
Theory of Correspondence Analysis



CA is based on fairly straightforward, classical results in matrix theory. The central result is the singular value decomposition (SVD), which is the basis of many multivariate methods such as principal component analysis, canonical correlation analysis, all forms of linear biplots, discriminant analysis and metric multidimensional scaling. In this appendix the theory of CA is summarized, as well as the theory of related methods discussed in the book. Matrix–vector notation is preferred because it is more compact, but also because it is closer to the implementation of the method in the R computing language.

Contents

| | |
|---|-----|
| <i>Correspondence matrix and preliminary notation</i> | 201 |
| <i>Basic computational algorithm</i> | 202 |
| <i>A note on the singular value decomposition (SVD)</i> | 203 |
| <i>The bilinear CA model</i> | 204 |
| <i>Transition equations between rows and columns</i> | 204 |
| <i>Supplementary points</i> | 204 |
| <i>Total inertia and χ^2-distances</i> | 204 |
| <i>Contributions of points to principal inertias</i> | 205 |
| <i>Contributions of principal axes to point inertias (squared correlations)</i> . | 205 |
| <i>Ward clustering of row or column profiles</i> | 206 |
| <i>Stacked tables</i> | 206 |
| <i>Multiple CA</i> | 206 |
| <i>Joint CA</i> | 207 |
| <i>Percentage of inertia explained in JCA</i> | 207 |
| <i>Contributions in JCA</i> | 208 |
| <i>Adjusted inertias in MCA</i> | 208 |
| <i>Subset CA and subset MCA</i> | 209 |
| <i>Analysis of square asymmetric tables</i> | 209 |
| <i>Canonical CA</i> | 210 |
| <i>Table for testing for significant clustering or significant dimensions</i> . . . | 211 |

Let \mathbf{N} denote the $I \times J$ data matrix, with positive row and column sums (almost always \mathbf{N} consists of nonnegative numbers, but there are some exceptions such as the one described at the end of Chapter 23). For notational simplicity the matrix is first converted to the *correspondence matrix* \mathbf{P} by dividing \mathbf{N} by its grand total $n = \sum_i \sum_j n_{ij} = \mathbf{1}^T \mathbf{N} \mathbf{1}$ (the notation $\mathbf{1}$ is used for a vector of ones of length that is appropriate to its use; hence the first $\mathbf{1}$ is $I \times 1$ and the second is $J \times 1$ to match the row and column lengths of \mathbf{N}).

Correspondence matrix and preliminary notation

Correspondence matrix:

$$\mathbf{P} = \frac{1}{n}\mathbf{N} \quad (\text{A.1})$$

The following notation is used:

Row and column masses:

$$\begin{aligned} r_i &= \sum_{j=1}^J p_{ij} & c_j &= \sum_{i=1}^I p_{ij} \\ \text{i.e., } \mathbf{r} &= \mathbf{P}\mathbf{1} & \mathbf{c} &= \mathbf{P}^T\mathbf{1} \end{aligned} \quad (\text{A.2})$$

Diagonal matrices of row and column masses:

$$\mathbf{D}_r = \text{diag}(\mathbf{r}) \quad \text{and} \quad \mathbf{D}_c = \text{diag}(\mathbf{c}) \quad (\text{A.3})$$

Note that all subsequent definitions and results are given in terms of these relative quantities $\mathbf{P} = \{p_{ij}\}$, $\mathbf{r} = \{r_i\}$ and $\mathbf{c} = \{c_j\}$, whose elements add up to 1 in each case. Multiply these by n to recover the elements of the original matrix \mathbf{N} : $np_{ij} = n_{ij}$, $nr_i = i$ -th row sum of \mathbf{N} , $nc_j = j$ -th column sum of \mathbf{N} .

*Basic
computational
algorithm*

The computational algorithm to obtain coordinates of the row and column profiles with respect to principal axes, using the singular value decomposition (SVD), is as follows:

CA Step 1 — Calculate the matrix \mathbf{S} of standardized residuals:

$$\mathbf{S} = \mathbf{D}_r^{-\frac{1}{2}}(\mathbf{P} - \mathbf{r}\mathbf{c}^T)\mathbf{D}_c^{-\frac{1}{2}} \quad (\text{A.4})$$

CA Step 2 — Calculate the SVD of \mathbf{S} :

$$\mathbf{S} = \mathbf{U}\mathbf{D}_\alpha\mathbf{V}^T \quad \text{where} \quad \mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I} \quad (\text{A.5})$$

where \mathbf{D}_α is the diagonal matrix of (positive) singular values in descending order: $\alpha_1 \geq \alpha_2 \geq \dots$

CA Step 3 — Standard coordinates Φ of rows:

$$\Phi = \mathbf{D}_r^{-\frac{1}{2}}\mathbf{U} \quad (\text{A.6})$$

CA Step 4 — Standard coordinates Γ of columns:

$$\Gamma = \mathbf{D}_c^{-\frac{1}{2}}\mathbf{V} \quad (\text{A.7})$$

CA Step 5 — Principal coordinates \mathbf{F} of rows:

$$\mathbf{F} = \mathbf{D}_r^{-\frac{1}{2}}\mathbf{U}\mathbf{D}_\alpha = \Phi\mathbf{D}_\alpha \quad (\text{A.8})$$

CA Step 6 — Principal coordinates \mathbf{G} of columns:

$$\mathbf{G} = \mathbf{D}_c^{-\frac{1}{2}}\mathbf{V}\mathbf{D}_\alpha = \Gamma\mathbf{D}_\alpha \quad (\text{A.9})$$

CA Step 7 — Principal inertias λ_k :

$$\lambda_k = \alpha_k^2, \quad k = 1, 2, \dots, K \quad \text{where } K = \min\{I - 1, J - 1\} \quad (\text{A.10})$$

The rows of the coordinate matrices in (A.6)–(A.9) refer to the rows or columns, as the case may be, of the original table, while the columns of these matrices refer to the principal axes, or dimensions, of which there are $\min\{I-1, J-1\}$, i.e., one less than the number of rows or columns, whichever is smaller. Notice how the principal and standard coordinates are scaled:

$$\mathbf{F}\mathbf{D}_r\mathbf{F}^\top = \mathbf{G}\mathbf{D}_c\mathbf{G}^\top = \mathbf{D}_\lambda \quad (\text{A.11})$$

$$\mathbf{\Phi}\mathbf{D}_r\mathbf{\Phi}^\top = \mathbf{\Gamma}\mathbf{D}_c\mathbf{\Gamma}^\top = \mathbf{I} \quad (\text{A.12})$$

That is, the weighted sum-of-squares of the principal coordinates on the k -th dimension (i.e., their inertia in the direction of this dimension) is equal to the principal inertia (or eigenvalue) $\lambda_k = \alpha_k^2$, the square of the k -th singular value, whereas the standard coordinates have weighted sum-of-squares equal to 1. All coordinate matrices have orthogonal columns, where the masses are always used in the calculation of the (weighted) scalar products.

The SVD is the fundamental mathematical result for CA, as it is for other dimension reduction techniques such as principal component analysis, canonical correlation analysis and linear discriminant analysis. This matrix decomposition expresses any rectangular matrix as a product of three matrices of simple structure, as in (A.5) above: $\mathbf{S} = \mathbf{U}\mathbf{D}_\alpha\mathbf{V}^\top$. The columns of the matrices \mathbf{U} and \mathbf{V} are the *left* and *right singular vectors* respectively, and the positive values α_k down the diagonal of \mathbf{D}_α , in descending order, are the *singular values*. The SVD is related to the more well-known eigenvalue–eigenvector decomposition (or *eigendecomposition*) of a square symmetric matrix as follows: $\mathbf{S}\mathbf{S}^\top$ and $\mathbf{S}^\top\mathbf{S}$ are square symmetric matrices which have eigendecompositions $\mathbf{S}\mathbf{S}^\top = \mathbf{U}\mathbf{D}_\alpha^2\mathbf{U}^\top$ and $\mathbf{S}^\top\mathbf{S} = \mathbf{V}\mathbf{D}_\alpha^2\mathbf{V}^\top$, so the singular vectors are also eigenvectors of these respective matrices, and the singular values are the square roots of their eigenvalues. The practical utility of the SVD is that if one constructs another $I \times J$ matrix $\mathbf{S}_{(m)}$ from the the first m columns of $\mathbf{U}_{(m)}$ and $\mathbf{V}_{(m)}$ and the first m singular values in $\mathbf{D}_{\alpha(m)}$: $\mathbf{S}_{(m)} = \mathbf{U}_{(m)}\mathbf{D}_{\alpha(m)}\mathbf{V}_{(m)}^\top$, then $\mathbf{S}_{(m)}$ is the least-squares rank m approximation of \mathbf{S} (this result is known as the *Eckart-Young theorem*). Since the objective of finding low-dimensional best-fitting subspaces coincides with the objective of finding low-rank matrix approximations by least-squares, the SVD solves our problem completely and in a very compact way. The only adaptation needed is to incorporate the weighting of the rows and columns by the masses into the SVD so that the approximations are by weighted least squares. If a generalized form of the SVD were defined, where the singular vectors are normalized with weighting by the masses, then the CA solution can be obtained in one step. For example, the generalized SVD of the *contingency ratios* $p_{ij}/(r_i c_j)$, elements of the matrix $\mathbf{D}_r^{-1}\mathbf{P}\mathbf{D}_c^{-1}$, centred at the constant value 1, leads to the standard row and column coordinates directly:

$$\mathbf{D}_r^{-1}\mathbf{P}\mathbf{D}_c^{-1} - \mathbf{1}\mathbf{1}^\top = \mathbf{\Phi}\mathbf{D}_\alpha\mathbf{\Gamma}^\top \quad \text{where} \quad \mathbf{\Phi}^\top\mathbf{D}_r\mathbf{\Phi} = \mathbf{\Gamma}^\top\mathbf{D}_c\mathbf{\Gamma} = \mathbf{I} \quad (\text{A.13})$$

A note on the singular value decomposition (SVD)

The bilinear CA model From steps 1 to 4 of the basic algorithm, the data in \mathbf{P} can be written as follows (see also (13.4) on page 101 and (14.9) on page 109):

$$p_{ij} = r_i c_j \left(1 + \sum_{k=1}^K \sqrt{\lambda_k} \phi_{ik} \gamma_{jk} \right) \quad (\text{A.14})$$

(also called the *reconstitution formula*). In matrix notation:

$$\mathbf{P} = \mathbf{D}_r (\mathbf{1}\mathbf{1}^\top + \mathbf{\Phi} \mathbf{D}_\lambda^{1/2} \mathbf{\Gamma}^\top) \mathbf{D}_c \quad (\text{A.15})$$

Because of the simple relations (A.8) and (A.9) between the principal and standard coordinates, this bilinear model can be written in several alternative ways — see also (14.10) and (14.11) on pages 109–110.

Transition equations between rows and columns The left and right singular vectors are related linearly, for example by multiplying the SVD on the right by \mathbf{V} : $\mathbf{S}\mathbf{V} = \mathbf{U}\mathbf{D}_\alpha$. Expressing such relations in terms of the principal and standard coordinates gives the following variations of the same theme, called transition equations (see (14.1) & (14.2) and (14.5) & (14.6) for the equivalent scalar versions):

Principal as a function of standard (barycentric relationships):

$$\mathbf{F} = \mathbf{D}_r^{-1} \mathbf{P} \mathbf{\Gamma} \quad \mathbf{G} = \mathbf{D}_c^{-1} \mathbf{P}^\top \mathbf{\Phi} \quad (\text{A.16})$$

Principal as a function of principal:

$$\mathbf{F} = \mathbf{D}_r^{-1} \mathbf{P} \mathbf{G} \mathbf{D}_\lambda^{-1/2} \quad \mathbf{G} = \mathbf{D}_c^{-1} \mathbf{P}^\top \mathbf{F} \mathbf{D}_\lambda^{-1/2} \quad (\text{A.17})$$

The equations (A.16) are those that were mentioned as early as Chapter 3, which express the profile points as weighted averages of the vertex points, where the weights are the profile elements. These are the equations that govern the *asymmetric maps*. The equations (A.17) show that the two sets of principal coordinates, which govern the *symmetric map*, are also related by a barycentric (weighted average) relationship, but with scale factors (the inverse square roots of the principal inertias) that are different on each axis.

Supplementary points The transition equations are used to situate supplementary points on the map. For example, given a supplementary column point with values in \mathbf{h} ($I \times 1$), divide by its total $\mathbf{1}^\top \mathbf{h}$ to obtain the column profile $\tilde{\mathbf{h}} = (\mathbf{1}/\mathbf{1}^\top \mathbf{h}) \mathbf{h}$ and then use the profile transposed as a row vector in the second equation of (A.16), for example, to calculate the coordinates \mathbf{g} of the supplementary column:

$$\mathbf{g} = \tilde{\mathbf{h}}^\top \mathbf{\Phi} \quad (\text{A.18})$$

Total inertia and χ^2 -distances The total inertia of the data matrix is the sum of squares of the matrix \mathbf{S} in (A.4):

$$\text{inertia} = \text{trace}(\mathbf{S}\mathbf{S}^\top) = \sum_{i=1}^I \sum_{j=1}^J \frac{(p_{ij} - r_i c_j)^2}{r_i c_j} \quad (\text{A.19})$$

The inertia is also the sum of squares of the singular values, i.e., the sum of the eigenvalues:

$$\text{inertia} = \sum_{k=1}^K \alpha_k^2 = \sum_{k=1}^K \lambda_k \quad (\text{A.20})$$

The χ^2 -distances between row profiles and between column profiles are:

$$\chi^2\text{-distances between rows } i \text{ and } i' : \sum_{j=1}^J \left(\frac{p_{ij}}{r_i} - \frac{p_{i'j}}{r_{i'}} \right)^2 / c_j \quad (\text{A.21})$$

$$\chi^2\text{-distances between columns } j \text{ and } j' : \sum_{i=1}^I \left(\frac{p_{ij}}{c_j} - \frac{p_{ij'}}{c_{j'}} \right)^2 / r_i \quad (\text{A.22})$$

To write the full set of χ^2 -distances in the form of a square symmetric matrix requires a bit more work. First, calculate the matrix \mathbf{A} of “ χ^2 scalar products” between row profiles, for example, as:

$$\chi^2 \text{ scalar products between rows : } \mathbf{A} = \mathbf{D}_r^{-1} \mathbf{P} \mathbf{D}_c^{-1} \mathbf{P}^T \mathbf{D}_r^{-1} \quad (\text{A.23})$$

Then define the vector \mathbf{a} as the elements on the diagonal of this matrix (i.e., the scalar products of the row profiles with themselves):

$$\mathbf{a} = \text{diag}(\mathbf{A}) \quad (\text{A.24})$$

Then the $I \times I$ matrix of squared χ^2 -distances is:

$$\text{squared } \chi^2\text{-distance matrix between rows : } \mathbf{a} \mathbf{1}^T + \mathbf{1} \mathbf{a}^T - 2\mathbf{A} \quad (\text{A.25})$$

To calculate the $J \times J$ matrix of squared χ^2 -distances between column profiles, interchange rows with columns in (A.23), defining \mathbf{A} as $\mathbf{D}_c^{-1} \mathbf{P}^T \mathbf{D}_r^{-1} \mathbf{P} \mathbf{D}_c^{-1}$ and then following with (A.24) and (A.25).

The contributions of the row and columns points to the inertia on the k -th dimension are the inertia components:

$$\text{for row } i: \frac{r_i f_{ik}^2}{\lambda_k} = r_i \phi_{ik}^2 \quad \text{for column } j: \frac{c_j g_{jk}^2}{\lambda_k} = c_j \gamma_{jk}^2 \quad (\text{A.26})$$

recalling the relationship between principal and standard coordinates given in (A.8) and (A.9): $f_{ik} = \sqrt{\lambda_k} \phi_{ik}$, $g_{jk} = \sqrt{\lambda_k} \gamma_{jk}$ (notice that the square roots of the values in (A.26) are exactly the coordinates proposed for the standard CA biplot of Chapter 13, which shows that the squared lengths of these coordinates are the contributions to the principal axes).

The contributions of the dimensions to the inertia of the i -th row and j -th column points (i.e., the squared cosines or squared correlations) are:

$$\text{for row } i: \frac{f_{ik}^2}{\sum_k f_{ik}^2} \quad \text{for column } j: \frac{g_{jk}^2}{\sum_k g_{jk}^2} \quad (\text{A.27})$$

As shown in Chapter 11, the denominators in (A.27) are the squared χ^2 -distances between the corresponding profile point and the average profile.

Contributions of points to principal inertias

Contributions of principal axes to point inertias (squared correlations)

Ward clustering of row or column profiles The clustering of Chapter 15 is described here in terms of the rows; exactly the same applies to the clustering of the columns. The rows are clustered at each step of the algorithm to minimise the decrease in the χ^2 statistic (equivalently, the decrease in the inertia since inertia = χ^2/n , where n is the total of the table). This clustering criterion is equivalent to Ward clustering, where each cluster is weighted by the total mass of its members. The measure of difference between rows can be shown to be the weighted form of the squared chi-squared distance between profiles. Suppose \mathbf{a}_i and r_i , $i = 1, \dots, I$, denote the I row profiles of the data matrix, and their masses, respectively. Then identifying the pair that gives the least decrease in inertia is equivalent to looking for the pair of rows (i, i') which minimize the following measure:

$$\frac{r_i r_{i'}}{r_i + r_{i'}} \|\mathbf{a}_i - \mathbf{a}_{i'}\|_c^2 \quad (\text{A.28})$$

The two rows are then merged by summing their frequencies, and the profile and mass are recalculated. The same measure of difference as (A.28) is calculated at each stage of the clustering for the row profiles at that stage (see (15.2) on page 120 for the equivalent formula based on profiles of clusters), and the two profiles with the least difference are merged. So (A.28) is the level of clustering in terms of the inertia decrease, or if multiplied by n the decrease in χ^2 . In the case of a contingency table the level of clustering can be tested for significance using the tables at the end of this Appendix.

Stacked tables Suppose tables \mathbf{N}_{qs} , $q = 1, \dots, Q$, $s = 1, \dots, S$ are concatenated row- and/or columnwise to make a block matrix \mathbf{N} . If the marginal frequencies are the same in each row and in each column (as is the case when the same individuals are cross-tabulated separately in several tables), then the inertia of \mathbf{N} is the average of the separate inertias of the tables \mathbf{N}_{qs} :

$$\text{inertia}(\mathbf{N}) = \frac{1}{QS} \sum_{q=1}^Q \sum_{s=1}^S \text{inertia}(\mathbf{N}_{qs}) \quad (\text{A.29})$$

Multiple CA Suppose the original matrix of categorical data is $N \times Q$, i.e., N cases and Q variables. Classical multiple CA (MCA) has two forms. The first form converts the cases-by-variables data to an indicator matrix \mathbf{Z} where the categorical data have been recoded as dummy variables. If the q -th variable has J_q categories, this indicator matrix will have $J = \sum_q J_q$ columns (see Chapter 18, Exhibit 18.1 for an example). Then the indicator version of MCA is the application of the basic CA algorithm to the matrix \mathbf{Z} , resulting in coordinates for the N cases and the J categories. The second form of MCA calculates the Burt matrix $\mathbf{B} = \mathbf{Z}^T \mathbf{Z}$ of all two-way cross-tabulations of the Q variables (see Chapter 18, Exhibit 18.4 for an example). Then the Burt version of MCA is the application of the basic CA algorithm to the matrix \mathbf{B} , resulting in coordinates for the J categories (\mathbf{B} is a symmetric matrix). The standard coordinates of the categories are identical in the two versions of MCA, and the

principal inertias in the Burt version are the squares of those in the indicator version.

Joint CA (JCA) is the fitting of the off-diagonal cross-tabulations of the Burt matrix, ignoring the cross-tabulations on the block diagonal. The algorithm we use is an alternating least-squares procedure which successively applies CA to the Burt matrix which has been modified by replacing the values on the block diagonal with estimated values from the CA of the previous iteration. The algorithm itself is explained in more detail in the Computational Appendix. On convergence of the JCA algorithm, the CA is performed on the last modified Burt matrix, $\tilde{\mathbf{B}}$, which has its diagonal blocks perfectly fitted by construction. In other words, supposing that the solution requested is two-dimensional, then the modified diagonal blocks satisfy (A.14) exactly using just two terms in the bilinear CA model (or reconstitution formula).

Joint CA

Hence the total inertia of $\tilde{\mathbf{B}}$ includes a part Δ for these diagonal blocks, and so do the first two principal inertias, $\tilde{\lambda}_1$ and $\tilde{\lambda}_2$, which perfectly explain the part Δ . To obtain the percentage of inertia explained by the two-dimensional solution, the amount Δ has to be discounted both from the total and from the sum of the two principal inertias. The value of Δ can be obtained via the difference between the inertia of the original Burt matrix \mathbf{B} (whose diagonal inertias are known) and the modified one $\tilde{\mathbf{B}}$, as follows (here we use the result (A.29) which applies to the subtables of \mathbf{B} , denoted by \mathbf{B}_{qs} , and those of $\tilde{\mathbf{B}}$, whose off-diagonal tables are the same):

Percentage of inertia explained in JCA

$$\begin{aligned} \text{inertia}(\mathbf{B}) &= \frac{1}{Q^2} \left(\sum \sum_{q \neq s} \text{inertia}(\mathbf{B}_{qs}) + \sum_q \text{inertia}(\mathbf{B}_{qq}) \right) \\ &= \frac{1}{Q^2} \left(\sum \sum_{q \neq s} \text{inertia}(\mathbf{B}_{qs}) + (J - Q) \right) \\ \text{inertia}(\tilde{\mathbf{B}}) &= \frac{1}{Q^2} \left(\sum \sum_{q \neq s} \text{inertia}(\mathbf{B}_{qs}) \right) + \Delta \end{aligned}$$

Subtracting the above leads to:

$$\text{inertia}(\mathbf{B}) - \text{inertia}(\tilde{\mathbf{B}}) = \frac{J - Q}{Q^2} - \Delta \quad (\text{A.30})$$

which gives the value of Δ :

$$\Delta = \frac{J - Q}{Q^2} - \left(\text{inertia}(\mathbf{B}) - \text{inertia}(\tilde{\mathbf{B}}) \right) \quad (\text{A.31})$$

Discounting this amount from the total and the sum of the principal inertias (assuming a two-dimensional solution) gives the percentage of inertia explained by the JCA solution:

$$100 \times \frac{\tilde{\lambda}_1 + \tilde{\lambda}_2 - \Delta}{\text{inertia}(\tilde{\mathbf{B}}) - \Delta} \quad (\text{A.32})$$

Contributions in JCA The previous section showed how to discount the extra inertia as a result of the modified diagonal blocks of the Burt matrix in JCA. There is an identical situation at the level of each point. Each category point j has an additional amount of inertia, δ_j , due to the modified diagonal blocks. In the case of the original Burt matrix \mathbf{B} we know exactly what this extra amount is due to the diagonal matrices in the diagonal blocks: for the j -th point it is $(1 - Qc_j)/Q^2$, where c_j is the j -th mass (summing these values for $j = 1, \dots, J$, we obtain $(J - Q)/Q^2$ which was the total additional amount due to the diagonal blocks of \mathbf{B}). Therefore, just as above, we can derive how to obtain contributions of the two-dimensional solution to the point inertias as follows:

$$\begin{aligned} \text{inertia}(j\text{-th category of } \mathbf{B}) &= \text{off-diagonal components} + \frac{1 - Qc_j}{Q^2} \\ \text{inertia}(j\text{-th category of } \tilde{\mathbf{B}}) &= \text{off-diagonal components} + \delta_j \end{aligned}$$

Subtracting the above (the “off-diagonal components” are the same) leads to:

$$\text{inertia}(j\text{th category of } \mathbf{B}) - \text{inertia}(j\text{-th category of } \tilde{\mathbf{B}}) = \frac{1 - Qc_j}{Q^2} - \delta_j$$

which gives the value of δ_j :

$$\delta_j = \frac{1 - Qc_j}{Q^2} - \left(\text{inertia}(j\text{-th category of } \mathbf{B}) - \text{inertia}(j\text{-th category of } \tilde{\mathbf{B}}) \right) \quad (\text{A.33})$$

Discounting this amount from the j -th category’s inertia and similarly from the sum of the components of inertia in two dimensions gives the relative contributions (qualities) with respect to the two-dimensional JCA solution:

$$\frac{c_j \tilde{g}_{j1}^2 + c_j \tilde{g}_{j2}^2 - \delta_j}{(\sum_k c_j \tilde{g}_{jk}^2) - \delta_j} \quad (\text{A.34})$$

where \tilde{g}_{jk} is the principal coordinate of category j on axis k in the CA of $\tilde{\mathbf{B}}$ (JCA solution), and the summation in the denominator is for all the dimensions. Notice that $\sum_j \delta_j = \Delta$ (i.e., summing (A.33) gives (A.31)).

Adjusted inertias in MCA The MCA solution can be adjusted to optimize the fit to the off-diagonal tables (this could be called a JCA *conditional on* the MCA solution). The optimal adjustments can be determined by weighted least-squares, as described in Chapter 19, but the problem is that the solution is not nested. So we prefer slightly sub-optimal adjustments which retain the nesting property and are very easy to compute from the MCA solution of the Burt matrix. The adjustments are made as follows (see Chapter 19, pages 148–149, for an illustration):

Adjusted total inertia of Burt matrix:

$$\text{adjusted total inertia} = \frac{Q}{Q-1} \times \left(\text{inertia of } \mathbf{B} - \frac{J-Q}{Q^2} \right) \quad (\text{A.35})$$

Adjusted principal inertias (eigenvalues) of Burt matrix:

$$\lambda_k^{\text{adj}} = \left(\frac{Q}{Q-1} \right)^2 \times \left(\sqrt{\lambda_k} - \frac{1}{Q} \right)^2, \quad k = 1, 2, \dots \quad (\text{A.36})$$

Here λ_k refers to the k -th principal inertia of the Burt matrix; hence $\sqrt{\lambda_k}$ is the k -th principal inertia of the indicator matrix. The adjustments are made only to those dimensions for which $\sqrt{\lambda_k} > \frac{1}{Q}$ and no further dimensions are used — hence percentages of inertia do not add up to 100%. It can be proved that these percentages are lower bound estimates of those that are obtained in a JCA, and in practice they are close to the JCA percentages.

Subset CA is simply the application of the same CA algorithm to a selected part of the standardized residual matrix \mathbf{S} in (A.4) (*not* to the subset of the original matrix). The masses of the full matrix are thus retained and all subsequent calculations are the same, except they are applied to the subset. Suppose that the columns are subsetted, but not the rows. Then the rows still maintain the centring property of CA; i.e., their weighted averages are at the origin of the map, whereas the columns are no longer centred. Subset MCA is performed by applying subset CA on a submatrix of the indicator matrix or the Burt matrix. In the case of the Burt matrix, a selection of categories implies that this subset has to be specified for both the rows and columns.

Subset CA and
subset MCA

If the data matrix \mathbf{N} is square asymmetric, where both rows and columns refer to the same objects, then \mathbf{N} can be written as the sum of symmetric and skew-symmetric parts:

$$\begin{aligned} \mathbf{N} &= \frac{1}{2}(\mathbf{N} + \mathbf{N}^T) + \frac{1}{2}(\mathbf{N} - \mathbf{N}^T) \\ &= \text{symmetric} + \text{skew-symmetric} \end{aligned} \quad (\text{A.37})$$

Analysis of square
asymmetric tables

CA is applied to each part separately, but with a slight variation for the skew-symmetric part. The analysis of the symmetric part $\frac{1}{2}(\mathbf{N} + \mathbf{N}^T)$ is the usual CA — this provides one set of coordinates, and the masses are the averages of the row and column masses corresponding to the same object: $w_i = \frac{1}{2}(r_i + c_i)$. The analysis of the skew-symmetric part $\frac{1}{2}(\mathbf{N} - \mathbf{N}^T)$ is the application of the CA algorithm *without centring* and using the same masses as in the symmetric analysis; i.e., the “standardized residuals” matrix of (A.4) is rather the “standardized differences” matrix

$$\mathbf{S} = \mathbf{D}_w^{-\frac{1}{2}} \left[\frac{1}{2}(\mathbf{P} - \mathbf{P}^T) \right] \mathbf{D}_w^{-\frac{1}{2}} \quad (\text{A.38})$$

where \mathbf{P} is the correspondence matrix and \mathbf{D}_w is the diagonal matrix of the masses w_i . As described in Chapter 22, both these analyses are subsumed in the ordinary CA of the block matrix

$$\begin{bmatrix} \mathbf{N} & \mathbf{N}^T \\ \mathbf{N}^T & \mathbf{N} \end{bmatrix} \quad (\text{A.39})$$

If \mathbf{N} is an $I \times I$ matrix, then the $2I - 1$ dimensions which emanate from this CA can be easily allocated to the symmetric and skew-symmetric solutions since the symmetric dimensions have unique principal inertias while the skew-symmetric dimensions occur in pairs with equal principal inertias. Similarly the coordinate vectors for each dimension have two parts: for dimensions corresponding to the symmetric analysis these are simple repeats on each other, while for dimensions corresponding to the skew-symmetric analysis these are repeats with a change of sign (see Chapter 22 for an example).

Canonical CA In canonical CA (CCA) an additional matrix \mathbf{X} of explanatory (independent) variables is available, and the requirement is that the dimensions be linearly related to \mathbf{X} . The total inertia is split into two parts: a part that is linearly related to the independent variables, called the inertia in the *constrained space*, and a part that is not, the inertia in the *unconstrained space*. CCA is necessarily an “asymmetric” method since \mathbf{X} is an additional set of either rows or columns. The usual data structure is that the rows are sampling units and \mathbf{X} is an additional set of M columns, i.e., $I \times M$. There is a regression step in CCA which calculates the $I \times J$ constrained matrix, whose columns are linearly related to \mathbf{X} . The difference between \mathbf{P} and the constrained matrix is the unconstrained matrix, whose columns are not linearly related to \mathbf{X} . CCA thus consists of applying CA to the constrained matrix and (optionally) to the unconstrained matrix. In each application the original row and column masses are maintained for all computations, and the various results such as coordinates, principal inertias, contributions, reconstruction formula, etc., are the same as in a regular application of CA. We assume that the columns of \mathbf{X} are standardized, using the row masses as weights in the calculation of means and variances. If there are some categorical independent variables, these are coded as dummy variables, dropping one category of each variable as in a conventional regression analysis. The retained dummy variables are then also standardized in the same way as the columns of \mathbf{X} .

The steps in CCA are as follows:

CCA Step 1 — Calculate the standardized residuals matrix \mathbf{S} as in CA:

$$\mathbf{S} = \mathbf{D}_r^{-\frac{1}{2}}(\mathbf{P} - \mathbf{r}\mathbf{c}^\top)\mathbf{D}_c^{-\frac{1}{2}} \quad (\text{A.40})$$

CCA Step 2 — Calculate the $I \times I$ projection matrix, of rank M , which projects onto the constrained space:

$$\mathbf{Q} = \mathbf{D}_r^{\frac{1}{2}}\mathbf{X}(\mathbf{X}^\top\mathbf{D}_r\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{D}_r^{\frac{1}{2}} \quad (\text{A.41})$$

CCA Step 3 — Project the standardized residuals to obtain the constrained matrix:

$$\mathbf{S}^* = \mathbf{Q}\mathbf{S} \quad (\text{A.42})$$

CCA Step 4 — Apply CA Steps 1–6 (page 202) to \mathbf{S}^* :

CCA Step 5 — Principal inertias λ_k^* in constrained space:

$$\lambda_k^* = \alpha_k^2, \quad k = 1, 2, \dots, K \text{ where } K = \min\{I - 1, J - 1, M\} \quad (\text{A.43})$$

CCA Step 5 (optional) — Project the standardized residuals onto the unconstrained space:

$$\mathbf{S}^\perp = (\mathbf{I} - \mathbf{Q})\mathbf{S} = \mathbf{S} - \mathbf{S}^* \tag{A.44}$$

CCA Step 6 (optional) — Apply CA Steps 1-6 to \mathbf{S}^\perp .

As described in Chapter 24, the principal inertias in (A.43) can be expressed as percentages of the total inertia, or as percentages of the constrained inertia, which is the sum of squares of the elements in \mathbf{S}^* , equal to $\sum_k \lambda_k^*$.

In the case of a contingency table based on a random sample, the first principal inertia can be tested for statistical significance. This is the same test as was used in the case of the Ward clustering of Chapter 15. In that case a critical level for clustering, on the χ^2 scale, can be determined from the table in Exhibit A.1 below, according to the size of the table (see page 119 for the food store example, a 5×4 table for which the critical point in Exhibit A.1 is 15.24). These critical points are the same for testing the first principal inertia for significance. For example, in the same example of the food stores, given in Exhibit 15.3, the first principal inertia was 0.02635, which if expressed as a χ^2 component is $0.02635 \times 700 = 18.45$. Since 18.45 is greater than the critical point 15.24, the first principal inertia is statistically significant (at the 5% level).

Table for testing for significant clustering or significant dimensions

| <i>I</i> | <i>J</i> | | | | | | | | | |
|----------|----------|-------|-------|-------|-------|-------|-------|-------|-------|--|
| | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | |
| 3 | 8.59 | | | | | | | | | |
| 4 | 10.74 | 13.11 | | | | | | | | |
| 5 | 12.68 | 15.24 | 17.52 | | | | | | | |
| 6 | 14.49 | 17.21 | 19.63 | 21.85 | | | | | | |
| 7 | 16.21 | 19.09 | 21.62 | 23.95 | 26.14 | | | | | |
| 8 | 17.88 | 20.88 | 23.53 | 25.96 | 28.23 | 30.40 | | | | |
| 9 | 19.49 | 22.62 | 25.37 | 27.88 | 30.24 | 32.48 | 34.63 | | | |
| 10 | 21.06 | 24.31 | 27.15 | 29.75 | 32.18 | 34.50 | 36.70 | 38.84 | | |
| 11 | 22.61 | 25.96 | 28.90 | 31.57 | 34.08 | 36.45 | 38.72 | 40.91 | 43.04 | |
| 12 | 24.12 | 27.58 | 30.60 | 33.35 | 35.93 | 38.36 | 40.69 | 42.93 | 45.10 | |
| 13 | 25.61 | 29.17 | 32.27 | 35.09 | 37.73 | 40.22 | 42.60 | 44.90 | 47.12 | |

Exhibit A.1:
Critical values for multiple comparisons test on a $I \times J$ (or $J \times I$) contingency table. The same critical points apply to testing the significance of a principal inertia. Significance level is 5%.

Source: Pearson, E.S. & Hartley, H.O. (1972). *Biometrika Tables for Statisticians, Volume 2: Table 51*. Cambridge University Press, UK.