Journal of Classification 19:7-43 (2002)
DOI: 10.1007/s00357-001-0031-0

SPECIAL INVITED ARTICLE

# Visualization Methodology for Multidimensional Scaling

Andreas Buja

University of Pennsylvania

Deborah F. Swayne

AT&T Labs

**Abstract:** We discuss the application of interactive visualization techniques to multidimensional scaling (MDS). MDS in its conventional batch implementations is prone to uncertainties with regard to (a) local minima in the underlying optimization, (b) sensitivity to the choice of the optimization criterion, (c) artifacts in point configurations, and (d) local inadequacy of the point configurations.

These uncertainties will be addressed by the following interactive techniques: (a) algorithm animation, random restarts, and manual editing of configurations, (b) interactive control over parameters that determine the criterion and its minimization, (c) diagnostics for pinning down artifactual point configurations, and (d) restricting MDS to subsets of objects and subsets of pairs of objects.

A system, called "XGvis", which implements these techniques, is freely available with the "XGobi" distribution. XGobi is a multivariate data visualization system that is used here for visualizing point configurations.

**Keywords:** Proximity data; Multivariate analysis; Data visualization; Interactive graphics.

We wish to thank Jon Kettenring and Daryl Pregibon, the two managers who made this work possible. We are indebted to Brian Ripley for his port of the software to Windows. We finally thank the editor, Phipps Arabie, for many suggestions and for improving the scholarship of the paper.

Authors' Addresses: Andreas Buja, Statistics Department, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104-6302, USA; http://www-stat.wharton.upenn.edu/~buja/; this work was done while the author was at AT&T Labs; Deborah Swayne, Senior Technical Staff Member, AT&T Labs, 180 Park Ave, P.O. Box 971, Florham Park, NJ 07932-0971, USA; e-mail: dfs@research.att.com; http://www.research.att.com/~dfs/.

## 1. Introduction

We describe methodology for multidimensional scaling based on interactive data visualization. This methodology was enabled by software in which MDS is integrated in a multivariate data visualization system. The software, called "XGvis", is described in a companion paper (Buja, Swayne, Littman, Dean and Hofmann 2001), that lays out the implemented functionality in some detail; in the current paper we focus on the use of this functionality in the analysis of proximity data. We therefore do not dwell on the mechanics of creating certain plots; instead we deal with problems that arise in the practice of proximity analysis: issues relating to the problem of multiple local minima in MDS optimization, to the detection and interpretation of artifacts, and to the examination of local structure.

The paper is organized as follows: Section 2 introduces the famous Rothkopf (1957) Morse code data and gives a detailed analysis that illustrates the reach of data visualization and direct manipulation through graphical interaction. Section 3 discusses the advantages of visual stopping of MDS optimization. Section 4 illustrates the problem of multiple local minima and shows ways to diagnose its nature and severity. Section 5 explains the fundamental problem of indifferentiation, that is, the tendency of proximity data to assign too similar distances to too many pairs of objects. Sections 6 and 7 demonstrate two ways of uncovering local structure: within-groups MDS, and MDS with truncated or down-weighted dissimilarities. The final Section 8 introduces a novel use of non-Euclidean Minkowski metrics for the rotation of configurations.

Multidimensional scaling is the subject of several books, among them a recent one by Borg and Groenen (1997) and an older one by Kruskal and Wish (1978). The latter is concise and gives sufficient background for this article. For the advanced reader there exist overview articles by, for example, Carroll and Arabie (1980, 1998) and Carroll and Green (1997). The collection edited by Davies and Coxon (1982) contains some of the seminal articles in the field, including Kruskal's (1964a,1964b), and so does the overview by Green, Carmone, and Smith (1989). An older book chapter we found still useful is Greenacre and Underhill (1982). Many books on multivariate analysis include chapters on multidimensional scaling, such as Gnanadesikan (1997) and Seber (1984).

## 2. The Rothkopf Morse Code Data

To illustrate the techniques described in this paper, we use the classic Rothkopf (1957) Morse code data as our running example. While these data may seem stale to those who are familiar with some of the MDS literature, there is merit in using a well-known dataset exactly because of the fact that so many prior analyses have appeared in print. This fact offers comparisons and it avoids distractions from the main point of the paper, which is methodology.

The Rothkopf Morse code data originated in an experiment where pairs of Morse codes were shown to subjects who had to decide whether the two codes in a pair were identical. The resulting data were summarized in a table of confusion rates.

To apply MDS, we first symmetrized the data ($s_{i,j}$) in the simplest possible way: $s_{i,j} \rightarrow s_{i,j} + s_{j,i}$ (see Arabie and Soli (1982) for a discussion of alternative symmetrization formulae). We then converted the symmetrized data to dissimilarities ($d_{i,j}$) using the formula

$$d_{i,j}^2 = s_{i,i} + s_{j,j} - 2s_{i,j} \ .$$

In principle any monotone descending transformation could be used for conversion, but we approached the confusion rates using an inner product model, $s_{i,j} \approx < \mathbf{x}_i, \mathbf{x}_j >$, which suggests the above conversion formula by mimicking the identity $\|\mathbf{x}_i - \mathbf{x}_j\|^2 = \|\mathbf{x}_i\|^2 + \|\mathbf{x}_j\|^2 - 2 < \mathbf{x}_i, \mathbf{x}_j >$.

Some properties of the resulting dissimilarities are the following:

(1) For the Morse code data, all dissimilarities are well-defined because of $s_{i,i} + s_{j,j} - 2s_{i,j} \geq 0$, which follows from the diagonal dominance of the symmetrized confusion matrix $\mathbf{S}$. (Nonnegativity is not guaranteed by the formula, however. Even the Morse code data have a close call: before symmetrization there exists an off-diagonal value that is larger than the smallest diagonal value.)

(2) The similarities of codes with themselves ($s_{i,i}$) are not ignored.

(3) Dissimilarities of codes with themselves are zero: $d_{i,i} = 0$.

(4) Classical MDS of the dissimilarities $d_{i,j}$ amounts to an eigenanalysis of a doubly-centered version of the symmetrized matrix $\mathbf{S}$.

After subjecting the resulting dissimilarity matrix to nonmetric Kruskal-Shepard scaling using Kruskal's stress formula 1 in two, three and four dimensions, we obtained the configurations shown in Figure 1. We interactively decorated the configurations with labels and lines to aid interpretation [one of the benefits of a visualization system; Swayne, Cook, and Buja (1998)]. In particular, we connected groups of codes of the same length, except for codes of length four which we broke up into three groups and a singleton. In the 2-D solution, one observes that the *code length* increases left to right, and (with the exception of the codes of length one) the *fraction of dots* increases from the bottom up, in agreement with published accounts, for example, in Shepard (1962, 1963), Kruskal and Wish (1978, p. 13), and Borg and Groenen (1997, p. 59). The 2-D

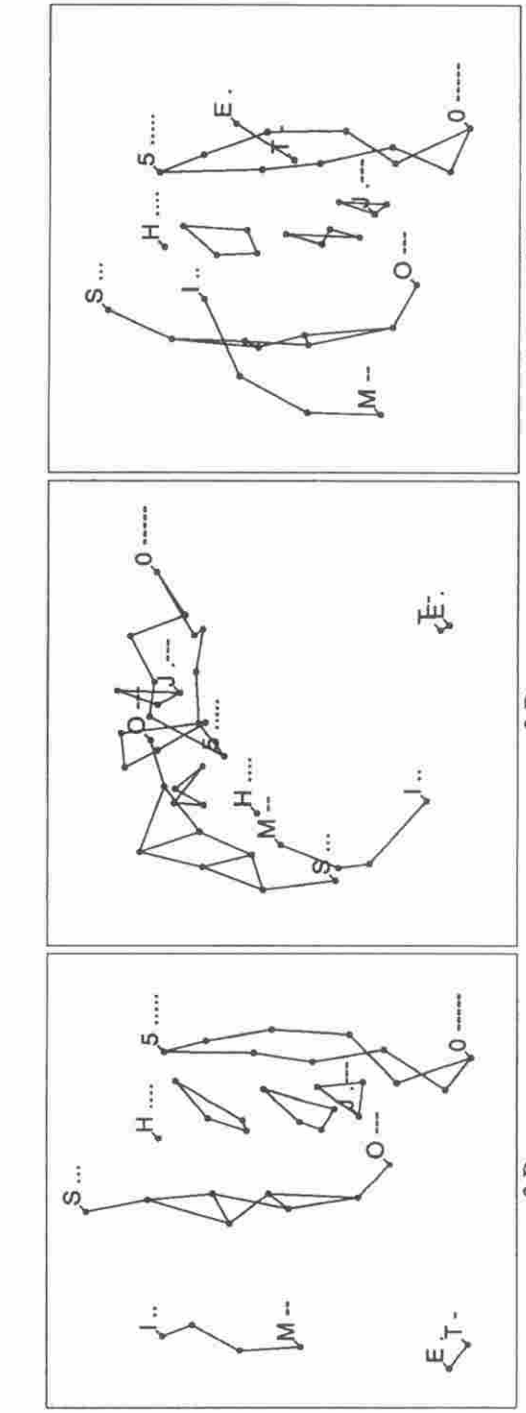


Figure 1. Views of converged nonmetric scaling solutions in 2-D and 3-D. The stress values are 0.1874 and 0.1254, respectively. Two projections of a 3-D configuration are shown. The 3-D projections were obtained with 3-D rotations. (Figure 1 continued on next page.)

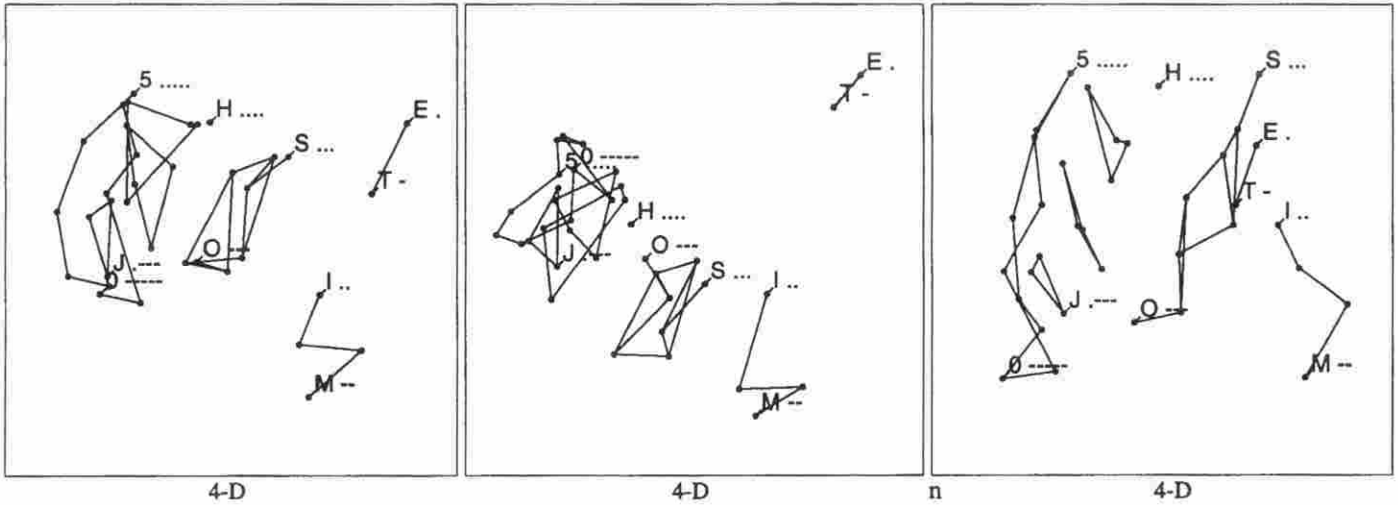


Figure 1. (continued) Views of converged nonmetric scaling solutions in 4-D. The stress value is 0.0974. Three projections of a 4-D configuration are shown. The 4-D projections were obtained with so-called grand tours and manual tours.

plot is of course rotation invariant, and it has been rotated to align *code length* with the horizontal axis and *fraction of dots* with the vertical axis. Expressions such as "left to right" and "bottom up" have to be interpreted accordingly with regard to a desirable rotation.

As a first application of visualization methodology to MDS, we examine the 3-D and 4-D solutions. The methods we use are 3-D rotations and their generalizations to higher dimensions, grand tours and manual tours [implemented in XGobi: Swayne et al. (1998); Cook and Buja (1997); Buja, Cook, and Swayne (1996)]. Making the usual caveats that the insights gained by viewing dynamic rotations and tours cannot be captured in a series of still pictures, we report what we were able to see:

(1) The 3-D solution is only seemingly more complex than the 2-D solution. Roughly speaking, the 3-D solution is the 2-D solution wrapped around the surface of an approximate sphere, with the difference that the codes of length one, "E" and "T", are further removed from the codes of length two and higher. This is the main insight: the 2-D solution has the defect that it has no good place for the codes of length one. The true distinctness of the shortest codes cannot be properly reflected in 2-D, but it can in 3-D. Thus, the additional dimension did not reveal a new dimension in the usual sense; it revealed an odd subset that should be separated from the rest by a dummy variable. Below we will also show that the pair {E,T} is extremely influential in the following sense: it inhibits an additional dimension inherent in the longer codes.

(2) The 4-D solution, when viewed in a grand tour, reveals a rigidity of the codes of length three, four, and five in their positions relative to each other. They form three roughly parallel sheets with low and high fractions of dots aligned across the sheets. The codes of length two form a line that tries to align itself with the longer siblings, but it seems to suffer from a strong attraction by the codes of length one.

This last finding suggests a simple diagnostic: remove the codes of lengths one and two, and analyze the longer codes separately. The result is in Figure 2 where we show two views of a nonmetric 3-D solution. The configuration was interactively rotated for optimal interpretation. The views share the vertical axis in 3-space, while the horizontal axes are orthogonal to each other. Here are the findings:

(3) The left view shows the layers of codes of constant length, as well as the matching trends from low to high fractions of dots within the layers. We note that the layers lean to the left, suggesting that *code length* and *fraction of dots* are slightly confounded. If the axis for *code length* is horizontal from left to right, then the axis for *fraction of dots* runs roughly

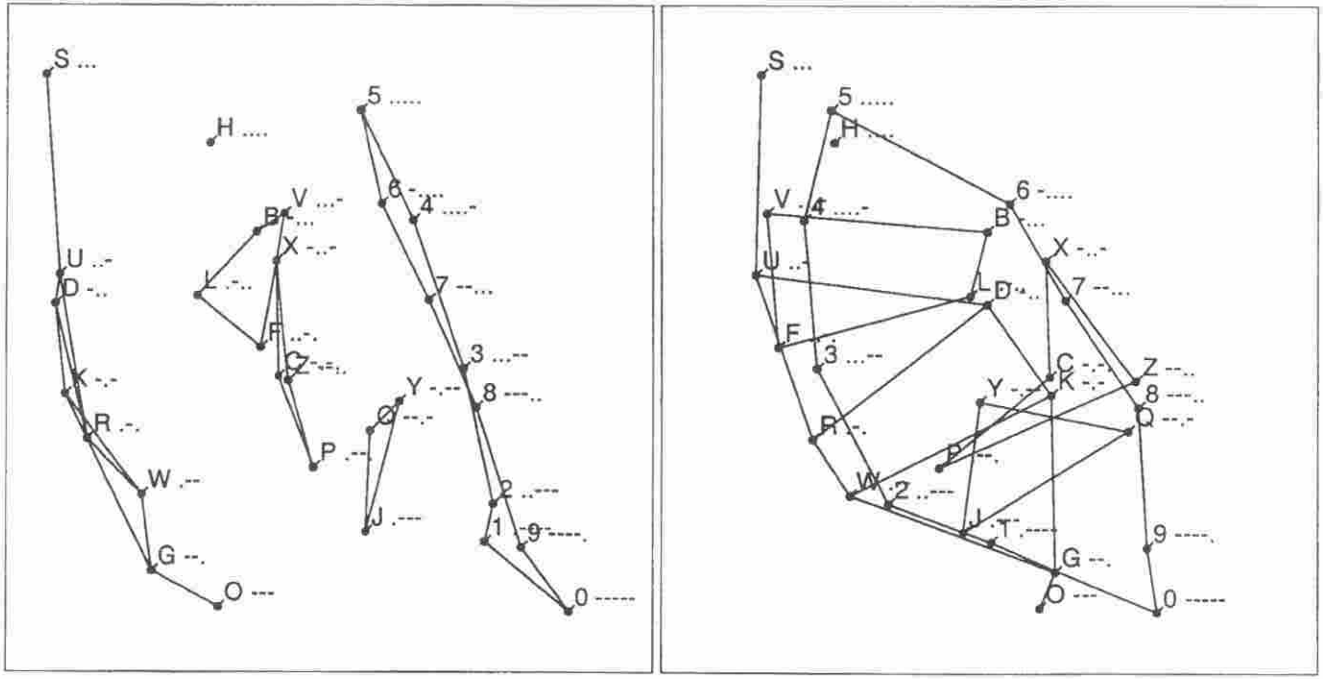


Figure 2. Views of a nonmetric 3-D solution of the subset of Morse codes of lengths three, four, and five. The stress is 0.1170.

from south-southeast to north-northwest. There is some intuitive meaning in this type of confounding according to physical duration of a code: long codes that have many dots are more often confused with shorter codes that have many dashes, than vice versa; for example, “5 = ·····” and “O = −−−” are more often confused than “S = ···” and “0 = −−−−−−”. One could therefore interpret the horizontal axis as physical duration and the strictly vertical axis as fraction of dashes. As a consequence, the duration of “5 = ·····” would be about the same as that of “J = · −−−” because they have about the same horizontal position.

(4) The right view of Figure 2 can be interpreted as follows: the codes fall into two subsets, one corresponding to the arc that runs from the left side to the top, the other subset to the arc that runs from the right side to the bottom. The two arcs differ in one aspect: codes in the upper left all start with a dot, the codes in the lower right all start with a dash. Therefore, the direction from the bottom right to the top left corresponds to a dimension that reflects the exposed initial position of the codes: initial dots and dashes correspond to a separate dimension. The fact that this dimension runs in the descending diagonal direction shows that it is slightly confounded both with fraction of dots (an initial dot contributes to the fraction of dots) and duration (an initial dot contributes to a shorter physical duration).

In summary, we have found four dimensions in the Morse code dissimilarities: (a) *code length*, (b) *fraction of dots*, (c) a dummy for the codes of *length one*, and (d) a dummy for *initial exposure position* for the long codes. A methodological message from this exercise is that dimensions can be local. Insisting on global dimensions for all objects may obscure the presence of local dimensions in meaningful subsets.

To close this section, we consider a still smaller subset: the codes of length five, representing the digits “0”,...,”9”. These codes have an obvious circular structure:

$$\begin{array}{cccccccc}
0 & = & - & - & - & - & - \\
1 & = & \cdot & - & - & - & - \\
2 & = & \cdot & \cdot & - & - & - \\
3 & = & \cdot & \cdot & \cdot & - & - \\
4 & = & \cdot & \cdot & \cdot & \cdot & - \\
5 & = & \cdot & \cdot & \cdot & \cdot & \cdot \\
6 & = & - & \cdot & \cdot & \cdot & \cdot \\
7 & = & - & - & \cdot & \cdot & \cdot \\
8 & = & - & - & - & \cdot & \cdot \\
9 & = & - & - & - & - & \cdot
\end{array}$$

This structure is reflected in a loop-shaped arrangement of MDS configurations, as shown in Figure 3. Of the two configurations in the figure, the metric version appears cleaner than the nonmetric version. This result should not be a surprise as the isotonic transformation of nonmetric scaling becomes tenuous to estimate for small numbers of objects. [For an approach to the Morse code digits that imposes circularity as a model, see Hubert, Arabie, and Meulman (1997).]

## 3. Visual Checks of Convergence of Optimization

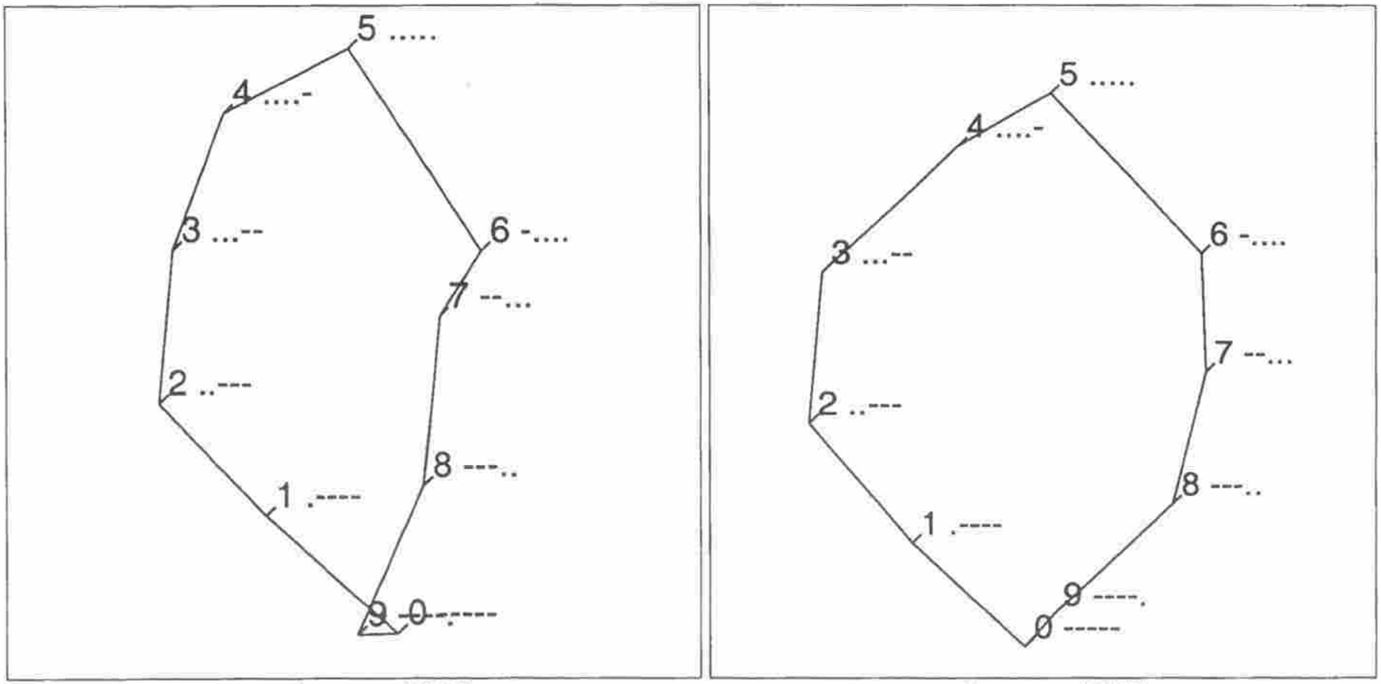
We start by way of illustration: Figure 4 shows a sequence of snapshots of an animation starting with a random configuration and ending with a locally converged nonmetric MDS configuration in $k = 2$ dimensions for the Morse code dissimilarities.

Animation of stress minimization gives users a way to check convergence of the configuration. The stress function alone is sometimes not a good numerical indicator of convergence because the stress can be quite flat near a local minimum. Numerical stopping criteria, as used in KYST-2 (Kruskal, Young, and Seery 1978), for example, may kick in when gradient steps may still be visually noticeable. In such situations it is highly desirable to have the ability to check convergence visually and stop the algorithm interactively.

It is difficult to demonstrate the benefits of visual convergence checks in print because the motions near a local minimum tend to be small and difficult to convey by comparing two static plots, yet trivial to pick up by eye. We therefore omit further illustrations and close this brief section with a general remark: Human vision is extremely acute at detecting motion throughout the field of vision, including the periphery. As a consequence, there is no need for a user to focus on any particular area of a dynamic plot: motion can be picked up literally out of the corner of the eye. Motion detection is therefore quite robust to the unpredictability of users' eye motions.

## 4. Local Minima

Most versions of MDS have trivially multiple minimum configurations because of symmetries in the stress function. Stress functions are invariant under rotations when the metric in configuration space is Euclidean; and they are invariant under reflections on the axes when the metric is general Minkowski, which includes Euclidean and city block metrics. Therefore, in discussions of local minima in MDS it is always implicit that two configurations are "different" only if they are not images of each other under transformations that leave the stress function invariant. To facilitate such comparisons, one would really need configuration matching with the Procrustes method. Matching of configurations is sometimes difficult in three and higher dimensions, but in two dimensions it can usually be done visually.



nonmetric, stress = 0.0569

metric, stress = 0.1447

Figure 3. Configurations of the Morse codes of length five in 2-D, obtained with nonmetric and metric MDS, respectively.

Examples of truly different local minima are shown in Figure 5, where the Morse code dissimilarities are scaled into two dimensions. The six locally minimal configurations are sorted in ascending order of stress. The first two differ mostly only in a local inversion of the placement of the two shortest codes in the bottom left, "E = ·" and "T = –". The third configuration places the shortest codes at the top, which implies a slight deformation of the rest of the configuration when compared to the first two plots. The fourth configuration is very similar to the first two, but this time the shortest codes are placed in the top left. The fifth configuration is more conspicuously different from the preceding ones in that the codes of length two together with those of length one are trapped at the top; the code "S = · · ·" forms a barrier that is impenetrable for the shorter codes. The sixth and last configuration is the most deformed in that both the codes of length one and two are trapped to the right of the digits.

In all six configurations of Figure 5 the codes of length three, four, and five attempt to reflect the dimensions of *code length* and *fraction of dots*. In fact, we were never able to achieve stronger rearrangements of the longer codes than those seen in Figure 5. We have therefore another indication that in 2-D the placement of the short codes of length 1 and 2 is problematic, whereas the placement of the long codes of length 3, 4, and 5 is quite robust.

In Figure 5 we showed only nonmetric solutions. It is known that different varieties of MDS suffer from local minima to differing degrees: Classical MDS produces essentially unique configurations because it is solved by an eigendecomposition; comparing between metric and nonmetric Kruskal-Shepard MDS, the former is sometimes thought to be less prone to multiple local minima, but this is not so. Metric MDS is less prone to degeneracies than nonmetric MDS, but metric MDS can actually be more prone to local minima than nonmetric MDS. This problem is particularly severe when the raw dissimilarities require a strongly nonlinear transformation to achieve a good fit, which is in fact the case for the Morse code dissimilarities. To give an idea of the extent of the problem, we show in Figure 6 two local minimum configurations. Although we were not able to upset the basic structure of the long Morse codes with nonmetric MDS, we could easily do so with metric MDS. Local barriers abounded, and almost any point could get trapped in implausible places. To force MDS to behave more reasonably, one needs a strongly nonlinear transformation of the dissimilarities. Nonmetric MDS will find such a transformation, but our interactive experiments showed that metric MDS applied to a third power of the dissimilarities will do almost as well. This point will be illustrated in Section 5 where we also give a reason for the problem of multiple minima as it manifests itself in the application of metric MDS to the raw Morse code dissimilarities.

In practice local minima are easily diagnosed if the software used offers

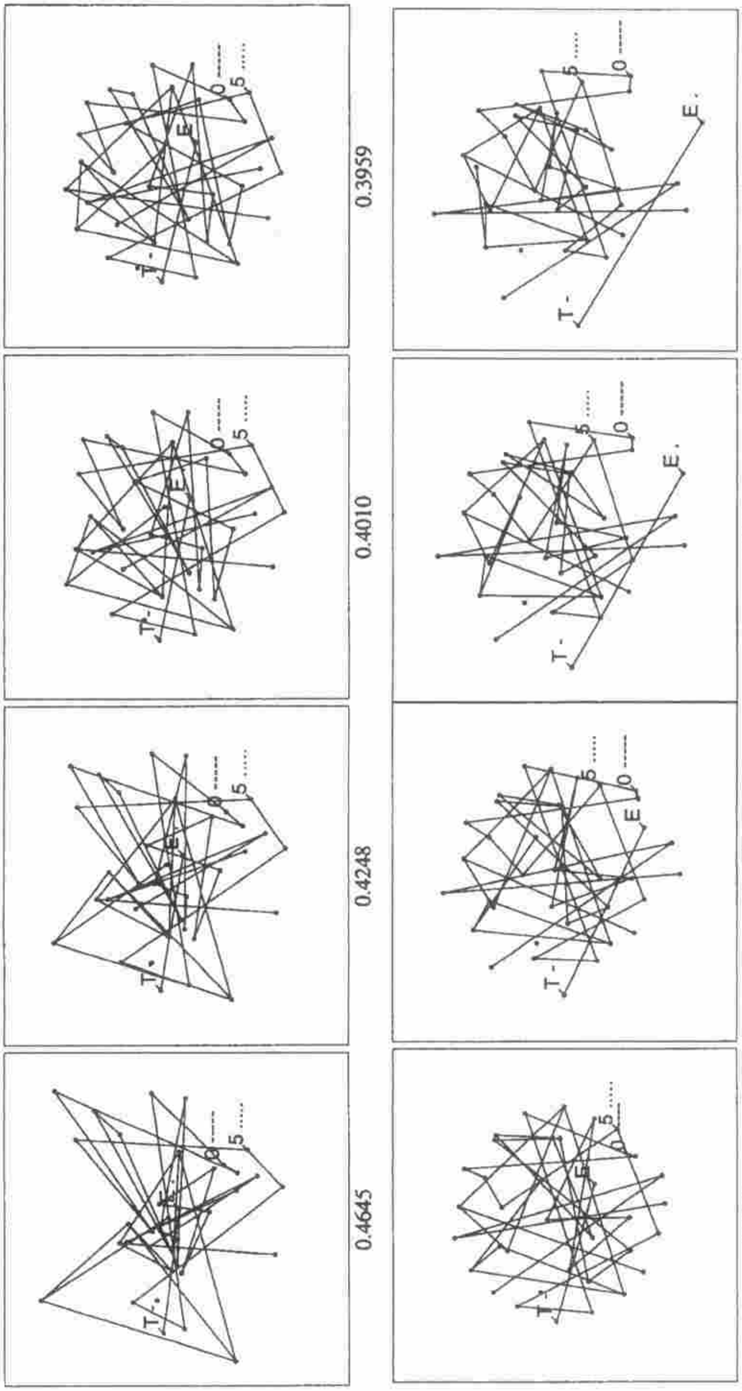
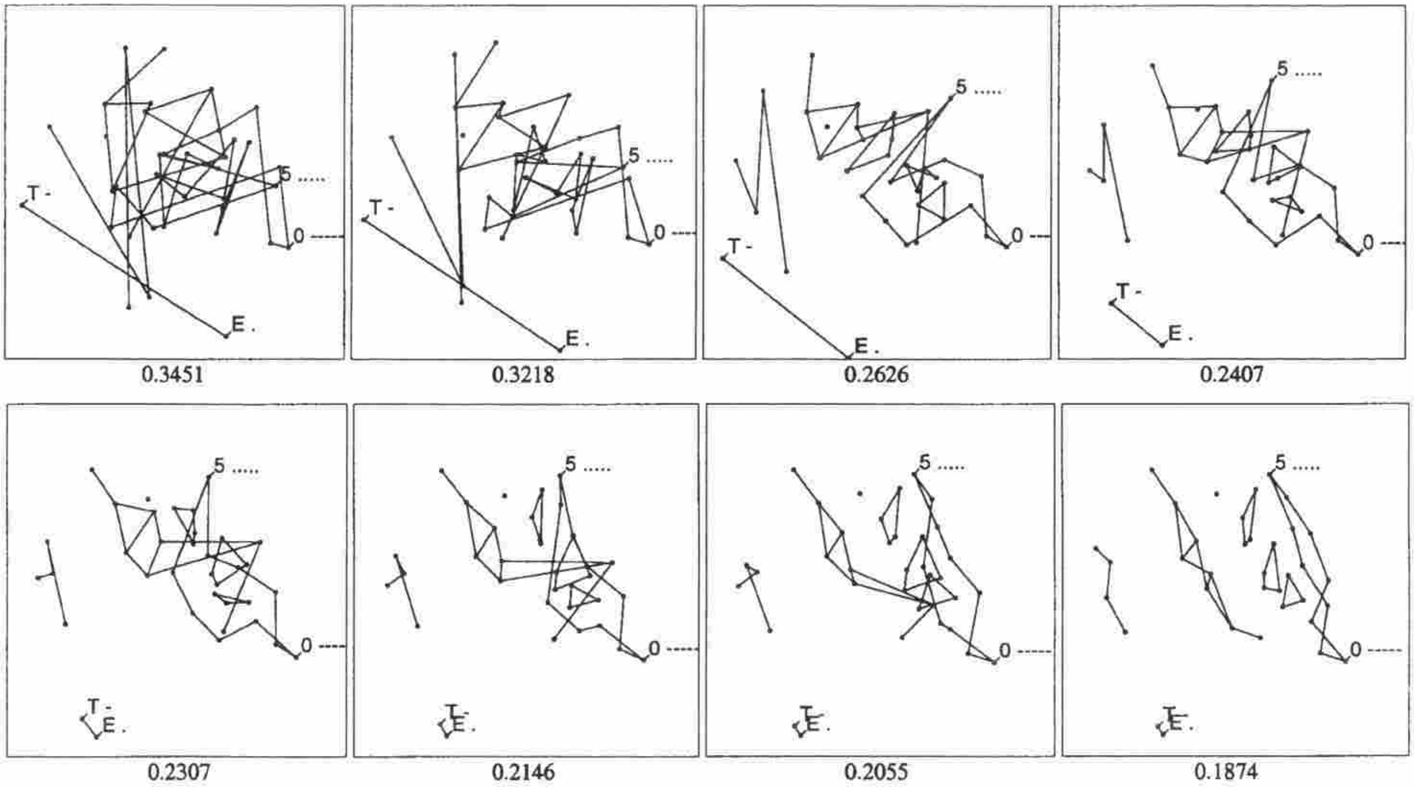


Figure 4. Snapshots from an animation of the stress minimization for nonmetric scaling of the Morse code data in 2-D. The numbers below the frames are the stress values. (Figure 4 continued on next page.)



0.3451 0.3218 0.2626 0.2407

0.2307 0.2146 0.2055 0.1874

Figure 4. (continued) Snapshots from an animation of the stress minimization for nonmetric scaling of the Morse code data in 2-D. The numbers below the frames are the stress values.

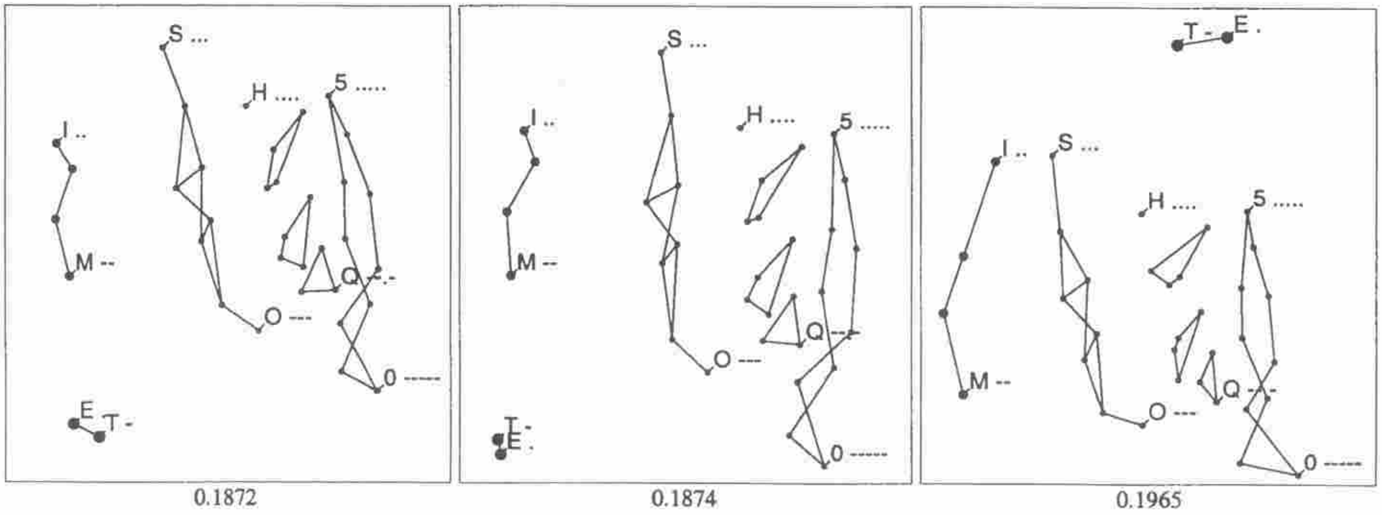


Figure 5. Multiple local minima of the *nonmetric* MDS stress function. The figure shows three converged configurations in two dimensions for the Morse code data. The stress value appears below each frame. (Figure 5 continued on next page.)

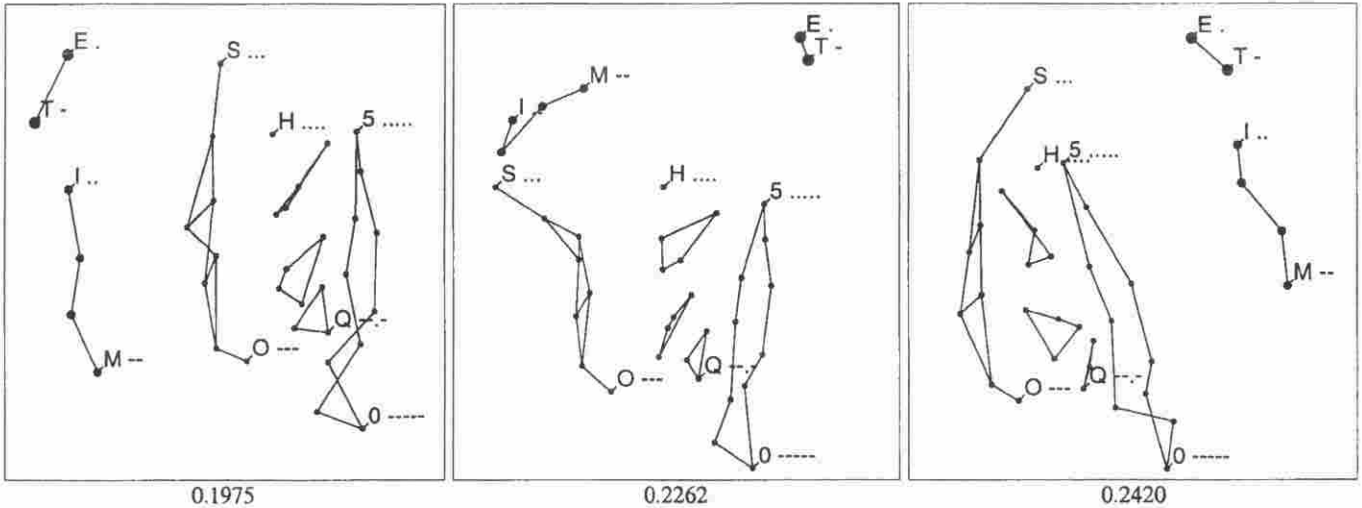


Figure 5. (continued) Multiple local minima of the *nonmetric* MDS stress function. The figure shows three converged configurations in two dimensions for the Morse code data. The stress value appears below each frame.

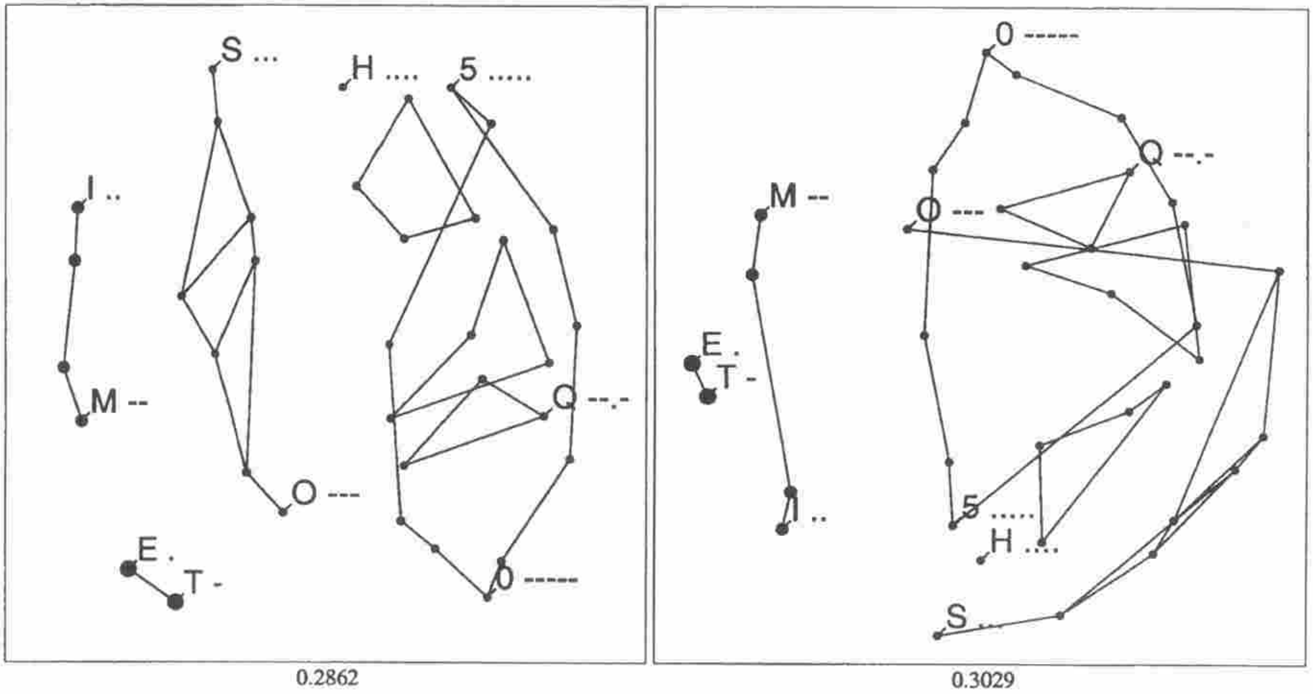


Figure 6. Multiple local minima of the *metric* MDS stress function. The figure shows two converged configurations in two dimensions for the Morse code data, none of which is an absolute minimum; the second frame of Figure 7 shows a solution whose stress is lower than either of the two shown here.

a few basic techniques. The three techniques we found most helpful are the following:

(1) Repeated stress minimization starting from *random configurations*: The metric solutions shown in Figure 6 were created in this way. An early recommendation for restarts from random configurations was made by Arabie (1973), who criticized published simulations for their dependence on particular starting configurations, usually the classical MDS solution. If conventional optimizers such as gradient descent are used, a number of solutions obtained from random restarts should be checked.

(2) Stress minimization starting from *systematically constructed configurations*: The most popular systematic starting configuration is the classical MDS solution (also available in our system), but sole reliance on it should be discouraged. An example of a different kind of systematic starting configuration based on prior insight is the following: for the Morse code dissimilarities, we form a starting configuration in 2-D by plotting the *number of dots* against *code length*. That is, we start from a configuration that is a perfect representation of the two major dimensions approximately recovered by MDS. The solutions are shown in the first frame of Figure 5 (nonmetric) and in the top right frame of Figure 7 (metric). It is no surprise that their stress values are the lowest we could find among local minima.

(3) Extensive experimentation is possible if the software at hand permits *interactive editing of configurations*. Users can then modify solutions by moving points or groups of points into suspected locations of local stability and rerun the optimizer to check the guess. This strategy is indeed how we generated the local minima in all except the first frame of Figure 5. In the first four frames we dragged the codes of length 1 into various positions while continuing to run the optimizer; in the fifth and the sixth frame we dragged the codes of length 1 and 2 to the top and to the right, respectively.

All three approaches are implemented in the XGvis/XGobi software: random restarting with a mouse click, importing precomputed configurations from files, and manually dragging points and groups of points. Point dragging was simultaneously and independently implemented by McFarlane and Young (1994) in their ViSta-MDS software. In the XGobi software, dragging points and groups of points is possible in rotated and toured views as well: dragging on the screen is translated into motion parallel to the projection plane in data space. Projection planes are implicit in all data rotations and tours.

## 5. The Problem of Indifferentiation

The problem of indifferentiation arises when dissimilarity data cluster around a positive constant. Such clustering is easily diagnosed with a histogram of the dissimilarities, an example of which is shown in the histogram of the raw Morse code dissimilarities in the top right frame of Figure 7. Data of this type approximate an extreme case in which the dissimilarities are all identical: $d_{i,j} = c \; \forall i \neq j$, where $c > 0$. This situation is illustrated by the histogram in the top left frame of Figure 7.

### 5.1 Constant Dissimilarities as the Extreme of Indifferentiation

Constant dissimilarities are a form of null data in which every object is equally dissimilar to every other object – hence our term "indifferentiation". The tighter a histogram of dissimilarities clusters around a nonzero value, the more the data suffer from indifferentiation.

Constant dissimilarities call for a configuration that is a regular simplex in $(N-1)$-dimensional space. A simplex re-creates constant dissimilarities exactly, with zero stress. When one flattens the $(N-1)$-D simplex with MDS into lower dimensions, the stress increases as the dimension decreases (following the intuitions behind Shepard's (1962) approach to MDS).

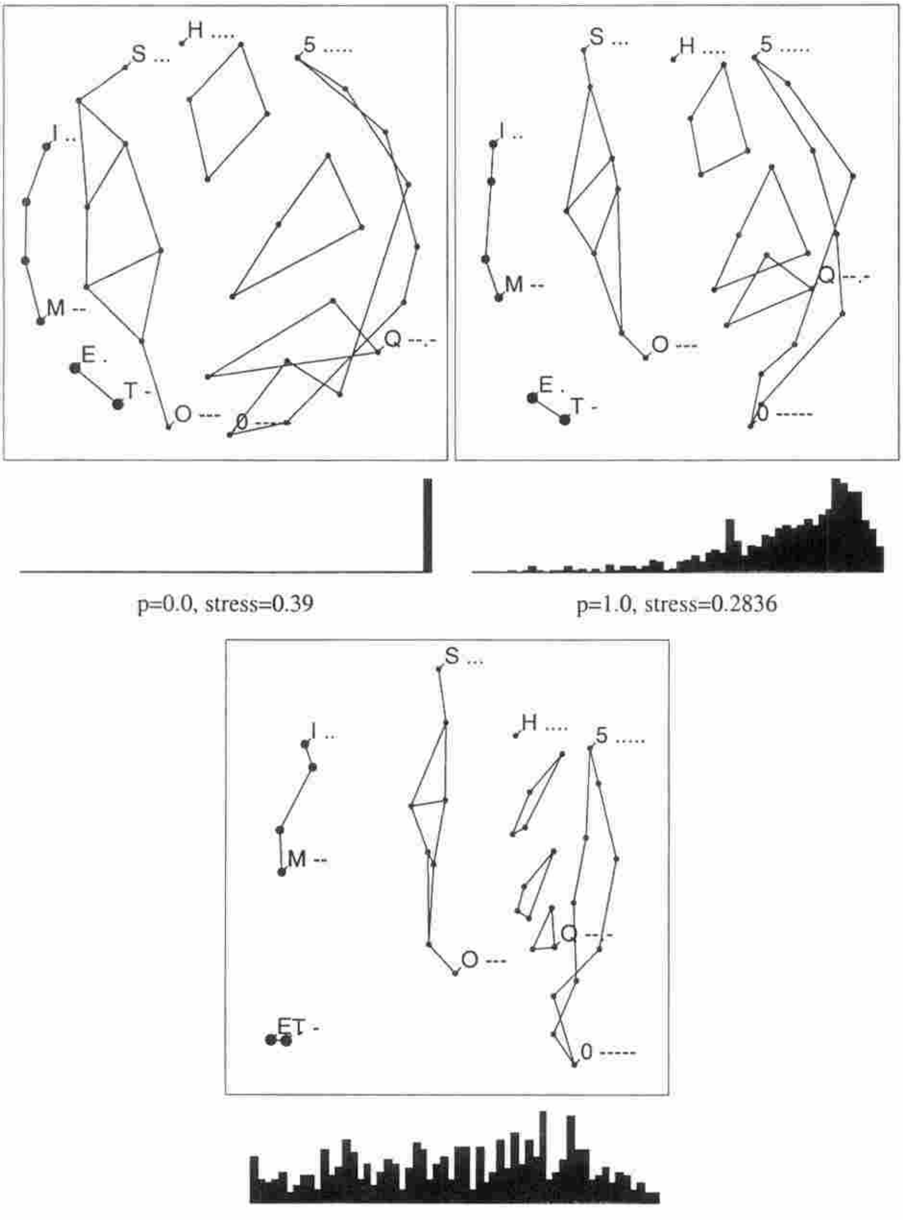
Whatever the configuration, though, the stress for constant dissimilarities is invariant under permutation of the objects:

$$Stress_D(\mathbf{x}_1, ..., \mathbf{x}_N) \;=\; Stress_D(\mathbf{x}_{\pi(1)}, ..., \mathbf{x}_{\pi(N)})$$

As a consequence, permutation of the labels of a minimum configuration yields another minimum configuration: There may exist as many as $N!$ different minimum configurations (actually: equivalence classes of solutions, modulo transformations that leave $Stress_D$ invariant, such as rotations under the Euclidean metric and axis reflections under general Minkowski metrics). Permutation symmetry under indifferentiation lends itself as an explanation for the abundance of multiple local minima in the application of metric MDS to data sets that exhibit approximate indifferentiation, such as the raw Morse code dissimilarities.

### 5.2 Power Transformations for the Analysis of Indifferentiation

Approximate indifferentiation does not necessarily mean that the dissimilarities are uninformative. We know, for example, from the application of nonmetric MDS that the Morse code dissimilarities are indeed highly structured and hence informative after the application of a monotone transformation. To make

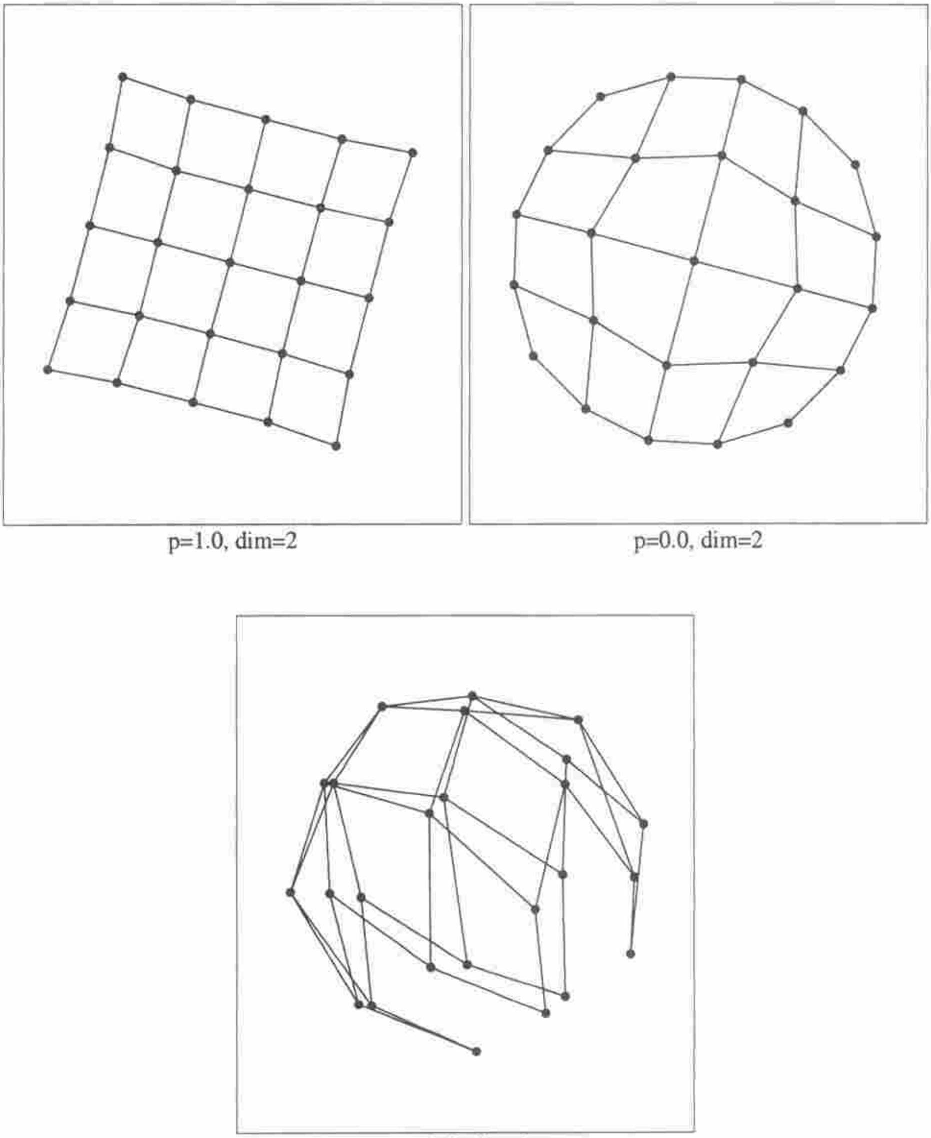


p=0.0, stress=0.39

p=1.0, stress=0.2836

p=3.2, stress=0.2095

Figure 7. Metric MDS solutions of the Morse code dissimilarities after power transformations $d^p_{i,j}$. Below the configurations are histograms of the transformed dissimilarities $d^p_{i,j}$, and the powers $p$ and the stress values.



p=1.0, dim=2

p=0.0, dim=2

p=0.0, dim=3

Figure 8. An example of the effects of complete indifferentiation on metric MDS. The raw dissimilarities describe a 5×5 grid as reflected in the 2-D configuration in the top left frame. When subjected to the power zero, the dissimilarities become constants. The results are the 2-D configuration in the top right frame and the 3-D configuration in the bottom frame.

metric MDS more competitive with nonmetric MDS, we implemented power transformations in XGvis. Their exponent is controlled by a slider widget that permits users to change the exponent interactively with the mouse. Our typical mode of operation is to drag the slider and therefore change the exponent while running the stress minimizer at the same time. We then get to see the immediate effect of changes in the exponent on the stress value and the configuration. The two main purposes of user-controlled power transformations are the following:

(1) Exploring the effect of transforming the dissimilarities to indifferentiation, by lowering the exponent to zero while running the stress optimizer. This operation indicates how close the raw dissimilarities are to indifferentiation. An example is shown in the left frame of Figure 7. Comparison of the two top frames of Figure 7 shows that the two are indeed quite close: the rounding of the configuration in the right frame approximates the circular configuration in the left frame.

(2) Searching for the lowest stress value by sliding up and down the scale of exponents. This is how we found that the exponent $p = 3.2$ is approximately optimal. In the bottom frame of Figure 7 we notice that the histogram of the transformed dissimilarities is flat, in particular, it is not clustered around a positive constant, and the configuration is very similar to the nonmetric configurations in the left and the center frame of Figure 5.

### 5.3 The Structure of Null Configurations

Minimum configurations of constant dissimilarity data are highly structured. For a first impression, see the left frame of Figure 7 and the center and right frame of Figure 8. Knowledge of this "null structure" is of considerable importance for the practice of MDS because this is structure in the output of MDS that indicates the absence of real structure in the input data, an example of the unlikely case of "garbage in, structure out". For real structure to be completely absent is rare, but it is often weak, which puts such data in the vicinity of indiscrimination with approximate null structure in the configurations. This observation is the methodologically important point: null structure appears to a variable degree, and it must be recognized as such to avoid overinterpretation of the data.

We first describe the null structure of MDS solutions for the case of perfectly constant dissimilarities as seen in computer experiments:

(1) In two dimensions, a minimum configuration often arranges the points on a set of concentric circles, as in the top left frame of Figure 7 and the top right frame of Figure 8. This fact has been widely noted and described

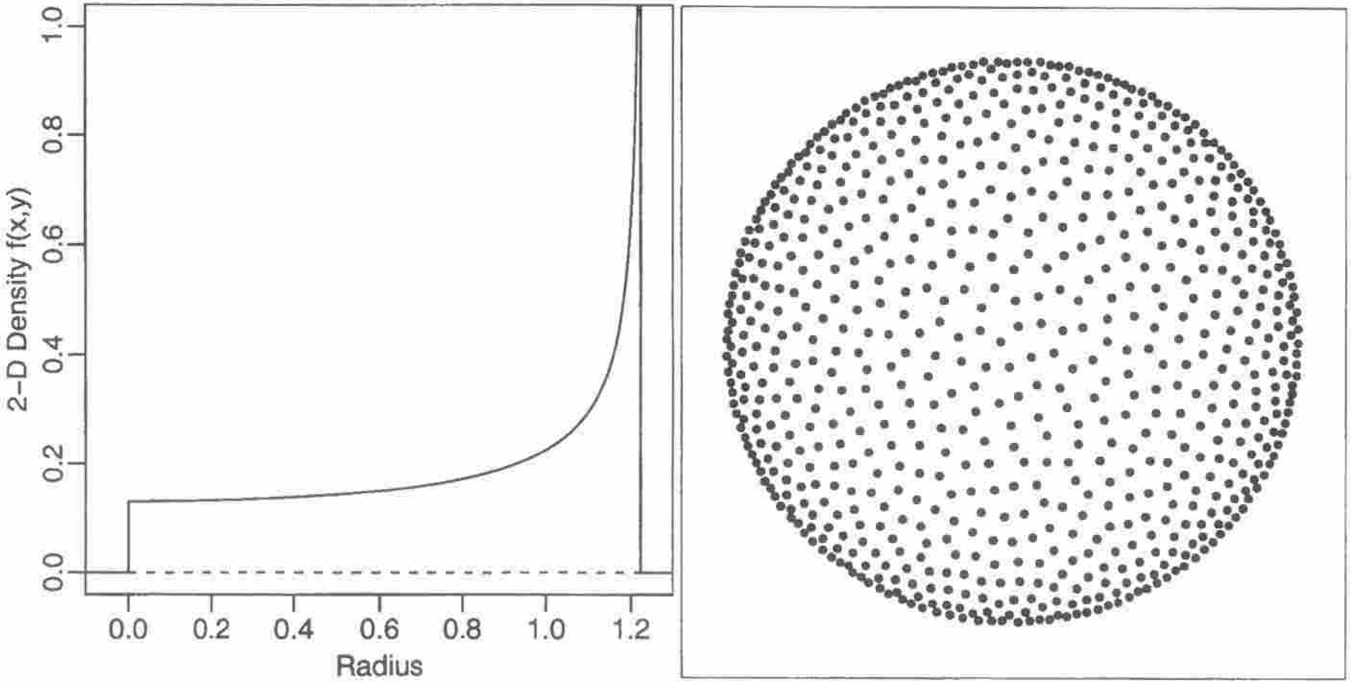


Figure 9: Null analysis in 2-D: All configurations are of size 750. Left: the null density for $N \to \infty$ as a function of radius; its form is $f(x_1, x_2) = (1 - (r/R)^2)^{-1/2}/(2\pi R)$ where $r = (x_1^2 + x_2^2)^{1/2} < R = (3/2)^{1/2}$. Right: a null configuration obtained from constant dissimilarities, both with metric and nonmetric MDS. (Figure 9 continued on next page.)

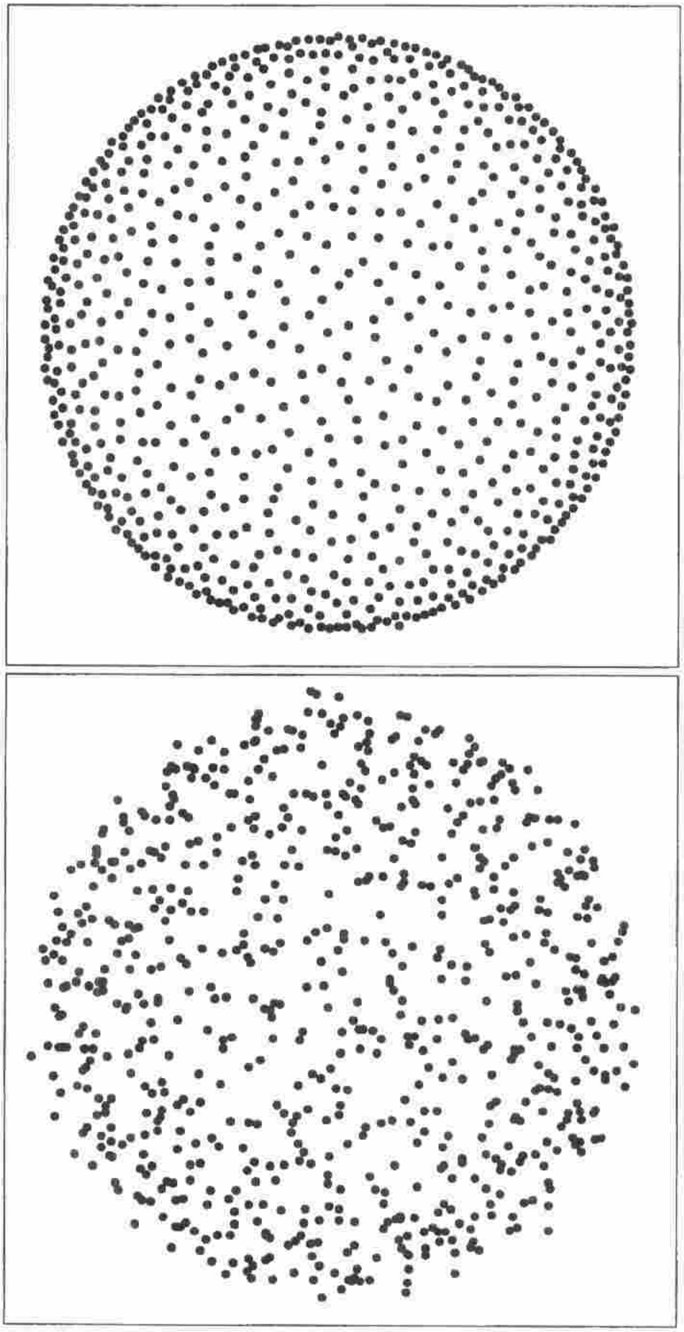


Figure 9. (continued) Null analysis in 2-D: All configurations are of size 750. Two configurations obtained from uniformly random dissimilarities; left: metric; right: nonmetric.

by, for example, de Leeuw and Stoop (1984, p. 397). The concentric circles, however, are somewhat inessential to a null configuration in 2-D. In light of theoretical results described below, the essential aspect of a null configuration is that it shows a point density that fills a circular disk with sharp boundary; the density is circularly symmetric with lowest density in the center and increasing density towards the boundary.

(2) In 3-D and higher dimensions, a minimum configuration arranges the points so as to approximate a uniform distribution on a sphere. This result is harder to illustrate in the printed medium, but the bottom frame of Figure 8 gives an impression of this effect for data that describe a $5 \times 5$ grid in their raw form, after having been made constant with a zero-th power transformation. In an interactive data visualization system such as XGvis/XGobi, one uses data rotations and sections to verify that the configurations are indeed spherically symmetric and hollow in the center (Furnas and Buja 1994; a data section is a "slice" or subset of data points selected with a narrow rectangle in a data projection).

These types of dimension-dependent null structure have a mathematical basis described in Buja, Logan, Reeds, and Shepp (1994), under an idealization in which the number of objects $N \to \infty$. The analysis suggests the following:

(1) In two dimensions the minimum configurations approximate a circularly symmetric distribution on a disk with a density that increases radially from the center to the periphery of the disk. The top left of Figure 9 shows the theoretical density as a function of radius.

(2) The same analysis suggests that in 3-D and higher dimensions the minimum configurations approximate a uniform distribution on a sphere.

### 5.4 Noisy Dissimilarities and Indifferentiation

The problem of indifferentiation arises not only when dissimilarities accrue around a positive value. The same effect occurs when the dissimilarities are noisy. The reason is essentially that for sufficiently large $N$, noise washes out and only its expected value is of relevance. In short: if dissimilarities $\{d_{i,j}\}_{i<j}$ are independently and identically distributed, then they are asymptotically equivalent to constant dissimilarities $d_{i,j} = E(d)$, that is, to indiscrimination. This observation holds more so for nonmetric than metric MDS because the isotonic regression in nonmetric MDS smooths random $d_{i,j}$ very nearly to a constant.

An illustration of these facts is shown in Figure 9: The top right shows a null solution for perfectly constant dissimilarities; the bottom row shows a metric and a nonmetric solution for uniform random dissimilarities. The nonmetric solution matches the null configuration, while the metric solution is a fuzzy version thereof. This comparison indicates that metric MDS has greater problems with noise than nonmetric MDS because the latter has a certain ability to average out noise.

### 5.5 Null Structure in Empirical Configurations

In light of these facts, it is now possible to interpret the 2-D metric configuration of the raw Morse code dissimilarities shown in the top right frame of Figure 7: The circular shape is in all likelihood a consequence of a fair amount of indifferentiation in the raw dissimilarities. The histogram of the dissimilarities below the plot confirms this impression with the values accruing near the maximum value.

A similar diagnosis is possible for Figure 4, the leftmost frame in the second row: there, the initial configuration is random and its pairwise distances largely uncorrelated with the dissimilarities; hence the isotonic regression maps the latter very nearly to a constant, which generates an approximate null configuration as a transition phase of the optimization. This result is probably a general fact for MDS minimizations that start from a random configuration. Because of slight deviations from the perfect null situation, gradient descent still finds its way to a true local minimum. This observation is just a diagnosis, not a criticism of random starts.

A general conclusion from these considerations is that the following properties of local minimum configurations are usually artifacts:

(1) in 2-D: circular disk shape with a sharp edge and low density in the center;

(2) in 3-D and higher: sphericity and holes in the center.

These features should not be interpreted as properties of the data, but as hinting at a degree of indifferentiation, possibly stemming from noise.

### 5.6 Horseshoes

We suspect that the well-known "horseshoe effect" (Kendall 1970) derives from the null structure just described. Some empirical evidence can be found in Figure 10: The dissimilarities in these plots were generated from 50 perfectly ordered equispaced points on a line, as in the leftmost frames of Figure 10. These dissimilarities were subjected to several power transformations with exponents starting at $p = 1.0$ and ending at $p = 0.0$. In 2-D, the intermediate exponents exhibit very clear horseshoe shapes. They turn scraggly for exponents near zero, so as to approximate a 2-D null configuration. In 3-D, there are more possibilities for a curve to bend; hence the shape is no longer so simple that it could be described as a horseshoe. However, data that have intrinsic 2-D structure such as the $5 \times 5$ grid of Figure 8 will be bent in a simple way because they must approach a sphere in the limit, as in the bottom frame of the figure.

The horseshoe effect in metric and nonmetric MDS of the Kruskal-Shepard variety should not be confused with the horseshoe effect in multivariate methods that are based on eigendecompositions, such as correspondence analysis and nonlinear principal component analysis. In these methods the horseshoes have a much simpler explanation: they arise in the same way as certain systems of orthogonal polynomials arise as solutions of spectral decompositions of certain linear operators. Horseshoes are here just an expression of the fact that the two dominant eigenfunctions are linear and quadratic. In noisy data this situation translates to two dominant eigensolutions, one which is ascending and one which is broken into two pieces, one piece descending and the other ascending. We leave things intentionally vague as they are the topic of another literature (see for example Buja 1990; Donnell, Buja, and Stuetzle 1994, and the references therein). We only note that the mathematics of null configurations in MDS (idealized for $N \to \infty$) is very different: it requires the solution of a variational problem that has no relation to eigendecompositions. A closer relative is potential theory because of the similarity in the roles of dimensions one and two versus three and higher, but the variational problem can not be reduced to potential theory either.

## 6. Localization with Groups

By "localization" we mean the examination of structure contained in relatively small dissimilarities. Localization is a difficult problem because of the following received wisdom about MDS:

> *The global shape of MDS configurations is determined by the large dissimilarities; consequently, small distances should be interpreted with caution: they may not reflect small dissimilarities.*

These statements are based on a study by Graef and Spence (1979) who ran simulations in which they removed, respectively, the largest third and the smallest third of the dissimilarities. Those authors found devastating effects when removing the largest third, but relatively benign effects when removing the smallest third. We will qualify these conclusions in the next section.

A refinement of the received wisdom is the following: if the points representing the Morse codes "T" and "E" lie close together, it does not follow that they are perceptually similar, that is, often confused. It may much rather mean that there exists a large set of codes from which they are both roughly

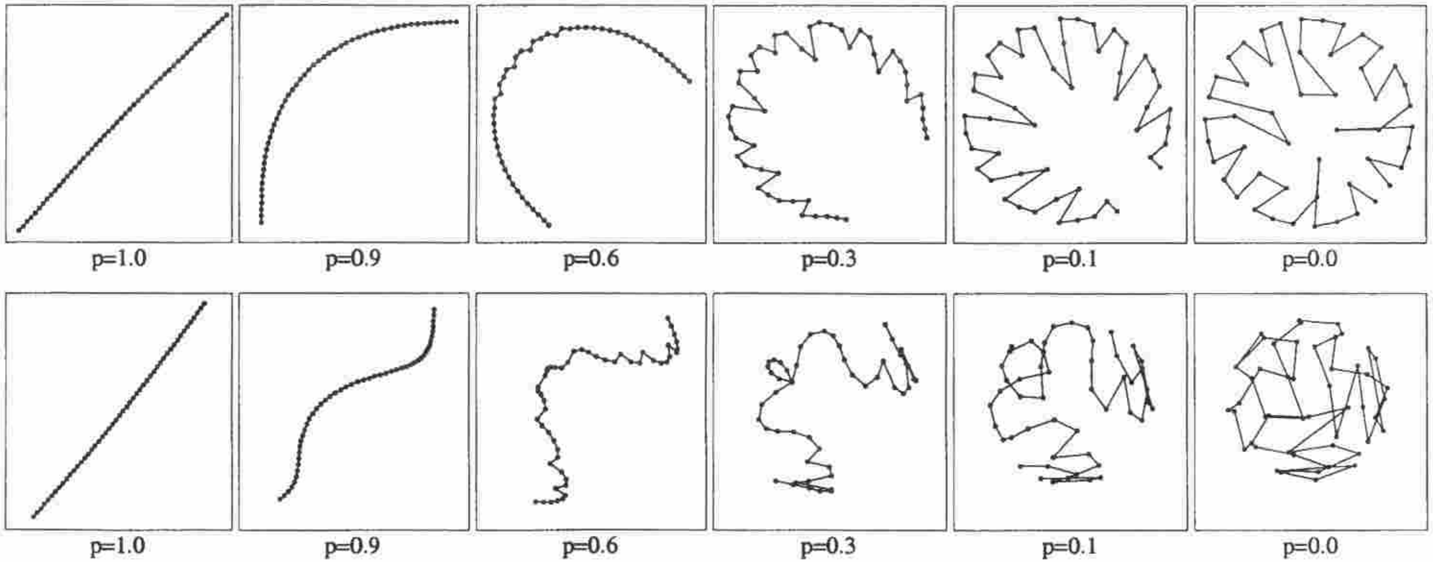


Figure 10. The horseshoe effect for metric MDS as an approximation to the null structure of indiscrimination. Top row: 2-D; bottom row: 3-D.

equally dissimilar. *Shared patterns in large dissimilarities are vastly more powerful contributors to the stress function than is the single dissimilarity* between the codes "T" and "E". Thus, it would be of interest to diagnose whether "T" and "E" are indeed close.

To answer these and similar questions, one may want to use some form of localized MDS. One proposal is as follows: Assume the objects have been partitioned into disjoint groups, where the groups have presumably been chosen to be homogeneous in some sense, as in the Morse code data the groups of codes of the same length. Then perform MDS with a stress function that has a reduced set of dissimilarities, namely, those for *pairs of objects in the same group*; omitted are the terms in the stress function that contain dissimilarities for objects in different groups. We have dubbed this method "within-groups MDS". The resulting locally minimal configurations have the following properties:

(1) The relative positions of points in the same group are meaningful because they are constrained by the within-group dissimilarities.

(2) The relative positions of the group configurations are *not* meaningful because they are unconstrained because of the removal of the between-groups dissimilarities. Similarly, orientations of groups configurations are *not* meaningful. Users of XGvis can experience this lack of constraint by dragging the groups interactively. [There exists one overall constraint, though: we always center the overall configuration at the origin and we always normalize the overall configuration size. This constraint does not seem to pose problems for users.]

(3) In metric MDS the relative sizes of the groups are meaningful and can be compared because the within-group configuration distances approximate dissimilarities that exist on the same scale. In nonmetric MDS it is not clear how to handle the isotonic transformation: in XGvis we estimate one shared transformation across groups, which again puts the transformed dissimilarities on a shared scale so that group sizes can be compared. The alternative of estimating a separate isotonic transformation in each group would decouple the group sizes, but this version is problematic because groups are often small and degenerate transformations become likely.

XGvis uses groups defined by colors and glyphs. The groups can be precomputed and entered in color and glyph files, or they can be continually redefined with brushing operations in the XGvis window or in auxiliary linked XGobi windows.

Figure 11 shows within-groups configurations for the Morse code dissimilarities partitioned into groups of constant code length. We rearranged and rotated the groups by dragging them interactively to new locations and orientations to facilitate comparisons. One recognizes that the shortest codes, "E"

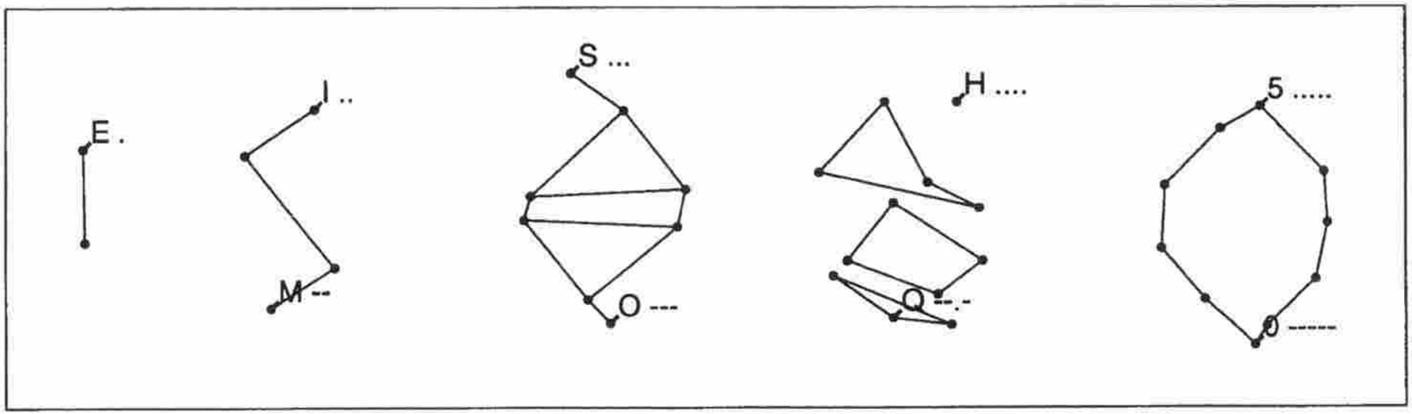


Figure 11. Metric within-groups MDS of the Morse code data, with groups defined by code length.

and "T", are relatively far apart from each other, farther than previous figures would have made us believe. In the circular structure of the codes of length five adjacent codes are clearly closer to each other than "E" and "T", which seems intuitive: we would expect for example "$- - \cdot \cdot \cdot$" and "$- - - \cdot \cdot$" to be more often confused with each other than "$\cdot$" and "$-$".

## 7. Localization by Truncation and Weighting

Within-groups MDS is a powerful way of exploring local structure. Yet there is another method of localization that is often proposed for MDS: dropping the large dissimilarities from the stress function and approximating only the small dissimilarities with configuration distances. The intuition behind this idea is that small dissimilarities reflect local properties, and that by building configurations from local information one obtains more faithful global configurations. This proposal is generally a great disappointment, which is not too surprising in light of Graef and Spence's (1979) work quoted earlier. Minimization of stress without the large dissimilarities does often not converge to meaningful global configurations. The approach fails mathematical intuitions trained on differential equations, where infinitesimal information is successfully integrated up to global solutions. Attempts at integrating local to global structure are often not successful in MDS.

Just the same, it is of interest to know whether removal of large dissimilarities actually fails MDS for a particular dataset. In XGvis we implemented two mechanisms for assessing the influence of large *and* small dissimilarities:

(1) *Truncation* to drop large (or small) dissimilarities from the stress function. Both truncation thresholds can be interactively controlled.

(2) *Weighting* to smoothly change the influence of small and large dissimilarities. The weights we provide are powers of the dissimilarities: $w_{i,j} = d_{i,j}^q$. For $q < 0$ large dissimilarities are down-weighted; for $q > 0$ they are up-weighted. (The default is identical weights: $q = 0$.) The exponent $q$ can be interactively controlled.

In our experience, both truncation and weighting have a data-dependent range in which they produce useful alternative configurations. Outside this range, MDS minimization tends to disintegrate. Figure 12 shows configurations obtained by truncating successively larger numbers of largest dissimilarities. For each configuration, stress minimization was started from the previous configuration. This minimization scheme masks the full scale of instability that would be apparent if one started each minimization from a random configuration. With the stability-favoring scheme, almost half the largest dissimilarities can be removed and the configurations are still meaningful. The Graef and Spence (1979) result should therefore be taken with a grain of salt: the fraction of dissimilarities that can be removed depends very much on the data and on the type of starting configurations used.

The interpretation of local features even in seemingly meaningful configurations requires caution, though, because of potential decoupling of distant objects. In the bottom row of Figure 12, for example, all dissimilarities between the codes {"E", "T"} and the rest had been truncated, and the placement of these codes is inherited from the last configuration in which a constraint to the rest existed.

The detection of decoupling is therefore a necessity. In XGvis there exist two different approaches to the problem: (a) instant local feedback can be obtained by interactively dragging a point of interest while stress minimization is in progress; if the point is constrained, it will instantly snap back into its position on release. (b) A global overview of the constraints can be obtained in a scatterplot of the indices $i$ and $j$ for which the dissimilarity $d_{i,j}$ is present in the current stress function; such a plot is accessible in a diagnostics window that shows the included dissimilarities $d_{i,j}$, their fitted distances, and their indices $i$ and $j$.

We end this section by noting that the problem of decoupling does not arise with power-weighting of dissimilarities: large dissimilarities are only downweighted, but they never disappear from the stress function.

## 8. The Use of Minkowski Distances for Rotation of Configurations

General Minkowski (or Lebesgue) distances on configuration space are sometimes used as alternatives to Euclidean distances. This family of metrics is parametrized by a parameter $m$ which ranges between 1 and $\infty$, both limits included. For $m = 2$ one obtains the Euclidean metric as a special case. Of particular importance are the two extremes of the Minkowski family: for $m = 1$ the $L_1$ (or city block or Manhattan) metric, and for $m = \infty$ the $L_\infty$ (or maximum) metric. Greater plausibility of $L_1$ or $L_\infty$ over the $L_2$ metric has been argued (Arabie 1991). MDS with these metrics is not properly solved by gradient descent because the corresponding minimization problem is more akin to a sorting problem than a smooth optimization problem (Hubert, Arabie, and Meulman 2001). Just the same, we incorporated the $L_1$ metric in XGvis as a limiting case when $m$ decreases smoothly to 1. Just like the power exponent $p$ in Section 5.2, the Minkowski parameter $m$ is controlled by a slider widget that permits users to change $m$ interactively with the mouse. The mathematical limit $\lim_{m \downarrow 1} \min Stress$ can therefore be approximated by a process in which a user slowly lowers $m$ to 1 on the slider while gradient descent is running. This process is a variant of the Arnold-Kruskal strategy which uses a ladder of

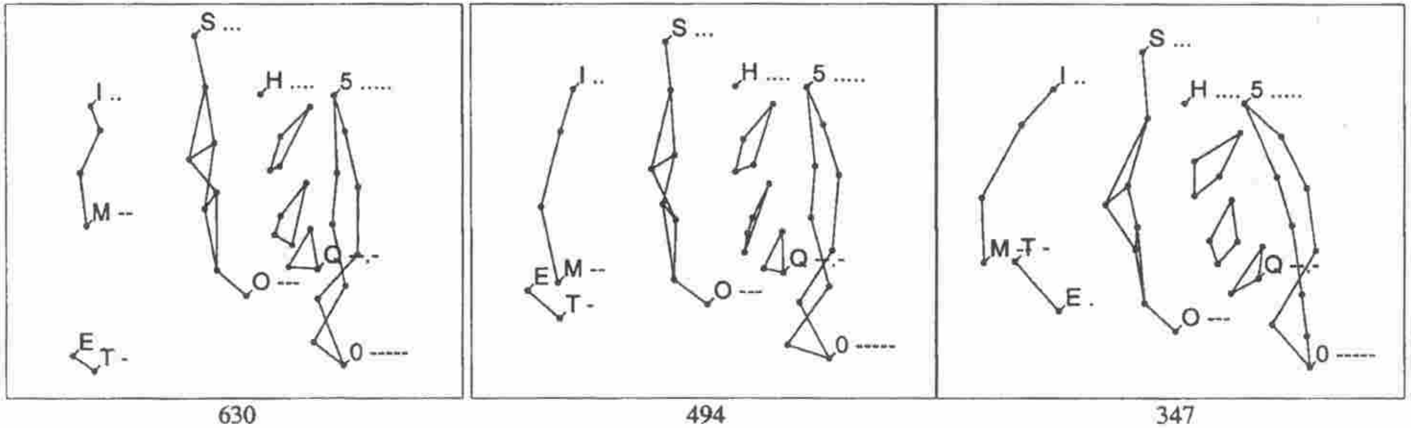


Figure 12. The effects of truncating large dissimilarities on nonmetric MDS in 2-D, applied to the Morse code data. Below the frames are the numbers of retained dissimilarities. The leftmost configuration stems from the complete set of dissimilarities. (Figure 12 continued on next page.)

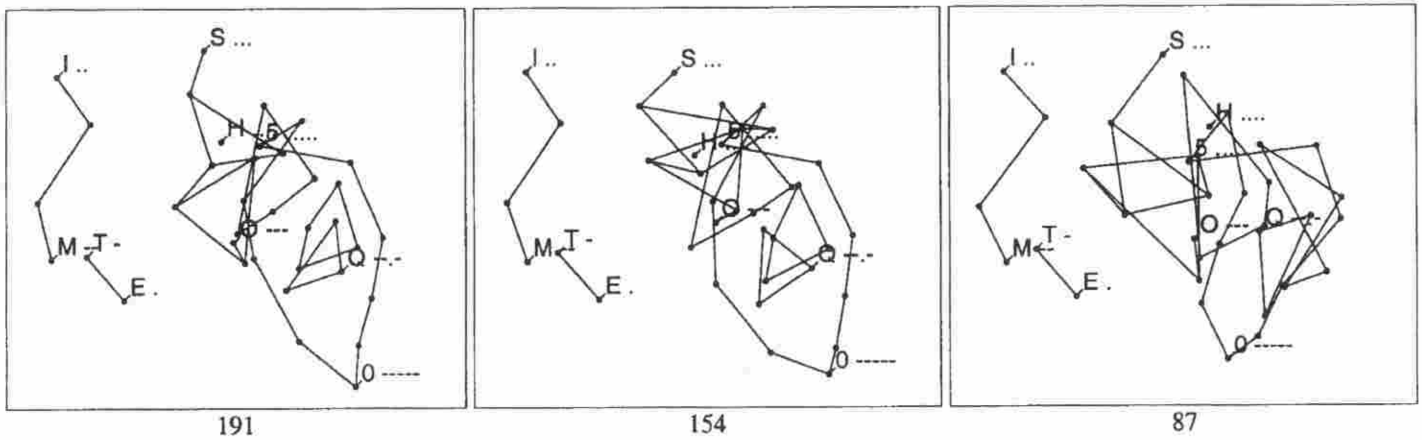
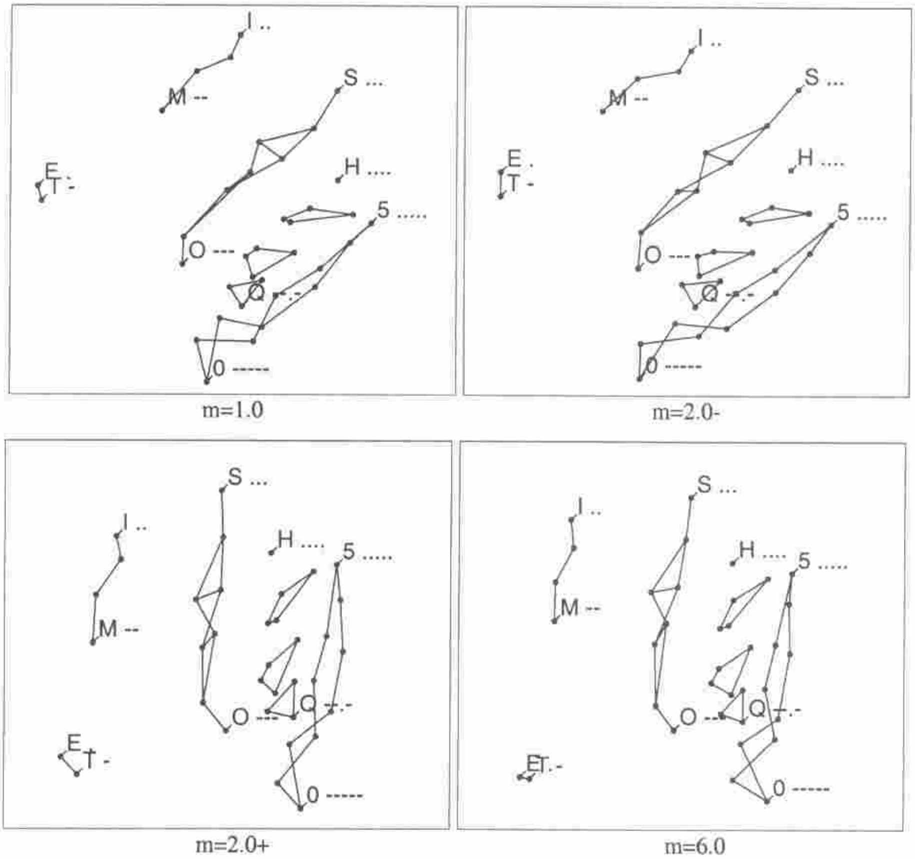


Figure 12. (continued) The effects of truncating large dissimilarities on nonmetric MDS in 2-D, applied to the Morse code data. Below the frames are the numbers of retained dissimilarities.



m=1.0

m=2.0-

m=2.0+

m=6.0

Figure 13. Minkowski metrics fitted to the Morse code dissimilarities. The Minkowski parameters are shown below the configurations. The stress values are 0.1984 ($m = 1.0$), 0.1872 ($m = 2.0$), 0.1842 ($m = 6.0$).

discrete values for $m$ to successively approximate the $L_1$ or $L_\infty$ metrics. Arabie (1991) gives empirical evidence that the strategy works well in 3-D and higher dimensions but often fails in 2-D. If this empirical finding of a discrepancy between 2-D and higher dimensions has a theoretical basis, one may wonder with Arabie (1991, p. 578) whether there exists a connection with our analysis of indifferentiation in Section 5.

In our experiments with interactive MDS, our interest in Minkowski metrics arose not from psychometric considerations but from the practical discovery that these metrics can be used for rotation of configurations for interpretation, similar to factor rotation in factor analysis. A standard method for rotating MDS configurations is principal component analysis, but non-Euclidean Minkowski metrics can be used for the same purpose because they break the rotation invariance of the stress function: For $m \neq 2$, optimal configurations must align themselves in particular ways with the coordinate axes (modulo reflections on them). This alignment often leads to interpretable axes. For example, if there exist certain axial (near-)symmetries in a configuration, non-Euclidean metrics may force the axes of symmetry to line up with the coordinate axes. To find these special alignments, the following simple recipe can be used:

> *Temporarily raise or lower the Minkowski parameter $m$ above or below 2, respectively, and return to $m = 2$ while running the stress minimizer all the while.*

One therefore approximates the mathematical limits $\lim_{m\downarrow 2} \min Stress$ and $\lim_{m\uparrow 2} \min Stress$ with interactive manipulation of $m$. The two processes should produce the same configurations up to rotation (unless the configurations get trapped in substantially different local minima, which can be avoided by moving $m$ not far above or below 2). The result in either case is a rotated version of the original $L_2$ configuration. Typically, for configurations in 2-D, the major difference between solutions based on $L_{m>2}$ and $L_{m<2}$ is a 45 degree rotation.

Figure 13 illustrates this use of Minkowski metrics with an application to the Morse code dissimilarities in 2-D: the top right and bottom left frames show the Euclidean solution in two orientations, the top right frame obtained by approaching $m = 2$ from $m = 1$, the bottom left frame by approaching $m = 2$ from $m = 6$ (the highest implemented value in XGvis). Clearly the latter frame has the more desirable solution because it roughly aligns the horizontal and vertical axes with the dimensions *code length* and *fraction of dots*.

## 9. Conclusions

The intention of this paper was to describe a rich methodology for visualizing, diagnosing, and manipulating MDS configurations. The list of techniques introduced here is by no means complete and other ideas should be tried, but we hope to have shown that MDS has much to gain from contemporary interactive data visualization.

The XGvis/XGobi software, in which this methodology can be realized, is freely available from the following web site:

http://www.research.att.com/areas/stat/xgobi/

## References

ARABIE, P. (1973), "Concerning Monte Carlo evaluations of nonmetric multidimensional scaling algorithms", *Psychometrika, 38*, 607.

ARABIE, P. (1991), "Was Euclid an unnecessarily sophisticated psychologist?", *Psychometrika, 56*, 567-587.

ARABIE, P., and SOLI, S.D. (1982), "The interface between the types of regression and methods of collecting proximity data," in: R. G. Golledge and J. N. Rayner (eds.), *Proximity and Preference*, Minneapolis: University of Minnesota Press, 90-115.

BORG, I., and GROENEN, P. (1997), *Modern Multidimensional Scaling: Theory and Applications*, New York: Springer-Verlag.

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

BUJA, A., SWAYNE, D. F., LITTMAN, M. L., DEAN, N., and HOFMANN, H. (2001), "XGvis: Interactive data visualization with multidimensional scaling," *Journal of Computational and Graphical Statistics*, tentatively accepted.

BUJA, A., COOK, D., and SWAYNE, D.F. (1996), "Interactive high-dimensional data visualization," *Journal of Computational and Graphical Statistics, 5*, 78-99. A companion video tape can be borrowed from the lending library of the Statistical Graphics Section of the American Statistical Association.

BUJA, A., LOGAN, B.F., REEDS, J.R., and SHEPP, L.A. (1994), "Inequalities and positive-definite functions arising from a problem in multidimensional scaling," *Annals of Statistics, 22*, 406-438.

CARROLL, J.D., and ARABIE, P. (1980), "Multidimensional scaling," in: M. R. Rosenzweig and L. W. Porter (eds.), *Annual Review of Psychology, 31*, 607-649.

CARROLL, J.D., and ARABIE, P. (1998), "Multidimensional scaling," in: M.H. Birnbaum (ed.), *Handbook of Perception and Cognition. Volume 3: Measurement, Judgment and Decision Making*, San Diego, CA: Academic Press, 179-250.

CARROLL, J.D., and GREEN, P. (1997), "Psychometric methods in marketing research: Part II, Multidimensional Scaling," *Journal of Marketing Research, 34*, 193-204.

COOK, D., and BUJA, A. (1997), "Manual controls for high-dimensional data projections," *Journal of Computational and Graphical Statistics, 6*, 464-480.

DAVIES, P.M., and COXON, A.P.M. (eds.) (1982), *Key Texts in Multidimensional Scaling*, Exeter, New Hampshire: Heinemann.

DE LEEUW, J., and STOOP, I. (1984), "Upper bounds for Kruskal's stress," *Psychometrika, 49*, 391-402.

DONNELL, D.J., BUJA, A., and STUETZLE, W. (1994), "Analysis of additive dependencies and concurvities using smallest additive principal components," *Annals of Statistics, 22*, 1635-1673.

GNANADESIKAN, R. (1997), *Methods for Statistical Data Analysis of Multivariate Observations*, New York: Wiley.

FURNAS, G.W., and BUJA, A. (1994), "Prosection Views: Dimensional Inference through Sections and Projections," *Journal of Computational and Graphical Statistics, 3*, 323-385.

GRAEF, J., and SPENCE, I. (1979), "Using distance information in the design of large multi-dimensional scaling experiments," *Psychological Bulletin, 86*, 60-66.

GREEN, P.E., CARMONE, F.J., Jr., and SMITH, S.M. (1989), *Multidimensional scaling: Concepts and applications*. Boston: Allyn and Bacon.

GREENACRE, M.J., and UNDERHILL, L.G. (1982), "Scaling a data matrix in a low-dimensional Euclidean space," in D.M. Hawkins (ed.), *Topics in Applied Multivariate Analysis*, Cambridge UK: Cambridge University Press, 183-268.

HUBERT, L., ARABIE, P., and MEULMAN, J. (1997), "Linear and circular unidimensional scaling for symmetric proximity matrices," *British Journal of Mathematical and Statistical Psychology, 50*, 253-284.

HUBERT, L., ARABIE, P., and MEULMAN, J. (2001), *Combinatorial Data Analysis: Optimization by Dynamic Programming*, Philadelphia: Society for Industrial and Applied Mathematics.

KENDALL, D.G. (1970), "A mathematical approach to seriation," *Philosophical Transactions of the Royal Society of London A, 269*, 125-135.

KRUSKAL, J.B., and WISH, M. (1978), *Multidimensional Scaling*, Beverly Hills and London: Sage.

KRUSKAL, J.B. (1964a), "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika, 29*, 1-27.

KRUSKAL, J.B. (1964b), "Nonmetric multidimensional scaling: a numerical method," *Psychometrika, 29*, 115-129.

KRUSKAL, J.B., YOUNG, F.W., and SEERY, J.B. (1978), "How to use KYST-2, a very flexible program to do multidimensional scaling and unfolding," (Tech. Rep.) Murray Hill, NJ: Bell Labs.

MCFARLANE, M., and YOUNG, F.W. (1994), "Graphical Sensitivity Analysis for Multidimensional Scaling," *Journal of Computational and Graphical Statistics, 3*, 23-33.

ROTHKOPF, E.Z. (1957), "A measure of stimulus similarity and errors in some paired-associate learning tasks," *Journal of Experimental Psychology, 53*, 94-101.

SEBER, G.A.F. (1984), *Multivariate Observations*, New York: Wiley.

SHEPARD, R.N. (1962), "The analysis of proximities: multidimensional scaling with an unknown distance function," I and II, *Psychometrika, 27*, 125-140 and 219-246.

SHEPARD, R.N. (1963), "Analysis of proximities as a technique for the stydy of information processing in man," *Human Factors, 5*, 33-48.

SWAYNE, D.F., COOK, D., and BUJA, A. (1998), "XGobi: Interactive Data Visualization in the X Window System," *Journal of Computational and Graphical Statistics, 7* 1, 113-130.