Chapter 15
Cluster Analysis and Correspondence Analysis

15.1. Introduction

We will employ the same notations as in the previous chapters. Lower-case letters $x, y, \ldots$ will denote real scalar variables, whether mathematical or random. Capital letters $X, Y, \ldots$ will be used to denote real matrix-variate mathematical or random variables, whether square or rectangular matrices are involved. A tilde will be placed on top of letters such as $\tilde{x}, \tilde{y}, \tilde{X}, \tilde{Y}$ to denote variables in the complex domain. Constant matrices will for instance be denoted by A, B, C. A tilde will not be used on constant matrices unless the point is to be stressed that the matrix is in the complex domain. The determinant of a square matrix A will be denoted by $|A|$ or $\det(A)$ and, in the complex case, the absolute value or modulus of the determinant of A will be denoted as $|\det(A)|$. When matrices are square, their order will be taken as $p \times p$, unless specified otherwise. When A is a full rank matrix in the complex domain, then AA^* is Hermitian positive definite where an asterisk designates the complex conjugate transpose of a matrix. Additionally, $\mathrm{d}X$ will indicate the wedge product of all the distinct differentials of the elements of the matrix X. Thus, letting the $p \times q$ matrix $X = (x_{ij})$ where the x_{ij}'s are distinct real scalar variables, $\mathrm{d}X = \wedge_{i=1}^{p} \wedge_{j=1}^{q} \mathrm{d}x_{ij}$. For the complex matrix $\tilde{X} = X_1 + iX_2,\ i = \sqrt{(-1)}$, where X_1 and X_2 are real, $\mathrm{d}\tilde{X} = \mathrm{d}X_1 \wedge \mathrm{d}X_2$.

15.1.1. Clusters

A cluster means a group or a cloud of items close together with reference to one or more characteristics. For instance, in a countryside, there are villages which are clusters of houses. In a city, there are clusters of high-rise buildings or clusters of apartment blocks. If we have 2-dimensional data points marked on a sheet of paper, then there may be several places where the points are grouped together in large crowds, at other places the points may be bunched together in smaller clumps and somewhere else, there may be singleton points. In a classification problem, we have a number of preassigned populations and we

A. M. Mathai et al., *Multivariate Statistical Analysis in the Real and Complex Domains*,
https://doi.org/10.1007/978-3-030-95864-0_15

want to assign a point at hand to one of those populations. This cannot be achieved in the context of cluster analysis as we do not know beforehand how many clusters there are in the data at hand or which data point belongs to which cluster. Cluster analysis is akin to pattern recognition whereas classification is a sort of taxonomy. Suppose that a new plant is to be classified as belonging to one of the known species of plants; if it does not fall into any of the known species, then we have a member from a new species. In cluster analysis, we are, in a manner of speaking, going to create various 'species'. To start with, we have only a cloud of items and we do not know how many categories or clusters there exist.

Cluster analysis techniques are widely utilized in many fields such as psychiatry, sociology, anthropology, archeology, medicine, criminology, engineering and geology, to mention only a few areas. If real scalar variables are to be classified as belonging to a certain category, one way of achieving this is to ascertain their joint dispersion or joint variation as measured in terms of scale-free covariance or correlation. Those variables that are similarly correlated may be grouped together.

We will consider the problem of cluster analysis involving n points $X_1, \ldots, X_n$ where each X_j is a real p-dimensional vector, that is, we have a $p \times n$ data matrix

$$\mathbf{X} = [X_1, X_2, \ldots, X_n] = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1n} \\ x_{21} & x_{22} & \ldots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p1} & x_{p2} & \ldots & x_{pn} \end{bmatrix}. \tag{15.1.1}$$

15.1.2. Distance measures

Two real p-vectors are close together if the "distance" between them is small. Many types of distance measures can be defined. Let X_r and X_s be two real p-vectors. These are the r-th and s-th members or columns in the data matrix (15.1.1). Then, the following are some distance measures:

$$d_m(X_r, X_s) = \Big[\sum_{i=1}^{p} |x_{ir} - x_{is}|^m\Big]^{\frac{1}{m}};$$

for $m = 2$, we have the Euclidean distance $d_2(X_r, X_s) = [\sum_{i=}^{p} |x_{ir} - x_{is}|^2]^{\frac{1}{2}}$, or, denoting d_2^2 as d^2, we have

$$d^2(X_r, X_s) = \sum_{i=1}^{p}(x_{ir} - x_{is})^2, \tag{15.1.2}$$

where the absolute value sign can be replaced by parentheses since we are dealing with real elements. We will utilize this convenient quantity d^2 for comparing observation vectors. There may be joint variation or covariances among the coordinates in each of the vectors, in which case, $\mathrm{Cov}(X_r) = \Sigma > O$. If all the X_j's, $j = 1, \ldots, n$, have the same covariance matrix, then $\mathrm{Cov}(X_j) = \Sigma,\ j = 1, \ldots, n$, and a statistician might wish to consider the generalized distance between X_r and X_s, or its square, $d^2_{(g)}(X_r, X_s) = (X_r - X_s)' \Sigma^{-1}(X_r - X_s)$, the subscript g designating the generalized distance. Since Σ is unknown, we may wish to estimate it. However, if there are clusters, it may not be appropriate to make use of the entire data set of all n points, since the joint variation or the covariance within each cluster is likely to be different. And as we do not know beforehand whether clusters are present, securing a proper estimate of Σ turns out to prove problematic. As a result, this problem is usually circumvented by resorting to the ordinary Euclidean distance instead of the generalized distance.

Let us examine the effect of scaling a vector. If the unit of measurement in one vector is changed, what will be the effect on the squared distance? Consider the following vectors:

$$X_1 = \begin{bmatrix} -1 \\ 0 \\ -2 \end{bmatrix} \text{ and } X_2 = \begin{bmatrix} -3 \\ 2 \\ 4 \end{bmatrix} \Rightarrow d^2(X_1, X_2) = (X_1 - X_2)'(X_1 - X_2)$$
$$= [(-1) - (-3)]^2 + [(0) - (2)]^2 + [(-2) - (4)]^2 = 44.$$

The squared distances between the vectors when (1) X_1 is multiplied by 2; (2) X_2 is multiplied by 2; (3) X_1 and X_2 are each multiplied by 2, are

$$\begin{aligned} d^2(2X_1, X_2) &= (-2+3)^2 + (0-2)^2 + (-4-4)^2 = 69 \\ d^2(X_1, 2X_2) &= (-1+6)^2 + (0-4)^2 + (-2-8)^2 = 141 \\ d^2(2X_1, 2X_2) &= 4[(X_1 - X_2)'(X_1 - X_2)] = 4 \times 44 = 176. \end{aligned}$$

Note that they are fully distorted as $69 \neq 4(44)$ and $141 \neq 4(44)$. Thus, the scaling of individual vectors can fully alter the nature of the clusters when there are clusters in the original data. As well, members of the original clusters need not be members of the same clusters in the scaled data and the number of clusters may also change. Accordingly, it is indeed inadvisable to make use of the generalized distance. Nor is re-scaling the individual vectors a good idea if we are seeking clusters. Accordingly, the recommended procedure consists of utilizing the original data without modifying them. It may also happen that the components in each p-vector are recorded in different units of measurement. Then, how to eliminate the location and scale effect on the components in each vector? This can be achieved by standardizing them individually, that is, by subtracting the average value of the components from the components of each vector and dividing the result by the sample standard deviation. Let us see what happens in the case of our numerical example. Letting $\bar{x}_1$ and $\bar{x}_2$ be the averages of the components in X_1 and X_2, and s_1^2 and s_2^2 be the associated sums of products, we have

$$\begin{aligned}
\bar{x}_1 &= \frac{1}{3}[(-1)+(0)+(-2)] = -1,\ \bar{x}_2 = \frac{1}{3}[(-3)+(2)+(4)] = 1,\\
s_1^2 &= \sum_{i=1}^{p}(x_{i1}-\bar{x}_1)^2 = [(-1)-(-1)]^2+[(0)-(-1)]^2+[(-2)-(-1)]^2 = 2,\\
s_2^2 &= \sum_{i=1}^{p}(x_{i2}-\bar{x}_2)^2 = 26.
\end{aligned}$$

Thus, the standardized vectors X_1 and X_2, denoted by Y_1 and Y_2, are the following:

$$Y_1 = \frac{\sqrt{3}}{\sqrt{2}}\begin{bmatrix}-1-(-1)\\ 0-(-1)\\ -2-(-1)\end{bmatrix} = \frac{\sqrt{3}}{\sqrt{2}}\begin{bmatrix}0\\ 1\\ -1\end{bmatrix} \text{ and } Y_2 = \frac{\sqrt{3}}{\sqrt{26}}\begin{bmatrix}-4\\ 1\\ 3\end{bmatrix},$$

and $d^2(Y_1, Y_2) = (Y_1-Y_2)'(Y_1-Y_2) = 7.6641$. However, Y_1 and Y_2 are very distorted and the distance between X_1 and X_2 is also modified. Hence, such procedures will change the clustering aspect as well, with new clusters possibly differing from the original clusters.

Let us consider the matrix of squared distances, denoted by D:

$$D = \begin{bmatrix}0 & d_{12}^2 & \dots & d_{1n}^2\\ d_{21}^2 & 0 & \dots & d_{2n}^2\\ \vdots & \vdots & \ddots & \vdots\\ d_{n1}^2 & d_{n2}^2 & \dots & d_{nn}^2\end{bmatrix} = D'. \tag{15.1.3}$$

For example, letting

$$X_1 = \begin{bmatrix}1\\ 0\\ -1\end{bmatrix},\ X_2 = \begin{bmatrix}2\\ 1\\ 3\end{bmatrix},\ X_3 = \begin{bmatrix}0\\ 1\\ 2\end{bmatrix} \text{ and } X_4 = \begin{bmatrix}3\\ 1\\ 2\end{bmatrix},$$

we have $d_{12}^2 = (1-2)^2 + (0-1)^2 + (-1-3)^2 = 18$, $d_{13}^2 = 11$, $d_{14}^2 = 14$, $d_{23}^2 = 5$, $d_{24}^2 = 2$, $d_{34}^2 = 9$, so that

$$D = \begin{bmatrix} 0 & 18 & 11 & 14 \\ 18 & 0 & 5 & 2 \\ 11 & 5 & 0 & 9 \\ 14 & 2 & 9 & 0 \end{bmatrix}.$$

The question of interest is the following: Given a set of n vectors of order p, how can one determine the number of clusters and then, classify them into these clusters?

15.2. Different Methods of Clustering

The main methods are hierarchical in nature, the other ones being non-hierarchical. We will begin with non-hierarchical techniques. In this category, the most popular one involves optimization or partitioning.

15.2.1. Optimization or partitioning

With this approach, we have to come up with two numbers: k, a probable number of clusters, and r, the maximum separation between the members of each prospective cluster. Based on the distances or on the dissimilarity matrix, D, one should be able to determine the likely number of clusters, that is, k. Then, one has to find a set of k vectors among the n given vectors, which will be taken as seed members or starting members within the k potential clusters. Several methods have been proposed for determining this k, including the following:

1. Examine the closeness of the original vectors as indicated by the dissimilarity matrix D and, to start with, decide on an initial numbers for k and the likely distance between members within a cluster denoted by r.

2. Examine the original data points or original p-vectors and, based on the comparative magnitudes of the components of the observed p-vectors, ascertain whether there is any grouping possible and predict a value for each of k and r.

3. Evaluate the sample sum of products matrix S from the original data matrix. Compute the two main principal components associated with this S. Substitute X_j, the j-th observation vector, in the two principal components. This provides a pair of numbers or one point in a two-dimensional space. Compute n such points for $j = 1, \ldots, n$. Plot these points. From the graph, assess the clustering pattern, the number k of possible clusters, estimates for r, the maximum distance between two members within a cluster as well as the minimum distance between the clusters.

4. Choose any number k, select k vectors at random from the set of n vectors; then, preselect a number r and use it as a measure of maximum separation between vectors.

5. Take any number k and select as seed vectors the first k vectors whose separation is at least two units among the set of n vectors.

6. Look at the farthest points. Select k of them that are separated by at least r units for preselected values of k and r.

If the dissimilarity matrix D is utilized, then the separation number r must be measured in d_{ij}^2 units, whereas r should be in d_{ij} units if the actual distances d_{ij} are used. After the seed vectors are selected, the remaining $n - k$ points are to be associated to these seed points to form clusters. Assign the vectors closest to each of the seed vectors and form the initial k clusters of two or more vectors. For example, if there are three closest members at equal distance to a seed vector then, that cluster comprises 4 members, including the seed vector. Then, compute the centroids of all initial clusters. The centroid of a cluster is the simple average of the vectors included in that cluster. Thus, the centroid is a p-vector. Then, measure the distances of all the points belonging to the same cluster from each centroid, and incorporate all points within the distance of r from a centroid to that cluster. This process will create the second stage of k clusters. Now, evaluate the centroid of each of these k clusters. Again, repeat the process of computing the distances of all points from each centroid. If a member in a cluster is found to be closer to the centroid of another cluster than to its own cluster's centroid, then redirect that vector to the cluster to which it belongs. Rearrange all vectors in such a manner, assigning each one to a cluster whose centroid is the closest. Note that the number k can increase or decrease in the course of this process. Continue the procedure until no more improvement is possible. At this stage, that final k is the number of clusters in the data and the final members in each cluster are set. This procedure is also called *k-means approach*.

This k-means approach has a serious shortcoming: if one starts with a different set of seed vectors, then it is possible to end up with a different set of final clusters. On the other hand, this method has the appreciable advantage that it allows a member provisionally assigned to a cluster to be moved to another cluster where it really belongs, that is, it allows the transfer of points. The following example should clarify the procedure.

Example 15.2.1. Ten volunteers are given an exercise routine in an experiment that monitors systolic pressure, diastolic pressure and heart beat. These are measured after adhering to the exercise routine for four weeks. The data entries are systolic pressure minus

120 (SP), diastolic pressure minus 80 (DP) and heart beat minus 60 (HB), where 120, 80 and 60 are taken as the standard readings of systolic pressure, diastolic pressure and heart beat, respectively. Carry out a cluster analysis of the data. The data matrix is the following where (1), ..., (10) represent the data vectors $A_1, \ldots, A_{10}$ for the 10 volunteers, the first row represents SP, the second row, DP, and the third, HB:

↓→	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
SP :	0	1	1	2	3	4	6	8	5	10
DP :	1	0	−1	3	2	3	8	10	6	8
HB :	−1	−1	−1	−2	5	2	7	8	9	4

Solution 15.2.1. Let us compute the dissimilarity matrix D:

$$D = \left[\begin{array}{ccccccccccc}
\downarrow\rightarrow & (1) & (2) & (3) & (4) & (5) & (6) & (7) & (8) & (9) & (10) \\
(1) & 0 & 2 & 5 & 9 & 46 & 29 & 149 & 226 & 150 & 174 \\
(2) & 2 & 0 & 1 & 11 & 44 & 27 & 153 & 230 & 152 & 170 \\
(3) & 5 & 1 & 0 & 18 & 49 & 34 & 170 & 251 & 165 & 187 \\
(4) & 9 & 11 & 18 & 0 & 51 & 20 & 122 & 185 & 139 & 125 \\
(5) & 46 & 44 & 49 & 51 & 0 & 11 & 49 & 98 & 36 & 86 \\
(6) & 29 & 27 & 34 & 20 & 11 & 0 & 54 & 101 & 59 & 65 \\
(7) & 149 & 153 & 170 & 122 & 49 & 54 & 0 & 9 & 9 & 25 \\
(8) & 226 & 230 & 251 & 185 & 98 & 101 & 9 & 0 & 26 & 24 \\
(9) & 150 & 152 & 165 & 139 & 36 & 59 & 9 & 26 & 0 & 54 \\
(10) & 174 & 170 & 187 & 125 & 86 & 65 & 25 & 24 & 54 & 0
\end{array}\right].$$

The data matrix suggests the possibility of three clusters. Accordingly, we may begin with the vectors A_2, A_5 and A_8 as seed vectors and take the separation width as $r = 15$ units. From D, we find $d^2_{23} = 1$, the smallest number, and hence A_2 and A_3 form the cluster: $\{A_2, A_3\}$. Note that $d^2_{56} = 11$, so that A_6 and A_5 form the cluster: $\{A_5, A_6\}$. Since $d^2_{87} = 9$, A_7 and A_8 form a cluster: $\{A_7, A_8\}$. Now, consider the centroids. Letting C_{11}, C_{21} and C_{31} denote the centroids, $C_{11} = \frac{1}{2}(A_2 + A_3)$, $C_{21} = \frac{1}{2}(A_5 + A_6)$ and $C_{31} = \frac{1}{2}(A_7 + A_8)$, that is,

$$C_{11} = \begin{bmatrix} 1 \\ -1/2 \\ -1 \end{bmatrix}, \quad C_{21} = \begin{bmatrix} 7/2 \\ 5/2 \\ 7/2 \end{bmatrix} \text{ and } C_{31} = \begin{bmatrix} 7 \\ 9 \\ 15/2 \end{bmatrix}.$$

Let us calculate the distances of $A_1,\ldots,A_{10}$ from C_{11}, C_{21}, C_{31}:

$$d^2(C_{11},A_1)=\frac{13}{4},\ d^2(C_{11},A_2)=\frac{1}{4},\ d^2(C_{11},A_3)=\frac{1}{4},\ d^2(C_{11},A_4)=\frac{57}{4},$$
$$d^2(C_{11},A_5)=\frac{185}{2},\ d^2(C_{11},A_6)=\frac{121}{4},\ d^2(C_{11},A_7)=\frac{645}{4},\ d^2(C_{11},A_8)=\frac{961}{4},$$
$$d^2(C_{11},A_9)=\frac{633}{4},\ d^2(C_{11},A_{10})=\frac{713}{4},\ d^2(C_{21},A_1)=\frac{139}{4},\ d^2(C_{21},A_2)=\frac{131}{4},$$
$$d^2(C_{21},A_3)=\frac{155}{4},\ d^2(C_{21},A_4)=\frac{131}{4},\ d^2(C_{21},A_5)=\frac{11}{4},\ d^2(C_{21},A_6)=\frac{11}{4},$$
$$d^2(C_{21},A_7)=\frac{195}{4},\ d^2(C_{21},A_8)=\frac{387}{4},\ d^2(C_{21},A_9)=\frac{179}{4},\ d^2(C_{21},A_{10})=\frac{291}{4},$$
$$d^2(C_{31},A_1)=\frac{741}{4},\ d^2(C_{31},A_2)=\frac{757}{4},\ d^2(C_{31},A_3)=\frac{833}{4},\ d^2(C_{31},A_4)=\frac{605}{4},$$
$$d^2(C_{31},A_5)=\frac{285}{4},\ d^2(C_{31},A_6)=\frac{301}{4},\ d^2(C_{31},A_7)=\frac{9}{4},\ d^2(C_{31},A_8)=\frac{9}{4},$$
$$d^2(C_{31},A_9)=\frac{61}{4},\ d^2(C_{31},A_{10})=\frac{89}{4}.$$

We include all the points located within 15 units of distance to the nearest cluster. Then, the second set of clusters are the following: Cluster 1: $\{A_1, A_2, A_3, A_4\}$, cluster 2: $\{A_5, A_6\}$, cluster 3: $\{A_7, A_8\}$. Note that A_9 is quite close to Cluster 3. We may either include it in Cluster 3 or treat it as a singleton. Since the next stage calculations do not change the composition of the clusters, we may take the final clusters as $\{A_1, A_2, A_3, A_4\}$, $\{A_5, A_6\}$, $\{A_7, A_8, A_9\}$ and $\{A_{10}\}$ where Cluster 4 consists of a single element. This completes the computations.

Let us examine the principal components of the sample sum of products matrix and plot the points to see whether any cluster can be detected. The sample matrix denoted by $\mathbf{X}$ and the sample average, denoted by $\bar{X}$, are the following:

$$\mathbf{X}=\begin{bmatrix}0&1&1&2&3&4&6&8&5&10\\1&0&-1&3&2&3&8&10&6&8\\-1&-1&-1&-2&5&2&7&8&9&4\end{bmatrix},\quad \bar{X}=\begin{bmatrix}4\\4\\3\end{bmatrix}.$$

Let the matrix of sample averages be $\bar{\mathbf{X}}=[\bar{X},\bar{X},\ldots,\bar{X}]$ and the deviation matrix be $\mathbf{X}_d=\mathbf{X}-\bar{\mathbf{X}}$. Then,

$$\mathbf{X}_d=\begin{bmatrix}-4&-3&-3&-2&-1&0&2&4&1&6\\-3&-4&-5&-1&-2&-1&4&6&2&4\\-4&-4&-4&-5&2&-1&4&5&6&1\end{bmatrix},$$

and the sample sum of products matrix is $S = \mathbf{X}_d\mathbf{X}_d'$, that is,

$$S = \begin{bmatrix} 96 & 101 & 88 \\ 101 & 128 & 112 \\ 88 & 112 & 156 \end{bmatrix}.$$

The eigenvalues of S are $\lambda_1 = 330.440$, $\lambda_2 = 40.522$ and $\lambda_3 = 9.039$. An eigenvector corresponding to $\lambda_1 = 330.440$ and an eigenvector corresponding to $\lambda_2 = 40.522$, respectively denoted by U_1 and U_2 are the following:

$$U_1 = \begin{bmatrix} 0.782 \\ 0.943 \\ 1.000 \end{bmatrix} \ \text{and} \ U_2 = \begin{bmatrix} -0.676 \\ -0.500 \\ -1.000 \end{bmatrix}.$$

Then the first two principal components are $U_1'Y$ and $U_2'Y$ with $Y' = [y_1, y_2, y_3]$. We substitute our sample points $A_1, \ldots, A_{10}$ to obtain 10 pairs of numbers. For example,

$$U_1'A_1 = [0.782, 0.943, 1]\begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} = -0.057,\ U_2'A_1 = [-0.676, -0.5, 1]\begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} = -1.5$$

and hence the first pair of numbers or the first point is $P_1 : (-0.057, -1.500)$. Similar calculations yield the remaining 9 points as: $P_2 : (-0.218, -1.676)$, $P_3 : (-1.161, -1.176)$, $P_4 : (2.393, -4.852)$, $P_5 : (9.232, 1.972)$, $P_6 : (7.957, -2.204)$, $P_7 : (19.236, -1.056)$, $P_8 : (23.686, -2.408)$, $P_9 : (18.568, 2.620)$, $P_{10} : (19.364, -6.760)$. It is seen that these points which are plotted in Fig. 15.2.2 form the same clusters as the original points shown in Fig. 15.2.1, that is, Cluster 1: $\{A_1, A_2, A_3, A_4\}$; Cluster 2: $\{A_5, A_6\}$; Cluster 3: $\{A_7, A_8, A_9\}$; Cluster 4: $\{A_{10}\}$.

Other non-hierarchical methods are currently in use. We will mention these procedures later, after discussing the main hierarchical technique known as *single linkage* or nearest neighbor method.

15.3. Hierarchical Methods of Clustering

Hierarchical procedures are of two categories. In one of them, we begin with all the n data points as n different clusters of one element each. Then, by applying certain rules, we start combining these single-member clusters into larger clusters, the process being halted when a desired number of clusters are obtained. If the process is continued, we ultimately end up with a single cluster containing all of the n points. In the second category, we initially consider one cluster that comprises the n elements. We then start splitting this cluster into two clusters by making use of some criteria. Next, one or both of these subclusters are divided again by applying the same criteria. If the process is continued, we

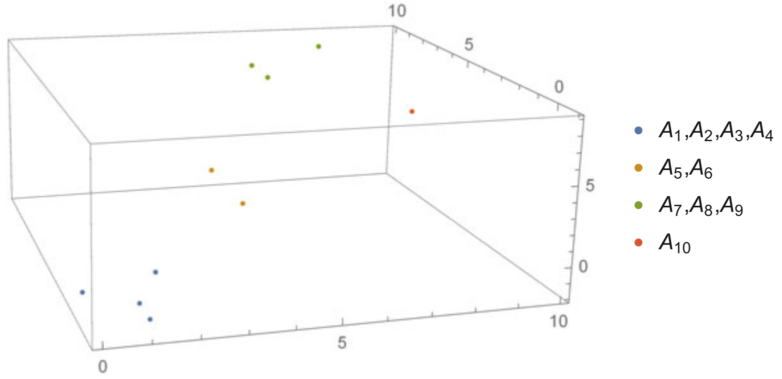

Figure 15.2.1 The original 10 data points

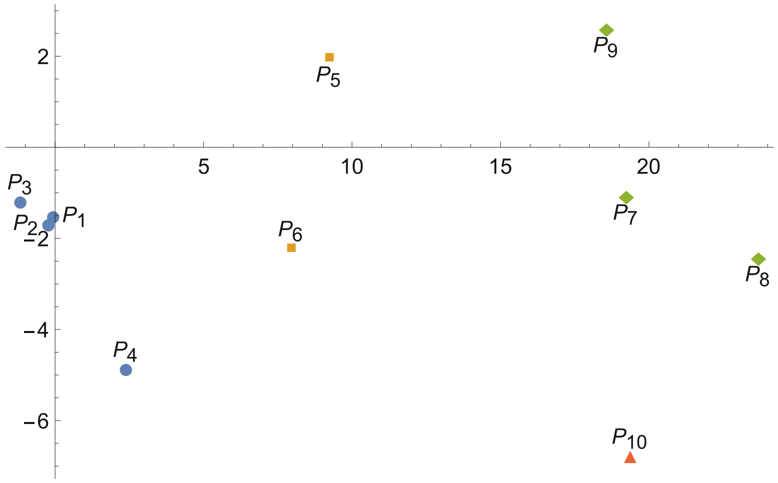

Figure 15.2.2 Second versus first principal component evaluated at the A_i's

finally end up with n clusters of one element each. The process is halted when a desired number of clusters are obtained. In all these procedures, one cannot objectively determine when to stop the process or how many distinct clusters are present. We have to specify some stopping rules as a means of selecting a suitable number of clusters.

15.3.1. Single linkage or nearest neighbor method

In this single linkage procedure, we begin by assuming that there are n clusters consisting of one item each. We then combine these clusters by applying a minimum distance rule. At the initial stage, we have only one element in each 'cluster', but at the following steps, each cluster will potentially contain several items and hence, the rule is stated for general clusters. Consider two clusters A and B whose elements are denoted by X_j and

Y_j, that is, $X_j \in A$ and $Y_j \in B$, the X_j's and Y_j's being p-vectors belonging to the data set at hand. In the minimum distance rule, we define the distance between two clusters, denoted by $d(A, B)$, as follows:

$$d(A, B) = \min\{d(X_i, Y_j),\ \text{for all } X_i \in A, Y_j \in B\}. \tag{15.3.1}$$

This distance is measured in the units of the definition of the distance being utilized. We will illustrate the single linkage hierarchical procedure by making use of the data set provided in Example 15.2.1 and its associated dissimilarity matrix D. We will utilize the dissimilarity matrix D to represent various "distances". Since the matrix D will be repeatedly referred to at every stage, it is duplicated next for ready reference:

$$D = \left[\begin{array}{ccccccccccc}
\downarrow\rightarrow & (1) & (2) & (3) & (4) & (5) & (6) & (7) & (8) & (9) & (10) \\
(1) & 0 & 2 & 5 & 9 & 46 & 29 & 149 & 226 & 150 & 174 \\
(2) & 2 & 0 & 1 & 11 & 44 & 27 & 153 & 230 & 152 & 170 \\
(3) & 5 & 1 & 0 & 18 & 49 & 34 & 170 & 251 & 165 & 187 \\
(4) & 9 & 11 & 18 & 0 & 51 & 20 & 122 & 185 & 139 & 125 \\
(5) & 46 & 44 & 49 & 51 & 0 & 11 & 49 & 98 & 36 & 86 \\
(6) & 29 & 27 & 34 & 20 & 11 & 0 & 54 & 101 & 59 & 65 \\
(7) & 149 & 153 & 170 & 122 & 49 & 54 & 0 & 9 & 9 & 25 \\
(8) & 226 & 230 & 251 & 185 & 98 & 101 & 9 & 0 & 26 & 24 \\
(9) & 150 & 152 & 165 & 139 & 36 & 59 & 9 & 26 & 0 & 54 \\
(10) & 174 & 170 & 187 & 125 & 86 & 65 & 25 & 24 & 54 & 0
\end{array}\right].$$

To start with, we have 10 clusters $\{A_j\},\ j = 1, \ldots, 10$. At the initial stage, each cluster has one element. Then $d(A, B)$ as defined in (15.3.1) is the smallest distance (dissimilarity) appearing in D, that is, 1 which occurs between the elements corresponding to A_2 and A_3. These two clusters of one vector each are combined and replaced by B_1 by taking the smaller entries in each column of the combined representation of the dissimilarity measures corresponding A_2 and A_3. For illustration, we now list the dissimilarity measures corresponding to the original A_2 and A_3 and the new B_1 as rows:

$$\begin{array}{rlllllllll}
A_2: & (2) & [0\quad 1] & (11) & (44) & (27) & (153) & (230) & (152) & (170) \\
A_3: & 5 & [1\quad 0] & 18 & 49 & 34 & 170 & 251 & 165 & 187 \\
B_1: & 2 & [0] & 11 & 44 & 27 & 153 & 230 & 152 & 170
\end{array}$$

The rows representing A_2 and A_3 are combined and replaced by B_1 as shown above. The second and third columns in D are combined into one column, namely, the B_1 column. The elements in B_1 are the smaller elements in each column of A_2 and A_3. The bracketed elements in A_2 and A_3, namely $[0, 1]$ and $[1, 0]$, are combined into one element $[0]$ in B_1, the updated dissimilarity matrix having one fewer row and one fewer column. These are the intersections of the two rows and columns. Other smaller elements in the two original columns, which make up B_1, are displayed in parentheses. This process will be repeated at each stage. At the first stage of the procedure, we end up with 9 clusters: $C_1 = \{A_2, A_3\}, \{A_j\},\ j = 1, 4, \dots, 10$, the resulting configuration of the dissimilarity matrix, denoted by D_1, being

$$D_1 = \left[\begin{array}{cccccccccc}
\downarrow\rightarrow & A_1 & B_1 & A_4 & A_5 & A_6 & A_7 & A_8 & A_9 & A_{10} \\
A_1 & 0 & 2 & 9 & 46 & 29 & 149 & 226 & 150 & 174 \\
B_1 & 2 & 0 & 11 & 44 & 27 & 153 & 230 & 152 & 170 \\
A_4 & 9 & 11 & 0 & 51 & 20 & 122 & 185 & 139 & 125 \\
A_5 & 46 & 44 & 51 & 0 & 11 & 49 & 98 & 36 & 86 \\
A_6 & 29 & 27 & 20 & 11 & 0 & 54 & 101 & 59 & 65 \\
A_7 & 149 & 153 & 122 & 49 & 54 & 0 & 9 & 9 & 25 \\
A_8 & 226 & 230 & 185 & 98 & 101 & 9 & 0 & 26 & 24 \\
A_9 & 150 & 152 & 139 & 36 & 59 & 9 & 26 & 0 & 54 \\
A_{10} & 174 & 170 & 125 & 86 & 65 & 25 & 24 & 54 & 0
\end{array}\right].$$

Now, the next smallest dissimilarity is 2 which occurs at (A_1, B_1). Thus, the rows (columns) corresponding to A_1 and B_1 are combined into one row (column) B_2. The original rows corresponding to A_1 and B_1 and the new row corresponding to B_2 are the following:

$$\begin{array}{rcccccccc}
A_1: & [0\ \ 2] & (9) & 46 & 29 & (149) & (226) & (150) & 174 \\
B_1: & [2\ \ 0] & 11 & (44) & (27) & 153 & 230 & 152 & (170) \\
B_2: & [0] & 9 & 44 & 27 & 149 & 226 & 150 & 170\,.
\end{array}$$

The new configuration, denoted by D_2, is the following:

$$D_2 = \left[\begin{array}{ccccccccc}
\downarrow\rightarrow & B_2 & A_4 & A_5 & A_6 & A_7 & A_8 & A_9 & A_{10} \\
B_2 & 0 & 9 & 44 & 27 & 149 & 226 & 150 & 170 \\
A_4 & 9 & 0 & 51 & 20 & 122 & 185 & 139 & 125 \\
A_5 & 44 & 51 & 0 & 11 & 49 & 98 & 36 & 86 \\
A_6 & 27 & 20 & 11 & 0 & 54 & 101 & 59 & 65 \\
A_7 & 149 & 122 & 49 & 54 & 0 & 9 & 9 & 25 \\
A_8 & 226 & 185 & 98 & 101 & 9 & 0 & 26 & 24 \\
A_9 & 150 & 139 & 36 & 59 & 9 & 26 & 0 & 54 \\
A_{10} & 170 & 125 & 86 & 65 & 25 & 24 & 54 & 0
\end{array}\right],$$

the resulting clusters being $C_2 = \{A_1, A_2, A_3\}, \{A_j\},\ j = 4, \dots, 10$. The next smallest dissimilarity is 9, which occurs at (B_2, A_4). Hence these are combined, that is, the first two columns (rows) are merged as explained. The combined row, denoted by B_3, is the following, its transpose becoming the first column:

$$B_3 = [0, 44, 20, 122, 185, 139, 125],$$

and the new configuration is the following:

$$D_3 = \left[\begin{array}{cccccccc} \downarrow\rightarrow & B_3 & A_5 & A_6 & A_7 & A_8 & A_9 & A_{10} \\ B_3 & 0 & 44 & 20 & 122 & 185 & 139 & 125 \\ A_5 & 44 & 0 & 11 & 49 & 98 & 36 & 86 \\ A_6 & 20 & 11 & 0 & 54 & 101 & 59 & 65 \\ A_7 & 122 & 49 & 54 & 0 & 9 & 9 & 25 \\ A_8 & 185 & 98 & 101 & 9 & 0 & 26 & 24 \\ A_9 & 139 & 36 & 59 & 9 & 26 & 0 & 54 \\ A_{10} & 125 & 86 & 65 & 25 & 24 & 54 & 0 \end{array}\right].$$

At this stage, the clusters are $C_2 = \{A_1, A_2, A_3, A_4\},\ \{A_j\},\ j = 5, \ldots, 10$. The next smallest number is 9, which is occurring at (A_7, A_8), (A_7, A_9). Accordingly, we combine A_7, A_8 and A_9, and the resulting configuration is the following where the resultant of the replacement rows (columns) is denoted by B_4:

$$D_4 = \left[\begin{array}{cccccc} \downarrow\rightarrow & B_3 & A_5 & A_6 & B_4 & A_{10} \\ B_3 & 0 & 44 & 20 & 122 & 125 \\ A_5 & 44 & 0 & 11 & 36 & 86 \\ A_6 & 20 & 11 & 0 & 54 & 65 \\ B_4 & 122 & 36 & 54 & 0 & 24 \\ A_{10} & 125 & 86 & 65 & 24 & 0 \end{array}\right],$$

the clusters being $C_2 = \{A_1, A_2, A_3, A_4\},\ C_3 = \{A_7, A_8, A_9\},\ \{A_i\},\ i = 5, 6, 10$. The next smallest dissimilarity measure is 11 at (A_5, A_6). Combining these, the replacement row is $B_5 = [20, 0, 36, 65]$, and the new configuration, denoted by D_5 is as follows:

$$D_5 = \left[\begin{array}{ccccc} \downarrow\rightarrow & B_3 & B_5 & B_4 & A_{10} \\ B_3 & 0 & 20 & 122 & 125 \\ B_5 & 20 & 0 & 36 & 65 \\ B_4 & 122 & 36 & 0 & 24 \\ A_{10} & 125 & 65 & 24 & 0 \end{array}\right],$$

the resulting clusters being $C_2 = \{A_1, A_2, A_3, A_4\},\ C_3 = \{A_7, A_8, A_9\}$, $C_4 = \{A_5, A_6\}$, $C_5 = \{A_{10}\}$.

We may stop at this stage since the clusters obtained from the other methods coincide with C_2, C_3, C_4, C_5. At the following step of the procedure, C_4 would combine with C_3, with the next final stage resulting in a single cluster that would encompass all 10 points.

15.3.2. Average linking as a modified distance measure

An alternative distance measure involving all the items in pairs of clusters is considered in this subsection. As one proceeds from any stage to the next one in a hierarchical procedure, a decision is based on the next smallest distance between two clusters. At the initial stage, this does not pose any problem since the dissimilarity matrix D is available and each cluster contains only a single element. However, further on in the process, as there are several elements in the clusters, a more suitable definition of "distance" is required in order to proceed to the next stage. Several types of methods have been proposed in the literature. One such procedure is the *average linkage method* under which the distance between two clusters A and B, denoted again by $d(A, B)$, is defined as follows:

$$d(A, B) = \frac{1}{n_1 n_2}\sum_{j=1}^{n_2}\sum_{i=1}^{n_1} d(X_i, Y_j) \text{ for all } X_i \in A, Y_j \in B \tag{15.3.2}$$

where the X_i's and Y_j's are all p-vectors from the given set of data points. In this case, the rule being applied is that two clusters having the smallest distance, as measured in terms of (15.3.2), are combined before initiating the next stage.

15.3.3. The centroid method

In a hierarchical single linkage procedure, another way of determining the distance between two clusters before proceeding to the next stage is referred to as the centroid method under which the Euclidean distance between the centroids of clusters A and B is defined as follows:

$$d(A, B) = d(\bar{X}, \bar{Y}) \text{ with } \bar{X} = \frac{1}{n_1}\sum_{j=1}^{n_1} X_j \text{ and } \bar{Y} = \frac{1}{n_2}\sum_{j=1}^{n_2} Y_j, \tag{15.3.3}$$

where $\bar{X}$ is the centroid of the cluster A and $\bar{Y}$ is the centroid of the cluster B, $X_i \in A,\ i = 1, \ldots, n_1,\ Y_j \in B,\ j = 1, \ldots, n_2$. In this case, the process involves combining two clusters with the smallest $d(A, B)$ as specified in (15.3.3) into a single cluster. After combining them, or equivalently, after taking the union of A and B, the centroid of the combined cluster, denoted by $\bar{Z}$, is

$$\bar{Z} = \frac{n_1\bar{X} + n_2\bar{Y}}{n_1 + n_2} = \frac{1}{n_1 + n_2}\sum_{j=1}^{n_1+n_2} Z_j,\ Z_j \in A \cup B,$$

where the Z_j's are the original vectors that were included in A or B.

15.3.4. The median method

A main shortcoming of the centroid method of joining two clusters is that if n_1 is very large compared to n_2, then $\bar{Z}$ is likely to be closer of $\bar{X}$, and vice versa. In order to avoid this type of imbalance, a method based on the median is suggested, under which the median of the combined clusters A and B is defined as

$$\text{Median}_{A\cup B} = \frac{1}{2}(\bar{X} + \bar{Y}) \ \text{ with } \ X_i \in A \ \text{ and } \ Y_j \in B, \tag{15.3.4}$$

for all $i,\ j$ and r. In this process, the clusters A and B for which $\text{Median}_{A\cup B}$ is the smallest are combined to form the next cluster whose elements are the Z_r's, $Z_r \in A \cup B$.

15.3.5. The residual sum of products method

From the one-way MANOVA layout, a residual or within group (within cluster) sum of products for clusters $A,\ B$ and $A\cup B$, denoted by $R_A,\ R_B$ and $R_{A\cup B}$, are the following:

$$R_A = \sum_{i=1}^{n_1}(X_i - \bar{X})'(X_i - \bar{X}),\ \ R_B = \sum_{j=1}^{n_2}(Y_j - \bar{Y})'(Y_j - \bar{Y})$$

$$R_{A\cup B} = \sum_{r=1}^{n_1+n_2}(Z_r - \bar{Z})'(Z_r - \bar{Z}),\ \ Z_j \in A \cup B,\ \ \bar{Z} = \frac{n_1\bar{X} + n_2\bar{Y}}{n_1 + n_2}.$$

Once those sums of squares have been evaluated, we compute the quantity

$$T_{A\cup B} = R_{A\cup B} - (R_A + R_B), \tag{15.3.5}$$

which can be interpreted as the increase in residual sum of products due to the process of merging the clusters A and B. Then, the procedure consists of combining those clusters A and B for which $T_{A\cup B}$ as defined in (15.3.5) is the minimum. This method is also called *Ward's method*.

There exist other methods for combining clusters such as the *flexible beta method*, and several comparative studies point out the merits and drawbacks of the various methods.

In the hierarchical procedures considered in Sect. 15.3, we begin with the n data points as n distinct clusters of one element each. Then, by applying certain "minimum distance" methods, "distance" being defined in different ways, we combined the clusters one by one. We may also consider a hierarchical procedure wherein the n data points are treated as one cluster of n elements. At this stage, by making use of some rules, we break up this cluster into two clusters. Then, one of these or both are split again as two clusters by applying the same rule. We continue the process and stop it when it is determined that there is a sufficient number of clusters. If the process is not halted at a certain stage, we will end up with a single cluster containing all of the n elements or points. We will not elaborate further on such procedures.

15.3.6. Other criteria for partitioning or optimization

In Sect. 15.2, we considered a non-hierarchical procedure known as the *k-means method*, which is the most popular in this area. After discussing this, we described the most widely utilized non-hierarchical procedure in Sect. 15.3. We will now examine other non-hierarchical procedures in common use. Some of these are connected with the MANOVA or multivariate analysis of variation of a one-way classification. In a multivariate one-way layout, let X_{ij} be the j-th vector in the i-th group or i-th cluster, all vectors being p-vectors or $p \times 1$ real vectors. Let there be k groups (k clusters) of sizes $n_1, \ldots, n_k$ with $n_1 + n_2 + \cdots + n_k = n_. = n$, that is, the cluster sizes are $n_1, \ldots, n_k$, respectively. Let the residual sum of products or sum of squares and cross products matrix be denoted by U, which is $p \times p$. This matrix U is also called *within group or within cluster* variation matrix. Let the *between groups or between clusters* variation matrix be V. In this setup, U and V are the following:

$$U = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)(X_{ij} - \bar{X}_i)', \ \ \bar{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}, \tag{15.3.6}$$

$$V = \sum_{ij}(\bar{X}_i - \bar{X})(\bar{X}_i - \bar{X})' = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})(\bar{X}_i - \bar{X})', \ \ \bar{X} = \frac{1}{n_.}\sum_{ij} X_{ij}\,. \tag{15.3.7}$$

Then, under the hypothesis that the group effects or cluster effects are the same, and under the normality assumption on the X_{ij}'s, U and V are independently distributed Wishart matrices with $n_. - k$ and $k - 1$ degrees of freedom, respectively, where $\Sigma > O$ is the parameter matrix in the Wishart densities as well as the common covariance matrix of the X_{ij}'s, referring to Chap. 5. Thus, $W_1 = (U + V)^{-\frac{1}{2}}U(U + V)^{-\frac{1}{2}}$ is a real matrix-variate type-1 beta random variable with the parameters $(\frac{n_.-k}{2}, \frac{k-1}{2})$, $W_2 = U^{-\frac{1}{2}}VU^{-\frac{1}{2}}$ is a real matrix-variate type-2 beta random variable with the parameters $(\frac{k-1}{2}, \frac{n_.-k}{2})$ and $W_3 = U + V$ follows a real Wishart distribution having $n_. - 1$ degrees of freedom and parameter matrix $\Sigma > O$, again referring Chap. 5. Observe that both U and V are real positive definite matrices, so that all of their eigenvalues are positive. The likelihood ratio criterion λ for testing the hypothesis that the group effects are the same is the following:

$$\lambda^{\frac{2}{n.}} = \frac{|U|}{|U+V|} = |W_1| = \frac{1}{|I + U^{-\frac{1}{2}} V U^{-\frac{1}{2}}|} = \frac{1}{|I + W_2|}. \tag{15.3.8}$$

We are aiming to have the within cluster variation small and the between cluster variation large, which means, in some sense, that U will be small and V will be large, in which case λ as given in (15.3.8) will be small. This also means that the trace of U must be small and trace of W_2 must be large. Accordingly, a few criteria for merging clusters are based on $\mathrm{tr}(U)$, $|U|$ and $\mathrm{tr}(W_2)$. The following are some commonly utilized criteria for combining clusters:

(1) Minimizing $\mathrm{tr}(U)$;

(2) Minimizing $|U|$;

(3) Maximizing $\mathrm{tr}(W_2)$.

These criteria are applied as follows: One of the n observation vectors is moved to a selected cluster if $\mathrm{tr}(U)$ is a minimum ($|U|$ is a minimum and $\mathrm{tr}(W_2)$ is a maximum for the other criteria). Then, $\mathrm{tr}(U)$ is evaluated after moving the observation vectors one by one to the selected cluster and, each time, $\mathrm{tr}(U)$ is noted; the vector for which $\mathrm{tr}(U)$ attains a minimum value belongs to the selected cluster, that is, it is combined with the selected cluster. Observe that

$$\begin{aligned}\mathrm{tr}(U) &= \mathrm{tr}\Big(\sum_{ij}(X_{ij}-\bar{X}_i)(X_{ij}-\bar{X}_i)'\Big)\\ &= \mathrm{tr}\Big(\sum_{j=1}^{n_1}(X_{1j}-\bar{X}_1)(X_{1j}-\bar{X}_1)'\Big)+\cdots+\mathrm{tr}\Big(\sum_{j=1}^{n_k}(X_{kj}-\bar{X}_k)(X_{kj}-\bar{X}_k)'\Big)\\ &= \sum_{j=1}^{n_1}(X_{1j}-\bar{X}_1)'(X_{1j}-\bar{X}_1)+\cdots+\sum_{j=1}^{n_k}(X_{kj}-\bar{X}_k)'(X_{kj}-\bar{X}_k),\end{aligned} \tag{15.3.9}$$

owing to the property that, for two matrices P and Q, $\mathrm{tr}(PQ) = \mathrm{tr}(QP)$ as long as PQ and QP are defined. As well, observe that since $(X_{ij}-\bar{X}_i)'(X_{ij}-\bar{X}_i)$ is a scalar quantity for every i and j, it is equal to its trace. How does this criterion work in practice? Consider moving a member from the s-th cluster to the selected cluster, namely, the r-th cluster. The original centroids are $\bar{X}_r$ and $\bar{X}_s$, and when one element is added to the r-th cluster from the s-th cluster, both centroids will respectively change to, say, $\bar{X}_{r+1}$ and $\bar{X}_{s-1}$. Compute the updated sums of squares in the new r-th and s-th clusters. Then, add up all the sums of squares in all the clusters and obtain a new $\mathrm{tr}(U)$. Carry out this process for every member in every other cluster and compute $\mathrm{tr}(U)$ each time. Take the smallest value of $\mathrm{tr}(U)$ thus calculated, including the original value of $\mathrm{tr}(U)$, before considering transferring any point. That vector for which $\mathrm{tr}(U)$ is minimum really belongs to the r-th cluster and so, is included in it. Repeat the process until no more improvement can be made, at which point no more transfer of points is necessary.

Simplification of the computations of $\mathrm{tr}(U)$

As will be explained, computing $\mathrm{tr}(U)$ can be simplified. Consider the new sum of squares in the r-th cluster. Let the new and old sums of squares be denoted by $(\mathrm{New})_r$, $(\mathrm{New})_s$, and $(\mathrm{Old})_r$, $(\mathrm{Old})_s$, respectively. Let the vector transferred from the s-th cluster to the r-th cluster be denoted by Y. Then,

$$\begin{aligned}(\mathrm{New})_r &= \sum_{j=1}^{r}(X_{rj}-\bar{X}_{r+1})'(X_{rj}-\bar{X}_{r+1})+(Y-\bar{X}_{r+1})'(Y-\bar{X}_{r+1})\\ &= \sum_{j=1}^{r}(X_{rj}-\bar{X}_r+(\bar{X}_r-\bar{X}_{r+1}))'(X_{rj}-\bar{X}_r+(\bar{X}_r-\bar{X}_{r+1}))\\ &\quad +(Y-\bar{X}_{r+1})'(Y-\bar{X}_{r+1})\\ &= \sum_{j=1}^{r}(X_{rj}-\bar{X}_r)'(X_{rj}-\bar{X}_r)+r(\bar{X}_r-\bar{X}_{r+1})'(\bar{X}_r-\bar{X}_{r+1})\\ &\quad +(Y-\bar{X}_{r+1})'(Y-\bar{X}_{r+1})\\ &= (\mathrm{Old})_r+r(\bar{X}_r-\bar{X}_{r+1})'(\bar{X}_r-\bar{X}_{r+1})+(Y-\bar{X}_{r+1})'(Y-\bar{X}_{r+1}).\end{aligned}$$

The difference between the new sum of squares and the old one is

$$\delta_1 = r(\bar{X}_r-\bar{X}_{r+1})'(\bar{X}_r-\bar{X}_{r+1})+(Y-\bar{X}_{r+1})'(Y-\bar{X}_{r+1}).$$

Noting that

$$\bar{X}_r-\bar{X}_{r+1}=\bar{X}_r-\frac{r\bar{X}_r+Y}{r+1}=\frac{1}{r+1}[\bar{X}_r-Y]\ \text{and}\ Y-\bar{X}_{r+1}=\frac{r}{r+1}[Y-\bar{X}_r],$$

δ_1 simplifies to

$$\delta_1=\frac{r}{r+1}(Y-\bar{X}_r)'(Y-\bar{X}_r).$$

A similar procedure can be used for the s-th cluster. In that case, the new sum of squares can be written as

$$
\begin{aligned}
(\text{New})_s &= \sum_{j=1}^{s-1}(X_{sj}-\bar{X}_{s-1})'(X_{sj}-\bar{X}_{s-1})\\
&= \sum_{j=1}^{s}(X_{ij}-\bar{X}_{s-1})'(X_{sj}-\bar{X}_{s-1})-(Y-\bar{X}_{s-1})'(Y-\bar{X}_{s-1}).
\end{aligned}
$$

Then, proceeding as in the case of the r-th cluster and denoting the difference between the new and the old sums of squares as δ_2, we have

$$
\delta_2 = -\frac{s}{s-1}(Y-\bar{X}_s)'(Y-\bar{X}_s),\ s>1,
$$

so that the sum of the differences between the new and old sums of squares, denoted by δ, is the following:

$$
\delta = \delta_1+\delta_2 = \frac{r}{r+1}(Y-\bar{X}_r)'(Y-\bar{X}_r)-\frac{s}{s-1}(Y-\bar{X}_s)'(Y-\bar{X}_s) \qquad (15.3.10)
$$

for $s>1$, where $\bar{X}_r$ and $\bar{X}_s$ are the original centroids of the r-th and s-th clusters, respectively. As such, computing δ is very simple. Evaluate the quantity specified in (15.3.10) for all the points outside the r-th cluster and look for the minimum of δ, including the original value of $\delta = 0$. If the minimum occurs at a point Y_1 outside of the r-th cluster, then transfer that point to the r-th cluster. Continue the process for every vector in the s-th cluster and then, for $r = 1,\ldots,k$, assuming there are k clusters, until $\delta = 0$. In the end, all the clusters are stabilized, and k may take on another value.

Among the three statistics $\text{tr}(U)$, $|U|$ and $\text{tr}(W_2)$, $\text{tr}(U)$ is the easiest to compute, as was just explained. However, if we consider a non-singular transformation, other than an orthonormal transformation, then $|U|$ and $\text{tr}(W_2)$ are invariant, but $\text{tr}(U)$ is not.

We have discussed one hierarchical methodology of single linkage nearest neighbor method and one non-hierarchical procedure consisting of the k-means method. These seem to be the most widely utilized. We also mentioned other hierarchical and non-hierarchical methods without going into the details. All these procedures are not well-defined mathematical procedures. None of the procedures can uniquely determine the clusters if there are some clusters in the multivariate data at hand, and none of the methods can uniquely determine the number of clusters. The advantages and shortcomings of the various methods will not be discussed so as not to confound the reader.

Exercises 15

15.1. For the $p \times 1$ vectors $X_1, \ldots, X_n$, let the dissimilarity measures be (1) $d_{ij}^{(1)} = \sum_{k=1}^{n} |x_{ik} - x_{jk}|$, (2) $d_{ij}^{(2)} = \sum_{k=1}^{n}(x_{ik} - x_{jk})^2$, $X_i' = [x_{1i}, x_{2i}, \ldots, x_{pi}]$. Compute the matrices (1) $(d_{ij}^{(1)})$; (2) $(d_{ij}^{(2)})$, for the following vectors:

$$X_1 = \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix},\ X_2 = \begin{bmatrix} -1 \\ 1 \\ 2 \end{bmatrix},\ X_3 = \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix},\ X_4 = \begin{bmatrix} 2 \\ 1 \\ -1 \end{bmatrix}.$$

15.2. Nine test runs $T - 1, \ldots, T - 9$ are done to test the breaking strengths of three alloys. The following data are the deviations from the respective expected strengths:

↓→	$T-1$	$T-2$	$T-3$	$T-4$	$T-5$	$T-6$	$T-7$	$T-8$	$T-9$
Alloy-1	0	−1	1	2	−1	2	5	4	5
Alloy-2	1	1	1	1	3	4	7	−4	8
Alloy-3	−1	0	1	2	2	3	8	4	−7

Carry out a cluster analysis by applying the following methods: (1) The single linkage or nearest neighbor method; (2) The average linkage method; (3) The centroid method; (4) The residual sum of products method.

15.3. Using the data provided in Exercise 15.2, carry out a cluster analysis by utilizing the following methods: (1) Partitioning or optimization; (2) Minimization of $\text{tr}(U)$; (3) Minimization of $|U|$; (4) Maximization of $\text{tr}(W_2)$ where U and W are given in Sect. 15.3.6.

15.4. Compare the results from the different methods in (1) Exercise 15.2; (2) Exercise 15.3, and make your observations.

15.5. Compare the results from the different methods in Exercises 15.2 and 15.3, and comment on the similarities and differences.

15.4. Correspondence Analysis

If the data at hand are classified according to two attributes, these characteristics may be of the same type, that is, both quantitative or both qualitative, or of different types, and whatever the types may be, we may construct a two-way contingency table. In a contingency table, the entries in the cells are frequencies or the number of times various combinations of the attributes appear. Correspondence Analysis is a process of identifying, quantifying, separating and plotting associations among the characteristics and relationships among the various levels. In a two-way contingency table, we identify, separate and plot associations between the two characteristics and attempt to identify relationships between row and column labels.

15.4.1. Two-way contingency table

Consider the following example. A random sample of 100 persons from a certain township are classified according to their educational level and their liberal disposition. In the frequency Table 15.4.1, the A_j's represent their dispositions and the B_j's, their educational levels, with $A_1 \equiv$ tolerant, $A_2 \equiv$ indifferent, $A_3 \equiv$ intolerant, $B_1 \equiv$ primary school education level, $B_2 \equiv$ high school education level, $B_3 \equiv$ bachelor's degree education level, $B_4 \equiv$ master's and higher degree education level.

Table 15.4.1: A two-way contingency table

$\downarrow \rightarrow$	B_1	B_2	B_3	B_4	Total
A_1	6	14	16	4	40
A_2	17	5	8	10	40
A_3	7	6	6	1	20
Total	30	25	30	15	100

There are 6 persons having a tolerant disposition and primary school level of education. There is one person with an intolerant disposition and a master's degree or a higher level of education, and so on. The marginal sums are also provided in the table. For example, the total number of persons having a primary school level of education is 30, the total number of persons having an intolerant disposition is 20, and so on. The corresponding relative frequencies (a given frequency divided by 100, the total frequency) are as follows (Table 15.4.2):

Table 15.4.2: Relative frequencies f_{ij} in the two-way contingency table

$\downarrow \rightarrow$	B_1	B_2	B_3	B_4	Total
A_1	$0.06(f_{11})$	$0.14(f_{12})$	$0.16(f_{13})$	$0.04(f_{14})$	$0.40(f_{1.})$
A_2	$0.17(f_{21})$	$0.05(f_{22})$	$0.08(f_{23})$	$0.10(f_{24})$	$0.40(f_{2.})$
A_3	$0.07(f_{31})$	$0.06(f_{32})$	$0.06(f_{33})$	$0.01(f_{34})$	$0.20(f_{3.})$
Total	$0.30(f_{.1})$	$0.25(f_{.2})$	$0.30(f_{.3})$	$0.15(f_{.4})$	$1.00(f_{..})$

The relative frequencies are denoted in parentheses by f_{ij} where the summation with respect to a subscript is designated by a dot, that is, $f_{i.} = \sum_j f_{ij}$, $f_{.j} = \sum_i f_{ij}$ and $f_{..} = \sum_i \sum_j f_{ij}$. Note that $f_{..} = 1$. In a general notation, a two-way contingency table and the corresponding relative frequencies are displayed as follows (Table 15.4.3):

Table 15.4.3: A two-way contingency table and a table of relative frequencies

$\downarrow\ \rightarrow$	B_1	B_2	$\cdots$	B_s	*Total*
A_1	n_{11}	n_{12}	$\cdots$	n_{1s}	$n_{1.}$
A_2	n_{21}	n_{22}	$\cdots$	n_{2s}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_r	n_{r1}	n_{r2}	$\cdots$	n_{rs}	$n_{r.}$
Total	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.s}$	$n_{..} = n$

,

$\downarrow\ \rightarrow$	B_1	B_2	$\cdots$	B_s	*Total*
A_1	f_{11}	f_{12}	$\cdots$	f_{1s}	$f_{1.}$
A_2	f_{21}	f_{22}	$\cdots$	f_{2s}	$f_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_r	f_{r1}	f_{r2}	$\cdots$	f_{rs}	$f_{r.}$
Total	$f_{.1}$	$f_{.2}$	$\cdots$	$f_{.s}$	$f_{..} = 1$

Letting the true probability of the occurrence of an observation in the $(i,\ j)$-th cell be p_{ij}, the following is the table of true probabilities:

Table 15.4.4: True probabilities p_{ij} in a two-way contingency table

$\downarrow\ \rightarrow$	B_1	B_2	$\cdots$	B_s	*Total*
A_1	p_{11}	p_{12}	$\cdots$	p_{1s}	$p_{1.}$
A_2	p_{21}	p_{22}	$\cdots$	p_{2s}	$p_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_r	p_{r1}	p_{r2}	$\cdots$	p_{rs}	$p_{r.}$
Total	$p_{.1}$	$p_{.2}$	$\cdots$	$p_{.s}$	$p_{..} = 1$

These are multinomial probabilities and, in this case, the n_{ij}'s become multinomial variables. An estimate of p_{ij}, denoted by $\hat{p}_{ij}$, is $\hat{p}_{ij} = f_{ij}$, the corresponding relative frequency. The marginal sums in Table 15.4.4 can be interpreted as follows: $p_{1.}$ = the probability of finding an item in the first row or the probability of an event will have the attribute A_1; $p_{.j}$ = the probability that an event will have the characteristic B_j, and so on. Thus,

$$\hat{p}_{ij} = f_{ij} = \frac{n_{ij}}{n},\ \ \hat{p}_{i.} = \frac{n_{i.}}{n},\ \ \hat{p}_{.j} = \frac{n_{.j}}{n},\ \ i = 1, \ldots, r,\ \ j = 1, \ldots, s.$$

If A_i and B_j are respectively interpreted as the event that an observation will belong to the i-th row or the event of the occurrence of the characteristic A_i, and the event that an observation will belong to the j-th column or the event of the occurrence of the attribute B_j, and if we let $p_{i.} = P(A_i)$ and $p_{.j} = P(B_j)$, then $p_{ij} = P(A_i \cap B_j)$, where $P(A_i)$ is the probability of the event A_i, $P(B_j)$ is the probability of the event B_j, and $(A_i \cap B_j)$ is the intersection or joint occurrence of the events A_i and B_j. If A_i and B_j are independent events, $P(A_i \cap B_j) = P(A_i)P(B_j)$ or $p_{ij} = p_{i.}p_{.j}$, the product of the marginal probabilities or the marginal totals in the table of probabilities. That is,

$$P(A_i \cap B_j) = P(A_i)P(B_j) \ \Rightarrow \ p_{ij} = p_{i.}p_{.j}, \ \hat{p}_{ij} = \Big(\frac{n_{i.}}{n}\Big)\Big(\frac{n_{.j}}{n}\Big) = \frac{n_{i.}n_{.j}}{n^2} \tag{15.4.1}$$

for all i and j. In a multinomial distribution, the expected frequency in the (i, j)-th cell is np_{ij} where n is the total frequency. Then, the expected frequency, denoted by $E[\cdot]$, the maximum likelihood estimate (MLE) of the expected frequency, denoted by $\hat{E}[\cdot]$, and the MLE of the expected frequency under the hypothesis H_o of independence of events A_i and B_j, are the following:

$$E[n_{ij}] = np_{ij}, \ \hat{E}[n_{ij}] = n\hat{p}_{ij} = n\Big(\frac{n_{ij}}{n}\Big), \ n\hat{p}_{ij}|H_o = n\hat{p}_{i,}\hat{p}_{.j} = n\Big(\frac{n_{i.}}{n}\Big)\Big(\frac{n_{.j}}{n}\Big) = \frac{n_{i.}n_{.j}}{n}. \tag{15.4.2}$$

Now, referring to our numerical example and the first row of Table 15.4.1, the estimated expected frequencies, under H_o are: $E[n_{11}|H_o] = \frac{n_{1.}n_{.1}}{n} = \frac{40\times 30}{100} = 12$, $E[n_{12}|H_o] = \frac{n_{1.}n_{.2}}{n} = \frac{40\times 25}{100} = 10$, $E[n_{13}|H_o] = \frac{40\times 30}{100} = 12$, $E[n_{14}|H_o] = \frac{40\times 15}{100} = 6$. All the estimated expected frequencies are shown in parentheses next to the observed frequencies in Table 15.4.5:

Table 15.4.5: A two-way contingency table

↓ →	B_1	B_2	B_3	B_4	Total
A_1	6(12)	14(10)	16(12)	4(6)	40(40)
A_2	17(12)	5(10)	8(12)	10(6)	40(40)
A_3	7(6)	6(5)	6(6)	1(3)	20(20)
Total	30(30)	25(25)	30(30)	15(15)	100(100)

15.4.2. Some general computations

Let J_r and J_s be respectively $r \times 1$ and $s \times 1$ vectors of unities and P be the true probability matrix, that is,

$$J_r = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \ J_s = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \ P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1s} \\ p_{21} & p_{22} & \cdots & p_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ p_{r1} & p_{r2} & \cdots & p_{rs} \end{bmatrix}. \tag{15.4.3}$$

Letting the marginal totals be denoted by R and C', we have

$$R = PJ_s = \begin{bmatrix} p_{1.} \\ p_{2.} \\ \vdots \\ p_{r.} \end{bmatrix}, \ \begin{aligned} J_r'P &= [p_{.1}, p_{.2}, \ldots, p_{.s}] = C' \\ J_r'R &= 1 = C'J_s \ . \end{aligned} \tag{15.4.4}$$

Referring to the initial numerical example, we have the following:

$$\hat{R} = \begin{bmatrix} \hat{p}_{1.} \\ \hat{p}_{2.} \\ \vdots \\ \hat{p}_{r.} \end{bmatrix} = \begin{bmatrix} n_{1.}/n \\ n_{2.}/n \\ \vdots \\ n_{r.}/n \end{bmatrix} = \begin{bmatrix} 40/100 \\ 40/100 \\ 20/100 \end{bmatrix} = \begin{bmatrix} 0.4 \\ 0.4 \\ 0.2 \end{bmatrix}$$

$$\hat{C}' = [\hat{p}_{.1}, \hat{p}_{.2}, \ldots, \hat{p}_{.s}] = \Big[\frac{n_{.1}}{n}, \ldots, \frac{n_{.s}}{n}\Big]$$
$$= [\tfrac{30}{100}, \tfrac{25}{100}, \tfrac{30}{100}, \tfrac{15}{100}] = [0.30, 0.25, 0.30, 0.15].$$

Writing the bordered matrix P as

$$\begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1s} & p_{1.} \\ p_{21} & p_{22} & \cdots & p_{2s} & p_{2.} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ p_{r1} & p_{r2} & \cdots & p_{rs} & p_{r.} \\ p_{.1} & p_{.2} & \cdots & p_{.s} & 1 \end{bmatrix} = \begin{bmatrix} P & R \\ C' & 1 \end{bmatrix}, \tag{15.4.5}$$

in the numerical example, these quantities are

$$\begin{bmatrix} \hat{P} & \hat{R} \\ \hat{C}' & 1 \end{bmatrix} = \begin{bmatrix} 0.06 & 0.14 & 0.16 & 0.04 & 0.40 \\ 0.17 & 0.05 & 0.08 & 0.10 & 0.40 \\ 0.07 & 0.06 & 0.06 & 0.01 & 0.20 \\ 0.30 & 0.25 & 0.30 & 0.15 & 1.00 \end{bmatrix}.$$

Let D_r and D_c be the following diagonal matrices corresponding respectively to the row and column marginal probabilities:

$$D_r = \begin{bmatrix} p_{1.} & 0 & \cdots & 0 \\ 0 & p_{2.} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p_{r.} \end{bmatrix}, \ D_c = \begin{bmatrix} p_{.1} & 0 & \cdots & 0 \\ 0 & p_{.2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & p_{.s} \end{bmatrix} \ \text{or}$$
$$D_r = \text{diag}(p_{1.}, p_{2.}, \ldots, p_{r.}), \ D_c = \text{diag}(p_{.1}, p_{.2}, \ldots, p_{.s}). \tag{15.4.6}$$

In the numerical example, these quantities are

$$\hat{D}_r = \text{diag}(0.4, 0.4, 0.2) \ \text{and} \ \hat{D}_c = \text{diag}(0.30, 0.25, 0.30, 0.15).$$

Now, consider $D_r^{-1}P$ and PD_c^{-1}:

$$D_r^{-1}P = \begin{bmatrix} \frac{p_{11}}{p_{1.}} & \frac{p_{12}}{p_{1.}} & \dots & \frac{p_{1s}}{p_{1.}} \\ \frac{p_{21}}{p_{2.}} & \frac{p_{22}}{p_{2.}} & \dots & \frac{p_{2s}}{p_{2.}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{p_{r1}}{p_{r.}} & \frac{p_{r2}}{p_{r.}} & \dots & \frac{p_{rs}}{p_{r.}} \end{bmatrix} \equiv \begin{bmatrix} R_1' \\ R_2' \\ \vdots \\ R_r' \end{bmatrix}, \ R_j = \begin{bmatrix} \frac{p_{j1}}{p_{j.}} \\ \frac{p_{j2}}{p_{j.}} \\ \vdots \\ \frac{p_{js}}{P_{j.}} \end{bmatrix}, \tag{15.4.7}$$

$$PD_c^{-1} = \begin{bmatrix} \frac{p_{11}}{p_{.1}} & \frac{p_{12}}{p_{.2}} & \dots & \frac{p_{1s}}{p_{.s}} \\ \frac{p_{21}}{p_{.1}} & \frac{p_{22}}{p_{.2}} & \dots & \frac{p_{2s}}{p_{.s}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{p_{r1}}{p_{.1}} & \frac{p_{r2}}{p_{.2}} & \dots & \frac{p_{rs}}{p_{.s}} \end{bmatrix} \equiv [C_1, \dots, C_s], \ C_j = \begin{bmatrix} \frac{p_{1j}}{p_{.j}} \\ \frac{p_{2j}}{p_{.j}} \\ \vdots \\ \frac{p_{rj}}{p_{.j}} \end{bmatrix}. \tag{15.4.8}$$

Referring to the numerical example, we have

$$\hat{D}_r^{-1}\hat{P} = \begin{bmatrix} n_{11}/n_{1.} & n_{12}/n_{1.} & \cdots & n_{1s}/n_{1.} \\ n_{21}/n_{2.} & n_{22}/n_{2.} & \cdots & n_{2s}/n_{2.} \\ \vdots & \vdots & \ddots & \vdots \\ n_{r1}/n_{r.} & n_{r2}/n_{r.} & \cdots & n_{rs}/n_{r.} \end{bmatrix} = \begin{bmatrix} 6/40 & 14/40 & 16/40 & 4/40 \\ 17/40 & 5/40 & 8/40 & 10/40 \\ 7/20 & 6/20 & 6/20 & 1/20 \end{bmatrix}$$

$$\hat{D}_r^{-1}\hat{P}J_s = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

$$\hat{P}\hat{D}_c^{-1} = \begin{bmatrix} n_{11}/n_{.1} & n_{12}/n_{.2} & \cdots & n_{1s}/n_{.s} \\ n_{21}/n_{.1} & n_{22}/n_{.2} & \cdots & n_{2s}/n_{.s} \\ \vdots & \vdots & \ddots & \vdots \\ n_{r1}/n_{.1} & n_{r2}/n_{.2} & \cdots & n_{rs}/n_{.s} \end{bmatrix} = \begin{bmatrix} 6/30 & 14/25 & 16/30 & 4/15 \\ 17/30 & 5/25 & 8/30 & 10/15 \\ 7/30 & 6/25 & 6/30 & 1/15 \end{bmatrix}$$

$$J_r'\hat{P}\hat{D}_c^{-1} = [1, 1, 1, 1].$$

For computing the test statistics in vector/matrix notation, we need (15.4.7) and (15.4.8).

15.5. Various Representations of Pearson's χ^2 Statistic

Now, let us consider Pearson's χ^2 statistic for testing the hypothesis that there is no association between the two characteristics of classification or the hypothesis $H_o : p_{ij} = p_{i.}p_{.j}$. The χ^2 statistic is the following:

$$\chi^2=\sum_{ij}\frac{(\text{observed frequency}-\text{expected frequency})^2}{(\text{expected frequency})}=\sum_{ij}\frac{(n_{ij}-\frac{n_{i.}n_{.j}}{n})^2}{\frac{n_{i.}n_{.j}}{n}} \tag{15.5.1}$$

$$=\sum_{ij}n\frac{(\frac{n_{ij}}{n}-\frac{n_{i.}}{n}\frac{n_{.j}}{n})^2}{\frac{n_{i.}}{n}\frac{n_{.j}}{n}}=n\sum_{ij}\frac{(\hat{p}_{ij}-\hat{p}_{i.}\hat{p}_{.j})^2}{\hat{p}_{i.}\hat{p}_{.j}} \tag{15.5.2}$$

$$=\sum_{i=1}^{r}n\hat{p}_{i.}\sum_{j=1}^{s}\Big[\Big(\frac{\hat{p}_{ij}}{\hat{p}_{i.}}-\hat{p}_{.j}\Big)^2/\hat{p}_{.j}\Big] \tag{15.5.3}$$

$$=\sum_{j=1}^{s}n\hat{p}_{.j}\sum_{i=1}^{r}\Big[\Big(\frac{\hat{p}_{ij}}{\hat{p}_{.j}}-\hat{p}_{i.}\Big)^2/\hat{p}_{i.}\Big]. \tag{15.5.4}$$

In order to simplify the notation, we shall omit placing a hat on top of the estimates of $R_i,\ C_j,\ R,\ C,\ D_c$ and D_r. We may then express the χ^2 statistic as the following quadratic forms:

$$\chi^2=\sum_{i=1}^{r}np_{i.}(R_i-C)'D_c^{-1}(R_i-C) \tag{15.5.5}$$

$$=\sum_{j=1}^{s}np_{.j}(C_j-R)'D_r^{-1}(C_j-R). \tag{15.5.6}$$

The forms given in (15.5.5) and (15.5.6) are very convenient for extending the theory to multi-way classifications.

It is well known that, under H_o, Pearson's χ^2 statistic is asymptotically distributed as a chisquare random variable having $(r-1)(s-1)$ degrees of freedom as $n\to\infty$. One can also express (15.4.8) as a generalized distance between the observed frequencies and the expected frequencies, which is a quadratic form involving the inverse of the true covariance matrix of the multinomial distribution of the n_{ij}'s. Then, on applying the multivariate version of the central limit theorem, it can be established that, as $n\to\infty$, Pearson's χ^2 statistic has a χ^2 distribution with $(r-1)(s-1)$ degrees of freedom. For the representation of Pearson's χ^2 goodness-of-fit statistic as a generalized distance and as a quadratic form, and for the proof of its asymptotic distribution, the reader may refer to Mathai and Haubold (2017). There exist other derivations of this result in the literature.

The quadratic forms specified in (15.5.5) and (15.5.6) can also be interpreted as comparing the generalized distance between the vectors R_i and C in (15.5.5) and between the vectors C_j and R in (15.5.6), respectively. These will also be equivalent to testing the hypothesis $H_o: p_{ij}=p_{i.}p_{.j}$. As well, an interpretation can be provided in terms of profile analysis: then, the test will correspond to testing the hypothesis that the weighted row profiles are similar; analogously, using (15.5.6) corresponds to testing the hypothesis that the column profiles in a two-way contingency table are similar. Now, examine the following item:

$$P - RC' = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1s} \\ p_{21} & p_{22} & \cdots & p_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ p_{r1} & p_{r2} & \cdots & p_{rs} \end{bmatrix} - \begin{bmatrix} p_{1.} \\ p_{2.} \\ \vdots \\ p_{r.} \end{bmatrix} [p_{.1}, p_{.2}, \ldots, p_{.s}]$$

$$= \begin{bmatrix} p_{11} - p_{1.}p_{.1} & p_{12} - p_{1.}p_{.2} & \cdots & p_{1s} - p_{1.}p_{.s} \\ p_{21} - p_{2.}p_{.1} & p_{22} - p_{2.}p_{.2} & \cdots & p_{2s} - p_{2.}p_{.s} \\ \vdots & \vdots & \ddots & \vdots \\ p_{rs} - p_{r.}p_{.1} & p_{r2} - p_{r.}p_{.2} & \cdots & p_{rs} - p_{r.}p_{.s} \end{bmatrix}.$$

Referring to our numerical example, these quantities are the following:

$$\hat{P} - \hat{R}\hat{C}' = \begin{bmatrix} \frac{n_{11}}{n} - \frac{n_{1.}n_{.1}}{n^2} & \cdots & \frac{n_{1s}}{n} - \frac{n_{1.}n_{.s}}{n^2} \\ \vdots & \ddots & \vdots \\ \frac{n_{r1}}{n} - \frac{n_{r.}n_{.1}}{n^2} & \cdots & \frac{n_{rs}}{n} - \frac{n_{r.}n_{.s}}{n^2} \end{bmatrix} = \frac{1}{100} \times$$

$$\begin{bmatrix} 6 - (40)(30)/100 & 14 - (40)(25)/100 & 16 - (40)(30)/100 & 4 - (40)(15)/100 \\ 17 - (40)(30)/100 & 5 - (40)(25)/100 & 8 - (40)(30)/100 & 10 - (40)(15)/100 \\ 7 - (20)(30)/100 & 6 - (20)(25)/100 & 6 - (20)(30)/100 & 1 - (20)(15)/100 \end{bmatrix}$$

$$= \begin{bmatrix} -6 & 4 & 4 & -2 \\ 5 & -5 & -4 & 4 \\ 1 & 1 & 0 & -2 \end{bmatrix}.$$

15.5.1. Testing the hypothesis of no association in a two-way contingency table

The observed value of Pearson's χ^2 statistic is

$$\chi^2 = \left[\frac{(-6)^2}{12} + \frac{(5)^2}{12} + \frac{(1)^2}{6}\right] + \left[\frac{(4)^2}{10} + \frac{(-5)^2}{10} + \frac{(1)^2}{5}\right]$$
$$+ \left[\frac{(4)^2}{12} + \frac{(-4)^2}{12} + \frac{(0)^2}{6}\right] + \left[\frac{(-2)^2}{6} + \frac{(4)^2}{6} + \frac{(-2)^2}{3}\right]$$
$$= 16.88.$$

Given our data, $(r - 1)(s - 1) = (2)(3) = 6$, and the tabulated critical value is $\chi^2_{6,0.05} = 12.59$ at the 5% significance level. Since $12.59 < 16.88$, the hypothesis of no association between the classification attributes is rejected as per the evidence provided by the data. This χ^2 approximation may be questionable since one of the expected cell frequencies is less that 5. For a proper application of this approximation, the cell frequencies ought to be at least 5.

15.6. Plot of Row and Column Profiles

Now, $(P - RC')D_c^{-1}$ means that the columns of $(P - RC')$ are multiplied by $\frac{1}{p_{.1}}, \ldots, \frac{1}{p_{.s}}$, respectively. Then, $(P - RC')D_c^{-1}(P - RC')'$ is a matrix of all square and cross product terms involving $p_{ij} - p_{i.}p_{.j}$ for all i and j, where the s columns are weighted by $\frac{1}{p_{.j}}$, and if pre-multiplied by D_r^{-1}, the rows are weighted by $\frac{1}{p_{1.}}, \ldots, \frac{1}{p_{r.}}$, respectively. Looking at the diagonal elements, we note that Pearson's χ^2 statistic is nothing but

$$\chi^2 = n\,\mathrm{tr}[D_r^{-1}(P - RC')D_c^{-1}(P - RC')'] \tag{15.6.1}$$

$$= n\sum_{ij}\frac{(p_{ij} - p_{i.}p_{.j})^2}{p_{i.}p_{.j}} \tag{15.6.2}$$

$$= n(\lambda_1^2 + \cdots + \lambda_k^2) \tag{15.6.3}$$

where $\lambda_1^2, \ldots, \lambda_k^2$ are the nonzero eigenvalues of the matrix $D_r^{-1}(P - RC')D_c^{-1}(P - RC')'$ or of the matrix $D_r^{-\frac{1}{2}}(P - RC')D_c^{-1}(P - RC')'D_r^{-\frac{1}{2}}$ with k being the rank of $P - RC'$. For the numerical example, the observed value of the matrix $Y = (y_{ij})$ with $y_{ij} = \frac{p_{ij} - p_{i.}p_{.j}}{\sqrt{p_{i.}p_{.j}}}$, is obtained as follows, observing that

$$\sqrt{n}\,\frac{\hat{p}_{ij} - \hat{p}_{i.}\hat{p}_{.j}}{\sqrt{\hat{p}_{i.}\hat{p}_{.j}}} = \Big[n_{ij} - \frac{n_{i.}n_{.j}}{n}\Big]\Big/\sqrt{n_{i.}n_{.j}/n}.$$

From the representation of $\hat{P} - \hat{R}\hat{C}'$, we already have the matrix $n_{ij} - n_{i.}n_{.j}/n$, that is,

$$\left(\frac{(n_{ij} - n_{i.}n_{.j}/n)}{\sqrt{n_{i.}n_{.j}/n}}\right)$$
$$= \begin{bmatrix} (6-12)/\sqrt{12} & (14-10)/\sqrt{10} & (16-12)/\sqrt{12} & (4-6)/\sqrt{6} \\ (17-12)/\sqrt{12} & (5-10)/\sqrt{10} & (8-12)/\sqrt{12} & (10-6)/\sqrt{6} \\ (7-6)/\sqrt{6} & (6-5)/\sqrt{5} & (6-6)/\sqrt{6} & (1-3)/\sqrt{3} \end{bmatrix}$$
$$= \begin{bmatrix} -6/\sqrt{12} & 4/\sqrt{10} & 4/\sqrt{12} & -2/\sqrt{6} \\ 5/\sqrt{12} & -5/\sqrt{10} & -4/\sqrt{12} & 4/\sqrt{6} \\ 1/\sqrt{6} & 1/\sqrt{5} & 0/\sqrt{6} & -2/\sqrt{3} \end{bmatrix}.$$

Then,

$$nYY' = \begin{bmatrix} \frac{33}{5} & -\frac{43}{6} & \frac{17\sqrt{2}}{30} \\ -\frac{43}{6} & \frac{103}{12} & -\frac{17\sqrt{2}}{12} \\ \frac{17\sqrt{2}}{30} & -\frac{17\sqrt{2}}{12} & \frac{17}{10} \end{bmatrix}. \tag{15.6.4}$$

The representation in (15.6.1) has the advantage that

$$\begin{aligned}
&\mathrm{tr}[D_r^{-\frac{1}{2}}(P-RC')D_c^{-1}(P-RC')'D_r^{-\frac{1}{2}}]\\
&=\mathrm{tr}[YY'],\quad Y=D_r^{-\frac{1}{2}}(P-RC')D_c^{-\frac{1}{2}}=(y_{ij}),\\
y_{ij}&=\frac{p_{ij}-p_{i.}p_{.j}}{\sqrt{p_{i.}p_{.j}}},\quad \sum_j n\hat{y}_{ij}^2=\chi^2=n\mathrm{tr}(\hat{Y}\hat{Y}'). \qquad (15.6.5)
\end{aligned}$$

Note that Y is $r\times s$ and the rank of Y is equal to the rank of $P-RC'$, which is k, referring to (15.6.3). Thus, there are k nonzero eigenvalues associated with the $r\times r$ matrix YY' as well as with the $s\times s$ matrix $Y'Y$, which are $\lambda_1^2,\ldots,\lambda_k^2$. Since $\mathrm{tr}(YY')=\lambda_1^2+\cdots+\lambda_k^2$, we can represent Pearson's χ^2 statistic as follows, substituting the estimates of $p_{ij},\ p_{i.}$ and $p_{.j}$, etc:

$$\begin{aligned}
\frac{\chi^2}{n}&=\mathrm{tr}(YY')=\lambda_1^2+\cdots+\lambda_k^2\\
&=\sum_{i=1}^{k}\hat{p}_{i.}(\hat{R}_i-\hat{C})'\hat{D}_c^{-1}(\hat{R}_i-\hat{C}) \qquad (15.6.6)\\
&=\sum_{j=1}^{s}\hat{p}_{.j}(\hat{C}_j-\hat{R})'\hat{D}_r^{-1}(\hat{C}_j-\hat{R}). \qquad (15.6.7)
\end{aligned}$$

The expressions given in (15.6.6) and (15.6.7) and the sum of the λ_j^2's are called the *total inertia* in a two-way contingency table. We can also define the squared distance between two rows as

$$d_{ij(r)}^2=(R_i-R_j)'D_c^{-1}(R_i-R_j) \qquad (15.6.8)$$

and the squared distance between two columns as

$$d_{ij(c)}^2=(C_i-C_j)'D_r^{-1}(C_i-C_j). \qquad (15.6.9)$$

When the distance as specified in (15.6.8) is very small, we may combine the i-th and j-th rows, if necessary. Sometimes, the cell frequencies are small and we may wish to combine the small frequencies with other cell frequencies so that the χ^2 approximation of Pearson's χ^2 statistic be more accurate. Then, one can rely on (15.6.8) and (15.6.9) to determine whether it is indicated to combine rows and columns.

For convenience, let $r\le s$. Let $U_1,\ldots,U_r$ be the $r\times 1$ normalized eigenvectors of YY' and let the $r\times k$ matrix $U=[U_1,U_2,\ldots,U_k],\ k\le r$. Let $V_1,\ldots,V_s$ be the normalized eigenvectors of $Y'Y$ and let the $r \times k$ matrix $V = [V_1, \ldots, V_k]$, $k \le s$. Now, consider the singular value decomposition

$$Y = D_r^{-\frac{1}{2}}(P - RC')D_c^{-\frac{1}{2}} = U\Lambda V' \tag{15.6.10}$$

where $UU' = I_k = V'V$ and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$. Then, we can write

$$P - RC' = D_r^{\frac{1}{2}} U\Lambda V' D_c^{\frac{1}{2}} = W\Lambda Z' \tag{15.6.11}$$

where $W = D_r^{\frac{1}{2}}U$ and $Z = D_c^{\frac{1}{2}}V$. Let $W_j,\ j = 1, \ldots, k$, denote the columns of $W = [W_1, W_2, \ldots, W_k]$ and let $Z_j,\ j = 1, \ldots, k$, denote the columns of $Z = [Z_1, Z_2, \ldots, Z_k]$. Then, we can write

$$P - RC' = \sum_{j=1}^{k} \lambda_j W_j Z_j' \tag{15.6.12}$$

where $W'D_r^{-1}W = U'U = I_k = V'V = Z'D_c^{-1}Z$. Note that $P - RC'$ is the deviation matrix under the hypothesis $H_o : p_{ij} = p_{i.}p_{.j}$ or

$$P - RC' = (p_{ij} - p_{i.}p_{.j}) \ \text{ and } \ Y = (y_{ij}) = D_r^{-\frac{1}{2}}(P - RC')D_c^{-\frac{1}{2}} = \left(\frac{p_{ij} - p_{i.}p_{.j}}{\sqrt{p_{i.}p_{.j}}}\right).$$

Thus, the procedure is as follows: If $r \le s$, then compute the r eigenvalues of the $r \times r$ matrix YY'. If Y is of rank r, $YY' > O$ (positive definite), otherwise YY' is positive semi-definite. Let the nonzero eigenvalues of YY' be $\lambda_1^2, \ldots, \lambda_k^2$, assuming that k is the number of nonzero eigenvalues of YY'. These will also be the nonzero eigenvalues of $Y'Y$. Compute the normalized eigenvectors from YY' and denote those corresponding to the nonzero eigenvalues by $U = [U_1, \ldots, U_k]$ where U_j is the j-th column of U. Letting the normalized eigenvectors obtained from $Y'Y$, which correspond to the same nonzero eigenvalues, be denoted by $V = [V_1, \ldots, V_k]$, we have

$$Y = U\Lambda V',\ \ \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_k),\ \ YY' = U\Lambda^2U' \ \text{ and } \ Y'Y = V\Lambda^2V'. \tag{15.6.13}$$

Example 15.6.1. Construct a singular value decomposition of the following matrix Q:

$$Q = \begin{bmatrix} -1 & 1 & -1 & 0 \\ 1 & 1 & 0 & 2 \end{bmatrix}.$$

Solution 15.6.1. Let us compute QQ' as well as $Q'Q$ and the eigenvalues of QQ'. Since

$$QQ' = \begin{bmatrix} -1 & 1 & -1 & 0 \\ 1 & 1 & 0 & 2 \end{bmatrix} \begin{bmatrix} -1 & 1 \\ 1 & 1 \\ -1 & 0 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} 3 & 0 \\ 0 & 6 \end{bmatrix},$$

the eigenvalues of QQ' are $\lambda_1 = 3$ and $\lambda_2 = 6$. Let us determine the normalized eigenvectors of QQ'. Consider the equation $[QQ' - \lambda I]X = O$ for $\lambda = 3$ and 6, and let $X' = [x_1, x_2]$ and $O' = [0, 0]$. Then, for $\lambda = 3$, we see that $x_2 = 0$ and for $\lambda = 6$, we note that $x_1 = 0$. Thus, the normalized solutions are

$$U_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \text{ and } U_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \Rightarrow U = [U_1, U_2] = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Note that $-U_1$ or $-U_2$ or $-U_1, -U_2$ will also satisfy all the conditions, and we could take any of these forms for convenience. Now, consider the equation $(Q'Q - \lambda I)X = O$, where $X' = [x_1, x_2, x_3, x_4]$ and $O' = [0, 0, 0, 0]$ for $\lambda = 3, 6$. For $\lambda = 3$, the coefficient matrix is

$$Q'Q - 3I = \begin{bmatrix} -1 & 0 & 1 & 2 \\ 0 & -1 & -1 & 2 \\ 1 & -1 & -2 & 0 \\ 2 & 2 & 0 & 1 \end{bmatrix} \rightarrow \begin{bmatrix} -1 & 0 & 1 & 2 \\ 0 & -1 & -1 & 2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 9 \end{bmatrix}$$

by elementary transformations. Observe that $x_4 = 0$ so that $-x_1 + x_3 = 0$ and $-x_2 - x_3 = 0$. Thus, one solution or an eigenvector corresponding to $\lambda = 3$ and their normalized form are

$$\begin{bmatrix} 1 \\ -1 \\ 1 \\ 0 \end{bmatrix} \Rightarrow V_1 = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 \\ -1 \\ 1 \\ 0 \end{bmatrix}.$$

Now, take $\lambda = 6$ and consider the equation $(Q'Q - 6I)X = O$; the coefficient matrix and its reduced form obtained through elementary transformations are the following:

$$\begin{bmatrix} -4 & 0 & 1 & 2 \\ 0 & -4 & -1 & 2 \\ 1 & -1 & -5 & 0 \\ 2 & 2 & 0 & -2 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & -1 & -5 & 0 \\ 0 & -4 & -1 & 2 \\ 0 & 0 & -21 & 0 \\ 0 & 0 & 9 & 0 \end{bmatrix},$$

which shows that $x_3 = 0$, so that $x_1 - x_2 = 0$ and $-4x_2 + 2x_4 = 0$. Hence, an eigenvector and its normalized form are

$$\begin{bmatrix} 1 \\ 1 \\ 0 \\ 2 \end{bmatrix} \Rightarrow V_2 = \frac{1}{\sqrt{6}} \begin{bmatrix} 1 \\ 1 \\ 0 \\ 2 \end{bmatrix}.$$

Thus, $V = [V_1, V_2]$. As mentioned earlier, we could have $-V_1$ or $-V_2$ or $-V_1, -V_2$ as the normalized eigenvectors. As per our notation,

$$\Lambda = \mathrm{diag}(\sqrt{3}, \sqrt{6}) \text{ and } Q = U\Lambda V'.$$

Let us verify this last equality. Since

$$\begin{aligned} U\Lambda V' &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \sqrt{3} & 0 \\ 0 & \sqrt{6} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & 0 & \frac{2}{\sqrt{6}} \end{bmatrix} \\ &= \begin{bmatrix} 1 & -1 & 1 & 0 \\ 1 & 1 & 0 & 2 \end{bmatrix}, \end{aligned}$$

we should take $-V_1$ to obtain Q. Then,

$$[U_1, U_2] \begin{bmatrix} \sqrt{3} & 0 \\ 0 & \sqrt{6} \end{bmatrix} \begin{bmatrix} -V_1' \\ V_2' \end{bmatrix} = Q,$$

which verifies the result and completes the computations.

Now, we shall continue with our row and column profile plots. From (15.6.4), we have

$$nYY' = \begin{bmatrix} \frac{33}{5} & -\frac{43}{6} & \frac{17\sqrt{2}}{30} \\ -\frac{43}{6} & \frac{103}{12} & -\frac{17\sqrt{2}}{12} \\ \frac{17\sqrt{2}}{30} & -\frac{17\sqrt{2}}{12} & \frac{17}{10} \end{bmatrix} \text{ and } nY'Y = \begin{bmatrix} \frac{63}{12} & -\frac{47\sqrt{30}}{60} & -\frac{11}{3} & \frac{7\sqrt{2}}{3} \\ -\frac{47\sqrt{30}}{60} & \frac{43}{10} & \frac{3\sqrt{30}}{5} & -\frac{16\sqrt{15}}{15} \\ -\frac{11}{3} & \frac{3\sqrt{30}}{5} & \frac{8}{3} & -2\sqrt{2} \\ \frac{7\sqrt{2}}{3} & -\frac{16\sqrt{15}}{15} & -2\sqrt{2} & \frac{14}{3} \end{bmatrix}.$$

The eigenvalues of nYY' are $\lambda_1 = 15.1369$, $\lambda_2 = 1.7471$ and $\lambda_3 = 0$ and the normalized eigenvectors from nYY', corresponding to λ_1, λ_2 and λ_3 are U_1, U_2 , U_3, so that $U = [U_1, U_2, U_3]$ where

$$U = \begin{bmatrix} 4.28 & -0.49 & 1.41 \\ -4.99 & -0.22 & 1.41 \\ 1 & 1 & 1 \end{bmatrix} \text{ and } \Lambda = \mathrm{diag}(\sqrt{15.1369}, \sqrt{1.7471}, 0).$$

For the same eigenvalues λ_1, λ_2 and λ_3, the normalized eigenvectors determined from $nY'Y$, which correspond to the nonzero eigenvalues, are V_1 and V_2, with

$$V = [V_1, V_2] = \begin{bmatrix} 1.10 & -0.84 \\ -1.06 & -0.16 \\ -0.83 & 0.28 \\ 1 & 1 \end{bmatrix}.$$

Since $\lambda_3 = 0$, $k = 2$, and we can take the $r \times k$, that is, 3×2 matrix $G = (g_{ij}) = D_r^{-\frac{1}{2}} U \Lambda$ to represent the row deviation profiles and the $s \times k = 4 \times 2$ matrix $H = (h_{ij}) = D_c^{-\frac{1}{2}} V \Lambda$ to represent the column deviation profiles. For our numerical example, it follows from (15.4.6) that

$$D_r = \text{diag}(0.4, 0.4.0.2) \Rightarrow D_r^{-\frac{1}{2}} = \text{diag}\Big(\frac{1}{0.63}, \frac{1}{0.63}, \frac{1}{0.45}\Big)$$
$$D_c = \text{diag}(0.30, 0.25, 0.30, 0.15) \Rightarrow D_c^{-\frac{1}{2}} = \text{diag}(\frac{1}{0.55}, \frac{1}{0.50}, \frac{1}{0.55}, \frac{1}{0.39})$$
$$\Lambda = \text{diag}(\sqrt{15.1369}, \sqrt{1.7471}, 0) = \text{diag}(3.89, 1.32, 0).$$

We only take the first two columns of U and V since $\lambda_3 = 0$; besides, only the first two vectors are required for plotting. Let $U_{(1)}$ and $V_{(1)}$ represent the first two columns of U and V, respectively. Then, $D_r^{-\frac{1}{2}} U_{(1)} \Lambda$ will be equivalent to multiplying the first and second columns by 3.89 and 1.32, respectively, and multiplying the first and second rows by $\frac{1}{0.63}$ and the third row by $\frac{1}{0.45}$. Then, we have

$$U_{(1)} = \begin{bmatrix} 4.28 & -0.49 \\ -4.99 & -0.22 \\ 1 & 1 \end{bmatrix}, \; D_r^{-\frac{1}{2}} U_{(1)} \Lambda = \begin{bmatrix} 26.42 & -1.03 \\ -30.81 & -0.46 \\ 6.17 & 2.09 \end{bmatrix} \equiv G_2$$

where G_2 is the matrix consisting of the first two columns of G. Hence, the points required for plotting the row profile are: $(26.42, -1.03)$, $(-30.81, -0.46)$, $(6.17, 2.09)$. These points being far apart, no two rows should be combined. Now, consider the column profiles: the effect of $D_c^{-\frac{1}{2}} V_{(1)} \Lambda$ is to multiply the columns of $V_{(1)}$ by 3.89 and 1.32, respectively, and to multiply the rows by $\frac{1}{0.55}, \frac{1}{0.50}, \frac{1}{0.55}, \frac{1}{0.39}$, respectively. Thus,

$$V_{(1)} = \begin{bmatrix} 1.10 & -0.84 \\ -1.06 & -0.16 \\ -0.83 & 0.28 \\ 1 & 1 \end{bmatrix}, \; D_c^{-\frac{1}{2}} V_{(1)} \Lambda = \begin{bmatrix} 7.78 & -2.02 \\ -8.25 & -0.42 \\ -5.87 & -0.67 \\ 9.97 & 3.38 \end{bmatrix} \equiv H_2$$

where H_2 is the matrix consisting of the first two columns of H. The row profile and the column profile points are plotted in Fig. 15.6.1 where r next to a point indicates a row point and c designates a column point. That is, $i\ r$ indicates the i-th row point and $j\ c$, the j-th column point. It can be seen from this plot that the row points are far apart while the second and third column points are somewhat close; accordingly, if necessary, the second and third columns could be combined.

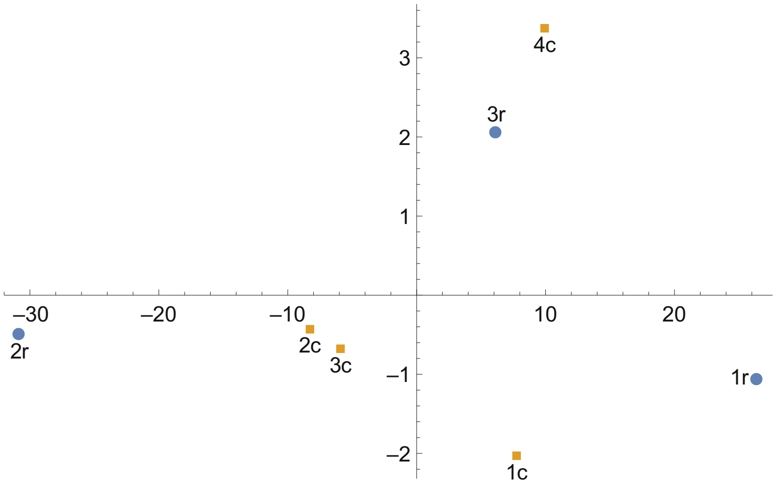

Figure 15.6.1 Row profile and column profile points

15.7. Correspondence Analysis in a Multi-way Contingency Table

When the data is classified under a number of variables, each variable having a number of categories, the resulting frequency table is referred to as a multi-way classification. Correspondence analysis for a multi-way classification involves converting data in a multi-way classification setting into a two-way classification framework and then, employing the techniques developed in Sects. 15.5 and 15.6. The first step in this regard consists of creating an indicator matrix C. In order to illustrate the steps, we will first present an example. Suppose that 10 persons selected at random from a community, are classified according to three variables. Variable 1 is gender. Under this variable, we shall consider the categories male and female. Variable 2 is weight. Under this variable, we are considering three categories: underweight, normal and overweight. The third variable is education which is assumed to have four levels: level 1, level 2, level 3 and level 4. Thus, there are three variables and 9 categories. The actual data are provided in Table 15.7.1.

Table 15.7.1: Ten persons classified under three variables

	Variables		
Person #	Gender	Weight	Educational level
1	Female	Overweight	Level 2
2	Female	Normal	Level 4
3	Male	Underweight	Level 1
4	Female	Normal	Level 3
5	Male	Overweight	Level 1
6	Male	Normal	Level 2
7	Female	Overweight	Level 3
8	Female	Underweight	Level 4
9	Male	Normal	Level 3
10	Female	Overweight	Level 1

Table 15.7.2: Entries of the indicator matrix of the data included in Table 15.7.1

	Variables								
	Gender		Weight			Educational level			
Person #	M	F	U	N	O	L1	L2	L3	L4
1	0	1	0	0	1	0	1	0	0
2	0	1	0	1	0	0	0	0	1
3	1	0	1	0	0	1	0	0	0
4	0	1	0	1	0	0	0	1	0
5	1	0	0	0	1	1	0	0	0
6	1	0	0	1	0	0	1	0	0
7	0	1	1	0	0	0	0	1	0
8	0	1	1	0	0	0	0	0	1
9	1	0	0	1	0	0	0	1	0
10	0	1	0	0	1	1	0	0	0

Next, we construct the indicator matrix C—distinct from C as defined in (15.4.4)—of the data displayed in Table 15.7.1. If an item is present, we write 1 in the corresponding location in Table 15.7.2, and if it is absent, we write 0, thus populating this table where M ≡ Male, F ≡ Female, U ≡ underweight, N ≡ Normal, O ≡ overweight, L1 ≡ Level 1, L2 ≡ Level 2, L3 ≡ Level 3 and L4 ≡ Level 4. The resulting indicator matrix C is

$$C = \begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \end{bmatrix}.$$

Note that since a person will belong to a single category of every variable, the row sum of every row will always be equal to the number of variables, which is 3 in the example. The sum of *all* the column entries under *each* variable is the number of items classified (10 in the example). We now convert the data into a two-way classification, which is achieved by converting C into a *Burt matrix* B, where $B = C'C$. In our example,

$$B = C'C = \begin{bmatrix} 4 & 0 & 1 & 2 & 1 & 2 & 1 & 1 & 0 \\ 0 & 6 & 1 & 2 & 3 & 1 & 1 & 2 & 2 \\ 1 & 1 & 2 & 0 & 0 & 1 & 0 & 0 & 1 \\ 2 & 2 & 0 & 4 & 0 & 0 & 1 & 2 & 1 \\ 1 & 3 & 0 & 0 & 4 & 2 & 1 & 1 & 0 \\ 2 & 1 & 1 & 0 & 2 & 3 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 2 & 0 & 0 \\ 1 & 2 & 0 & 2 & 1 & 0 & 0 & 3 & 0 \\ 0 & 2 & 1 & 1 & 0 & 0 & 0 & 0 & 2 \end{bmatrix}$$

Observe that the diagonal blocks in $C'C$ correspond to the variables, gender, weight and educational level or gender versus gender, weight versus weight, educational level versus educational level. These blocks are the following:

$$\begin{bmatrix} 4 & 0 \\ 0 & 6 \end{bmatrix}, \quad \begin{bmatrix} 2 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix}, \quad \begin{bmatrix} 3 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}.$$

Various two-way contingency tables, namely gender versus weight, gender versus educational level, weight versus educational level, are combined into one two-way table displaying category versus category. The observed Pearson's χ^2 statistic from $C'C$ is seen to be 79.85. In this case, the number of degrees of freedom is $8 \times 8 = 64$ and at 5% level, the tabulated $\chi^2_{64,0.05} \approx 84 > 79.85$; hence, the hypothesis of no association in $C'C$ is not rejected. Note that this χ^2 approximation is unreliable since the expected frequencies are small. The most relevant parts in the Burt matrix $C'C$ are the non-diagonal blocks of frequencies. The two non-diagonal blocks of the first two rows represent the two-way contingency tables for gender versus weight and gender versus educational level. Similarly, the non-diagonal block in the third to fifth rows represent the two-way contingency table for weight versus educational level. These are the following, denoted by A_1, A_2, A_3 respectively, where A_1 is the two-way contingency table of gender versus weight, A_2 is the contingency table of gender versus educational level and A_3 is the table of weight versus educational level:

$$A_1 = \begin{bmatrix} 1 & 2 & 1 \\ 1 & 2 & 3 \end{bmatrix}, \; A_2 = \begin{bmatrix} 2 & 1 & 1 & 0 \\ 1 & 1 & 2 & 2 \end{bmatrix}, \; A_3 = \begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 2 & 1 \\ 2 & 1 & 1 & 0 \end{bmatrix}.$$

The corresponding matrices of expected frequencies, under the hypothesis of no association between the characteristics of classification, denoted by $E(A_i),\ i = 1, 2, 3$ are

$$E(A_1) = \begin{bmatrix} 0.8 & 1.6 & 1.6 \\ 1.2 & 2.4 & 2.4 \end{bmatrix}, \; E(A_2) = \begin{bmatrix} 1.2 & 0.8 & 1.2 & 0.8 \\ 1.8 & 1.2 & 1.8 & 1.2 \end{bmatrix},$$
$$E(A_3) = \begin{bmatrix} 0.6 & 0.4 & 0.6 & 0.4 \\ 1.2 & 0.8 & 1.2 & 0.8 \\ 1.2 & 0.8 & 1.2 & 0.8 \end{bmatrix}.$$

The observed values of Pearson's χ^2 statistic under the hypothesis of no association in the contingency table, and the corresponding tabulaled χ^2 critical values at the 5% significance level, are the following: $A_1 :\ \chi^2 = 0.63,\ \chi^2_{2,0.05} = 5.99 > 0.63$; $A_2 :\ \chi^2 = 2.36,\ \chi^2_{3,0.05} = 7.81 > 2.36$; $A_3 :\ \chi^2 = 5.42,\ \chi^2_{6,0.05} = 12.59 > 5.42$; hence the hypothesis would not be rejected in any of the contingency table if Pearson's statistic were applicable. Actually, the χ^2 approximation is not appropriate in any of these cases since the expected frequencies are quite small. Hence, decisions cannot be made on the basis of Pearson's statistic in these instances.

Observe that the first column of the matrix C corresponds to the count on "Male", the second to the count on "Female", the third to "Underweight", the fourth to "Normal", the fifth to "Overweight", the sixth to "Level 1", the seventh to "Level 2", the eighth to "Level

3" and the ninth to "Level 4". Thus, the columns represent the various characteristics or the various variables and their categories. So, if we were to plot one column as one point in the two-dimensional space, then by looking at the points we could determine which points are close to each other. For example, if the "Overweight" column point is close to the "Male" column point, then there is possibility of association between "Overweight" and "Male". Thus, our aim will be to plot each column of C or each column of $C'C$ as a point in two dimensions. For this purpose, we may make use of the plotting technique described in Sects. 15.5 and 15.6. Consider a singular value decomposition of $C = U\Lambda V'$, $U'U = I_k$, $V'V = I_k$. If C is $r \times s$, $s < r$, then U is $r \times k$ and V is $s \times k$ where k is the number of nonzero eigenvalues of CC' as well as those of $C'C$, and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$ where λ_j^2, $j = 1, \ldots, k$, are the nonzero eigenvalues of CC' and $C'C$. In the numerical example, $r = 10$ and $s = 9$. Consider the eigenvalues of $C'C$ since in this case, the order is smaller than the order of CC'. Let the nonzero eigenvalues of $C'C$ be $\lambda_1^2 \geq \cdots \geq \lambda_k^2$. From $C'C$, compute the normalized eigenvectors corresponding to these nonzero eigenvalues. This $s \times k$ matrix of normalized eigenvectors is V in the singular value decomposition. By using the same nonzero eigenvalues, compute the normalized eigenvectors from CC'. This $r \times k$ matrix is U in the singular value decomposition. Since the columns of C and $C'C$ represent the various variables and their subdivisions, only the columns are useful for our geometrical representation, that is, only V will be relevant for plotting the points. Consider $H = V\Lambda$ and let $\lambda_1^2 \geq \lambda_2^2 \geq \cdots \geq \lambda_k^2$. Observe that $C = U\Lambda V' \Rightarrow C' = V\Lambda U' = HU'$. The rows of C' represent the various variables and their categories. Let $h_1, \ldots, h_s$ be the rows of H. Then, we have

$$\begin{aligned} h_1U' &= \text{Men-row} \\ h_2U' &= \text{Women-row} \\ &\vdots \\ h_sU' &= \text{Level 4-row}. \end{aligned}$$

This shows that the rows $h_1, \ldots, h_s$ represent the various variables and their categories. Since the first two eigenvalues are the largest ones and V_1, V_2 are the corresponding eigenvectors, we can take it for granted that most of the information about the various variables and their categories is contained in the first two elements in $h_1, \ldots, h_s$ or in the first two columns weighted by λ_1 and λ_2. Accordingly, take the first two columns from H and denote this submatrix by $H_{(2)}$ where

$$H_{(2)} = \begin{bmatrix} h_{11} & h_{12} \\ h_{21} & h_{22} \\ \vdots & \vdots \\ h_{s1} & h_{s2} \end{bmatrix}.$$

Plot the points $(h_{11}, h_{12}), (h_{21}, h_{22}), \ldots, (h_{s1}, h_{s2})$. These s points correspond to the s columns in the $r \times s$ matrix C or the s rows in C'.

Referring to our numerical example, the eigenvalues are

$$\lambda_1^2 = 11.66,\ \lambda_2^2 = 5.57,\ \lambda_3^2 = 5.28,\ \lambda_4^2 = 3.47,\ \lambda_5^2 = 2.34,$$
$$\lambda_6^2 = 1, 14,\ \lambda_7^2 = 0.54,\ \lambda_8 = \lambda_9 = 0,$$

so that $k = 7$ and the nonzero eigenvalues, $\sqrt{\lambda_j^2},\ j = 1, \ldots, 7$, are

$$\lambda_1 = 3.41,\ \lambda_2 = 2.36,\ \lambda_3 = 2.30,\ \lambda_4 = 1.86,\ \lambda_5 = 1.53,\ \lambda_6 = 1.07,\ \lambda_7 = 0.73.$$

Thus, the matrix Λ is

$$\Lambda = \mathrm{diag}(3.41, 2.36, 2.30, 1.86, 1.53, 1.07, 0.73),$$

and the

$$\text{total inertia} = 11.66 + 5.57 + 5.28 + 3.47 + 2.34 + 1.14 + 0.54 = 30 = \mathrm{tr}(C'C).$$

Noting that $\frac{11.66}{30} = 0.39$ and $\frac{(11.66+5.57+5.28)}{30} = 0.75$, we can assert that 75% of the inertia is accounted for by the first three eigenvalues of $C'C$.

The normalized eigenvectors of $C'C$, which correspond to the nonzero eigenvalues and are denoted by $V = [V_1, \ldots, V_7]$, are the following:

$$V_1 = \begin{bmatrix} 0.293826 \\ 0.615045 \\ 0.138401 \\ 0.36352 \\ 0.406951 \\ 0.248836 \\ 0.173839 \\ 0.306906 \\ 0.179291 \end{bmatrix},\ V_2 = \begin{bmatrix} -0.711194 \\ 0.512432 \\ -0.0995834 \\ -0.106977 \\ 0.00779839 \\ -0.386116 \\ -0.0833586 \\ 0.041766 \\ 0.228947 \end{bmatrix},\ V_3 = \begin{bmatrix} 0.123362 \\ -0.0941084 \\ -0.0879546 \\ 0.638327 \\ -0.521119 \\ -0.428293 \\ 0.0446188 \\ 0.302598 \\ 0.110329 \end{bmatrix},\ V_4 = \begin{bmatrix} 0.0732966 \\ 0.107206 \\ 0.601988 \\ -0.03252 \\ -0.388966 \\ 0.167134 \\ -0.164402 \\ -0.357001 \\ 0.534772 \end{bmatrix},$$

$$V_5 = \begin{bmatrix} 0.0406003 \\ 0.0238456 \\ -0.124608 \\ 0.110633 \\ 0.0784214 \\ -0.206691 \\ 0.754879 \\ -0.584142 \\ 0.1004 \end{bmatrix},\ V_6 = \begin{bmatrix} -0.115587 \\ -0.0128351 \\ -0.572877 \\ 0.408319 \\ 0.0361355 \\ 0.400243 \\ -0.367304 \\ -0.382451 \\ 0.22109 \end{bmatrix},\ V_7 = \begin{bmatrix} 0.320998 \\ -0.262335 \\ -0.162227 \\ -0.195339 \\ 0.416228 \\ -0.427087 \\ -0.191702 \\ 0.0724589 \\ 0.604993 \end{bmatrix}.$$

Then, the first two eigenvectors weighted by λ_1 and λ_2 and the points to be plotted are

$$\lambda_1 V_1 = 3.41472 \begin{bmatrix} 0.293826 \\ 0.615045 \\ 0.138401 \\ 0.36352 \\ 0.406951 \\ 0.248836 \\ 0.173839 \\ 0.306906 \\ 0.179291 \end{bmatrix} = \begin{bmatrix} 1.00333 \\ 2.10021 \\ 0.4726 \\ 1.24132 \\ 1.38962 \\ 0.849704 \\ 0.593611 \\ 1.048 \\ 0.612228 \end{bmatrix},$$

$$\lambda_2 V_2 = 2.36098 \begin{bmatrix} -0.711194 \\ 0.512432 \\ -0.0995834 \\ -0.106977 \\ 0.00779839 \\ -0.386116 \\ -0.0833586 \\ 0.041766 \\ 0.228947 \end{bmatrix} = \begin{bmatrix} -1.67911 \\ 1.20984 \\ -0.235114 \\ -0.25257 \\ 0.0184118 \\ -0.911613 \\ -0.196808 \\ 0.0986087 \\ 0.540539 \end{bmatrix};$$

$$\text{Points to be plotted :} \begin{bmatrix} (1.00333, -1.67911) \\ (2.10021, 1.20984) \\ (0.4726, -0.235114) \\ (1.24132, -0.25257) \\ (1.38962, 0.0184118) \\ (0.849704, -0.911613) \\ (0.593611, -0.196808) \\ (1.048, 0.0986087) \\ (0.612228, 0.540539) \end{bmatrix} \leftrightarrow \begin{bmatrix} \text{Men} \\ \text{Women} \\ \text{Underweight} \\ \text{Normal} \\ \text{Overweight} \\ \text{Level 1} \\ \text{Level 2} \\ \text{Level 3} \\ \text{Level 4} \end{bmatrix}.$$

The plot of these points is displayed in Fig. 15.7.1.

It is seen from the points plotted in Fig. 15.7.1 that the categories underweight and educational level 2 are somewhat close to each other, which is indicative of a possible association, whereas the categories underweight and women are the farthest apart.

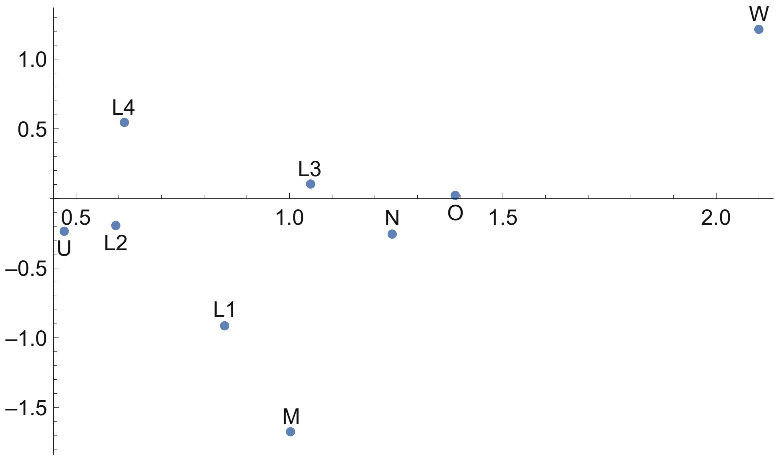

Figure 15.7.1 Multiple contingency plot

Exercises 15 (continued)

15.6. In the following two-way contingency table, where the entries in the cells are frequencies, (1) calculate Pearson's χ^2 statistic and give the representations in (15.5.1)–(15.5.6); (2) plot the row profiles; (3) plot the column profiles:

$\downarrow \rightarrow$	B_1	B_2	B_3	B_4
A_1	10	15	20	15
A_2	15	10	10	5

15.7. Repeat Exercise 15.6 for the following two-way contingency table:

$\downarrow \rightarrow$	B_1	B_2	B_3	B_4
A_1	10	5	15	5
A_2	5	10	10	20
A_3	10	5	10	5
A_4	15	10	5	10

15.8. For the data in (1) Exercise 15.6, (2) Exercise 15.7, and by using the notations defined in Sects. 15.5 and 15.6, compute the following items: Estimates of (i) $A = D_r^{-\frac{1}{2}}(P - RC')D_c^{-1}(P - RC')'D_r^{-\frac{1}{2}}$; (ii) Eigenvalues of A and $\mathrm{tr}(A)$; (iii) Total inertia and proportions of inertia accounted for by the eigenvalues; (iv) The matrix of row-profiles; (v) The matrix of column-profiles, and make comments.

15.9. Referring to Exercises 15.6 and 15.7, plot the row profiles and column profiles and make comments.

15.10. In a used car lot, there are high price, average price and low price cars, the cars come in the following colors: red, white, blue and silver, and the paint finish is either mat or shiny. Fourteen customers bought vehicles from this car lot. Their preferences are given next. (1) Carry out a multiple correspondence analysis, plot the column profiles and make comments; (2) Create individual two-way contingency tables, analyze these tables and make comments. The following is the data where the first column indicates the customer's serial number:

1	Low price	white color	mat finish
2	Low price	red color	shiny finish
3	Average price	silver color	shiny finish
4	High price	red color	shiny finish
5	High price	blue color	shiny finish
6	Average price	white color	mat finish
7	Average price	blue color	mat finish
8	High price	blue color	shiny finish
9	High price	red color	mat finish
10	Average price	silver color	mat finish
11	Low price	white color	shiny finish
12	Average price	white color	mat finish
13	Average price	silver color	shiny finish
14	Low price	white color	shiny finish