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)

The now familiar psychometric topic of nonmetric multidimensional scaling (NMDS) was introduced into the quantitative psychology literature by the publication of two highly-cited companion papers from Roger Shepard:

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. II. *Psychometrika*, *27*, 125–140; 219—246. (2309 citations in Google Scholar as of 4/1/2016)

These two papers discuss the task of embedding a set of n objects, say, $\{O_1, \ldots, O_n\}$, into a Euclidean space of K dimensions, based on a given numerical proximity measure, p_{ij}, defined for each object pair, O_i and O_j. It will be assumed for convenience that these proximities are keyed as dissimilarities so that larger values for p_{ij} reflect more dissimilar objects. The embedding was to be done in such a way that: (1) the rank-ordering of the proximities would reflect as closely as possible the rank-ordering of the induced (Euclidean) distances in the K-dimensional space, and (2), the number of dimensions used (that is, the value of K), would be as small as possible. The term "nonmetric" arises from a reliance on only the rank-ordering of the proximities and not on their actual ("metric") numerical values.

To implement his ideas for nonmetric multidimensional scaling, Shepard developed a FORTRAN program to carry out an iterative process of embedding n objects into a space of minimum dimensionality that would have a satisfactory rank-order correspondence between the proximities and the induced Euclidean distances. In Shepard's hands, some dramatic data analysis examples were given in the *Psychometrika* paper labeled as II (for example, the famous two-dimensional color circle illustration based on data from Ekman). What was less clear from these two Shepard papers was whether a naive practitioner could get the same type of results using Shepard's program, which apparently was guided by some trial-and-error adjustments during the iterative process. This is the point where Roger Shepard's colleague from Bell Labs, Joe Kruskal, enters the picture, and which led to the two highly-cited

Psychometrika papers that are the main purpose of this current note. In comments that Shepard made on the occasion of his two *Psychometika* papers being named "Citation Classics" (May 31, 1979), Kruskal's crucial role is emphasized in making nonmetric multidimensional scaling a widely-used and viable technique:

> After a period of trial-and-error adjustment of the parameters of the iterative process, success came with dramatic suddenness on March 17, 1961. According to the computer log, it was at precisely 2:33 p.m. EST on that day that the iterative process first converged to a stationary configuration, revealing a remarkably exact recovery of an underlying test configuration. The excitement of that moment was rivaled only by the birth of my daughter on the very next day. Since then my daughter has developed into a fine young woman; and, thanks in part to the subsequent contributions of my mathematical colleague Joseph Kruskal, nonmetric multidimensional scaling is now finding wide application throughout the cognitive, behavioral, and biomedical sciences.

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 loss function to minimize, called stress; he then proposed a numerical strategy for its optimization through the "method of steepest descent." To effect a sole reliance on the rank-order of the proximities, a number of "monotone regression" steps, based on a separately important "pool adjacent violators algorithm" (PAVA), were interspersed within the steepest descent mechanism.

To define the stress of a given fixed configuration, suppose the n objects are embedded in a K-dimensional space with an interpoint distance, d_{ij}, between objects O_i and O_j. Stress, S, is then given by

$$S = \sqrt{\frac{\sum_{i<j}(d_{ij} - \widehat{d}_{ij})^2}{\sum_{i<j} d_{ij}^2}} \;;$$

here, the $\widehat{d}_{ij}$ are called *disparities* and are those numbers that minimize S subject to the constraint that the $\widehat{d}_{ij}$ have the same rank-order as the p_{ij}: $\widehat{d}_{ij} \leq \widehat{d}_{i'j'}$ whenever $p_{ij} < p_{i'j'}$. This later minimization of S is the monotone regression step carried out using PAVA. At this point, the iterative method of steepest descent is used to improve S, which then leads to another monotone regression step, and so on until convergence (that is, until S can no longer be improved).

A prime contribution of the Kruskal papers is that they came with a practical way of carrying out a nonmetric multidimensional scaling through an available FORTRAN program, `mdscal.f`. This specific program as well as its various Bell Labs successors, such as KYST2a, are still available on various `netlib.org` web sites. An additional advantage of using one of these routines was the availability of various built-in graphics and plots, such as what has become known as the "Shepard diagram." The latter is a scatterplot of final interpoint distances on the y-axis and dissimilarities on the x-axis. The disparities, $\hat{d_{ij}}$, are plotted as "predicted values" from the monotone regression, tracing out the monotone regression line that either stays at the same horizontal level or increases when moving toward that more positive dissimilarity values along the x-axis. The degree to which the values cluster compactly around the monotone regression line, gives an indication of how well the multidimensional scaling reflects the original dissimilarities.

As we have noted, the main contribution of the two highly-cited Kruskal papers (second only in *Psychometrika* to the massively cited coefficient alpha paper of Cronbach (1951) and the heavily miscited paper of Kaiser (1974)), was to make NMDS viable by proposing an explicit loss function plus a means for its optimization that included an interweaving of the monotone regression steps. This Kruskal codification of NMDS is well-noted by Shepard in his 1974 Psychometric Society Presidential Address published in *Psychometrika*: "Representation of structure in similarity data: Problems and prospects" (*39*, p. 376):

> Although my original method generally yielded spatial configurations that appeared indistinguishable from those furnished by subsequent methods, my mathematical colleague Joseph Kruskal soon noted that the precise measure of departure from monotonicity that was being minimized by the method was neither explicitly defined nor even known to exist in an explicitly definable form. Thus, despite the intuitive plausibility and practical success of the method, it lacked the conceptual advantage of a strict mathematical specification of exactly what problem was being solved. Moreover, in the absence of an explicitly defined loss function, general techniques for the minimization of such functions (notably, gradient methods) were not apparently applicable. The iterative method that was used consequently appeared somewhat *ad hoc*.

In closing, we might raise a few side issues with the two seminal Kruskal papers, some of which have been addressed subsequently in the literature:

1) Kruskal proposed a verbal scale for evaluating the numerical value of stress that should not be used to assess how good or bad a particular scaling might be. For example, .20 was labeled as "poor," .10 was labeled "fair," and so on. What is most relevant for studying the adequacy of a spatial representation, is substantive interpretability, and that is never reducible to a single number.

2) The local minimum issue for the method of steepest descent is addressed by Kruskal but not in the sufficient depth the problem deserves. For one dimension, in particular (that is, for $K = 1$), and for the city-block distance function more generally, local minima are such a major problem that alternative combinatorial optimization strategies may be the preferred approaches.

3) Degenerate solutions in NMDS are not that uncommon, where stress is reduced to zero but the resulting configurations are confined to just several points. These noninformative solutions occur whenever monotone regression is used and there is a group structure in the dissimilarity data; that is, when all within group dissimilarities are smaller than those between groups. Provision should be made to use some alternative to monotone regression, such as monotone regression splines or convex/concave regression.

4) In choosing the number of dimensions for a spatial representation (that is, in choosing K), Kruskal suggests the use of a "scree plot" based on stress and looking for the proverbial "elbow." This rarely works in any context where one is forced to identify an "elbow." A better strategy may be to restrict K to be less than or equal to three, thus allowing only the types of realizable spatial representations that the method was designed to provide.