Appendix A

The Mathematics of Correspondence Analysis

In the following we give the computational details and rationale of correspondence analysis (CA) of item response data. The formulation is adapted from Greenacre (1993).

Let \( Z \) denote the original data matrix, where the entries \( z_{ij} \) indicate the observed response of subject \( i \) (\( i = 1, \ldots, n \)) to item \( j \) (\( j = 1, \ldots, k \)). The responses are considered measures of association strength between the row entry (here: subject \( i \)) and column entry (here: item \( j \)). The association measure is assumed to be some non-negative quantity, where lack of association (for instance a strongly disagree response to an attitude item with a graded response scale) is indicated by a zero entry.

It is algebraically simpler to work with the so-called correspondence matrix \( P \), with elements \( p_{ij} = z_{ij}/z_{++} \), where the index + indicates the sum over the omitted index. From \( P \) we compute the matrix \( D \), with standardized deviations from independence, \( d_{ij} \), where

\[
d_{ij} = \frac{p_{ij} - p_{i+}p_{+j}}{\sqrt{p_{i+}p_{+j}}}. \tag{A.1}
\]

Note that if the subjects and items are independent (which means that the subjects’ ratings of the various items cannot be explained from their mutual distances on one or only a few latent scales(s)), an element \( p_{ij} \) equals the product \( p_{i+}p_{+j} \). By weighing the deviations from independence with the respective marginals \( p_{i+} \) and \( p_{+j} \) as in (A.1), we obtain the matrix \( D \) of standardized deviations from independence. A rationale for this approach is that the weighing is “variance-standardizing”, which compensates for the larger variance in relatively popular items and the smaller variance in relatively “rare” items. If no such standardization is performed, the differences between larger proportions, would tend to be large compared to the differences between smaller proportions, and hence dominate the solution. The weighing factors are used to equalize these differences.
A. The Mathematics of Correspondence Analysis

For $D$ we compute the singular value decomposition: $D = U\Delta V^T$, where $U$ is the matrix of left singular vectors, with elements $u_{is}, s = 1,\ldots,q$, with $q = \min(n-1,k-1)$; $\Delta$ is a diagonal matrix with positive singular values $l_s$, in descending order along the diagonal; and $V$ is the matrix with right singular vectors, with elements $v_{js}$.

The aim of CA is to find a lower-dimensional representation of $D$. The CA estimated subject location $\hat{\theta}_i$ and estimated item location $\hat{\delta}_{js}$ on dimension $s$ can be expressed as, respectively,

$$\hat{\theta}_i = l_s^{1-a} \cdot u_{is}/\sqrt{p_{i+}},$$  \hspace{1cm} (A.2)

and

$$\hat{\delta}_{js} = l_s^a \cdot v_{js}/\sqrt{p_{+j}}.$$  \hspace{1cm} (A.3)

There are three choices for $a$ in (A.2) and (A.3) in common usage, namely $a = 0, 1, \text{ or } 1/2$ (also referred to as, respectively, row principal, column principal, and symmetrical normalization). With $a=0$ the subject locations $\hat{\theta}_i$ are weighted averages of the sample locations $\hat{\delta}_j$ (which is called by Benzécri, 1973, “le principe barycentrique”), which is the choice of normalization in the current thesis, as it corresponds to the notion of the subject scaling in Thurstone’s (1928) method (where each subject’s scale score is the weighted average of the item scale scores, with the ratings used as weights). In the unfolding literature this representation of subject locations is also referred to as the ideal point representation (cf. Heiser, 1981).

Note that the current thesis focuses on one-dimensional data, in which case only the first left and right singular vectors and the first singular value are used to determine respectively, the subject and item location estimates.

The quality of the lower-dimensional representation of the data is derived from the singular values $l_s$ and is expressed as the percentage of the total inertia that is explained by each dimension. The total inertia of the data table is the chi-square statistic divided by $n$, which can be written as

$$\chi^2/n = \sum_{i=1}^n p_{i+} \sum_{j=1}^k (z_{ij}/z_{i+} - p_{+j})^2/p_{+j}.$$  \hspace{1cm} (A.4)

The total inertia of the data table can be regarded as the weighted average of the squared deviations between the subjects’ profiles (the subjects’ scores proportional
A. The Mathematics of Correspondence Analysis

to their total score) and the average score profile. Hence, it can be thought of the amount of variation among the subjects’ score patterns (See Greenacre, 1993, p. 28-29, for a more thorough explanation of the concept of inertia).

The total inertia of the data table is identical to

\[ \sum_{s=1}^{q} l_s^2, \]

where \(l_s^2\) (which equals the eigenvalue \(\lambda_s\)) is referred to as the principal inertia of dimension \(s\). The percentage of inertia explained by dimension \(s\) is

\[ 100 \times \frac{l_s^2}{\sum_{s=1}^{q} l_s^2}. \]

The contribution of item point \(\hat{\delta}_{js}\) to the inertia of dimension \(s\) is

\[ p_{+j} \hat{\delta}_{js}^2 / l_s^2, \]

or, equivalently, of subject point \(\hat{\theta}_{is}\), the contribution to the inertia of dimension \(s\) is

\[ p_{i+} \hat{\theta}_{is}^2 / l_s^2. \]
References


References

Chernyshenko, O. S., Stark, S., Drasgow, F., & Roberts, B. W. (2007). Constructing personality scales under the assumptions of an ideal-point response pro-
cess: Toward increasing the flexibility of personality measures. *Psychological Assessment, 19*, 88-106.


References


References


References

Measurement, 25, 77-87.
Maraun, M. D., Slaney, K., & Jalava, J. (2005). Dual scaling for the analysis of
factor analysis. Psychometrika, 32, 77-112.
Netherlands: DSWO Press.
Inc.
De Gruyter (Mouton).
Nishisato, S. (1996). Gleaning in the field of dual scaling. Psychometrika, 61,
559-599.
Noel, Y. (1999). Recovering unimodal latent patterns of change by unfolding
analysis: Application to smoking cessation. Psychological Methods, 4, 173-
191.
Noy-Meir, I., & Austin, M. P. (1970). Principal component ordination and simu-
lated vegetational data. Ecology, 51, 551-552.
Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., O’Hara, R. B., Simp-
son, G. L., et al. (2010). Vegan: Community Ecology Package [Com-
puter software manual]. Available from http://CRAN.R-project.org/
package=vegan (R package version 1.17-3)
of the Menninger Clinic, 26, 120-128.
Interrater reliability in unbalanced designs: A comment on Scholte et al.
Polak, M. G., De Rooij, M., & Heiser, W. J. (2010a). Diagnostics for single-
peakedness of item responses with ordered conditional means (OCM).
Manuscript submitted for publication.
Polak, M. G., De Rooij, M., & Heiser, W. J. (2010b). The psychometric evaluation
of bipolar measurement scales: Correspondence analysis as an alternative to
unfolding IRT models). Manuscript submitted for publication.
data: Correspondence analysis as an alternative to principal component
analysis. Computational Statistics and Data Analysis, 53, 3117-3128.


References


Thurstone, L. L., & Chave, E. J. (1929). *The measurement of attitude: A psychophysical method and some experiments with a scale for measuring attitude toward the church*. Chicago: University of Chicago Press.


References


