MULTIDIMENSIONAL SCALING AND UNFOLDING

JAN DE LEEUW

ABSTRACT. We give an overview of the distance-based interpretation of multivariate data, and the corresponding techniques and algorithms to derive distance-based representations in low-dimensional Euclidean space.

1. INTRODUCTION

Multidimensional scaling (MDS) techniques are statistical techqnies that convert information about distances between a number of objects into a spatial representation of these objects. These techniques were first discussed systematically in psychometrics [Torgerson, 1958; Coombs, 1964] as multidimensional extensions of univariate psychophysics and sensory scaling. Because they needed considerable computational resources, they did not become widely used until he early seventies, when digital computers became available.

Marketing scientists were among the first users outside psychology. In particular, the books written by Paul Green and his co-workers [Green and Carmone, 1970; Green and Rao, 1972; Green et al., 1989] were among the very first applied multidimensional scaling books. In general terms, MDS is useful in marketing because, as Cooper [1983, p. 427] says in his useful review paper:

> Understanding the choices people make in the marketplace is the main goal of marketing research.

Date: September 18, 2005.

2000 *Mathematics Subject Classification.* 62H25.

Key words and phrases. Multivariate Analysis, Correspondence Analysis, Multidimensional Scaling, Unidimensional Scaling, Unfolding.

More specifically, Green [1975, page 27] indicates that MDS in the form of perceptual and preference mapping has been used to answer one or more of the following questions:

(1) What are the major perceptual and evaluative dimensins of the product class ?
(2) What existing brands are perceived as similar to what other exiting brands ?
(3) What are the major perceptual points of views among consumers ?
(4) What new brand possibilities are suggested by the configuration of existing brands ?
(5) How are respondent ideal points of preference vectors distributed in the various perceptual spaces ?
(6) How compatible are various advertising messages, slogans, or other types of promotional materils with brand perceptions ?

1.1. **Techniques.** Statistical techniques are defined on sets of possible data structures, and they map these data structures into numerical representations. The techniques in this chapter are used to construct geometrical representations from multivariate data structures.

The representations we consider are, in all cases, sets of points on the line, in the plane, or in some other low-dimensional Euclidean space. Representations are chosen in such a way that observed numerical or relational aspects of the data correspond with geometrical aspects of the representation. In this chapter we shall concentrate on mappings that portray relations in the data as Euclidean distances between the points in the representation. This defines an important class of techniques, much larger than one would expect, mostly because the notion of (dis)similarity is very general, and the notion of distance is quite intuitive.

The motivation for finding a geometrical representation of data is that "a picture is worth a thousand numbers". Or, equivalently, that summarizing

a large amount of data in a geometrical representation such as a scatterplot, tends to be more informative than just looking at the numerical data or computing a large number of descriptive statistics.

1.2. **Omissions.** The choices we have made exclude a number of related techniques, which are more properly discussed under the heading of "cluster analysis". In those technique the representation is in the form of a tree or a nested set of partitionings. Such representations are more algebraic and combinatorial, although they may have geometrical aspects. We also exclude the use of non-Euclidean distances, which have been used (on a limited scale) in some forms of multidimensional scaling. And, perhaps most importantly, we exclude techniques that portray relations in the data by other geometrical aspects such as inner products of vectors. These types of techniques are more appropriately discussed under the heading of "principal component analysis" or "factor analysis".

2. Data

Following Coombs [1964]; Shepard [1972]; Carroll and Arabie [1980] we can classify the various multidimensional scaling (or MDS) techniques by using a taxonomy of the various types of data they can be applied to. In the context of this chapter, all data are dissimilarity or similarity data, because data always provide us with information about the distance between the objects in our study. Entries in the data matrix, which we will write as Δ, are interpreted as approximate distances or as transformations of distances. In some case the entries only convey ordinal information about the distances, or we only want to use the ordinal information in the numerical data.

2.1. **Symmetric Matrices.** Distances are *symmetric*, and consequently dissimilarity data are often symmetric. They are also often *hollow*, which means Δ has a zero diagonal.

Data come directly in the form of a symmetric matrix of dissimilarities if we measure distances approximately, as in geodesics or in the conformation

of large molecules. They also occur if we ask individuals for similarity or dissimilarity estimates between pairs of objects. In both cases small deviations of similarity and hollowness are possible, but they are attributed to chance or measurement error.

In many experiments with human subjects, we do not ask for numerical dissimilarity estimates, but we ask the subjects to compare dissimilarities. Thus judgments are ordinal. In the methods of *triads*, for example, objects are presented as triples and subjects are asked to select the largest and/or the smallest dissimilarity. If dissimilarity is only used to determine order of the distances, then of course the diagonal is only required to contain the smallest elements (or, in the case of similarity, the largest ones).

2.1.1. *Derived Distance Matrices.* Symmetric matrices of distances are often computed from more basic data structures, such as multivariate data. If we have a rectangular matrix, we can compute distances between either the rows and the columns in various ways. Since multivariate data can have mixed measurement level, with numerical, ordinal, and nominal variables, we may have to use measure of distance which take this information into account [Gower and Legendre, 1986].

Computing *derived distance matrices* is the first step in many techniques. The most common example of a square symmetric similarity matrix, for example, is a correlation matrix. Usually, correlations are computed between variables (the columns of the multivariate data matrix), but in some cases it may make sense to compute them between individuals (the rows). Subtracting all correlations from one makes them into distances. MDS techniques can then be applied to these derived distance matrices.

There are other important examples in which we preprocess data into distance matrices before proceeding with further analysis. For paired comparison data, for example, in which p_{ij} is the probability that stimulus i is preferred to stimulus j, the Bradley-Terry-Luce model [Luce, 1959] gives

$$\delta_{ij} = \left| \log \frac{p_{ij}}{p_{ji}} \right| \approx |x_i - x_j|$$

Thurstone's Case V model [Thurstone, 1959] gives the same result by applying the inverse cumulative normal to the p_{ij}.

For binary item response data, in which p_{iJ} is the probability that item i is answered correctly and j is not and p_{Ij} is the probability that item j is answered correctly and i is not, the Rasch model [Fischer and Molenaar, 1995] gives

$$\delta_{ij} = \left| \log \frac{p_{iJ}}{p_{Ij}} \right| \approx |x_i - x_j|$$

In both cases estimating the probabilities by aggregating over individuals and computing the corresponding transformations gives symmetric data matrices that can be used as input for MDS.

2.2. **Square Asymmetric Matrices.** Distance information can also be derived from square asymmetric matrices, in which row and column objects are the same, but data are collected in such a way that there is no symmetry.

A common example is confusion matrices in identification experiments. Another one are input-output matrices, such as tourist traffic or import/-export between a number of countries. Yet another one is social interaction between a number of individuals in a classroom, or the interlocking board structure of a number of companies. In most of these example the data are frequency matrices, indicating frequency of interaction.

For frequency matrices $F = \{f_{ij}\}$ we often assume *quasi-symmetry*[Caussinus, 1965], which means in this context $f_{ij} \approx \alpha_i \beta_j \exp(-d_{ij}(X))$. This implies that if we define

$$\delta_{ij} = -\frac{1}{2} \log \frac{f_{ij} f_{ji}}{f_{ii} f_{jj}}$$

then $\delta_{ij} \approx d_{ij}(X)$ and we can apply ordinary metric MDS. The same transformation from frequencies to distances has also been proposed in psychology by Shepard [1957] and Luce [1963].

Asymmetry also occurs because of *row-conditionality*, in which each row has its own scale to measure distances, and there is no comparability between different rows. This happens, for instance, if we collect similarity

rankings for a number of consumer products, using each of the products in turn as a standard. If we have 20 cars, for instance, we produce 20 similarity rankings using each car as a pivot, and clearly the 20×20 matrix of rank orders is asymmetric.

In the case of asymmetric square matrices we can actually choose between two different types of representations. Consider the import/export example, for instance. We can compute a single configuration for the countries, or we can construct two different ones. There is an export configuration and an import configuration, and we only have information about the distances between those two. The distances within each of the two configurations are missing.

2.3. **Rectangular Matrices.** In non-square rectangular matrices the row objects and the column objects are two different sets. We collect distance information between sets, the within-set information is missing. Preference and voting data are a prominent example. From the preference information we know how close a consumer is to a product, but we have no direct information how close consumers are to each other, or how close products are to each other.

As we have seen above, it is quite common to apply MDS techniques to the derived distance matrices computed from multivariate data. This is not, however, what we have in mind in this section. It is also possible to apply MDS using the numerical or relational information in the rectangular data matrix directly, without aggregating to distances first.

2.4. **Three-way Data.** A somewhat less common, but still very important situation, involved more than one dissimilarity matrix [Arabie et al., 1987]. We could have a matrix Δ_k for each individual or group k. Two obvious ways to deal with such data are to aggregate them to a single matrix or to analyze each of them separately. The first alternative gets rid of all individual or group differences, and this may not be desirable. We throw away too much information, as it were, and do not introduce enough parameters. Separate analysis does not reflect the fact that the data presumably come

from related groups or individuals, and thus the solutions will tend to have structural properties in common. So here we do not incorporate enough prior information, and we have too many parameters. Special techniques intermediate between these two extremes will be discussed below.

3. Techniques

MDS involves minimizing complicated multivariate loss functions. In the section we discuss and compare some of the more common loss functions that have been proposed.

3.1. **Basics.** The computational MDS problem, in its simplest form, is to minimize the loss function

$$\sigma(X) = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(\delta_{ij} - d_{ij}(X))^2 \tag{1}$$

over X. Here $\Delta = \{\delta_{ij}\}$ are the data, collected in a matrix of dissimilarities, the matrix $D(X) = \{d_{ij}(X)$ has the corresponding Euclidean distances, and X is the $n \times p$ *configuration matrix*. Thus X is the set of n points in low-dimensional Euclidean space $\mathbb{R}^p$ that we are solving for. Clearly the δ_{ij} are interpreted as approximate distances and they are approximated by actual distances between n points in $\mathbb{R}^p$. This defines the simplest form of *metric* MDS, it uses the *stress* loss function (1), which was first introduced by Kruskal [1964a,b].

The w_{ij} in (1) are *weights*. They can be used for various purposes. If we have information about the relative measurement errors of the dissimilarities, then weights can be used to take this into account. If there are missing dissimilarities, then the corresponding weights can be set equal to zero.

3.2. **Transformation.** If we have similarities, or if the data can only be interpreted as approximate distances after a transformation, then we must minimize

$$\sigma(X) = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(\phi(\delta_{ij}) - d_{ij}(X))^2$$

For similarities between zero and one, for instance, we can use the transformation $\phi(\delta) = -\log(\delta)$. Of course after we have applied the transformation to the data, we are back to simple metric MDS.

But matters are more complicated if we do not have a single fixed transformation such as the negative logarithm to work with. In some cases we want to allow for a class of possible transformations Φ such as polynomials, splines, or monotone transformations. In that case the problem, in the Kruskal formulation, becomes to minimize

$$\sigma(X, \phi) = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(\phi(\delta_{ij}) - d_{ij}(X))^2$$

over the configuration matrices X and the transformations $\phi \in \Phi$. The case in which Φ is the class of all monotone transformations, suitably normalized, is usually called *non-metric* MDS. If Φ is the class of linear transformations, we often refer to the corresponding MDS problem as the *additive constant problem*.

The result of our analysis will not only be the optimal configuration of points in $\mathbb{R}^p$, but it will also be the optimal transformation from Φ of the dissimilarities or similarities. Finding a transformation of the data which optimizes the fit of a (usually geometric) model is often called *optimal scaling* [Young, 1981]. In non-metric MDS the optimal transformation is computed by *monotone regression*. Plotting the optimal transformation of the data in the scatterplot of dissimilarties versus the optimal distances gives us the *Shepard diagram*, after Shepard [1962a,b].

The straightforward way of dealing with row-conditionality is to introduce, and to optimize over, separate transformations for each row.

$$\sigma(X, \phi_1, \cdots, \phi_n) = \sum_{i=1}^{n}\sum_{j=1}^{n}(\phi_i(\delta_{ij}) - d_{ij}(X))^2$$

Each row will have its own Shepard diagram. Observe, however, that this will tend to introduce many additional parameters and can easily lead to instability of the solutions.

3.3. **Alternative Loss Functions.** Sometimes, for computational or other reasons, it may make sense to minimize

$$\sigma_\psi(X) = \sum_{i=1}^n \sum_{j=1}^n w_{ij}(\psi(\delta_{ij}) - \psi(d_{ij}(X)))^2$$

Here we do not fit distances to data, but we fit transformed distances to transformed data. The most common special case has $\psi(d) = d^2$, creating a loss function $\sigma_2(X)$ which is often called *sstress* [Takane et al., 1977]. The reason for squaring the distances is both conceptual and computational. Squared distances are quadratic functions, and thus *sstress* is a multivariate quartic polynomial [Browne, 1987]. Multivariate polynomials are smoother and in some ways easier to handle computationally than *stress*, which is not even differentiable in configurations for which one or more of the $d_{ij}(X)$ are equal to zero.

Another important special case has $\psi(d) = \log(d)$. This transformation is mainly used for statistical reasons [Ramsay, 1977]. One can argue that the logarithm will stabilize the variance of the dissimilarities under some fairly reasonable scenarios, and thus using log-dissimilarities will lead to techniques with simpler statistical properties.

Observe that if the dissimilarities and the distances are close, which means that we have a good fit, then

$$\sigma_\psi(X) \approx \sum_{i=1}^n \sum_{j=1}^n w_{ij}\{\psi'(\delta_{ij})\}^2(\delta_{ij} - d_{ij}(X))^2.$$

Thus we can approximate the minimizing solution by a simple modification of the weights.

3.4. **Torgerson Transform.** For the historical point of view, the most important type of transformation of the dissimilarities is the Torgerson tranformation $\tau(\Delta)$ [Torgerson, 1958; Critchley, 1988]. It is apply to the whole matrix of squared dissimilarities, which we write as $\Delta \star \Delta$, and is consequently more complicated than the simple element-by-element transformations ψ

we have discussed so far.

$$\tau_{ij}(\Delta \star \Delta) = -\frac{1}{2}(\delta_{ij}^2 - \delta_{i\bullet}^2 - \delta_{\bullet j}^2 + \delta_{\bullet\bullet}^2)$$

Bullets are used for averages. In words, we square the dissimilarities, and then we subtract from each element the row and column mean of the squared matrix, and add the overall mean. In other words, we double-center the matrix of squared dissimilarities. Clearly τ is linear in the squared distances.

It may seem slightly mysterious why this transformation is applied, but the motivation becomes clear if we apply the Torgerson transformation to the squared distance matrix $E(X) = D(X) \star D(X)$. Then some algebra gives $\tau_{ij}(E(X)) = \tilde{X}\tilde{X}'$, where $\tilde{X}$ is centered X, or X in deviations from its column means. That means τ (linearly) transforms squared distance matrices to inner product matrices.

The loss function

$$\sigma_\tau(X) = \sum_{i=1}^n \sum_{j=1}^n w_{ij}(\tau_{ij}(\Delta \star \Delta) - \tau_{ij}(E(X)))^2 \tag{2}$$

is often called *strain*. In the case of equal weights, we can write *strain* in matrix notation as

$$\sigma_\tau(X) = \mathbf{tr}\ (\tau(\Delta \star \Delta) - XX')^2$$

and this can be minimized [Householder and Young, 1938] by spectral (or eigen) decomposition. In fact, if $\tau(\Delta \star \Delta) = K\Lambda K'$ is this eigen-decomposition, with the eigenvalue decreasing along the diagonal, then the solution for the best p-dimensional configuration is $X_p = K_p[\Lambda_p]_+^{1/2}$. Here K_p contains the first p columns of K and $[\Lambda]_p$ is the positive part of the leading order p submatrix of Λ. Eigen decompositions are comparatively inexpensive to compute, and thus in this respect *strain* has some clear advantages over *stress* and *sstress*. This remains true even in the case of arbitrary non-constant weights [Gabriel and Zamir, 1979].

If Δ is a Euclidean distance matrix, then $\tau(\Delta \star \Delta)$ is positive semi-definite. Thus the diagonal of Λ in $\tau(\Delta \star \Delta) = K\Lambda K'$ is non-negative. Define $X_p = K_p\Lambda_p^{1/2}$ and define $d_{ij(p)}^2 = (e_i - e_j)'X_pX_p'(e_i - e_j)$ as the squared distances

between the rows of X_p. Here e_i and e_j are unit vectors (columns of the identity matrix). Then

$$d^2_{ij(1)} \leq d^2_{ij(2)} \leq \cdots \leq d^2_{ij(r)} = d^2_{ij}(X),$$

with r the rank of $\tau(\Delta \star \Delta)$. Thus we see that the squared distances, and thus the distances themselves, are approximated from below, and that each successive dimension brings the fitted distances closer to the actual distances.

Now suppose that we compute dissimilarities as derived distances using a formula of the form

$$\delta^2_{ij} = (e_i - e_j)'A(e_i - e_j),$$

with A positive semi-definite. We want to find X such that $d^2_{ij}(X) = \delta^2_{ij}$. The previous construction, with $A = K\Lambda K'$ and $X = K\Lambda^{1/2}$ gives the solution, and in the same way we can approximate the dissimilarities from below.

A simple example are squared Euclidean distances between the rows of a multivariate data matrix Z, because

$$\delta^2_{ij} = (x_i - x_j)'(x_i - x_j) = (e_i - e_j)'ZZ'(e_i - e_j).$$

Gower and Legendre [1986] give more examples of derived distance matrices of this form that apply even if the variables in Z are not numerical. Below, we shall also discuss the derived distances used in correspondence analysis.

4. Special Forms of MDS

4.1. **Metric and Nonmetric Unfolding.** Unfolding simply applies MDS to an off-diagonal matrix. Thus in the metric case we minimize

$$\sigma(X, Y) = \sum_{i=1}^{n}\sum_{j=1}^{m} w_{ij}(\delta_{ij} - d_{ij}(X, Y))^2$$

over the row-configuration X and the column-configuration Y. This does not seem to introduce any new aspects, in fact it could be incorporated simply in our previous MDS techniques by using zero weights for the diagonal blocks.

Nevertheless, unfolding is often discussed separately, because it is important in practice. It is perhaps the most common way to analyze unaggregated preference data and rank orders. And because the sparsity of the data, which large blocks of missing information, unfolding creates some unique computational problems, which are still being studied.

For the purposes of this chapter, however, it suffices to observe that there is metric and non-metric unfolding, that we can unfold using *stress, sstress* and *strain*. There is both unidimensional and multidimensional unfolding, and most of our other previous remarks and distinctions also apply in this context.

In most cases, unfolding techniques are used row-conditionally. Since the data are already sparse, this often leads to unstable or even degenerate solutions. Not only do we miss the two diagonal blocks, we also miss the comparisons between different rows of the off-diagonal block.

As we have indicated above, square asymmetric can also be unfolded. We then have different configurations for the row and columns points, even though they may refer to the same objects.

4.2. **Unidimensional Scaling.** Much of classical psychophysics and psychometrics uses aggregated data to construct one-dimensional scales [Guilford, 1954]. MDS methods can also be used in this context. We have seen that Bradley-Terry-Luce, Thurstone, and Rasch models specify transformations of choice probabilities to distance on a one-dimensional scale.

The basic loss function for Unidimensional Scaling (UDS) is

$$\sigma(x) = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(\delta_{ij} - |x_i - x_j|)^2.$$

If we write $|x_i - x_j| = s_{ij}(x)(x_i - x_j)$, where $s_{ij}(x) = \textbf{sign}(x_i - x_j)$, then we see that if the rank-order, and thus the sign-matrix, of x is fixed, then minimizing *stress* is a linear regression problem. To find the minimum, we have to

look at all possible rank orders and minimize *stress* over all x in the appropriate order, which means we have to solve a simple *monotone regression* problem [Barlow et al., 1972] for each of the possible rank orders.

Of course if n is at all large, the number of possible rank orders is simply to large for this recipe to be practical. In this case methods combining search with clever heuristics are needed [Hubert et al., 2002].

4.3. **Full-dimensional Scaling.** In MDS we usually look for the best fitting configuration in a given dimensionality p. Suppose we drop this last constraint and consider the problem of minimizing *stress* over all Euclidean distance matrices. It is slightly more convenient to minimize over all matrices E which are matrices of *squared* Euclidean distances. Thus

$$\sigma(E) = \sum_{i=1}^n \sum_{j=1}^n w_{ij}(\delta_{ij} - \sqrt{e_{ij}})^2.$$

Now observe that

$$\sigma(E) = \sum_{i=1}^n \sum_{j=1}^n w_{ij}\delta_{ij}^2 - 2\sum_{i=1}^n \sum_{j=1}^n w_{ij}\delta_{ij}\sqrt{e_{ij}} + \sum_{i=1}^n \sum_{j=1}^n w_{ij}e_{ij}.$$

The first term is independent of E, the second term is convex in E, and the last term is linear in E. Thus $\sigma(E)$ is convex. Because the set of all squared Euclidean distance matrices is convex cone in the space of all nonnegative, symmetric, hollow matrices, we see that the Full Dimensional Scaling (FDS) problem is a convex programming problem, which has no non-global local minima.

4.4. **Three-way Scaling.** With dissimilarity matrices $\Delta_1, \cdots, \Delta_m$ the obvious generalization of *stress* is

$$\sigma(X_1, \cdots, X_m) = \sum_{k=1}^m \sum_{i=1}^n \sum_{j=1}^n w_{ij}(\delta_{ijk} - d_{ij}(X_k))^2 \tag{3}$$

As we indicated before, we can minimize this over all X_k, or we can minimize it under the constraint $X_1 = \cdots = X_m$, which amounts to aggregating the dissimilarities.

But the problem only becomes interesting if we find parametrizations of the configurations that make them partially identical. The most common one is $X_k = X\Phi_k$, with Φ_k diagonal. This are the constraints proposed, in the *strain* context, by Carroll and Chang [1970]. This means configurations are the same for all k, except for the fact that the axes are differentially shrunken and stretched. This eliminates rotational indeterminacy from MDS, because rotated configurations do not satisfy the constraints $X_k = X\Phi_k$ any more. Two less popular three-way specifications, with more parameters, are $X_k = X\Phi_k$, with Φ_k full, and $X_k = X\Phi_k Z$, with Φ_k diagonal.

4.5. **Constrained Scaling.** In ordinary MDS we minimize *stress* over all $n \times p$ configurations. But in many practical applications, reviewed by De Leeuw and Heiser [1980a], it make sense to constrain the configuration. We might want to require that the points are on a circle, or a regular grid, that the configuration matrix is in the linear span of a given matrix, or that certain points are kept fixed in certain locations. We have seen that three-way scaling also imposes constraints on the configurations.

Another important example is when the $n \times (q + n)$ matrix X can be partitioned as $X = (Z \mid \Phi)$, with Φ diagonal. It follows that $d_{ij}^2(X) = d_{ij}^2(Z) + \phi_i^2 + \phi_j^2$, which can be thought of as the MDS version of factor analysis. In the same way we can impose various simplex and circumplex constraints on the configuartion. See De Leeuw and Heiser [1980a] for details, additional examples, and for algorithms.

4.6. **Correspondence Analysis.**

4.7. **Multiple Correspondence Analysis.**

5. Algorithms

Computationally we solve an MDS problem by minimizing the loss function over the configurations. This is a high-dimensional optimization problem, and most computer packages have general-purpose routines that can be used to compute local minimizers of arbitrary functions. It is generally

more efficient, however, to devise special purpose optimization routines that take the specific properties of the loss function into account.

Much of the published research on minimizing *stress* and its variations is technical, and outside the scope of this chapter. We shall only discuss some of the basic results, and refer to the literature for the details.

We know that $d_{ij}^2(X) = (e_i - e_j)'XX'(e_i - e_j)$ and this can also be written as $d_{ij}^2(X) = \mathbf{tr}\ X'A_{ij}X$, where A_{ij} is the matrix $(e_i - e_j)(e_i - e_j)'$. Thus A_{ij} has the (i, i) and (j, j) diagonal elements equal to $+1$ and the (i, j) and (j, i) off-diagonal elements equal to -1. All other elements are zero. Define

$$V = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}A_{ij},$$

$$B(X) = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\frac{\delta_{ij}}{d_{ij}(X)}A_{ij}.$$

The basic algorithm to minimize *stress* is to iteratively replace X by the update $V^+B(X)X$, with V^+ the generalized inverse of V. The global convergence of this algorithm was proved by De Leeuw [1977], and the algorithm was shown to have a linear convergence rate by De Leeuw [1988]. It may seem that the algorithm should run into problems when some of the distances become zero, but there is an easy modification which avoids the problem, and De Leeuw [1984] shows that at a local minimum of *stress* the distances are always positive anyway. This algorithm can be generalized to unfolding, to unidimensional scaling, to full-dimensional scaling, to three-way scaling, and to constrained MDS. See Borg and Groenen [1997] for details.

Minimizing *sstress* can be done in many ways. In Takane et al. [1977] we find the best fitting configuration by cycling over all np coordinates of the configuration. Because *sstress* is a quartic, finding an optimal coordinate with all other coordinates fixed at their current value means finding a root of a one-dimensional cubic. More efficient algorithms have been studied by Browne [1987], and for unfolding by Greenacre and Browne [1986], but further improvements may still be possible.

We have seen that minimizing *strain*, at least in the unweighted case, means solving an eigen problem. Unfortunately that is no longer true in the weighted case, or in unfolding, or in three-way scaling. Alternating least squares algorithms, cycling over subsets of the parameters, are relatively easy to construct, but *strain* loses its computational advantage over *stress* and *sstress*.

6. Criticism

6.1. **Assumptions.** The role of assumptions in MDS techniques is minimal. We are making a picture of a data structure. The picture can be informative or it can be disappointing. The distances can approximate the numerical or relational information in the data well or poorly. But none of this has anything to do with assumptions.

It is true that there have been some attempts to embed MDS into one of the standard statistical replication frameworks. In Ramsay [1977], for instance, maximum likelihood techniques for non-linear regression are used. In the derived distance methods for frequency tables and paired comparison we compute distances from the frequencies, and delta method calculations can be used to compute their standard errors. In fact, frequency models can be formulated as multinomial models for contingency tables, and they can be fitted by minimizing the corresponding deviances.

6.2. **Appropriateness.** MDS methods are especially appropriate if there are external reasons, or if there is theory that makes it natural to think of distance to represent aspects of the data. If we are measuring actual distances with error, then MDS seems natural. If there is a psychological theory of stimulus generalization based on distance, such as Shepard [1957], then again using MDS makes perfect sense. The same thing is true for choice theories based on single peaked preference functions, especially if they are cast in the geometrical framework of Coombs [1964].

If the objects we are scaling have an underlying geometric structure, then MDS will tend to be more informative. This is true for regions of the brain,

for countries exchanging tourists, for imports and exports, for locations between which we measure travel times, and for genes located on a chromosome. See De Leeuw and Heiser [1980b] for more discussion of this and related issues.

6.3. **Validation.** In most applications there is no complete statistical replication framework, so the usual forms of inference are not relevant. We have seen that derived distances from frequency tables are an exception, but we must emphasize that a lot of parameters are involved and the asymptotic results may not be even close. And even if they were, the main function of MDS remains data reduction and visualization, not formal inference.

Despite of this, the more general notion of *stability* remains of great importance. We can look at the second derivatives of our various loss functions to draw stability ellipsoids around our points. We can use influence function techniques to study the effects of single dissimilarities on the solution. If we have multivariate data, or some other notion of replication, then resampling becomes possible and the bootstrap can be used to study stability. Even in straightforward MDS it is possible to construct special versions of the Jackknife [De Leeuw and Meulman, 1986] to derive global stability information about the recovered configuration.

7. Examples

There are many succesfull examples of mutlidimensional scaling, unfolding, and correspondence analysis in marketing. For scaling and unfolding this is ampy illustrated in the classic books by Paul Green and his coworkers [Green and Carmone, 1970; Green and Rao, 1972; Green et al., 1989]. For simple correspondence analysis we refer to the equally classical paper by Hoffman and Franke [1986] and for multiple correspondence analysis to the chapter by Hoffman et al. [1994].

References

P. Arabie, J. D. Carroll, and W.S. DeSarbo. *Three-way Scaling and Clustering*, volume 65 of *Quantitative Applications in the Social Sciences*. Sage Publications, 1987.

R.E. Barlow, R.J. Bartholomew, J.M. Bremner, and H.D. Brunk. *Statistical Inference under Order Restrictions*. Wiley, New York, 1972.

I. Borg and P.J.F. Groenen. *Modern Multidimensional Scaling: Theory and Applications*. Springer, New York, 1997.

M.W. Browne. The Young-Householder Algorithm and the Least Squares Multdimensional Scaling of Squared Distances. *Journal of Classification*, 4:175–190, 1987.

J.D. Carroll and P. Arabie. Multidimensional Scaling. *Annual Review of Psychology*, 31:607–649, 1980.

J.D. Carroll and J.J. Chang. Analysis of Individual Differences in Multidimensional Scaling Via an N-way Generalization of Eckart-Young Decomposition. *Psychometrika*, pages 283–319, 1970.

H. Caussinus. Contribution à l'Analyse Statistique des Tableaux de Corrélation. *Annals Faculté Sciences Université de Toulouse*, 29:77–182, 1965.

C. H. Coombs. *A Theory of Data*. Wiley, 1964.

L. G. Cooper. A Review of Multidimensional Scaling in Masrketing Research. *Applied Psychological Measurement*, 7:427–450, 1983.

F. Critchley. On Certain Linear Mappings Between Inner Product and Squared Distance Matrices. *Linear Algebra and its Applications*, 105: 91–107, 1988.

J. De Leeuw. Convergence of the Majorization Method for Multidimensional Scaling. *Journal of Classification*, 5:163–180, 1988.

J. De Leeuw. Differentiability of Kruskal's Stress at a Local Minimum. *Psychometrika*, 49:111–113, 1984.

J. De Leeuw. Applications of Convex Analysis to Multidimensional Scaling. In J.R. Barra, F. Brodeau, G. Romier, and B. Van Cutsem, editors, *Recent developments in statistics*, pages 133–145, Amsterdam, The Netherlands, 1977. North Holland Publishing Company.

J. De Leeuw and W. J. Heiser. Multidimensional Scaling with Restrictions on the Configuration. In P.R. Krishnaiah, editor, *Multivariate Analysis, Volume V*, pages 501–522, Amsterdam, The Netherlands, 1980a. North Holland Publishing Company.

J. De Leeuw and W. J. Heiser. Theory of Multidimensional Scaling. In P.R. Krishnaiah, editor, *Handbook of Statistics, Volume II*. North Holland Publishing Company, Amsterdam, The Netherlands, 1980b.

J. De Leeuw and J. Meulman. A Special Jackknife for Multidimensional Scaling. *Journal of Classification*, 3:97–112, 1986.

G.H. Fischer and I.W. Molenaar, editors. *Rasch Models. Foundations, Recent Developments, and Applications*. Springer, 1995.

K.R. Gabriel and S. Zamir. Lower Rank Approximation of Matrices by Least Squares with Any Choize of Weights. *Technometrics*, 21:489–498, 1979.

J.C. Gower and P. Legendre. Metric and Euclidean Properties of Dissimilarity Coefficients. *Journal of Classification*, 3:5–48, 1986.

P. E. Green. Marketing Applications of MDS: Assessment and Outlook. *Journal of Marketing*, 39:24–31, 1975.

P.E. Green and F.J. Carmone. *Multidimensional Scaling and Related Techniques in Marketing Analysis*. Allyn and Bacon, 1970.

P.E. Green and V.R. Rao. *Applied Multidimensional Scaling*. Holt, Rinehart & Winston, 1972.

P.E. Green, F.J. Carmone, and S. Smith. *Multidimensional Scaling: Concepts and Applications*. Allyn and Bacon, 1989.

M.J. Greenacre and M.W. Browne. An Efficient Alternating Least-Squares Algorithm to Perform Multidimensional Unfolding. *Psychometrika*, 51: 241–250, 1986.

J.P. Guilford. *Psychometric Methods*. McGrawHill, second edition, 1954.

D.L. Hoffman and G. Franke. Correspondence Analysis: The Graphical Representation of Categorical Data in Marketing Research. *Journal of Marketing Research*, 23:213–227, 1986.

D.L. Hoffman, J. De Leeuw, and R.V. Arjunji. Multiple Correspondence Analysis. In R.P. Bagozzi, editor, *Advanced Methods of Marketing Research*. Blackwell, 1994.

A.S. Householder and G. Young. Matrix Approximation and Latent Roots. *American Mathematical Monthly*, 45:165–171, 1938.

L.J. Hubert, P. Arabie, and J.J. Meulman. Linear Unidimensional Scaling in the L_2-Norm: Basic Optimization Methods Using MATLAB. *Journal of Classification*, 19:303–328, 2002.

J. B. Kruskal. Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis. *Psychometrika*, 29:1–27, 1964a.

J.B. Kruskal. Nonmetric Multidimensional Scaling: a Numerical Method. *Psychometrika*, 29:115–129, 1964b.

R.D. Luce. Detection and Recognition. In R.D. Luce, R.R. Bush, and E. Galanter, editors, *Handbook of Mathematical Psychology*, volume 1, chapter 3, pages 103–189. Wiley, 1963.

R.D. Luce. *Individual Choice Behavior*. Wiley, 1959.

J. O. Ramsay. Maximum Likelihood Estimation in MDS. *Psychometrika*, 42:241–266, 1977.

R.N. Shepard. The Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function (Part I). *Psychometrika*, 27:125–140, 1962a.

R.N. Shepard. The Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function (Part II). *Psychometrika*, 27:219–246, 1962b.

R.N. Shepard. Stimulus and Response Generalization: A Stochastic Model Relating Generalization to Distance in Psychological Space. *Psychometrika*, 22:325–345, 1957.

R.N. Shepard. A Taxonomy of Some Principal Types of Data and of the Multidimensional Methods for their Analysis. In R.N. Shepard, A.K. Romney, and S.B. Nerlove, editors, *Multidimensional Scaling. Volume I, Theory*, pages 23–47. Seminar Press, 1972.

Y. Takane, F.W. Young, and J. De Leeuw. Nonmetric Individual Differences in Multidimensional Scaling: An Alternating Least Squares Method with

Optimal Scaling Features. *Psychometrika*, 42:7–67, 1977.

L.L. Thurstone. *The Measurement of Values*. University of Chicago Press, 1959.

W.S Torgerson. *Theory and Methods of Scaling*. Wiley, New York, 1958.

F. W. Young. Quantitative Analysis of Qualitative Data. *Psychometrika*, 46: 357–388, 1981.

Department of Statistics, University of California, Los Angeles, CA 90095-1554

E-mail address, Jan de Leeuw: deleeuw@stat.ucla.edu

URL, Jan de Leeuw: http://gifi.stat.ucla.edu