Chapter 15
Correspondence Analysis

Correspondence analysis provides tools for analysing the associations between rows and columns of contingency tables. A contingency table is a two-entry frequency table where the joint frequencies of two qualitative variables are reported. For instance a (2×2) table could be formed by observing from a sample of n individuals two qualitative variables: the individual's sex and whether the individual smokes. The table reports the observed joint frequencies. In general $(n \times p)$ tables may be considered.

The main idea of correspondence analysis is to develop simple indices that will show the relations between the row and the columns categories. These indices will tell us simultaneously which column categories have more weight in a row category and vice versa. Correspondence analysis is also related to the issue of reducing the dimension of the table, similar to principal component analysis in Chap. 11, and to the issue of decomposing the table into its factors as discussed in Chap. 10. The idea is to extract the indices in decreasing order of importance so that the main information of the table can be summarised in spaces with smaller dimensions. For instance, if only two factors (indices) are used, the results can be shown in two-dimensional graphs, showing the relationship between the rows and the columns of the table.

Section 15.1 defines the basic notation and motivates the approach and Sect. 15.2 gives the basic theory. The indices will be used to describe the χ^2 statistic measuring the associations in the table. Several examples in Sect. 15.3 show how to provide and interpret, in practice, the two-dimensional graphs displaying the relationship between the rows and the columns of a contingency table.

W.K. Härdle, L. Simar, *Applied Multivariate Statistical Analysis*,
DOI 10.1007/978-3-662-45171-7_15

15.1 Motivation

The aim of correspondence analysis is to develop simple indices that show relations between the row and columns of a contingency tables. Contingency tables are very useful to describe the association between two variables in very general situations. The two variables can be qualitative (nominal), in which case they are also referred to as categorical variables. Each row and each column in the table represents one category of the corresponding variable. The entry x_{ij} in the table $\mathcal{X}$ (with dimension $(n \times p)$) is the number of observations in a sample which simultaneously fall in the ith row category and the jth column category, for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. Sometimes a "category" of a nominal variable is also called a "modality" of the variable.

The variables of interest can also be discrete quantitative variables, such as the number of family members or the number of accidents an insurance company had to cover during 1 year, etc. Here, each possible value that the variable can have defines a row or a column category. Continuous variables may be taken into account by defining the categories in terms of intervals or classes of values which the variable can take on. Thus contingency tables can be used in many situations, implying that correspondence analysis is a very useful tool in many applications.

The graphical relationships between the rows and the columns of the table $\mathcal{X}$ that result from correspondence analysis are based on the idea of representing all the row and column categories and interpreting the relative positions of the points in terms of the weights corresponding to the column and the row. This is achieved by deriving a system of simple indices providing the coordinates of each row and each column. These row and column coordinates are simultaneously represented in the same graph. It is then clear to see which column categories are more important in the row categories of the table (and the other way around).

As was already eluded to, the construction of the indices is based on an idea similar to that of PCA. Using PCA the total variance was partitioned into independent contributions stemming from the principal components. Correspondence analysis, on the other hand, decomposes a measure of association, typically the total χ^2 value used in testing independence, rather than decomposing the total variance.

Example 15.1 The French "baccalauréat" frequencies have been classified into regions and different baccalauréat categories, see Chap. 22, Table 22.8. Altogether $n = 202{,}100$ baccalauréats were observed. The joint frequency of the region *Ile-de-France* and the modality *Philosophy*, for example, is 9,724. That is, 9,724 baccalauréats were in Ile-de-France and the category Philosophy.

The question is whether certain regions prefer certain baccalauréat types. If we consider, for instance, the region *Lorraine*, we have the following percentages:

A	B	C	D	E	F	G	H
20.5	7.6	15.3	19.6	3.4	14.5	18.9	0.2

The total percentages of the different modalities of the variable baccalauréat are as follows:

A	B	C	D	E	F	G	H
22.6	10.7	16.2	22.8	2.6	9.7	15.2	0.2

One might argue that the region *Lorraine* seems to prefer the modalities E, F, G and dislike the specialisations A, B, C, D relative to the overall frequency of baccalauréat type.

In correspondence analysis we try to develop an index for the regions so that this over- or underrepresentation can be measured in just one single number. Simultaneously we try to weight the regions so that we can see in which region certain baccalauréat types are preferred.

Example 15.2 Consider n types of companies and p locations of these companies. Is there a certain type of company that prefers a certain location? Or is there a location index that corresponds to a certain type of company?

Assume that $n = 3$, $p = 3$, and that the frequencies are as follows:

$$\mathcal{X} = \begin{pmatrix} 4 & 0 & 2 \\ 0 & 1 & 1 \\ 1 & 1 & 4 \end{pmatrix} \quad \begin{matrix} \leftarrow \text{Finance} \\ \leftarrow \text{Energy} \\ \leftarrow \text{HiTech} \end{matrix}$$

$$\begin{matrix} \uparrow \text{Frankfurt} \\ \uparrow \text{Berlin} \\ \uparrow \text{Munich} \end{matrix}$$

The frequencies imply that four type three companies (HiTech) are in location 3 (Munich), and so on. Suppose there is a (company) weight vector $r = (r_1, \ldots, r_n)^\top$ such that a location index s_j could be defined as

$$s_j = c \sum_{i=1}^{n} r_i \frac{x_{ij}}{x_{\bullet j}}, \tag{15.1}$$

where $x_{\bullet j} = \sum_{i=1}^{n} x_{ij}$ is the number of companies in location j and c is a constant. s_1, for example, would give the average weighted frequency (by r) of companies in location 1 (Frankfurt).

Given a location weight vector $s^* = \left(s_1^*, \ldots, s_p^*\right)^\top$, we can define a company index in the same way as

$$r_i^* = c^* \sum_{j=1}^{p} s_j^* \frac{x_{ij}}{x_{i\bullet}}, \tag{15.2}$$

where c^* is a constant and $x_{i\bullet} = \sum_{j=1}^{p} x_{ij}$ is the sum of the ith row of $\mathcal{X}$, i.e. the number of type i companies. Thus r_2^*, for example, would give the average weighted frequency (by s^*) of energy companies.

If (15.1) and (15.2) can be solved simultaneously for a "row weight" vector $r = (r_1, \ldots, r_n)^\top$ and a "column weight" vector $s = (s_1, \ldots, s_p)^\top$, we may represent each row category by r_i, $i = 1, \ldots, n$ and each column category by s_j, $j = 1, \ldots, p$ in a one-dimensional graph. If in this graph r_i and s_j are in close proximity (far from the origin), this would indicate that the ith row category has an important conditional frequency $x_{ij}/x_{\bullet j}$ in (15.1) and that the jth column category has an important conditional frequency $x_{ij}/x_{i\bullet}$ in (15.2). This would indicate a positive association between the ith row and the jth column. A similar line of argument could be used if r_i was very far away from s_j (and far from the origin). This would indicate a small conditional frequency contribution, or a negative association between the ith row and the jth column.

Summary

↪ The aim of correspondence analysis is to develop simple indices that show relations among qualitative variables in a contingency table.

↪ The joint representation of the indices reveals relations among the variables.

15.2 Chi-Square Decomposition

An alternative way of measuring the association between the row and column categories is a decomposition of the value of the χ^2-test statistic. The well-known χ^2-test for independence in a two-dimensional contingency table consists of two steps. First the expected value of each cell of the table is estimated under the hypothesis of independence. Second, the corresponding observed values are compared to the expected values using the statistic

$$t = \sum_{i=1}^{n} \sum_{j=1}^{p} (x_{ij} - E_{ij})^2 / E_{ij}, \tag{15.3}$$

where x_{ij} is the observed frequency in cell (i, j) and E_{ij} is the corresponding estimated expected value under the assumption of independence, i.e.

$$E_{ij} = \frac{x_{i\bullet}\, x_{\bullet j}}{x_{\bullet\bullet}}. \tag{15.4}$$

Here $x_{\bullet\bullet} = \sum_{i=1}^{n} x_{i\bullet}$. Under the hypothesis of independence, t has a $\chi^2_{(n-1)(p-1)}$ distribution. In the industrial location example introduced above the value of $t = 6.26$ is almost significant at the 5 % level. It is therefore worth investigating the special reasons for departure from independence.

The method of χ^2 decomposition consists of finding the SVD of the matrix $\mathcal{C}$ $(n \times p)$ with elements

$$c_{ij} = (x_{ij} - E_{ij})/E_{ij}^{1/2}. \tag{15.5}$$

The elements c_{ij} may be viewed as measuring the (weighted) departure between the observed x_{ij} and the theoretical values E_{ij} under independence. This leads to the factorial tools of Chap. 10 which describe the rows and the columns of $\mathcal{C}$.

For simplification define the matrices $\mathcal{A}\,(n \times n)$ and $\mathcal{B}\,(p \times p)$ as

$$\mathcal{A} = \mathrm{diag}(x_{i\bullet}) \text{ and } \mathcal{B} = \mathrm{diag}(x_{\bullet j}). \tag{15.6}$$

These matrices provide the marginal row frequencies $a(n \times 1)$ and the marginal column frequencies $b(p \times 1)$:

$$a = \mathcal{A}1_n \text{ and } b = \mathcal{B}1_p. \tag{15.7}$$

It is easy to verify that

$$\mathcal{C}\sqrt{b} = 0 \text{ and } \mathcal{C}^\top\sqrt{a} = 0, \tag{15.8}$$

where the square root of the vector is taken element by element and $R = \mathrm{rank}(\mathcal{C}) \leq \min\{(n-1), (p-1)\}$. From (10.14) of Chap. 10, the SVD of $\mathcal{C}$ yields

$$\mathcal{C} = \Gamma\Lambda\Delta^\top, \tag{15.9}$$

where Γ contains the eigenvectors of $\mathcal{C}\mathcal{C}^\top$, Δ the eigenvectors of $\mathcal{C}^\top\mathcal{C}$ and $\Lambda = \mathrm{diag}(\lambda_1^{1/2}, \ldots, \lambda_R^{1/2})$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_R$ (the eigenvalues of $\mathcal{C}\mathcal{C}^\top$). Equation (15.9) implies that

$$c_{ij} = \sum_{k=1}^{R} \lambda_k^{1/2} \gamma_{ik} \delta_{jk}. \tag{15.10}$$

Note that (15.3) can be rewritten as

$$\mathrm{tr}(\mathcal{C}\mathcal{C}^\top) = \sum_{k=1}^{R} \lambda_k = \sum_{i=1}^{n}\sum_{j=1}^{p} c_{ij}^2 = t. \tag{15.11}$$

This relation shows that the SVD of $\mathcal{C}$ decomposes the total χ^2 value rather than, as in Chap. 10, the total variance.

The duality relations between the row and the column space (10.11) are now for $k = 1, \ldots, R$ given by

$$\begin{aligned} \delta_k &= \tfrac{1}{\sqrt{\lambda_k}} \mathcal{C}^\top \gamma_k, \\ \gamma_k &= \tfrac{1}{\sqrt{\lambda_k}} \mathcal{C} \delta_k. \end{aligned} \tag{15.12}$$

The projections of the rows and the columns of $\mathcal{C}$ are given by

$$\begin{aligned} \mathcal{C}\delta_k &= \sqrt{\lambda_k}\gamma_k, \\ \mathcal{C}^\top \gamma_k &= \sqrt{\lambda_k}\delta_k. \end{aligned} \tag{15.13}$$

Note that the eigenvectors satisfy

$$\delta_k^\top \sqrt{b} = 0, \quad \gamma_k^\top \sqrt{a} = 0. \tag{15.14}$$

From (15.10) we see that the eigenvectors δ_k and γ_k are the objects of interest when analysing the correspondence between the rows and the columns. Suppose that the first eigenvalue in (15.10) is dominant so that

$$c_{ij} \approx \lambda_1^{1/2} \gamma_{i1} \delta_{j1}. \tag{15.15}$$

In this case when the coordinates γ_{i1} and δ_{j1} are both large (with the same sign) relative to the other coordinates, then c_{ij} will be large as well, indicating a positive association between the ith row and the jth column category of the contingency table. If γ_{i1} and δ_{j1} were both large with opposite signs, then there would be a negative association between the ith row and jth column.

In many applications, the first two eigenvalues, λ_1 and λ_2, dominate and the percentage of the total χ^2 explained by the eigenvectors γ_1 and γ_2 and δ_1 and δ_2 is large. In this case (15.13) and (γ_1, γ_2) can be used to obtain a graphical display of the n rows of the table ((δ_1, δ_2) play a similar role for the p columns of the table). The interpretation of the proximity between row and column points will be interpreted as above with respect to (15.10).

In correspondence analysis, we use the projections of weighted rows of $\mathcal{C}$ and the projections of weighted columns of $\mathcal{C}$ for graphical displays. Let $r_k(n \times 1)$ be the projections of $\mathcal{A}^{-1/2}\mathcal{C}$ on δ_k and $s_k(p \times 1)$ be the projections of $\mathcal{B}^{-1/2}\mathcal{C}^\top$ on γ_k $(k = 1, \ldots, R)$:

$$\begin{aligned} r_k &= \mathcal{A}^{-1/2}\mathcal{C}\delta_k = \sqrt{\lambda_k}\mathcal{A}^{-1/2}\gamma_k, \\ s_k &= \mathcal{B}^{-1/2}\mathcal{C}^\top \gamma_k = \sqrt{\lambda_k}\mathcal{B}^{-1/2}\delta_k. \end{aligned} \tag{15.16}$$

These vectors have the property that

$$\begin{aligned} r_k^\top a &= 0, \\ s_k^\top b &= 0. \end{aligned} \tag{15.17}$$

The obtained projections on each axis $k = 1, \ldots, R$ are centred at zero with the natural weights given by a (the marginal frequencies of the rows of $\mathcal{X}$) for the row coordinates r_k and by b (the marginal frequencies of the columns of $\mathcal{X}$) for the column coordinates s_k (compare this to expression (15.14)). As a result, the origin is the centre of gravity for all of the representations. We also know from (15.16) and the SVD of $\mathcal{C}$ that

$$\begin{aligned} r_k^\top \mathcal{A} r_k &= \lambda_k, \\ s_k^\top \mathcal{B} s_k &= \lambda_k. \end{aligned} \tag{15.18}$$

From the duality relation between δ_k and γ_k (see (15.12)) we obtain

$$\begin{aligned} r_k &= \tfrac{1}{\sqrt{\lambda_k}} \mathcal{A}^{-1/2} \mathcal{C} \mathcal{B}^{1/2} s_k, \\ s_k &= \tfrac{1}{\sqrt{\lambda_k}} \mathcal{B}^{-1/2} \mathcal{C}^\top \mathcal{A}^{1/2} r_k, \end{aligned} \tag{15.19}$$

which can be simplified to

$$\begin{aligned} r_k &= \sqrt{\tfrac{x_{\bullet\bullet}}{\lambda_k}} \mathcal{A}^{-1} \mathcal{X} s_k, \\ s_k &= \sqrt{\tfrac{x_{\bullet\bullet}}{\lambda_k}} \mathcal{B}^{-1} \mathcal{X}^\top r_k. \end{aligned} \tag{15.20}$$

These vectors satisfy the relations (15.1) and (15.2) for each $k = 1, \ldots, R$ simultaneously.

As in Chap. 10, the vectors r_k and s_k are referred to as factors (row factor and column factor respectively). They have the following means and variances:

$$\begin{aligned} \overline{r}_k &= \tfrac{1}{x_{\bullet\bullet}} r_k^\top a = 0, \\ \overline{s}_k &= \tfrac{1}{x_{\bullet\bullet}} s_k^\top b = 0, \end{aligned} \tag{15.21}$$

$$\begin{aligned} \mathsf{Var}(r_k) &= \tfrac{1}{x_{\bullet\bullet}} \textstyle\sum_{i=1}^n x_{i\bullet} r_{ki}^2 = \tfrac{r_k^\top \mathcal{A} r_k}{x_{\bullet\bullet}} = \tfrac{\lambda_k}{x_{\bullet\bullet}}, \\ \mathsf{Var}(s_k) &= \tfrac{1}{x_{\bullet\bullet}} \textstyle\sum_{j=1}^p x_{\bullet j} s_{kj}^2 = \tfrac{s_k^\top \mathcal{B} s_k}{x_{\bullet\bullet}} = \tfrac{\lambda_k}{x_{\bullet\bullet}}. \end{aligned} \tag{15.22}$$

Hence, $\lambda_k / \sum_{k=1}^j \lambda_j$, which is the part of the kth factor in the decomposition of the χ^2 statistic t, may also be interpreted as the proportion of the variance explained by the factor k. The proportions

$$C_a(i, r_k) = \frac{x_{i\bullet} r_{ki}^2}{\lambda_k}, \text{ for } i = 1, \ldots, n, \ k = 1, \ldots, R \tag{15.23}$$

are called the absolute contributions of row i to the variance of the factor r_k. They show which row categories are most important in the dispersion of the kth row factors. Similarly, the proportions

$$C_a(j, s_k) = \frac{x_{\bullet j} s_{kj}^2}{\lambda_k}, \text{ for } j = 1, \ldots, p, \ k = 1, \ldots, R \tag{15.24}$$

are called the absolute contributions of column j to the variance of the column factor s_k. These absolute contributions may help to interpret the graph obtained by correspondence analysis.

15.3 Correspondence Analysis in Practice

The graphical representations on the axes $k = 1, 2, \ldots, R$ of the n rows and of the p columns of $\mathcal{X}$ are provided by the elements of r_k and s_k. Typically, two-dimensional displays are often satisfactory if the cumulated percentage of variance explained by the first two factors, $\Psi_2 = \frac{\lambda_1 + \lambda_2}{\sum_{k=1}^{R} \lambda_k}$, is sufficiently large.

The interpretation of the graphs may be summarised as follows:

- The proximity of two rows (two columns) indicates a similar profile in these two rows (two columns), where "profile" refers to the conditional frequency distribution of a row (column); those two rows (columns) are almost proportional. The opposite interpretation applies when the two rows (two columns) are far apart.
- The proximity of a particular row to a particular column indicates that this row (column) has a particularly important weight in this column (row). In contrast to this, a row that is quite distant from a particular column indicates that there are almost no observations in this column for this row (and vice versa). Of course, as mentioned above, these conclusions are particularly true when the points are far away from 0.
- The origin is the average of the factors r_k and s_k. Hence, a particular point (row or column) projected close to the origin indicates an average profile.
- The absolute contributions are used to evaluate the weight of each row (column) in the variances of the factors.
- All the interpretations outlined above must be carried out in view of the quality of the graphical representation which is evaluated, as in PCA, using the cumulated percentage of variance.

Remark 15.1 Note that correspondence analysis can also be applied to more general $(n \times p)$ tables $\mathcal{X}$ which in a "strict sense" are not contingency tables.

As long as statistical (or natural) meaning can be given to sums over rows and columns, Remark 15.1 holds. This implies, in particular, that all of the variables are measured in the same units. In that case, $x_{\bullet\bullet}$ constitutes the total frequency

of the observed phenomenon, and is shared between individuals (n rows) and between variables (p columns). Representations of the rows and columns of $\mathcal{X}$, r_k and s_k, have the basic property (15.19) and show which variables have important weights for each individual and vice versa. This type of analysis is used as an alternative to PCA. PCA is mainly concerned with covariances and correlations, whereas correspondence analysis analyses a more general kind of association. (See Exercises 15.3 and 15.11.)

Example 15.3 A survey of Belgium citizens who regularly read a newspaper was conducted in the 1980s. They were asked where they lived. The possible answers were ten regions: seven provinces (Antwerp, Western Flanders, Eastern Flanders, Hainant, Liège, Limbourg, Luxembourg) and three regions around Brussels (Flemish-Brabant, Wallon-Brabant and the city of Brussels). They were also asked what kind of newspapers they read on a regular basis. There were 15 possible answers split up into three classes: Flemish newspapers (label begins with the letter v), French newspapers (label begins with f) and both languages together (label begins with b). The data set is given in Table 22.9. The eigenvalues of the factorial correspondence analysis are given in Table 15.1.

Two-dimensional representations will be quite satisfactory since the first two eigenvalues account for 81 % of the variance. Figure 15.1 shows the projections of the rows (the 15 newspapers) and of the columns (the ten regions).

As expected, there is a high association between the regions and the type of newspapers which is read. In particular, v_b (Gazet van Antwerp) is almost exclusively read in the province of Antwerp (this is an extreme point in the graph). The points on the left all belong to Flanders, whereas those on the right all belong to Wallonia. Notice that the Wallon-Brabant and the Flemish-Brabant are not far from Brussels. Brussels is close to the centre (average) and also close to the bilingual newspapers. It is shifted a little to the right of the origin due to the majority of French speaking people in the area.

The absolute contributions of the first three factors are listed in Tables 15.2 and 15.3. The row factors r_k are in Table 15.2 and the column factors s_k are in Table 15.3.

Table 15.1 Eigenvalues and percentages of the variance (Example 15.3)

λ_j	Percentage of variance	Cumulated percentage
183.40	0.653	0.653
43.75	0.156	0.809
25.21	0.090	0.898
11.74	0.042	0.940
8.04	0.029	0.969
4.68	0.017	0.985
2.13	0.008	0.993
1.20	0.004	0.997
0.82	0.003	1.000
0.00	0.000	1.000

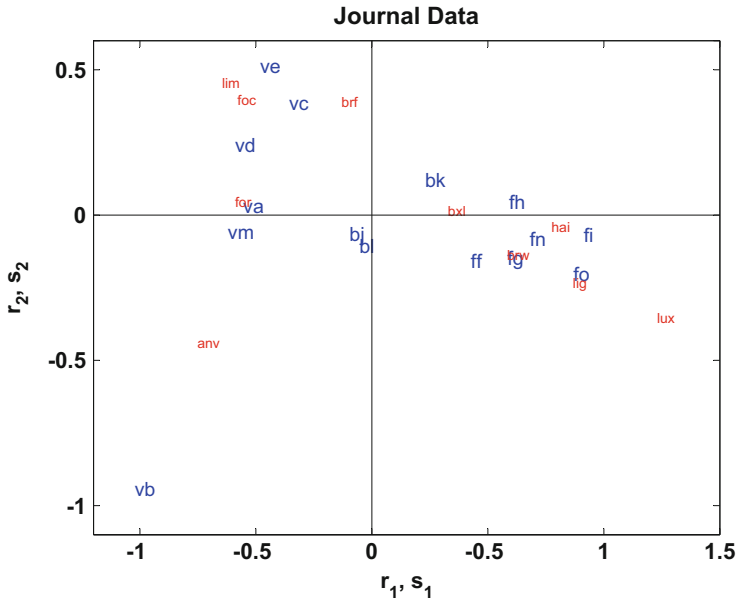

Fig. 15.1 Projection of rows (the 15 newspapers) and columns (the ten regions) `MVAcorrjourn`

Table 15.2 Absolute contributions of row factors r_k

	$C_a(i,r_1)$	$C_a(i,r_2)$	$C_a(i,r_3)$
v_a	0.0563	0.0008	0.0036
v_b	0.1555	0.5567	0.0067
v_c	0.0244	0.1179	0.0266
v_d	0.1352	0.0952	0.0164
v_e	0.0253	0.1193	0.0013
f_f	0.0314	0.0183	0.0597
f_g	0.0585	0.0162	0.0122
f_h	0.1086	0.0024	0.0656
f_i	0.1001	0.0024	0.6376
b_j	0.0029	0.0055	0.0187
b_k	0.0236	0.0278	0.0237
b_l	0.0006	0.0090	0.0064
v_m	0.1000	0.0038	0.0047
f_n	0.0966	0.0059	0.0269
f_0	0.0810	0.0188	0.0899
Total	1.0000	1.0000	1.0000

They show, for instance, the important role of Antwerp and the newspaper v_b in determining the variance of both factors. Clearly, the first axis expresses linguistic differences between the three parts of Belgium. The second axis shows a larger dispersion between the Flemish region than the French speaking regions.

Table 15.3 Absolute contributions of column factors s_k

	$C_a(j, s_1)$	$C_a(j, s_2)$	$C_a(j, s_3)$
brw	0.0887	0.0210	0.2860
bxl	0.1259	0.0010	0.0960
anv	0.2999	0.4349	0.0029
brf	0.0064	0.2370	0.0090
foc	0.0729	0.1409	0.0033
for	0.0998	0.0023	0.0079
hai	0.1046	0.0012	0.3141
lig	0.1168	0.0355	0.1025
lim	0.0562	0.1162	0.0027
lux	0.0288	0.0101	0.1761
Total	1.0000	1.0000	1.0000

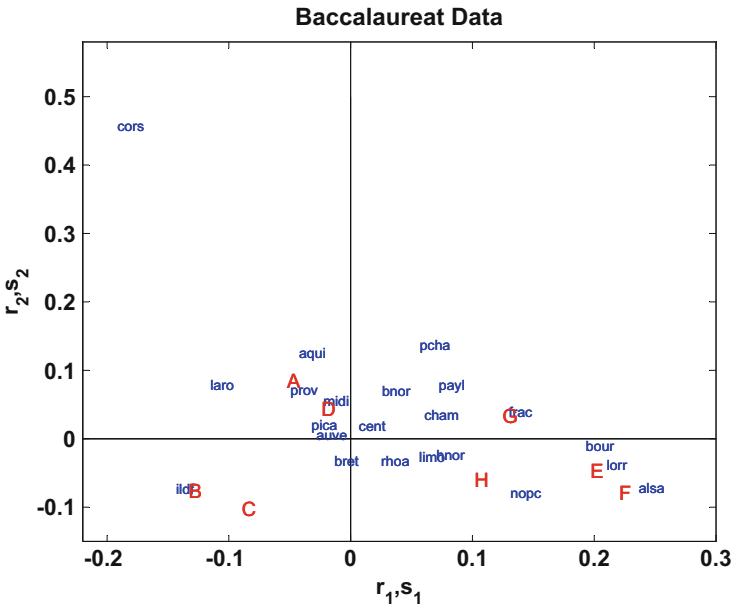

Fig. 15.2 Correspondence analysis including Corsica `MVAcorrbac`

Note also that the third axis shows an important role of the category "f_i" (other French newspapers) with the Wallon-Brabant "brw" and the Hainant "hai" showing the most important contributions. The coordinate of "f_i" on this axis is negative (not shown here) so are the coordinates of "brw" and "hai". Apparently, these two regions also seem to feature a greater proportion of readers of more local newspapers.

Example 15.4 Applying correspondence analysis to the French baccalauréat data (Table 22.8) leads to Fig. 15.2. Excluding Corsica we obtain Fig. 15.3. The different

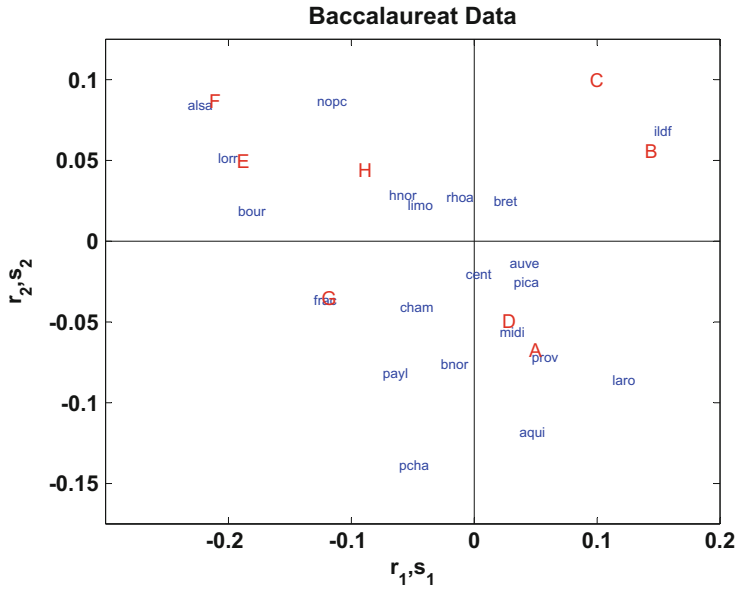

Fig. 15.3 Correspondence analysis excluding Corsica `MVAcorrbac`

Table 15.4 Eigenvalues and percentages of explained variance (including Corsica)

Eigenvalues λ	Percentage of variances	Cumulated percentage
2,436.2	0.5605	0.561
1,052.4	0.2421	0.803
341.8	0.0786	0.881
229.5	0.0528	0.934
152.2	0.0350	0.969
109.1	0.0251	0.994
25.0	0.0058	1.000
0.0	0.0000	1.000

modalities are labeled `A`, ..., `H` and the regions are labeled `ILDF`, ..., `CORS`. The results of the correspondence analysis are given in Table 15.4 and Fig. 15.2.

The first two factors explain 80 % of the total variance. It is clear from Fig. 15.2 that Corsica (in the upper left) is an outlier. The analysis is therefore redone without Corsica and the results are given in Table 15.5 and Fig. 15.3. Since Corsica has such a small weight in the analysis, the results have not changed much.

The projections on the first three axes, along with their absolute contribution to the variance of the axis, are summarised in Table 15.6 for the regions and in Table 15.7 for baccalauréats.

The interpretation of the results may be summarised as follows. Table 15.7 shows that the baccalauréats B on one side and F on the other side are most strongly

Table 15.5 Eigenvalues and percentages of explained variance (excluding Corsica)

Eigenvalues λ	Percentage of variances	Cumulated percentage
2,408.6	0.5874	0.587
909.5	0.2218	0.809
318.5	0.0766	0.887
195.9	0.0478	0.935
149.3	0.0304	0.971
96.1	0.0234	0.994
22.8	0.0056	1.000
0.0	0.0000	1.000

Table 15.6 Coefficients and absolute contributions for regions, Example 15.4

Region	r_1	r_2	r_3	$C_a(i, r_1)$	$C_a(i, r_2)$	$C_a(i, r_3)$
ILDF	0.1464	0.0677	0.0157	0.3839	0.2175	0.0333
CHAM	−0.0603	−0.0410	−0.0187	0.0064	0.0078	0.0047
PICA	0.0323	−0.0258	−0.0318	0.0021	0.0036	0.0155
HNOR	−0.0692	0.0287	0.1156	0.0096	0.0044	0.2035
CENT	−0.0068	−0.0205	−0.0145	0.0001	0.0030	0.0043
BNOR	−0.0271	−0.0762	0.0061	0.0014	0.0284	0.0005
BOUR	−0.1921	0.0188	0.0578	0.0920	0.0023	0.0630
NOPC	−0.1278	0.0863	−0.0570	0.0871	0.1052	0.1311
LORR	−0.2084	0.0511	0.0467	0.1606	0.0256	0.0608
ALSA	−0.2331	0.0838	0.0655	0.1283	0.0439	0.0767
FRAC	−0.1304	−0.0368	−0.0444	0.0265	0.0056	0.0232
PAYL	−0.0743	−0.0816	−0.0341	0.0232	0.0743	0.0370
BRET	0.0158	0.0249	−0.0469	0.0011	0.0070	0.0708
PCHA	−0.0610	−0.1391	−0.0178	0.0085	0.1171	0.0054
AQUI	0.0368	−0.1183	0.0455	0.0055	0.1519	0.0643
MIDI	0.0208	−0.0567	0.0138	0.0018	0.0359	0.0061
LIMO	−0.0540	0.0221	−0.0427	0.0033	0.0014	0.0154
RHOA	−0.0225	0.0273	−0.0385	0.0042	0.0161	0.0918
AUVE	0.0290	−0.0139	−0.0554	0.0017	0.0010	0.0469
LARO	0.0290	−0.0862	−0.0177	0.0383	0.0595	0.0072
PROV	0.0469	−0.0717	0.0279	0.0142	0.0884	0.0383

responsible for the variation on the first axis. The second axis mostly characterises an opposition between baccalauréats A and C. Regarding the regions, Ile de France plays an important role on each axis. On the first axis, it is opposed to Lorraine and Alsace, whereas on the second axis, it is opposed to Poitou-Charentes and Aquitaine. All of this is confirmed in Fig. 15.3.

On the right side are the more classical baccalauréats and on the left, more technical ones. The regions on the left side have thus larger weights in the technical

Table 15.7 Coefficients and absolute contributions for baccalauréats, Example 15.4

Baccal	s_1	s_2	s_3	$C_a(j, s_1)$	$C_a(j, s_2)$	$C_a(j, s_3)$
A	0.0447	−0.0679	0.0367	0.0376	0.2292	0.1916
B	0.1389	0.0557	0.0011	0.1724	0.0735	0.0001
C	0.0940	0.0995	0.0079	0.1198	0.3556	0.0064
D	0.0227	−0.0495	−0.0530	0.0098	0.1237	0.4040
E	−0.1932	0.0492	−0.1317	0.0825	0.0141	0.2900
F	−0.2156	0.0862	0.0188	0.3793	0.1608	0.0219
G	−0.1244	−0.0353	0.0279	0.1969	0.0421	0.0749
H	−0.0945	0.0438	−0.0888	0.0017	0.0010	0.0112

Table 15.8 Eigenvalues and explained proportion of variance, Example 15.5

λ_j	Percentage of variance	Cumulated percentage
4,399.0	0.4914	0.4914
2,213.6	0.2473	0.7387
1,382.4	0.1544	0.8932
870.7	0.0973	0.9904
51.0	0.0057	0.9961
34.8	0.0039	1.0000
0.0	0.0000	0.0000

baccalauréats. Note also that most of the southern regions of France are concentrated in the lower part of the graph near the baccalauréat A.

Finally, looking at the third axis, we see that it is dominated by the baccalauréat E (negative sign) and to a lesser degree by H (negative) (as opposed to A (positive sign)). The dominating regions are HNOR (positive sign), opposed to NOPC and AUVE (negative sign). For instance, HNOR is particularly poor in baccalauréat D.

Example 15.5 The US crime data set (Table 22.10) gives the number of crimes in the 50 states of the US classified in 1985 for each of the following seven categories: murder, rape, robbery, assault, burglary, larceny and auto-theft. The analysis of the contingency table, limited to the first two factors, provides the following results (see Table 15.8).

Looking at the absolute contributions (not reproduced here, see Exercise 15.6), it appears that the first axis is robbery (+) versus larceny (−) and auto-theft (−) axis and that the second factor contrasts assault (−) to auto-theft (+). The dominating states for the first axis are the North-Eastern States MA (+) and NY (+) contrasting the Western States WY (−) and ID (−). For the second axis, the differences are seen between the Northern States (MA (+) and RI (+)) and the Southern States AL (−), MS (−) and AR (−). These results can be clearly seen in Fig. 15.4 where all the states and crimes are reported. The figure also shows in which states the proportion of a particular crime category is higher or lower than the national average (the origin).

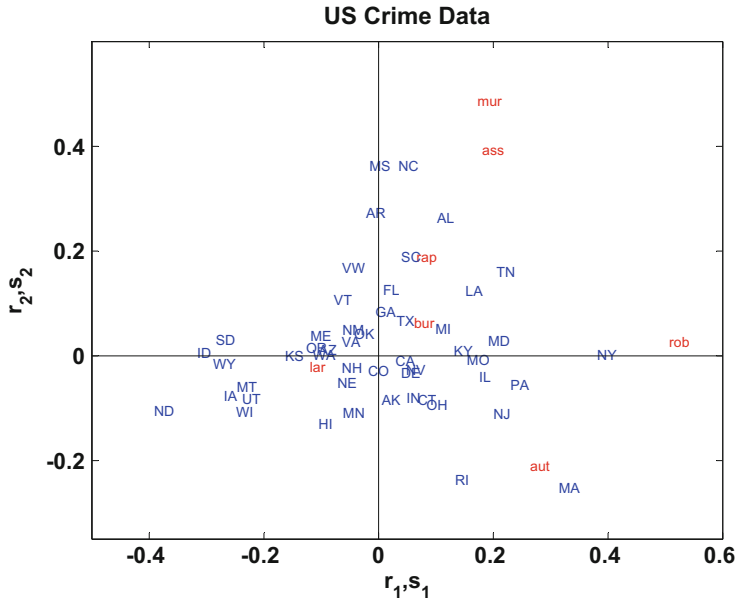

Fig. 15.4 Projection of rows (the 50 states) and columns (the seven crime categories) `MVAcorrcrime`

Biplots

The biplot is a low-dimensional display of a data matrix $\mathcal{X}$ where the rows and columns are represented by points. The interpretation of a biplot is specifically directed towards the scalar products of lower dimensional factorial variables and is designed to approximately recover the individual elements of the data matrix in these scalar products. Suppose that we have a (10×5) data matrix with elements x_{ij}. The idea of the biplot is to find 10 row points $q_i \in \mathbb{R}^k$ $(k < p,\ i = 1, \ldots, 10)$ and 5 column points $t_j \in \mathbb{R}^k$ $(j = 1, \ldots, 5)$ such that the 50 scalar products between the row and the column vectors closely approximate the 50 corresponding elements of the data matrix $\mathcal{X}$. Usually we choose $k = 2$. For example, the scalar product between q_7 and t_4 should approximate the data value x_{74} in the seventh row and the fourth column. In general, the biplot models the data x_{ij} as the sum of a scalar product in some low-dimensional subspace and a residual "error" term:

$$\begin{aligned} x_{ij} &= q_i^\top t_j + e_{ij} \\ &= \sum_k q_{ik} t_{jk} + e_{ij}. \end{aligned} \tag{15.25}$$

To understand the link between correspondence analysis and the biplot, we need to introduce a formula which expresses x_{ij} from the original data matrix (see (15.3)) in terms of row and column frequencies. One such formula, known as the "reconstitution formula", is (15.10):

$$x_{ij} = E_{ij}\left(1 + \frac{\sum_{k=1}^{R} \lambda_k^{\frac{1}{2}} \gamma_{ik} \delta_{jk}}{\sqrt{\frac{x_{i\bullet} x_{\bullet j}}{x_{\bullet\bullet}}}}\right) \tag{15.26}$$

Consider now the row profiles $x_{ij}/x_{i\bullet}$ (the conditional frequencies) and the average row profile $x_{i\bullet}/x_{\bullet\bullet}$. From (15.26) we obtain the difference between each row profile and this average:

$$\left(\frac{x_{ij}}{x_{i\bullet}} - \frac{x_{i\bullet}}{x_{\bullet\bullet}}\right) = \sum_{k=1}^{R} \lambda_k^{\frac{1}{2}} \gamma_{ik} \left(\sqrt{\frac{x_{\bullet j}}{x_{i\bullet} x_{\bullet\bullet}}}\right) \delta_{jk}. \tag{15.27}$$

By the same argument we can also obtain the difference between each column profile and the average column profile:

$$\left(\frac{x_{ij}}{x_{\bullet j}} - \frac{x_{\bullet j}}{x_{\bullet\bullet}}\right) = \sum_{k=1}^{R} \lambda_k^{\frac{1}{2}} \gamma_{ik} \left(\sqrt{\frac{x_{i\bullet}}{x_{\bullet j} x_{\bullet\bullet}}}\right) \delta_{jk}. \tag{15.28}$$

Now, if $\lambda_1 \gg \lambda_2 \gg \lambda_3 \ldots$, we can approximate these sums by a finite number of K terms (usually $K = 2$) using (15.16) to obtain

$$\left(\frac{x_{ij}}{x_{\bullet j}} - \frac{x_{i\bullet}}{x_{\bullet\bullet}}\right) = \sum_{k=1}^{K} \left(\frac{x_{\bullet i}}{\sqrt{\lambda_k} x_{\bullet\bullet}} r_{ki}\right) s_{kj} + e_{ij}, \tag{15.29}$$

$$\left(\frac{x_{ij}}{x_{i\bullet}} - \frac{x_{\bullet j}}{x_{\bullet\bullet}}\right) = \sum_{k=1}^{K} \left(\frac{x_{\bullet j}}{\sqrt{\lambda_k} x_{\bullet\bullet}} s_{kj}\right) r_{ki} + e'_{ij}, \tag{15.30}$$

where e_{ij} and e'_{ij} are error terms. Equation (15.30) shows that if we consider displaying the differences between the row profiles and the average profile, then the projection of the row profile r_k and a rescaled version of the projections of the column profile s_k constitute a biplot of these differences. Equation (15.29) implies the same for the differences between the column profiles and this average.

Summary

- ↪ Correspondence analysis is a factorial decomposition of contingency tables. The p-dimensional individuals and the n-dimensional variables can be graphically represented by projecting onto spaces of smaller dimension.
- ↪ The practical computation consists of first computing a spectral decomposition of $\mathcal{A}^{-1}\mathcal{X}\mathcal{B}^{-1}\mathcal{X}^\top$ and $\mathcal{B}^{-1}\mathcal{X}^\top\mathcal{A}^{-1}\mathcal{X}$ which have the same first p eigenvalues. The graphical representation is obtained by plotting $\sqrt{\lambda_1}r_1$ vs. $\sqrt{\lambda_2}r_2$ and $\sqrt{\lambda_1}s_1$ vs. $\sqrt{\lambda_2}s_2$. Both plots maybe displayed in the same graph taking into account the appropriate orientation of the eigenvectors r_i, s_j.
- ↪ Correspondence analysis provides a graphical display of the association measure $c_{ij} = (x_{ij} - E_{ij})^2/E_{ij}$.
- ↪ Biplot is a low-dimensional display of a data matrix where the rows and columns are represented by points

15.4 Exercises

Exercise 15.1 *Show that the matrices* $\mathcal{A}^{-1}\mathcal{X}\mathcal{B}^{-1}\mathcal{X}^\top$ *and* $\mathcal{B}^{-1}\mathcal{X}^\top\mathcal{A}^{-1}\mathcal{X}$ *have an eigenvalue equal to* 1 *and that the corresponding eigenvectors are proportional to* $(1,\ldots,1)^\top$.

Exercise 15.2 *Verify the relations in (15.8), (15.14) and (15.17).*

Exercise 15.3 *Do a correspondence analysis for the car marks data (Table 22.7)! Explain how this table can be considered as a contingency table.*

Exercise 15.4 *Compute the* χ^2*-statistic of independence for the French baccalauréat data.*

Exercise 15.5 *Prove that* $\mathcal{C} = \mathcal{A}^{-1/2}(\mathcal{X} - E)\mathcal{B}^{-1/2}\sqrt{x_{\bullet\bullet}}$ *and* $E = \frac{ab^\top}{x_{\bullet\bullet}}$ *and verify (15.20).*

Exercise 15.6 *Do the full correspondence analysis of the US crime data (Table 22.10), and determine the absolute contributions for the first three axes. How can you interpret the third axis? Try to identify the states with one of the four regions to which it belongs. Do you think the four regions have a different behaviour with respect to crime?*

Exercise 15.7 *Repeat Exercise 15.6 with the US health data (Table 22.16). Only analyse the columns indicating the number of deaths per state.*

Exercise 15.8 *Consider a* $(n \times n)$ *contingency table being a diagonal matrix* $\mathcal{X}$. *What do you expect the factors* r_k, s_k *to be like?*

Exercise 15.9 *Assume that after some reordering of the rows and the columns, the contingency table has the following structure:*

$$\mathcal{X} = \begin{array}{|c|c|c|} \hline & J_1 & J_2 \\ \hline I_1 & * & 0 \\ \hline I_2 & 0 & * \\ \hline \end{array}$$

That is, the rows I_i *only have weights in the columns* J_i, *for* $i = 1, 2$. *What do you expect the graph of the first two factors to look like?*

Exercise 15.10 *Redo Exercise 15.9 using the following contingency table:*

$$\mathcal{X} = \begin{array}{|c|c|c|c|} \hline & J_1 & J_2 & J_3 \\ \hline I_1 & * & 0 & 0 \\ \hline I_2 & 0 & * & 0 \\ \hline I_3 & 0 & 0 & * \\ \hline \end{array}$$

Exercise 15.11 *Consider the French food data (Table 22.6). Given that all of the variables are measured in the same units (Francs), explain how this table can be considered as a contingency table. Perform a correspondence analysis and compare the results to those obtained in the NPCA analysis in Chap. 11.*