Topics in correspondence analysis
van de Velden, M.

Citation for published version (APA):
Chapter 2

Correspondence analysis

2.1 Introduction

The aim of correspondence analysis is to depict multivariate numerical data in a graphical manner. In particular, rows and columns of a contingency matrix are depicted as points in a plot of few (usually two) dimensions. Insight in the complex multivariate data is then obtained simply by looking at the plot. The main objective of correspondence analysis is to explore rather than to analyze the data. No assumptions are necessary to carry out correspondence analysis. As a result of the exploratory nature of the method, the role of statistical inference is limited.

As was mentioned in the previous chapter, correspondence analysis is closely related to several other well-known multivariate methods. In this chapter the relationship between component analysis and correspondence analysis is employed to describe correspondence analysis.

In section 2.3 we will derive the essential equations encountered in correspondence analysis using the relationship between correspondence analysis and component analysis. For this purpose a summary of component analysis, a method closely related to principal component analysis, is given in section 2.2. The close relationship between correspondence analysis and biplots will be clarified in section 2.3.1. In section 2.4 correspondence analysis is described in a more geometrical fashion. Rows and columns of the data matrix are treated as points in a space of high dimensionality. These points are then projected into
Chapter 2. Correspondence analysis

a space of few (usually two) dimensions in such a way that the approximated points are “closest”, in a least-squares sense, to the original points. The distance measure with respect to which the least-squares approximation takes place is referred to as chi-squared distance. Greenacre (1988) notes that this distance is in fact variance standardizing.

No formal derivation of this result is however provided. In section 2.4 the least-squares approximation of the rows and columns is considered with respect to the Mahalanobis, i.e. variance standardizing, distance. The equivalence between the chi-squared distance and the Mahalanobis distance, in the treatment of a contingency matrix, will be apparent.

We conclude this chapter with an example based on data concerning the role of emotions in economic decision making.

2.2 Component analysis

2.2.1 One-mode component analysis

Suppose we have an \( n \times p \) observed data matrix \( X \). The aim of one-mode component analysis is to approximate the \( p \) columns (usually representing the \( p \) variables) of \( X \) by linear combinations of a smaller number of vectors, say \( k \) (\( k \leq \text{rank}(X) \)), such that the residuals are as small as possible. Thus,

\[
x_{(j)} = \sum_{h=1}^{k} b_{jh} a_{(h)} + e_{(j)} \quad \text{(for } j = 1, \ldots p),
\]

where, generically, \( x_{(j)} \) denotes the \( j \)th column of \( X \) and \( e_{(j)} \) is the \( j \)th vector of residuals. Alternatively we can write

\[
X = AB' + E. \quad (2.1)
\]

Component analysis is closely related to sample principal component analysis where the aim is to find linear combinations of the variables such that the sample variance of the linear combinations is as high as possible. See, e.g., Magnus and Neudecker (1999).

In matrix notation, we can formulate as least-squares objective for one-mode component analysis

\[
\min_{A,B} \text{trace} \left( X - AB' \right)' \left( X - AB' \right) \quad (2.2)
\]
2.2. Component analysis

where $A$ is of order $n \times k$ and $B$ is of order $p \times k$. Without loss of generality we restrict $B$ in such a manner that $B'B = I_k$. Eckart and Young (1936) solved this approximation problem using the singular value decomposition of $X$. For later purposes, however, we will solve the minimization problem in (2.2) using matrix derivatives. An extensive treatment of matrix derivatives can be found in Magnus and Neudecker (1999).

By dropping the constant term and reversing the sign we can reformulate the problem as

$$\max_{A,B} \psi = 2 \text{trace} X'AB' - \text{trace} A'A$$

s.t. $B'B = I_k$.

This yields the Lagrangian expression

$$\phi = 2 \text{trace} X'AB' - \text{trace} AA' - \text{trace} L(B'B - I_k).$$

Due to the symmetry of the constraints we can, without loss of generality, take the Lagrangian multiplier matrix $L$ to be symmetric. Using this symmetry we consider variations in $A$ and $B$. Then,

$$d\phi = 2 \text{trace} \left( X'A - BL \right) dB' + 2 \text{trace} \left( B'X' - A' \right) dA.$$

Setting $d\phi = 0$ for arbitrary $dA$ and $dB$ yields together with the given constraint,

$$A = XB \quad \text{(2.3)}$$

$$X'A = BL \quad \text{(2.4)}$$

$$B'B = I_k. \quad \text{(2.5)}$$

From (2.3) and (2.4) it follows that

$$X'XB = BL. \quad \text{(2.6)}$$

Moreover, we see that

$$B'X'A = A'A = L,$$
whence $L$ is positive definite. From these expressions it follows that

$$\psi = 2 \text{trace } X'AB' - \text{trace } A'A = \text{trace } L$$

which is to be a maximum.

Let

$$L = U\Lambda U',$$

where $U$ is orthogonal and $\Lambda$ is diagonal, and define

$$\tilde{A} \equiv AU, \quad \tilde{B} \equiv BU.$$ 

Then

$$\tilde{A}'\tilde{A} = \Lambda,$$

and

$$\tilde{B}'\tilde{B} = I_k.$$ 

Equations (2.3) through (2.6) then become

$$\tilde{A} = X\tilde{B}$$

$$X'\tilde{A} = \tilde{B}\Lambda$$

$$\tilde{B}'\tilde{B} = I_k$$

$$X'X\tilde{B} = \tilde{B}\Lambda.$$ 

Thus, $\Lambda$, whose trace is to be maximized, is a diagonal matrix containing the eigenvalues of $X'X$. Moreover, as $\tilde{A}\tilde{B}' = AB'$ we can, without loss of generality drop the tilde in the above expressions. Hence, $B$ can be obtained from the eigenequation

$$X'XB = B\Lambda.$$ 

It is clear that the objective function in (2.2) is minimized when the $k$ largest eigenvalues, and corresponding eigenvectors, are selected.
2.2. Component analysis

2.2.2 Two-mode component analysis

If we are not only interested in reducing the $p$ columns of $X$, but also the $n$ rows we can apply two-mode component analysis\(^1\). The elements of $X$ are then expressed as

$$x_{ij} = \sum_{h=1}^{k_1} \sum_{l=1}^{k_2} a_{ih} z_{hl} b_{lj} + e_{ij},$$

where $k_1 \leq n$ and $k_2 \leq p$. In matrix notation we have

$$X = AZB' + E,$$

where $A$ is an $n \times k_1$ component matrix for the columns of $X$, $B$ is a $p \times k_2$ component matrix for the rows of $X$ and $Z$ is a $k_1 \times k_2$ matrix, usually referred to as the core matrix, relating the two component matrices. The aim of two-mode component analysis is to find such $A$, $B$ and $Z$ that the sum of squared residuals, i.e. $\sum_{i=1}^n \sum_{j=1}^p e_{ij}^2$, is as small as possible.

We can formulate as objective, together with two identification constraints,

$$\min_{A,Z,B} \text{trace}(X - AZB')(X - AZB)' \quad (2.7)$$

$$s.t. \quad A'A = I_{k_1}, \quad B'B = I_{k_2}.\$$

Here we will only treat the case where $k_1 = k_2 = k$, (the general case does not produce a better approximation and is easily derived from the case $k_1 = k_2$). Moreover, we take the core matrix $Z$ to be nonsingular, i.e. $\text{rank}(Z) = k$. The solution to this problem can be obtained immediately using the singular value decomposition of $X$, see Eckart and Young (1936). For the sake of completeness we will, in a similar fashion as before, give a complete derivation using matrix derivatives.

Derivation of the two-mode component analysis solution

Dropping the constant term $XX'$ and reversing the sign in (2.7) yields as equivalent objective

$$\max_{A,Z,B} \ 2 \text{trace} AZB'X' - \text{trace} ZZ' \quad (2.8)$$

\(^1\)In the context of three-mode factor analysis Levin (1965) describes a similar approach which he refers to as “a modification of two-mode factor analysis".
Chapter 2. Correspondence analysis

subject to the aforementioned constraints. Then, proceeding in a similar fashion as in section 2.2.1, we formulate as Lagrangian function

$$\phi = 2 \text{trace} AZB'X' - \text{trace} ZZ' - \text{trace} L_1(A'A - I_k) - \text{trace} L_2(B'B - I_k),$$

where $L_1$ and $L_2$ represent matrices of Lagrange multipliers. Note that, due to the symmetry of the constraints these matrices may, without loss of generality, also be taken to be symmetric. Taking derivatives and equating them to zero yields as first-order conditions

$$XBZ' = AL_1 \quad (2.9)$$
$$X'AZ = BL_2 \quad (2.10)$$
$$Z = A'XB, \quad (2.11)$$

where we have used the symmetry of $L_1$ and $L_2$. Inserting (2.11) in (2.10) yields

$$X'AA'XB = BL_2, \quad (2.12)$$

and similarly we obtain

$$XBB'X'A = AL_1 \quad (2.13)$$

after inserting (2.11) in (2.9). Also, pre-multiply (2.9) by $A'$ and (2.10) by $B'$ to get

$$L_1 = A'XBZ' = ZZ', \quad$$

and

$$L_2 = B'X'AZ = Z'Z.$$  

Although (2.12) and (2.13) resemble eigenequations, they have as disadvantage that no solution for one of the component matrices can be obtained without a solution for the other component matrix. However, there is a way around this problem. Consider a singular value decomposition of $Z$, i.e.,

$$Z = U\Lambda^\frac{1}{2}V',$$
2.2. Component analysis

where $U'U = V'V = I_k$. Then

$$L_1 = UAU'$$

and

$$L_2 = VAV'.$$

Inserting these expressions in (2.9) and (2.10) yields

$$XBV\Lambda^{\frac{1}{2}}U' = AU\Lambda U',$$  \hspace{1cm} (2.14)

and

$$X'AU\Lambda^{\frac{1}{2}}V' = BV\Lambda V'.$$  \hspace{1cm} (2.15)

Postmultiply (2.14) by $UA^{\frac{1}{2}}$ and (2.15) by $VA^{\frac{1}{2}}$ to get

$$XBV = AU\Lambda^{\frac{1}{2}},$$  \hspace{1cm} (2.16)

and

$$X'AU = BV\Lambda^{\frac{1}{2}}.$$  \hspace{1cm} (2.17)

Now, upon defining $\tilde{A} = AU$ and $\tilde{B} = BV$, premultiplication of (2.16) with $X'$ yields, after inserting the expression for $X'AU$, the eigenequation

$$X'X\tilde{B} = \tilde{B}\Lambda.$$

Similarly we obtain

$$XX'\tilde{A} = \tilde{A}\Lambda.$$

It is clear that

$$\tilde{A}'\tilde{A} = \tilde{B}'\tilde{B} = I_k,$$

and

$$AZB' = AU\Lambda^{\frac{1}{2}}V'B' = \tilde{A}\Lambda^{\frac{1}{2}}\tilde{B}'.$$
Chapter 2. Correspondence analysis

The objective (2.8) can be written as

$$\max_{A,Z,B} 2 \text{trace} ZB'X'A - \text{trace} ZZ' = \text{trace} ZZ' = trA.$$

Hence, the solution of two-way component analysis can be obtained by taking the $k$ largest eigenvalues and corresponding eigenvectors. As mentioned before, this solution can be obtained in one step by means of the singular value decomposition of $X$, i.e.

$$X \simeq AA'B'.$$

2.3 Two-mode component analysis of categorical data: correspondence analysis

Suppose we have as data matrix an $n \times p$ contingency matrix $F$, i.e. the entries $f_{ij}$ of $F$ denote the number of times an observation falls simultaneously into the $i$th category of the first, and the $j$th category of the second variable.

If one assumes independence between the row and column variables, the expected number of observations in each cell can be calculated as

$$e_{ij} = ss_{i} \pi_{.j}$$

where $s$ denotes the total number of observations, i.e.

$$s \equiv \sum_{i=1}^{n} \sum_{j=1}^{p} f_{ij} = 1_{n}'F1_{p},$$

where $1_{n}$ denotes an $n \times 1$ vector of ones and $\pi_{.i}$ is the probability of an observation falling into the $i$th category of the first variable, whereas $\pi_{.j}$ denotes the probability of an observation falling into the $j$th category of the second variable. Typically $\pi_{.i}$ and $\pi_{.j}$ are unknown, however, they can be estimated by taking the means of the rows and columns. Let

$$r \equiv F1_{p}, \text{ and } c \equiv F'1_{n},$$

i.e. $r$ is an $n \times 1$ vector of row totals, and $c$ is a $p \times 1$ vector of column totals. Then,

$$\bar{\pi}_{r} = \frac{1}{s}r, \text{ and } \bar{\pi}_{c} = \frac{1}{s}c.$$

(2.18)
2.3. Two-mode component analysis of categorical data: correspondence analysis

Thus, \( \pi_r \) is the \( n \times 1 \) vector of row means and \( \pi_c \) is the \( p \times 1 \) vector of column means.

To test the hypothesis that the rows and columns of the contingency matrix \( F \) are independent we define

\[
T \equiv \sum_{i=1}^{n} \sum_{j=1}^{p} \frac{(f_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}}, \tag{2.20}
\]

where \( \hat{e}_{ij} \) denotes the estimated expected frequency under the assumption of independence between rows and columns, i.e.

\[
\hat{e}_{ij} \equiv s \hat{\pi}_i \hat{\pi}_j = \frac{1}{s} r_i c_j,
\]

or, expressed in matrix notation,

\[
E \equiv \frac{1}{s} rc'.
\]

This statistic is well known and is referred to as the Pearson chi-squared statistic for testing independence.

We can express \( T \) in matrix notation as

\[
T = s \times \text{trace} D_c^{-1} \left( F - \frac{1}{s} rc' \right)' D_r^{-1} \left( F - \frac{1}{s} rc' \right), \tag{2.21}
\]

where \( D_r \) and \( D_c \) are diagonal matrices with as elements the elements of the vectors \( r \) and \( c \) respectively. Hence, \( D_r \) is an \( n \times n \) matrix and \( D_c \) is of the order \( p \times p \). It can be shown that, under the assumption of independence, \( T \) is asymptotically chi-squared distributed with \((n - 1) \times (p - 1)\) degrees of freedom. (For a proof see Kendall, Stuart and Ord, 1987).

Correspondence analysis is concerned with the analysis of the matrix of deviations between the observed frequencies and the expected frequencies, i.e. the analysis of \( F - \frac{1}{s} rc' \). Correspondence analysis is in fact the two-way component analysis of the matrix of deviations where weights have been assigned to the rows and columns in a similar way as was done in the calculation of the chi-squared statistic in (2.21), i.e. the entries are divided by the expected frequencies.

The weighted matrix of residuals is

\[
\tilde{F} \equiv D_r^{-\frac{1}{2}} \left( F - \frac{1}{s} rc' \right) D_c^{-\frac{1}{2}}. \tag{2.22}
\]
By assigning weights in this fashion a relatively large weight is assigned to rows and
columns that have a small number of occurrences, whereas a relatively small weight is
assigned to rows and columns that have a large number of occurrences. It should be
noted that Greenacre (1984, 1993b) bases his analysis on the matrix of proportions, i.e.
$P = \frac{1}{2} F$, rather than on $F$. However, as he also defines scaling matrices $D_r$ and $D_c$ based
on this matrix of proportions, all differences cancel out.

Consider the two-way component analysis of $\tilde{F}$. As was shown in section 2.2.2 the
solution of such an analysis can be obtained through the singular value decomposition of
this weighted residual matrix, i.e.

$$\tilde{F} = U \Lambda^{\frac{1}{2}} V', \quad (2.23)$$

where

$$U'U = V'V = I_\kappa, \quad (2.24)$$

$\Lambda^{\frac{1}{2}}$ is the diagonal matrix of singular values and $\kappa$ denotes the rank of $\tilde{F}$. The matrices $U$
$(n \times \kappa)$ and $V (p \times \kappa)$ are the component matrices whereas $\Lambda^{\frac{1}{2}}$ is the core-matrix relating
the two component matrices.

As the aim of two-way component analysis is to reduce the dimensionality of the
original data, one would like to approximate the weighted residual matrix by a matrix
of lower rank. This can be done by selecting the $k$ ($k < \kappa$) largest singular values
and corresponding columns of $U$ and $V$. For convenience however we will consider the
full decomposition to derive some important correspondence analysis equations. It is
understood that, as all definitions are based on the singular value decomposition (2.23),
an approximation is obtained simply by selecting the first $k$ columns of $U$ and $V$ and the
corresponding $k$ largest singular values.

Principal axes

We define matrices $A$ and $B$ as

$$A = D_r^{\frac{1}{2}} U \quad (2.25)$$
2.3. Two-mode component analysis of categorical data: correspondence analysis

and

\[ B \equiv D^\frac{1}{2} V. \]  

(2.26)

The columns of A and B are called the *principal axes* for the columns and rows of F respectively. It follows from (2.24) that they are standardized as

\[ A'D_r^{-1}A = B'D_c^{-1}B = I_k. \]

Principal coordinates

Greenacre (1984) defines so-called *principal coordinate* matrices G and H as

\[ G \equiv \sqrt{s}D_r^{-\frac{1}{2}}U\Lambda^\frac{1}{2}, \]  

(2.27)

and

\[ H \equiv \sqrt{s}D_c^{-\frac{1}{2}}V\Lambda^\frac{1}{2}, \]  

(2.28)

so that

\[ G'D_rG = H'D_cH = s\Lambda. \]  

(2.29)

The matrix G is referred to as the matrix of principal row coordinates, whereas the matrix H is referred to as the matrix of principal column coordinates. By choosing the solution to be of low rank (usually \( k = 2 \)) one can plot the principal coordinates, i.e. the rows of G and H, as points in a graph.

Rewriting (2.23) as

\[ U = D_r^{-\frac{1}{2}} \left( F - \frac{1}{s} rc' \right) D_c^{-\frac{1}{2}} V \Lambda^{-\frac{1}{2}} \]  

(2.30)

and inserting this expression in (2.27) yields after premultiplying by \( r' \)

\[ r'G = \sqrt{s}1' \left( F - \frac{1}{s} rc' \right) D_c^{-\frac{1}{2}} V = \sqrt{s} \left( c' - c' \right) D_c^{-\frac{1}{2}} V = 0. \]  

(2.31)

A similar argument leads to

\[ c'H = 0. \]  

(2.32)
Standard coordinates

In addition to the principal coordinate matrices $G$ and $H$, Greenacre (1984) defines so-called standard coordinate matrices, say $X$ and $Y$, as

$$X \equiv \sqrt{sD_r^{-\frac{1}{2}}} U = GA^{-\frac{1}{2}}, \quad (2.33)$$

and

$$Y \equiv \sqrt{sD_c^{-\frac{1}{2}}} V = HA^{-\frac{1}{2}} \quad (2.34)$$

so that

$$X' D_r X = Y' D_c Y = sI_k. \quad (2.35)$$

It is not difficult to see that both the principal and the standard coordinate matrices can be obtained from eigenequations. In the case of the standard coordinate matrices (2.33) and (2.34), we have

$$D_r^{-1} \left( F - \frac{1}{s} rc' \right) D_c^{-1} \left( F - \frac{1}{s} rc' \right)' X = XL, \quad (2.36)$$

and

$$D_c^{-1} \left( F - \frac{1}{s} rc' \right)' \ D_r^{-1} \left( F - \frac{1}{s} rc' \right) Y = YL. \quad (2.37)$$

both of which follow from (2.23), (2.33) and (2.34). Replacing $X$ by $GA^{-\frac{1}{2}}$ and $Y$ by $HA^{-\frac{1}{2}}$ yields the same eigenequations for $G$ and $H$.

Transition formulae

The row and column coordinate matrices obtained in correspondence analysis are related through so-called transition formulae. For the standard coordinate matrices $X$ and $Y$ we have

$$X = D_r^{-1} FYL^{-\frac{1}{2}}, \quad (2.38)$$

and

$$Y = D_c^{-1} F' XL^{-\frac{1}{2}}. \quad (2.39)$$
2.3. Two-mode component analysis of categorical data: correspondence analysis

(Similar formulae can be obtained for the principal coordinate matrices using (2.33) and (2.34)).

To obtain (2.38) insert (2.30) in (2.33), i.e.

\[ X = \sqrt{s} D_r^{-1} \left( F - \frac{1}{s} rc' \right) D_c^{-\frac{1}{2}} V \Lambda^{-\frac{1}{2}}. \]

Then, from (2.32) and (2.34) it immediately follows that

\[ X = D_r^{-1} FY \Lambda^{-\frac{1}{2}}. \]

In a similar fashion we obtain (2.39).

Inertia

From (2.29) it is clear that the weighted squared lengths of the principal coordinates are equal to \( s \) times the squared singular values. As these squared singular values are the sums of squared deviations from independence, they are the correspondence analysis analogues of variance. They are referred to as inertias.

Note that the test statistic \( T \) for testing the hypothesis of independence is equal to \( s \) times the sum of all inertias, i.e.

\[ T = s \times \text{trace } \Lambda. \]

The inertias are often used to assess the quality of a \( k \)-dimensional approximation. For example, \( \sum_{j=1}^{k} \gamma_j \) where \( \gamma_j \) is defined as

\[ \gamma_j \equiv \frac{\lambda_j}{\sum_{i=1}^{s} \lambda_i}, \quad (2.40) \]

indicates how much inertia is accounted for by the \( k \)-dimensional approximation.

Distances

Distances between the principal coordinates are so-called chi-squared distances. To see this we first need to define chi-squared distance, which is related to the chi-squared statistic \( T \) defined in (2.20).
Gifi (1990, p. 266) defines the chi-squared distance $\delta_{ij}$ between row $i$ and $j$ of a contingency matrix $F$ as

$$\delta_{ij}^2 \equiv s \times \sum_{l=1}^{p} \frac{1}{c_l} \left( \frac{f_{il}}{r_i} - \frac{f_{jl}}{r_j} \right)^2. \quad (2.41)$$

An alternative definition of chi-squared distance, given by Gower and Hand (1996, p. 176), does not contain the constant $s$. As this is clearly of no influence we will, for convenience, use the definition in (2.41).

We can express $\delta_{ij}^2$ as

$$\delta_{ij}^2 = s \times (e_i - e_j)' D_r^{-1} \left( F - \frac{1}{s} r c' \right) D_c^{-1} \left( F - \frac{1}{s} r c' \right)' D_r^{-1} (e_i - e_j),$$

where $e_i$ is the $i$th unit vector of order $n \times 1$. Then, from (2.23) and (2.27) it immediately follows that

$$\delta_{ij}^2 = s \times (e_i - e_j)' D_r^{-\frac{1}{2}} U \Lambda U' D_r^{-\frac{1}{2}} (e_i - e_j)' = (e_i - e_j)' GG' (e_i - e_j). \quad (2.42)$$

Hence, the Euclidean distances between the principal row coordinates are chi-squared distances. If $k$ is chosen smaller than $\kappa$ the equality in (2.42) does not hold and the distances between the principal row coordinates are approximated chi-squared distances. Following a similar approach we can show that, for $k = \kappa$, the chi-squared distances between the columns of $F$ are equal to the distances between the principal column coordinates, i.e. the rows of $H$.

Note that distances between rows of $G$ and rows of $H$ are not defined.

### 2.3.1 Biplot

In a biplot as introduced by Gabriel (1971), rows and columns of a matrix are plotted in low (usually two) dimensional space such that the inner products of the coordinate vectors for the rows and columns are least-squares approximations of the original elements of the data matrix. For example, consider the $n \times p$ data matrix $X$, then

$$X = AB' + E, \quad (2.43)$$
where $A$ is an $n \times k$ matrix of row coordinates, $B$ is an $p \times k$ matrix of column coordinates and $E$ is an $n \times k$ matrix of residuals. The rows of $A$ give $n$ coordinate vectors for the rows, and the rows of $B$ are the $p$ coordinate vectors for the columns. If $k = 2$, a scatter plot can be obtained representing a two-dimensional approximation of the original data. Clearly the choice of $A$ and $B$ is not unique. For example, postmultiply $A$ by a non-singular matrix $T$. Then, postmultiplication of $B$ by $T^{-1}$ does not affect the approximation.

Comparing (2.1) and (2.43) immediately shows the equivalence between one-mode component analysis and a biplot. In fact, a biplot can be obtained by plotting the rows of $A$ and $B$ obtained in the one-mode component analysis of $X$.

Recall the singular value decomposition essential in correspondence analysis, i.e.

$$D_r^{-\frac{1}{2}} \left( F - \frac{1}{s} rc' \right) D_c^{-\frac{1}{2}} = U \Lambda^\frac{1}{2} V'. $$

Using (2.33) and (2.34) we can write

$$s \times D_r^{-\frac{1}{2}} \left( F - \frac{1}{s} rc' \right) D_c^{-\frac{1}{2}} = D_r^{\frac{1}{2}} X H' D_c^{\frac{1}{2}} = D_r^{\frac{1}{2}} G Y' D_c^{\frac{1}{2}},$$

or, alternatively,

$$s \times \tilde{f}_{ij} = \sqrt{\tilde{r}_i \tilde{c}_j} x_i h_j = \sqrt{\tilde{r}_i \tilde{c}_j} g_i y_j.$$

Thus, the correspondence analysis plot of $X$ and $H$ is a biplot where the coordinates have been weighted by the square roots of the row and column totals respectively. A similar argument applies to the correspondence analysis plot of $G$ and $Y$. Greenacre (1993a) describes the relationship between biplots and correspondence analysis in more detail. Gower and Hand (1996) introduce correspondence analysis as a type of biplot.

### 2.4 Correspondence analysis using Mahalanobis distance

In the previous section we derived the correspondence analysis solution using two-mode component analysis. We found that the distances between principal coordinates were chi-squared distances. Greenacre (1984) describes correspondence analysis as the simultaneous low-rank approximation of row and column profiles, which we will define later,
by minimization of weighted chi-squared distance between the original row and column profiles and their low-rank approximations. For a definition of weighted chi-squared distance, as well as an extensive argumentation for its applicability in the analysis of the row and column profiles, we refer to Greenacre (1984). In this section we will follow a similar approach. However, instead of minimizing the weighted chi-squared distances we will consider the minimization of the squared Mahalanobis, i.e. variance standardizing, distance. The results we obtain are identical to the ones derived in the previous section.

2.4.1 The analysis of the row profiles

Consider the $n \times p$ contingency matrix $F$. The rows of $F$ can be seen as observations from a multinomial distribution with

$$E(f_i) = r_i \pi_c,$$  \hspace{1cm} (2.44)

$$\text{Var}(f_i) = r_i (D_{\pi_c} - \pi_c \pi'_c),$$  \hspace{1cm} (2.45)

where $D_{\pi_c}$ is a diagonal matrix with as entries on the diagonal the elements of $\pi_c$, and $r_i$ -the $i$th element of $r$ as defined in (2.18)-- denotes the sample size.

Thus, it is assumed that the $i$th row of $F$ represents a sample of size $r_i$ from a multinomial distribution with as expectation a $p \times 1$ vector $\pi_c$ (where $\pi'_c 1_p = 1$) multiplied by the sample size $r_i$.

A matrix of proportions, say $R$, can be obtained by dividing the rows of $F$ through their sample sizes, i.e.

$$R \equiv D^{-1}_r F,$$  \hspace{1cm} (2.46)

so that $R 1_p = 1_n$. These proportions are often referred to as row profiles. For each row they give the distribution over the $p$ columns. Let $R_i$ denote the $i$th row profile, i.e. the $i$th row of $R$ written as a column, then it follows immediately from (2.44), (2.45) and (2.46) that

$$E(R_i) = \pi_c,$$
and

\[ \text{Var}(R_i) = \frac{1}{r_i} (D_{\pi_c} - \pi_c \pi_c'). \]

A squared Mahalanobis distance can be defined for the distances between the rows of \( R \).

However, as the variance matrix of \( R_i \) is singular (post-multiply \( \text{Var}(R_i) \) by \( 1_p \) to see this) we have to use the Moore-Penrose inverse. It is easily verified that the Moore-Penrose inverse of \( D_{\pi_c} - \pi_c \pi_c' \) is \( MD_{\pi_c}^{-1}M \), where \( M \) is the centering matrix of the appropriate order, i.e.

\[ M = I_p - \frac{1}{p} 1_p 1_p', \]

see e.g. Tanabe and Sagae (1992) and Neudecker (1995). Hence, as expression for the squared Mahalanobis distance between row profiles \( i \) and \( j \), e.g. see Mardia et al. (1979) we have

\[ d_{ij}^2 = \frac{r_ir_j}{r_i + r_j} (e_i - e_j)' RD_{\pi_c}^{-1} R' (e_i - e_j). \]

Note that the centering matrix \( M \) does not occur in this expression because the rows of \( R \) sum to 1.

We are interested in approximating the matrix \( R \) by linear combinations, say \( \hat{R} = WB' \), where the matrices \( W \) and \( B \), both of full column rank, are of order \( n \times k \) and \( p \times k \) \((k \leq \kappa)\) respectively. By doing so we manage to reduce the dimensionality \( p \) of the data and we simplify their interpretation. Considering the distribution of the rows of \( F \) we formulate as objective for approximating \( R \) by \( \hat{R} \) the minimization of their squared Mahalanobis distance. However, as \( D_{\pi_c}^{-1} \) depends on the unknown vector \( \pi_c \), an estimator needs to be introduced. We will use the maximum-likelihood estimator, i.e. the sample average as defined in (2.19). Hence

\[ \hat{D}_{\pi_c} = \frac{1}{s} D_{\pi_c}. \] (2.47)

As expression for the Moore-Penrose inverse for the estimated covariance matrix of the \( i \)th row profile we get

\[ \text{Var}(R_i)^+ = sr_i M D_{\pi_c}^{-1} M. \] (2.48)
Chapter 2. Correspondence analysis

As only the relative positions of the row-profiles are of interest to us we will approximate the matrix of row-profiles in deviations from the mean profile, i.e.

\[ \tilde{R} \equiv R - \frac{1}{s} 1c', \]

instead of the non-centered row-profile matrix \( R \).

Using (2.48) we can formulate as objective

\[
\min_{B, W} \quad \text{trace} \left( D_r (\tilde{R} - WB') MD_c^{-1} M (\tilde{R} - WB')' \right)
\]

\[ \text{s.t. } B'MD_c^{-1}MB = I_k, \]

(2.49)

where we have dropped the constant \( s \). Upon defining \( \tilde{B} \equiv MB \) and because \( \tilde{R}M = \tilde{R} \) we can rewrite this objective as

\[
\min_{\tilde{B}, W} \quad \text{trace} \left( D_r (\tilde{R} - \tilde{WB}') D_c^{-1} (\tilde{R} - \tilde{WB}')' \right)
\]

\[ \text{s.t. } \tilde{B}'D_c^{-1}\tilde{B} = I_k. \]

For convenience, we will hereafter drop the tilde on \( B \).

Let

\[ \tilde{R} \equiv D_c^{\frac{1}{2}} \tilde{R} D_c^{-\frac{1}{2}}, \]

\[ U \equiv D_c^{\frac{1}{2}} W, \]

and

\[ V \equiv D_c^{-\frac{1}{2}} B. \]

Then (2.49) can be rewritten as

\[
\min_{U, V} \quad \text{trace} \left( \tilde{R} - UV \right) \left( \tilde{R} - UV \right)' \]

\[ \text{s.t. } V'V = I_k. \]
This problem is equivalent to the one-mode component analysis problem described in section 2.2.1. Hence, as solution we have

$$\tilde{R}' \tilde{R} V = V \Lambda,$$

$$U = \tilde{R} V,$$

$$V' V = I_k.$$

Inserting the expressions for $\tilde{R}$, $U$ and $V$ yields as first-order conditions

$$\tilde{R}' D_r \tilde{R} D_c^{-1} B = B \Lambda, \quad (2.50)$$

$$W = \tilde{R} D_c^{-1} B, \quad (2.51)$$

$$B' D_c^{-1} B = I_k. \quad (2.52)$$

As

$$D_r \tilde{R} = D_r \left( R - \frac{1}{s} \text{rc}' \right) = F - \frac{1}{s} \text{rc}', \quad (2.53)$$

(2.50) becomes

$$\left( F - \frac{1}{s} \text{rc}' \right)' D_r^{-1} \left( F - \frac{1}{s} \text{rc}' \right) D_c^{-1} B = B \Lambda,$$

and from (2.23) and (2.26) it immediately follows that the matrix $B$ obtained from the minimization of the Mahalanobis distance is equivalent to the matrix of principal axes for the rows as defined in (2.26). Moreover,

$$W = D_r^{-1} \left( F - \frac{1}{s} \text{rc}' \right) D_c^{-1} B = D_r^{-\frac{1}{2}} U \Lambda^{\frac{1}{2}},$$

where we used (2.23) and (2.53). Hence, the rows of $W$, which can be seen as coordinates with respect to the principal axes $B$, are equivalent to the principal row coordinates $G$, defined in (2.27), multiplied by a constant $\frac{1}{\sqrt{s}}$. 

2.4. Correspondence analysis using Mahalanobis distance
2.4.2 The analysis of the column profiles

Instead of assuming that the rows of $F$ are observations from a multinomial distribution we could also consider the columns $f_{(j)}$ of $F$ to be observations from a multinomial distribution. Then

$$E(f_{(j)}) = c_j \pi_r,$$

$$Var(f_{(j)}) = c_j (D_{\pi_r} - \pi_r \pi_r'),$$

where $D_{\pi_r}$ is an $n \times n$ diagonal matrix with as entries on the diagonal the elements of the $n \times 1$ vector $\pi_r$. Like before we will estimate this expected value by the sample average, i.e.

$$\hat{\pi}_r = \frac{1}{s} r.$$

Proceeding in a similar fashion as before an $n \times p$ matrix $P_c$ of so-called column profiles, can be constructed by dividing each column of $F$ through its total, i.e.

$$C = D_c^{-1}F',$$

so that $C 1_n = 1_p$. Following the same procedure as before the objective for approximating $\bar{C} = C - \frac{1}{s} 1_p r'$ by $QA'$ becomes

$$\min_{A,Q} \text{trace} D_c (\bar{C}_c - QA') D_c^{-1} (\bar{C} - QA')',$$

s.t. $A'D_r^{-1}A = I_k,$

where the rank $k$ matrix $Q$ is of the order $p \times k$ and $A$ is an $n \times k$ matrix of rank $k$ ($k \leq \kappa$). Then, following the same procedure outlined in the previous section, we obtain

$$\left( F - \frac{1}{s} rc' \right) D_c^{-1} \left( F - \frac{1}{s} rc' \right)' D_r^{-1} A = AA,'$$

where $A$ is standardized as in (2.56) and $\Lambda$ is a diagonal matrix containing the $k$ largest eigenvalues of $\left( F - \frac{1}{s} rc' \right) D_c^{-1} \left( F - \frac{1}{s} rc' \right)' D_r^{-1}$, and

$$Q = \bar{C} D_r^{-1} A.$$
From (2.23) and (2.25) it follows immediately that the principal axes obtained in the analysis of the column profiles, i.e. A obtained form (2.57), is equivalent to the matrix of principal axes for the columns as defined in (2.25). Furthermore, it is not difficult to show that $\sqrt{s}Q$ is equivalent to the matrix of principal column coordinates H as defined in (2.28), viz

$$Q = D_c^{-1} \left( F' - \frac{1}{s} cr' \right) D_r^{-1} A = D_c^{-\frac{1}{2}} VA^\frac{1}{2},$$

where we used (2.23).

Combining these results it is clear that correspondence analysis as described in section 2.3 is equivalent to the simultaneous approximation of row- and column-profiles through the minimization of the appropriate squared Mahalanobis distance. Moreover, the two problems described in this section, can be solved in one step by means of the singular value decomposition of $D_r^{-\frac{1}{2}} (F - \frac{1}{s} cr') D_c^{-\frac{1}{2}}$.

### 2.5 An example

In order to illustrate the use of correspondence analysis we consider a data set kindly made available to us by Ronald Bosman and Frans van Winden of the Universiteit van Amsterdam. Bosman and van Winden (1999) studied the role of emotions in economic decision making. An experiment was conducted in which participants had to play a one-shot two player power-to-take game. The participants were randomly divided into pairs consisting of a take authority and a responder. The game consisted of two stages. In the first stage, the take authority had to decide how much income\(^2\) was to be transferred from the responder to the take authority after the second stage (the so-called take rate). In the second stage the responder was given the opportunity to punish the take authority by destroying own income. For a complete description of the experiment as well as an overview of the role played by emotions in economics see Bosman and van Winden (1999).

After the game the participants were asked to answer several questions concerning their emotions during the experiment. In one of these questions the responders were

\(^2\)The income in this experiment was earned by the participants in an individual real-effort decision making experiment preceeding the power-to-take game.
Table 2.1: Frequency matrix emotion data

<table>
<thead>
<tr>
<th></th>
<th>not at all</th>
<th>Irritated 2</th>
<th>Irritated 3</th>
<th>Irritated 4</th>
<th>Irritated 5</th>
<th>Irritated 6</th>
<th>very intensely</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-20</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>21-40</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>41-60</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>61-80</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>81-100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

asked to think back of the moment they were confronted with the take rate proposed by the take authority. They then had to indicate, on a 7 point scale reaching from not at all to very intensely, to what extent they experienced irritation.

In table 2.1 a cross-tabulation of the 7 irritation categories and 5 take rate categories is given. For this purpose the take rates, which could reach from 0-100, where coded into 5 categories: 0-20, 21-40, 41-60, 61-80 and 81-100.

Applying correspondence analysis to the frequency matrix in table 2.1 yields a two-dimensional approximation that accounts for 90.31% of the total inertia. In figure 2.1 we have plotted the rows, representing the take rates, in standard coordinates, and the columns, representing the irritation levels, in principal coordinates. We immediately see that the take rates are separated along the first axes from high (left) to low (right). The experienced irritation is, following a similar pattern, separated along the first axes from very intensely irritated (left) to not at all irritated (right). The second axis, the y-axis, separates the “intermediate” take-rates from the two extremes. In a similar fashion the categories very intensely and, to a lesser extent, not at all are separated from the less outspoken experienced emotions.

From figure 2.1 it appears that there is a positive correlation between the take rate proposed by the take authority and the irritation level of the responder. In other words, on average, responders confronted with a high take rate had a tendency to experience more irritation than those confronted with a lower take rate and vice versa.
Figure 2.1: Correspondence analysis biplot. Take rates in standard coordinates, irritation categories in principal coordinates.