setosa
5.2, 3.4, 1.4, 0.2, setosa
4.7, 3.2, 1.6, 0.2, setosa
4.8, 3.1, 1.6, 0.2, setosa
5.4, 3.4, 1.5, 0.4, setosa
5.2, 4.1, 1.5, 0.1, setosa

5.5, 4.2, 1.4, 0.2, setosa
4.9, 3.1, 1.5, 0.1, setosa
5.0, 3.2, 1.2, 0.2, setosa
5.5, 3.5, 1.3, 0.2, setosa
4.9, 3.1, 1.5, 0.1, setosa
4.4, 3.0, 1.3, 0.2, setosa
5.1, 3.4, 1.5, 0.2, setosa
5.0, 3.5, 1.3, 0.3, setosa
4.5, 2.3, 1.3, 0.3, setosa
4.4, 3.2, 1.3, 0.2, setosa
5.0, 3.5, 1.6, 0.6, setosa
5.1, 3.8, 1.9, 0.4, setosa
4.8, 3.0, 1.4, 0.3, setosa
5.1, 3.8, 1.6, 0.2, setosa
4.6, 3.2, 1.4, 0.2, setosa
5.3, 3.7, 1.5, 0.2, setosa
5.0, 3.3, 1.4, 0.2, setosa
7.0, 3.2, 4.7, 1.4, versicolor
6.4, 3.2, 4.5, 1.5, versicolor
6.9, 3.1, 4.9, 1.5, versicolor
5.5, 2.3, 4.0, 1.3, versicolor
6.5, 2.8, 4.6, 1.5, versicolor
5.7, 2.8, 4.5, 1.3, versicolor
6.3, 3.3, 4.7, 1.6, versicolor
4.9, 2.4, 3.3, 1.0, versicolor
6.6, 2.9, 4.6, 1.3, versicolor
5.2, 2.7, 3.9, 1.4, versicolor
5.0, 2.0, 3.5, 1.0, versicolor
5.9, 3.0, 4.2, 1.5, versicolor
6.0, 2.2, 4.0, 1.0, versicolor
6.1, 2.9, 4.7, 1.4, versicolor
5.6, 2.9, 3.6, 1.3, versicolor
6.7, 3.1, 4.4, 1.4, versicolor
5.6, 3.0, 4.5, 1.5, versicolor
5.8, 2.7, 4.1, 1.0, versicolor
6.2, 2.2, 4.5, 1.5, versicolor

5.6, 2.5, 3.9, 1.1, versicolor
5.9, 3.2, 4.8, 1.8, versicolor
6.1, 2.8, 4.0, 1.3, versicolor
6.3, 2.5, 4.9, 1.5, versicolor
6.1, 2.8, 4.7, 1.2, versicolor
6.4, 2.9, 4.3, 1.3, versicolor
6.6, 3.0, 4.4, 1.4, versicolor
6.8, 2.8, 4.8, 1.4, versicolor
6.7, 3.0, 5.0, 1.7, versicolor
6.0, 2.9, 4.5, 1.5, versicolor
5.7, 2.6, 3.5, 1.0, versicolor
5.5, 2.4, 3.8, 1.1, versicolor
5.5, 2.4, 3.7, 1.0, versicolor
5.8, 2.7, 3.9, 1.2, versicolor
6.0, 2.7, 5.1, 1.6, versicolor
5.4, 3.0, 4.5, 1.5, versicolor
6.0, 3.4, 4.5, 1.6, versicolor
6.7, 3.1, 4.7, 1.5, versicolor
6.3, 2.3, 4.4, 1.3, versicolor
5.6, 3.0, 4.1, 1.3, versicolor
5.5, 2.5, 4.0, 1.3, versicolor
5.5, 2.6, 4.4, 1.2, versicolor
6.1, 3.0, 4.6, 1.4, versicolor
5.8, 2.6, 4.0, 1.2, versicolor
5.0, 2.3, 3.3, 1.0, versicolor
5.6, 2.7, 4.2, 1.3, versicolor
5.7, 3.0, 4.2, 1.2, versicolor
5.7, 2.9, 4.2, 1.3, versicolor
6.2, 2.9, 4.3, 1.3, versicolor
5.1, 2.5, 3.0, 1.1, versicolor
5.7, 2.8, 4.1, 1.3, versicolor
6.3, 3.3, 6.0, 2.5, virginica
5.8, 2.7, 5.1, 1.9, virginica
7.1, 3.0, 5.9, 2.1, virginica
6.3, 2.9, 5.6, 1.8, virginica
6.5, 3.0, 5.8, 2.2, virginica

7.6, 3.0, 6.6, 2.1, virginica
4.9, 2.5, 4.5, 1.7, virginica
7.3, 2.9, 6.3, 1.8, virginica
6.7, 2.5, 5.8, 1.8, virginica
7.2, 3.6, 6.1, 2.5, virginica
6.5, 3.2, 5.1, 2.0, virginica
6.4, 2.7, 5.3, 1.9, virginica
6.8, 3.0, 5.5, 2.1, virginica
5.7, 2.5, 5.0, 2.0, virginica
5.8, 2.8, 5.1, 2.4, virginica
6.4, 3.2, 5.3, 2.3, virginica
6.5, 3.0, 5.5, 1.8, virginica
7.7, 3.8, 6.7, 2.2, virginica
7.7, 2.6, 6.9, 2.3, virginica
6.0, 2.2, 5.0, 1.5, virginica
6.9, 3.2, 5.7, 2.3, virginica
5.6, 2.8, 4.9, 2.0, virginica
7.7, 2.8, 6.7, 2.0, virginica
6.3, 2.7, 4.9, 1.8, virginica
6.7, 3.3, 5.7, 2.1, virginica
7.2, 3.2, 6.0, 1.8, virginica
6.2, 2.8, 4.8, 1.8, virginica
6.1, 3.0, 4.9, 1.8, virginica
6.4, 2.8, 5.6, 2.1, virginica
7.2, 3.0, 5.8, 1.6, virginica
7.4, 2.8, 6.1, 1.9, virginica
7.9, 3.8, 6.4, 2.0, virginica
6.4, 2.8, 5.6, 2.2, virginica
6.3, 2.8, 5.1, 1.5, virginica
6.1, 2.6, 5.6, 1.4, virginica
7.7, 3.0, 6.1, 2.3, virginica
6.3, 3.4, 5.6, 2.4, virginica
6.4, 3.1, 5.5, 1.8, virginica
6.0, 3.0, 4.8, 1.8, virginica
6.9, 3.1, 5.4, 2.1, virginica
6.7, 3.1, 5.6, 2.4, virginica

6.9, 3.1, 5.1, 2.3, virginica
5.8, 2.7, 5.1, 1.9, virginica
6.8, 3.2, 5.9, 2.3, virginica
6.7, 3.3, 5.7, 2.5, virginica
6.7, 3.0, 5.2, 2.3, virginica
6.3, 2.5, 5.0, 1.9, virginica
6.5, 3.0, 5.2, 2.0, virginica
6.2, 3.4, 5.4, 2.3, virginica
5.9, 3.0, 5.1, 1.8, virginica

9 Appendix II

R code for Example 6:

```
BIC <- mclustBIC(X)
mod <- Mclust(X, x = BIC) summary(mod, parameters = TRUE)
plot(mod, what = "classification")
table(class, mod$classification)
plot(mod, what = "uncertainty")
```

SAS code for Example 7:

```
data T12_14;
input Mental_Health_status A B C D E;
cards;
Well 121 57 72 36 21
MildSymptoms 188 105 141 97 71
ModerateSymptoms 112 65 77 54 54
Impaired 86 60 94 78 71
run;
proc corresp data=T12_14 out=E12_14 short;
var A B C D E;
id Mental_Health_status;
run;
```