MULTIDIMENSIONAL SCALING
BY OPTIMIZING GOODNESS OF FIT TO A SMOOTH HYPOTHESIS

Willem J. Heiser
Department of Data Theory
University of Leiden
Middelstegracht 4
2312 TW Leiden
The Netherlands

Acknowledgment

The author is indebted to Jacqueline Meulman for providing many helpful comments and suggestions that contributed significantly to the development of this paper.

Multidimensional Scaling by optimizing goodness of fit to a smooth hypothesis

Abstract

Multidimensional Scaling is the problem of representing n objects geometrically by n points, so that the interpoint distances correspond optimally to empirical proximities between objects. Existing approaches can be characterized by their particular definition of optimality, and such a definition in turn reflects what the hypothesized class of representations is. The fundamental hypothesis examined in this paper is that - apart from being monotonically related to proximities - distances are smoothly distributed. A measure of departure from smooth monotonicity is defined, including provisions for weights and tied data. Next the most important considerations for optimizing this measure are discussed. One of the consequences of the hypothesis of smoothness is that degeneracies cannot occur. Analyses of specifically selected, degeneracy-prone proximity tables show that the proposed method behaves satisfactorily in this respect.

Keywords

Multidimensional Scaling, monotone regression, ultrametric trees, nonmetric methods, degeneracy, departure from bimodality.

Introduction

The establishment of a spatial structure in which distance accounts for empirical proximity has become one of the central goals in the study of complex relationships. A spatial model can offer transparancy where an arbitrarily arranged array of numbers fails to do so; in addition, it presumably catches enduring trends and disregards what is accidental (Shepard, 1974). The basic notion of spatial modeling is to assign a point x_i to each and every empirical object o_i from some given collection O of objects (i=1,...,n). Depending on the method of selecting points and the way of assigning objects to points a number of different approaches can be distinguished. In order to obtain background for what follows, the situation is first considered as a falsification problem.

In a falsification framework the selection and assignment of points is done a priori, after which the specified model can be put into challenge against the data directly. For example, the investigator may use the Munsell color system if O is a batch of color patches, or may carry through a facet analysis (Guttman, 1959; Levy, 1981) if O is a set of psychological tasks, or he may choose as coordinate values for point i the positions on a number of controversial issues of O is a group of political actors. Let $d(x_i,x_j)$ denote the distance between points i and j in the spatial model, and $\delta(o_i,o_j)$ the empirically determined proximity between objects i and j (Note 1). Then the remaining data analytical problem is to ascertain the regression of δ on d. Under the rules of univariate statistics the hypothesized spatial model is claimed to be satisfactory as long as the proportion of variance of δ accounted for by d (or some function of d) is satisfactory.

Now suppose that the variance accounted for is not satisfactory, so that the a priori model is falsified, whereas there still is strong evidence that the

variability in proximity is systematic. Such evidence can be informal, e.g., a subject confidently rates or ranks $\delta(o_i,o_j)$ higher than $\delta(o_k,o_\ell)$ for many of such pairs of pairs. It could be statistical as well; e.g., replicated observations of proximity are available for all pairs i,j and a test like the Kruskal-Wallis multi-sample test indicates that these samples do not come from the same population. In response to the rejection of $d(x_i,x_j)$ as a predictor, one might of course try to come up with an independently conceived alternate set of points $y_1, \ldots, y_n$ and repeat the procedure with $d(y_i,y_j)$. The prospects of such an endeavour are not favorable, however, since presumably the best thoughts and experiences had already been spent on the first proposal. Therefore, suppose we ask guidance from the data, and consider the idea of trying to construct a spatial model by capitalizing, to a reasonable extent, on δ's variability. It is this notion that is the central feature of Multidimensional Scaling (MDS), and different specifications of what is regarded as a satisfactory data driven construction define different MDS methods. Among these, the nonmetric approach asks as little as possible from the data, relying on rank orders only, and this will be our point of departure as well.

The first rigorous nonmetric MDS specification was formulated by Kruskal (1964a,b), and runs as follows. By hypothesis there exists a configuration $x_1, \ldots, x_n$ in p dimensions with distances d that are monotonically increasing with δ. Define γ as any set of numbers (not necessarily distances) that satisfies monotonicity with δ, and denote the set of all those sets of numbers by Γ. Thus $\gamma \in \Gamma$ implies

$$\gamma(o_i,o_j) \geq \gamma(o_k,o_\ell) \quad \text{if} \quad \delta(o_i,o_j) > \delta(o_k,o_\ell). \tag{1}$$

Then for any candidate configuration $x_1, \ldots, x_n$ it is possible to determine departure from monotonicity by evaluating the so-called STRESS function (Note 2)

$$\sigma(x_1, \ldots, x_n) = \min_{\gamma \in \Gamma} \left\{ \frac{\sum(\gamma(o_i,o_j) - d(x_i,x_j))^2}{\sum(d(x_i,x_j) - \bar{d})^2} \right\}^{\frac{1}{2}}. \qquad (2)$$

The evaluation of STRESS is formally equivalent to a monotone (or isotone) regression problem (Barlow et al., 1972), lacking however its statistical rationale (for it is not the intention to regard d as a random variable, and Γ generated by the data as a fixed predictor). The spatial model is now selected as the configuration that minimizes STRESS. Before anything else, there is an optimization problem to solve. If the monotonicity hypothesis is correct, the solution of the optimization problem is a $d \in \Gamma$ and STRESS becomes zero. If this is not achieved, we still end up with a configuration that is optimal in the sense of having least departure from monotonicity.

The distinction between falsification and optimization shows up very clearly if the former is cast in monotone regression form as well (this possibility was explored by Heiser and Meulman (1984), in the context of cross-validation). This time the a priori configuration provides a set Π of monotone functions, and $\pi \in \Pi$ whenever π satisfies

$$\pi(x_i,x_j) \geq \pi(x_k,x_\ell) \qquad \text{if} \qquad d(x_i,x_j) > d(x_k,x_\ell) \; . \qquad (3)$$

Only the ordinal relations derived from the model are at issue, and monotonicity of δ with respect to d can be examined by computing the statistic

$$\mu(\delta) = \min_{\pi \in \Pi} \left\{ \frac{\sum(\pi(x_i,x_j) - \delta(o_i,o_j))^2}{\sum(\delta(o_i,o_j) - \bar{\delta})^2} \right\}^{\frac{1}{2}}. \qquad (4)$$

Observe that μ is explicitly written as a function of the data, which are treated as a numerical variable. Assuming in addition that δ is normally distributed, a significance test for the related variance-accounted-for quantity

$1-\mu^2$ is available (Barlow et al., 1972, Chapter 3). Of course there should preferably be independent replications of δ for each i,j in order to comply with the rationale of this test. After the choice of the regression function there is nothing left to optimize.

The present study was inspired by the following difficulty in the nonmetric approach. The statistic μ can be unsatisfactorily high, in which case a specific spatial model is dismissed. But μ can never be "too low", at least not on the conceptual level. Of course σ can be unsatisfactorily high too, in which case the whole idea of a spatial model has to be reconsidered. However, in addition there are situations for which σ is low and yet $x_1, \ldots, x_n$ is unsatisfactory. Such solutions are called degenerate, and arise from a flaw in the specification - not from a technical deficiency. In the falsification approach the spatial model is fully under control, so that triviality is avoided implicitly. In the optimization approach it has to be avoided explicitly. For a specific proximity modeling problem there are oftentimes simple, perfectly natural, but nevertheless ad-hoc actions conceivable to escape from degeneration. This paper proposes a general strategy to reinforce the modeling process that adds as little as possible restraints to the basic principle of monotonicity. The additional principle is to assume that distances are smoothly distributed. After a short review of the conditions under which degeneracies are likely to occur, a definition of smooth monotonicity will be developed that resolves the inherent drawbacks of earlier approaches. Next a consistent treatment of tied data values is offered, followed by a discussion of the computational aspects of smooth monotone regression. Finally, the proposed MDS method is tested by analyzing two degeneracy-prone proximity tables.

Prototypical degeneracies under simple nonmetricity

Shepard (1974) has reported nine elementary four-point configurations in two dimensions, for which the six interpoint distances take on just two distinct values. He remarks: "Of the many strong degenerate two-dimensional configurations that I have obtained in the analysis of real and artificial data, regardless of the number of points all have taken one or another of the nine forms shown in Fig. 6. In cases of extreme degeneracy, apparently, the additional points usually collapse onto these same few vertices." (Shepard, 1974, 393-394). Thus degeneracy is conceived as a property of the distance distribution (in the sense that all mass is concentrated at two points), and not as a uniform type of configuration.

It is of interest to know under what conditions, in terms of the proximities, such concentrated distance distributions are to be expected to prevail. At least a partial answer to this question is given in the following proposition:

Proposition 1

> If δ is ultrametric, then there always exists a two-point configuration with STRESS equal to zero.

Proof

It is not difficult to give a constructive proof. An ultrametric δ corresponds with a structure in which the objects are the nodes of a binary tree. The highest level split in the tree defines two subsets for which all within distances are smaller than all between distances. Assign all objects in one subset to one point. Now all within distances are equal to zero, and all between distances are equal to a positive constant, so that monotonicity is satisfied and STRESS attains its minimum value. □

Proposition 1 does not imply that the two-point solution is unique; there may be other zero STRESS solutions, depending on further details in the specification. Yet it does give a sharp contrast with Holman's (1972) assertion that an ultrametric requires at least n-2 dimensions in Euclidean space. While in this idealized case optimizing σ results in an oversimplified model without warning, the oversimplification would be signalled by μ if the two-point model were advanced to predict the very same δ. Table 1 illustrates that real data

==========
Insert Table 1 about here
==========

can be close enough to being ultrametric to get a strong degeneracy. Actually, the distances in the upper triangular part of Table 1 correspond with a three-point solution; because a subset of a tree is again a tree, such a two-dimensional zero STRESS configuration can be understood by repeated application of proposition 1.

There is an important special case of the general MDS situation for which degeneracy is not a tendency towards extreme bimodality, but towards a concentrated unimodal distance distribution. This notorious special case is the Unfolding technique, which scales two distinct sets of objects jointly, and capitalizes on the rank order of the between-set proximities only. Depending on the precise definition of STRESS, the degeneracy again can take various forms in terms of the configuration (Heiser, 1981). Attempts to remedy the situation always focus on avoiding extreme unimodality. However, even if unimodality would be ruled out in the specification we must expect trouble, as is illustrated in the following example. Suppose n_1 judges give a complete rank order of n_2 options, and suppose their option of last choice is either o_A or o_B or o_C. Then regardless the selection of existing nonmetric loss functions and

regardless all remaining properties of the data there exists a perfect two-dimensional solution for any value of n_1 and n_2 (see figure 1). In this solution,

==========
Insert Figure 1 about here
==========

all judges who choose o_A (or o_B, or o_C, resp.) last are collapsed into one point X/A (or X/B, or X/C, resp.), at equal distance from all object points D, ... , Z except A (or B, or C, resp.). To appreciate that having precisely three options of last choice is not such a contrived situation as it may seem to be at first sight, notice that data obeying a one-dimensional Unfolding model perfectly will contain only two options of last choice. Indeed, in such a case a zero STRESS, one-dimensional, five-point solution always exists.

The existence of degeneracies, or - in practice - partial degeneracies, degrades the beautiful perspective put forward by the early originators of nonmetric MDS; i.e., the possibility to construct spatial models by merely requiring monotonicity. It is true that in the examples given the unsatisfactory solutions are not completely independent of the data. Nevertheless, if so little of the rank order information is effectively used in the construction, the resulting configurations will be of no practical or theoretical value. The question of how to ensure that the rank order information does get transferred to the model will be the subject of the next paragraph.

Smooth monotonicity.

There is nothing in the "quantitative" part of STRESS (i.e., the ratio of sums of squares) to prevent the interpoint distances from attaining only a very few values, even if we would switch from squares to absolute values or other weighting schemes. Therefore, what must be changed is the "qualitative" part

of the loss function, i.e. the definition of the set Γ. It is convenient to assume at this point that there are no ties in δ. Let $r_1, \ldots, r_m$ denote the ranked object pairs that would order the dissimilarties in strictly ascending order (Note 3). Monotonicity requirement (1) can now be written as:

$$\gamma(r_1) \leq \gamma(r_2) \leq \ \ldots \ \leq \gamma(r_m) \ . \tag{5}$$

Whenever there is no need to give explicit reference to the data, the simpler notational device will be used of subscripting γ in the prescribed order. The minimizing values of γ in the loss function are denoted by $\hat{d}(r_1), \ldots, \hat{d}(r_m)$ and are called pseudo-distances (cf. Kruskal, 1977; the somewhat unfortunate terminology "disparities" is also in use). To distinguish between γ and $\hat{d}$, the former are called here admissible pseudo-distances. In terms of the pseudo--distances a degeneracy is characterized by, for some s:

$$\hat{d}(r_1) = \hat{d}(r_2) = \ \ldots \ = \hat{d}(r_s) \leq \hat{d}(r_{s+1}) = \ \ldots \ = \hat{d}(r_m) \ , \tag{6}$$

and, in less severe cases, by a dominance of equalities over inequalities among $\hat{d}$. Obviously, (6) is not excluded by (5).

Before defining smooth monotonicity a number of earlier approaches will be discussed. They all focus on the steps between consecutive pseudo-distances. For s=1,...,m the admissible steps are defined as

$$\theta_s = \gamma_s - \gamma_{s-1} \ , \tag{7}$$

where $\gamma_0 = 0$. The possibility of equalities among the pseudo-distances can be excluded radically by introducing non-negative bounds α_s and β_s on the steps. Thus (5) is replaced by

$$\alpha_s \leq \theta_s \leq \beta_s \ . \tag{8}$$

This formulation gives both flexibility and arbitrariness: by using a priori

bounds the values of the pseudo-distances can be manipulated in a very detailed manner. Such a strategy might be attractive in one particular application, but as a general procedure so much freedom is left that it can hardly be called a "strategy". In Heiser (1981) some of the arbitrariness was removed by defining

$$\begin{aligned} \alpha_s &= \alpha\{\delta(r_s) - \delta(r_{s-1})\} \ , \\ \beta_s &= 1/\alpha_s \ . \end{aligned} \tag{9}$$

So the bounds are a constant fraction of the consecutive proximity differences. Because α_s must be larger than β_s, the only freedom left is the choice of $0 < \alpha \leq 1$; for $\alpha \to 0$ ordinary monotonicity is approximated, whereas for $\alpha = 1$ the approach becomes completely metric. Quite satisfactory results were obtained for various Unfolding analyses with α in the range from .30 to .70. Moreover, the bounded monotone regression problem that must be solved repeatedly during the iterative minimization of the loss function is well understood, and presents no special computational difficulties (Reid, 1968; Barlow et al., 1972; Heiser, 1981).

The bounded approach with (9) has three major drawbacks: it is difficult to give a general rule for the choice of α; invariance of results under monotone transformations of δ is not assured, because the bounds are a function of the numerical values of δ; and the pseudo-distances are precluded from becoming equal. The last objection may come as a surprise, because equalities were designated earlier as a sign of degeneracy. Observe, however, that a number of regular configurations are conceivable - such as points equally spaced on a line, a grid or a circle - for which a considerable amount of equal distances must be anticipated. In terms of "possible configurations", generality is lost if all steps are constrained to be non-zero. Shepard (1974) has reported an approach which is very similar in spirit, and seems to avoid reliance on an external parameter. He proposed to use "convexity" (or "concavity") as an

additional constraint, presumably by requiring (Note 4):

$$\theta_1/\alpha_1 \leq \theta_2/\alpha_2 \leq \ldots \leq \theta_m/\alpha_m \qquad (10)$$

with α_s defined as in (9); (10) says that the piecewise slopes of γ as a function of δ should be non-decreasing. This requirement circumvents the first drawback (choice of α), but leaves the other two untouched, since an analysis which requires convexity of γ with respect to δ will generally be different from an analysis starting with $\phi(\delta)$, where δ is monotone. In addition, once a convex function starts increasing it can never become flat anymore.

The basic idea of the smooth approach proposed here is to control the acceleration of the pseudo-distances through an internally determined bound. This is achieved by requiring, for s=1,...,m:

$$| \theta_s - \theta_{s-1} | \leq \bar{\theta} , \qquad (11)$$

where $\bar{\theta}$ is the mean step $1/m \sum \theta_t$, and $\theta_0 = 0$. Thus each step must not deviate from the previous one by an amount larger than the mean step. Not the first order, but the second order differences are bounded, to such an extent that a considerable amount of nonlinearity is still allowed. For example, quadratically increasing values, or logarithmically increasing values, or the ordinates of the cumulative normal at equally spaced intervals all satisfy condition (11). We now do obtain invariance under monotone transformations of δ, for the only way δ enters into (11) is through the rank information defining θ_s (which in turn is based on the rank order of δ only). Also, some pseudo-distances may become equal, as long as there is enough variability amongst the others. Notice that if we regard γ (or $\hat{d}$) as a function of δ, there may be quantifications of δ for which the graph of the function does not look smooth at all. Smoothness - as used here - refers to the "non-lumpiness" of the pseudo-distances, con-

sidered as a distribution. The aim of MDS is to maximize the variance of this distribution, and smooth MDS requires in addition that there are no excessive concentrations of mass. The last point is illustrated by the two pseudo-distance distributions in Figure 2, resulting from an ordinary MDS and a smooth MDS analysis of the data in Table 1. The inherent bimodality in the data is shown

==========
Insert Figure 2 about here
==========

in the lower part of Figure 2 as an upside-down histogram, but it should again be emphasized that nonmetric MDS does not rely on this type of information. Apparently, bimodality is recovered in both analyses, but the smooth pseudo-distances obviously retain much more of the original rank order distinctions.

Before going further into the details and technicalities of implementing (11) in a MDS procedure, it is worthwhile to express the smoothness condition directly in terms of the admissible pseudo-distances:

$$-\frac{1}{m}\gamma_m \leq \gamma_s - 2\gamma_{s-1} + \gamma_{s-2} \leq \frac{1}{m}\gamma_m . \qquad (12)$$

Whenever a set of γs satisfies (12) this will be denoted by $\gamma \in \Gamma_S$. It is evident now that the smoothness condition implies that the distance of γ_{s-1} from the midway interpolated value $\frac{1}{2}(\gamma_{s-2}+\gamma_s)$ must be less than $(1/2m)\gamma_m$. Perhaps it can be argued that there remains some arbitrariness in choosing the bounding value $(1/2m)\gamma_m$ (or, for that matter, $\overline{\theta}$ in (11)). A number of try-outs on a wide range of examples suggest that we cannot use something very much larger without running the risk of admitting degeneracies again. And of course it is not desirable to make it smaller, for then the case of equal steps would be excluded (in which the first step equals the mean step).

Finally note that it is not true that Γ_S is contained in Γ; so the optimal

regression function must be selected from $\Gamma \cap \Gamma_S$. This combined set is homogeneous, i.e. if $\gamma \in \Gamma \cap \Gamma_S$, then also $a\gamma \in \Gamma \cap \Gamma_S$ for any scalar $a \geq 0$. It is also closed under addition, so it is in fact a convex cone. But the constant vector, i.e. the m-vector with all elements equal to unity, is not contained in Γ_S, so that the regression is no longer as rich as linear (or polynomial) regression with an intercept. Not having the constant vector in the set of admissible pseudo-distances is in accordance with the basic MDS assumption that not all proximities are equal.

Smooth monotone regression and the treatment of ties.

Now that the qualitative part of STRESS has been redefined, let's return to its quantitative part. Of primary interest is the numerator: the difference between distances and pseudo-distances, measured as a residual sum of squares. In the present subproblem the elements of d are regarded as fixed quantities, and therefore it is possible and convenient to make a switch from double subscripted summations to vector notation, as anticipated in the previous section through the introduction of the record-keeping ranked object pairs $r_1, \ldots, r_m$. Besides, the formulation will also be valid for the ALSCAL loss function called SSTRESS (Takane et al., 1977), which involves a residual sum of squares in terms of the squared distances. Even more generality is obtained by introducing a set of non-negative weights $w(r_s)$, assembled in a diagonal matrix W. The smooth monotone regression problem is to find a $\hat{d}$ that minimizes the function

$$\tau(\gamma) = \| \gamma - d \|_W^2 = \mathrm{tr}(\gamma - d)'W(\gamma - d) \qquad (13)$$

over all γ in the cone defined by smooth monotonicity. In some situations a slightly different version of the problem is needed, defined by the normalized loss function

$$\tau^*(\gamma) = \| \gamma - d \|^2_W \; / \; \| \gamma \|^2_W \; . \qquad (14)$$

It is not difficult to show that the solutions of the two problems are identical up to the length of the solution vector (Kruskal and Carroll, 1969). If $\hat{d}$ solves the unnormalized problem, then $(d'd/\hat{d}'d)\,\hat{d}$ solves the normalized one. The only condition for this result to be true is the homogeneity of the cone of regression functions. Other normalizations - such as on the variance of γ - are more problematical in the present case, because $\Gamma \cap \Gamma_S$ does not contain the constant vector.

The assumption that the ordinal information is a simple order will now be relaxed. Suppose some of the proximities are equal (or are regarded as equal for the sake of the MDS analysis). The ranked object pairs can be grouped into tieblocks $T_1, \ldots, T_q, T_{q*}, \ldots, T_b$. The tieblocks form a partition of $r_1, \ldots, r_m$ into b ordered classes, each having n_q elements. The simple order on the blocks induces a partial order on the admissible pseudo-distances; for $r_s \in T_q$ and $r_t \in T_{q*}$ we have

$$\gamma(r_s) \leq \gamma(r_t) \quad \text{if} \quad T_q \text{ precedes } T_{q*} \qquad (15)$$

When (15) is the only requirement imposed, pseudo-distances within tieblocks may become unequal; Kruskal (1964a) called this the primary approach to ties, because it seemed the most natural one to him. Sticking to the convention to indicate a simple ascending order by attaching subscripts to γ, there evidently is a number of ways to assign the object pairs to subscripted γs; for within each block the object pairs can be arbitrarily ranked in $n_q!$ ways, so that there are $\Pi n_q!$ assignments consistent with (15). Therefore the complexity of the regression problem is enlarged. Before dealing with this combinatorial complication in detail, a simpler treatment of ties will be discussed first.

In the so-called secondary approach to ties the pseudo-distances within tieblocks are simply required to be equal. Thereby the problem of assignment is removed, only the issue of approximation remains. Suppose the common within-tieblock values are collected in the b-vector γ_T, and let the m×b matrix G be defined by $g_{sq} = 1$ if r_s is in tieblock T_q, and $g_{sq} = 0$ otherwise. Thus G has row sums equal to unity and column sums equal to n_q. Then the smooth monotone regression problem with secondary approach to ties is:

$$\min_{\gamma_T \in \Gamma \cap \Gamma_S} \| G\gamma_T - d \|^2_W . \tag{16}$$

The within tieblock ordering of d and W is immaterial. Moreover, we can reduce problem (16) to an ordinary smooth monotone regression problem of size b by splitting the residuals into two orthogonal parts. Let $\bar{d}_T = (G'WG)^{-1}G'Wd$ (because of the special structure of G, these are simply the weighted mean distances for each tieblock). Then the residual sum of squares can be decomposed as

$$\begin{aligned} \| G\gamma_T - d \|^2_W &= \| (G\bar{d}_T - d) + (G\gamma_T - G\bar{d}_T) \|^2_W \\ &= \| G\bar{d}_T - d \|^2_W + \| \gamma_T - \bar{d}_T \|^2_{G'WG} . \end{aligned} \tag{17}$$

The first term on the right-hand side of (17) is the loss due to the equality constraints, whereas the second term measures the loss due to departure from smooth monotonicity of the weighted mean distances $\bar{d}_T$. It is sufficient to minimize the second term of (17) in order to solve (16). If the number of tieblocks is small with respect to the number of distances an enormous gain in computational efficiency is accomplished this way.

When the equality constraints are dropped, the original primary approach has to be adjusted to stay in line with the basic idea of smooth monotonicity. For without precautions the majority of non-zero steps might be found within tieblocks, and a minority (or even none) between tieblocks. In terms of the distri-

bution of $\hat{d}$ it would mean that the highest densities occur at locations where tieblocks meet. However, it is much more desirable, and consistent with the notion that the inequality information must be preserved, to have high density within groups of object pairs that share the same proximity. The within-tieblock variance should be small with respect to the total variance. Such a characteristic can be ensured as follows. The steps among consecutive values of γ are divided into two groups: a leading step extends from the largest element of a tieblock to the smallest element of the next tieblock; the other steps are called successive. Now the definition of the smooth primary approach to ties involves an adjustment of (11) on three points:

a. Instead of the mean step $\bar{\theta}$, the mean of the leading steps is used: θ^*.
b. Leading steps are compared with the preceding leading step.
c. Successive steps are compared with the same step with which their preceding leading step is compared.

As an illustration, suppose we have the following small ordered sequence of tieblocks: $\{r_1\}$, $\{r_2,r_3\}$, $\{r_4\}$. By an assignment is meant that

$$\begin{aligned} \text{either} \quad & \gamma_2 = \gamma(r_2) \quad \text{and} \quad \gamma_3 = \gamma(r_3), \\ \text{or} \quad & \gamma_2 = \gamma(r_3) \quad \text{and} \quad \gamma_3 = \gamma(r_2), \end{aligned}$$

and for both cases $\theta_1 = \gamma_1$, $\theta_2 = \gamma_2 - \gamma_1$ and $\theta_4 = \gamma_4 - \gamma_3$ are the leading steps, whereas $\theta_3 = \gamma_3 - \gamma_2$ is the successive step. For whatever assignment (which one is optimal depends on the distances), the smooth primary approach requires:

$$\begin{aligned} \theta_1 &\leq \theta^* \\ |\theta_2 - \theta_1| &\leq \theta^* \\ |\theta_3 - \theta_1| &\leq \theta^* \\ |\theta_4 - \theta_2| &\leq \theta^* \end{aligned}$$

where $\theta^* = (\theta_1 + \theta_2 + \theta_4)/3$. To indicate that a slightly different cone is at issue the notation $\gamma \in \Gamma_{SP}$ will be used for the adjusted smoothness conditions.

The next result provides the key to simplification of the original regression problem (see de Leeuw, 1977a, for a proof of the correctness of Kruskal's (1964b) way of handling the situation; the slightly different proof given here is valid for both the ordinary and the smooth primary approach). Suppose $\hat{d}$ is the optimum vector for minimizing $\tau(\gamma)$ over $\gamma \in \Gamma \cap \Gamma_{SP}$; consider any two elements r_s, r_t from the same tieblock T_q.

Proposition 2

If $d(r_s) < d(r_t)$ then $\hat{d}(r_s) \leq \hat{d}(r_t)$.

Proof

Suppose $\hat{d}(r_s) > \hat{d}(r_t)$. Then define $\bar{d}$ by $\bar{d}(r_s) = \hat{d}(r_t)$, $\bar{d}(r_t) = \hat{d}(r_s)$ and $\bar{d}(r_u) = \hat{d}(r_u)$ for all $u \neq s,t$. It is still true that $\bar{d} \in \Gamma \cap \Gamma_{SP}$, because $\bar{d}$ deviates from $\hat{d}$ only in assignment of object pairs. The loss difference is

$$\tau(\hat{d}) - \tau(\bar{d}) = \{w(r_s) + w(r_t)\}\ \{d(r_t) - d(r_s)\}\ \{\hat{d}(r_s) - \hat{d}(r_t)\}$$

Because the three factors on the right-hand side are all strictly larger than zero, we find $\tau(\hat{d}) > \tau(\bar{d})$, which contradicts the optimality of $\hat{d}$. Therefore, $\hat{d}(r_s) \leq \hat{d}(r_t)$. □

By repeatedly using this result a simple assignment rule is obtained: solve the minimization of $\tau(\gamma)$ with respect to a d and W that are (re)ordered within tieblocks in such a way that the numerical values of d are non-decreasing. This settles the primary approach.

Summarizing, there are two major ways of handling ties in smooth monotone regression: one in which the within tieblock pseudo-distances are kept close

together, the other in which they are kept exactly equal. For both there are simple preliminary operations to arrive at the same general type of computational problem, to be discussed in the next section.

Algorithmic considerations.

The general principles and tactics to minimize a function like STRESS are well-known and need not be discussed extensively here (cf. Kruskal, 1977; de Leeuw, 1977b). In all experimental runs so far, the so-called explicit normalization on the pseudo-distances was used; this implies that as soon as a new $\hat{d}$ has been found for some fixed configuration of points its sum of squares is set equal to a constant, and a new configuration is sought by making at least one step in the right direction for solving

$$\min_{x_1, \cdots, x_n} \sum w(o_i,o_j) \{ \hat{d}(o_i,o_j) - d(x_i,x_j) \}^2 . \tag{18}$$

Repeating this sequence of operations is bound to converge to a local minimum (de Leeuw and Heiser, 1977).

The subproblem posed by the smooth regression requires some special care, since it turns out that a straight-forward Alternating Least Squares (ALS) algorithm gets into trouble. Experimentation with such an algorithm was based on a parametrization of the problem in terms of the steps. Let L be the unit lower triangular matrix, and let S be its left-inverse: $SL = I$. The operator S takes successive differences, and L cumulates; in vector notation (7) becomes $\theta = S\gamma$, and $\gamma = L\theta$. The regression problem (13) becomes:

$$\min_{\substack{\theta_s \geq 0 \\ |\theta_s - \theta_{s-1}| \leq \bar{\theta} \\ \text{for } s=1,\ldots,m}} \| L\theta - d \|_W^2 . \tag{19}$$

Notice that the constraints on the step parameters are no longer "rectangular", as is the case in ordinary monotone regression (where there are only non-negativity constraints), or in bounded monotone regression (where (8) is the constraint set, which is exactly the definition of rectangular). An element-wise ALS algorithm (or cyclic-coordinate-descent method) proceeds by minimizing over each θ_s separately, while holding the θ_t $(t \neq s)$ fixed at their current values. If the constraints are rectangular such a procedure converges to the (global) minimum (D'Esopo, 1959). This result holds for a more general type of loss function as well (as long as it is convex), but not necessarily for the present type of constraint set. Indeed, the "jamming" phenomenon occurred several times in extreme cases, such as d equal to the constant vector (the sequence of solution points figuratively jams into a corner formed by the boundaries of the constraint set). Although the regression subproblem need not be solved completely for the master scaling program to remain convergent (de Leeuw and Heiser, 1977), local minima should be avoided whenever possible. Consequently, a somewhat more general quadratic programming strategy has to be considered.

The matter can be brought into standard quadratic programming form by switching from the minimization of $\tau(\gamma)$ to the equivalent problem

$$\max_{D\gamma \geq 0} \quad 2\,\gamma'Wd - \gamma'W\gamma\,, \tag{20}$$

where D is a $(3m-1)\times m$ coefficient matrix whose first m rows are defined by the monotonicity requirements (5), and the remaining rows are defined by the smoothness conditions (12). In the primary approach to ties, the coefficients are adjusted according to the specifications given in the previous section. By setting up the Lagrangean the Kuhn-Tucker conditions for problem (20) are readily derived. They imply that there must exist a $(3m-1)$ - vector of non-negative numbers λ such that the solution $\hat{d}$ can be written as:

$$\hat{d} = d + W^{-1}D'\lambda \ . \tag{21}$$

Defining $E = W^{-\frac{1}{2}}D'$ and $f = -W^{\frac{1}{2}}d$, the dual problem of finding the right λ is

$$\min_{\lambda \geq 0} \ \| E\lambda - f \|^2 \ . \tag{22}$$

Thus by solving the simpler non-negative least squares problem (22) the solution of the primal problem (20) is obtained immediately through (21).

After the initial failure of the direct ALS approach to (19), further experiments have all been done through solving (22) with the general non-negative least squares algorithm described in Lawson and Hanson (1974, ch. 23). This algorithm considers subsets of coordinates in a systematic way, keeping the others equal to zero, rather than mechanically cycling through the coordinates. It is possible to show that it converges in a finite number of steps. In this context it is interesting to note that Pool-Adjacent-Violators algorithms for ordinary monotone regression (cf. Barlow et al., 1972), including Kruskal's (1964b) Up-and-Down-Blocks version, can be viewed as alternate subset selection strategies for solving the dual (22). Moreover, since in that case D is exclusively the differencing operator S, which is bidiagonal, the subset minimizations are particularly simple and various shortcuts are possible. The present application took no account of the special characteristics of E, which is somewhat more complicated than S, but nevertheless still sparse (Note 5).

So there may be some room for improvement in the dual approach. It might also be that a smart feasible directions method to attack the primal problem directly is even better. For the number of variables in (20) is smaller, and - as was remarked earlier - it is sufficient to find slightly improved values for γ rather than optimal ones. Notice that it is not sufficient to make a few function decreasing steps in the dual, as relationship (21) does not

guarantee feasibility of $\hat{d}$ if λ is merely feasible, but non-optimal. If the dual is used, it has to be solved completely. These remarks are made because in the present set-up the computational burden of the smooth regression subroutine becomes overwhelming already for moderately sized MDS analyses (about 25 objects), except when there are many ties and the secondary approach is applied.

Illustrations: mutation distances and a perfect tree.

The distribution of smooth pseudo-distances in Figure 2 was obtained from the dissimilarity data in Table 1, which has been taken from Hartigan (1975). Originally, Fitch and Margoliash (1967) illustrated the construction of phylogenetic trees on the basis of these data. Dissimilarity between two species is defined as the mutation distance between certain protein molecules (Note 6). The smooth analysis has been performed with both the primary and the secondary approach to ties (number of tieblocks: 52). The optimal configuration of the primary analysis is given in Figure 3 (the configuration resulting from the secondary approach is virtually the same). Clearly, SKIN FUNGUS, BREAD YEAST and BAKER'S MOULD occupy remote positions, as they should according to the data, but in contrast to the degenerated ordinary nonmetric

==========
Insert Figure 3 about here
==========

MDS solution (three clumps) it remains possible to discern some phylogenetic structure. Figure 4 displays the pseudo-distances in relation to the original mutation distances. The gap in the middle extends only in the horizontal

==========
Insert Figure 4 about here
==========

direction (the data in numerical form); not in the vertical direction, due to smoothness. Notice that there are some flat parts and some rather large steps within tieblocks, but neither of these phenomena is of decisive importance in the analysis. The residual scatter plot (not shown) is homoscedastic.

Because the smoothness conditions will always cause Stress to be higher then it would be in an ordinary monotone analysis, a number of additional evaluative measures are helpful for assessing the quality of the model. Shepard (1974) suggested as a rough index of the non-degeneracy of a solution the ratio of the number of distinct distance values to the total number of distances. In practice, the problem with this index is that it will nearly always be equal to one due to the fallible precision and iterative character of the computations. Therefore it seems better to work with the percentage of distinct pseudo-distances, as these quantities are possibly made exactly equal during the computational process. Moreover, the pseudo-distances represent "the information extracted from the data", and their distinctness represents "the information retained". Table 2 list this index for a number of

==========

Insert Table 2 about here

==========

analyses; the ordinary monotone regression approach is indeed strongly disqualified.

Also given in Table 2 are the values of a measure called departure from bimodality. This measure is proposed here in order to be able to catch a degenerate shape of the pseudo-distance distribution with somewhat more subtlety (it was inspired by a similar index discussed in Hartigan, 1975, ch. 4.8). Departure from bimodality is defined as

$$\omega = \frac{\frac{1}{m}\sum (z_s - \bar{z})^2 - (\frac{1}{m}\sum |z_s - \bar{z}|)^2}{(\frac{1}{m}\sum |z_s - \bar{z}|)^2} \qquad (23)$$

for any set of quantities $z_1,\ldots,z_s,\ldots,z_m$ ($\bar{z}$ being their arithmetic mean). Of particular interest are the following properties of ω:

a. It is scale- and location invariant, and nonnegative.

b. It becomes zero for a symmetric distribution in which all mass is concentrated at precisely two points.

c. It converges quickly (with growing m) to 1/3 if the z_s are equally spaced.

d. It becomes large when the distribution is strongly unimodal.

In order to get a rough idea about the distribution of ω as a statistic for distances among points in intuitively "regular" configurations, a small Monte Carlo experiment has been performed. A hundred random configurations of 20 points in two dimensions were generated. The distribution of ω, computed for the 100 so obtained distance distributions, turned out to be approximately unimodal and symmetric with a mean of .458 and a standard deviation of .055. A tentative conclusion can be that configurations associated with an ω smaller than .35 or larger than .57 are special. Indeed, the degenerated solution gets the smallest value (not zero, the distribution is not symmetric). The other small values for the mutation distance analyses reflect that even the smooth solutions are still unusually dense on the right and sparse on the left (see Figure 3). Exceptionally large values of ω must be expected in the Unfolding situation, where degeneracy is a tendency towards all distances becoming equal.

Confirmation of the results of the analyses was sought by repeating them on a set of perfectly ultrametric dissimilarities, for a problem of the same size (see Table 3). An arbitrary tree was selected under the provision it were approximately balanced, i.e. the splits were always in about equal groups.

==========
Insert Table 3 about here
==========

Such a tree perhaps best reveals the prototypical MDS behaviour for "tree-like" data in general. The ordinary monotone regression analyses behaved as indicated in Proposition 1; this time with an ω very close to zero because the sets of "between" and "within" pseudo-distances are of about equal size (see Table 2). The metric analysis resulted in a rather reasonable configuration with respect to the small dissimilarities, but it did not represent the large dissimilarities very well; this discrepency is reflected in the poor value of Stress. A departure from bimodality of 1.2234 indicates that the ultrametric dissimilarity distribution is strongly unimodal (and very skew, of course). The configuration resulting from the smooth analysis with primary approach to ties is shown in

==========
Insert Figure 5 about here
==========

Figure 5. The clusters formed at different heights of the tree are embedded in the configuration as closed curves. Figure 6 displays the very strongly accelerating transformation of the rank numbers needed to get Stress as low as .005.

==========
Insert Figure 6 about here
==========

It is clear that the smooth spatial model preserves all the essential information from the tree in the sense that the complete hierarchy of nested clusters is represented. By contrast, the smooth secondary approach preserves less of the fine structure: all subdivisions below level 14 disappear, so that 6 clusters remain at locations that one would expect from inspection of

Figure 5. The percentage of distinct pseudo-distances becomes poor, but the Stress value is not really bad (.084).

In conclusion, it turns out that smooth two-dimensional spatial models for trees or tree-like data can be constructed without much Stress. The smoothness conditions effectively prevent the construction from collapsing. Especially encouraging are the good results with the adjusted primary approach to ties; for the insistence on exact equality within tieblocks - as is done in the secondary approach - seems to be at variance with the mere reliance on inequalities between tieblocks. After all, the proximity information is usually empirical, and therefore it will almost always carry a trace of uncertainty.

Discussion.

The occurrence of undesirable irregularities in nonmetric spatial modeling, in particular of degeneracies, has often been interpreted as a major weakness of nonmetricity. However, not nonmetricity itself is to blame, but the way it is implemented. The work reported in this paper shows that a fully nonmetric method can be devised that does not suffer from the most obvious degenerating tendencies. Apart from the advantage of protection against triviality there might be other benefits in terms of stability and improved local precision. A lot more experience is needed before anything can be established with certainty on that point; the expectation is that in many situations the difference between ordinary nonmetric and smooth MDS will be small, especially when n grows. Until now there seem to be two clear exceptions to this rule: the case of proximity data with the characteristics of an ultrametric, emphasized in the present paper, and the Unfolding case, which will be dealt with in a forthcoming paper. In a broader perspective, smoothness might be profitably introduced in Nonmetric Principal Components analysis and Canonical Correlation analysis as well.

Of course there are other ways to strengthen the modeling procedure. We can set up a parametric family of transformations for the data, and estimate the transformation parameters along with the estimation of the configuration. Within the MDS framework power transforms, low degree polynomials and negative exponentials are already in standard use (see Kruskal, 1977; Ramsay, 1982). Shepard's (1974) hints to obtain convexity or smoothness were also inspired by the idea of transforming a fixed set of datavalues. Winsberg and Ramsay (1983, and work cited therein) have introduced spline transformations for a variety of data analysis techniques. Splines are piecewise polynomials that are easy to manipulate; they can be constrained to be monotone, and they are almost by definition smooth. Yet none of these approaches enjoys the property that characterizes the present proposal: invariance of all results under monotonic transformations of the data. Or, to be more cautious, such invariance is a case to be proven continually for the other methods, whereas it holds by definition for the method proposed here.

Pushing this argument to the limit, it can be maintained that the optimization of goodness-of-fit to a smooth hypothesis does not involve the idea of some unique, optimal transformation of the data at all. It is true that after the analysis a plot can be made of the pseudo-distances against the data in some quantitative form. However, an ordinal treatment of the data implies that they could as well have been in another quantitative form - including discontinuities -, so that a second plot of the same pseudo-distances against an ordinally equivalent set of data values would show a different transformation. Thus the shape of the transformation plots can be entirely different, while the configuration remains the same. Recovering and comparing transformation shapes was one of the major objectives in Shepard's pioneering work (Shepard, 1962). In sofar as the (parametric) function fitting approach capitalizes on

the quantitative aspects of the data (polynomials, splines, or arbitrary convex functions all do this to a different degree), it can never claim to achieve the same generality as the truly nonmetric approach. A related, but maybe somewhat less essential point is that there is no such thing as a primary approach to ties when functions are fitted to the data. If tied data values indicate indifference, the idea to require equality of distance is quite radical, and may easily lead to a pessimistic judgement on scalability.

From a computational point of view, smooth MDS is still heavy and slow. Possibilities to speed things up are under investigation. If these attempts succeed, it will also become feasible to perform extensive sensitivity studies and to make smoothness a regular option in nonmetric MDS programs. A further point of investigation is the idea to drop explicit monotonicity alltogether, and to base the method on smoothness alone. Small and zero steps are already "unattractive", because they tend to diminish the size of the mean step, thereby reducing the freedom for taking larger steps when necessary. Negative steps will be even more unattractive in this sense, but they could improve the fit by allowing local departures from monotonicity.

Notes.

(1) Throughout the paper the Euclidean metric is assumed; some arguments will need slight modifications to remain valid for a wider class of metrics. Whenever there is no reason to distinguish dissimilarity from similarity the generic name proximity (Shepard, 1962) is adopted for any measure of confusibility, correlation, mutual animosity, and so on, that serves to give operational meaning to the concept of psychological distance.

(2) This is Kruskal's so-called Stress-formula 2, which has a variance term in the denominator. It was chosen for easy comparison with μ to be defined in (4). Choice of normalization is not critical for the argument at this stage.

(3) Thus r_1 is that particular object pair (o_i,o_j) associated with the smallest dissimilarity, r_2 the next one, and so on. For similarity data, the object pairs are ranked so that a strictly descending order would be obtained. Missing data are simply skipped, it being understood that m can be smaller than the full number of pairs $\frac{1}{2}n(n-1)$.

(4) This is only a guess, because Shepard does not give a detailed statement of the constraints used.

(5) In the MDS program, (22) must be solved repeatedly with only slightly different versions of f; therefore, the previous λ can be used to start the procedure, which gives about 50% reduction of execution time.

(6) In particular, the protein molecule cytochrome-c was used, and the mutation distance is defined as the minimal number of nucleotides that would need to be altered in order for the gene for one cytochrome to code for the other. Hartigan's table slightly deviates from the Fitch and Margoliash table in four of its entries.

REFERENCES

Barlow, R.E., Bartholomew, D.J., Bremner, J.M. & Brunk, H.D. (1972). Statistical Inference under Order Restrictions. New York: Wiley.

De Leeuw, J. (1977a). Correctness of Kruskal's algorithms for monotone regression with ties. Psychometrika, 42, 141-144.

De Leeuw, J. (1977b). Applications of convex analysis to multidimensional scaling. In J.R. Barra et al. (Eds.), Recent Developments in Statistics. Amsterdam: North-Holland, 133-145.

De Leeuw, J. & Heiser, W.J. (1977). Convergence of correction matrix algorithms for multidimensional scaling. In J. Lingoes et al. (Eds.), Geometric representations of relational data. Ann Arbor, Mich.: Mathesis Press, 735-752.

D'Esopo, D.A. (1959). A convex programming procedure. Naval Research Logistics Quarterly, 6, 33-42.

Guttman, L. (1959). A structural theory for intergroup beliefs and action. American Sociological Review, 24, 318-328.

Fitch, W.M. & Margoliash, E. (1967). Construction of phylogenetic trees. Science, 155, 279-284.

Hartigan, J.A. (1975). Clustering Algorithms. New York: Wiley.

Heiser, W.J. (1981). Unfolding Analysis of Proximity Data. Doctoral Thesis, University of Leiden, the Netherlands.

Heiser, W.J. & Meulman, J. (1984). Cross validation of geometric models by isotone regression. Research Report, Department of Data Theory, Leiden, the Netherlands.

Holman, E.W. (1972). The relation between hierarchical and Euclidean models for psychological distances. Psychometrika, 37, 417-423.

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

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

Kruskal, J.B. (1977). Multidimensional scaling and other methods for discovering structure. In K. Enslein et al. (Eds.), Statistical Methods for digital computers, Vol III. New York: Wiley, 296-339.

Kruskal, J.B. & Carroll, J.D. (1969). Geometrical models and badness-of-fit functions. In P.R. Krishnaiah (Ed.), Multivariate Analysis II. New York: Academic Press, 639-671.

Lawson, C.L. & Hanson, R.J. (1974). Solving Least Squares Problems. Englewood Cliffs, NJ: Prentice Hall.

Levy, S. (1981). Lawful roles of facets in social theories. In I. Borg (Ed.), Multivariate Data Representations: When & Why. Ann Arbor, Mich.: Mathesis Press, 65-107.

Ramsay, J.O. (1982). Some statistical approaches to multidimensional scaling data. J.Royal Stat. Soc. A, 145, 285-312.

Reid, W.T. (1968). A simple optimal control problem involving approximation by monotone functions. J. Optimization Theory and Applications, 2, 365-377.

Shepard, R.N. (1962). The analysis of proximities, I and II. Psychometrika, 27, 125-140 and 219-246.

Shepard, R.N. (1974). Representation of structure in similarity data: problems and prospects. Psychometrika, 39, 373-421.

Takane, Y., Young, F.W. & De Leeuw, J. (1977). Nonmetric individual differences multidimensional scaling: an alternating least squares method with optimal scaling features. Psychometrika, 42, 7-67.

Winsberg, S. & Ramsay, J.O. (1983). Monotone spline transformations for dimension reduction. Psychometrika, 48, 575-595.

TABLE 1

Mutation Distances[a]

Protein	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
MAN		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	62	62	62
MONKEY	1		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	62	62	62
DOG	13	12		0	0	0	0	0	0	0	0	0	0	0	0	0	0	62	62	62
HORSE	17	16	10		0	0	0	0	0	0	0	0	0	0	0	0	0	62	62	62
DONKEY	16	15	8	1		0	0	0	0	0	0	0	0	0	0	0	0	62	62	62
PIG	13	12	4	5	4		0	0	0	0	0	0	0	0	0	0	0	62	62	62
RABBIT	12	11	6	11	10	6		0	0	0	0	0	0	0	0	0	0	62	62	62
KANGAROo	12	13	7	11	12	7	7		0	0	0	0	0	0	0	0	0	62	62	62
PEKIN Duck	17	16	12	16	15	13	10	14		0	0	0	0	0	0	0	0	62	62	62
PIGEON	16	15	12	16	15	13	8	14	3		0	0	0	0	0	0	0	62	62	62
CHICKEN	18	17	14	16	15	13	11	15	3	4		0	0	0	0	0	0	62	62	62
KING PEnguin	18	17	14	17	16	14	11	13	3	4	2		0	0	0	0	0	62	62	62
SNAPPINg turtle	19	18	13	16	15	13	11	14	7	8	8	8		0	0	0	0	62	62	62
RATTLESnake	20	21	30	32	31	30	25	30	24	24	28	28	30		0	0	0	62	62	62
TUNA	31	32	29	27	26	25	26	27	27	27	26	27	27	38		0	0	62	62	62
SCREWWOrm fly	33	32	24	24	25	26	23	26	26	26	26	28	30	40	34		0	62	62	62
MOTH	36	35	28	33	32	31	29	31	30	30	31	30	33	41	41	16		62	62	62
BAKER'S mould	63	62	64	64	64	64	62	66	59	59	61	62	65	61	72	58	59		62	62
BREAD Yeast	56	57	61	60	59	59	59	58	62	62	62	61	64	61	66	63	60	57		0
SKIN FUngus	66	65	66	68	67	67	67	68	66	66	66	65	67	69	69	65	61	61	41	

[a]Values in the lower left half of the table are the mutation distances as determined from aminoacid sequences. Values in the upper right half of the table are reconstructed distances found by a nonmetric MDS method.

TABLE 2

Analysis Results for Ordinary and Smooth MDS

Analysis	Mutation distances			Perfect ultrametric		
	Stress	Percentage distinct pseudo distances	Departure from bimodality	Stress	Percentage distinct pseudo distances	Departure from bimodality
Monotone	.000	1	.2429	.000	1	.0001
Smooth primary	.068	94	.3364	.005	99	.3883
Smooth secondary	.076	26/94[a]	.3165	.084	4/37[a]	.4049
Metric	.135	27/100[a]	.3776	.274	10/100[a]	1.2234

[a]The first figure is based on the total number of elements (190), the second one on the number of tieblocks (52 for the mutation distances, 19 for the perfect ultrametric).

TABLE 3

Perfect Ultrametric[a]

Terminal node	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1		0	0	0	11	11	11	29	29	29	29	29	41	41	41	41	41	41	41	41
2	5		0	0	11	11	11	29	29	29	29	29	41	41	41	41	41	41	41	41
3	13	13		0	11	11	11	29	29	29	29	29	41	41	41	41	41	41	41	41
4	13	13	4		11	11	11	29	29	29	29	29	41	41	41	41	41	41	41	41
5	16	16	16	16		0	2	29	29	29	29	29	41	41	41	41	41	41	41	41
6	16	16	16	16	8		2	29	29	29	29	29	41	41	41	41	41	41	41	41
7	16	16	16	16	14	14		29	29	29	29	29	41	41	41	41	41	41	41	41
8	18	18	18	18	18	18	18		0	0	0	0	41	41	41	41	41	41	41	41
9	18	18	18	18	18	18	18	11		0	0	0	41	41	41	41	41	41	41	41
10	18	18	18	18	18	18	18	12	12		0	0	41	41	41	41	41	41	41	41
11	18	18	18	18	18	18	18	12	12	6		0	41	41	41	41	41	41	41	41
12	18	18	18	18	18	18	18	12	12	7	7		41	41	41	41	41	41	41	41
13	19	19	19	19	19	19	19	19	19	19	19	19		0	5	5	19	19	19	19
14	19	19	19	19	19	19	19	19	19	19	19	19	3		5	5	19	19	19	19
15	19	19	19	19	19	19	19	19	19	19	19	19	15	15		0	19	19	19	19
16	19	19	19	19	19	19	19	19	19	19	19	19	15	15	10		19	19	19	19
17	19	19	19	19	19	19	19	19	19	19	19	19	17	17	17	17		0	0	0
18	19	19	19	19	19	19	19	19	19	19	19	19	17	17	17	17	2		0	0
19	19	19	19	19	19	19	19	19	19	19	19	19	17	17	17	17	9	9		0
20	19	19	19	19	19	19	19	19	19	19	19	19	17	17	17	17	9	9	1	

[a]Values in the lower left half of the table are the ultrametric distances from an arbitrarily selected hierarchical tree. Values in the upper left half of the table are pseudo-distances found by smooth MDS with secondary approach to ties.

FIGURE CAPTIONS

Figure 1. Unfolding degeneracy whenever there are three options of last choice.

Figure 2. Distribution of mutation distances (below, upside down), degenerated pseudo-distances (above, dashed) and smooth pseudo-distances (above, solid).

Figure 3. Smooth spatial model of mutation distances (primary).

Figure 4. Smooth pseudo-distances (primary) against mutation distances.

Figure 5. Smooth spatial model of a hierarchical tree (primary).

Figure 6. Smooth pseudo-distances (primary) against ultrametric distances.

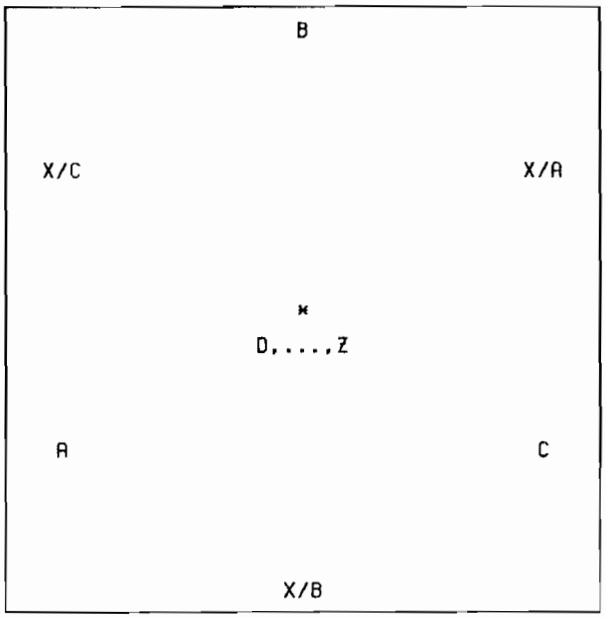

Figure 1.

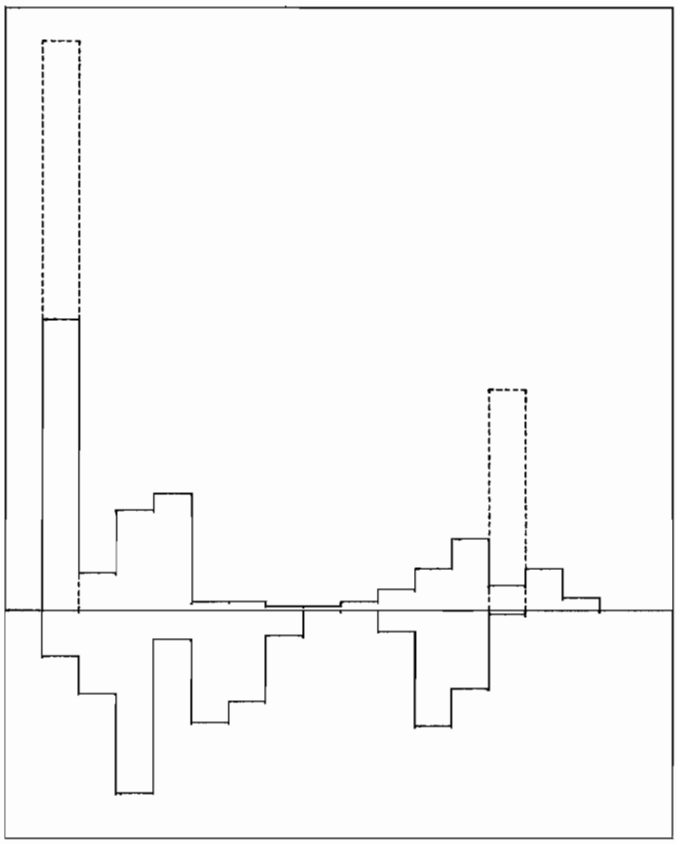

Figure 2.

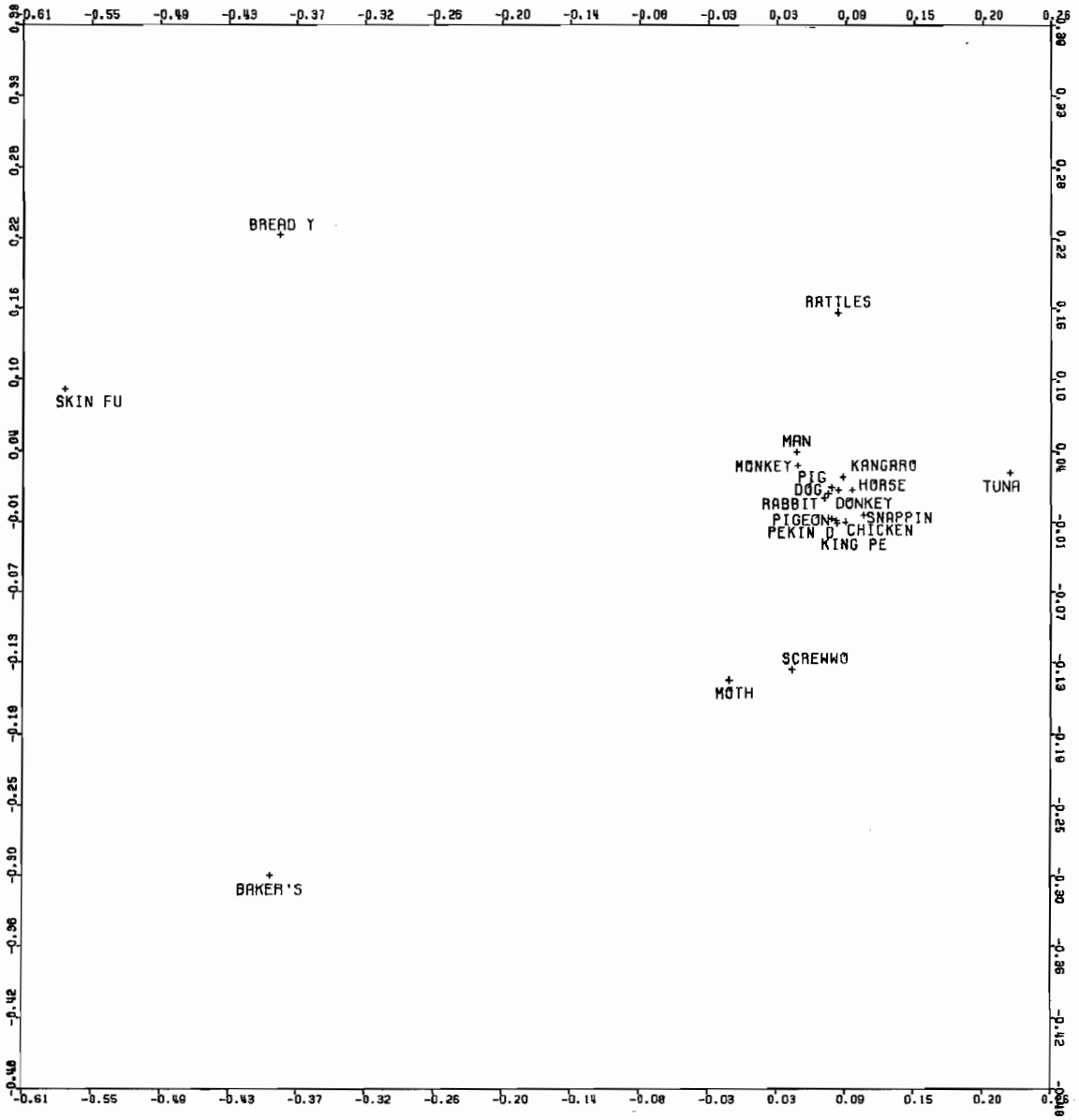
BREAD Y
RATTLES
SKIN FU
MAN
MONKEY
KANGARO
PIG
DOG
HORSE
TUNA
RABBIT
DONKEY
PIGEON
SNAPPIN
PEKIN D
CHICKEN
KING PE
SCREWWO
MOTH
BAKER'S

Figure 3.

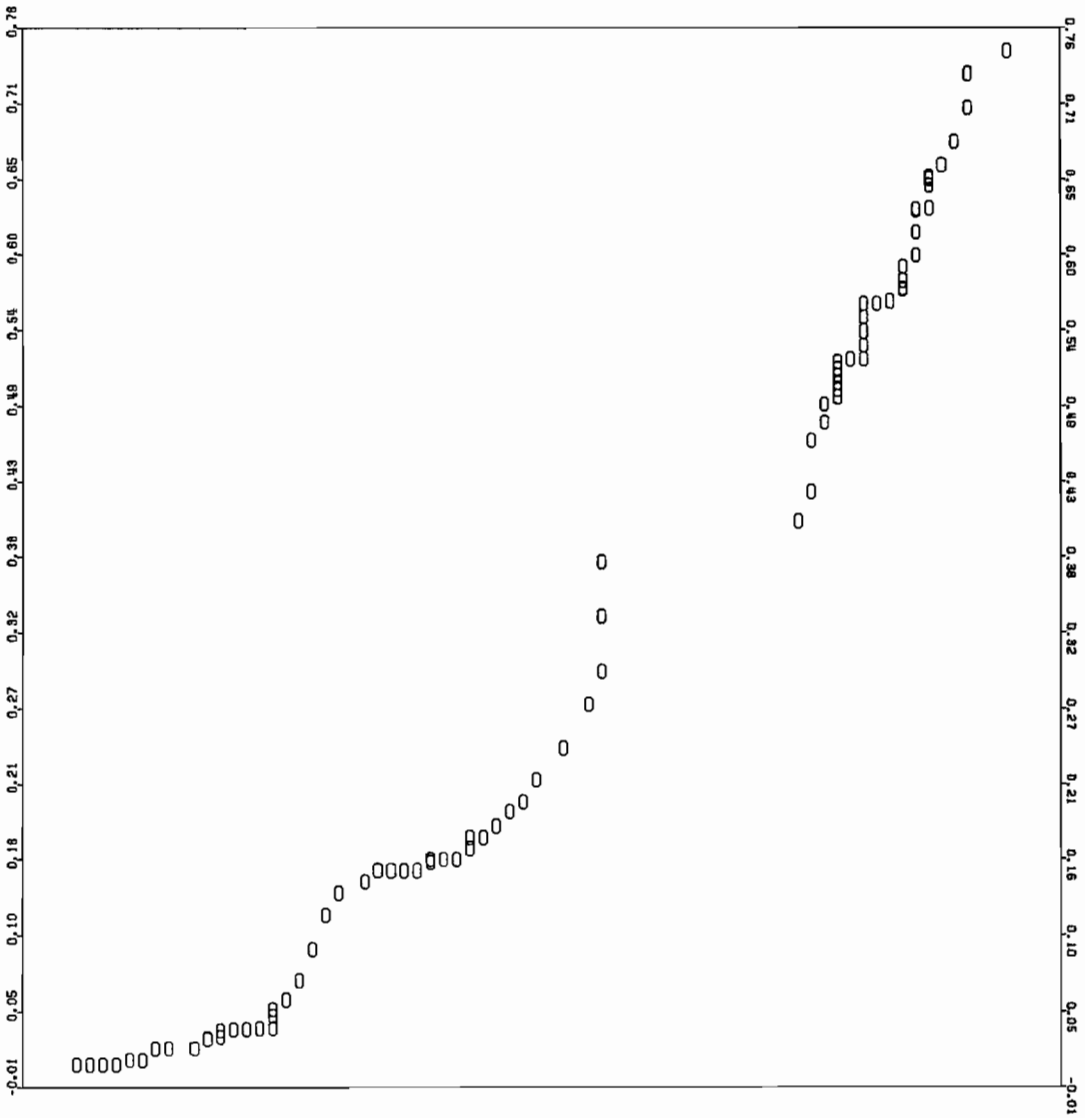

Figure 4.

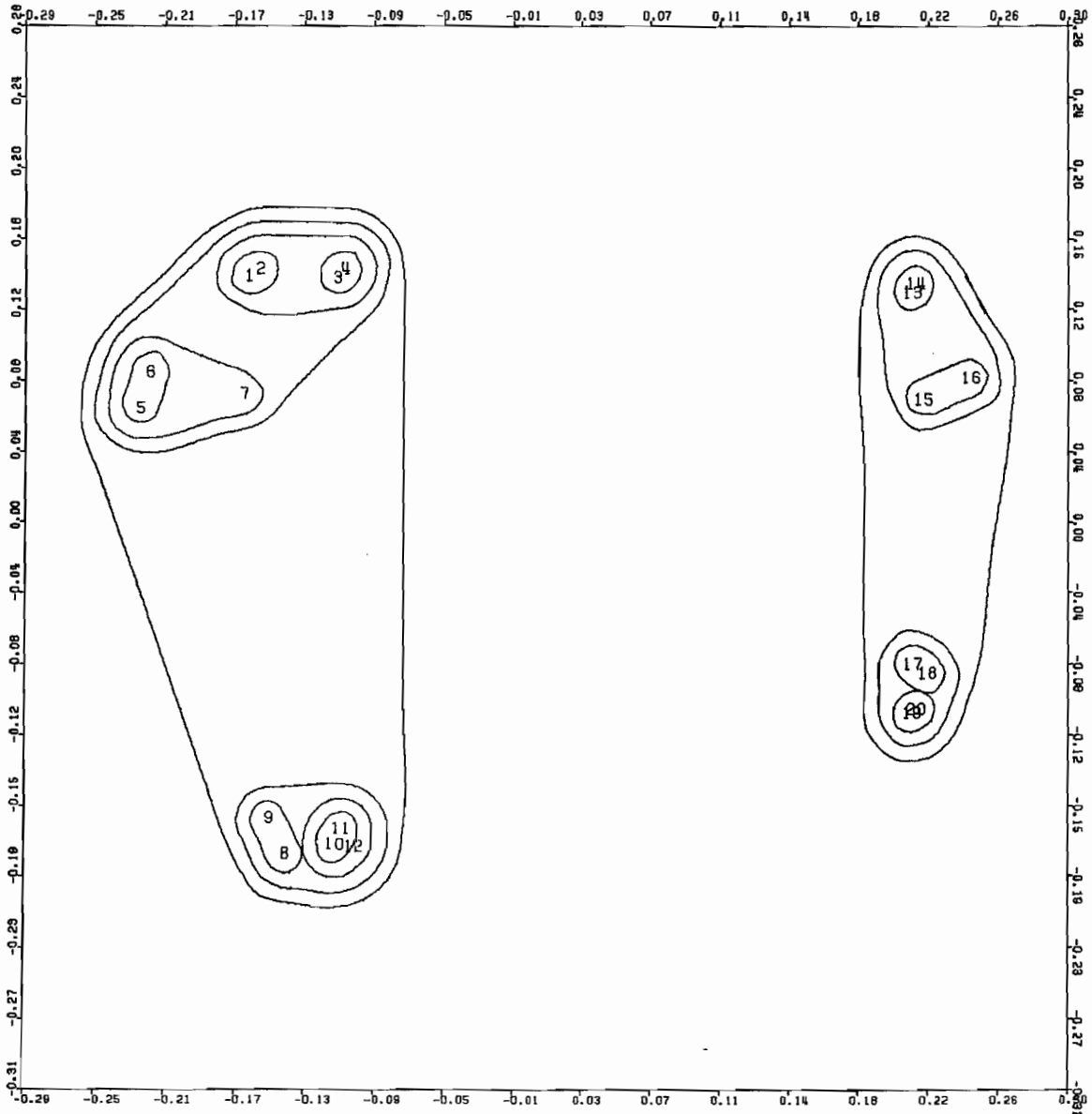
-0.29 -0.25 -0.21 -0.17 -0.13 -0.09 -0.05 -0.01 0.03 0.07 0.11 0.14 0.18 0.22 0.26
0.28 0.24 0.20 0.16 0.12 0.08 0.04 0.00 -0.04 -0.08 -0.12 -0.15 -0.19 -0.23 -0.27 -0.31
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Figure 5.

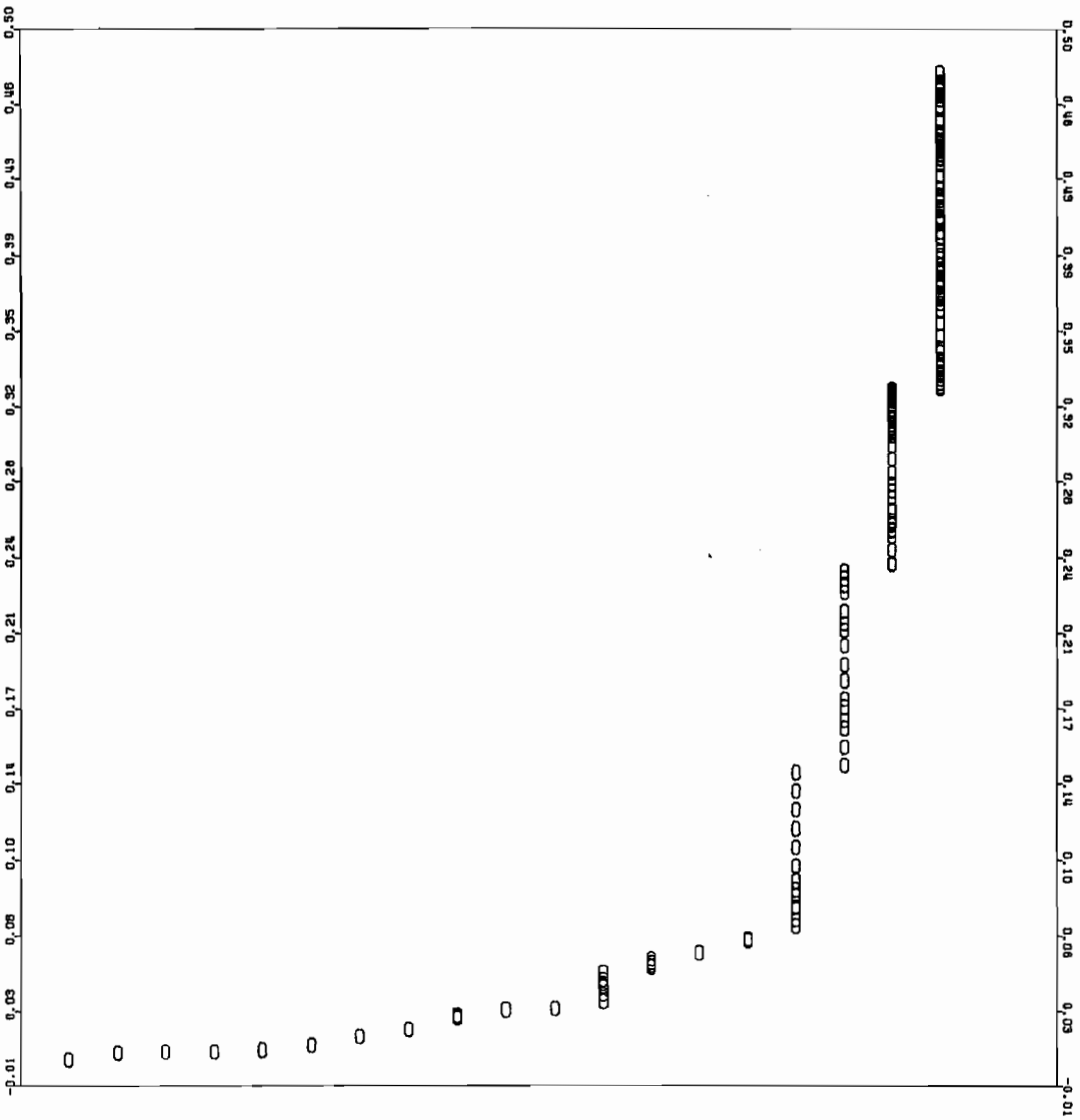

Figure 6.