The Formulation and Solution of Multidimensional Scaling Problems

Michael W. Trosset

November, 1993

TR93-55
(TR93-13 revised)

The Formulation and Solution of Multidimensional Scaling Problems

Michael W. Trosset*

November 17, 1993

Abstract

Numerous experiments in a variety of applied disciplines involve measuring distances between pairs of objects. The statistical problem posed by such experiments is that of fitting the observed data with a model defined to be the Euclidean distances between an abstract configuration of points. Techniques for solving this problem are collectively known as multidimensional scaling. These techniques have a long history in psychometrics and multivariate statistics, a much shorter one in the application of distance geometry to problems of molecular conformation. This review attempts to integrate these two traditions, which presently exist almost unaware of each other. Emphasis is placed on the rigorous formulation of the defining optimization problems, and on the computational practices that have been developed for solving these problems. Recent developments suggest that multidimensional scaling has entered a new and exciting era, as researchers begin to apply the tools of modern computational mathematics.

Key words and phrases: Multidimensional scaling, ordination, distance geometry, molecular conformation, principal coordinate analysis, computational statistics, graphical statistics.

*Adjunct Associate Professor, Department of Computational and Applied Mathematics, Rice University (email: trosset@rice.edu); Adjunct Associate Professor, Department of Psychology, University of Arizona (email: trosset@ccit.arizona.edu); and Consultant, P. O. Box 40993, Tucson, AZ 85717-0993.

1 Introduction

Multidimensional scaling (MDS) is a collection of data analytic techniques for constructing configurations of points from information about interpoint distances. This process is sometimes called ordination. An heuristic understanding of MDS may be acquired by considering a road map accompanied by a table of intercity driving distances. The table is merely a convenience, for it is usually but a simple matter to compute the intercity distances from the map itself. In contrast, it would be a formidable undertaking for a motorist possessing only the table to reconstruct the spatial relations between the cities. It is this inverse problem that is the province of MDS.

As early as Fechner[1] (1860) and Stumpf[2] (1880), psychologists became interested in matrix data sets that resembled the interpoint distances of a set of objects. Typically, each datum δ_{ij} represents a judgement about the dissimilarity of a pair of objects, e.g. colors, and it is desired to represent the objects as a configuration of points in a low-dimensional Euclidean space. Because it may not be possible to exactly realize an observed set of dissimilarities as interpoint distances in a Euclidean space of the specified dimension, the distinction between dissimilarities and distances is crucial.

Formally, a symmetric $n \times n$ matrix $\Delta = (\delta_{ij})$ is called a dissimilarity matrix if $\delta_{ij} \geq 0$ (nonnegative elements) and $\delta_{ii} = 0$ (zero diagonal elements). Instead of "dissimilarity matrix," some authors use the term "pre-distance matrix." Whatever the terminology, it is on such data matrices that MDS procedures operate.

From a given dissimilarity matrix Δ, an MDS algorithm constructs a configuration of points in a Euclidean space of specified dimension p. For a configuration $x_1, \ldots, x_n \in R^p$, the $n \times p$ configuration matrix X is the matrix whose rows are the $x_i^t, i = 1, \ldots, n$. From X it is easy to compute the Euclidean interpoint distance matrix $D(X) = (d_{ij})$, defined by

$$d_{ij}^2 = \sum_{k=1}^{n} (x_{ik} - x_{jk})^2. \tag{1}$$

The objective of MDS is to construct a configuration for which the interpoint distances d_{ij} closely approximate the given dissimilarities δ_{ij}.

For statisticians, a natural interpretation of the scenario just described is that dissimilarities are distances measured subject to error. MDS is then concerned with fitting a Euclidean distance model to a dissimilarity data set, e.g.

$$\delta_{ij} = d_{ij} + e_{ij}; \tag{2}$$

or more generally,

$$\delta_{ij} = g(d_{ij}, e_{ij}). \tag{3}$$

This is a very highly structured sort of regression problem: since the $m = n(n-1)/2$ observed dissimilarities and fitted distances are obtained from just n common objects, they enjoy a very complicated interdependent relationship with each other. Thus, the problems of MDS cannot be solved by general regression procedures — they require their own, highly specialized techniques.

In fact, the seemingly simple objective of MDS has resulted in a bewildering proliferation of such specialized techniques. The influential researcher J. de Leeuw (1988, pp. 163–164) began a recent article on MDS by expressing the following concern:

> "Recent research in multidimensional scaling has moved in the direction of proposing more and more complicated models, often with a very large number of parameters, and sometimes even with severe discontinuities in the model. The emphasis has been on producing computer programs that work, and comparatively little attention has been paid to theoretical problems associated with the loss functions and the algorithms used to minimize them. We think that such a more theoretical study is long overdue. In fact we think that at the moment an in-depth study of some

[1] The German scientist and philosopher G. W. Fechner invented *psychophysics*, the study of the quantitative relations between sensations and the stimuli that produce them. His monumental *Elemente der Psychophysik* (Fechner, 1860) founded the modern science of experimental psychology. One of Fechner's many contributions was the *method of paired comparisons*, which produces data describing the proportion of times stimulus i is judged greater than stimulus j.

[2] According to de Leeuw and Heiser (1982, p. 288), whose review includes an excellent history of MDS, Stumpf explicitly defined psychological distance as degree of dissimilarity, and made several important observations about the judgement of dissimilarity.

of the more simple models and techniques is more urgent than the development of even more complicated ones."

I share de Leeuw's concern, and accordingly will emphasize only the simplest, most fundamental problems of MDS. For example, this review restricts attention to *two-way* MDS, which addresses the case of a single dissimilarity matrix. It does not consider *three-way* (or *individual differences*) MDS, which addresses the case of multiple dissimilarity matrices. Nor does this review consider MDS procedures for constructing configurations in non-Euclidean spaces.

Recent developments in other disciplines have helped to inspire a renewed interest in certain fundamental aspects of MDS. As I will endeavor to illustrate, the theory and practice of MDS have historically advanced in response to dialogues between different disciplines (one of which was usually psychometrics). Over the last decade, the interest of various chemists in molecular conformation problems involving the measurement of interatomic distances has begun to attract the attention of the computational mathematics community. Although the optimization problems associated with MDS are purely deterministic, numerical analysts have never studied MDS, with the consequence that MDS techniques using state-of-the-art numerical algorithms have yet to be devised. Thus, a primary motivation for this work is to facilitate communication between psychometricians, statisticians, chemists, and computational mathematicians, about a subject that is (or should be) of mutual interest.

This review begins, in Section 2, with a detailed examination of several different applications of MDS. These applications were selected to help motivate why MDS has been formulated in certain ways, but they also sample its diverse uses. Sections 3–6 explore how MDS techniques actually fit distances to dissimilarities. Here the emphasis is on formulating MDS as a collection of well-defined optimization problems. These sections are organized according to the assumptions that are commonly placed on the dissimilarity matrix. Section 3 presents the fundamental problem of determining if a fixed dissimilarity matrix is a distance matrix, and of determining a configuration in the special case that it is. Section 4 presents the problem of approximating a fixed dissimilarity matrix with a distance matrix (the case of metric MDS), and introduces the standard objective functions for evaluating the approximation. Next, the dissimilarities are varied subject to bound constraints (Section 5) and order constraints (Section 6, the case of nonmetric MDS). The topics addressed in these four sections cover the vast majority of past research on these MDS problems; in contrast, Section 7 considers the relatively neglected topics of sensitivity analyses and probability models for MDS. The review concludes with a discussion of the role of MDS in modern multivariate analysis.

2 Applications of MDS

In recent years, chemists have developed a technique known as (2-dimensional) nuclear magnetic resonance (NMR) spectroscopy, from which can be obtained a matrix data set that contains measurements δ_{ij} of the interatomic distances in a molecule.[3] On the basis of such data, it is desired to represent the molecule as a configuration of atoms in R^3. Although such a configuration obviously exists, the measurements are subject to error and it may not be possible to exactly realize them as Euclidean interpoint distances.

There are several fundamental characteristics of the molecular conformation problem that distinguish it from more traditional applications of MDS. First, the desired configuration is a physically real entity, not the sort of hypothetical construct typically encountered in psychophysical experiments. Not only is it known that this entity exists in a Euclidean space, it is futher known that the dimension of that space is $p = 3$. Furthermore, it can plausibly be presumed that the measurement process, although subject to random error, does not systematically distort interatomic distances in the manner that human perception often distorts "psychological distances." Thus, the simple regression equation (2) suffices for modelling NMR data. This was the first MDS problem for which effective techniques were developed, by psychometricians in the 1950s.

In contrast to the ordination problems of molecular conformation, the ordination problems of psychophysics are concerned with locating perceived objects in a hypothetical psychological space. This raises several potentially troubling questions.

First, why should we assume that such a space is Euclidean? In an extremely provocative paper, Attneave (1950, p. 521) commented on "the exceedingly precarious assumption that psychological space is Euclidean in

[3] Details of NMR experiments can be found in the reviews by Wider et al. (1984), Wemmer and Reid (1985), Bax and Lerner (1986), and Wüthrich (1986).

its character." Indeed, many researchers have recommended non-Euclidean distance models.[4] Such work is beyond the scope of this review; however, in practice MDS is most commonly used to permit the visualization of the most obvious relations between the stimuli. Because very few individuals have a highly developed intuition about non-Euclidean geometries, it is often desirable to *represent* psychological space as Euclidean, even if that representation necessarily introduces some degree of distortion.

Second, what is the dimension of psychological space? Very rarely is even a theory available to guide the choice of p. Furthermore, as we shall see, it is often desirable to choose p larger than the number of psychological dimensions, in order to effectively represent a curvilinear configuration in Euclidean space. In practice, therefore, one typically applies an MDS technique to a set of problems indexed by different values of p. Often, the fit is evaluated and plotted for each value, and a value is selected for which the fit is dramatically better than for smaller values.[5] For some MDS techniques, rules of thumb exist to assist the user in deciding what constitutes an adequate fit. Historically, the desire to visualize the configuration has often limited the choice to $p \leq 3$. Following the emphasis of most MDS research to date, this review will ignore these concerns and focus on how to formulate and solve an MDS problem for a specified choice of p.

Third, does the simple regression equation (2) suffice when the measurement process is something as complex as human perception? The answer to this question appears to be an emphatic *No!*, and various attempts to understand how human perception functions played an extremely important role in shaping the development of MDS. There is a long psychophysical tradition of trying to discover functional descriptions of human perception. For example, Fechner's law states that the intensity of a sensation is a linear function of the logarithm of the strength of the stimulus. In contrast, Stevens (1957) has argued strongly on behalf of a *power law*, which states that the logarithm of the intensity is a linear function of the logarithm of the strength of the stimulus.[6]

It is certainly very tempting to apply either Fechner's law or the power law of Stevens to the modelling of dissimilarity data. In the case of Fechner's law, regression equation (3) becomes

$$\delta_{ij} = a + b \log d_{ij} + e_{ij}; \tag{4}$$

in the case of the power law, it becomes

$$\log \delta_{ij} = a + b \log d_{ij} + e_{ij}. \tag{5}$$

Equation (4) was used by Attneave (1950); equation (5) forms the basis of an entire theory of MDS, proposed by Ramsay (1977), to which we shall return in Section 7.

The use of Fechner's law in modelling dissimilarity data allows the statistician to account for the well-known fact that large stimulus distances are typically underestimated by human observers.[7] Alternatively, one might transform the dissimilarities in a way that exaggerates larger values, e.g. by replacing δ_{ij} with $\delta'_{ij} = \exp(\delta_{ij})$ in equation (2). This transformation was successfully used by Helm (1959), who discovered that large stimulus distances had been underestimated in a data set of color measurements. When he applied Torgerson's metric MDS procedure (see Section 4.1) to the actual dissimilarities he found twelve dimensions, but when he applied it to the exponentially transformed dissimilarities he found only two. Other researchers have discovered other relations in other data sets, although comparisons are difficult because different experiments used different measures of dissimilarity.

A very comprehensive program to determine the correct transformation from dissimilarities to distances was undertaken by R. N. Shepard.[8] In a series of papers based on his doctoral dissertation, Shepard (1957, 1958a,b) studied the phenomena of stimulus and response generalization in paired-associate learning

[4] Attneave cited the book by Householder and Landahl (1945), who were evidently the first researchers to study configurations in non-Euclidean spaces. They metrized R^p by endowing it with the L^1, rather than the L^2 (Euclidean) norm. MDS researchers have traditionally referred to L^1 distance as the "city block" metric. Arabie's (1991) superb review of MDS using the city block metric includes a very extensive bibliography.

[5] In the context of factor analysis, this rather subjective method for choosing a dimension is sometimes called screeing.

[6] The Harvard psychophysicist S. S. Stevens has written extensively on the superiority of the power law. For example, he has argued (Stevens, 1970) that it is a mathematical consequence of the biological properties of the neurons used for sensory processing. However, Stevens is better known to statisticians for his infamous taxonomy of data measurement scales (Stevens, 1946), recently critiqued by Velleman and Wilkinson (1993).

[7] According to de Leeuw and Heiser (1982), this tendency was first identified by Stumpf (1880).

[8] R. N. Shepard received his Ph.D. in psychology from Yale University in 1955. Three years later, he accepted a position at Bell Telephone Laboratories, where he was joined by J. B. Kruskal in 1959.

experiments.[9] Specifically, he sought to discover the functional form of the "gradient of generalization," the function that transforms the conditional probabilities of stimulus confusion into psychological distances. Shepherd (1958b, p. 255) concluded:

> "3. Under conditions of frequent and regular reinforcement (as in the typical paired-associate experiment), the gradient of generalization is closely approximated by an exponential decay function (concave upward).
>
> "4. Under conditions of infrequent or intermittent reinforcement, this gradient departs from the exponential function in that it is convex upward in the vicinity of the reinforced stimulus or response."

By the late 1950s, it had became apparent that the only acceptable generalization that could be stated about the relation between observed dissimilarity and psychological distance was that the function g in regression equation (3) is nondecreasing. This realization fundamentally altered the course of MDS research, deflecting its emphasis from parametric to *isotonic* regression equations. Historically, the former techniques have been termed *metric*, whereas the latter techniques have been termed *nonmetric*. Shepard (1962a,b) subsequently demonstrated the feasibility of constructing configurations of points from dissimilarities while requiring only a monotonic relation between distances and dissimilarities. Although his techniques were somewhat *ad hoc*, his examples remain among the most famous triumphs in the history of MDS. It will be instructive to examine one of these examples in some detail.

One of Shepard's most outstanding successes concerned an experiment performed by Ekman (1954), in which each of 31 subjects rated the similarity of every pair of 14 color stimuli, which varied primarily in hue. Each rating was on a five-point scale. The ratings were averaged over subject and scaled to range from 0 ("no similarity at all") to 1 ("identical"). This resulted in a 14×14 *similarity* matrix, which Ekman treated as a correlation matrix and subjected to factor analysis. He concluded that no less than five factors, which he identified as violet, blue, green, yellow, and red, were required to adequately account for variation in the empirical similarity measures.

In contrast, Shepard (1962b, p. 235) observed that, "Since variations in hue are usually represented by the familiar 'color circle' in two-dimensional space, Ekman's finding that five dimensions are required is rather surprising." He subjected the same similarity matrix to nonmetric MDS (notice that dissimilarities can be obtained from the similarities by subtracting each similarity from 1), and found that two dimensions sufficed to account for most of the variation. The configuration of 14 points constructed by nonmetric MDS in R^2 was almost perfectly circular, and a clockwise ordering of the points placed the colors in order of increasing wavelength. This is a beautiful example of a curvilinear configuration that can be effectively represented in a Euclidean space of higher dimension.

Finally, we consider the application of MDS to dissimilarities of objects known to be ordered in time. The mathematical reconstruction of an unknown chronology is a technique known as seriation.[10] Seriation appears to have been introduced by Petrie (1899), who attempted to sequence groups of 900 prehistoric Egyptian graves using characteristics of the pottery (and other artifacts) found in them. Hole and Shaw (1967) proposed seriation techniques that were predicated on systematically searching (subsets of) all possible orderings of the objects. As an alternative, Kendall (1969a,b) proposed the use of nonmetric MDS.

Following Kendall's (1971) more general review article, let $a_{ij} \geq 0$ denote the *abundance* of the jth *variety* (e.g. of pottery) in the ith object (e.g. grave). For example, a_{ij} might be either the number of occurences of pottery variety j in grave i, or the proportion of the contents of grave i that are pottery variety j. Define the similarity between graves i and j by

$$s_{ij} = \sum_h w_h \min(a_{ih}, a_{jh}),$$

[9] In a paired-associate learning experiment, there is a one-to-one assignment of "correct" responses $\{R_1, \ldots, R_N\}$ to stimuli $\{S_1, \ldots, S_N\}$. Each trial consists of the presentation to the subject of a stimulus and the elicitation of a response. *Stimulus generalization* refers to the tendency of the conditional probability of eliciting the erroneous response R_i to the presentation of stimulus S_j to increase with the similarity of S_i and S_j.

[10] Dunnell (1970) has emphasized that seriation is a formal scaling *technique*, not a *method* for establishing chronology. He reviewed the assumptions of the method, i.e. the conditions under which a chronology may be validly inferred from a seriation, in some detail. Our concern is indeed with MDS as a seriation technique.

where the $w_h > 0$ are arbitrary weights. (Of course, other definitions of grave similarity are also possible.) Then the similarity matrix $S = (s_{ij})$ is symmetric and diagonally dominant, and we need only convert the similarities to dissimilarities in order to apply MDS.

Different authors occasionally prefer somewhat different definitions of similarity. Mardia, Kent, and Bibby (1979) defined a similarity matrix $C = (c_{ij})$ to be any symmetric matrix for which the entries satisfy

$$c_{ii} \geq c_{ij} \geq 0.$$

They further defined the *standard transformation* from a similarity matrix C to a dissimilarity matrix Δ to be

$$\delta_{ij} = (c_{ii} - 2c_{ij} + c_{jj})^{1/2}.$$

If C is positive semidefinite, then dissimilarity matrices so obtained are necessarily Euclidean interpoint distance matrices for some $q \leq n$.

Assuming that the similarity matrix S has somehow been transformed into a dissimilarity matrix Δ, MDS can then be applied to obtain a configuration of points from which a chronology of the corresponding objects may possibly be inferred. Although it would seem natural to construct this configuration in a Euclidean space of dimension $p = 1$, experience has revealed that choosing $p = 2$ usually produces more useful results. The reason is that the typical configuration constructed in R^2 is not linear but U-shaped, a phenomenon sometimes called the "horseshoe effect."[11] This phenomenon occurs "because pairs of fairly remote graves will then be lumped together with pairs of extremely remote graves by reason of the fact that in general neither such pair of graves will have any 'varieties' in common." (Kendall, 1971, p. 225). Kendall observed that dissimilarities obtained from abundances were likely to resemble distances defined by a bounded metric and casually reported, "I am told (E. C. Zeeman: personal communication) that [the horseshoe effect] is to be expected on general Lie-group-theoretical grounds." (p. 227).

3 The Embedding Problem

For an arbitrary dissimilarity matrix Δ, the embedding problem of classical distance geometry is to determine if there exists an $n \times p$ configuration matrix X such that $D(X) = \Delta$. Cayley (1841) anonymously communicated a necessary condition for the case of $n = 5$ and $p = 3$, but it was not until the 1930s that the problem was completely solved. Perhaps the most elegant treatment was given by Schoenberg (1935). As a discussant of an article by Fréchet (1935) that addressed the question of when a linear metric space is isometric with a generalized Hilbert space, Schoenberg seized on developments in Fréchet's last section and demonstrated that the embedding problem could be solved by considering the properties of a certain quadratic form. He also noted that Menger (1931a,b) had already provided a characterization in terms of equations and inequalities involving certain determinants.

Three years later, Schoenberg's solution to the embedding problem was independently discovered by Young and Householder (1938), two young researchers at the University of Chicago.[12] Although they acknowledged Cayley's necessary condition, Young and Householder were apparently unaware that their necessary and sufficient condition was already known. It was, however, their good fortune to publish, not in the *Annals of Mathematics*, but in *Psychometrika*. In assessing who was influenced by whom, there can be no doubt that Young's and Householder's work is properly regarded as the mathematical foundation of MDS.

Young's and Householder's solution of the embedding problem produces configurations in which the nth point is placed at the origin. Torgerson (1952) subsequently observed that such configurations depend on the (presumably arbitrary) labelling of the points. He proposed a (now standard) construction that is independent of the labelling of the points, and which places the center of the configuration at the origin. Derivations of the following constructive solution to the embedding problem appear in the excellent textbooks by Mardia, Kent, and Bibby (1979) and Seber (1984).

[11] This terminology was so popular in the early 1970s that seriation by MDS was called the horseshoe method. It was institutionalized by Kendall's HORSHU program for seriation, which comprised several modifications and additions to the nonmetric MDS program MDSCAL.

[12] G. Young received his M.S. in physics in 1936 and remained as a research assistant in mathematical biophysics from 1936–40; A. S. Householder received his Ph.D. in mathematics in 1937 and remained as a Rockefeller Foundation Fellow from 1937–39. As we shall see, interdisciplinary discussions between various researchers at the University of Chicago played a crucial role in the early development of MDS.

Theorem 1 *Let $\Delta * \Delta$ denote the Hadamard product of Δ with itself, so that the elements of $\Delta * \Delta$ are the squared dissimilarities. Let $\tau : A \rightarrow B$ denote the linear operator defined by*

$$b_{ij} = -\frac{1}{2}(a_{ij} - \bar{a}_{i\cdot} - \bar{a}_{\cdot j} + \bar{a}_{\cdot\cdot}), \tag{6}$$

where we adopt the usual convention of averaging over the dotted subscripts. Then, the $n \times p$ configuration matrix X satisfies $D(X) = \Delta$ if and only if $\tau(\Delta * \Delta) = XX^t$, so that an appropriate configuration matrix exists if and only if $\tau(\Delta * \Delta)$ is a symmetric positive semidefinite matrix of rank $\leq p$.

The linear operators τ, often called double centering by psychometricians, and κ, the inverse operator that is defined by

$$a_{ij} = b_{ii} - 2b_{ij} + b_{jj},$$

were extensively analyzed by Critchley (1988). Because the existence of the mutually inverse operators τ and κ permits the straightforward recovery of configuration matrices from distance matrices (and vice versa), one useful consequence of Theorem 1 is that MDS problems can be parametrized by X or by D or by B, whichever matrix is conceptually (or computationally) more convenient.

Although Theorem 1 is of fundamental importance for MDS, it is not the only characterization of distance matrices on which MDS can be based. Distance (or metric) geometry comprises a considerable collection of results, representing a rich mathematical tradition that is most closely identified with K. Menger and L. M. Blumenthal.[13] When chemists became interested in problems of molecular conformation, it was to this tradition that they turned for assistance. For example, the remarkable monograph by Crippen and Havel (1988), *Distance Geometry and Molecular Conformation*, acknowledges a special debt to Blumenthal's (1953) classic *Theory and Applications of Distance Geometry*.

Much of the distance geometry that appears in the study of molecular conformation is used to manage *chirality* constraints (used to specify the sterochemical structure of the molecule) and is beyond the scope of this review.[14] Nevertheless, the study of molecular conformation has revived interest in the embedding problem and attracted the attention of researchers not previously familiar with MDS. Thus, Hayden, Wells, Liu, and Tarazaga (1991) obtained a new solution of the embedding problem, and recent work by Hayden and Tarazaga (1993) and Tarazaga, Hayden, and Wells (1993) has produced further insights into the geometry of distance matrices.

4 MDS With Fixed Dissimilarities

The first systematic method of psychological scaling was introduced by Thurstone[15] (1927a) for analyzing judgements about a set of stimuli. His *psychophysical analysis* presented a mathematical model for inferring a quantitative ordering of the stimuli from knowledge of their pairwise relations (obtained by Fechner's method of paired comparisons). In a subsequent expository article, Thurstone (1927b, p. 273) explained:

> "This attribute which may be assigned, as it were, in differing amounts to each specimen defines what we shall call the *psychological continuum* for that particular project in measurement."

[13] In the preface to his University of Missouri monograph, Blumenthal (1938, p. 3) attributed his "interest in abstract metrics" to "lectures that Karl Menger gave at the Rice Institute the spring of 1931" and to "the second year [1934–35] of [a National Research Fellowship] spent with Professor Menger at the University of Vienna".

[14] However, the problems that are considered in this review are fundamental. Crippen and Havel (1988, p. 343) observed:

> "As a general rule, in these calculations we have found it helpful to first minimize the distance error function, and then add on the chirality error function and further minimize the total error function. Although this can cause the optimizer to get trapped in local minima in certain idealized cases ..., in general the force of the chirality error function is large enough to ensure that all tetrahedra with the wrong chirality invert readily on minimization of the total error. The error with respect to the distance constraints alone, however, can be reduced more easily when the additional flexibility allowed by leaving out the error function C is present."

More general reviews of distance geometry methods for analyzing NMR data were provided by Braun (1987) and Havel (1991).

[15] From 1927, the writings of the University of Chicago psychologist L. L. Thurstone, which culminated in *The Vectors of Mind* (Thurstone, 1935), were instrumental in the development of psychometrics. Thurstone also helped to found the Psychometric Society, which began publishing *Psychometrika* in 1936.

Thurstone's judgement scaling method for psychophysical analysis is intrinsically unidimensional. This limitation was apparently first addressed by Richardson[16] (1938), at the 46th annual meeting of the American Psychological Association, in a talk titled *Multidimensional Psychophysics.* But although Richardson presented "the theory and method for extending the scaling of psychophysical judgments to n dimensions" (p. 659), *multidimensional scaling* was not immediately embraced by the psychometric community.[17] One reason for this neglect was the difficulty of deducing the dissimilarities from judgement experiments such as Richardson's *method of triadic combinations* or Klingberg's *method of rank order.* Another reason may have been the complete reliance of early MDS techniques on the embedding theorem of Young and Householder (1938), which requires the dissimilarities to be actual Euclidean distances.

In most psychometric applications of MDS, the user selects a low-dimensional Euclidean target space and does not seriously entertain the possibility that a configuration matrix can be found for which the interpoint distances exactly match the dissimilarities. Even with NMR data, which are obtained by measuring the interpoint distances of an actual 3-dimensional configuration, measurement error typically precludes the dissimilarities from being Euclidean distances. Thus, emphasis is properly placed not on solving the embedding problem, but on measuring how close one can come to solving it. Equivalently, the statistician tries to minimize the measurement error according to some criterion. This requires the specification of a loss function, which we will denote by $\rho(D, \Delta)$, that measures how well the interpoint distances approximate the dissimilarities.

Stated somewhat differently, we define metric MDS as techniques for solving optimization problems of the form

$$\begin{aligned} &\text{minimize} && \rho(D, \Delta) \\ &\text{subject to} && D \in \mathcal{D}_n(p), \end{aligned}$$

where $\mathcal{D}_n(p)$ is the set of all $n \times n$ matrices whose elements can be realized as the interpoint distances of n points in R^p. There are three popular choices of the loss (or objective) function ρ. The first choice, the STRAIN criterion, was derived from the tradition of distance geometry, specifically from the embedding problem discussed in Section 3. The other two choices, the STRESS and SSTRESS criteria, were motivated by the statistical tradition of fitting a model to data, as in regression equation (2). These choices will be described in the following two subsections.

4.1 STRAIN

The first MDS technique for *approximating* dissimilarities with Euclidean distances was published by Torgerson[18] (1952, p. 414), who emphasized that, "*We would probably never obtain a positive semidefinite B^* matrix with a rank less than the number of stimuli minus two from fallible data.*" Torgerson formalized the methods of Richardson (1938) and Klingberg (1941), identifying the "three basic steps" of MDS as (1) obtaining scale values from judgement data, (2) determining an additive constant that converts the scale values to dissimilarities,[19] and (3) constructing a configuration from the dissimilarities. However, his most fundamental contribution was the following observation:

> "This means that with fallible data the condition that B^* be positive semidefinite as a criterion for the points' existence in real space is not to be taken too seriously. What we would like to obtain is a B^*-matrix whose latent roots consist of

[16] The psychometricians M. W. Richardson and H. Gulliksen were employed by the Examiner's Office at the University of Chicago. Young and Householder (1938, p. 19) prefaced their solution to the embedding problem by stating, "This paper was written in response to suggestions by Harold Gulliksen and by M. W. Richardson. The latter is working on a psychophysical problem in which the dimensionality of a set of points whose mutual distances are available is a central idea."

[17] For more than a decade, the only application of Richardson's method was by Klingberg (1941). An excellent review was subsequently published by Gulliksen (1946).

[18] W. S. Torgerson received his Ph.D. from Princeton University in 1951. His subsequent article in *Psychometrika* was based on his dissertation, which was directed by M. W. Richardson's former colleague, H. Gulliksen. Gulliksen also chaired the Social Science Research Council's Committe on Scaling Theory and Methods, which engaged Torgerson to prepare a comprehensive monograph on psychological scaling (Torgerson, 1958).

[19] The so-called additive constant problem is beyond the scope of this review. The idea is to improve the fit of regression equation (2) by replacing δ_{ij} with $\delta'_{ij} = \delta_{ij} + a$. Although the subsequent shift of emphasis from metric to nonmetric MDS devalued the importance of the additive constant problem, a relatively recent analysis was undertaken by Saito (1978).

1. A few large positive values (the 'true' dimensions of the system), and
2. The remaining values small and distributed about zero (the 'error' dimensions)."

(p. 414).

Implicit in this observation is the notion of comparing D and Δ by comparing $\tau(D * D)$ and $\tau(\Delta * \Delta)$. This leads to the MDS objective function that J. D. Carroll has dubbed STRAIN,

$$\rho(D, \Delta) = \|\tau(D * D) - \tau(\Delta * \Delta)\|_F^2, \tag{7}$$

where $\|\cdot\|_F$ denotes the Frobenius norm, i.e. the L^2 norm on $R^{n \times n}$.

One striking advantage of using the STRAIN objective function is that an explicit solution to the metric MDS problem is then available in terms of the spectral decomposition of $B = \tau(\Delta * \Delta)$. A proof of the following result was given by Keller (1962).

Theorem 2 *Let $\lambda_1 \geq \cdots \geq \lambda_n$ denote the ordered eigenvalues of B and let $v_1, \ldots, v_n$ denote corresponding orthogonal eigenvectors. Let $\bar{\lambda}_j = \max(\lambda_j, 0)$. Then a configuration matrix X whose interpoint distances minimize the STRAIN objective function is obtained by setting the columns of X equal to $\bar{\lambda}_j^{1/2} v_j$ for $j = 1, \ldots, p$.*

Metric MDS using the STRAIN criterion is sometimes called the classical method of MDS, and is also called *principal coordinate analysis*. A very elegant relation between principal coordinate analysis and principal component analysis, in which a configuration of points in R^q is projected into the best fitting affine subspace of specified dimension p, was first established in a very penetrating paper by Gower (1966).

Suppose that a multivariate data set consists of q measurements on each of n objects, stored in the $n \times q$ data matrix Y. This data set represents a configuration of n points in R^q, so that Y is also a configuration matrix. By convention, multivariate techniques based on comparisons of the n objects are called Q techniques, and the $n \times n$ dissimilarity matrix $\Delta = D(Y)$ is an example of a Q matrix. Thus, principal coordinate analysis is a Q technique. Techniques based on comparisons of the q variables are called R techniques, and the $q \times q$ sample covariance matrix is an example of an R matrix. Principal component analysis is an R technique.

Gower (1966) defined a (Q, R) pair of multivariate techniques to be dual if they produce configurations of points having the same interpoint distances. He proved that principal coordinate analysis and principal component analysis are dual, a relation that had been overlooked by previous researchers. In fact, if Y has been centered at the origin, then principal coordinate analysis and principal component analysis produce the identical $n \times p$ configuration matrix X. Thus, metric MDS using the STRAIN criterion turns out to be a technique with which most statisticians are very familiar.

Despite its elegant optimality properties and computational advantages, the STRAIN criterion is currently not an extremely popular objective function for MDS. Carroll and Chang (1970) relied on STRAIN in their techniques for three-way metric MDS, but Takane, Young, and de Leeuw (1977) rejected it as the basis for three-way nonmetric MDS. Some of the objections to STRAIN are technical. For example, it has not been clear how to accommodate missing or variable dissimilarities. These issues will be addressed in subsequent sections; however, the primary objection to STRAIN appears to be simply that it is perceived as a less natural criterion than others that more directly compare distances and dissimilarities, in the spirit of regression equation (2).

4.2 STRESS and SSTRESS

A very natural choice for an MDS objective function is the criterion

$$\rho(D, \Delta) = \|D - \Delta\|_F^2.$$

This quantity is precisely twice the "residual sum of squares" obtained by fitting the m distances d_{ij} to the m dissimilarities δ_{ij} in regression equation (2). The use of the residual sum of squares as a measure of goodness of fit in the context of (nonmetric) MDS was proposed by Kruskal (1964a), who dubbed it the *raw stress*.

Actually, Kruskal did not endorse the use of raw stress as an objective function. He observed that this quantity is not invariant under dilations of a configuration, and therefore proposed dividing it by a scaling factor that would eliminate the dependence on scale. This device, which will be examined in some detail in Section 6.2, has come to be regarded as intrinsic to most approaches to nonmetric MDS. However, it is not appropriate for metric MDS, which seeks to find a configuration whose interpoint distances quantitatively approximate a fixed matrix of dissimilarities, and for which, therefore, it is necessary to distinguish between scaled versions of the same configuration. Consequently, and in accordance with current convention, I will indulge in a slight abuse of terminology and appropriate the term STRESS for the present objective function.

To date, all of the algorithms that have been proposed for minimizing STRESS parametrize the metric MDS problem in terms of the configuration matrix by writing $D = D(X)$. Thus, a constrained optimization problem with a relatively simple objective function is traded for an unconstrained optimization problem with a much more complicated objective function. This is an almost universal device in the MDS literature for managing the constraint that $D \in \mathcal{D}_n(p)$; however, as is evident from equation (1), it sacrifices the smoothness and the convexity of the STRESS objective function.

One popular way of recovering a somewhat more pleasant objective function is based on the observation that the interpoint distances of a configuration are not so simply derived from the configuration coordinates as are the squares of those distances. Accordingly, Obenchain (1971) and Hayashi (1974) proposed replacing the comparison of distances and dissimilarities with the comparison of squared distances and squared dissimilarities. This results in the MDS objective function

$$\rho(D, \Delta) = \|D * D - \Delta * \Delta\|_F^2.$$

The first effective use of this objective function was by Takane, Young, and de Leeuw (1977), who introduced the terminology SSTRESS. These authors were primarily concerned with nonmetric MDS, and they normalized SSTRESS in a manner analogous to Kruskal's (1964a) normalization of raw stress.

Most researchers not only use the parametrization $D = D(X)$, but also adopt the greater generality of weighted comparisons of distances and dissimilarities. Thus, a standard representation of the STRESS ($r = 1/2$) and SSTRESS ($r = 1$) objective functions is

$$\sigma_r(X) = \sum_{i<j} w_{ij} \left[\left(\sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \right)^r - (\delta_{ij}^2)^r \right]^2, \tag{8}$$

where the w_{ij} are the nonnegative weights. A particular advantage of this greater generality is that missing dissimilarities can be accommodated by setting the corresponding weights equal to zero. Because of the averaging effects of double centering, no analogous device is possible for the STRAIN criterion.

In contrast to the metric STRAIN problem, no explicit solution is known for the metric STRESS or SSTRESS problems,

$$\text{minimize} \quad \sigma_r(X). \tag{9}$$

Instead, configurations must be optimized by the application of iterative methods for finding (local) minimizers. Unfortunately, the straightforward application of standard numerical algorithms to Problem (9) is problematized by the fact that $D(X)$, hence $\sigma_r(X)$, is invariant under isometric transformations of the configuration. As discussed by Kearsley, Tapia, and Trosset (1993), the translational and rotational indeterminancy thus introduced by the reparametrization $D = D(X)$ results in connected sets of minimizing configurations. It therefore precludes the Hessian matrix of σ_r from being positive definite at minimizers, a condition strongly desired for algorithmic efficiency. Different MDS researchers have attempted to remove this indeterminancy in different ways, resulting in a plethora of numerical algorithms for the metric STRESS and SSTRESS problems.

Most of the algorithms for minimizing STRESS are based on a fixed point method proposed by Guttman (1968) and rigorously analyzed by de Leeuw (1977), de Leeuw and Heiser (1980), and de Leeuw (1988). The method is globally convergent, i.e. it converges from any initial configuration, to a connected set of local minimizers. Translational indeterminancy is easily removed by centering the initial configuration, a property that is preserved by the *Guttman transform* on which the method is based. De Leeuw's (1988) implementation of the method, sometimes called the *majorization algorithm*, removes rotational indeterminancy by rotating to principal components; however, convergence is still very slow — linear, with a constant near

unity. The slow convergence is not surprising, since this method is essentially a weighted gradient method. Glunt, Hayden, and Raydan (1993a,b) have attempted to improve on this gradient method. Their numerical experiments indicate that their *spectral gradient algorithm* has considerably faster local convergence than the majorization algorithm, despite the fact that it does not remove rotational indeterminancy. However, it is not known to be globally convergent.

The Guttman transform exploits the special structure of the STRESS criterion. At present, no analogous device is available for the SSTRESS criterion. For this reason, MDS researchers have historically proposed very different methods for minimizing STRESS and SSTRESS. Because σ_1 is smoother than $\sigma_{.5}$, there has been less reluctance to use second-order methods for SSTRESS than for STRESS; however, the devices used to remove indeterminancy have been extremely cumbersome. For example, Browne (1987) introduced a second configuration and penalized configurations for departing from it. This resulted in two sets of configuration coordinate variables and a more complicated objective function, which Browne minimized by the method of alternating least squares (see Section 6.3). Glunt, Hayden, and Liu (1991) removed translational indeterminancy by penalizing configurations for not being centered at the origin, minimizing the resulting objective function by a quasi-Newton method. Their numerical experiments indicate that their algorithm is considerably faster than Browne's, despite the fact that it does not remove rotational indeterminancy.

Each of the algorithms described in the preceding two paragraphs was extensively explicated and critiqued by Kearsley, Tapia, and Trosset (1993), who interpreted Problem (9) as a nonlinear least squares problem. To remove translational and rotational indeterminancy in a way that preserves this interpretation, they employed an elementary device that has been widely used in the literature on minimizing the Lennard-Jones potential energy of molecular configurations, e.g. by Hoare and Pal (1971). For $p = 3$, one simply constrains one point in the configuration to lie at the origin, specifies two coordinate axes, constrains a second point to lie along the first specified axis, and constrains a third point to lie in the plane determined by the two axes. Unless the selected points are collinear, this parametrization of Problem (9) removes both translational and rotational indeterminancy.

In general, the parametrization just described reduces Problem (9) to a nonlinear least squares problem with $N = np - p(p+1)/2$ free variables. For either $r = 1/2$ (STRESS) or $r = 1$ (SSTRESS), this problem can be solved reasonably efficiently by any good general nonlinear least squares algorithm, e.g. Moré, Garbow, and Hillstrom (1980) or Dennis, Gay, and Welsch (1981). However, Kearsley, Tapia, and Trosset (1993) were able to improve on the performance of general algorithms by exploiting several special characteristics of Problem (9), specifically that the nonlinearity of σ_r is relatively mild, that analytic derivatives are easily computed, and that the Hessian matrix is very dense. Their algorithm, a standard globalization of Newton's method, is extremely efficient for *both* the STRESS and SSTRESS problems.

It should be emphasized that all of the algorithms discussed in the preceding four paragraphs are methods for finding *local* minimizers. Naturally, we would prefer to find global minimizers, as we can for the metric STRAIN problem. Historically, MDS researchers have addressed this desire by trying to find a "good" initial configuration from which to start iterating, hoping that this will lead to a good local minimizer — perhaps even the global minimizer. Typically, an initial configuration is constructed that solves an easier MDS problem; for example, Browne (1987) used the solution of the metric STRAIN problem. Kearsley, Tapia, and Trosset (1993) discussed some of the issues involved in constructing an initial configuration.

Finally, there is an important distinction, which may influence the choice of an objective function, between optimal STRESS and optimal SSTRESS configurations. Notice that the STRESS criterion, $\sigma_{.5}$, is not differentiable at a configuration in which two points *coalesce* (have the same coordinates), since the distance between such points vanishes and the square root function is not differentiable at zero. However, de Leeuw (1984) established that $\sigma_{.5}$ *is* differentiable at every local minimizer. One consequence of this fact is that local derivative information is guaranteed to exist for the algorithms discussed above. A second consequence is that optimal STRESS configurations necessarily have distinct points, in contrast to optimal SSTRESS configurations in which points sometimes coalesce. For some applications, this property of STRESS may be decisive. In molecular conformation, in which it is physically impossible for the points (atoms) to coalesce, STRESS is indeed the preferred criterion, e.g. in Havel's (1991) DG-II package.

Nevertheless, the choice of STRAIN, STRESS, or SSTRESS needs to be informed by a better understanding of the consequences of adopting each. A decade ago, de Leeuw and Heiser (1982, p. 309) commented that, "There are as yet ... no detailed comparisons of the three types of loss functions," and that situation

does not appear to have changed. However, if it was established that different objective functions tend to produce essentially equivalent configurations, then certainly it would be folly to eschew the most tractable function. On the other hand, if it was established that different objective functions tend to produce dramatically different configurations, then surely knowledge of those differences would be crucial to the intelligent choice of an objective function. That is, one ought to at least choose an objective function based on what one is trying to accomplish by using it, not based on prejudices about what is natural.

5 MDS With Bound Constrained Dissimilarities

For most psychometricians and statisticians, there is a dichotomy between metric and nonmetric MDS. Because I believe that this dichotomy is dated, and also that it confuses the question of which issues are truly fundamental to nonmetric MDS, this review adopts a somewhat different classification scheme. The present section considers an MDS problem that is neither metric nor nonmetric.[20] This problem is technically simpler than nonmetric MDS, but is conceptually similar to it.

Bound constrained dissimilarities derive from an important type of NMR spectroscopy known as two-dimensional Nuclear Overhauser Enhancement (NOE) spectroscopy. Bax and Lerner (1986, p. 964) observed that, "The accuracy of distances obtained from 2-D NOE spectra is a highly controversial issue," and enumerated several potential sources of error. In consequence, Wüthrich (1986, p. 189) commented:

> "In structural interpretations with distance geometry the experimental NOE constraints ...are best interpreted as a range of *equally probable* values for the distances in question .... This provides for a proper treatment of the NMR data, since the uncertainties in the experimental distance measurements would presently be difficult to account for by a probability distribution. 'Pure' distance geometry would then attempt to find an exact fit to these *imprecise but completely correct* data, and ideally a structure obtained by a distance geometry calculation should have no violations of the distance constraints imposed by the experimental data. In practice, finding these structures involves also a numerical optimization versus an error function, and each solution typically includes a certain number of small, residual distance violations ...."

Wüthrich's interpretation of the NOE constraints can be formalized by replacing the observed dissimilarities Δ^0 with a set of dissimilarities $\square(\Delta^0)$ defined by inequalities $l_{ij} \leq \delta_{ij} \leq u_{ij}$. This results in the constrained optimization problem

$$\begin{aligned} \text{minimize} \quad & \rho(D, \Delta) \\ \text{subject to} \quad & D \in \mathcal{D}_n(p), \\ & \Delta \in \square(\Delta^0). \end{aligned}$$

If one chooses the objective function ρ to be STRESS, then this is precisely the problem addressed by the Data Box Algorithm of Glunt, Hayden, and Raydan (1993a).

The Data Box Algorithm treats D and Δ as separate sets of variables and successively minimizes the objective function in each. Thus, when minimizing D, Δ is fixed and one solves a metric STRESS subproblem; and when minimizing Δ, D is fixed and one projects into the data box $\square(\Delta^0)$. This is the method of variable alternation for reducible nonlinear programming, which is discussed in Section 6.3.

A second application of bound constrained dissimilarities derives from the fact that NMR experiments do not produce interatomic dissimilarities for all pairs of atoms. Nevertheless, it is desired to determine if a molecule with the specified set of interatomic distances can exist (embedding problem with missing data), or to find a configuration whose corresponding interpoint distances approximate the specified set of interatomic distances (metric MDS with missing data). If STRESS or SSTRESS is the objective function, then the metric MDS problem can be solved by an appropriate choice of weights in Equation (8), and an embedding will exist if and only if the minimum value of the objective function is zero. If STRAIN is the objective function, then define the bounds by setting $l_{ij} = u_{ij} = \delta_{ij}$ for the specified dissimilarities, and $l_{ij} = 0$ and $u_{ij} = \infty$ for the missing dissimilarities. The solution of this bound constrained MDS problem solves the metric MDS problem, and the existence of an embedding is determined as indicated above.

[20] It is tempting to argue that "nonmetric" should be interpreted to mean "not metric," and to advocate its use for any MDS problem in which the dissimilarities are not fixed. However, the phrase "nonmetric MDS" is so strongly identified with what I call "MDS with order constrained dissimilarities" that I prefer to avoid confusion and use it with its traditional connotations.

6 MDS With Order Constrained Dissimilarities

In Section 2, we observed that R. N. Shepard (1962a,b) had offered a number of extremely persuasive examples in support of the viability of nonmetric MDS, i.e. of fitting Euclidean distances to an unspecified nondecreasing transformation of the dissimilarities. However, Shepard's formulation and computational techniques were somewhat heuristic. A rigorous framework for Shepard's ideas was supplied by J. B. Kruskal, whose two original papers (Kruskal, 1964a,b) are perhaps the most important and influential in the MDS literature.[21] Kruskal's fundamental contribution to the theory of nonmetric MDS was the suggestion that the monotonicity constraint on the transformation could be realized by replacing the fixed dissimilarity matrix Δ^0 with the set of all dissimilarity matrices having the same rank ordering of their elements as Δ^0.

Specifically, suppose that the subdiagonal elements of Δ^0 are ordered as

$$\delta^0_{i_1 j_1} \leq \cdots \leq \delta^0_{i_m j_m},$$

and let $\mathcal{M}(\Delta^0)$ denote the set of all matrices Δ for which

$$\delta_{i_1 j_1} \leq \cdots \leq \delta_{i_m j_m}.$$

The δ_{ij} are sometimes called *disparities* (Young, 1972) or *pseudodistances* (Kruskal, 1977). Notice that $\mathcal{M}(\Delta^0) = \mathcal{M}(\Delta^0 * \Delta^0)$. Nonmetric MDS can now be defined as techniques for solving optimization problems of the form

$$\begin{array}{ll} \text{minimize} & \rho(D, \Delta) \\ \text{subject to} & D \in \mathcal{D}_n(p), \\ & \Delta \in \mathcal{M}(\Delta^0). \end{array} \tag{10}$$

The following four subsections consider the fundamental issues involved in formulating and solving such problems.

6.1 STRESS and SSTRESS

In this section we state the usual formulations of the nonmetric STRESS problem proposed by Kruskal (1964a,b) and the nonmetric SSTRESS problem proposed by Takane, Young, and de Leeuw (1977). The relations of these formulations to the general formulation of Problem (10) are not immediately apparent — it is from studying these relations that one gains an appreciation of the fundamental issues of nonmetric MDS.

The nonmetric STRESS problem of Kruskal (1964a,b) is

$$\begin{array}{ll} \text{minimize} & \|D(X) - \pi\, D(X)\|_F^2 \div \|D(X)\|_F^2 \\ \text{subject to} & X \in R^{n \times p}, \end{array} \tag{11}$$

where π denotes projection into the closed and convex set $\mathcal{M}(\Delta^0)$. This is the problem for which the popular KYST computer program (Kruskal, Young, and Seery, 1973) was devised. It is an unconstrained minimization problem in the configuration coordinate variables.

The nonmetric SSTRESS problem of Takane, Young, and de Leeuw (1977) is

$$\begin{array}{ll} \text{minimize} & \|D(X) * D(X) - \Delta * \Delta\|_F^2 \div \|D(X) * D(X)\|_F^2 \\ \text{subject to} & X \in R^{n \times p}, \\ & \Delta \in \mathcal{M}(\Delta^0). \end{array} \tag{12}$$

This is the problem for which the popular ALSCAL computer program was devised. It is a minimization problem in two sets of variables: unconstrained in the configuration coordinate variables, constrained in the disparity variables.

[21] This is another example of MDS research benefitting from the interaction of different disciplines. The statistician J. B. Kruskal was a colleague of R. N. Shepard at Bell Telephone Laboratories. In the words of de Leeuw and Heiser (1982, p. 290), he "introduced psychometricians to loss functions, monotone regression, and (gradient) minimization techniques." Commented Kruskal (1964a, p. 25), "Indeed, this whole paper is the outcome of the author's attempt to rationalize Shepard's successful iterative procedure."

We begin by noting that both Problems (11) and (12) have parametrized Euclidean distances with configuration coordinate variables, writing $D = D(X)$. This device for removing the constraint $D \in \mathcal{D}_n(p)$ was discussed in Section 4.2. It is therefore apparent that the numerator of the objective function in Problem (12), the squared Frobenius (L^2) distance between the squared distances and the squared disparities, is the metric SSTRESS criterion. Furthermore, since $\pi D(X) \in \mathcal{M}(\Delta^0)$, it is apparent that the numerator of the objective function in Problem (11) compares distances and certain disparities using the metric STRESS criterion. (For simplicity, we are ignoring the possibility of arbitrary nonnegative weights in the STRESS or SSTRESS criteria.) But what of the denominators of these objective functions?

The objective functions of both Problems (11) and (12) scale the metric criterion in the numerator, dividing by the squared Frobenius norm of the distances in the case of STRESS and by the squared Frobenius norm of the squared distances in the case of SSTRESS. This widely imitated device, introduced by Kruskal (1964a), will be examined in some detail in Section 6.2. To briefly anticipate, it turns out that this device is equivalent to imposing a constraint on the distances. This means that one can recover the formulation, suggested by Problem (10), of nonmetric MDS as minimizing a metric MDS objective function subject to constraints on distance variables and constraints on disparity variables.

Formulations of nonmetric MDS such as Problem (12) are characterized by a natural distinction between two sets of variables, the distances and the disparities. If one fixes either set of variables, then one obtains an easier subproblem in the other set. Historically, there have been two ways of exploiting this sort of structure. One strategy, exemplified by Problem (12) and the corresponding ALSCAL algorithm, is to retain both sets of variables and successively solve alternating subproblems. The other strategy, exemplified by Problem (11) and the corresponding KYST algorithm, is to parametrize one set of variables in terms of the other. Specifically, one defines a function by assigning, to each fixed value of the first set, a solution to the subproblem in the second set. This is precisely what Kruskal (1964a) did when he replaced Δ with $\pi D(X)$. Each of these strategies will be carefully surveyed in Section 6.3.

The fundamental components of nonmetric MDS are thus identified to be: the metric objective function, the parametrizations of the distance and disparity variables, the constraints imposed on the distances and the disparities, and the optimization strategy. To illustrate how these fundamental components may be effectively juxtaposed, Section 6.4 concludes our discussion of nonmetric MDS by describing Trosset's (1993b) recent formulation of the nonmetric STRAIN problem.

6.2 Nondegeneracy Constraints

As was noted in Section 4.2, Kruskal (1964a) included a scaling factor in the STRESS criterion in order to make it invariant under dilations of the configuration. Besides further complicating the objective function to be minimized, the use of a scaling factor raises fundamental questions about the arbitrariness of the usual formulations of nonmetric MDS. It is to these questions that we now turn.

Kruskal (1964a, p. 8) argued for a dilation-invariant objective function as follows:

> "Surely sheer enlargement of a configuration should not change how well it fits the data, for the relationships between the distances do not change. An obvious way to cure this defect in the raw stress is to divide it by a scaling factor ..."

This argument is predicated on two assumptions: first, that the scale of the *minimizing configuration* is irrelevant; and second, that the *minimum value* of the objective function should not depend on the scale of the data. Because the fundamental purpose of MDS is, after all, the construction of satisfactory configurations, it is the first assumption that is crucial. If one sacrifices Kruskal's desire to interpret the objective function in a way that is not problem-specific,[22] then the first assumption states that it should not matter if the dilations of the minimizing configuration are not minimizers themselves — all such configurations are considered equivalent, and the objective function specifies one of them for output. Thus, Kruskal's preference for a dilation-invariant objective function, however appealing, should not be construed as an intrinsic requirement of MDS.

[22] It must be conceded that this is not an entirely negligible sacrifice. The comparison of minimum objective function values across problems is necessary to establish rules of thumb, as in Kruskal and Wish (1978), for guiding the user to an appropriate choice of dimension.

However, even if one does not require a dilation-invariant objective function, it is not possible to simply revert to minimizing the raw stress, as in nonmetric MDS. The reason is that the sets $\mathcal{D}_n(p)$ and $\mathcal{M}(\Delta^0)$ always intersect at the origin of $R^{n \times n}$. Thus, a minimum value of STRESS (or SSTRESS, or STRAIN) equal to zero can always be obtained by simply placing all of the points in the configuration at the same location. It is evident, therefore, that any meaningful nonmetric MDS optimization problem must somehow incorporate a *nondegeneracy constraint* that precludes such trivial solutions.

Trosset (1990) observed that Kruskal's use of a scaling factor could be viewed as a device for enforcing a nondegeneracy constraint by means of a penalty function. Consider the optimization problems

$$\begin{aligned} &\text{minimize} && f(D) \\ &\text{subject to} && \|D\| \geq c \end{aligned} \tag{13}$$

and

$$\text{minimize} \quad f(D)/\|D\|^2. \tag{14}$$

For Δ fixed, Problem (11) is a special case of Problem (14) with $D = D(X)$ and Problem (12) is a special case of Problem (14) with $D = D(X) * D(X)$. Trosset proved the following result, which casts the use of a scaling factor in nonmetric MDS in a somewhat different light than others have viewed it.

Theorem 3 *If the constraint $\|D\| \geq c$ is active for every $c > 0$, then there exists a particular $c > 0$ for which Problems (13) and (14) have identical solutions.*

It follows that Problem (11) is equivalent to minimizing the raw stress, but demanding that the norm of the distance matrix be at least a certain positive value. A corresponding characterization is true of Problem (12).

Requiring that $\|D\|_F \geq c > 0$ is certainly one way of precluding degenerate solutions to Problem (10), the general nonmetric MDS problem. However, because it is not convex, it is not an appealing constraint. Moreover, it further complicates the already problematic, nonconvex constraint $D \in \mathcal{D}_n(p)$. In consequence, Trosset (1990, 1993b) suggested the possibility of imposing a convex nondegeneracy constraint on the disparities.

For the sake of simplicity, Trosset (1990) assumed that all of the dissimilarities were strictly positive. He suggested requiring the disparities to be no smaller than the smallest dissimilarity, which can be shown to preclude degenerate solutions ($D = 0$) for the STRESS, SSTRESS, and STRAIN criteria. In practice, however, this nondegeneracy constraint tends to result in rather uninteresting configurations. The reason for this phenomenon is highly instructive.

It is transparent that $D = 0$ is not an acceptable solution to Problem (10). What is somewhat less obvious is that other solutions may also be unacceptable. The difficulty is that the set $\mathcal{M}(\Delta^0)$ is so large that it contains disparity matrices that preserve very little of the structure of the dissimilarity matrix. For a nonmetric MDS procedure to be useful in practice, it cannot allow the disparities to stray too far from the dissimilarities, i.e. it must preserve *some* of the metric information in the dissimilarity data. The difficulty with Trosset's (1990) nondegeneracy constraint is that, far too often, the disparities all decrease to the value of the smallest dissimilarity, in which case *none* of the metric information in the dissimilarity data is preserved.

Trosset (1993b) consequently suggested a "conservation of total squared dissimilarity" nondegeneracy constraint, requiring the sum of the squared disparities to be no smaller than the sum of the squared dissimilarities. This appears to work reasonably well. If one parametrizes Problem (10) by the squared disparities (recall that $\mathcal{M}(\Delta^0) = \mathcal{M}(\Delta^0 * \Delta^0)$), then this is a linear (hence convex) constraint. If one parametrizes Problem (10) by the disparities, then this constrains the norm of the disparity matrix be at least a certain positive value — a precise analog of the constraint on the distance matrix implied by Kruskal's scaling factor.

In comparison with Kruskal's penalty function approach, Trosset's (1993b) approach to avoiding nondegeneracy simplifies the objective function while only slightly complicating the constraints. From a computational perspective, this is a very promising development that certainly warrants further exploration. However, it is evident from the preceding discussion that constructing a useful nondegeneracy constraint is an art that probably involves more than a little luck. Whatever its drawbacks, the use of a scaling factor has been resulting in useful configurations for three decades, and the device should not be abandonned precipitously.

6.3 Reducible Nonlinear Programming

We have already observed that a structural characteristic of the nonmetric MDS problem is that there is a natural distinction between the distance variables and the disparity variables. If one fixes the disparity variables, then optimizing in the distance variables is a metric MDS subproblem. If one fixes the distance variables, then optimizing in the disparity variables is the subproblem of projecting into the closed and convex set $\mathcal{M}(\Delta^0)$. This is accomplished by a technique known as isotonic regression, which we will discuss later in this section. For the moment, it suffices to remark that each of these subproblems is considerably easier to solve than the original nonmetric MDS problem. The present section reviews two techniques for exploiting this structure.

Nonlinear programs with the characteristic just described are sometimes called *problems whose variables separate*, e.g. by Golub and Pereyra (1973). Parks (1985), however, has argued persuasively for the superiority of the term *reducible*. One standard technique for exploiting the structure of a reducible nonlinear program is *variable alternation*, in which one successively cycles through the distinguished subsets of variables, optimizing each in turn. We begin by surveying the history of this technique.

Variable alternation has most commonly been used when each subproblem involves computing a projection into a subspace. In this context, variable alternation is often called the method of alternating (orthogonal) projections (MAP). MAP was first studied, in 1933, by von Neumann (1950), who considered the problem of projecting into the intersection of two closed linear subspaces in Hilbert space. In this setting, alternately projecting into each subspace converges to the desired solution. The first use of MAP in statistics was by Wiener and Masani (1957), for the linear prediction of a stochastic process. A more recent example is the alternating conditional expectation (ACE) algorithm of Breiman and Friedman (1985). A recent survey of the MAP literature was made by Deutsch (1992).

Closely related to MAP is the variable alternation algorithm of Dykstra (1983) for projecting into the intersection of closed convex cones in Euclidean space. Dykstra observed that this problem, which he called restricted least squares regression, included a number of important statistical applications as special cases, and demonstrated that cyclically projecting into the cones converges to the desired solution.

A somewhat different use of variable alternation was made by Cheney and Goldstein (1959), who considered the problem of minimizing the distance between K_1 and K_2, two closed convex sets in Hilbert space. This is somewhat closer to the problems of nonmetric MDS than were the previous examples, in that the sets are not required to intersect. Let P_i denote projection into K_i. Cheney and Goldstein gave sufficient conditions for the sequence $(P_1 P_2)^n x$ to converge to a point in K_1 nearest K_2.

In general, a sequence of points constructed by the method of variable alternation is not guaranteed to converge to a solution. However, under very weak conditions, the convergence theory of Zangwill (1969) can be used to establish that every accumulation point of the sequence remains fixed under variable alternation. Stronger conditions must be imposed to guarantee convergence of the sequence and to obtain convergence rates. For example, Zangwill stated conditions that imply the convergence of an unconstrained optimization method that optimizes cyclically in each of the coordinate directions. When variable alternation methods do converge, the rate of convergence is typically linear. Thus, variable alternation methods are often quite slow.

Despite the theoretical limitations of variable alternation, its simplicity and the fact that it produces a monotonic sequence of objective function values have appealed to many statisticians. It has been systematically used in *optimal scaling*, originally by Kruskal (1965). Extensive discussions of optimal scaling were provided by Young (1981) and Gifi (1990). The EM algorithm of Dempster, Laird, and Rubin (1977) is an example of variable alternation, as is the iterative proportional fitting method of loglinear models proposed by Bishop, Fienberg, and Holland (1975). Variable alternation has also been suggested for the solution of certain nonlinear least squares problems by Walling (1968), Wold and Lyttkens (1969), and Barham and Drane (1972). An elementary survey of variable alternation methods in statistics was made by de Leeuw (1993).

In Section 5, we noted the use of variable alternation by Glunt, Hayden, and Raydan (1993a). Its use in nonmetric MDS, in which context it is usually called the method of alternating least squares (ALS), was pioneered by Guttman (1968) and Young and Torgerson (1968). Although Guttman's rather extravagant assertion that use of ALS ensures convergence to a global minimizer is unsubstantiated, the method does appear to work well in practice. It is most commonly identified with the ALSCAL algorithm of Takane,

Young, and de Leeuw (1977) for solving Problem (12).

The other standard technique for exploiting the structure of a reducible nonlinear program is variable reduction. Given an optimization problem of the form

$$\text{minimize} \quad f(x, \alpha), \tag{15}$$

suppose that, for $\bar{\alpha}$ fixed, the subproblem

$$\text{minimize} \quad f(x, \bar{\alpha})$$

has solution $x(\bar{\alpha})$. Then a reduced form of Problem (15) is

$$\text{minimize} \quad \bar{f}(\alpha) = f(x(\alpha), \alpha). \tag{16}$$

The new objective function, $\bar{f}$, is sometimes called the *variable projection functional*, and $x(\cdot)$ is sometimes called the *value function*.

The potential difficulty with variable reduction is that the value function may not be differentiable, although when it is the derivatives of $\bar{f}$ usually have a very simple relation to the derivatives of f. Generalized differentiability of value functions has been the subject of extensive investigation in nonsmooth optimization, e.g. by Clarke (1983, Section 6.5). For the special case of the *semilinear* least squares problem

$$\text{minimize} \quad f(x, \alpha) = \tfrac{1}{2}\|A(\alpha)x - b\|^2, \tag{17}$$

the value function is $x(\alpha) = A(\alpha)^+ b$, where A^+ is the Moore-Penrose pseudoinverse of the matrix A. Golub and Pereyra (1973) successfully differentiated the pseudoinverse and compared the performance of Gauss-Newton algorithms on the full and reduced forms of this problem. Since then, a number of researchers have studied variable reduction methods for Problem (17). Parks (1985) also considered more general nonlinear programs.

The use of variable reduction in nonmetric MDS is due to Kruskal (1964a,b) and is most commonly identified with the KYST algorithm (Kruskal, Young, and Seery, 1973) for solving Problem (11). Instead of letting both D and Δ vary within their respective constraint sets and minimizing the (weighted) L^2 distance between them, Kruskal replaced Δ with the projection of D into $\mathcal{M}(\Delta^0)$. This projection, which we have denoted by π, is easily computed by a technique known as isotonic regression. In fact, in proposing an algorithm for nonmetric MDS, Kruskal (1964b) also introduced the "Up-and-Down Blocks" algorithm, an implementation of the "Pool Adjacent Violators" algorithm for computing the isotonic regression of a simply ordered set. A comprehensive survey of various isotonic regression algorithms was made by Barlow, Bartholomew, Bremner, and Brunk (1972). At present, the most efficient implementation appears to be that of Grotzinger and Witzgall (1984).

Various studies suggest that both variable alternation and variable reduction inevitably improve on algorithms that fail to exploit the structure of a reducible nonlinear program. However, comparisons of variable alternation and variable reduction are less common. Ruhe and Wedin (1980) determined that alternation on the semilinear least squares problem exhibits linear convergence, whereas Gauss-Newton on the reduced problem exhibits superlinear convergence if Gauss-Newton on the full problem does. In general, a useful oversimplification may be that variable alternation tends to be slow but simple, variable reduction tends to be faster but more complicated. A detailed study comparing the performance of these two techniques for solving nonmetric MDS problems would be most welcome.

6.4 STRAIN

We have now identified several fundamental components of nonmetric MDS: the metric objective function (typically STRESS or SSTRESS), the parametrizations of the distances (typically by the configuration coordinates) and the disparities, the nondegeneracy constraint (typically in the form of a scaling factor), and the choice of an optimization strategy for solving the resulting nonlinear program (alternation or reduction). Decomposing nonmetric MDS in this way allows us to recognize the arbitrariness of traditional formulations and, by making different choices, to propose new formulations. As an example of the latter, we briefly summarize Trosset's (1993b) recent formulation of nonmetric MDS using the STRAIN criterion.

In comparison with the metric STRAIN, STRESS, and SSTRESS problems, and with the nonmetric STRESS and SSTRESS problems, the nonmetric STRAIN problem has been almost totally ignored. Takane, Young and de Leeuw (1977) and Trosset (1993b) critiqued the few *ad hoc* approaches proposed in the 1970s. However, by juxtaposing the fundamental components of nonmetric MDS in an appropriate manner, it is not difficult to derive a satisfying formulation of the nonmetric STRAIN problem.

The objective function for the nonmetric STRAIN problem is specified by equation (7), where Δ is reinterpreted as a disparity matrix that can vary subject to constraints. Instead of parametrizing the distances by the configuration coordinates, $D = D(X)$, we can use Theorem 1 to write $\tau(D * D) = B$, where $B \in \Omega_n(p)$. Furthermore, because $\mathcal{M}(\Delta^0) = \mathcal{M}(\Delta^0 * \Delta^0)$, we can substitute Δ for $\Delta * \Delta$. (This substitution means that the δ_{ij} are to be reinterpreted as squared disparities.) Finally, if we impose the "conservation of total squared dissimilarity" nondegeneracy constraint discussed in Section 6.2, we obtain Trosset's (1993b) nonmetric STRAIN problem,

$$\begin{aligned} \text{minimize} \quad & \|B - \tau(\Delta)\|_F^2 \\ \text{subject to} \quad & B \in \Omega_n(p), \\ & \Delta \in \mathcal{M}(\Delta^0) \cap \mathcal{N}(\Delta^0), \end{aligned} \tag{18}$$

where $\mathcal{N}(\Delta^0)$ is the nondegeneracy constraint, defined by

$$\Delta \in \mathcal{N}(\Delta^0) = \left\{ \Delta : \sum_{i,j=1}^{n} \delta_{ij} \geq \sum_{i,j=1}^{n} (\delta_{ij}^0)^2 \right\}.$$

Trosset analyzed both variable alternation (ALS) and variable reduction techniques for solving Problem (18). An elementary argument, applicable to most MDS problems, establishes that all of the accumulation points of the ALS sequence are ALS fixed points. A somewhat more technical argument establishes that all ALS fixed points are *global* minimizers of Problem (18). Because the sequence of objective function values produced by ALS is nondecreasing, it follows that ALS is guaranteed to find configurations with STRAIN values arbitrarily close to the global minimum.

In contrast to Kruskal (1964a), variable reduction was used to parametrize Problem (18) entirely in terms of the *disparity* variables. This was accomplished by exploiting the explicit representation of the global minimizer of the metric STRAIN problem provided by Theorem 2. Every *local* minimizer of the resulting variable projection functional corresponds to an ALS fixed point, hence to a global minimizer of Problem (18). Except in rare instances, the value function is differentiable and the gradient of the variable projection functional is easily computed. A gradient projection algorithm for constrained optimization successfully solved the reduced problem, although Trosset noted that more efficient algorithms were available.

Although there is no compelling reason to prefer the STRAIN criterion for nonmetric MDS, solutions to Problem (18) appear to be perfectly reasonable configurations. For example, the optimal configuration for Ekman's (1954) color data is almost identical to the configuration obtained by Shepard (1962b). What is especially appealing, of course, is that this formulation of nonmetric MDS ensures the discovery of globally optimal configurations.

7 Statistical Properties of MDS

In Section 4, the dissimilarities were fixed. In Sections 5 and 6, we distinguished between dissimilarities and disparities. The disparities were free to vary, subject to constraints that are typically determined by the fixed dissimilarities, as in Section 6. Thus, each of the MDS problems formulated in Sections 4–6 is a deterministic optimization problem with parameters determined by observing a fixed matrix of dissimilarity data. However, because data observation is subject to error, the statistician is (or should be) inclined to inquire how MDS is affected by this error, i.e. to inquire about the *statistical* properties of MDS.

In Section 2, MDS was motivated as a method of fitting Euclidean distances to dissimilarity data. The deterministic optimization problems formulated in Sections 4–6 specify how that fitting is to be done. However, because these optimization problems are not easy to solve, MDS researchers have tended to neglect other concerns. As Ramsay (1977, p. 241) explained,

"Implicit in almost all data analyses is some statement about the manner in which the observation varies about its fitted value. Often this model for the distribution of residuals or errors is only implicitly defined, not easy to display, and only of secondary interest. This is especially true when the fitting process is so challenging to the data analyst that he congratulates himself on having fit the data at all. In this case he can be excused for his disinterest in the error distribution assumptions that he has implied, and he is likely to respond to doubts about their plausibility by pointing out that an implausible fit is better than none at all."

Several potentially useful distributions for dissimilarity data were motivated in Section 2. Later in this section, we will return to one of them. However, one need not specify a distribution to investigate how perturbations to the observed dissimilarities affect the configurations constructed by MDS. Such investigations are the province of *sensitivity analysis*, the study of how the solutions to a parametrized family of optimization problems vary with changes in the parameters.

To analyze the sensitivity of an MDS procedure, it is necessary to measure the discrepancy between the configurations constructed from the original and the perturbed dissimilarities. The comparison of two (or more) configurations is usually termed *procrustes analysis*.[23] Such comparisons were originally devised in order to facilitate factor analyses, and can be traced back to Mosier (1939). Cliff (1966), Schönemann (1968), Schönemann and Carroll (1970), and especially Gower (1971) were the first researchers to recognize the applicability of procrustes analysis to comparing configurations constructed by MDS. An elegant introduction to the subject, motivated by the concerns of MDS sensitivity analyses, was provided by Sibson (1978). A more general review, with much discussion and many references, was made by Goodall (1991).

Given configurations X and Y, Sibson (1978) defined the procrustes statistic $G_\Gamma(X,Y)$ by

$$G_\Gamma(X,Y) = \inf\{G(X,\phi Y) : \phi \in \Gamma\},$$

where

$$G(X,Y) = \|X-Y\|_F^2 = \text{trace}\left[(X-Y)^t(X-Y)\right]$$

and Γ is a group of transformations of R^p. If one takes $\Gamma = E$ to be the group of isometric transformations (translations, rotations, and reflections), with respect to which the distances $D(X)$ are invariant, then

$$G_E(X,Y) = \text{trace}\left[XX^t\right] + \text{trace}\left[YY^t\right] - 2\,\text{trace}\left[M^{1/2}\right], \tag{19}$$

where $M^{1/2}$ is the unique positive semidefinite symmetric square root of the symmetric positive semidefinite matrix $M = YX^tXY^t$. This measure of the discrepancy between the configurations X and Y was originally proposed by Schönemann and Carroll (1970) and by Gower (1971).

Mardia (1978) and Sibson (1979) each employed (19) to investigate the stability of classical MDS under small perturbations of the dissimilarities. Mardia's result is a special case of the more definitive treatment by Sibson. As in Sections 3 and 4.1, let

$$B = \tau(\Delta * \Delta) = XX^t.$$

For any symmetric matrix C, let

$$B_\epsilon = B + \epsilon C + O(\epsilon^2) = YY^t$$

be the inner product matrix obtained by perturbing the dissimilarities. Then Sibson demonstrated that

$$G_E(X,Y) = O(\epsilon^2),$$

[23] The terminology is borrowed from the Procrustes program of Hurley and Cattell (1962) for rotating factor structures, named for the villainous innkeeper of Greek mythology. According to Diodorus Siculus (Book IV.59.5; English translation by Oldfather, 1939),

"After this he [Theseus] put to death Procrustes, as he was called, who dwelt in what was known as Corydallus in Attica; this man compelled the travellers who passed by to lie down upon a bed, and if any were too long for the bed he cut off the parts of their body which protruded, while in the case of such as were too short for it he stretched (*prokrouein*) their legs, this being the reason why he was given the name Procrustes."

concluding that "there is ample theoretical justification for the view that classical scaling is a method which is robust against errors which leave observed dissimilarities still approximately linearly related to distance." (p. 228). Of course, from (7) it is easily inferred (and widely appreciated) that classical MDS is quite sensitive to a large perturbation of even a single dissimilarity; an example was provided by Spence and Lewandowsky (1989).

Formal sensitivity analyses of other MDS problems have not been forthcoming, no doubt due to the formidable mathematical challenges posed by such investigations. Nevertheless, it is possible to draw some qualitative conclusions about the relative stability of different MDS problems simply by inspecting their objective functions. Let us write the STRAIN, STRESS, and SSTRESS criteria as

$$\rho(D,\Delta) = \sum_{i,j=1}^{n} \left[h(d_{ij}^2) - h(\delta_{ij}^2)\right]^2 .$$

In this representation, all three criteria are defined as the squared error between a transformation h of the squared distances and the same transformation of the squared dissimilarities. On this basis, one would certainly be inclined to anticipate some sensitivity to perturbations of the dissimilarities. For SSTRESS, $h(a) = a$, so that SSTRESS is the L^2 distance between $D * D$ and $\Delta * \Delta$. For STRAIN, h is defined by equation (6). Since this is a linear transformation, we might plausibly conjecture that comparable MDS techniques using STRAIN or SSTRESS have comparable stability properties. In contrast, for STRESS, $h(a) = \sqrt{a}$. This means that STRESS is the Hellinger distance between $D * D$ and $\Delta * \Delta$. In the context of minimum distance parameter estimation, Donoho and Liu (1988) have established favorable stability properties of the Hellinger distance between probability density functions; accordingly, we might plausibly conjecture that MDS techniques using STRESS are more stable than comparable MDS techniques using STRAIN or SSTRESS. Of course, nonmetric MDS techniques, which depend on the dissimilarities only through there ranks, will tend to be more stable than corresponding metric MDS techniques.

In the absence of formal sensitivity analyses, several researchers have proposed resampling schemes for investigating the stability of MDS configurations. This is straightforward for three-way MDS, in which there are independent replications of the dissimilarity matrix available for resampling. For two-way MDS, Kruskal and Wish (1978, p. 59) suggested that "stimuli could be eliminated from the data matrix, and solutions determined for the remaining stimuli, using the 'jacknife' idea of J. W. Tukey ...." This suggestion was taken up by de Leeuw and Meulman (1986), who concluded that the stability results thus obtained justified the computational expense. For future reference, we record their admonition that, "*our results are not supposed to have inferential applications.*" (p. 99).

It is almost irresistible to describe the stability properties of MDS as robustness properties, as indeed many MDS researchers have done. This usage, however, is somewhat problematic. For example, Donoho and Liu (1988) were obliged to distinguish between two notions of robustness, "stability of variance" and "stability of quantity estimated." The former is the most common notion of robustness, whereas the latter is the notion implied by sensitivity analysis (and the notion studied by Donoho and Liu). Furthermore, the usual connotation associated with robustness is that a procedure is (or is not) robust against certain departures from some assumption, e.g. about the form of the error distribution. In reviewing what is known about the stability of MDS, we have not yet introduced any distributional assumptions that might be violated. We now proceed to do precisely that.

De Leeuw and Meulman (1986, p. 97) observed that "statistical analysis, in the the form of computation of standard errors or confidence intervals for example, is a particular kind of sensitivity analysis, which studies the *perturbations* induced by random sampling." In Section 2, we considered a variety of situations in which dissimilarity data might be sampled. Because different situations entail different dissimilarity distributions, a plethora of probability models for MDS have been proposed. Some of these models were motivated primarily by the need for mathematical tractability; many were proposed for highly specialized situations that are unlikely to be of general interest. Therefore, this review does not attempt a comprehensive survey of probability models for MDS; we consider only the MULTISCALE procedure of Ramsay (1977, 1978, 1980, 1982, 1983), the single most important example of probability modelling in MDS.

In discussing the work of Shepard (1957, 1958a,b) in Section 2, we traced the logic for permitting monotone transformations of the observed dissimilarities. It is this logic that led Kruskal (1964a) to propose the order constraints imposed in nonmetric MDS. However, in Section 6.2 we observed that these constraints

sometimes allow the disparities too much freedom to vary, i.e. one may not really want to permit *all* monotone transformations. This point was also made by Ramsay (1977), whose work represents a return to the 1950s tradition of applying metric MDS to transformed dissimilarity data. Partially motivated by psychophysical phenomena, Ramsay argued for a model in which the standard deviation of each (independent) dissimilarity is proportional to its expected value, which led him to propose the lognormal distribution for dissimilarity data. This is equivalent to assuming regression equation (5), with $a = 0$, $b = 1$, and the e_{ij} independently drawn from a normal distribution with mean zero and fixed variance σ^2. He also considered the possibility of retaining b as a parameter, thereby proposing to estimate the best of a parametric family of transformations. Ramsay (1982) subsequently argued that even the family of power transformations is inadequate for many types of data, and proposed replacing it with a parametric family of monotone splines.

Ramsay's distributional assumptions were motivated by more than a simple desire for MDS models that utilize metric information about the dissimilarities. The defining characteristic of his work is not so much his particular sets of assumptions as his estimation of the configuration coordinates by the method of maximum likelihood. This leads to optimization problems that are different from, and perhaps less tractable than, the ones that have been explicated in this review. Ramsay's (1977, p. 255) motivation for formulating MDS in this way was "that the log likelihood can be used to construct test statistics for a wide range of hypotheses for which the asymptotic distribution is known." Thus, what distinguishes Ramsay's MULTISCALE approach to MDS from the others that we have considered is his interest in formal statistical inference.

Statistical inference, especially *asymptotic* statistical inference, requires replications of the data. In the present context, these replications are obtained by observing multiple dissimilarity matrices. Thus, Ramsay's work is most instructively placed in the category of three-way MDS, and MULTISCALE is most naturally a competitor of Carroll's and Chang's (1970) INDSCAL procedure for metric individual differences MDS using the STRAIN criterion. Ramsay (1978) developed analytic techniques for constructing asymptotic confidence ellipsoids of the individual points in a configuration. In contrast, Weinberg, Carroll, and Cohen (1984) used bootstrap and jackknife techniques to estimate corresponding confidence ellipsoids for configurations constructed by INDSCAL and MULTISCALE. Their Monte Carlo experiments led them to suggest "that MULTISCALE asymptotic estimates of standard errors based on small samples provide an optimistic view of the actual statistical reliability of the solution." (p. 490). Moreover, these experiments used data simulated to have the statistical properties assumed by MULTISCALE. In practical applications of MDS, these assumptions will almost certainly be violated, and the nonrobustness of most maximum likelihood procedures with respect to departures from the assumptions on which the asymptotic theory depends is well-known.

Complementing Ramsay's work on asymptotic maximum likelihood inference for metric MDS, Brady (1985) has developed a corresponding theory for nonmetric MDS. Brady exploited a formidable probabilistic framework to establish the set-valued consistency of his estimation procedure, citing three well-known texts on the importance of statistical consistency. His blatant insistence on this point foregrounds a central issue in the controversy over probability models for MDS, for it is not at all clear that statistical consistency is a useful notion in this context. For one thing, the number of replications (dissimilarity matrices) is usually *much* smaller than the number of parameters (configuration coordinate variables). In the case of two-way MDS, only one dissimilarity matrix is observed. More fundamentally, the notion of statistical consistency can only be meaningful if there is an underlying statistical model to be estimated. The range of examples discussed in Section 2 suggest that this is sometimes the case, sometimes not. Even when it is the case, it may be impossible to adequately specify what that model is. Thus, while I can only applaud the development of specialized MDS procedures for those specialized models that do arise in practice, I am more generally inclined to sympathize with Chatfield's feeling that "model-based MDS is a technique in search of a suitable problem rather than the reverse." (Ramsay, 1982, p. 306).

The controversy over probability models for MDS raises fundamental questions. What is the role of MDS in statistics? What, indeed, is the scope of modern statistics? The final section of this review attempts to address these questions.

8 Discussion

The most profound exploration of the role of MDS in statistics that I have encountered occurs in the discussion of the paper read by Ramsay (1982) before the Royal Statistical Society. Ramsay described most

approaches to MDS as "non-statistical employing as a rule least squares estimation." (p. 287). Several discussants took issue with the implications of this characterization. Chatfield (p. 306) responded,

> "In Statistics, it is sometimes useful to distinguish between two broad approaches, the one being data-analytic and the other probabilistic. In the former a probability model is either completely absent or plays a very subordinate role (see Cox and Snell, 1981, Section 1.5). Of course the distinction is not clear cut and the good statistician will often use a combination of both approaches. An alternative distinction is that between exploratory techniques and confirmatory techniques. ...
>
> "In his Section 1.2, the author describes some approaches to MDS as being 'non-statistical', and this description is one that I would take issue with. It is my belief that data-analytic techniques are just as important a part of Statistics as are probabilistic methods."

The same belief was expressed even more succinctly by de Leeuw (p. 310):

> "My thesis would be that in many cases MDS is not a statistical, in the sense of inferential or inductive, technique, but simply a graphical method. My second thesis is that there is nothing inferior about graphical methods."

Thus, the issue of probability-based MDS presents the statistician with a sort of Rorshach test: what one sees in MDS reveals one's more general perceptions of statistics, especially multivariate statistics.

Recent decades have witnessed a revolution in multivariate statistics. Classical multivariate analysis, exemplified by Anderson (1958), was primarily concerned with formal inference and an accompanying distribution theory. Implicit in this approach was a tendency to think of multivariate problems as straightforward generalizations of univariate problems. However, as advances in computer technology began to permit the analysis of large data sets in many dimensions, it became increasingly clear that multivariate problems often pose difficulties that are uniquely multivariate. Although classical multivariate analysis will always remain important, modern multivariate analysis, exemplified by Gnanadesikan (1977), has emphasized the detection of multivariate structure, especially through the use of sophisticated graphical methods. An especially visible symbol of the increasing importance of such methods was the establishment, in 1992, of the *Journal of Computational and Graphical Statistics.*

Because many widely known graphical methods are extremely elementary, the theoretical challenges posed by such methods are sometimes overlooked. The computational challenges posed by exploratory methods like projection pursuit and MDS are very different from the distributional challenges posed by inferential procedures for probabilistic models, but they are no less formidable. The formulation of meaningful optimization problems, the development of efficient algorithms for their solution, and the analysis of the stability of the solutions, are unarguably central concerns of modern computational statistics.

The present review has strongly emphasized the formulation of the optimization problems that define different approaches to MDS. In my view, an appreciation of these problems is essential to an understanding of what it is that MDS actually does. Moreover, I strongly believe that computational statisticians should address the issue of how a problem can best be formulated, not merely how a given formulation can best be solved. If an imprecisely specified problem can be formulated in various ways, and the choice of a formulation is arbitrary, then it is natural to prefer a formulation that is easy to solve. This philosophy has a long and honored tradition in mathematical statistics. The best known example is the conventional choice of the squared error loss function, usually for the purpose of achieving mathematical tractability and not for the purpose of modelling any intrinsic characteristic of the original problem. Another example was provided by Bickel and Lehmann (1975), who considered the problem of estimating the "center" of an asymmetric distribution. Because many functionals are satisfactory measures of location, the authors suggested choosing a parameter that can be efficiently estimated.[24]

As developed in this review, the intrinsic characteristics of MDS involve the approximation of a set of dissimilarity matrices with a Euclidean interpoint distance matrix. The requirement of a distance matrix specifies the constraint $D \in \mathcal{D}_n(p)$, and the set of dissimilarity matrices (this review has considered fixed, bound constrained, and order constrained dissimilarities) is usually specified by the application. However,

[24] For a very amusing dismissal of this philosophy, the reader is referred to Johansen's witty discussion (pp. 164–165) of the review by Bickel (1976).

the requirement that the former approximate the latter does not specify any one objective function (or even squared error loss), nor does it specify any one way of avoiding degenerate solutions. Nevertheless, certain arbitrary specifications have indeed come to be regarded by many researchers as intrinsic to MDS. Perhaps it is time for computational statisticians to free themselves from these conventions.

Acknowledgements

John Dennis and especially David Scott encouraged me to undertake this project. Pat Jones, Al Kaszniak, and Darrell Sabers supplied me with fascinating information about the early history of psychometric research. John Dennis, Tony Kearsley, Jorge Moré, Marcos Raydan, and especially Richard Tapia greatly supplemented my own knowledge of optimization theory and numerical analysis. Tom Hayden, Marcos Raydan, and Pablo Tarazaga graciously provided and discussed with me their unpublished research. Virginia Torczon and Richard Tapia read an early version of the original technical report (Trosset, 1993a) and made many helpful suggestions. The transition from the somewhat more narrow focus of the technical report to the more general concerns of the present article was suggested by the editor. In the process of making these revisions, I benefitted enormously from my interactions with Jan de Leeuw, whose many valuable suggestions are far too numerous to be listed here.

References

Anderson, T. W. (1958). *An Introduction to Multivariate Statistical Analysis.* John Wiley & Sons, New York.

Arabie, P. (1991). Was Euclid an unnecessarily sophisticated psychologist? *Psychometrika*, 56:567–587.

Attneave, F. (1950). Dimensions of similarity. *American Journal of Psychology*, 63:516–556.

Barham, R. H. and Drane, W. (1972). An algorithm for least squares estimation of nonlinear parameters when some of the parameters are linear. *Technometrics*, 14:757–766.

Barlow, R. E., Bartholomew, J. M., Bremner, J. M., and Brunk, H. D. (1972). *Statistical Inference Under Order Restrictions.* John Wiley & Sons, New York.

Bax, A. and Lerner, L. (1986). Two-dimensional nuclear magnetic resonance spectroscopy. *Science*, 232:960–967.

Bickel, P. J. (1976). Another look at robustness: A review of reviews and some new developments. *Scandinavian Journal of Statistics*, 3:145–168. Includes comments by six discussants and reply by author.

Bickel, P. J. and Lehmann, E. L. (1975). Descriptive statistics for nonparametric models. II. Location. *Annals of Statistics*, 3:1045–1069.

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice.* MIT Press, Cambridge.

Blumenthal, L. M. (1938). Distance geometries. *University of Missouri Studies*, 13.

Blumenthal, L. M. (1953). *Theory and Applications of Distance Geometry.* Clarendon Press, Oxford.

Brady, H. E. (1985). Statistical consistency and hypothesis testing for nonmetric multidimensional scaling. *Psychometrika*, 50:509–537.

Braun, W. (1987). Distance geometry and related methods for protein structure determination from NMR data. *Quarterly Reviews of Biophysics*, 19:115–157.

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

Browne, M. (1987). The Young-Householder algorithm and the least square multidimensional scaling of squared distance. *Journal of Classification*, 4:175–190.

Carroll, J. D. and Chang, J.-J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. *Psychometrika*, 35:283–319.

Cayley, A. (1841). On a theorem in the geometry of position. *Cambridge Mathematics Journal*, 2:267–271.

Cheney, W. and Goldstein, A. A. (1959). Proximity maps for convex sets. *Proceedings of the American Mathematical Society*, 10:448–450.

Clarke, F. H. (1983). *Optimization and Nonsmooth Analysis*. John Wiley & Sons, New York.

Cliff, N. (1966). Orthogonal rotation to congruence. *Psychometrika*, 31:33–42.

Cox, D. R. and Snell, E. J. (1981). *Applied Statistics*. Chapman and Hall, London.

Crippen, G. M. and Havel, T. F. (1988). *Distance Geometry and Molecular Conformation*. John Wiley & Sons, New York.

Critchley, F. (1988). On certain linear mappings between inner-product and squared-distance matrices. *Linear Algebra and Its Applications*, 105:91–107.

de Leeuw, J. (1977). Applications of convex analysis to multidimensional scaling. In Barra, J. R., Brodeau, F., Romier, G., and van Cutsem, B., editors, *Recent Developments in Statistics*, pages 133–145. North-Holland Publishing Company, Amsterdam.

de Leeuw, J. (1984). Differentiability of Kruskal's Stress at a local minimum. *Psychometrika*, 49:111–113.

de Leeuw, J. (1988). Convergence of the majorization method for multidimensional scaling. *Journal of Classification*, 5:163–180.

de Leeuw, J. (1993). Block-relaxation algorithms in statistics. Preprint, UCLA Statistics, Los Angeles.

de Leeuw, J. and Heiser, W. (1980). Multidimensional scaling with restrictions on the configuration. In Krishnaiah, P. R., editor, *Multivariate Analysis*, volume 5. North-Holland Publishing Company, Amsterdam.

de Leeuw, J. and Heiser, W. (1982). Theory of multidimensional scaling. In Krishnaiah, P. R. and Kanal, I. N., editors, *Handbook of Statistics*, volume 2, chapter 13, pages 285–316. North-Holland Publishing Company, Amsterdam.

de Leeuw, J. and Meulman, J. (1986). A special jackknife for multidimensional scaling. *Journal of Classification*, 3:97–112.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38.

Dennis, J. E., Gay, D. M., and Welsch, R. E. (1981). An adaptive nonlinear least-squares algorithm. *ACM Transactions on Mathematical Software*, 7:348–368.

Deutsch, F. (1992). The method of alternating orthogonal projections. In Singh, S. P., editor, *Approximation Theory, Spline Functions and Applications*, pages 105–121. Kluwer, Netherlands.

Donoho, D. L. and Liu, R. C. (1988). The "automatic" robustness of minimum distance functionals. *Annals of Statistics*, 16:552–586.

Dunnell, R. C. (1970). Seriation method and its evaluation. *American Antiquity*, 35:305–319.

Dykstra, R. L. (1983). An algorithm for restricted least squares regression. *Journal of the American Statistical Association*, 78:837–842.

Ekman, G. (1954). Dimensions of color vision. *Journal of Psychology*, 38:467–474.

Fechner, G. T. (1860). *Elemente der Psychophysik*. Breitkopf und Hartel, Leipzig. Two volumes.

Fréchet, M. (1935). Sur la définition axiomatique d'une classe d'espaces distanciés vectoriellement applicable sur l'espace de Hilbert. *Annals of Mathematics*, 36:705–718.

Gifi, A. (1990). *Nonlinear Multivariate Analysis*. John Wiley & Sons, New York.

Glunt, W., Hayden, T. L., and Liu, W.-M. (1991). The embedding problem for predistance matrices. *Bulletin of Mathematical Biology*, 53:769–796.

Glunt, W., Hayden, T. L., and Raydan, M. (1993a). Molecular conformations from distance matrices. *Journal of Computational Chemistry*, 14:114–120.

Glunt, W., Hayden, T. L., and Raydan, M. (1993b). Preconditioners for distance matrix algorithms. Manuscript submitted for publication.

Gnanadesikan, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations*. John Wiley & Sons, New York.

Golub, G. H. and Pereyra, V. (1973). The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. *SIAM Journal on Numerical Analysis*, 10:413–432.

Goodall, C. (1991). Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society, Series B*, 53:285–339. Includes comments by 18 discussants and reply by author.

Gower, J. C. (1966). Some distance properties of latent root and vector methods in multivariate analysis. *Biometrika*, 53:315–328.

Gower, J. C. (1971). Statistical methods of comparing different multivariate analyses of the same data. In Hodson, F. R., Kendall, D. G., and Tăutu, P., editors, *Mathematics in the Archaeological and Historical Sciences*, pages 138–149. Edinburgh University Press, Edinburgh.

Grotzinger, S. J. and Witzgall, C. (1984). Projections onto order simplexes. *Applied Mathematics and Optimization*, 12:247–270.

Gulliksen, H. (1946). Paired comparisons and the logic of measurement. *Psychological Review*, 53:199–213.

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

Havel, T. F. (1991). An evaluation of computational strategies for use in the determination of protein structure from distance constraints obtained by nuclear magnetic resonance. *Progress in Biophysics and Molecular Biology*, 56:43–78.

Hayashi, C. (1974). Minimum dimension analysis (MDA): One of the methods of multidimensional quantification. *Behaviormetrika*, 1:1–24.

Hayden, T. L. and Tarazaga, P. (1993). Distance matrices and regular figures. *Linear Algebra and Its Applications*, 195:9–16.

Hayden, T. L., Wells, J., Liu, W., and Tarazaga, P. (1991). The cone of distance matrices. *Linear Algebra and Its Applications*, 144:153–169.

Helm, C. E. (1959). A multidimensional ratio scaling analysis of color relations. Technical report, Educational Testing Service, Princeton.

Hoare, M. R. and Pal, P. (1971). Physical cluster mechanics: Statics and energy surfaces for monatomic systems. *Advances in Physics*, 20:161–196.

Hole, F. and Shaw, M. (1967). Computer analysis of chronological seriation. *Rice University Studies*, 53.

Householder, A. S. and Landahl, H. D. (1945). *Mathematical Biophysics of the Central Nervous System.* Principia Press, Bloomington, IN.

Hurley, J. R. and Cattell, R. B. (1962). The Procrustes program: Producing direct rotation to test a hypothesised factor structure. *Behavioral Science*, 7:258–262.

Kearsley, A. J., Tapia, R. A., and Trosset, M. W. (1993). The solution of the metric STRESS and STRESS problems in multidimensional scaling using Newton's method. Technical Report TR93-?, Department of Computational and Applied Mathematics, Rice University, Houston, TX 77251-1892.

Keller, J. B. (1962). Factorization of matrices by least squares. *Biometrika*, 49:239–242.

Kendall, D. G. (1969a). Incidence matrices, interval graphs, and seriation in archaeology. *Pacific Journal of Mathematics*, 28:565–570.

Kendall, D. G. (1969b). Some problems and methods in statistical archaeology. *World Archaeology*, 1:68–76.

Kendall, D. G. (1971). Seriation from abundance matrices. In Hodson, F. R., Kendall, D. G., and Tăutu, P., editors, *Mathematics in the Archaeological and Historical Sciences*, pages 215–251. Edinburgh University Press, Edinburgh.

Klingberg, F. L. (1941). Studies in measurement of the relations among sovereign states. *Psychometrika*, 6:335–352.

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*, 29:28–42.

Kruskal, J. B. (1965). Analysis of factorial experiments by estimating monotone transformations of the data. *Journal of the Royal Statistical Society, Series B*, 27:251–263.

Kruskal, J. B. (1977). Multidimensional scaling and other methods for discovering structure. In Enslein, K., Ralston, A., and Wilf, H. S., editors, *Statistical Methods for Digital Computers*, chapter 12, pages 296–339. John Wiley & Sons, New York. Volume III of *Mathematical Methods for Digital Computers.*

Kruskal, J. B. and Wish, M. (1978). *Multidimensional Scaling.* Sage Publications, Beverly Hills.

Kruskal, J. B., Young, F. W., and Seery, J. B. (1973). How to use KYST, a very flexible program to do multidimensional scaling and unfolding. Technical report, Bell Laboratories.

Mardia, K. V. (1978). Some properties of classical multi-dimensional scaling. *Communications in Statistics — Theory and Methods*, A7:1233–1241.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis.* Academic Press, Orlando.

Menger, K. (1931a). Bericht über metrische geometrie. *Jahresber. der Deutschen Math.-Ver.*, 40:201–219.

Menger, K. (1931b). New foundations of Euclidean geometry. *American Journal of Mathematics*, 53:721–745.

Moré, J. J., Garbow, B. S., and Hillstrom, K. E. (1980). User's guide for MINPACK-1. Technical Report ANL 80-74, Argonne National Laboratories.

Mosier, C. I. (1939). Determining a simple structure when loadings for certain tests are known. *Psychometrika*, 4:149–162.

Obenchain, R. L. (1971). Squared distance scaling as an alternative to principal components analysis. Research note, Bell Laboratories, Holmdell.

Oldfather, C. H. (1939). *The Library of History of Diodorus of Sicily.* William Heinemann Ltd, London.

Parks, T. A. (1985). Reducible nonlinear programming problems. Technical Report 85-8, Department of Mathematical Sciences, Rice University, Houston, TX. Author's Ph.D. thesis.

Petrie, W. M. F. (1899). Sequences in prehistoric remains. *Journal of the Anthropological Institute of Great Britain and Ireland*, 29:295–301.

Ramsay, J. O. (1977). Maximum likelihood estimation in multidimensional scaling. *Psychometrika*, 42:241–266.

Ramsay, J. O. (1978). Confidence regions for multidimensional scaling analysis. *Psychometrika*, 43:145–160.

Ramsay, J. O. (1980). Some small sample results for maximum likelihood estimation in multidimensional scaling. *Psychometrika*, 45:139–144.

Ramsay, J. O. (1982). Some statistical approaches to multidimensional scaling data. *Journal of the Royal Statistical Society, Series A*, 145:285–312. Includes comments by 14 discussants and reply by author.

Ramsay, J. O. (1983). MULTISCALE: A multidimensional scaling program. *American Statistician*, 37:326–327.

Richardson, M. W. (1938). Multidimensional psychophysics. *Psychology Bulletin*, 35:659–660. Abstract.

Ruhe, A. and Wedin, P. A. (1980). Algorithms for separable nonlinear least squares problems. *SIAM Review*, 22:318–337.

Saito, T. (1978). The problem of the additive constant and eigenvalues in metric multidimensional scaling. *Psychometrika*, 43:193–201.

Schoenberg, I. J. (1935). Remarks to Maurice Fréchet's article "Sur la définition axiomatique d'une classe d'espaces distanciés vectoriellement applicable sur l'espace de Hilbert". *Annals of Mathematics*, 38:724–732.

Schönemann, P. H. (1968). On two-sided orthogonal procrustes problems. *Psychometrika*, 33:19–33.

Schönemann, P. H. and Carroll, R. M. (1970). Fitting one matrix to another under choice of a central dilation and a rigid motion. *Psychometrika*, 35:245–255.

Seber, G. A. F. (1984). *Multivariate Observations.* John Wiley & Sons, New York.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, 22:325–345.

Shepard, R. N. (1958a). Stimulus and response generalization: Deduction of the generalization gradient from a trace model. *Psychological Review*, 65:242–256.

Shepard, R. N. (1958b). Stimulus and response generalization: Tests of a model relating generalization to distance in psychological space. *Journal of Experimental Psychology*, 55:509–523.

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

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

Sibson, R. (1978). Studies in the robustness of multidimensional scaling: Procrustes statistics. *Journal of the Royal Statistical Society, Series B*, 40:234–238.

Sibson, R. (1979). Studies in the robustness of multidimensional scaling: Perturbationa; analysis of classical scaling. *Journal of the Royal Statistical Society, Series B*, 41:217–229.

Spence, I. and Lewandowsky, S. (1989). Robust multidimensional scaling. *Psychometrika*, 54:501–513.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103:677–680.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, 64:153–181.

Stevens, S. S. (1970). Neural events and the psychophysical power law. *Science*, 170:1043–1050.

Stumpf, C. (1880). *Tonpsychologie.* Teubner, Leipzig. Two volumes.

Takane, Y., Young, F. W., and de Leeuw, J. (1977). Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features. *Psychometrika*, 42:7–67.

Tarazaga, P., Hayden, T. L., and Wells, J. (1993). Spherical distance matrices and faces. Manuscript submitted for publication.

Thurstone, L. L. (1927a). A law of comparative judgment. *Psychological Review*, 34:273–286.

Thurstone, L. L. (1927b). Psychophysical analysis. *American Journal of Psychology*, 38:368–389.

Thurstone, L. L. (1935). *The Vectors of Mind.* University of Chicago Press, Chicago.

Torgerson, W. S. (1952). Multidimensional scaling: I. Theory and method. *Psychometrika*, 17:401–419.

Torgerson, W. S. (1958). *Theory and Methods of Scaling.* John Wiley & Sons, New York.

Trosset, M. W. (1990). A constrained optimization approach to nonmetric multidimensional scaling. In *1990 Proceedings of the Statistical Computing Section*, pages 117–121. American Statistical Association.

Trosset, M. W. (1993a). Optimization problems associated with multidimensional scaling. Technical Report TR93-13, Department of Computational and Applied Mathematics, Rice University, Houston, TX 77251-1892.

Trosset, M. W. (1993b). A new formulation of the nonmetric STRAIN problem in multidimensional scaling. Technical Report TR93-30, Department of Computational and Applied Mathematics, Rice University, Houston, TX 77251-1892. Submitted for publication; in revision.

Velleman, P. F. and Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *American Statistician*, 47:65–72.

von Neumann, J. (1950). *Functional Operators, Volume II. The Geometry of Orthogonal Spaces.* Annals of Mathematical Studies, #22. Princeton University Press, Princeton, NJ. Reprinted from mimeographed lecture notes first distributed in 1933.

Walling, D. (1968). Non-linear least squares curve fitting when some parameters are linear. *Texas Journal of Science*, 20:119–124.

Weinberg, S. L., Carroll, J. D., and Cohen, H. S. (1984). Confidence regions for INDSCAL using the jackknife and bootstrap techniques. *Psychometrika*, 49:475–491.

Wemmer, D. E. and Reid, B. R. (1985). High resolution NMR studies of nucleic acids and proteins. *Annual Review of Physical Chemistry*, 36:105–137.

Wider, G., Macura, S., Kumar, A., Ernst, R. R., and Wüthrich, K. (1984). Homonuclear two-dimensional ^{1}H NMR of proteins. Experimental procedures. *Journal of Magnetic Resonance*, 56:207–234.

Wiener, N. and Masani, P. (1957). The prediction theory of multivariate stochastic processes. II: The linear predictor. *Acta Mathematica*, 93:95–137.

Wold, H. and Lyttkens, E. (1969). Nonlinear iterative partial least squares (NIPALS) estimation procedures. *Bulletin of the International Statistics Institute*, 43:29–51.

Wüthrich, K. (1986). *NMR of Proteins and Nucleic Acids*. John Wiley & Sons, New York.

Young, F. W. (1972). A model for polynomial conjoint analysis algorithms. In Shepard, R. N., Romney, A. K., and Nerlove, S. B., editors, *Multidimensional Scaling: Theory and Applications in the Behavioral Sciences, Volume 1*, pages 69–104. Seminar Press, New York.

Young, F. W. (1981). Quantitative analysis of qualitative data. *Psychometrika*, 46:357–388.

Young, F. W. and Torgerson, W. S. (1968). TORSCA, a Fortran-IV program for nonmetric multidimensional scaling. *Behavioral Science*, 13:343–344.

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

Zangwill, W. I. (1969). *Nonlinear Programming: A Unified Approach*. Prentice-Hall, Englewood Cliffs, NJ.