SPSS®

white paper

Optimal scaling methods for multivariate categorical data analysis

By

Jacqueline J. Meulman, Ph.D., Data Theory Group
Faculty of Social and Behavioral Sciences
Leiden University

Over more than 60 years, classical statistical methods have been adapted in various ways to suit the particular characteristics of social and behavioral science research, with important areas of applications such as market research. Research in these branches often results in data that are non-numerical, with measurements recorded on scales having an uncertain unit of measurement. Data would typically consist of qualitative or categorical variables that describe the units (objects, subjects) in a limited number of categories.

The zero point of these scales is uncertain, the relationships among the different categories is often unknown, and although frequently it can be assumed that the categories are ordered, their mutual distances might still be unknown. The uncertainty in the unit of measurement is not just a matter of measurement error, because its variability may have a systematic component.

An important development in multidimensional data analysis has been the optimal assignment of quantitative values to such qualitative scales. This form of optimal quantification (scaling, scoring) is a general approach to treat multivariate (categorical) data. For example, in the simple linear regression model we wish to predict a response variable **z** from m predictor variables in **X**. This objective is achieved by finding a particular linear combination **Xb** that correlates maximally with **z**. Incorporating optimal scaling amounts to the minimization of $\mathbf{||X^*b - z^*||^{**}2}$ over regression weights **b**, and nonlinear functions $\mathbf{z^* = \theta(Z)}$ and $\mathbf{X_j^* = \varphi_j(X_j), j = 1,..., \mathit{m}}$. Thus, optimal scaling maximizes the correlation between $\theta(z)$ and $\sum_j^m (b_j \varphi_j (X_j))$, over feasible nonlinear functions. These functions are called transformations for quantitative variables, and scalings, scorings or quantifications for categorical variables.

Categorical variables are dealt with in this framework in the following way. A categorical variable $\mathbf{h}_j$ defines a binary indicator matrix $\mathbf{G}_j$ with n rows and l_j columns, where l_j denotes the number of categories. Elements h_{ij} then define elements $g_{ir(j)}$ as follows: $h_{ij} = r \rightarrow g_{ir(j)} = 1$; $h_{ij} \neq r \rightarrow g_{ir(j)} = 0$, where $r = 1,..., l_j$ is the running index indicating a category number in variable j. If category quantifications are denoted by $\mathbf{y}_j$, then a transformed variable can be written as $\mathbf{G_j y_j}$ and, for instance, a weighted sum of predictor variables as $\mathbf{\sum_j^m b_j G_j y_j = X^* b}$, which is as in the standard linear model.

The optimal scaling process turns qualitative variables into quantitative ones. Optimality is a relative notion, because it is always obtained with respect to the particular data. set that is analyzed and the particular criterion that is optimized. Some associated interpretations of these optimality features include discrimination among objects, maximization of homogeneity or internal consistency among variables, making pairwise relationships as linear as possible, maximization of variance accounted for (in the analysis of interdependence), and transformations toward additivity, maximization of r^2, canonical correlation, and the ratio of Between to Total dispersion (in the analysis of dependence). In the optimal scaling process an appropriate quantification level has to be chosen. In addition to a numerical (interval) level, we distinguish between the following levels and scaling methods:

- The ordinal scaling level, taking only rank-orders (among categories) into account, and using least squares monotonic regression or monotonic regression splines (rank = l optimal scaling)
- The nominal scaling level, taking only categorical information into account, and using rank = 1 optimal scaling (as above, either by least squares regression or regression splines) or the centroids approach (rank = p optimal scaling, where p denotes the chosen dimensionality in the solution).

A categorical variable is represented by a set of category points; rank = p optimal scaling locates a category point in the center of gravity (centroid) of the associated objects; rank = l optimal scaling (nominal and ordinal) fits category points on a straight line through the origin.

A short history and a selection of important references

The idea of optimal scaling originates with different sources. Looking first at rank = p optimal scaling, we find the history of the class of techniques that is nowadays usually called (multiple) correspondence analysis (Greenacre, 1984), a literal translation of Berzecri's "analyse des correspondances (multiple)." The class of techniques is also known under the names dual scaling (Nishisato, 1980; 1994), and homogeneity analysis (Gifi, 1981, 1990). Some famous early contributions are by Horst (1935), Fisher (1938,1940), Guttman (1941), Burt (1950), and Hayashi (1952).

Another major impetus to optimal scaling was given by work in the area of nonmetric multidimensional scaling (MDS), pioneered by Shepard (1962), Kruskal (1964) and Guttman (1968). In MDS, a set of proximities between n objects is approximated by a set of distances in some low-dimensional space, usually Euclidean. Optimal scaling of the proximities is typically performed by monotonic (rank =1) regression. Since the breakthrough in MDS in the early 1960s, optimal scaling has subsequently been incorporated in multivariate analysis techniques as well. Some early contributions include Kruskal (1965), Shepard (1966) and Roskam (1968).

In the 1970s and 1980s psychometric contributions to the area became numerous; attempts to systematize resulted in the ALSOS system by Young, De Leeuw and Takane (1976,1978), Young (1981), and the system by the Leiden "Albert Gifi" group. The Albert Gifi (1990) book "Nonlinear Multivariate Analysis" aimed to provide a comprehensive system, combining optimal scaling with multivariate analysis, including statistical developments from the 1970s and 1980s.

Since the middle 1980s, the principles of optimal scaling have gradually appeared in the mainstream statistical literature (Breiman and Friedman, 1985; Gilula and Haberman, 1988; Ramsay, 1989; Buja, 1990; Hastie, Buja, and Tibshirani, 1994). The Gifi system is discussed among the traditional statistical techniques in Krzanowski and Marriott (1994). In the 1990s, optimal scaling methods have been extended into a more general framework.

The data theory scaling system

Since the mission of the Data Theory Scaling System (DTSS) is to meet typical concerns in the social and behavioral sciences, both from a substantive perspective as well as the technical point of view, it has to deal with:

- Discrete multivariate data
- Ordinal data
- Incomplete data
- Nonlinear relationships between pairs of variables
- Non-normal distributions
- Ordering/scaling of response patterns
- Social network data and other proximity relations

DTSS focuses on the multivariate analysis of qualitative or categorical data, including:

- Dimension reduction by linear mapping (see, for example, Heiser and Meulman, 1995); the Gifi system is confined to this aspect of DTSS, hence DTSS can be viewed as a natural successor to the Gifi system, being much more general
- Distance approximation in multivariate data analysis (Meulman, 1992)
- Clustering objects (Heiser, 1993; Heiser and Groenen, 1997)
- Clustering variables (Meulman and Verboon, 1993; Meulman, 1997)
- Graphically displaying objects and variables in linear biplots as in Tucker (1960) and Gabriel (1971), and optimized for least squares multidimensional scaling of multivariate data (Meulman, 1998a)
- Nonlinear biplots as in Gower and Harding (1991), generalized for least squares MDS (Meulman and Heiser, 1993; Groenen and Meulman, 1997)
- Combinatorial data analysis (Hubert, Arabie, Meulman, 1997); for example, optimal sequencing of objects under optimal scaling of the variables (Meulman, Hubert, Arabie, 1997)
- Fitting graphs or networks (Heiser, 1997; Heiser and Meulman, 1997; 1998)
- Implementing procedures in SPSS Categories. SPSS Categories 8.0 includes a new procedure for categorical regression using optimal scaling (see Van der Kooij and Meulman, 1997, also for additional references). The SPSS Categories also includes a new procedure for correspondence analysis, including analysis with supplementary points, restrictions, and alternative standardizations/biplots.

Graphical display

Most DTSS analyses display the objects as points and variables as vectors (arrows) in the same low-dimensional space (biplots); this type of joint display is associated with rank = 1 optimal scaling. A variable, however, can also be viewed as a group of category points, which is associated with rank = p optimal scaling, and this is done in multiple correspondence analysis. The joint display of object points and category points has the following geometrical properties:

- In solutions with a decent fit, objects with similar response profiles are close together in the representation; the "average" object is located near the center of the plot
- Category points for a nominal variable are displayed as centroids of the subjects who share the same category
- The weighted mean squared distance of these category points toward the origin gives the discrimination measure (a measure of variance accounted for, but separate for each dimension due to the rank-p optimal scaling)
- In solutions with a decent fit, categories of different variables that are associated with the same objects are close together in the representation
- For each variable, categories partition the subject points into subclouds; overlapping subclouds correspond to a relatively badly discriminating vanable; well-separated subclouds to a good discriminator

In contrast to the prevailing belief that multiple correspondence analysis is radically different from loglinear analysis because the first would ignore higher-order interactions, Meulman and Heiser (1997) have shown that distances in the graphical display in multiple correspondence analysis are inverse functions of the odd-ratios that express higher-order interactions.

If objects are displayed as points, and variables as vectors, the orthogonal projection of the object points onto the variable vectors (the inner product of the row scores and the column scores) gives an approximation of the columns (and the rows) of the (optimally transformed) data matrix. Because the fit in a joint representation is defined on inner products, one has to make a coherent choice of normalization.

Usually, the object scores are normalized to have means of zero and variance equal to one (the cloud of object points is spherical or orthonormal). Then the coherent normalization identifies scores for the variables as correlations between the variables and the p dimensions of the space fitted to the objects.

If the cloud of the object points has been scaled to be orthonormal, the squared length of the vector representing a variable is proportional to the variance accounted for (fit). If two variables have a decent fit, the angle between their vectors approximates their correlation. When the row scores are normalized, however, one loses the classical scaling distance interpretation with respect to the objects (as in Gower, 1966). To attain the latter, one should rescale the row scores by using the square root of the eigenvalues, and normalize the column scores, keeping the inner product fixed.

When optimal scaling of the variables is included, the categories are located on the vector that represents the variable, and the spacing between the points corresponds to the optimal quantification of the variable. The locations (in a direction in space) are given by coordinates that are sometimes called single category coordinates; such points could also be called markers (Gabriel, 1971; Gower and Hand, 1996).

Some final remarks

If objects (rows) are represented in a principal components analysis (which is infrequently the case, particularly when principal components analysis is needlessly limited to the analysis of a correlation or covariance matrix), it is best done as points, with the variables (columns) as vectors. You should realize, however, that the rows of the data matrix are not necessarily always the subjects (respondents) in the data.

On the contrary, the first applications of Tucker's (1960) vector model (e.g., Carroll, 1972) involved preference data, where a number of subjects express their preference for a number of options. For such data, the appropriate analysis is to put the options as objects in the rows, and the judges (subjects) as variables in the columns.

In contrast to the vector model, the results of a correspondence analysis are usually represented in an alternative way: both row and column objects are represented as points. This way of representation is associated with the so-called unfolding/ideal point model, where relations between row and column entries are represented as distances (the closer a row object to a column object, the larger the association).

There are, however, many complications with the unfolding/ideal point interpretation of correspondence analysis due to the indeterminancy (freedom) in choosing the coherent normalization of row and column scores. This indeterminancy is directly associated with an inner product approximation, and therefore it would be much more appropriate to interpret the correspondence analysis results in terms of the vector model as well (see Meulman, 1998b).

References

Breiman, L., & Friedman, J.H. (1985). "Estimating optimal transformations for multiple regression and correlation." *Journal of the American Statistical Association,* 80, 580-598.

Buja, A. (1990). "Remarks on functional canonical variates, alternating least squares methods and ACE." *Annals of Statistics,* 18, 1032-1069.

Burt, C. (1950). "The factorial analysis of qualitative data." *British Journal of Psychology,* 3,166-185.

Carroll, J. D. (1972): "Individual differences and multidimensional scaling. *Multidimensional scaling: Theory and applications in the behavioral sciences,* R. N. Shepard, A. K. Romney, and S. B. Nerlove (eds.), Vol. 1,105-155, Seminar Press, New York and London.

Fisher, R.A. (1938). "Statistical methods for research workers." Edinburgh: Oliver and Boyd.

Fisher, R.A. (1940). "The precision of discriminant functions." *Annals of Eugenics,* 10, 422-429.

Gabriel, K.R. (1971). "The biplot graphic display of matrices with application to principal components analysis." *Biometrika*, 58, 453-467.

Gifi, A. (1990). "Nonlinear multivariate analysis." Chichester: John Wiley and Sons (First edition 1981, Department of Data Theory, University of Leiden].

Gilula, Z., & Haberman, S.J. (1988). "The analysis of multivariate contingency tables by restricted canonical and restricted association models." *Journal of the American Statistical Association,* 83, 760-771.

Gower, J.C. (1966). "Some distance properties of latent roots and vector methods used in multivariate analysis." *Biometrika,* 53, 325-338.

Gower, J.C., & Harding, S.A (1988). "Nonlinear biplots." *Biometrika,* 75, 445-455.

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

Groenen, P.J.F. & Meulman, J.J. (1997). "Fitting objects and variables simultaneously in smoothed nonlinear biplots by least-squares multidimensional scaling." Paper presented at the 51st 1.5.1. Session, Istanbul.

Guttman, L. (1941). "The quantification of a class of attributes: a theory and method of scale construction." In P. Horst et al. (Eds.), "The prediction of personal adjustment" (pp. 319-348). New York: Social Science Research. Council.

Guttman, L. (1968). "A general nonmetric technique for finding the smallest coordinate space for a configuration of points." *Psychometrika,* 33, 469, 506.Hayashi, C. (1952). "On the prediction of phenomena from qualitative data and the quantification of qualitative data from the mathematico--statistical point of view." *Annals of the Institute of Statistical Mathematics*, 2, 93-96.

Hastie, T., Tibshirani, R. & Buja, A. (1994). "Flexible discriminant analysis." *Journal of the American Statistical Association,* 89, 1255-1270.

Heiser, W.J. (1993). "Clustering in low-dimensional space." In: 0. Opitz, B. Lausen, & R. War (Eds.), Information and classification: Concepts, methods and applications (pp. 145-155). Berlin: Springer-Verlag.

Heiser, W.J. (1997). Fitting graphs and trees. In: C. Hayashi, N. Ohsumi & Y. Baba (Eds.), *Data Science, Classification, and related Methods.* Tokyo: Springer Verlag, (in press).

Heiser, W.J., & Groenen, P.J.F. (1997). "Cluster differences scaling with a within-clusters loss component and a fuzzy successive approximation strategy to avoid local minima." *Psychometrika,* 62, 63-83.

Heiser, W.J., & Meulman, J.J., (1994). "Homogeneity analysis: exploring the distribution of variables and their nonlinear relationships." In: M. Greenacre, & 3. Blasius.(Eds.), *Correspondence Analysis in the Social Sciences: Recent Developments and Applications.* New York: Academic Press, 179-209.

Heiser, W.J. & Meulman, J.J., (1995). "Nonlinear methods for the analysis of homogeneity and heterogeneity." In: W.J. Krzanowski (Ed.), *Recent Advances in Descriptive Multivariate Analysis.* Oxford: Oxford University Press, 51-89.

Heiser, W.J., & Meulman, J.J. (1997). "Representation of binary multivariate data by graph models using the Hamming distance." *Computing Science and Statistics,* 29, 517-525.

Heiser, W.J., & Meulman, J.J. (1998). "A clustering method for a distribution of points on a lattice." In: A. Rizzi and M. Vichi (Eds.), *Book of Short Papers of HFCS98,* Rome, (in press).

Horst, P. (1935). "Measuring complex attitudes." *Journal of Social Psychology,* 6, 369-374.

Hubert, L.J., Arabie, P., & Meulman, J.J. (1997). "Linear and circular unidimensional scaling for symmetric proximity matrices." *British Journal of Mathematical and Statistical Psychology,* 50 (in press).

Kruskal, J.B. (1964). "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis." *Psychometrika,* 29,1-28.

Kruskal, J.B. (1965). "Analysis of factorial experiments by estimating monotone transformations of the data." *Journal of the Royal Statistical Society Series B,* 27, 251-263.Kruskal, J.B. & Shepard, R.N;(1974). "A nonmetric variety of linear factor analysis." *Psychometrika,* 39, 123-157.

Krzanowski W.J. & F.H.C. Marriott (1994), "Multivariate Analysis, Part A Distributions, Ordination and Inference," Edward Arnold, London, 1994.

Meulman, J.J. (1986). "A distance approach to nonlinear multivariate analysis." Leiden: DSWO Press.

Meulman, J.J. (1992). "The integration of multidimensional scaling and multivariate analysis with optimal transformations of the variables." *Psychometrika,* 57, 539-565.

Meulman, J.J. (1997). "Fitting a distance model to homogeneous subsets of variables: Points of view analysis of categorical data." *Journal of Classification,* 13, 249-266.

Meulman, J.J., (1998a). "A distance-based biplot for multidimensional scaling of multivariate data." In: C. Hayashi, N. Ohsumi & Y. Baba (Eds.), *Data Science, Classification, and related Methods* (pp. 506-517). Tokyo: Springer Verlag.

Meulman, J.J. (1998b). "Optimal scaling methods for graphical multivariate data analysis." Paper to be presented at the XIII Symposium on Computational Statistics, Bristol, 1998.

Meulman, J.J., & Heiser, W.J. (1993). "Nonlinear biplots for nonlinear mappings." In: 0. Opitz, B. Lausen, & R. Klar (Eds), *Information and Classification: Concepts, Methods and Applications* (pp. 201-213).
Berlin: Springer Verlag.

Meulman, J.J., & Heiser, W.J., (1997). "Graphical display of interaction in multiway contingency tables by use of homogeneity analysis." In: M. Greenacre, & J. Blasius (Eds.), *Visual Display of Categorical Data.* New York: Academic Press (in press).

Meulman, J.J., Hubert. L.J., & Arabie, P. "Optimal sequencing of objects in multivariate analysis, including optimal scaling of variables." Paper presented at the Fifth Copper Mountain Conference on Iterative Methods, Copper Mountain, Colorado, 1998.

Meulman, J.J., & Verboon, P. (1993). "Points of view analysis revisited: fitting multidimensional structures to optimal distance components with cluster restrictions on the variables." *Psychometrika,* 58, 7, 35.

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

Nishisato, S. (1994). Elements of dual Scaling: An introduction to Practical Data Analysis. Hillsdale, NJ:Lawrence Erlbaum.

Ramsay, J.O. (1989). Monotone regression splines in action. Statistical Science, 4, 425-441.

Roskam. E.E.C.I. (1968). Metric analysis of ordinal data in psychology. Voorschoten: VAM.

Shepard, R.N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function I. Psychometrika, 27, 125-140. II. Psychometrika, 27, 219-246.

Shepard, R.N. (1966). Metric structures in ordinal data. Journal of Mathematical Psychology. 3, 287-315.

Tucker, L.R. (1960) Intra-individual and inter-individual multidimensionality. In: H. Gulliksen & S Messick (Eds.), Psychological Scaling: Theory and Applications. New York: Wiley.

Van der Kooij, A.J., and Meulman J.J., (1997). MURALS: Multiple regression and optimal scaling using alternating least squares. In: E. Faulbaum & W. Bandilla (Eds.), Softstat '97, pp. 99-106. Stuttgart: Lucius & Lucius.

Young, F.W. (1981). Quantitative analysis of qualitative data. Psychometrika, 46, 357-387.

Young, F.W., De Leeuw, J., & Takane, Y. (1976). Regression with qualitative and quantitative variables: An alternating least squares method with optimal scaling features. Psychometrika, 41, 505-528.

Young, F.W., Takane, Y., & De Leeuw, J. (1978). The principal components of mixed measurement level multivariate data: An alternating least squares method with optimal scaling features. Psychometrika, 49, 279-281.

About SPSS

SPSS Inc. is a multinational software company that delivers "Statistical Product and Service Solutions." Offering the world's best-selling desktop software for in-depth statistical analysis and data mining, SPSS also leads the markets for data collection and tabulation.

Customers use SPSS products in corporate, academic and government settings for all types of research and data analysis. The company's primary businesses are: SPSS (for business analysis, including market research and data mining, academic and government research); SPSS Science for scientific research; and SPSS Quality for quality improvement.

Based in Chicago, SPSS has offices, distributors and partners worldwide. Products run on leading computer platforms, and many are translated into Catalan, English, French, German, Italian, Japanese, Korean, Russian, Spanish and traditional Chinese. In 1997, the company employed nearly 800 people worldwide and generated net revenues of approximately $110 million.

Contacting SPSS

To place an order or to get more information, call your nearest SPSS office or visit our World Wide Web site at **www.spss.com**

SPSS Inc.		+1.312.651.3000
	Toll-free:	+1.800.543.2185
SPSS Argentina srl.		+541.814.5030
SPSS Asia Pacific Pte. Ltd.		+65.245.9110
SPSS Australasia Pty. Ltd.		+61.2.9954.5660
	Toll-free:	+1800.024.836
SPSS Belgium		+32.162.389.82
SPSS Benelux		+31.183.636711
SPSS Central and Eastern Europe		+44.(0)1483.719200
SPSS Czech Republic		+420.2.24813839
SPSS East Mediterranea & Africa		+972.9.9526701
SPSS Federal Systems		+1.703.527.6777
SPSS Finland Oy		+358.9.524.801
SPSS France SARL		+33.1.5535.2700
SPSS Germany		+49.89.4890740
SPSS Hellas SA		+30.1.7251925
SPSS Hispanoportuguesa S.L.		+34.91.447.37.00
SPSS Hong Kong Ltd.		+852.2.811.9662
SPSS Ireland		+353.1.496.9007
SPSS Israel Ltd.		+972.9.9526700
SPSS Italia srl		+39.51.252573
SPSS Japan Inc.		+81.3.5466.5511
SPSS Kenya Ltd.		+254.2.577.262
SPSS Korea KIC Co., Ltd.		+82.2.3446.7651
SPSS Latin America		+1.312.494.3226
SPSS Malaysia Sdn Bhd		+60.3.704.5877
SPSS Mexico Sa de CV		+52.5.682.87.68
SPSS Middle East & South Asia		+91.80.545.0582
SPSS Polska		+48.12.6369680
SPSS Russia		+7.095.125.0069
SPSS Scandinavia AB		+46.8.506.105.50
SPSS Schweiz AG		+41.1.266.90.30
SPSS Singapore Pte. Ltd.		+65.533.3190
SPSS South Africa		+27.11.706.7015
SPSS Taiwan		+886.2.25771100
SPSS UK Ltd.		+44.1483.719200