NONMETRIC MULTIDIMENSIONAL SCALING

JACQUELINE J MEULMAN

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*, 1–27; Nonmetric multidimensional scaling: A numerical method. *29*, 115–129. (5773 citations in Google Scholar as of 4/1/2016)

Although the assignment was to write a note about the famous, highly cited Kruskal 1964 papers, it would hardly be fair if the topic wasn't described in the context of two other papers, being Shepard's 1962 papers (with 2309 citations in Google Scholar as of 4/1/2016) that started the development of what is called *nonmetric multidimensional scaling.* Before getting into more detail, some of the earlier history of multidimensional scaling should be addressed as well, to place the breakthrough from the early 1960s in a proper context.

Multidimensional scaling (MDS) is the generic name of a class of techniques that aim to find a low-dimensional space, in which the distances between N points resemble as closely as possible a given or derived set of dissimilarities between N objects. Multidimensional scaling methods can be divided into two major approaches. Prior to the Shepard-Kruskal approach, MDS was mostly associated with Torgerson (1958). As Shepard wrote in his 1980 paper in Science:

> Proposals that stimuli be modeled by points in a space in such a way that perceived similarity is represented by spatial proximity go back to the suggestions of Isaac Newton....However, little progress was made toward the development of data-analytic methods for the construction of such spatial representations on the basis of psychological data until the efforts of a group of psychometricians, beginning in the late 1930's at Chicago and subsequently moving to Princeton, culminated in the 1952 development by Torgerson of the first fully workable method of metric multidimensional scaling.

The prime incentive of the Shepard-Kruskal approach was the notion of nonmetricity. The interest in nonmetricity was invoked by considerations from a number of perspectives, the most important one it being viewed as a property that could be attributed to data, especially psychological data like preference rankorders. The most important aspect of the so-called nonmetric breakthrough is usually recognized as the accomplishment of obtaining a *metric* representation, like a configuration of points, from essentially *nonmetric* information, like rankorders for pairs of objects. Shepard describes his work as the merging of what used to be two phases (i.e. first obtaining tentative distances from empirically given data by some predetermined distance function, and then applying classical MDS to obtain a configuration of points) into one phase. The innovating step was to state that no assumptions were made about the form of the distance functions, except that its values had to be monotonic with the data. Shepard observed that forces tending to flatten the configuration were counteracted by forces tending to maintain monotonicity. He proposed a procedure that seeks a low-dimensional configuration while working in an $N-1$ dimensional space, where monotonicity is exactly realized. In the final low-dimensional space, however, departure from monotonicity could arise. To implement his ideas, Shepard developed a FORTRAN program to carry out an iterative process of embedding N objects into

This note was written while the author was Visiting Professor at the Department of Statistics, Stanford University.

a space of minimum dimensionality that would have a satisfactory rank-order correspondence between the proximities and the induced Euclidean distances. From the Shepard (1980) Science paper:

> At the Bell Telephone Laboratories, I began in 1960 to explore a new approach to multidimensional scaling, called analysis of proximities, that proved capable of overcoming the limitations of the earlier approaches. I used a one-stage iterative method (i) to adjust the positions of points in a space until the rank order of the interpoint distances was as nearly as possible the inverse of the rank order of the corresponding similarities, and (ii) to find the space of the smallest number of dimensions for which the residual departure from a perfect inverse ranking was acceptably small... Following a few adjustments, on 17 March 1961 the iterative process with which I had been experimenting finally converged to its first stationary configuration (at just 2.33 p.m. EST, according to the computer log). From then on, results of surprising precision were regularly obtained.

The fact that departure from monotonicity did not play a central part in Shepard's procedure, inspired Joe Kruskal, his colleague from Bell Labs, to display MDS as a problem of statistical fitting, where dissimilarities are given and a configuration of points whose distances fit them best is aimed at, which is described in full detail in the two highly-cited 1964 *Psychometrika* papers. The genius of the two Kruskal papers is that they took Shepard's intuitive ideas and turned them into a reliable means of carrying out a nonmetric multidimensional scaling. Kruskal began by defining an explicit least squares loss function to minimize, called *stress*, proposed a gradient method to obtain successor configurations, and introduced monotone regression of distances upon dissimilarities to obtain perfect monotonicity.

As is often the case in major scientific developments, similar objectives were pursued elsewhere, most notably by Guttman. Guttman's important 1968 paper postulates the rank-image principle as the fundamental mechanism guarding nonmetricity, while at the same time introducing his correction-matrix method to obtain successor configurations. At the risk of oversimplifying, disregarding the very interesting discussion on the differences between the Kruskal and the Guttman approach, we could say that Guttman's nonmetric transformation involves the permutation of the distances such that they become monotonic with the dissimilarities, while Kruskal's transformation involves a monotone regression of the distances upon the dissimilarities.

The major contribution by Kruskal and Shepard to the development of MDS, which turned the technique into the elegant data analysis method as it is appreciated nowadays, has been primarily recognized as transforming multidimensional scaling into a nonmetric method. What is often not greatly acknowledged, is that both Kruskal and Guttman, proposed at the same time an alternative procedure for the *metric* multidimensional scaling problem, until then solved by the classical, metric, procedure mentioned in the beginning. In Classical MDS, the approximation of dissimilarities by distances is channeled through scalar products, and involves an eigenvalue decomposition of a matrix $\mathbf{S} \equiv -\frac{1}{2}\mathbf{J}\boldsymbol{\Delta}^2\mathbf{J}$. Here $\boldsymbol{\Delta} = \{\delta_{ij}\}$ is a matrix with given or derived dissimilarities δ_{ij} between N objects, and $\mathbf{J}$ is a centering operator that renders the resulting matrix $\mathbf{S}$ double-centered. From the eigenvalue decomposition of $\mathbf{S}$, with eigenvalues ordered from big to small, a low-dimensional configuration $\mathbf{X}$ is derived by taking the first p eigenvectors and the square root of the first p eigenvalues, with $\mathbf{S}$ approximated being by the scalar product matrix $\mathbf{X}\mathbf{X}'$. This procedure is also known als the Young-Householder (1938) process. Since the method was independently from Torgerson proposed by Gower (1966), Heiser came up with the acronym `YoHoToGo` scaling. An important property of the scalar product approximation approach is that the resulting distances $d_{ij}(\mathbf{X})$ will always be smaller than their associated dissimilarities δ_{ij}.

By contrast, instead of using scalar products, Kruskal introduced a least squares loss function defined on dissimilarities and distances directly; it can be written (in its raw, squared, form) as

$$STRESS = \min_{\delta_{ij} \epsilon \mathbf{\Gamma}} \sum_{i=1}^{N} \sum_{j=1}^{N} (\delta_{ij} - d_{ij}(\mathbf{X}))^2$$

where $\mathbf{\Gamma}$ denotes the set of all monotone transformations of the original dissimilarites $\{\delta_{ij}\}$. Finding the optimal transformations, usually called disparities and indicated by $\hat{d_{ij}}$, is done by *inner minimization* while the *outer minimization* solves the metric scaling problem by minimization over $\mathbf{X}$ for given dissimilarites δ_{ij}. The first Kruskal `FORTRAN` program was called `MDSCAL`, However, as Shepard (1980) writes:

> A variety of computer programs of this general type have subsequently been developed... Perhaps the currently most versatile such program is the one, available from the Bell Telephone Laboratories, named `KYST` (after Kruskal, Young, Shepard, and Torgerson, from whose earlier programs it derives).

What is also noteworthy, is that the nonmetric breakthrough inspired a whole succession of programs that applied the same idea of monotonic transformation to the variables in multivariate data (instead of to the dissimilarities in proximity data). According to Doug Carroll, at Shepard's suggestion, Kruskal introduced his `MFIT` monotone regression algorithm into the first procedure fitting a two-way ANOVA model with no interaction terms but with optimal monotonic transformation of the dependent variable (Kruskal 1965). This application was followed by Shepard (1966), Shepard & Kruskal (1966), and the Alternating Least Squares-Optimal Scaling system by Young, De Leeuw and Takane (1980). Independently, and outside the psychometric literature, this idea was applied in the Alternating Conditional Expectations algorithm (ACE) by Breiman and Friedman (1985). Last, but not least, we have to mention the Gifi system, developed by "the Dutch school of nonlinear multivariate analysis" (Buja, 1990), and the very successful development of software for nonlinear data analysis in the IBM/SPSS package `CATEGORIES`. The latter package also includes very elaborate extensions of the original Kruskal et al. ideas for nonmetric multidimensional scaling and unfolding, which shows that the influence of the 1960s nonmetric breakthrough still carries on.

REFERENCES

Breiman, L., and Friedman, J.H. (1985). Estimating optimal transformations for multiple regression and correlation (with discussion). *Journal of the American Statistical Association*, *80*, 580–619.

Buja (1990). Remarks on Functional Canonical Variates, Alternating Least Squares Methods and ACE.*Annals of Statistics*, *18*, 1032–1069.

Gifi, A. (1990). *Nonlinear multivariate analysis.* Chichester: Wiley.

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

Guttman, L. (1968). A general nonmetirc technique for finding the smalles coordinate space for a configuration of points. *Psychometrika*, *33*, 469–506.

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

Kruskal, J.B., and Shepard, R.N. (1978). A nonmetric variety of linear factor analysis. *Psychometrika*, *39*, 123–157.

Torgerson, W. S. (1952). Multidimensional scaling: 1. Theory and method. *Psychometrika*, *17*, 401–19.

Torgerson, W.S. (1958). *Theory and methods of scaling.* New York: Wiley.

Young,. G. and Householder, A.S. (1938). Discussion of a set of points in terms of their mutual distances. *Psychometrika*, *3*, 19–22.

Young, F.W., De Leeuw, J. and Takane, Y. (1980). Quantifying qualitative data. In: E.D. Lantermann and H. Feger (Eds.), *Similarity and Choice.* Bern: Huber.

FIGURE 1. An original KYST tape