ANACOR

The ANACOR 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).

Notation

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

k_1	Number of rows (row objects)
k_2	Number of columns (column objects)
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 ANACOR 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 ANACOR algorithm can be subdivided into five steps, as explained below.

1. Data scaling and centering

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

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

2. Singular value decomposition

Let the singular value decomposition of **Z** be denoted by

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

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:

$$r_{is} = k_{is} \Big/ \sqrt{f_{i+}/N}\,,$$

$$c_{js} = l_{js} \Big/ \sqrt{f_{+j}/N}\,.$$

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 (1981, 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 $p = \left\{ p_{ij} = f_{ij} / N \right\}$ 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, which implies a compensatory rescaling of the coordinate axes of the row scores and the column scores. The general formula for the weighted sum of squares that results from this rescaling is

row scores: $\sum_i f_{i+} r_{is}^2 = N \lambda_s (1+q)$

column scores: $\sum_j f_{+j} c_{js}^2 = N \lambda_s (1-q)$

The parameter q can be chosen freely or it can be specified according to the following designations:

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

There is a fifth possibility, choosing the designation “principal,” that does not fit into this scheme. It implies that the weighted sum of squares of both sets of scores becomes equal to $N\lambda_s^2$. The estimated variances and covariances are adjusted according to the type of normalization chosen.

Diagnostics

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

Singular Values, Maximum Rank and Inertia

All singular values λ_s defined in step 2 are printeopd up to a maximum of $\min\left\{ (k_1 - 1), (k_2 - 1) \right\}$. 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

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

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 .

Scores and Contributions

This output is given first for rows, then for columns, and always preceded by a column of marginal proportions (f_{i+}/N and f_{+j}/N , respectively). The table of scores is printed in p dimensions. The contribution to the inertia of each dimension is given by

$$\tau_{is} = \frac{f_{i+}}{N} \frac{r_{is}^2}{\lambda_s^2}$$

$$\tau_{js} = \frac{f_{+j}}{N} c_{js}^2$$

The above formula is true only under the row principal normalization option. For the other normalizations, similar relationships are again easily derived from step 5. The contribution of dimensions to the inertia of each point is given by, for $s, t = 1, \ldots, p$,

$$\sigma_{is} = r_{is}^2 \Big/ \sum_t r_{it}^2$$

$$\sigma_{js} = c_{js}^2 \Big/ \sum_t c_{jt}^2$$

Variances and Correlation Matrix 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 variances and covariances the 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 k_1 integers that would sort the sth column of $\{r_{is}\}$ in ascending order. Similarly, let $\rho(j|s)$ be the permutation of the first k_2 integers that would sort the sth column of $\{c_{js}\}$ in ascending order. Then the permuted data matrix is given by $\left\{ f_{\rho(i|s),\rho(j|s)} \right\}$.

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.