The Past, Present, and Future of Multidimensional Scaling

Patrick J.F. Groenen, Ingwer Borg

Econometric Institute Report EI 2013-07

Abstract

Multidimensional scaling (MDS) has established itself as a standard tool for statisticians and applied researchers. Its success is due to its simple and easily interpretable representation of potentially complex structural data. These data are typically embedded into a 2-dimensional map, where the objects of interest (items, attributes, stimuli, respondents, etc.) correspond to points such that those that are near to each other are empirically similar, and those that are far apart are different. In this paper, we pay tribute to several important developers of MDS and give a subjective overview of milestones in MDS developments. We also discuss the present situation of MDS and give a brief outlook on its future.

Multidimensional scaling (MDS) has become one of the core multivariate analysis techniques discussed in any standard data analysis, multivariate analysis, or computer science text book. A search in the Thomson Reuters Web of Science on the topic "multidimensional scaling" yielded 5,186 papers that were cited in 68,429 other papers (per January 2013). This clearly shows that MDS is an established multivariate analysis technique.

Several important milestones in the development of MDS can be distinguished and the present paper is a subjective interpretation of that. As the present authors have been working in the area since the 1970s, they have developed their own subjective view of what they consider to be milestones in the development of MDS. The emphasis here lies on algorithmic milestones as they have cleared the way for practical use. We do not intend to provide an exhaustive overview of the history of MDS as that could easily require a book by itself (for further details, see Borg and Groenen (2005)).

The remainder of this paper is organized both chronologically and per topic. We roughly distinguish three periods: past (until 1980), present (1980-2000), future (from 2000). Even though the future necessarily lies ahead, it always takes time for developments to be used by a wider audience, which explains the lag of about 15 years. Table 1 gives an overview of these subjective milestones of the authors, which are discussed in more detail in the subsequent sections.

1 The basic ideas of Multidimensional Scaling

The core idea of MDS is explained by the first sentence in Borg and Groenen (2005): "Multidimensional scaling (MDS) is a method that represents measurements of similarity

Table 1: Subjective overview of milestones in MDS.

Years	Main author(s)	Topic
Past		
1958, 1966	Torgerson,Gower	Classical MDS
1962	Shepard	First MDS heuristic
1964	Kruskal	Least-squares MDS through Stress with transformations
1964	Guttman	Facet theory and regional interpretations in MDS
1969, 1970	Horan, Carroll	Three-way MDS models (INDSCAL, IDIOSCAL)
1977-	De Leeuw and others	The majorization algorithm for MDS
Present		
1986-1998	Meulman	Distance-based MVA through MDS
1994	Buja	Constant dissimilarities
1978, 1995-	Various	Local minimum problem
1998	Buja	Smart use of weights in MDS
Future		
1999-,	Heiser, Meulman, Busing	Modern MDS software: Proxscal in SPSS (PASW)
2000	Tenenbaum, et al.	Large scale MDS ISOMAP heuristic
2002	Buja, Swayne, Cook	Dynamic MDS in GGvis (part of GGobi)
2003	Groenen	Dynamic MDS visualization through iMDS
2005-	Groenen, Trosset, Kagie	Large scale MDS through Stress
2002	Denœux, Masson, Groenen, Winsberg, Diday	Symbolic MDS of interval dissimilarities
2006	Groenen, Winsberg	Symbolic MDS of histograms
2009	De Leeuw, Mair	Smacof package in R

Table 2: Part of the confusion table of Morse signals (Rothkopf, 1957).

Morse		Sign					
Code	Sign	A	B	C	D	...	0
·–	A	92	4	6	13	...	3
–···	B	5	84	37	31	...	4
–·–·	C	4	38	87	17	...	12
–··	D	8	62	17	88	...	6
⋮	⋮	⋮	⋮	⋮	⋮	⋱	⋮
–––––	0	9	3	11	2	...	94

(or dissimilarity) among pairs of objects as distances among points of a low-dimensional multidimensional space." Thus, instead of the usual cases–by–variables data, the data of MDS consist of measurements of (dis)similarity among pairs of objects, collectively called "proximities". Objects could be persons, attributes, stimuli, countries, etc., and the measurements may be correlations of test items, similarity of politicians, dissimilarity of mobile telephones, etc. The overall goal is to represent these objects as points in a low-dimensional (usually 2-dimensional) space such that the distances among the points represent the (dis)similarities as good as possible. The motive for doing this is to visualize the data in a "picture" that makes the data structure much more accessible to the researcher than the data matrix with its many numbers.

As a classic example, consider how test persons confuse acoustic Morse signals (Rothkopf, 1957). The research question here is to detect psychological rules that govern what is and what is not confused. There are 36 Morse signals, 26 for the letters in the alphabet, and 10 for numbers. The task of test persons was to judge whether the Morse signals in a particular pair of signals seem to be the "same" or "different". Each pair was presented in two orders: first A (di-da or '·–') and then D (da-di-di or '– ···'), for example, and also first D and then A. Each of 598 subjects, who were unfamiliar with Morse codes, judged 351 pairs. Table 2 exhibits a part of the full 36×36 matrix of confusion rates. Note that these data are similarities. As distances are always symmetric, asymmetries are considered errors in a distance model and the MDS is done on the symmetrized data matrix. Moreover, because the distance of a point to itself is always zero, the diagonal of the data matrix is ignored. Figure 1 shows the (ordinal) MDS configuration representing these data. It exhibits that the confusion rates are systematically related to the signals' physical properties. The North-West vs. South-East direction is correlated with the signals' lengths, with long signals in the North-West corner and short signals in the South-East direction. The vertical scatter of the points is related the signals' compositions of di's and da's, with da's becoming more dominant as one moves upwards. Hence, MDS here succeeded to uncover two psychophysical regularities that are difficult, if not impossible, to discern in the numerical data.

This example shows that MDS essentially transforms a matrix of dissimilarities into a low-dimensional map that, as much as this is possible, approximates the dissimilarities

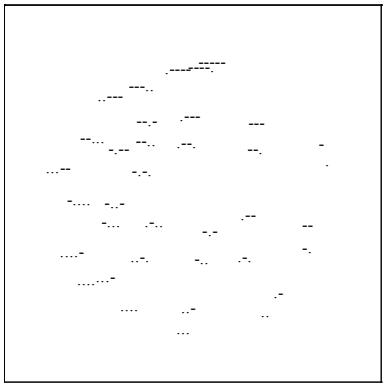

Figure 1: MDS solution of ordinal MDS on the Rothkopf confusion data (Table 2).

by distances of points representing the objects.

2 Motives for MDS: A historical account

MDS was not invented by statisticians. It was first developed to solve specific scaling problems that arose in practical and scientific contexts. In the following, we outline some of these developments.

2.1 Early MDS in geography

The first traces of MDS can be found in the 17th century. Figure 2 shows a small table of distances among several towns and villages in Durham county, England. The order of the row and column towns is reversed so that an unusual matrix of distances appears that is symmetric over the lower left upper right diagonal. This diagonal does not contain the zero distances of a town to itself, but contains the distances to London. Apart from this table, Figure 2 also shows the geographical map of Durham county. It is considered the first instance of showing both a table of distances and the map that corresponds to these distances in a single figure (Gower, personal communication). Therefore, this case can be seen as a predecessor of MDS. (Note that this map is one of a series covering the counties of England, made by the Dutch cartographer Jacob van Langren in 1635.)

Modern MDS is not concerned with cartography. Rather, similar to factor analysis, it evolved as a model for certain psychological phenomena, and only later became more and more popular as a general-purpose data-analytic tool. Historically, MDS can be related to at least four different purposes.

2.2 The distance formula as a psychological model of (dis)similarity judgments

The notion that (dis)similarity judgments can normally be modeled as distances has been around in psychology for quite some time. It seems obvious that persons generate

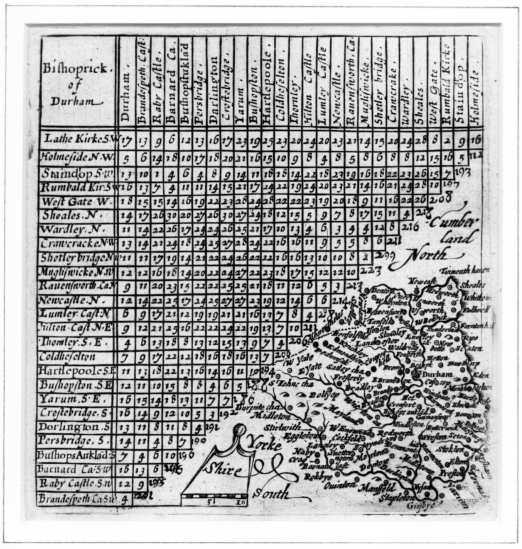

Figure 2: Map of Durham county by Jacob van Langren in 1635.

(dis)similarity judgments for pairs of objects in a process that closely mimics the natural distance function in a Cartesian space. That is, if a judgment is needed, the person forms a mental representation of the objects in "psychological space", a space spanned by the objects' attributes and with the objects corresponding to points in this coordinate system. Dissimilarity judgments are then formed by first assessing, dimension by dimension, the differences of each pair of points, and then summing these intra-dimensional differences. This generates a global distance, the basis for an overall (dis)similarity impression or rating of the respective objects. To the outside observer—and possibly also to the person him- or herself—the psychological space itself is unknown (underlying, latent), but MDS promises to "uncover" it—as Kruskal, one of the MDS pioneers, claimed—from the individual's overall (dis)dissimilarity judgments. To do this, classical MDS first assumes that the given judgments are ratings on a metric scale, and that the distance function of the psychological space is the Euclidean distance.

It was, however, not before MDS algorithms were developed that allowed to process not metric but ordinal data before this distance model received a lot of attention. This met with the Zeitgeist of the late 60's which emphasized ordinal data. Later, the Euclidean distance formula was also generalized to the more general Minkowski metric. This metric can be seen as a family of distances where the intra-dimensional differences are weighted in proportion to their absolute size before they are summed. Minkowski metrics range from weights of 1 for all intra-dimensional differences (city-block distance) to weights that are so extreme that the largest difference is essentially equal to the total sum (dominance metric). Series of studies were done to investigate what particular Minkowski metric was most suitable for what kind of context (e.g., judgments under time pressure, perception of analytic vs. integral stimuli). Then, it was argued that Minkowski spaces have local validity only because (dis)similarity judgments in psychol-

ogy are subadditive. To account for this empirical lawfulness, Schönemann proposed an MDS model with a bounded geometry. However, this idea was essentially ignored by psychological research and not pursued much further by psychometricians.

Besides such refinements of the similarities-explained-by-distances model, attention was also turned to modeling individual differences in MDS. This produced a hugely popular model, often identified with a particular computer program called INDSCAL. It assumes that different individuals differ in how they weight the same set of dimensions of a common space. Numerous applications in the social sciences used this model, since it promised to identify "the" dimensions uniquely, until step-by-step approaches (e.g. PINDIS) showed that rotations of the group space often are only slightly less successful explaining the data. Moreover, the individual dimension weights can be deceptive in the sense that they scatter a lot without explaining much more variance than unit weights. The biggest mistake when using INDSCAL modeling is, however, comparing individual weights all too loosely. These weights depend on the (arbitrary!) norming of the group space, so that only the order of the weights of different persons for the same dimension can be compared, while market researchers, in particular, had hoped that INDSCAL would show them what dimensions are most important in product perception. When these restrictions became clearer, INDSCAL became less important in applied research.

A further line of research used the MDS method to solve Coombs' unfolding model, where the data are not (dis)similarities but preferences of different individuals for the same set of objects. In unfolding, persons are represented as points in space, and choice objects by other points, and the distances among these two types of points represent the preference data. Each person point is taken as this person's point of maximal preference ("ideal point"), and circles about each ideal point as iso-preference contours. Unfolding was used a lot to model voting behavior, for example, but it was soon discovered that theory-guided multiple uni-dimensional unfolding for different subgroups can be superior to exploratory multi-dimensional unfolding of the total sample. Hence, unfolding's popularity as a model of preferential choice dropped in importance.

This type of research where the MDS geometry and its distance function are taken as psychological models has considerably advanced the understanding of human perception, judgment, and preference. In particular, it has become clear under what conditions such models yield good descriptions of empirical phenomena, and when they do not. Today, research where general MDS plays a major role as a psychological model are over.

2.3 Ordinal MDS as a response to the premise that measurement in psychology must build on non-metric data

In the late 60's, methodologists were much concerned with "Foundations of Measurement" (Krantz et al., 1971). The cardinal premise of this research initiative was that numerical judgments (mostly "ratings" on, say, a scale from 0 to 10) cannot automatically be assumed to be real numbers. Rather, it was argued that real-valued measurements must be constructed and, first of all, justified by testing typically large sets of pair-wise ordinal judgments that, together with some technical assumptions, establish structure-preserving maps of relational into numerical systems (homomorphisms). To respond to

this measurement philosophy, scaling methodologists felt driven to replace the classical MDS of Torgerson and Gower–which assumed metric data as input–by ordinal MDS (or, as it was called at that time) by "non-metric" MDS.

Kruskal and Guttman (with Lingoes) developed computer programs for ordinal MDS. They both used gradient-based minimization to optimize the point coordinates, but Kruskal did this in combination with ordinal regression of the data onto the distances, while Guttman invented "rank images" as targets, a method that is less likely to yield degenerate solutions. There are other technical differences (later harmonized in a best-of-both-world's program called MINISSA), but the main difference between Kruskal and Guttman was how they approached an MDS solution. Kruskal (as most users of MDS at that time) first of all asked: "What do the dimensions mean?" Guttman, in contrast, was content-driven. For him content came first and methods only served as tools to build substantive theory in a partnership with data. He called MDS "SSA" (Smallest Space Analysis, later reinterpreted as Similarity Structure Analysis), because he wanted to emphasize that the Cartesian dimensions of an MDS representation are but an algebraic scaffolding for solving a geometric problem. Hence, any geometric patterns (such as dimensions, directions, clusters, figures, and, in particular, regions and neighborhoods) that correspond to substantive knowledge about the objects can be meaningful. This perspective later developed into facet theory and led to other data-analytic methods such as partial order scalogram analysis.

Ordinal MDS stimulated a huge number of applications, but, over the years, interval and even ratio MDS recovered considerably in terms of utilization. This had statistical reasons on the one hand (interval MDS solutions, for example, are often less cluttered, with fewer tight point clusters), and theoretical reasons on the other hand (the emphasis of measurement foundations had shifted towards cumulative theory construction over replications, away from from an almost endless testing of single data sets, and metric MDS solutions often allow for simpler and more robust interpretations than "over-fitted" ordinal solutions). Today, ordinal MDS is but one of several MDS models. Advanced computer programs generate solutions for each of them in seconds, and so they can easily be tested against each other at virtually no costs.

2.4 Ordinal MDS as a method to study the shape of generalization gradients in learning

One historical motivation for MDS is closely linked to a special issue in the psychology of learning. Its main focus is not the MDS space itself, but the shape of the regression function of MDS distances to the data they represent. The theory that generates this interest is the following. If a response R is conditioned to a stimulus S, then stimuli similar to S also tend to trigger response R with a certain probability. This probability should be a monotonically decreasing function of the distance of S' to S. Yet, it is difficult to tell the shape the generalization gradient, because S' and S lie in perceptual not physical space. However, given a set of stimuli and measurements of the probabilities of giving the S_i response to stimulus S_k, the issue turns into an MDS problem: If the data are taken as similarity measures, one can first scale them via ordinal MDS; then, the

regression trend of the MDS distances onto the data shows the shape of the generalization gradient. So, rather than postulating that the gradient is an exponential or a linear decay function, Shepard wanted to let the data speak for themselves, finding the generalization gradient empirically through what is now known as the "Shepard diagram". Shepard struggled with the problem without really solving it, but his work motivated Kruskal to develop an ordinal MDS algorithm (M-D-Scal) as a statistician's answer to the scaling task.

Today, applications of MDS where the Shepard diagram is of primary interest are exceedingly rare. Rather, researchers almost always focus on the MDS space itself and its relationship to known or assumed properties of the objects represented in this space. Ordinal MDS can sometimes be useful to check certain model assumptions empirically. For example, in Thurstonian Case-5 scaling, dominance probabilities are mapped into scale differences by a cumulative Gaussian function. Rather than assuming such a mapping function, one can use ordinal MDS to scale the

data into scale distances, and then check empirically if the mapping function is indeed S-shaped.

2.5 MDS as a general data-analytic tool

As soon as ordinal MDS became possible, it was enthusiastically received by many disciplines outside of psychology, in particular by market researchers. Green and his coworkers, in particular, published scores of papers and books that showed how MDS can be used to "uncover" how consumers perceive products. Sociologists also used MDS to study social networks and, in particular, attitudes and values. Schwartz, for example, used MDS to develop his Theory of Universals in Values (TUV), an influential theory on social values that is well and alive today. The TUV is intimately related to a circumplex of regions (a wheel of regional sectors) in 2-dimensional MDS space. It partitions the MDS space into neighborhoods that each contain only points representing values of the same category (e.g., achievement values, security values, enjoyment values). While Green had used MDS in a purely exploratory way, Schwartz (as Guttman) was content-driven and hence looked for correspondences of content theory about the MDS objects and their representation in space.

Schwartz, however, never enforced such external constraints onto the MDS solutions using confirmatory MDS (CMDS), although CMDS had been around since the early 80's. DeLeeuw and Heiser, among others, developed certain forms of confirmatory MDS, and programs like Proxscal (in SPSS) or Smacof (in R) are able to handle most of them. However, many forms of external constraints onto the MDS configuration (except those on dimensions) are not easy (or simply impossible) to set up in the present MDS programs. Nor is it often clear how to assess the effects of those constraints statistically. Sometimes the present programs also yield incorrect solutions, which can be difficult to diagnose for the the applied researcher. From the substantive researcher's point-of-view, this is where more work is needed but recently work on CMDS has received more attention.

Another line of research deals with one general argument against distance models,

i.e. that distances are always symmetric but (dis)similarity data may not be symmetric. Various proposals were made on how to handle non-symmetric data. Most amount to first splitting the data information into a symmetric part (which can be modeled via MDS) and a skew-symmetric part (which can be added to the points in MDS space in form of small arrows leading to a vector field, for example). These models are not psychological models in the sense that they explain how a person generates asymmetric (dis)similarities. Rather, they are statistical tools that can be useful showing systematic trends in asymmetric (dis)similarity data. However, no user-friendly programs exist so far for these models, and hence their potential for statistical diagnostics has not been exploited.

2.6 Utilization of MDS today

Today, many of the original motives that led to the development of MDS, have become unimportant. What has survived, in particular, is using MDS in Guttman's sense, in particular in attitude and value research, where intercorrelations of survey items are studied for correspondences of the conceptual facets of the items to regions of their spatial representation. Yet, most applications of MDS today actually serve a much wider purpose, i.e. they are done to visualize tables of indices that can be interpreted as (dis)similarity data. For that purpose, MDS is highly useful as it can handle a vast variety of data as long as they are (dis)similarities (e.g., correlations, covariances, co-occurrence data, profile distances); it does not require interval-scaled data but also handles ordinal (and even nominal) data; it is robust against missing data and coarse data; it often serves as a data smoother, showing a structure that is replicable even under conditions of high error; it is easily explained to non-experts and allows them to explore the solutions without much risk (given that the Euclidean metric is employed!); it is easy to run for non-experts even though its solution algorithms are rather difficult (but: driving a car also does not require knowing how the engine works); and it does not impose a particular interpretation ("dimensions", in particular) onto the user but allows the data to speak for themselves.

3 Technical aspects of MDS: The past

3.1 Classical MDS

Classical MDS can be considered the first algebraic approach to MDS. It has been independently proposed by several authors: Torgerson (1958), Gower (1966), and Kloek and Theil (1965). Classical MDS rests on the following equation: Let $\mathbf{X}$ be the $n \times p$ matrix of point coordinates (assumed here to be column-centered for simplicity); then, the matrix of squared Euclidean distances with elements $d_{ij}^2(\mathbf{X}) = \sum_{s=1}^{p}(x_{is} - x_{js})^2$ is

$$\mathbf{D}^{(2)} = \mathbf{1}\boldsymbol{\alpha}' + \boldsymbol{\alpha}\mathbf{1}' - 2\mathbf{X}\mathbf{X}', \tag{1}$$

where $\mathbf{1}$ is a vector of ones of appropriate length and $\boldsymbol{\alpha}$ the vector with diagonal elements of $\mathbf{X}\mathbf{X}'$. Given $\mathbf{D}$, $\mathbf{X}$ is found as follows. Let $\mathbf{J} = \mathbf{I} - \mathbf{1}\mathbf{1}'/\mathbf{1}'\mathbf{1}$ be the centering matrix

with $\mathbf{I}$ the identity matrix. Then, multiplying the left- and right-hand side of (1) with $\mathbf{J}$ makes the terms with $\boldsymbol{\alpha}$ disappear as $\mathbf{J1} = \mathbf{0}$. An additional multiplication by $-1/2$ yields

$$-1/2\mathbf{J}\mathbf{D}^{(2)}\mathbf{J} = \mathbf{X}\mathbf{X}'. \tag{2}$$

Then, the eigendecomposition of $-1/2\mathbf{J}\mathbf{D}^{(2)}\mathbf{J}$ is $\mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}'$, and so $\mathbf{X} = \mathbf{Q}\boldsymbol{\Lambda}^{1/2}$. Classical MDS rests on the idea that if the matrix of dissimilarities $\boldsymbol{\Delta}$ is not a Euclidean distance matrix (which is almost always true with real data), it can be approximated by inserting $\boldsymbol{\Delta}^{(2)}$ for $\mathbf{D}^{(2)}$ in (2), and then retaining the first p positive eigenvalues of the eigendecomposition.

It can be proved that classical MDS minimizes the Strain loss function

$$\begin{aligned} \text{Strain}(\mathbf{X}) &= 1/4\text{tr } \mathbf{J}(\boldsymbol{\Delta}^{(2)} - \mathbf{D}^{(2)})\mathbf{J}(\boldsymbol{\Delta}^{(2)} - \mathbf{D}^{(2)})\mathbf{J} \\ &= \|(-1/2\mathbf{J}\boldsymbol{\Delta}^{(2)}\mathbf{J}) - \mathbf{X}\mathbf{X}'\|^2. \end{aligned}$$

Gower (1966) was the first to realize that the dimension reduction of principal component analysis (often seen as the eigendecomposition of a correlation matrix or the singular value decomposition of the data matrix $\mathbf{Z}$ itself) has a dual method that can be obtained by doing classical MDS on the Euclidean distances of the rows of the data matrix $\mathbf{Z}$. This method was coined principal coordinate analysis that emphasizes the representation of the rows (usually individuals or samples) of the data matrix $\mathbf{Z}$. For more on classical MDS, we refer to Chapter 12 of Borg and Groenen (2005).

3.2 Stress

Arguably the two most important breakthroughs in MDS were (1) modeling dissimilarities directly by distances in a loss function and (2) allowing very free transformations of the dissimilarities that in turn are estimated by distances. Shepard (1962a, 1962b) proposed heuristic methods to do both aspects but he did not provide a loss function. Kruskal (1964a, 1964b), then, suggested the loss function

$$\text{Stress}(\mathbf{X}, \hat{\mathbf{d}}) = \frac{\sum_{i<j}\left(\hat{d}_{ij} - d_{ij}(\mathbf{X})\right)^2}{\sum_{i<j} d_{ij}^2(\mathbf{X})}, \tag{3}$$

where $\hat{d}_{ij}$ is a transformed dissimilarity. For the moment assume that $\hat{d}_{ij} = \delta_{ij}$. Then, this Stress loss function fits the distance $d_{ij}(\mathbf{X})$ directly to the dissimilarities δ_{ij} and minimizes simply the squared errors over all combinations i, j. The minimization of (3) over $\mathbf{X}$ is not trivial as no analytical solution exists. Kruskal proposed a gradient-based minimization method to get the coordinates. The second breakthrough is to allow for transformations of the dissimilarities. One such transformation is the linear transformation $\hat{d}_{ij} = a + b\delta_{ij}$ for unknown a and b. With a large positive intercept a and a negative slope b the dissimilarities may be replaced by similarity measures, thereby opening up a large variety of applications that are based on similarity measurements (e.g.,

Table 3: Schematic overview of a dissimilarity matrix $\boldsymbol{\Delta}$ and a facet design.

Dissimilarity matrix $\boldsymbol{\Delta}$							Facet design			
			$\boldsymbol{\Delta}$					Facet		
	O_1	O_2	O_3	$\dots$	O_{n-1}	O_n		1	2	3
O_1	0						O_1	1	1	3
O_2	δ_{12}	0					O_2	1	2	3
O_3	δ_{13}	δ_{23}	0				O_3	2	1	3
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$			$\vdots$	$\vdots$	$\vdots$	$\vdots$
O_{n-1}	$\delta_{1,n-1}$	$\delta_{2,n-1}$	$\delta_{3,n-1}$	$\dots$	0		O_{n-1}	3	1	1
O_n	δ_{1n}	δ_{2n}	δ_{3n}	$\dots$	$\delta_{n-1,n}$	0	O_n	3	2	1

correlations) among the objects. Kruskal proposed an even more flexible transformation, that is, the ordinal transformation. This implies that the $\hat{d}_{ij}$s should be chosen such that whenever $\delta_{ij} \leq \delta_{kl}$ it must also hold that $\hat{d}_{ij} \leq \hat{d}_{kl}$ for any combination of pairs ij and kl. For fixed $\mathbf{X}$, the minimization of (3) over $\hat{\mathbf{d}}$ amounts to a quadratic program with linear inequality constraints on $\hat{\mathbf{d}}$. Kruskal provided a solution called monotone regression that provides a global minimum to this optimization problem. These two contributions can be seen as crucial milestones in the development of MDS as a statistical technique. When optimizing both over $\mathbf{X}$ and $\hat{\mathbf{d}}$, some adaptation is needed to avoid the trivial solution $\mathbf{X} = \mathbf{0}$ and $\hat{\mathbf{d}} = \mathbf{0}$. In (3), this trivial solution is avoided by dividing by the sum-of-squares of the $d_{ij}(\mathbf{X})$s.

3.3 Facet Theory and Regional Interpretation in MDS

In facet theory, "content" information is available for the objects in the form of external coding variables. These variables are called facets. The objects of observation are assigned to a certain level on each facet, as illustrated in Table 3. Guttman (1964) proposed to use such facets to form regions in MDS space. That is, it is hypothesized that if the facets are scientifically useful at all, then the points should fall into certain (non-overlapping and exhaustive) neighborhoods that correspond to the levels of a particular facet, facet by facet. Ordered facets should lead to correspondingly ordered regions, and this order can be linear ("stripes" in space) but also circular ("wedges"), with the usual "dimensions" as special cases of linearly ordered stripes. Three types of regional patterns are often observed with empirical data: axial, modular, and polar regions (see Figure 3).

In empirical research, regions are almost always found by hand (drawing and re-drawing partitioning lines on print-outs of MDS plots until the partitioning seems optimal), but Borg and Groenen (1997) were the first to minimize Stress while imposing axial constraints when the number of axial facets equals the number of dimensions. Groenen and Van der Lans (2004) extend this to the case where the number of axial facets exceeds the number of dimensions.

Let us return to the Morse signals example. Content information is available on the

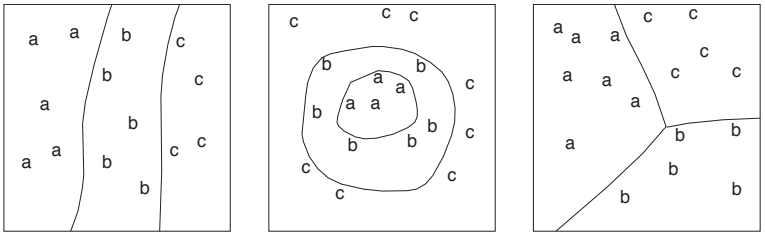

Figure 3: Three possible ways for regional partitioning by a facet: Panel a shows an axial partitioning by parallel lines, Panel b a modular partitioning by concentric circles, and Panel c an polar partitioning by rays emanating from a common origin.

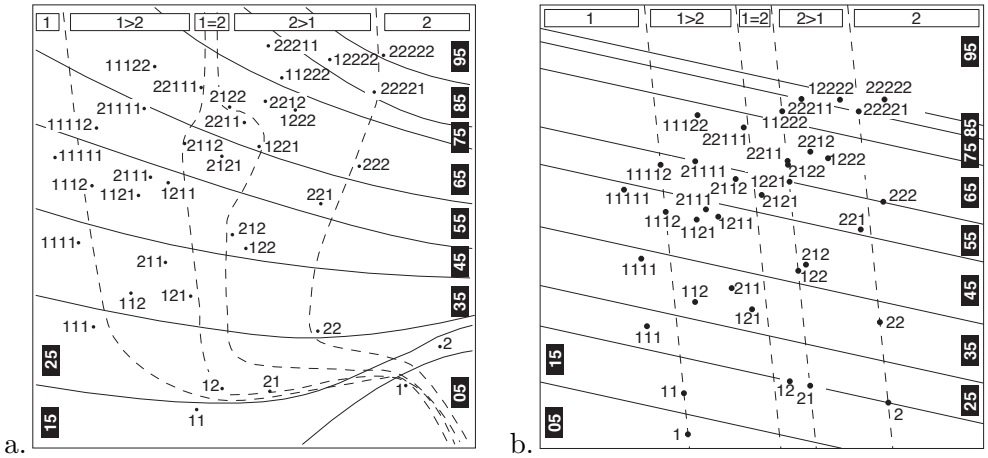

Figure 4: MDS solution of ordinal MDS on the Rothkopf confusion data with regions drawn by hand (Panel a) and regionally constrained MDS (Panel b).

Morse signals: Each signal has a temporal length (from .05 to .95 seconds) and a certain composition of long and short beeps. We code the latter as 'only short beeps' (1), 'more short beeps than long beeps' ($1 > 2$), 'an equal number of short and long beeps' ($1 = 2$), 'more long than short beeps' ($2 > 1$), and 'only long beeps' (2). These two external variables are facets of the signals. An approximate axial partitioning of the unconstrained solution is shown in Figure 4a (and Figure 1 without the axial partitioning) and the regionally constrained version in Figure 4b. The axially constrained ("confirmatory") solution has a slightly higher Stress (.21) compared to the unconstrained solution (.18), yet it gives a plot that is much easier to interpret in psychophysical terms. Moreover, the linearized structure is related to substantive laws of formation and, therefore, it can be expected to be more robust over replications than the possibly over-fitted exploratory MDS pattern with its partitioning lines that were inserted only afterwards.

In a recent application, Borg, Groenen, Jehn, Bilsky, and Schwartz (2011) impose two regional axial constraints with only two levels per facet (which effectively imposes a

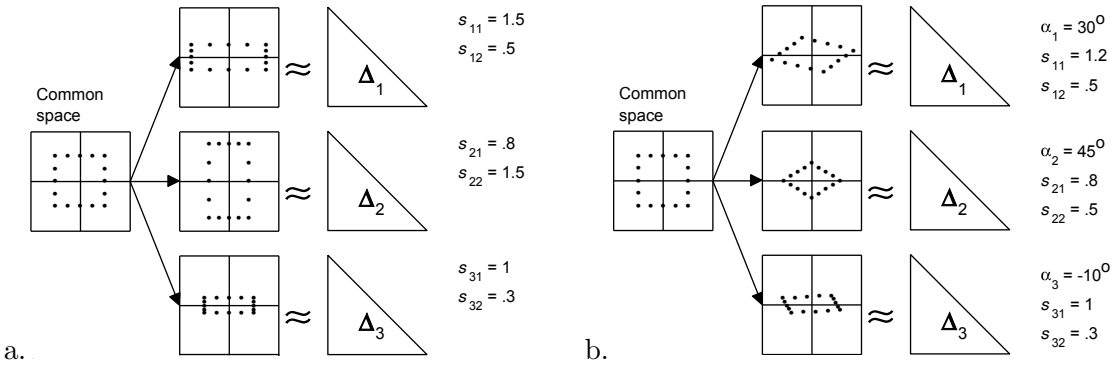

Figure 5: Illustration of the effect of dimension weighting for a common space $\mathbf{G}$ that forms a square of points and three examples for individual spaces $\mathbf{GS}_k$. Panel a shows the weighted Euclidean model with S_k diagonal and Panel b the generalized Euclidean model.

quadrant restriction), and then perform a permutation test on Stress to test if the axial constraints perform better than random assignment of points to quadrants. For more information on Facet Theory, we refer to the book of Borg and Shye (1995) and for some other applications, see, for example, Borg and Groenen (1997, 1998)

3.4 Three-way MDS models

In many applications of MDS, there is not just one data matrix but K dissimilarity matrices, $\mathbf{\Delta}_k$, for $k = 1, \ldots, K$. To model such data in one MDS representation, Horan (1969) and Carroll and Chang (1970) proposed an important extension of the basic MDS model. Their "weighted Euclidean model" assumes that each $\mathbf{\Delta}_k$ can be explained by distances of a single common space $\mathbf{G}$ transformed by the diagonal matrix $\mathbf{S}_k$ for each k. This means that each individual k simply stretches or compresses the common space along its dimensions as if each k attributes his or her own specific salience to each dimension. Figure 5a shows an artificial example of three individual spaces $\mathbf{X}_k = \mathbf{GS}_k$ that are derived from a single common space.

Carroll and Chang (1970) did not use Stress but Strain in their INDSCAL program. In this formulation, negative weights are a problem as they could lead to negative "distances". The constrained MDS approach of De Leeuw and Heiser (1980) that uses Stress avoids this problem and the sign of the dimension weight is formally unimportant. To eliminate a basic indeterminacy in the $\mathbf{GS}_k$ model, $\mathbf{G}$ is normalized so that $\mathbf{G}'\mathbf{G} = \mathbf{I}$. As this restricts the sum-of-squares of all columns in $\mathbf{G}$ to one, it becomes possible to compare the dimension weighting values among persons over dimensions, but only conditional to how $\mathbf{G}$ is normed (i.e., other norms lead to other dimension weights).

Three extensions exist of this approach. The first one also allows for a rotation before stretching (see Figure 5b) which means that $\mathbf{S}_k$ is allowed to be any square matrix. Using the Strain loss function, Carroll and Chang (1970) called this method IDIOSCAL. The second extension comes from the constrained MDS approach of De Leeuw and Heiser

(1980) using Stress that imposes a model that allows $\mathbf{S}_k$ to be of lower rank than the dimensionality of the common space $\mathbf{G}$. For example, one could model the common space to be 4-dimensional and the individual spaces as (rotated and stretched/compressed) subspaces of the 4-dimensional common space. A third possibility is to also pre-multiply the common-space coordinate matrix $\mathbf{G}$ by a weight matrix (Lingoes & Borg, 1978), but this leads to a model that has few applications.

Note that these three-way models tend to be very restrictive in case the dimensionality is low. The alternative of doing K separate MDS analyses allows the $\mathbf{X}_k$ to be estimated freely. On the other hand, the generalized Euclidean model or reduced rank model with very high dimensionality of the common space (close to n) will yield solutions close to K separate MDS analyses because the number of parameters in $\mathbf{S}_k$ becomes large. We believe the generalized Euclidean model or reduced rank model are most useful for common spaces whose dimensionality is not too small (say, larger than 3 and smaller than 6), and, in case of the reduced rank model, the rank of the individual spaces (thus the rank of $\mathbf{S}_k$) may be small (say, 2).

3.5 The Majorization Algorithm

A key contribution to MDS was made by De Leeuw (1977) when he first used the idea of majorization, albeit in the context of convex analysis in this paper. Up to then, the minimization of Stress was essentially done through gradient algorithms, such as the one proposed by Kruskal. The problem is that if only a single pair ij has a zero distance, then the gradient is not defined any more. De Leeuw (1977) proved that by using subgradients for those Euclidean distances that are zero, a convergent algorithm can be obtained. In De Leeuw and Heiser (1977), the idea of majorization was worked out further. The algorithm uses in each iteration an auxiliary function (called the majorizing function) that is simple (quadratic in $\mathbf{X}$); touches the original Stress function at the current estimate; and is located above the original Stress function anywhere else (or has the same value as the Stress function). Consequently, the update of the majorizing function must have a smaller (or equal) value as the majorizing function and as the Stress function at the current estimate (as these two functions touch there). Because the Stress function either touches or is smaller than the majorization function by construction, it must be so that at the update of the majorizing function, the Stress function also is smaller than (or equal to) the Stress function of the current estimate. Hence, making the $\mathbf{X}$ that minimizes the majorization function to be the next current estimate reduces Stress (or keeps it the same). In practice, the majorizing algorithm is fast and reduces Stress until the reductions in Stress become very small. This algorithm for minimizing Stress was coined SMACOF (Scaling by MAjorizing a COmplicated Function).

The SMACOF approach operates on a slightly different formulation of raw Stress

$$\sigma_r^2(\mathbf{X}, \hat{\mathbf{d}}) = \sum_{i<j} w_{ij} \left(\hat{d}_{ij} - d_{ij}(\mathbf{X}) \right)^2, \tag{4}$$

where w_{ij} are nonnegative weights indicating the importance of misrepresentation of a particular pair of objects ij. An obvious choice for $w_{ij} = 1$ for all ij so that all pairs

contribute equally. The w_{ij} can also be used for accommodating missing dissimilarities. In the SMACOF approach, degeneration to $\mathbf{X} = \mathbf{0}$ and $\hat{\mathbf{d}} = \mathbf{0}$ is avoided by imposing the explicit restriction that the sum of squared d-hats must equal some positive constant, for example, $\sum_{i<j} w_{ij}\hat{d}_{ij}^2 = n(n-1)/2$.

The strength of the majorization approach lies in its generalizability and the desirable properties of the algorithm. For example, it allows imposing constraints onto the configuration quite easily (De Leeuw & Heiser, 1980). This can be seen as follows. We focus on Stress as a function of $\mathbf{X}$ and assume a ratio transformation so that $\hat{d}_{ij} = \delta_{ij}$. Then, (4) can be expressed as

$$\sigma_r^2(\mathbf{X}) \quad = \quad \sum_{i<j} w_{ij}\delta_{ij}^2 + \sum_{i<j} w_{ij}d_{ij}^2(\mathbf{X}) - 2\sum_{i<j} w_{ij}\delta_{ij}^2 d_{ij}^2(\mathbf{X}) = \eta_\delta^2 + \eta^2(\mathbf{X}) - 2\rho(\mathbf{X}). \tag{5}$$

The core inequality of SMACOF is based on the Cauchy-Schwartz inequality yielding $-\sum_{i<j} w_{ij}\delta_{ij}d_{ij}(\mathbf{X}) \leq \sum_{i<j} w_{ij}\delta_{ij}/d_{ij}(\mathbf{Y})$, where $\mathbf{Y}$ is the estimate of $\mathbf{X}$ from the previous iteration (assuming $d_{ij}(\mathbf{Y}) > 0$). Then, $-\rho(\mathbf{X}) \leq \text{tr } \mathbf{X}'\mathbf{B}(\mathbf{Y})\mathbf{Y}$ with $\mathbf{B}(\mathbf{Y})$ a matrix function defined in, for example, De Leeuw (1988). There also exists the nice matrix expression for $\eta^2(\mathbf{X}) = \text{tr } \mathbf{X}'\mathbf{V}\mathbf{X}$ where the offdiagonal elements of $\mathbf{V}$ are equal to $-w_{ij}$ and the diagonal elements contain the row sums of matrix $\mathbf{W}$. Then,

$$\sigma_r^2(\mathbf{X}) \quad \leq \quad \eta_\delta^2 + \text{tr } \mathbf{X}'\mathbf{V}\mathbf{X} - 2\text{tr } \mathbf{X}'\mathbf{B}(\mathbf{Y})\mathbf{Y} \tag{6}$$

showing that the majorizing function at the right side of (6) has a constant, a quadratic, and a linear term in $\mathbf{X}$. Let $\overline{\mathbf{X}} = \mathbf{V}^-\mathbf{B}(\mathbf{Y})\mathbf{Y}$ be the (unconstrained) update with $\mathbf{V}^-$ the Moore-Penrose inverse of $\mathbf{V}$. Then, (6) can be expressed as

$$\begin{aligned} \sigma_r^2(\mathbf{X}) \quad &\leq \quad \eta_\delta^2 + \text{tr } \mathbf{X}'\mathbf{V}\mathbf{X} - 2\text{tr } \mathbf{X}'\mathbf{V}\overline{\mathbf{X}} \\ &= \quad \eta_\delta^2 + \text{tr } (\mathbf{X} - \overline{\mathbf{X}})'\mathbf{V}(\mathbf{X} - \overline{\mathbf{X}}) - \text{tr } \overline{\mathbf{X}}\mathbf{V}'\overline{\mathbf{X}}. \end{aligned} \tag{7}$$

This shows that the SMACOF algorithm can handle any constraint on $\mathbf{X}$ for which the function tr $(\mathbf{X} - \overline{\mathbf{X}})'\mathbf{V}(\mathbf{X} - \overline{\mathbf{X}})$ subject to the constraints can be minimized easily. For example, consider a given $n \times r$ matrix $\mathbf{H}$ with r additional attributes (external variables) on the objects. This information can be used easily by constraining the coordinates $\mathbf{X}$ to be a linear combination of the known external variables $\mathbf{H}$ thereby allowing the MDS dimensions to be interpreted in terms of the external variables. Then, in each iteration the minimum of tr $(\mathbf{H}\mathbf{C} - \overline{\mathbf{X}})'\mathbf{V}(\mathbf{H}\mathbf{C} - \overline{\mathbf{X}})$ over $\mathbf{C}$ needs to be found and that is obtained by $\mathbf{C}^+ = (\mathbf{H}'\mathbf{V}\mathbf{H})^{-1}\mathbf{H}'\mathbf{V}\overline{\mathbf{X}}$.

Three-way models can also be seen as a form of constrained MDS that can be handled by the majorizing approach. For example, three-way MDS models (such as the weighted Euclidean model, the generalized Euclidean model, and the reduced rank model) are expressed in the Stress framework as

$$\sigma_r^2(\mathbf{X}) \quad = \quad \sum_k \sum_{i<j} w_{ij} \left(\delta_{ijk} - d_{ij}(\mathbf{G}\mathbf{S}_k)\right)^2. \tag{8}$$

Let $\mathbf{\Delta}^*, \mathbf{W}^*$, and $\mathbf{X}^*$ be defined as

$$\mathbf{\Delta}^* = \begin{bmatrix} \mathbf{\Delta}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Delta}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{\Delta}_3 \end{bmatrix}, \mathbf{W}^* = \begin{bmatrix} \mathbf{W}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{W}_3 \end{bmatrix}, \text{ and } \mathbf{X}^* = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \mathbf{X}_3 \end{bmatrix}.$$

Now, three-way MDS can be viewed as doing a constrained MDS on $\mathbf{\Delta}^*, \mathbf{W}^*$, and $\mathbf{X}^*$ where $\mathbf{X}_k$ is constrained to be of the form $\mathbf{GS}_k$.

Other algorithmic properties on convergence of the algorithm were proved in De Leeuw (1988). Note that such convergence properties are not available for other MDS algorithms. The SMACOF algorithm is available in the SPSS module PROXSCAL (Meulman, Heiser, & SPSS, 1999) and as the SMACOF package in R (De Leeuw & Mair, 2009).

Although Euclidean distances are the easiest to visually interpret and therefore are predominantly used in MDS, there can be reasons to deviate from the Euclidean distance and use the more general Minkowski distance $d_{ij}(\mathbf{X}) = (\sum_s |x_{is} - x_{js}|^q)^{1/q}$ for $q \geq 1$. The well known special cases are the city-block ($q = 1$), the Euclidean ($q = 2$), and the dominance ($q = \infty$) distances. The majorization approach to MDS was extended to deal with these cases in Groenen, Mathar, and Heiser (1995) for $1 \leq q \leq 2$ and also for $q \geq 2$ in Groenen, Heiser, and Meulman (1999).

3.6 Other Algorithms

The property of undefined gradients for Stress led Takane, Young, and De Leeuw (1977) to propose the S-Stress loss function that minimizes the sum over all pairs of the squared differences of squared Euclidean distances and squared dissimilarities as its gradient is defined for all distances, even if they are zero. The disadvantage of S-Stress is that it will tend to over-represent large dissimilarities and that it can allow for relatively large errors for small dissimilarities.

Ramsay (1977) proposed the MULTISCALE loss function that equals the sum of squared differences of the logarithms of dissimilarities and distances, or, equivalently, the sum of squared logarithms of the ratio of a dissimilarity and its distance. If the ratio is one, the log is zero and there is a perfect representation of the dissimilarity by its distance. Note that zero dissimilarities or zero distances cannot be handled by MULTISCALE. The advantage of this method is that it can be seen within a maximum likelihood framework. In particular, its three-way extension with replications by individuals allows inference and confidence ellipses for the points.

4 Present

4.1 Distance-based Multivariate Analysis

In a series of publications in the period from 1986 to 1998, Meulman generalized the relation of principal coordinates analysis and classical MDS to a much wider range of multivariate analysis techniques such as (multiple) correspondence analysis, (generalized) canonical correlation analysis, and discriminant analysis. The emphasis in this

approach is on the representation of the objects, and much less on the variables. It allows for optimal scaling of the variables in the Gifi (1990) approach approximating distances through Stress or Strain. A comprehensive overview was outlined in her thesis (Meulman, 1986) and a series of papers were written on that topic (see, for example, Meulman, 1992).

4.2 Constant Dissimilarities

Even though the use of Stress is dominant in MDS, little was known on the "nullest of null models", that is, having constant dissimilarities without any variation (Buja, Logan, Reeds, & Shepp, 1994). Many classical multivariate analysis techniques assume centered data so that constant data do not occur. Buja et al. proved what kind of solutions occur when Stress with Euclidean distances is fed with constant dissimilarities. It turns out that in one dimension, the objects will be positioned equally spaced on a line, in two dimensions the points will be in concentric circles, and in three dimensions, they are positioned on the surface of a hypersphere. The 2-dimensional solution is used very often in MDS applications and near constant dissimilarity data can occur after transformations. The contribution Buja et al. is that they focused the attention on such noninformative solutions. Users of MDS should always check if their data have sufficient variation in the dissimilarities or d-hats before starting interpreting a solution.

4.3 Local Minima

The advantage of classical MDS is that there is an algebraic solution that yields a global minimum. Some of the disadvantages are that it cannot handle transformations of the dissimilarities and that the resulting distances often under-estimate the dissimilarities. The use of Stress avoids these disadvantages but introduces the problem of local minima. From 1978 until recently, several contributions have been made. Three cases should be distinguished.

First, De Leeuw and Heiser (1977) and Defays (1978) noted that unidimensional scaling becomes an NP-hard combinatorial problem. Hubert and Golledge (1981) and Hubert and Arabie (1986) used dynamic programming to globally optimize the combinatorial problem and hence the unidimensional Stress function up to about 22 objects. The approach by Pliner (1996) that smoothes small distances in $\rho(\mathbf{X})$ by a quadratic function is very effective in finding global minima for even larger unidimensional scaling problems. Brusco (2001) applied simulated annealing.

The second case is for $2 \leq p < n - 1$. For small $p \geq 2$, Groenen and Heiser (1996) found that local minima occur frequently, more so in low dimensionalities than for higher dimensionalities. They proposed the tunneling method that is indeed capable of finding a series of subsequent lower local minima that could end in a global minimum although there is no guarantee of finding it. In Groenen et al. (1999), the smoothing approach of Pliner (1996) was adapted and extended to deal with any Minkowski distance and in any dimensionality. For city-block distances and Euclidean distances, their smoothing approach was effective in locating a global optimum, but for Minkowski distances with

exponents $q > 2$, in particular the dominance distance ($q = \infty$), the smoothing approach was not effective. For the special case of city-block distances, combinatorial approaches have been proposed by Hubert, Arabie, and Hesson-McInnis (1992). More recently, there has been a series of articles on simulated annealing approaches of several MDS variants which work well when appropriately tuned (Murillo, Vera, & Heiser, 2005; Vera, Heiser, & Murillo, 2007).

The third case is full-dimensional scaling ($p = n - 1$). De Leeuw (1993) proved that there is a single local minimum for Stress that is global.

5 Future

5.1 MDS with Weights

Apart from handling missings, the weights w_{ij} in the Stress function 3 can be also be used to emphasize certain aspects of the dissimilarities. The first one to exploit this was Heiser (1988) who proposed to mimic certain other MDS loss functions by choosing appropriate weights. For example, the S-Stress loss function can be mimicked by $w_{ij} = \delta_{ij}^2$. Buja and Swayne (2002) emphasized choosing $w_{ij} = \delta_{ij}^q$ allows a more refined weighting of the errors depending on the size of δ_{ij}. If the objective is to have the large dissimilarities well-represented and it does not matter much to have larger errors for small dissimilarities, then q should be chosen large (for example, $q = 5$ or 10). Conversely, if the interest is in the proper representation of small dissimilarities and the larger ones are unimportant in the representation, then this can be assured by choosing q small (for example, $q = -5$ or -10). Making the weights w_{ij} in such a way dependent on the dissimilarities allows emphasizing the proper representation of certain selection of the dissimilarities that is dependent of their values. This can be particularly useful in the context of large scale MDS.

5.2 Dynamic MDS

For most of its life time, MDS has been a static method: dissimilarities are input to an MDS program producing a usually 2-dimensional solution that is shown as a map. The GGvis software of Buja and Swayne (2002) (see also Buja et al., 2008) that is a part of GGobi was the first comprehensive interactive software. The advantage is that in real time MDS options can be changed and its effects are immediately shown as the iterative process progresses. Such dynamics allow for a completely new, direct, and intuitive interaction with an MDS user on the interplay of the specific dissimilarities at hand and the possible MDS options.

5.3 Large Scale MDS

Traditional MDS tends to have a small number of objects (say, between 10-200). Both computations and interpretation completely changes when one is dealing with far more objects, e.g., 10.000-100.000 objects. With such large n the total number of pairs of

objects $n(n-1)/2$ increases quadratically, which generally is prohibitive for standard MDS algorithms.

The ISOMAP algorithm of Tenenbaum, De Silva, and Langford (2000) is based on classical MDS for large scale MDS problems. In particular, it focusses on nonlinear manifolds in higher dimensionalities. By k-nearest neighbours a network of connected pairs of objects is created by declaring the bigger dissimilarities as missing. This structure of nonmissing dissimilarities can be seen as a weighted graph with weights at the vertices being the dissimilarities. As classical MDS cannot handle missing dissimilarities, they are replaced by the shortest path on the graph. This forms a fully filled matrix of pseudo dissimilarities on which classical MDS is performed. In this way, ISOMAP is capable of recovering low dimensional manifolds that exists in higher dimensionality.

Another approach was taken by Trosset and Groenen (unpublished, see also the dissertation of Kagie, 2010). Trosset and Groenen proposed to allow for many missing values thereby creating a sparse dissimilarity matrix. In an adaptation of SMACOF, they provide an algorithm that can indeed handle large n provided there is sufficient sparseness. In joint work, the dissertation of Kagie expands on this approach. It was noticed that often large scale MDS solutions are dominated by the mode of the distribution of dissimilarities and that, therefore, solutions comparable to the constant dissimilarity case occur often, such as in two dimensions a circle filled with a blur of points. It was proposed to be solved by an a priori weighting of the dissimilarities to avoid a single mode becoming dominant and by appropriate transformations.

5.4 Symbolic MDS

The use of symbolic data and adapted multivariate analysis methods has been advocated in Bock and Diday (2000). Symbolic data can be seen as richer forms of data values. Here, we discuss two such forms, that is, (a) the case that for each pair ij not the dissimilarity is known but the interval of the dissimilarity, and (b) the case that a distribution (histogram) of the dissimilarity for each pair ij is know. Often, such symbolic data are obtained by aggregation or summary statistics over larger units such as geographic areas, countries, etc.

For interval dissimilarities, Denœux and Masson (2000) proposed to present the coordinates of object also as intervals yielding a rectangle to represent an object in two dimensions. Then, the smallest distance of rectangles i and j should match as closely as possible the lower boundary value $\delta_{ij}^{(L)}$ of the interval for dissimilarity ij and the largest distance of the rectangles should match as the upper boundary value $\delta_{ij}^{(U)}$ of the interval. A rectangle for object i can be specified by the coordinates of its center, that is, row i of $\mathbf{X}$, and by half of its width per dimension given by row i of $\mathbf{R}$ with $r_{is} \geq 0$ for all is. Now, the largest Euclidean distance between two rectangles i and j can be expressed as

$$d_{ij}^{(U)}(\mathbf{X}, \mathbf{R}) = \left(\sum_{s=1}^{p} [|x_{is} - x_{js}| + (r_{is} + r_{js})]^2 \right)^{1/2} \tag{9}$$

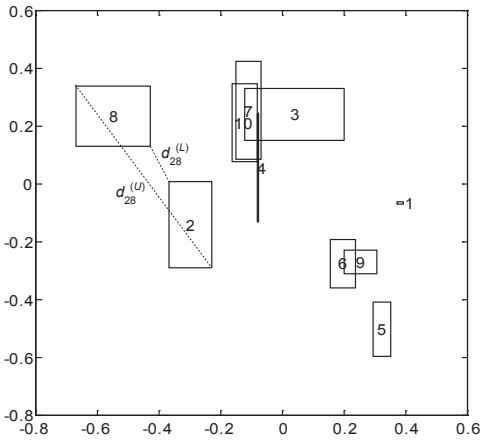

Figure 6: Example of an iMDS solution that approximates intervals of dissimilarities by the interval of the smallest and largest distance between two rectangles representing the the objects. For rectangles 2 and 8 the minimum and maximum Euclidean distances are shown.

and the smallest Euclidean is given by

$$d_{ij}^{(L)}(\mathbf{X},\mathbf{R}) = \left(\sum_{s=1}^{p} \max[0, |x_{is} - x_{js}| - (r_{is} + r_{js})]^2\right)^{1/2}. \tag{10}$$

The corresponding I-Stress function equals

$$\sigma_{\mathrm{I}}^2(\mathbf{X},\mathbf{R}) \;=\; \sum_{i<j}^{n} w_{ij}\left[\delta_{ij}^{(U)} - d_{ij}^{(U)}(\mathbf{X},\mathbf{R})\right]^2 + \sum_{i<j}^{n} w_{ij}\left[\delta_{ij}^{(L)} - d_{ij}^{(L)}(\mathbf{X},\mathbf{R})\right]^2.$$

Groenen, Winsberg, Rodriguez, and Diday (2006) provide the I-Scal algorithm for minimizing $\sigma_{\mathrm{I}}^2(\mathbf{X},\mathbf{R})$. Their algorithm is based on iterative majorization thereby guaranteeing a monotone decrease of I-Stress values until convergence.

Figure 6 gives an example of an I-Scal solution of rectangles to represent interval dissimilarities. The minimum Euclidean distance between rectangles 2 and 8, $d_{28}^{(L)}(\mathbf{X},\mathbf{R})$ and maximum Euclidean distance, $d_{28}^{(U)}(\mathbf{X},\mathbf{R})$ are explicitly shown. Note that rectangle 4 collapses into a line because r_{i1} is (almost) zero.

Groenen and Winsberg (2006) proposed to model histogram dissimilarities. In this case, the distribution of a dissimilarity is summarized by several quantiles, for example, by the percentiles 20, 30, 40, 60, 70, and 80. In this model, the percentiles should be chosen in pairs around the median, so 20-80, 30-70, and 40-60. Therefore, each such pair consists of an interval dissimilarity. The current choice of percentiles can be perceived as a three-way interval dissimilarity matrix with three replications (as there are three percentile pairs). Just as in regular three-way MDS, there will be one common matrix

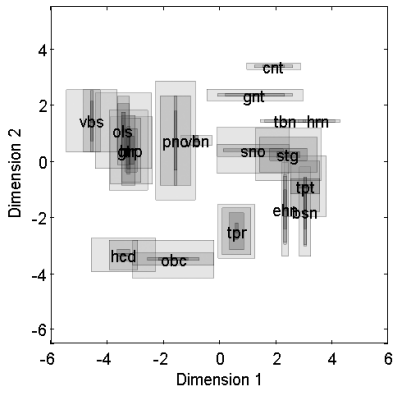

Figure 7: Example of a histogram MDS solution on synthetical musical instruments. The rectangles represent the percentile pairs 20-80, 30-70, and 40-60.

X with the rectangle centers for all three replications. The heights and widths of the rectangles over the three replications should be such that each rectangle representing a percentile range closer to the median should fit within a rectangle representing a wider percentile range around the median. These restrictions can be handled easily by an extension of the I-Scal algorithm with inequality restrictions on the r_{isk}. For more details, we refer to Groenen and Winsberg (2006). Figure 7 shows an example of solution of histogram dissimilarities of artificial musical instruments with the percentile pairs 20-80, 30-70, and 40-60.

5.5 What needs to be done?

Of the developments described above in the section "Future", what is most interesting for the typical MDS user is the possibility to interact more with MDS programs in a dynamic way (as in GGvis). Heady developed a powerful interactive MDS program called PERMAP, a stand-alone program that is available as free-ware in the Internet (`http://cda.psych.uiuc.edu/mds 509 2013/permap/permap 11 8pdf.pdf`). Unfortunately, this program is not supported anymore. For the user, it would be nice to see how MDS responds if he or she eliminates some objects/points from the solution; shifts some points in space; or draws in some partitioning lines that are then enforced (in some way such as linear axial constraints) onto the MDS solution. Such programs would be hard to write and test, of course, but often simpler programs are missing too. For example, programs that can handle asymmetric proximities and produce vector-fields over MDS plots would be helpful to diagnose asymmetric data for systematic trends. Even Procrustean transformations that are needed when comparing different MDS solutions for similarities and differences are missing (or are difficult to find) in many of the statistics packages. It is hoped that such programs will soon be generated within the R environment where they should also survive longer than it has been true for many of the old FORTRAN programs such as KYST or PINDIS, for example.

Another area where we expect more developments is large scale MDS. For this case, there are several technical and perceptual problems. When is large scale MDS interesting? How many dimensions should be used? How can points in such plots informatively be labeled? How should dissimilarities and weights be adapted such that MDS yields informative solutions. How to treat missing values that could lead to unconnected or only partially connected groups of objects? In this area, we expect that researchers from computer science and machine learning are and will contribute to new developments. One such development is the VOS approach for bibliometrics, see Van Eck and Waltman (2010) and for software and more references `http://www.vosviewer.com/`.

References

Bock, H., & Diday, E. (2000). *Analysis of symbolic data*. Berlin: Springer.

Borg, I., & Groenen, P. J. F. (1997). Multitrait-multimethod by multidimensional scaling. In F. Faulbaum & W. Bandilla (Eds.), *SoftStat '97* (pp. 59–66). Stuttgart: Lucius.

Borg, I., & Groenen, P. J. F. (1998). Regional interpretations in multidimensional scaling. In J. Blasius & M. Greenacre (Eds.), *Visualization of categorical data* (pp. 347–364). New York: Academic Press.

Borg, I., & Groenen, P. J. F. (2005). *Modern multidimensional scaling* (2. ed.). New York: Springer.

Borg, I., Groenen, P. J. F., Jehn, K. A., Bilsky, W., & Schwartz, S. H. (2011). Embedding the organizational culture profile into schwartz's theory of universals in values. *Journal of Personnel Psychology*, *10*, 1–12.

Borg, I., & Shye, S. (1995). *Facet theory: Form and content*. Newbury Park, CA: Sage.

Brusco, M. J. (2001). A simulation annealing heuristic for unidimensional and multidimensional (city-block) scaling of symmetric proximity matrices. *Journal of Classification*, *18*, 3–33.

Buja, A., Logan, B. F., Reeds, J. R., & Shepp, L. A. (1994). Inequalities and positive-definite functions arising from a problem in multidimensional scaling. *The Annals of Statistics*, *22*, 406–438.

Buja, A., & Swayne, D. F. (2002). Visualization methodology for multidimensional scaling. *Journal of Classification*, *19*, 7–44.

Buja, A., Swayne, D. F., Littman, M. L., Dean, N., Hofmann, H., & Chen, L. (2008). Data visualization with multidimensional scaling. *Journal of Computational and Graphical Statistics*, *17*, 444—472.

Carroll, J. D., & Chang, J. J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of 'Eckart-Young' decomposition. *Psychometrika*, *35*, 283–320.

De Leeuw, J., & Mair, P. (2009). Multidimensional scaling using majorization: SMACOF in R. *Journal of Statistical Software*, *31*(3), 1–30. Retrieved from `http://www.jstatsoft.org/v31/i03/`

Defays, D. (1978). A short note on a method of seriation. *British Journal of Mathematical and Statistical Psychology*, *31*, 49–53.

De Leeuw, J. (1977). Applications of convex analysis to multidimensional scaling. In J. R. Barra, F. Brodeau, G. Romier, & B. van Cutsem (Eds.), *Recent developments in statistics* (pp. 133–145). Amsterdam, The Netherlands: North-Holland.

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

De Leeuw, J. (1993). *Fitting distances by least squares* (Tech. Rep. No. 130). Los Angeles, CA: Interdivisional Program in Statistics, UCLA.

De Leeuw, J., & Heiser, W. J. (1977). Convergence of correction-matrix algorithms for multidimensional scaling. In J. C. Lingoes, E. E. Roskam, & I. Borg (Eds.), *Geometric Representations of Relational Data* (pp. 735–752). Ann Arbor, MI: Mathesis Press.

De Leeuw, J., & Heiser, W. J. (1980). Multidimensional scaling with restrictions on the configuration. In P. R. Krishnaiah (Ed.), *Multivariate Analysis* (Vol. V, pp. 501–522). Amsterdam, The Netherlands: North-Holland.

Denœux, T., & Masson, M. (2000). Multidimensinal scaling of interval-valued dissimilarity data. *Pattern Recognition Letters*, *21*, 83–92.

Gifi, A. (1990). *Nonlinear multivariate analysis.* Chichester, England: Wiley.

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, *53*, 325–338.

Groenen, P. J. F., & Heiser, W. J. (1996). The tunneling method for global optimization in multidimensional scaling. *Psychometrika*, *61*, 529–550.

Groenen, P. J. F., Heiser, W. J., & Meulman, J. J. (1999). Global optimization in least-squares multidimensional scaling by distance smoothing. *Journal of Classification*, *16*, 225–254.

Groenen, P. J. F., Mathar, R., & Heiser, W. J. (1995). The majorization approach to multidimensional scaling for Minkowski distances. *Journal of Classification*, *12*, 3–19.

Groenen, P. J. F., & Van der Lans, I. (2004). Multidimensional scaling with regional restrictions for facet theory: An application to Levy's political protest data. In M. Braun & P. P. Mohler (Eds.), *Beyond the horizon of measurement* (pp. 41–64). Mannheim: ZUMA.

Groenen, P. J. F., & Winsberg, S. (2006). Multidimensional scaling of histogram dissimilarities. In V. Batagelj, H. Bock, A. Ferligoj, & A. Ziberna (Eds.), *Data science and classification* (pp. 161–170). Berlin: Springer.

Groenen, P. J. F., Winsberg, S., Rodriguez, O., & Diday, E. (2006). I-scal: Multidimensional scaling of interval dissimilarities. *Computational Statistics and Data Analysis*, *51*, 360–378.

Guttman, L. (1964). The structure of interrelations among intelligence tests. In *Proceedings of the 1964 Invitational Conference on Testing Problems.* Princeton, N. J.: Educational Testing Service.

Heiser, W. J. (1988). Multidimensional scaling with least absolute residuals. In

H. H. Bock (Ed.), *Classification and related methods* (pp. 455–462). Amsterdam: North-Holland.

Horan, C. B. (1969). Multidimensional scaling: Combining observations when individuals have different perceptual structures. *Psychometrika*, *34*, 139–165.

Hubert, L. J., & Arabie, P. (1986). Unidimensional scaling and combinatorial optimization. In J. De Leeuw, W. J. Heiser, J. J. Meulman, & F. Critchley (Eds.), *Multidimensional data analysis* (pp. 181–196). Leiden, The Netherlands: DSWO-Press.

Hubert, L. J., Arabie, P., & Hesson-McInnis, M. (1992). Multidimensional scaling in the city-block metric: A combinatorial approach. *Journal of Classification*, *9*, 211–236.

Hubert, L. J., & Golledge, R. G. (1981). Matrix reorganisation and dynamic programming: Applications to paired comparison and unidimensional seriation. *Psychometrika*, *46*, 429–441.

Kloek, T., & Theil, H. (1965). International comparisons of prices and quantities consumed. *Econometrica*, *33*, 535–556.

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

Kruskal, J. B. (1964b). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, *29*, 115–129.

Lingoes, J. C., & Borg, I. (1978). A direct approach to individual differences scaling using increasingly complex transformations. *Psychometrika*, *43*, 491–519.

Meulman, J. J. (1986). *A distance approach to nonlinear multivariate analysis*. Leiden, The Netherlands: DSWO Press.

Meulman, J. J. (1992). The integration of multidimensional scaling and multivariate analysis with optimal transformations. *Psychometrika*, *57*, 539–565.

Meulman, J. J., Heiser, W. J., & SPSS. (1999). *SPSS Categories 10.0*. Chicago: SPSS.

Murillo, A., Vera, J. F., & Heiser, W. J. (2005). A permutation-translation simulated annealing algorithm for l_1 and l_2 unidimensional scaling. *Journal of Classification*, *22*, 119–138.

Pliner, V. (1996). Metric, unidimensional scaling and global optimization. *Journal of Classification*, *13*, 3–18.

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

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

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.

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

Tenenbaum, J. B., De Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimension reduction. *Science*, *290*, 2319–2323.

Torgerson, W. S. (1958). *Theory and methods of scaling.* New York: Wiley.

Van Eck, N. J., & Waltman, L. (2010). Software survey: Vosviewer, a computer program for bibliometric mapping. *Scientometrics*, *84*, 523-538.

Vera, J. F., Heiser, W. J., & Murillo, A. (2007). Global optimization in any minkowski metric: A permutation-translation simulated annealing algorithm for multidimensional scaling. *Journal of Classification*, *24*, 277–301.