CORRESPONDENCE

The CORRESPONDENCE algorithm consists of three major parts:

1. A singular value decomposition (SVD)
2. Centering and rescaling of the data and various rescalings of the results
3. Variance estimation by the delta method.

Other names for SVD are "Eckart-Young decomposition" after Eckart and Young (1936), who introduced the technique in psychometrics, and "basic structure" (Horst, 1963). The rescalings and centering, including their rationale, are well explained in Benzécri (1969), Nishisato (1980), Gifi (1981), and Greenacre (1984). Those who are interested in the general framework of matrix approximation and reduction of dimensionality with positive definite row and column metrics are referred to Rao (1980). The delta method is a method that can be used for the derivation of asymptotic distributions and is particularly useful for the approximation of the variance of complex statistics. There are many versions of the delta method, differing in the assumptions made and in the strength of the approximation (Rao, 1973, ch. 6; Bishop et al., 1975, ch. 14; Wolter, 1985, ch. 6).

Other characteristic features of CORRESPONDENCE are the ability to fit supplementary points into the space defined by the active points, the ability to constrain rows and/or columns to have equal scores, and the ability to make biplots using either chi-squared distances, as in standard correspondence analysis, or Euclidean distances.

Notation

The following notation is used throughout this chapter unless otherwise stated:

t_1	Total number of rows (row objects)
s_1	Number of supplementary rows
k_1	Number of rows in analysis (i $t_1 - s_1$)
t_2	Total number of columns (column objects)
s_2	Number of supplementary columns

k_2	Number of columns in analysis ($t_2 - s_2$)
p	Number of dimensions

Data-Related Quantities

f_{ij}	Nonnegative data value for row *i* and column *j*: collected in table *F*
f_{i+}	Marginal total of row *i*, $i = 1,\ldots,k_1$
f_{+j}	Marginal total of column *j*, $j = 1,\ldots,k_2$
N	Grand total of *F*

Scores and Statistics

r_{is}	Score of row object *i* on dimension *s*
c_{js}	Score of column object *j* on dimension *s*
I	Total inertia

Basic Calculations

One way to phrase the CORRESPONDENCE objective (cf. Heiser, 1981) is to say that we wish to find row scores $\{r_{is}\}$ and column scores $\{c_{js}\}$ so that the function

$$\sigma\left(\{r_{is}\};\{c_{js}\}\right) = \sum_i \sum_j f_{ij} \sum_s \left(r_{is} - c_{js}\right)^2$$

is minimal, under the standardization restriction either that

$$\sum_i f_{i+}\, r_{is} r_{it} = \delta^{st}$$

or

$$\sum_j f_{+j}\, c_{js} c_{jt} = \delta^{st}$$

where δ^{st} is Kronecker's delta and t is an alternative index for dimensions. The trivial set of scores ({1},{1}) is excluded.

The CORRESPONDENCE algorithm can be subdivided into five steps, as explained below.

1. Data scaling and centering

When rows and/or columns are specified to be equal, first the frequencies of the rows/columns to be equal are summed. The sums are put into the row/column with the smallest row/column number and the other rows/columns are set to zero.

1.1 Measure is Chi Square

The first step is to form the auxiliary matrix **Z** with general element

$$z_{ij} = \frac{f_{ij}}{\sqrt{f_{i+} f_{+j}}} - \frac{\sqrt{f_{i+} f_{+j}}}{N}$$

The standardization with Chi Square measure is always rcmean (both row and column means removed.

1.2 Measure is Euclidean

When Euclidean measure is choosen, the auxiliary matrix **Z** is formed in two steps:

1. $\tilde{f}_{ij} = f_{ij} - \dfrac{\tilde{f}_{i+}\tilde{f}_{+j}}{N}$

With $\tilde{f}_{ij}$, $\tilde{f}_{i+}$, and $\tilde{f}_{j+}$ depending on the standardization option.:

(a) standardization option rmean (remove row means)

$$\tilde{f}_{ij} = f_{ij},\ \tilde{f}_{i+} = f_{i+},\ \tilde{f}_{+j} = \frac{N}{k_2}$$

(b) standardization option cmean (remove column means)

$$f_{ij}^{\sim} = f_{ij},\ f_{i+}^{\sim} = \frac{N}{k_1},\ f_{j+}^{\sim} = f_{+j}$$

(c) rcmean (remove both row and column means)

$$f_{ij}^{\sim} = f_{ij},\ f_{i+}^{\sim} = f_{i+},\ f_{j+}^{\sim} = f_{+j}$$

(d) standardization option rsum (equalize row totals, then remove row means)

$$f_{ij}^{\sim} = \frac{f_{ij} f_{i+}^{\sim}}{f_{i+}},\ f_{i+}^{\sim} = \frac{N}{k_1},\ f_{+j}^{\sim} = \frac{N}{k_2}$$

(e) standardization option csum (equalize column totals, then remove column means)

$$f_{ij}^{\sim} = \frac{f_{ij} f_{+j}^{\sim}}{f_{+j}},\ f_{i+}^{\sim} = \frac{N}{k_1},\ f_{+j}^{\sim} = \frac{N}{k_2}$$

2. Then, if not computed yet in step 1, $f_{i+}^{\sim}$, or/and $f_{j+}^{\sim}$ are computed:

$$f_{i+}^{\sim} = \frac{N}{k_1},\ f_{+j}^{\sim} = \frac{N}{k_2}, \text{ and}$$

$$z_{ij} = \frac{f_{ij}^{\sim}}{\sqrt{f_{i+}^{\sim} f_{+j}^{\sim}}}$$

2. Singular value decomposition

When rows and/or columns are specified as supplementary, first these rows and/or colums of $\mathbf{Z}$ are set to zero, yielding $\underline{\mathbf{Z}}$

Let the singular value decomposition of $\underline{\mathbf{Z}}$ be denoted by

$$\underline{\mathbf{Z}} = \mathbf{K}\Lambda\mathbf{L}'$$

with $\mathbf{K}'\mathbf{K} = \mathbf{I}$, $\mathbf{L}'\mathbf{L} = \mathbf{I}$, and Λ diagonal. This decomposition is calculated by a routine based on Golub and Reinsch (1971). It involves Householder reduction to bidiagonal form and diagonalization by a QR procedure with shifts. The routine requires an array with more rows than columns, so when $k_1 < k_2$ the original table is transposed and the parameter transfer is permuted accordingly.

3. Adjustment to the row and column metric

The arrays of both the left-hand singular vectors and the right-hand singular vectors are adjusted row-wise to form scores that are standardized in the row and in the column marginal proportions, respectively:

$$\tilde{r}_{is} = k_{is} \Big/ \sqrt{\tilde{f}_{i+} / N}\,,$$

$$\tilde{c}_{js} = l_{js} \Big/ \sqrt{\tilde{f}_{+j} / N}\,.$$

This way, both sets of scores satisfy the standardization restrictions simultaneously.

4. Determination of variances and covariances

For the application of the delta method to the results of generalized eigenvalue methods under multinomial sampling, the reader is referred to Gifi (1990, ch. 12) and Israëls (1987, Appendix B). It is shown there that *N* time variance-covariance matrix of a function ϕ of the observed cell proportions $p = \left\{p_{ij} = \tilde{f}_{ij} / N\right\}$ asymptotically reaches the form

$$N \times \operatorname{cov}(\phi(p)) \cong \sum_i \sum_j \pi_{ij} \left(\frac{\partial \phi}{\partial p_{ij}}\right)\left(\frac{\partial \phi}{\partial p_{ij}}\right)' - \left(\sum_i \sum_j \pi_{ij} \frac{\partial \phi}{\partial p_{ij}}\right)\left(\sum_i \sum_j \pi_{ij} \frac{\partial \phi}{\partial p_{ij}}\right)'$$

Here the quantities π_{ij} are the cell probabilities of the multinomial distribution, and $\partial\phi/\partial p_{ij}$ are the partial derivatives of ϕ (which is either a generalized eigenvalue or a generalized eigenvector) with respect to the observed cell

proportion. Expressions for these partial derivatives can also be found in the above-mentioned references.

5. Normalization of row and column scores

Depending on the normalization option chosen, the scores are normalized. The normalization option q can be chosen to be any value in the interval [-1,1] or it can be specified according to the following designations:

$$q = \begin{cases} 0, & \textit{symmetrical} \\ 1, & \textit{row principal} \\ -1, & \textit{column principal} \end{cases}$$

There is a fifth possibility, choosing the designation “principal”, that does not correspond to a q-value.

When “principal” is chosen, normalization parameters α for the rows and β for the columns are both set to 1. When one of the other options is chosen, α and β are functions of q:

$\alpha = (1+q)/2,$

$\beta = (1-q)/2.$

The normalization implies a compensatory rescaling of the coordinate axes of the row scores and the column scores:

$$r_{is} = \tilde{r}_{is}\lambda_s^{\alpha},$$

$$c_{js} = \tilde{c}_{js}\lambda_s^{\beta}.$$

The general formula for the weighted sum of squares that results from this rescaling is

$$\text{row scores}: \quad \sum_i \tilde{f}_{i+} r_{is}^2 = N\lambda_s^{2\alpha}$$

$$\text{column scores}: \quad \sum_j \tilde{f}_{+j} c_{js}^2 = N\lambda_s^{2\beta}$$

The estimated variances and covariances are adjusted according to the type of normalization chosen.

Diagnostics

After printing the data, CORRESPONDENCE optionally also prints a table of row profiles and column profiles, which are $\{f_{ij} / f_{i+}\}$ and $\{f_{ij} / f_{+j}\}$, respectively.

Singular Values, Maximum Rank and Inertia

All singular values λ_s defined in step 2 are printed up to a maximum of $\min\{(k_1 - 1), (k_2 - 1)\}$. Small singular values and corresponding dimensions are suppressed when they don't exceed the quantity $(k_1 k_2)^{1/2} 10^{-7}$; in this case a warning message is issued. Dimensionwise inertia and total inertia are given by the relationships

$$I = \sum_s \lambda_s^2 = \sum_s \sum_i \frac{\tilde{f}_{i+} r_{is}^2}{N}$$

where the right-hand part of this equality is true only if the normalization is row principal (but for the other normalizations similar relationships are easily derived from step 5). The quantities "proportion explained" are equal to inertia divided by total inertia: λ_s^2 / I.

Supplementary Points

Supplementary row and column points are given by

$$r_{is}^{\text{sup}} = \sum_j \frac{\tilde{f}_{ij}}{\tilde{f}_{i+}} c_{js} \lambda_s^{2\alpha-2}$$

$$c_{js}^{\text{sup}} = \sum_i \frac{\tilde{f}_{ij}}{\tilde{f}_{+j}} r_{is} \lambda_s^{2\beta-2}$$

Mass, Scores, Inertia and Contributions

The mass, scores, inertia and contributions for the row and columns points (including supplementary points) are given in the Overview Row Points Table and the Overview Column Points Table. These tables are printed in *p* dimensions. The tables are given first for rows, then for columns. The masses are the marginal proportions ($\tilde{f}_{i+} / N$ and $\tilde{f}_{+j} / N$, respectively). The inertia of the rows/columns is given by:

$$I_i = \sum_j^{k_2} z_{ij}^2$$

$$I_j = \sum_i^{k_1} z_{ij}^2$$

For supplementary points, the contribution to the inertia of dimensions is zero. The contribution of the active points to the inertia of each dimension is given by

$$\tau_{is} = \frac{\tilde{f}_{i+}}{N} \frac{r_{is}^2}{\lambda_s^{2\alpha}}$$

$$\tau_{js} = \frac{\tilde{f}_{+j}}{N} \frac{c_{js}^2}{\lambda_s^{2\beta}}$$

The contribution of dimensions to the inertia of each point is given by

$$\sigma_{is} = \frac{\tilde{f_{i+}}}{N} \frac{r_{is}^2 \lambda_s^{2-2\alpha}}{I_i}$$

$$\sigma_{js} = \frac{\tilde{f_{+j}}}{N} \frac{c_{js}^2 \lambda_s^{2-2\beta}}{I_j}$$

Confidence Statistics of Singular Values and Scores

The computation of variances and covariances is explained in step 4. Since the row and column scores are linear functions of the singular vectors, an adjustment is necessary depending on the normalization option chosen. From these adjusted standard deviations and correlations are derived in the standard way.

Permutations of the Input Table

For each dimension *s*, let $\rho(i|s)$ be the permutation of the first t_1 integers that would sort the *s*th column of $\{r_{is}\}$ in ascending order. Similarly, let $\rho(j|s)$ be the permutation of the first t_2 integers that would sort the *s*th column of $\{c_{js}\}$ in ascending order. Then the permuted data matrix is given by $\{f_{\rho(i|s),\rho(j|s)}\}$.

References

Benzécri, J. P. 1969. Statistical analysis as a tool to make patterns emerge from data. In: *Methodologies of Pattern Recognition*, S. Watanabe, ed. New York: Academic Press.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. 1975. *Discrete multivariate analysis*: *Theory and practice*. Cambridge, Mass.: MIT Press.

Eckart, C., and Young, G. 1936. The approximation of one matrix by another one of lower rank. *Psychometrika*, 1: 211–218.

Gifi, A. 1981. *Nonlinear multivariate analysis*. Leiden: Department of Data Theory.

Golub, G. H., and Reinsch, C. 1971. Linear algebra, Chapter I.10. In: *Handbook for Automatic Computation*, Volume II, J. H. Wilkinson and C. Reinsch, eds. New York: Springer-Verlag.

Greenacre, M. J. 1984. *Theory and applications of correspondence analysis*. London: Academic Press.

Heiser, W. J. 1981. *Unfolding analysis of proximal data*. Doctoral dissertation. Department of Data Theory, University of Leiden.

Horst, P. 1963. *Matrix algebra for social scientists*. New York: Holt, Rinehart, and Winston.

Israëls, A. 1987. *Eigenvalue techniques for qualitative data*. Leiden: DSWO Press.

Nishisato, S. 1980. *Analysis of categorical data*: *dual scaling and its applications*. Toronto: University of Toronto Press.

Rao, C. R. 1973. *Linear statistical inference and its applications*, 2nd ed. New York: John Wiley & Sons, Inc.

Rao, C. R. 1980. Matrix approximations and reduction of dimensionality in multivariate statistical analysis. In: *Multivariate Analysis*, Vol. 5, P. R. Krishnaiah, ed. Amsterdam: North-Holland.

Wolter, K. M. 1985. *Introduction to variance estimation*. Berlin: Springer-Verlag.