24
Methods Related to MDS

In this chapter, two other techniques are discussed that have something in common with MDS. First, we discuss the analysis of a variables-by-objects data matrix by principal components analysis and show how it is related to MDS. Then, we discuss correspondence analysis, a technique particularly suited for the analysis of a contingency table of two categorical variables.

24.1 Principal Component Analysis

Principal component analysis (PCA) is a technique that goes back to Pearson (1901) and Hotelling (1933). It begins with a data matrix of n cases (often: persons) and k variables (often: items, tasks). The objective of the method is to explain the k variables by a much smaller set of m "new" variables that are linear combinations of the original variables. Thus, *new variable* $i = w_1 \cdot (\textit{variable } 1) + w_2 \cdot (\textit{variable } 2) + \cdots + w_k \cdot (\textit{variable } k)$, where the weights, w_j, are the unknowns. The hypothesis is that only a few ($m \ll k$) of these new variables suffice to explain most of the variance of the data. For example, in intelligence testing, the testees are typically asked to work through test batteries with many items. One assumes, however, that not every item requires a special ability to solve it. Rather, only a few abilities should be needed, and each item requires a different mixture of these abilities. Somewhat more formally, one thus wants to (a) find these underlying mixtures of more general components, and then (b) assign each case a score on them. For example, a test battery of an intelligence test may

require essentially only verbal and numerical reasoning (the components), and each testee is assigned a score on these components on the basis of his or her test results. The components are, of course, not identified directly: rather, PCA shows which variables combine with high weights to form one particular component, and then one has to infer from the content of these variables what the component means. This approach is similar to interpreting dimensions in MDS on the basis of the points that have the longest projections onto these dimensions.

Consider an example. Assume that $\mathbf{M}$ is the usual person-by-variable data matrix. We begin by standardizing $\mathbf{M}$ so that its columns all sum to zero and have norms equal to 1. This leads to matrix $\mathbf{Z}$; that is,

$$\mathbf{M} = \begin{bmatrix} 8 & 9 & 1 \\ 5 & 5 & 5 \\ 4 & 4 & 5 \\ 8 & 7 & 2 \\ 7 & 1 & 4 \\ 4 & 5 & 7 \\ 5 & 3 & 6 \\ 2 & 6 & 8 \end{bmatrix} \quad \rightarrow \quad \mathbf{Z} = \begin{bmatrix} .46 & .62 & -.60 \\ -.07 & .00 & .04 \\ -.24 & -.15 & .04 \\ .46 & .31 & -.44 \\ .29 & -.62 & -.12 \\ -.24 & .00 & .36 \\ -.07 & -.31 & .20 \\ -.60 & .15 & .52 \end{bmatrix}. \tag{24.1}$$

To see what PCA does *geometrically*, we plot in Figure 24.1a the persons (=rows) of $\mathbf{Z}$ as points in a 3D space. The axes are formed by the three variables (=columns) of $\mathbf{Z}$. If we rotate the axes, the variance of the projections of the points on the rotated axes will change in general. We know from Section 7.10 that there exists one particular rotation to *principal axes*. These axes are characterized by the property that they are closest to the points or, expressed differently, that the projections of all points onto the principal axes have maximal length, axis by axis in decreasing order. The principal axes give us what we are looking for: the coordinates of the points on the principal axes are the principal components. The variance of the elements of the first principal component (denoted by $\mathbf{k}_1$) is maximal. The second principal axis gives rise to the second PC, $\mathbf{k}_2$, and the third principal axis to the last PC, $\mathbf{k}_3$. Note that each principal axis $\mathbf{k}_a$ may be reflected without changing the variance of the corresponding PC. Thus, any PCA solution is unique up to reflections of its components.

To see how PCA works computationally, consider the (full rank) singular value decomposition $\mathbf{Z} = \mathbf{P}\mathbf{\Lambda}\mathbf{Q}'$; that is,

$$\mathbf{Z} = \begin{bmatrix} -.64 & .32 & -.10 \\ .05 & .03 & -.08 \\ .16 & -.05 & -.80 \\ -.49 & .04 & .22 \\ -.02 & -.75 & .03 \\ .27 & .16 & .48 \\ .20 & -.25 & .25 \\ .46 & .49 & -.00 \end{bmatrix} \begin{bmatrix} 1.46 & 0 & 0 \\ 0 & .92 & 0 \\ 0 & 0 & .20 \end{bmatrix} \begin{bmatrix} -.64 & -.38 & .67 \\ -.38 & .91 & .16 \\ .67 & .15 & .72 \end{bmatrix}, \tag{24.2}$$

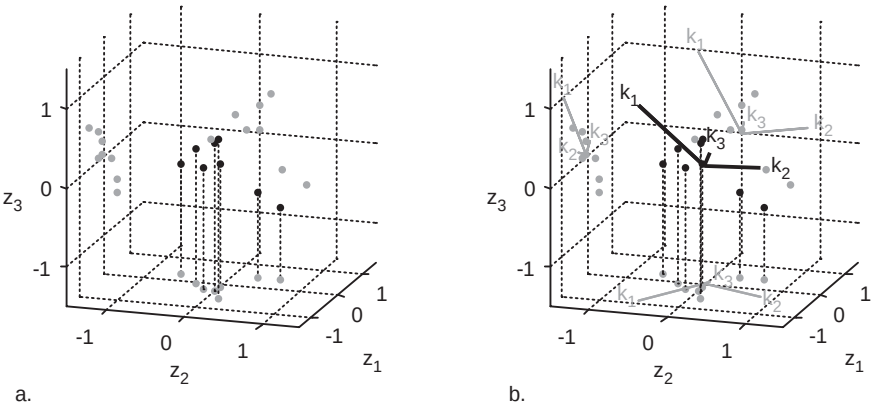

FIGURE 24.1. Plot of persons of $\mathbf{Z}$ in space spanned by variables of $\mathbf{Z}$ (panel a); same space with principal axes shown as tripod $\mathbf{K}$ (panel b).

where $\mathbf{P}$ contains the *standardized* principal components (PCs). The PCs are orthogonal to each other, because $\mathbf{P}'\mathbf{P} = \mathbf{I}$ in any singular value decomposition. The columns of $\mathbf{K} = \mathbf{P}\mathbf{\Lambda}$ are the unstandardized principal components. Figure 24.1b shows the principal axes that generate these PCs—$\mathbf{k}_1, \mathbf{k}_2$, and $\mathbf{k}_3$—in the space of the original variables, the columns of $\mathbf{Z}$. The PCs are related to the original $\mathbf{Z}$ by a rotation/reflection, $\mathbf{K} = \mathbf{Z}\mathbf{Q}$, because $\mathbf{Q}'\mathbf{Q} = \mathbf{Q}\mathbf{Q}' = \mathbf{I}$ in any singular value decomposition.

We can also directly look at the space spanned by the principal axes, where the elements of $\mathbf{K}$ are the coordinates of points that represent the persons. This view is shown in Figure 24.2a, where the tripod of $\mathbf{z}_1, \mathbf{z}_2$, and $\mathbf{z}_3$ indicates how the original variables are oriented in this principal axes space.

The norms of the principal components $\mathbf{k}_1, \mathbf{k}_2$, and $\mathbf{k}_3$ are equal to the respective singular values λ_a on the diagonal of $\mathbf{\Lambda}$. The squared singular values indicate how much variance is accounted for by the various principal components. In our small example, we see that the third PC is very small so that the various person points are almost all located at the same height on the third dimension of Figure 24.2a. Expressed algebraically, the data matrix $\mathbf{Z}$ is decomposed into a sum of matrices each with rank 1, $\lambda_1\mathbf{p}_1\mathbf{q}_1' + \lambda_2\mathbf{p}_2\mathbf{q}_2' + \lambda_3\mathbf{p}_3\mathbf{q}_3'$ (with $\mathbf{q}_a$ column a of $\mathbf{Q}$), so that the first $k < 3$ terms are the best approximation of $\mathbf{Z}$ by a matrix of lower rank k. The singular value λ_a is the weight of the information in the term $\lambda_a\mathbf{p}_a\mathbf{q}_a'$ (see Section 7.6, item 4).

Standardizing $\mathbf{K}$ amounts to adjusting the components $\mathbf{k}_1, \mathbf{k}_2$, and $\mathbf{k}_3$ to length one by dividing each column of $\mathbf{K}$ by the corresponding λ_a. That is, $\mathbf{P} = \mathbf{K}\mathbf{\Lambda}^{-1}$. Geometrically, this operation means that the configuration is stretched or compressed along the axes of Figure 24.2a. The result of this transformation is shown in Figure 24.2b.

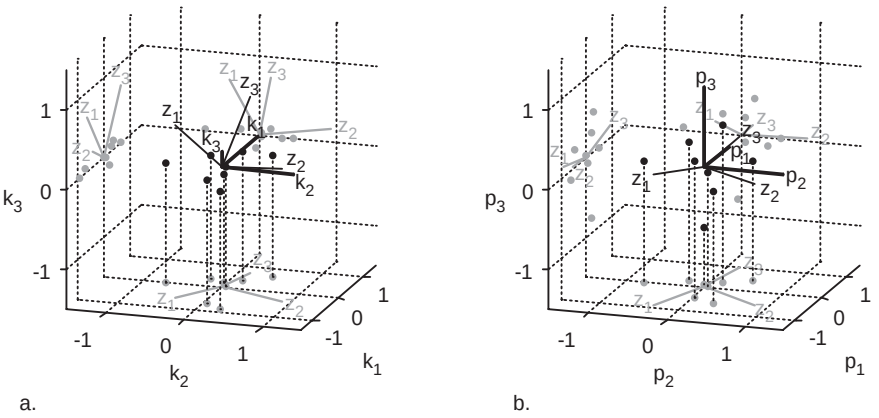

FIGURE 24.2. Persons as points in space that have a principal axes orientation; the axes that correspond to the original variables are shown as the tripod **Z** (panel a). Panel (b) shows same as panel (a), except that space is spanned here by the standardized principal components.

Figure 24.2b can be obtained from the original data as follows. Start with the original row-points configuration **Z**, rotate it to principal axes orientation by **Q**, and then stretch or compress the configuration along the principal axes by the weights in $\mathbf{\Lambda}^{-1}$. Algebraically, this corresponds to computing $\mathbf{P} = \mathbf{ZQ\Lambda}^{-1}$, where the columns of **P** are obviously weighted sums of **Z**'s columns, as intended.

The matrices **Q** and $\mathbf{\Lambda}$ can also be found from an eigendecomposition of the intercorrelation matrix of the original variables, $\mathbf{R} = \mathbf{Z}'\mathbf{Z}$, because $\mathbf{R} = \mathbf{Q\Lambda P}'\mathbf{P\Lambda Q}' = \mathbf{Q\Lambda}^2\mathbf{Q}'$. Thus, the eigenvalues of **R** are equal to the squared singular values of **Z**. A graphical representation of the first two principal axes for our small example is given in Figure 24.3.

Once the components **P** are found, one can reverse the perspective and ask how they explain the original variables. Assuming here that the decomposition has full rank, we can simply reverse the equation $\mathbf{P} = \mathbf{ZQ\Lambda}^{-1}$ to get **Z** from **P** via $\mathbf{Z} = \mathbf{PL}'$ with $\mathbf{L} = \mathbf{Q\Lambda}$. The coefficients in **L** are called *component loadings* and can be interpreted as the correlations between the variables and the components. This property can be seen as follows. The correlations between the variables (columns) of **Z** and **P** are $\mathbf{Z}'\mathbf{P}$, because both **Z** and **P** are standardized. Thus, we get $\mathbf{Z}'\mathbf{P} = \mathbf{Z}'\mathbf{ZQ\Lambda^{-1}} = \mathbf{RQ\Lambda^{-1}} = \mathbf{Q\Lambda}^2\mathbf{Q}'\mathbf{Q\Lambda}^{-1} = \mathbf{Q\Lambda}$. This yields for the example above

$$\mathbf{L} = \begin{bmatrix} -.93 & -.35 & .13 \\ -.55 & .84 & .03 \\ .98 & .14 & .14 \end{bmatrix}.$$

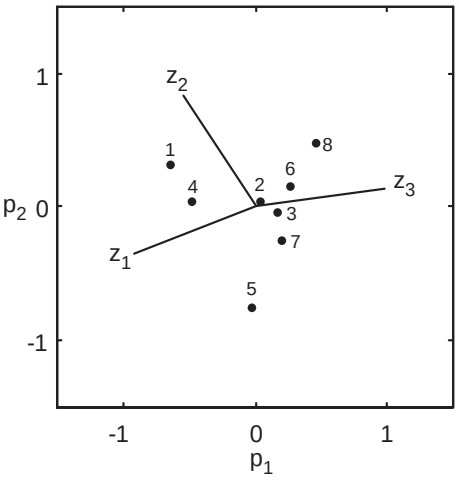

FIGURE 24.3. Persons (labeled by row numbers) in the space of the first two standardized principal components, together with projections of the original variables $\mathbf{z}_1, \mathbf{z}_2$, and $\mathbf{z}_3$ onto this space.

$\mathbf{L}$ shows that the column vector 1 of the data matrix correlates with the first PC with -0.93. It is therefore almost fully explained by this PC. The second variable of the data matrix correlates with the first PC with -0.55, and the third variable with 0.98. Overall, the component loadings make clear that the three variables of our data matrix are essentially only two-dimensional (as Figure 24.2a shows graphically). They correlate most with the first PC, and almost not at all with the third PC.

The loadings can also be interpreted geometrically as the lengths of the projections of the vectors that represent the variables onto the standardized PCs $\mathbf{P}$. The squares of the elements of the component loadings in $\mathbf{L}$ are a measure of fit for the variables (see Table 24.1). The sum of the squared component loadings for dimension a is equal to the eigenvalue λ_a^2. Because the sum-of-squares of the loadings in matrix $\mathbf{L}$ above is 1 in each row, each variable is fully accounted for in the 3D space spanned by the PCs and for 98.7% in 2D.

The simultaneous representation of objects and variables in one plot as in Figure 24.3 is called a *biplot* (Gabriel, 1971). The term *bi* in biplot refers to the representation of the two modes (the objects and the variables) in one plot but not to the dimensionality, although the plots are usually made in two dimensions. The two sets of points, the object points that correspond to the rows of $\mathbf{P}$ and the variables whose coordinates are the component loadings $\mathbf{L} = \mathbf{Q}\mathbf{\Lambda}$, are related as scalar products. This means that we can only interpret the projection of object points on the vector that represents a variable (similar to Figure 16.3), not the distance between an object point and the variable-vector. This projection predicts the value of the object on

TABLE 24.1. Squared component loadings of the example data in (24.1). The last row contains the proportion of variance accounted for (VAF).

	Dimension			Total	
Variable	1	2	3	$1+2$	$1+2+3$
1	.860	.122	.018	.982	1.000
2	.299	.700	.001	.999	1.000
3	.961	.019	.021	.980	1.000
λ_a^2	2.120	.841	.040	2.961	3.000
VAF	.707	.280	.013	.987	1.000

the variable. For more details and some examples of biplots, we refer to Gabriel (1971), Gower and Hand (1996), and Gifi (1990).

A Typical Application of PCA

In many applications, only the structure of the variables is of interest. Then, PCA becomes quite similar to (metric) MDS, because it then reduces to the question of analyzing the structure of a correlation matrix. As an illustration, consider the correlation matrix in Table 5.1. Rather than taking these numbers as similarities and attempting to represent them by distances in a Euclidean space, in PCA we look at the correlations as scalar products. An optimal solution for a PCA representation is easy to find, as was shown above. The loadings of the intelligence test items of Table 5.1 are exhibited in Table 24.2. Overall, these eight variables have a total variance of 8 (geometrically expressed: a total length of 8). Hence, for example, the first three PCs account for $\lambda_1^2 + \lambda_2^2 + \lambda_3^2 = 3.37 + 1.35 + 1.05 = 5.77$ or $(5.77/8) \cdot 100 = 72\%$ of the variance. This follows from the spectral decomposition theorem [see (7.11)], and the convention to norm the eigenvectors to length 1. Note also that the ath PC accounts for a maximum of the variance of the original variables that has not been explained already by the PCs $1, \ldots, a-1$.

Geometrically, we see that the configuration of the variables in the space spanned by the first three PCs, as shown in Figure 24.4, is similar to Figure 5.1. Both exhibit a circular arrangement of the points and vector endpoints, respectively. The PCA representation, however, is higher-dimensional. The (ordinal) MDS representation of Figure 5.1 essentially corresponds to a plane that captures the vector endpoints in Figure 24.4, because in MDS it is the distance of the vectors' endpoints that we want to represent, not the angles that the vectors subtend.

MDS and PCA (in the sense of metrically analyzing a correlation matrix) are, therefore, closely related. However, one cannot always expect similar results. PCA not only leads to higher-dimensional representation spaces than MDS. PCA is also almost always done metrically, whereas most MDS

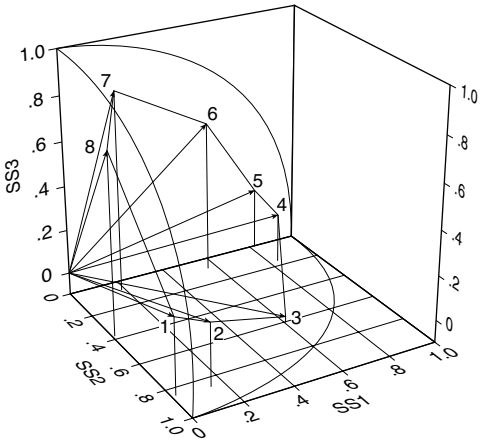

FIGURE 24.4. 3D principal component representation of correlations in Table 5.1, rotated to simple structure.

TABLE 24.2. Loading of variables in Table 5.1 on principal components (PC1, ..., PC8) and on dimensions rotated to simple structure in the space spanned by first three PCs (SS1, SS2, SS3).

Test	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	SS1	SS2	SS3
1	0.63	-0.59	0.15	-0.14	-0.22	-0.02	-0.22	0.33	0.02	0.85	0.24
2	0.69	-0.50	0.24	-0.13	-0.11	-0.04	0.06	-0.41	0.15	0.86	0.18
3	0.70	0.02	0.50	0.06	0.35	0.15	0.31	0.15	0.60	0.61	-0.07
4	0.68	0.48	0.22	0.18	0.11	0.17	-0.42	-0.10	0.84	0.16	0.12
5	0.60	0.57	0.10	-0.20	-0.10	-0.49	0.06	0.04	0.82	0.00	0.18
6	0.70	0.30	-0.29	0.15	-0.44	0.27	0.21	0.03	0.56	0.08	0.58
7	0.57	0.02	-0.61	-0.46	0.28	0.13	-0.02	-0.01	0.18	0.07	0.81
8	0.59	-0.31	-0.46	0.50	0.18	-0.25	0.01	0.00	0.01	0.38	0.72
λ_a^2	3.37	1.35	1.05	0.59	0.51	0.45	0.37	0.31			
Explained variance (%)	42.1	16.9	13.1	7.4	6.4	5.59	4.7	3.9	26.4	25.1	20.7

applications are ordinal ones, in particular those in exploratory data analysis where one wants data representations that are as simple as possible. Moreover, the PCA solution is seldom studied geometrically. Rather, typically only the loadings of the vectors on the components are interpreted, similar to traditional dimension-oriented MDS. In our illustrative application, that means that one would interpret the values of the various tests on the rotated components SS1, SS2, and SS3 but not the circular manifold that we see in Figure 24.4.

Principal Coordinates Analysis

A closely related technique with the same algebraic results as PCA, called *principal coordinates analysis* (PCO), emphasizes the representation of the objects (Gower, 1966). Consider the rows of the data matrix $\mathbf{Z}$ with k variables as points in the k-dimensional space. The aim is to approximate the distances $d_{ij}(\mathbf{Z})$ in a low-dimensional $m < k$ space $\mathbf{X}$. If this is done with classical scaling, then we have to do the following computations. First, compute the matrix of squared distances $\mathbf{D}^{(2)}(\mathbf{Z}) = \mathbf{1c}' + \mathbf{c1}' - 2\mathbf{ZZ}'$, with $\mathbf{c}$ the vector of the diagonal elements of $\mathbf{ZZ}'$; see (7.5). Then pre- and postmultiply $\mathbf{D}^{(2)}(\mathbf{Z})$ with the centering matrix $\mathbf{J}$ and multiply the result with $-\frac{1}{2}$. These operations lead to

$$\begin{aligned} -\tfrac{1}{2}\mathbf{J}\mathbf{D}^{(2)}(\mathbf{Z})\mathbf{J} &= -\tfrac{1}{2}\mathbf{J}(\mathbf{1c}' + \mathbf{c1}' - 2\mathbf{ZZ}')\mathbf{J} \\ &= -\tfrac{1}{2}\mathbf{J}(-2\mathbf{ZZ}')\mathbf{J} = \mathbf{ZZ}'. \end{aligned}$$

Then, the eigendecomposition of $\mathbf{ZZ}' = \mathbf{P}\mathbf{\Lambda}^2\mathbf{P}'$ is computed. The configuration $\mathbf{X}$ for the object points obtained by classical scaling equals the first m columns of $\mathbf{P}\mathbf{\Lambda}$. The configuration obtained from PCO is exactly the same as $\mathbf{K}$ obtained by PCA. Thus, using the normalization $\mathbf{P}\mathbf{\Lambda}$, this equivalence shows that PCA may be seen as MDS that tries to reconstruct distances in a high-dimensional space by a low-dimensional representation.

Of course, instead of using the classical MDS criterion, the high-dimensional distances can also be approximated by using the Stress function in MDS. This approach has been advocated by Meulman (1986, 1992) and is called distance-based PCA. It turns out that the Stress values at a minimum can also be interpreted as a ratio of variances, similar to PCA (Groenen & Meulman, 2004).

24.2 Correspondence Analysis

Correspondence analysis (CA) can be seen as an equivalent of PCA on a contingency table of two categorical variables. In such a table, every entry gives the frequency of each combination of categories of the two variables. The objective of CA is to show the interaction in this table graphically.

TABLE 24.3. A hypothetical contingency table of the distributions of seats by country and political faction (Groenen & Gifi, 1989).

	Political Faction			
Country	Christian Democrats	Socialists	Other	Total
Belgium	8	9	7	24
Germany	39	30	6	75
Italy	25	11	39	75
Luxembourg	3	2	1	6
The Netherlands	13	10	2	25
Total	88	62	55	205

Consider the following hypothetical example. Assume we are interested in the political similarity of some European countries. One set of data that speaks to this issue is the distribution of the seats of these countries in the European Parliament over the political factions. Let Table 24.3 be the hypothetical contingency table of sets for five countries and three political factions. Figure 24.5 shows the result of the correspondence analysis of Table 24.3. In the figure, both the row points (the countries) and the column points (political factions) are plotted. The distance between row points is a particular form of similarity of the countries. For example, The Netherlands and Germany have the same relative distribution of seats over the political factions (see Table 24.4). That is, they have the same data "profile" (Greenacre, 1984, p. 55). Zero distances in CA always occur for profiles that are exactly the same. The properties of these two countries are similar to the profile of Luxembourg and thus are located close to each other but not at zero distance. The centroid can be interpreted as the average country, so that the closer a country is located towards the centroid, the more similar the country is to the average country. Italy and Belgium differ from the other countries because they are not located close together. Note that the scatter of the country points is almost exclusively along the first dimension, indicating that the second dimension is of minor importance. The distance between column points along each axis can be interpreted in a similar way, but the distance between country points and party points has to be interpreted with some care. We return to this later when discussing the example at the end of this section.

Although CA is often applied to contingency tables, the method can in principle be used on any rectangular table with nonnegative similarity values. For example, CA can be used on preference rankings and could be used as an alternative to unfolding. (If used this way, the entries in the table should be similarities, though.)

CA is known under different names, such as reciprocal averaging, dual scaling, canonical correlation analysis (applied to qualitative data), and simultaneous regression, because it has been discovered independently in dif-

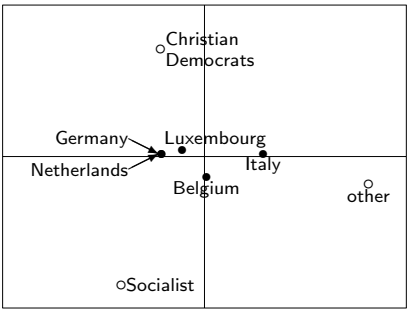

FIGURE 24.5. The correspondence analysis solution of Table 24.3. Note that the points for Germany and the Netherlands are located on top of each other.

TABLE 24.4. Row profiles of Table 24.3.

	Political Faction			
Country	Christian Democrats	Socialists	Other	Total
Belgium	.333	.375	.292	1
Germany	.520	.400	.080	1
Italy	.333	.147	.520	1
Luxembourg	.500	.333	.167	1
The Netherlands	.520	.400	.080	1
Mean row profile	.429	.302	.268	

ferent areas (Hotelling, 1933; Richardson & Kuder, 1933; Hirschfeld, 1935). Guttman (1941) presented a comprehensive treatment of the algebra of CA. The graphical and geometric emphasis in CA has been largely due to Benzécri et al. (1973), a book that also contains a historical overview.[1] There is a wide literature on CA, and standard textbooks are: Nishisato (1980, 1994), Lebart, Morineau, and Warwick (1984), Greenacre (1984, 1994), and Gifi (1990). Developments on CA can be found in Greenacre and Blasius (1994) and Blasius and Greenacre (1998). For a discussion of the relation of CA with MDS, we refer to Heiser and Meulman (1983a). Groenen and Van de Velden (2004) discuss the inverse CA problem, that is, given a CA solution what data sets would have produced the same CA solution.

The remainder of this section is organized as follows. First, we consider the geometry of CA following the example of Groenen and Gifi (1989), also discussed in SPSS (1990). Then, it is shown how the CA solution can be computed. Also, several algebraic properties of CA are discussed such as the inertia, the contribution of a point to the inertia of a dimension, and the proportion of total distance of a point shown in a dimension. Next, we apply CA to crime rates in 10 US states. Finally, we end with some remarks on the relation of CA and MDS.

Geometry of Correspondence Analysis

To measure the similarity between two countries, correspondence analysis uses (row) profiles normed to sum to one in each row. For example, the Christian Democrats occupy 52% (39/75 = .520) of Germany's seats in the European Parliament. Table 24.4 contains the row profiles of Table 24.3. From the row profiles, we see that the Netherlands and Germany have the same relative distribution of seats over the factions, irrespective of their difference in the total number of seats. Now, we discuss how to reconstruct geometrically the CA solution of Figure 24.5 in three steps.

1. Consider Table 24.4 as coordinates in a 3D space (Figure 24.6). The mean row profile is represented as the centroid. Because the profiles sum to one, all of the points lie in the 2D subspace spanned by the points representing the political factions: point (1, 0, 0) for Christian Democrats, point (0, 1, 0) for Socialist, and point (0, 0, 1) for Other. This 2D triangle is shown in Figure 24.7.

2. The next step in correspondence analysis is to assign weights to the dimensions. Let $\mathbf{F} = (f_{ij})$ be the contingency table, such as Table 24.3. In CA, a weighted Euclidean distance is used, where the dimen-

[1] Other historical overviews can be found in Nishisato (1980), Van Rijckevoorsel and Tijssen (1987), Van Rijckevoorsel (1987), and Gifi (1990).

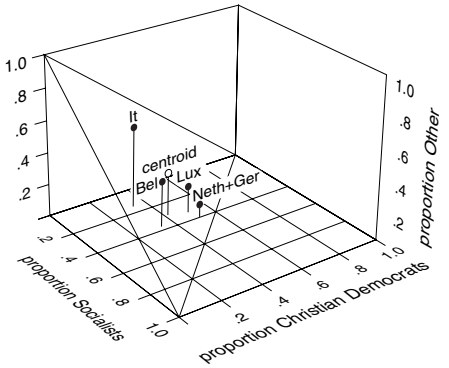

FIGURE 24.6. 3D representation of the row profiles from Table 24.4.

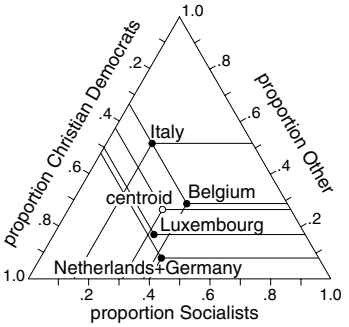

FIGURE 24.7. 2D representation of the row profiles from Table 24.4.

sion weights are equal to $(\sum_j f_{ij}/\sum_{ij} f_{ij})^{-1/2}$, that is, the inverse of the square root of the column means of Table 24.4. In CA, the columns with small means are considered to be more discriminating than the columns with large means. Hence, the weight for column 1 is $1/\sqrt{.429} = 1.527$, for column 2 it is $1/\sqrt{.302} = 1.820$, and for column 3 it is $\sqrt{.268} = 1.932$. The weighted configuration is shown in Figure 24.8. This configuration is the same as the solution obtained by CA in Figure 24.5, apart from the rotation.

3. The final step is to rotate to principal axes such that maximum variance is shown in the first dimension, the second dimension maximizing the remaining variance, and so on.

These three steps show geometrically how a correspondence analysis solution is obtained. The emphasis in these steps was on the row points. The role of the rows and columns can be reversed by simply transposing the correspondence table. Next, we discuss some of the algebraic properties of correspondence analysis.

Algebraic Properties

The weighted Euclidean distance used in CA has a close relation with the χ^2-statistic and so-called χ^2-distances, provided the entries in the correspondence table are frequencies. Let $f_{i+} = \sum_j f_{ij}$ be the row sum of $\mathbf{F}$, $f_{+j} = \sum_i f_{ij}$ the column sum, and $n = \sum_{ij} f_{ij}$ the total sum. The weighted Euclidean distance of row profiles k and l (the distances between the points in Figure 24.8) is given by

$$d_{kl} = \left(\sum_j \frac{(f_{kj}/f_{k+} - f_{lj}/f_{l+})^2}{f_{+j}/n} \right)^{1/2}, \tag{24.3}$$

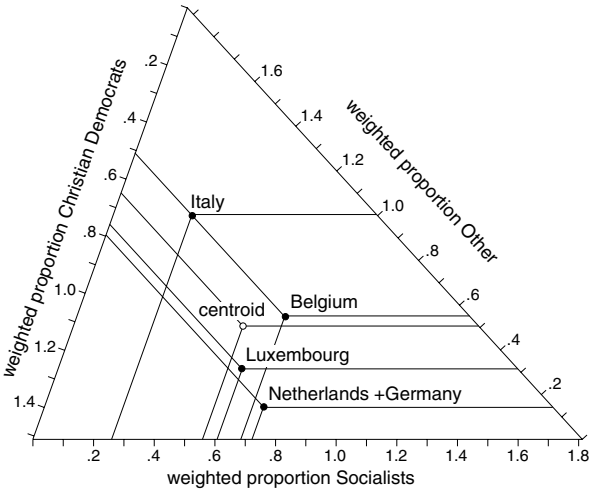

FIGURE 24.8. The weighted Euclidean space used by correspondence analysis.

and the weighted Euclidean distance of row profile k to the average profile z by

$$d_{kz} = \left(\sum_j \frac{(f_{kj}/f_{k+} - f_{+j}/n)^2}{f_{+j}/n} \right)^{1/2}. \tag{24.4}$$

These distances are called χ^2-*distances* because

$$\begin{aligned} \sum_i \frac{f_{i+}}{n} d_{iz}^2 &= \left(\sum_{i,j} \frac{(f_{i+}/n)(f_{ij}/f_{i+} - f_{+j}/n)^2}{f_{+j}/n} \right) \\ &= n^{-1} \left(\sum_{i,j} \frac{(f_{ij} - f_{i+}f_{+j}/n)^2}{f_{i+}f_{+j}/n} \right) = \frac{\chi^2}{n}. \end{aligned}$$

Thus, n times the weighted sum of the squared distances of the row points to their centroid (in full dimensionality) is equal to the χ^2-statistic. Expression (24.5) is called *total inertia*.

We continue discussing how the coordinates in correspondence analysis are obtained. Let $\mathbf{D}_r$ be the diagonal matrix of row marginals (with diagonal elements f_{i+}) and $\mathbf{D}_c$ the diagonal matrix of column marginals (with diagonal elements f_{+j}). Let matrix $\mathbf{E}$ be the matrix of expected values under the *independence model*, which has elements $e_{ij} = f_{i+}f_{+j}/n$. Then, correspondence analysis requires the singular value decomposition of

$$\mathbf{D}_r^{-1/2}(\mathbf{F} - \mathbf{E})\mathbf{D}_c^{-1/2} = \mathbf{P}\mathbf{\Phi}\mathbf{Q}', \tag{24.5}$$

with the usual properties $\mathbf{P}'\mathbf{P} = \mathbf{Q}'\mathbf{Q} = \mathbf{I}$ and $\mathbf{\Phi}$ the diagonal matrix of singular values. The rank of the decomposed matrix in (24.5) is at most $M = \min(\text{number of row points, number of column points}) - 1$. The row scores $\mathbf{R}$ and column scores $\mathbf{C}$ are given by

$$\mathbf{R} = n^{1/2}\mathbf{D}_r^{-1/2}\mathbf{P}\mathbf{\Phi} \quad \text{and} \quad \mathbf{C} = n^{1/2}\mathbf{D}_c^{-1/2}\mathbf{Q}. \tag{24.6}$$

This normalization implies $\mathbf{R}'\mathbf{D}_r\mathbf{R} = \mathbf{\Phi}^2$, $\mathbf{C}'\mathbf{D}_c\mathbf{C} = n\mathbf{I}$, and is called the *row principal* by SPSS (1990), because the squared singular values are the weighted sum of the squared row coordinates (after the principal coordinates normalization of Greenacre, 1984, p. 88). For a discussion of other normalizations, we refer to Greenacre (1984) and Gifi (1990).

Properties of this decomposition are:

- The weighted sum of the row scores (weights $\mathbf{D}_r$) and the weighted sum of the column scores (weights $\mathbf{D}_c$) are equal to zero. The origin is the average row (and column) profile.

- The term $\sum_a \phi_a^2$ is called the *inertia*. In our example, we have perfect fit, so that all of the inertia is shown in 2D. Inertia is related to the χ^2-statistic by $\chi^2/n = \sum_a \phi_a^2$. Therefore, the proportion of total inertia recovered in m dimensions equals $(\sum_{a=1}^m \phi_a^2)/(\chi^2/n)$.

- The contribution of row point i in recovering the inertia in dimension a is $(f_{i+}/n)r_{ia}^2/\phi_a^2$. For column points, this contribution is $(f_{+j}/n)c_{ja}^2$. The difference in formulas for row and column points stems from the row principal normalization that is used. These relative contributions are important to find those points that are important on dimension a.

- Another interesting measure is the proportion of the χ^2-distance of row i to the centroid that is represented by the coordinate in dimension a. This proportion is given by r_{ia}^2/d_{iz}^2 for the row objects, and $c_{ja}^2\phi_a^2/(\sum_l c_{jl}^2\phi_l^2)$ for the column objects.

- Using the normalization above, the row scores are the weighted centroid of the column scores, which is called the *barycentric principle* (Benzécri et al., 1973). The *transition formulas* allow the transformation of the column scores into row scores and the row scores into column scores by

$$\mathbf{R} = \mathbf{D}_r^{-1}\mathbf{F}\mathbf{C}, \tag{24.7}$$
$$\mathbf{C} = \mathbf{D}_c^{-1}\mathbf{F}'\mathbf{R}\mathbf{\Phi}^{-2}. \tag{24.8}$$

  Note that (24.7) computes the weighted centroid of the column points.

- Because (24.5) is a dimensionwise decomposition, the elements of $\mathbf{F}$ can be *reconstituted* in m dimensions by

$$\widehat{f}_{ij} = (f_{i+}f_{+j}/n)\left[1 + \sum_{a=1}^{m} r_{ia}c_{ja}\right]. \tag{24.9}$$

  If $m = M$ (full dimensionality), then (24.9) reconstitutes $\mathbf{F}$ perfectly.

- In our example, we saw that the Netherlands and Germany have the same row profile which gave yielding equal scores in CA. It turns out that CA also gives the same results if these two rows are aggregated. This principle is called *distributional equivalence* (Benzécri et al., 1973). For our example, this principle implies that the matrix with aggregated frequencies for the Netherlands and Germany,

$$\mathbf{F} = \begin{bmatrix} 8 & 9 & 7 \\ 52 & 40 & 8 \\ 25 & 11 & 39 \\ 3 & 2 & 1 \end{bmatrix},$$

  yields exactly the same correspondence analysis solution as the one obtained in Figure 24.5.

CA can be viewed as the residual analysis of the independence model for a contingency table. If the χ^2-value is significant (the independence model does not hold), then the residuals contain more than noise alone, so that it makes sense to analyze the remaining structure in the residuals by CA. However, if the χ^2-value of the independence model is *not* significant, then the residuals are simply the result of noise, so that CA should be avoided. The view of CA as residual analysis of loglinear models has been advocated by Van der Heijden and De Leeuw (1985) and Van der Heijden, De Falguerolles, and De Leeuw (1989). A maximum likelihood version of CA was proposed by Goodman (1985, 1986), and Gilula and Haberman (1986). For a comparison of these methods, see Van der Heijden, Mooijaart, and Takane (1994).

Crime Rates

To illustrate how CA works, consider Table 24.5 with crime rates of seven offenses of 10 U.S. states (U.S. Statistical Abstract 1970, Bureau of Census: Crime rates per 100,000 people). The 50 states were used in an MDS analysis in Chapter 1, but here we restrict ourselves to the 10 states reported in Table 24.5. The main question is how similar or different the states are with respect to their crime statistics. What criminal offenses characterize the states?

CA on Table 24.5 yields the inertia reported in Table 24.6. The first two dimensions show 72% of the total inertia, 47% in the first dimension

TABLE 24.5. Crime rates per 100,000 people for 10 U.S. states. The rows entries are the criminal offenses and the column entries are the states.

State	AK	AL	AR	HI	IL	MA	NE	NY	TN	WY
Murder	12.2	11.7	10.1	3.6	9.6	3.5	3.0	7.9	8.8	5.7
Rape	26.1	18.5	17.1	11.8	20.4	12.0	9.3	15.5	15.5	12.3
Robbery	71.8	50.3	45.6	63.3	251.1	99.5	57.3	443.3	82.0	22.0
Assault	168.0	215.0	150.0	43.0	187.0	88.0	115.0	209.0	169.0	73.0
Burglary	790.0	763.0	885.0	1456.0	765.0	1134.0	505.0	1414.0	807.0	646.0
Larceny	2183.0	1125.0	1211.0	3106.0	2028.0	1531.0	1572.0	2025.0	1025.0	2049.0
Auto theft	551.0	223.0	109.0	581.0	518.0	878.0	292.0	682.0	289.0	165.0

TABLE 24.6. Singular values ϕ_a and percentage of reconstructed inertia of correspondence analysis on crime rates in Table 24.5.

Dim.	ϕ_a	Inertia ϕ_a^2	Perc. Inertia	Cum. Inertia
1	.195	.038	46.8	46.8
2	.143	.020	25.1	72.0
3	.123	.015	18.6	90.6
4	.086	.007	9.1	99.7
5	.014	.000	0.3	100.0
6	.002	.000	0.0	100.0
Total		.081	100.0	

and 25% in the second dimension. The coordinates for the points are displayed in Figure 24.9. Because the row principal normalization is used, the crimes are the weighted average of the points representing the states. The predicted profile (or reconstructed profile) for a state consists of the projections of the criminal offenses points onto the line through the origin and a state. For example, the projections on the line through the origin and MA (Massachusetts) (see Figure 24.9) show that auto theft and robbery happen more often than average. Because larceny and burglary project almost on the origin, they occur at an average rate in Massachusetts, whereas murder, rape, and assault are below average. Robbery (and to a lesser extent assault) happens in New York (NY) more often than average, and larceny less than average. In contrast, Nebraska (NE), Wyoming (WY), and Hawaii (HI) have the opposite profile compared to NY. Murder happens more often than average in the Southern states of Arkansas (AR), Alabama (AL), and Tennessee (TN). The first dimension seems to be dominated by states with robbery (on the right) versus states with more than average larceny. The second axis shows crimes with physical violence (bottom) versus property crimes (top).

Detailed results for the row and column points are given in Table 24.7. The second column gives the so-called mass ($\mathbf{D}_r/n$ and $\mathbf{D}_c/n$ for the rows

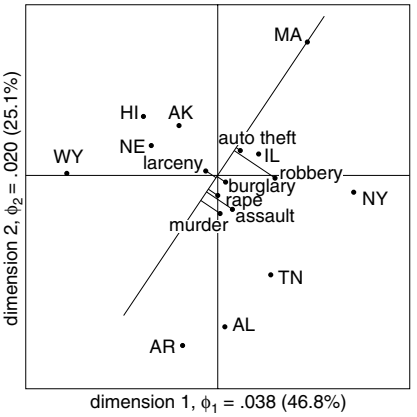

FIGURE 24.9. The correspondence analysis solution of the crime data for 10 U.S. states reported in Table 24.5.

and columns, respectively) weighting the importance of each point in the CA solution. The next two columns give the row scores **R** and column scores **C**. Then, column d_{iz} and d_{jz} shows how much of the each point contributes to the total inertia of .018 in the full dimensional space. We see that the crimes murder and rape and the states AK, IL, and NE hardly determine the CA solution as their contributions to the total inertia are very low. The next two columns show the contribution of each point to the total inertia of a dimension. For example, the first dimension is mostly determined by the crimes robbery and larceny in the states NY, WY, and to a lesser extent in HI and MA. The second dimension is mainly determined by the crimes assault and auto theft in the states AL, AR, and MA. Even though points may not determine the dimension, it may still be that a reasonable proportion of the inertia of a point is shown in that dimension. The last column shows the proportion of the inertia d_{iz} and d_{jz} that is shown in both dimensions. We see that the inertia of all crimes is reasonably well recovered in these two CA dimensions, because their total proportion of inertia d_{iz} and d_{jz} recovered in two dimensions is varying from 42.1% to 87.9%. The same is true for the states with the exception of IL of which only 19.2% of its inertia is shown in these two dimensions. Therefore, IL should be excluded from the interpretation of this CA solution.

The MDS analysis in Chapter 1 (Figure 1.1) yields similar results. In both analyses, we find that violent crimes (rape, assault, murder) are close together as opposed to the property crimes. These results seem more pronounced in the ordinal MDS solution. We have to bear in mind, though, that the MDS solution was based on the full data set, whereas the correspondence analysis solution was based on 10 states only.

TABLE 24.7. Results for row and column points of the correspondence analysis on crime rates in Table 24.5.

					Contr. to Dim.		Prop. d_{iz} Shown		
Crime	$\mathbf{D}_r/n$	$\mathbf{r}_1$	$\mathbf{r}_2$	d_{iz}	Dim 1	Dim 2	Dim 1	Dim 2	Total
Murder	.002	.012	-.470	.001	.000	.024	.000	.642	.642
Rape	.005	-.015	-.258	.001	.000	.015	.001	.420	.421
Robbery	.035	.656	-.055	.023	.393	.005	.648	.005	.653
Assault	.041	.157	-.421	.012	.027	.360	.086	.621	.708
Burglary	.268	.077	-.101	.011	.042	.133	.150	.253	.403
Larceny	.523	-.156	.029	.015	.333	.021	.848	.029	.876
Auto theft	.126	.249	.268	.019	.205	.442	.408	.471	.879
Total	1.000			.081	1.000	1.000			

					Contr. to Dim.		Prop. d_{jz} Shown		
State	$\mathbf{D}_c/n$	$\mathbf{c}_1$	$\mathbf{c}_2$	d_{jz}	Dim 1	Dim 2	Dim 1	Dim 2	Total
AK	.111	-.469	.561	.003	.024	.035	.273	.210	.483
AL	.070	.068	-1.797	.006	.000	.227	.002	.740	.742
AR	.071	-.427	-2.016	.008	.013	.289	.064	.772	.836
HI	.154	-.886	.668	.008	.121	.069	.547	.167	.713
IL	.111	.464	.228	.005	.024	.006	.170	.022	.192
MA	.110	1.035	1.540	.015	.118	.260	.303	.360	.663
NE	.075	-.794	.327	.003	.047	.008	.579	.053	.632
NY	.140	1.578	-.220	.017	.350	.007	.785	.008	.793
TN	.070	.606	-1.188	.004	.026	.099	.245	.507	.752
WY	.087	-1.784	.001	.011	.277	.000	.930	.000	.930
Total	1.000			.081	1.000	1.000			

Comparing CA and MDS

Correspondence analysis has several properties in common with MDS but differs on other aspects. Both techniques graphically display the objects as points in a low-dimensional space. In its basic form, MDS is a one-mode technique (only one set of objects is analyzed), whereas CA is a two-mode technique (row and column objects are displayed, as in unfolding). The data in CA are restricted to be nonnegative, whereas MDS can process more types of data: nonnegative or negative, frequencies, correlations, ratings, rankings, and so on. In addition, MDS can optimally transform the data. For contingency tables (the most widely used type of data for CA) the nonnegativity restriction does not pose a problem, because frequencies between two categorical variables are always nonnegative. CA uses the χ^2-distance as a dissimilarity measure, whereas MDS can accept any dissimilarity or similarity measures (see Chapter 6). In MDS (and unfolding), the distances between all points can be directly interpreted, but in CA this is so only for either the row or the column points. The relation between row and column points can only be assessed by projection (as in Figure 24.9). Therefore, a CA solution has to be interpreted with some care, analogous to non-Euclidean MDS solutions.

There exists a close relation between CA and Classical Scaling. Let $\mathbf{\Delta}$ contain the χ^2-distances between the rows. Let the centering matrix $\mathbf{J}$ be replaced by the weighted centering matrix $\mathbf{J}_w = \mathbf{I} - (\mathbf{1}\mathbf{D}_r\mathbf{1})^{-1}\mathbf{1}\mathbf{1}'\mathbf{D}_r$, so that $\mathbf{J}_w\mathbf{X}$ has weighted mean zero (see Section 12.3). Then, the eigendecomposition in classical scaling of $-(1/2)\mathbf{J}_w\mathbf{\Delta}^{(2)}\mathbf{J}_w'$ yields exactly the same solution for the row scores as does correspondence analysis.

Applying MDS to this $\mathbf{\Delta}$ gives an even higher proportion of explained inertia than CA. Gifi (1990) discusses the decomposition of the χ^2-distances with Stress for binary $\mathbf{F}$. For a general $\mathbf{F}$, setting $\delta_{ik} = d_{ik}$ as defined in (24.3) and setting the weight $w_{ik} = n/(f_{i+}f_{+k})$ gives a decomposition of the χ^2-distances by MDS. If the MDS algorithm uses the CA solution as a start configuration, then the final MDS solution always gives a better reconstruction of the χ^2-distances than CA. One drawback of using MDS (on the matrix of χ^2-distances) instead of CA is that the MDS solution only displays the row points, not the column points. If $\mathbf{X}$ has weighted sum zero, that is, $\sum_i f_{i+}x_{ia} = 0$ for each dimension a, then the origin represents the average row profile, just as in correspondence analysis.

24.3 Exercises

Exercise 24.1 Consider the matrix below. It shows correlations (multiplied by 100) among 13 work value items described in Table 5.2. The lower (upper) half of the matrix is based on a representative survey of the East

(West) German workforce. Note that in this study, no data were gathered on work value item 10.

No.	Work Value	1	2	3	4	5	6	7	8	9	11	12	13	14
1	Interesting job		47	43	38	28	37	29	28	27	16	15	21	28
2	Independent work	51		53	31	27	34	23	25	28	25	16	15	26
3	Much responsibility	42	57		39	32	42	38	38	41	24	16	09	25
4	Meaningful job	37	30	33		20	33	38	44	29	24	13	08	33
5	Chances for advancement	28	29	33	18		43	19	25	15	39	52	27	34
6	Respected job	18	23	34	24	43		37	39	29	37	29	21	35
7	Can help others	20	19	31	33	17	32		48	49	16	10	14	26
8	Useful job	20	17	28	40	18	37	56		32	23	16	18	30
9	Contact with other people	31	34	39	31	21	24	43	34		16	11	10	19
11	Secure position	14	17	18	19	39	37	24	25	17		40	18	38
12	High income	20	26	25	05	54	32	05	08	11	32		27	29
13	Much spare time	25	22	13	09	19	30	13	18	19	16	30		25
14	Healthy working cond.	32	31	23	37	25	20	25	23	24	33	16	23	

(a) Analyze both of these correlation matrices via PCA (varimax rotation) and interpret the resulting component loadings. Do the components correspond to any of the facets of Table 5.2?

(b) What type of facets—axial, modular, or polar facets—can or cannot be seen in a PCA of item correlations?

(c) Take the facet "Alderfer" in Table 5.2, for example. Imagine we had many "material" work values, but only very few "growth" and "relational" work values, respectively. How would this affect an attempt to verify a facet classification via PCA and via MDS analysis with regional interpretations, respectively? Which approach is less robust against uneven item sampling, and why?

(d) Fit the two PCA solutions by Procrustean methods to each other. Which transformations are admissible for PCA solutions and why are they? Which ones are not and why not?

Exercise 24.2 Assume that it takes two abilities, AR = "ability to read well" and AM = "ability for mathematics", to perform well in the tasks $T_1, \ldots, T_4$. The table below shows the ability scores for five persons, their respective performance in four tasks (T_i), and the measured performance in these tests (T_i^*). The scores were constructed as follows: let $T_i = a \cdot AR + b \cdot AM$, where $a + b = 1$; let $T_i^* = T_i + \text{error}$.

Person	AR	AM	T_1	T_2	T_3	T_4	T_1^*	T_2^*	T_3^*	T_4^*
1	10	4	8.80	8.20	5.80	4.00	8.06	7.50	7.01	3.69
2	5	3	4.60	4.40	3.60	3.00	3.93	3.86	4.63	0.93
3	2	7	3.00	3.50	5.50	7.00	3.24	3.91	6.43	5.11
4	5	9	5.80	6.20	7.80	9.00	4.19	4.48	8.70	8.76
5	9	10	9.20	9.30	9.70	10.00	9.51	9.93	11.04	9.92

(a) Take the observed task scores, $T_1^*, \ldots, T_4^*$, intercorrelate them, and find out by using PCA how these scores can best be explained through (possibly rotated) latent factors.

(b) Compute the components and compare them with the true AR- and AM-scores, respectively.

(c) Sketch the vector configuration of the four tasks in 2D, both in principal axes orientation and in varimax orientation.

(d) Intercorrelate $T_1^*, \ldots, T_4^*$ and do an MDS analysis of these correlations. Compare the result with the PCA solution.

Exercise 24.3 Bendixen (1996) reports frequencies of 14 statements on 8 breakfast items judged by a sample of 100 housewives (see the table below). The breakfast items are Cereals (CER), Muesli (MUE), Porridge (POR), Bacon and eggs (B&E), Toast and tea (T&T), Fresh fruit (FRF), Stewed fruit (STF), and Yoghurt (YOG). Note that each respondent could choose more than one statement for each breakfast item.

		Breakfast Item							
no.	Statement	CER	MUE	POR	B&E	T&T	FRF	STF	YOG
1	Healthy	14	38	25	18	8	31	28	34
2	Nutritious	14	28	25	25	7	32	26	31
3	Good in summer	42	22	11	13	7	37	16	35
4	Good in winter	10	10	32	26	6	11	19	8
5	Expensive	6	33	5	27	3	9	18	10
6	Quick and easy	54	33	8	2	15	26	8	20
7	Tasty	24	21	16	34	11	33	26	26
8	Economical	24	3	20	3	16	7	3	7
9	For a treat	5	3	3	31	4	4	16	17
10	For weekdays	47	24	15	9	13	11	6	10
11	For weekends	12	5	8	56	16	10	23	18
12	Tasteless	8	6	2	2	0	0	2	1
13	Takes too long to prepare	0	0	9	35	1	0	10	0
14	Family favorite	14	4	10	31	5	7	2	5

(a) What items and statements do you expect to influence the CA solution most? Why do you think so?

(b) Apply CA to the matrix of frequencies above. (You can use, for example, the CORRESPONDENCE program in SPSS.) How many dimensions do you choose? How much of the total inertia is accounted for by these dimensions?

(c) Interpret the most important relations in the CA solution. (Hint: focus on a statement and look which breakfast items are more and less than average characterized by this statement.)

(d) What items and statements are not well represented in this CA solution? Do you need to revise your interpretation at (c)?

(e) Remove the bad fitting items and statements and redo CA. Interpret the solution. Are the relations different from the CA solution in (b)? If so, explain the differences.

Exercise 24.4 Consider the data on the interpretations of Rorschach inkblot pictures reported in Exercise 15.3 on p. 332.

(a) Do CA on these data. How many dimensions do you choose?

(b) Identify good and bad fitting row and column points. What measures do you use for doing so?

(c) Interpret the CA solution. What are the most important relations in these data?

Exercise 24.5 PCA can also be attempted "by hand".

(a) Consider the correlation matrix 5.1. Convert the correlations into angles among pairs of vectors. [Hint: For example, for $r_{12} = .67$ in Table 5.1, the angle of the corresponding geometric vectors for items 1 and 2 is $\arccos(.67) = 47.9°$.]

(b) With this angle information, construct a vector representation of the correlations. First, take eight knitting needles, straws, sticks, or the like. Then stick needle 1 into a styrofoam ball and then needle 2 such that it forms an angle of 47.9° with 1. Then proceed with needle 3, and so on.

(c) Compare your result to a solution arrived at by computation.