Smacof at 50: A Manual
Part 1: Introduction

Jan de Leeuw

December 9, 2024

Table of contents

Note: This manual is a working manuscript which will be expanded/updated frequently. All suggestions for improvement are welcome. All Rmd, tex, html, pdf, R, and C files are in the public domain and can be copied, modified, and used by anybody in any way they see fit. Attribution will be appreciated, but is not required. The files can be found at https://github.com/deleeuw in the repositories smacofCode, smacofManual, and smacofExamples.

Conventions, Notations and Reserved Symbols

I number and label *all* displayed equations. Equations are displayed, instead of inlined, if and only if one of the following is true.

- They are important.
- They are referred to elsewhere in the text.
- Not displaying them messes up the line spacing.

All code chunks in the text are named. Theorems, lemmas, chapters, sections, subsections, and so on are also named and numbered. I use the serial comma.

The dilemma of whether to use "we" or "I" throughout the book is solved in the usual way. If I feel that a result is the work of a group (me, my co-workers, and the giants on whose shoulders we stand) then I use "we". If it's an individual decision, or something personal, then I use "I". The default is "we", as it always should be in scientific writing.

Most of the individual chapters also have some of the necessary mathematical background material, both notation and results, sometimes with specific elaborations that seem useful for the book. Sometimes this background material is quite extensive. Examples are splines, majorization, unweighting, monotone regression, and the basic Zangwill and Ostrowski fixed point theorems we need for convergence analysis of our algorithms.

Spaces

- $\mathbb{R}^n$ is the space of all real vectors, i.e. all n-element tuples of real numbers. Typical elements of $\mathbb{R}^n$ are x, y, z. The element of x in position i is x_i. Defining a vector by its elements is done with $x = \{x_i\}$.
- $\mathbb{R}^n$ is equipped with the inner product $\langle x, y\rangle = x'y = \sum_{i=1}^n x_iy_i$ and the norm $\|x\| = \sqrt{x'x}$.
- The canonical basis for $\mathbb{R}^n$ is the $n-$tuple $(e_1, cdots, e_n)$, where e_i has element i equal to $+1$ and all other elements equal to zero. Thus $\|e_i\| = 1$ and $\langle e_i, e_j\rangle = \delta^{ij}$, with δ^{ij} the Kronecker delta (equal to one if $i = j$ and zero otherwise). Note that $x_i = \langle e_i, x\rangle$.
- $\mathbb{R}$ is the real line and $\mathbb{R}_+$ is the half line of non-negative numbers. The postive reals are $\mathbb{R}_{++}$.
- $\mathbb{R}^{n\times m}$ is the space of all $n \times m$ real matrices. Typical elements of $\mathbb{R}^{n\times m}$ are A, B, C. The element of A in row i and column j is a_{ij}. Defining a matrix by its elements is done with $A = \{a_{ij}\}$.

- $\mathbb{R}^{n\times m}$ is equipped with the inner product $\langle A,B\rangle=\mathrm{tr}A'B=\sum_{i=1}^n\sum_{j=1}^m a_{ij}b_{ij}$ and the norm $\|A\|=\sqrt{\mathrm{tr}\ A'A}$.
- The canonical basis for $\mathbb{R}^{n\times m}$ is the $nm-$tuple $(E_{11},cdots,E_{nm})$, where E_{ij} has element (i,j) equal to $+1$ and all other elements equal to zero. Thus $\|E_{ij}\|=1$ and $\langle E_{ij},E_{kl}\rangle=\delta^{ik}\delta^{jl}$.

vec and vec^{-1}

Matrices

- $a_{i\bullet}$ is row i of matrix A, $a_{\bullet j}$ is column j.
- $a_{i\star}$ is the sum of row i of matrix A, $a_{\star j}$ is the sum of column j.
- A' is the transpose of A, and diag(A) is the diagonal matrix with the diagonal elements of A. The inverse of a square matrix A is A^{-1}, the Moore-Penrose generalized inverse of any matrix A is A^+. The transpose of the inverse, and the inverse of the transpose, are A^{-T}.
- If A and B are two $n\times m$ matrices then their Hadamard (or elementwise) product $C=A\times B$ has elements $c_{ij}=a_{ij}b_{ij}$. The Hadamard quotient is $C=A/B$, with elements $c_{ij}=a_{ij}/b_{ij}$. The Hadamard power is $A^{(k)}=A^{(p-1)}\times A$.
- DC matrices. Centering matrix. $J_n=I_n-n^{-1}E_n$. We do not use the subscripts if the order is obvious from the context.
- Matrices of matrices. Partitioned matrices. A_{ij} and thus $\{A_{ij}\}_{kl}$.
- Direct sum and Product

Functions

- $f,g,h,\cdots$ are used for functions or mappings. $f:X\rightarrow Y$ says that f maps X into Y.
- σ is used for all real-valued least squares loss functions.

MDS

- $\Delta = \{\delta_{ij\cdots}\}$ is a matrix or array of dissimilarities.
- $\langle\mathbb{X}, d\rangle$ is a metric space, with $d : \mathcal{X} \otimes \mathcal{X} \rightarrow \mathbb{R}_+$ the distance function. If X is is an ordered n-tuple $(x_1, \cdots, x_n)$ of elements of $\mathcal{X}$ then $D(X)$ is $\{d(x_i, x_j)\}$, the elements of which we also write as $d_{ij}(X)$.
- Summation over the elements of vector $x \in \mathbb{R}^n$ is $\sum_{i=1}^n x_i$. Summation over the elements of matrix $A \in \mathbb{R}^{n\times m}$ is $\sum_{i=1}^n \sum_{j=1}^m a_{ij}$. Summation over the elements above the diagonal of A is $\sum\sum_{1\leq i<j\leq n} a_{ij}$.
- Conditional summation is, for example, $\sum_{i=1}^n \{x_i \mid x_i > 0\}$.

Preface

This manual is definitely *not* an impartial and balanced review of all of multidimensional scaling (MDS) theory and history. It emphasizes computation, and the mathematics needed for computation. In addition, it is a summary of over 50 years of MDS work by me, either solo or together with my many excellent current or former co-workers and co-authors. It is heavily biased in favor of the smacof formulation of MDS (De Leeuw (1977), De Leeuw and Heiser (1977), De Leeuw and Mair (2009), Mair, Groenen, and De Leeuw (2022)), and the corresponding majorization (or MM) algorithms. And, moreover, I am shamelessly squeezing in as many references to my published and unpublished work as possible, with links to the corresponding pdf's if they are available. Thus this book is also a jumpstation into my bibliography.

I have not organized the book along historical lines because most of the early techniques and results have been either drastically improved or completely abandoned. Nevertheless, some personal historical perspective may be useful. I will put most of it in this preface, so uninterested readers can easily skip it.

I got involved in MDS in 1968 when John van de Geer returned from a visit to Clyde Coombs in Michigan and started the Department of Data Theory in the Division of Social Sciences at Leiden University. I was John's first hire, although I was still a graduate student at the time.

Remember that Clyde Coombs was running the Michigan Mathematical Psychology Program, and he had just published his remarkable book "A Theory of Data" (Coombs (1964)). The name of the new department in Leiden was taken from the title of that book, and Coombs was one of the first visitors to give a guest lecture there.

This is maybe the place to clear up some possible misunderstandings about the name "Data Theory". Coombs was mainly interested in a taxonomy of data types, and in pointing out that "data" were not limited to a table or data-frame of objects by variables. In addition, there were also similarity ratings, paired comparisons, and unfolding data. Coombs also emphasized that data were often non-metric, i.e. ordinal or categorical, and that it was possible to analyze these ordinal or categorical relationships directly, without first constructing numerical scales to which classical techniques could be applied. One of the new techniques discussed in Coombs (1964) was a ordinal form of MDS, in which not only the data but also the representation of the data in Euclidean space were non-metric.

John van de Geer had just published Van de Geer (1967). In that book, and in the subsequent book Van de Geer (1971), he developed his unique geometric approach to multivariate analysis. Relationship between variables, and between variables and individuals, were not just discussed using matrix algebra, but were also visualized in diagrams. This was related to the geometric representations in Coombs' Theory of Data, but it concentrated on numerical data in the form of rectangular matrices of objects by variables.

Looking back it is easy to see that both Van de Geer and Coombs influenced my approach to data analysis. I inherited the emphasis on non-metric data and on visualization. But, from the beginning, I interpreted “Data Theory” as “Data Analysis”, with my emphasis shifting to techniques, loss functions, implementations, algorithms, optimization, computing, and programming. This is of interest because in 2020 my former Department of Statistics at UCLA, together with the Department of Mathematics, started a bachelor’s program in Data Theory, in which “Emphasis is placed on the development and theoretical support of a statistical model or algorithmic approach. Alternatively, students may undertake research on the foundations of data science, studying advanced topics and writing a senior thesis.” This sounds like a nice hybrid of Data Theory and Data Analysis, with a dash of computer science mixed in.

Computing and optimization were in the air in 1968, not so much because of Coombs, but mainly because of Roger Shepard, Joe Kruskal, and Doug Carroll at Bell Labs in Murray Hill. John’s other student Eddie Roskam and I were fascinated by getting numerical representations from ordinal data by minimizing explicit least squares loss functions. Eddie wrote his dissertation in 1968 (Roskam (1968)). In 1973 I went to Bell Labs for a year, and Eddie went to Michigan around the same time to work with Jim Lingoes, resulting in Lingoes and Roskam (1973).

My first semi-publication was De Leeuw (1968), quickly followed by a long sequence of other, admittedly rambling, internal reports. Despite this very informal form of publication the sheer volume of them got the attention of Joe Kruskal and Doug Carroll, and I was invited to spend the academic year 1973-1974 at Bell Laboratories. That visit somewhat modified my cavalier approach to publication, but I did not become half-serious in that respect until meeting with Forrest Young and Yoshio Takane at the August 1975 US-Japan seminar on MDS in La Jolla. Together we used the alternating least squares approach to algorithm construction that I had developed since 1968 into a quite formidable five-year publication machine, with at its zenith Takane, Young, and De Leeuw (1977).

In La Jolla I gave the first presentation of the majorization method for MDS, later known as smacof, with the first formal convergence proof. The canonical account of smacof was published in a conference paper (De Leeuw (1977)). Again I did not bother to get the results into a journal or into some other more effective form of publication. The basic theory for what became known as smacof was also presented around the same time in another book chapter De Leeuw and Heiser (1977).

In 1978 I was invited to the Fifth International Symposium on Multivariate Analysis in Pittsburgh to present what eventually became De Leeuw and Heiser (1980). There I met Nan Laird, one of the authors of the basic paper on the EM algorithm (Dempster, Laird, and Rubin (1977)). I remember enthusiastically telling her on the conference bus that EM and smacof were both special case of the general majorization approach to algorithm construction, which was consequently born around the same time. But that is a story for a companion volume, which currently only exists in a very preliminary stage (https://github.com/deleeuw/bras).

My 1973 PhD thesis (De Leeuw (1973), reprinted as De Leeuw (1984)) was actually my second attempt at a dissertation. I had to get a PhD, any PhD, before going to Bell Labs, because of the difference between the Dutch and American academic title and reward systems. I started writing a dissertation on MDS, in the spirit of what later became De Leeuw and Heiser (1982). But halfway through I lost interest and got impatient, and I decided to switch to nonlinear multivariate analysis. This second attempt did produced a finished dissertation (De Leeuw (1973)), which grew over time, with the help of multitudes, into Gifi (1990). But that again is a different history, which I will tell some other time in yet another companion volume (https://github.com/deleeuw/gifi). For a long time I did not do much work on MDS, until the arrival of Patrick Mair and the R language led to a resurgence of my interest, and ultimately to De Leeuw and Mair (2009) and Mair, Groenen, and De Leeuw (2022).

I consider this MDS book to be a summary and extension of the basic papers De Leeuw (1977), De Leeuw and Heiser (1977), De Leeuw and Heiser (1980), De Leeuw and Heiser (1982), and De Leeuw (1988), all written 30-40 years ago. Footprints in the sands of time. It can also be seen as an elaboration of the more mathematical and computational sections of the excellent and comprehensive textbook of Borg and Groenen (2005). That book has much more information about the origins, the data, and the applications of MDS, as well as on the interpretation of MDS solutions. In this book I concentrate almost exclusively on the mathematical, computational, and programming aspects of MDS.

For those who cannot get enough of me, there is a data base of my published and unpublished reports and papers since 1965, with links to pdf's, at https://jansweb.netlify.app/publication/.

There are many, many people I have to thank for my scientific education. Sixty years is a long time, and consequently many excellent teachers and researchers have crossed my path. I will gratefully mention the academics who had a major influence on my work and who are not with us any more, since I will join them in the not too distant future: Louis Guttman (died 1987), Clyde Coombs (died 1988), Warren Torgerson (died 1999), Forrest Young (died 2006), John van de Geer (died 2008), Joe Kruskal (died 2010), Doug Carroll (died 2011), and Rod McDonald (died 2012).

I will also use this preface to thank Rstudio, in particular J.J. Allaire, Hadley Wickham, and Yihui Xi, for their contributions to the R universe, and for their promotion of open source software and open access publications. Not too long ago I was an ardent LaTeX user, firmly convinced I would never use anything else again in my lifetime. In the same way that I was convinced before that I would never use anything besides, in that order, FORTRAN, PL/I, APL, and (X)Lisp. And PHP/Apache/MySQL. But I lived too long. And then, in my dotage, lo and behold, R, Rstudio, (R)Markdown, Quarto, ggplot, bookdown, blogdown, Git, Github, and Netlify came along.

Figure 1: Forrest Young, Bepi Pinner, Jean-Marie Bouroche, Yoshio Takane, Jan de Leeuw at La Jolla, August 1975

In this manual we study the smacof family of *Multidimensional Scaling (MDS)* techniques. In MDS the data consist of some type of information about the *dissimilarities* between a pairs of *objects*. These objects can be anything: individuals, variables, colors, locations, chemicals, molecules, works of Plato, political parties, Morse code signals, and so on. The dissimilarities can be approximate or imprecise distances, dissimilarity judgments, import/export tables, sociometric choices, and so on. They generally are *distance-like*, but we do not expect them to satisfy the triangle inequality, and in general not even non-negativity and symmetry. *Similarities*, such as confusion probabilities, correlations, or preferences, are always converted in some way or another to dissimilarities before they can serve as data for MDS.

The information we have about these dissimilarities can be numerical, ordinal, or categorical. Thus we may have the actual values of some or all of the dissimilarities, we may know their rank order, or we may have a classification of them into a small number of qualitative bins.

Let's formalize this, and introduce some notation at the same time. The set of ojects is $\mathfrak{O}$. For example, it can be the set of all cities with more than 10,000 inhabitants. In our MDS analysis we only use $O := (o_1, \cdots, o_n)$, an n-tuple (i.e. a finite sequence) of n *different* elements of $\mathfrak{O}$, for example n capital cities selected from $\mathfrak{O}$. If you want to, you can call O a *sample* from $\mathfrak{O}$. It is entirely possible, however, that $\mathfrak{O}$ has only n elements, in which case O is just an permutation of the elements of $\mathfrak{O}$.

A dissimilarity is a function δ on all pairs of objects, with values in a set $\mathfrak{D}$. It can be, for example, the time in seconds for an airline flight from city one to city two. Thus $\delta : \mathfrak{O}\otimes\mathfrak{O} \Rightarrow \mathfrak{D}$. A dissimilaritry is *numerical* if $\mathfrak{D}$ is subset of real line, it is *ordinal* if $\mathfrak{D}$ is a partially ordered set, and it is *nominal* if $\mathfrak{D}$ is neither. Or a dissimilarty is nominal if $\mathfrak{D}$ is any set, and we choose to ignore the ordinal and numerical information if it is there. No matter what $\mathfrak{D}$ is, we suppose it always has the element *NA* to indicate missing dissimilarities. Cities may not have airports, for example, or we just don't have the information about the airline distances. Define $\delta_{ij} := \delta(o_i, o_j)$ and $\Delta := \delta(O \times O)$. We can think of Δ and an $n \times n$ matrix with elements in $\mathfrak{D}$.

MDS techniques map the objects o_i into *points* x_i in some metric space $\langle \mathfrak{X}, d \rangle$ in such a way that the distances between pairs of points approximate the dissimilarities of the corresponding pairs of objects. Thus we want to find a map $x : \mathfrak{O} \rightarrow \mathfrak{X}$ that produces an n-tuple $X = (x_1, \cdots, x_n)$ of elements of $\mathfrak{X}$, where $x_i := x(o_i)$. Also define $d_{ij} := d(x_i, x_j)$ and $D(X) := d(X \times X$. Unlike the dissimilarities the d_{ij} are always numerical, because distances are. So MDS finds X such that $D(X) \approx \Delta$.

For numerical dissimilarities it is clear what "approximation" means, we simply want the distances and the corresponding dissimilarities to be numerically close. Because there are generally many dissimilarities and distances a combined measure of closeness can still be defined in many different ways. For ordinal and nominal dissimilarities the notion of approximation is less clear, and we have to develop more specialized techniques to measure how well the distances fit the dissimilarities.

1 Brief History

De Leeuw and Heiser (1980)

This section has a different emphasis. We limit ourselves to developments in Euclidean MDS, and to contributions with direct computational consequences that have a direct or indirect link to psychometrics, and to work before 1960. This is reviewed ably in the presidential address of W. S. Torgerson (1965).

Our history review takes the form of brief summaries of what we consider to be milestone papers or books.

1.1 Prehistory

Young-Householder, etc.

1.2 Torgerson

W. S. Torgerson (1952) W. S. Torgerson (1965)

1.3 Bell Laboratories

Shepard (1962a) Shepard (1962b)

Kruskal (1964a) Kruskal (1964b)

1.4 Guttman-Lingoes

Guttman (1968)

1.5 Alternating Least Squares

1.6 Majorization

De Leeuw (1977) De Leeuw and Heiser (1977)

There was some early work by Richardson, Messick, Abelson and Torgerson who combined Thurstonian scaling of similarities with the mathematical results of Schoenberg (1935) and Young and Householder (1938).

Despite these early contributions it makes sense, certainly from the point of view of my personal history, but probably more generally, to think of MDS as starting as a widely discussed, used, and accepted technique since the book by W. S. Torgerson (1958). This was despite the fact that in the fifties and sixties computing eigenvalues and eigenvectors of a matrix of size 20 or 30 was still a considerable challenge.

A few years later the popularity of MDS got a large boost by developments centered at Bell Telephone Laboratories in Murray Hill, New Jersey, the magnificent precursor of Silicon Valley. First there was nonmetric MDS by Shepard (1962a), Shepard (1962b) and Kruskal (1964a), Kruskal (1964b), And later another major development was the introduction of individual difference scaling by Carroll and Chang (1970) and Harshman (1970). Perhaps even more important was the development of computer implementations of these new techniques. Some of the early history of nonmetric MDS is in De Leeuw (2017a).

Around the same time there were interesting theoretical contributions in Coombs (1964), which however did not much influence the practice of MDS. ….. And several relatively minor variations of the Bell Laboratories approach were proposed by Guttman (1968), but Guttman's influence on further MDS implementations turned out to be fairly localized and limited.

The main development in comptational MDS after the Bell Laboratories surge was probably smacof. Initially, in De Leeuw (1977), this stood for *Scaling by Maximizing a Convex Function*. Later it was also used to mean *Scaling by Majorizing a Complicated Function*. Whatever. In this book smacof just stands for smacof. No italics, no boldface, no capitals.

The first smacof programs were written in 1977 in FORTRAN at the Department of Data Theory in Leiden (Heiser and De Leeuw (1977)). Eventually they migrated to SPSS (for example, Meulman and Heiser (2012)) and to R (De Leeuw and Mair (2009)). The SPSS branch, now the IBM SPSS branch, and the R branch have diverged somewhat, and they continue to be developed independently.

Parallel to this book there is an attempt to rewrite the various smacof programs in C, with the necessary wrappers to call them from R (De Leeuw (2017b)). The C code, with makefiles and test routines, is at github.com/deleeuw/smacof

2 Basic MDS

Following Kruskal, and to a lesser extent Shepard, we measure the fit of distances to dissimilarities using an explicit real-valued *loss function* (or *badness-of-fit measure*), which is minimized over the possible maps of the objects into the metric space. This is a very general definition of MDS, covering all kinds of variations of the target metric space and of the way fit is measured. Obviously we will not discuss *all* these possible forms of MDS, which also includes various techniques more properly discussed as cluster analysis, classification, or discrimination.

To fix our scope we first define *basic MDS*, which is short for *Least Squares Euclidean Metric MDS*. It is defined as MDS with the following characteristics.

1. The metric space is a Euclidean space.
2. The dissimilarities are numerical, symmetric, and non-negative.
3. The loss function is a weighted sum of squares of the *residuals*, which are the differences between dissimilarities and Euclidean distances.
4. Weights are numerical, symmetric, and non-negative.
5. Self-dissimilarities are zero and the corresponding terms in the loss function also have weight zero.

By a *Euclidean space* we mean a finite dimensional vector space, with addition and scalar multiplication, and with an inner product that defines the distances. For the *inner product* of vectors x and y we write $\langle x, y\rangle$. The *norm* of x is defined as $\|x\| := \sqrt{\langle x, x\rangle}$, and the *distance* between x and y is $d(x, y) := \|x - y\|$.

The *loss function* we use is called *stress*. It was first explicitly introduced in MDS as *raw stress* by Kruskal (1964a) and Kruskal (1964b). We define stress in a slightly different way, because we want to be consistent over the whole range of the smacof versions and implementations. In smacof stress is the real-valued function σ, defined on the space $\mathbb{R}^{n\times p}$ of configurations, as

$$\sigma(X) := \frac{1}{4}\sum_{i=1}^n\sum_{j=1}^n w_{ij}(\delta_{ij} - d_{ij}(X))^2.(\#eq:stressall) \tag{1}$$

Note that we use $:=$ for definitions, i.e. for concepts and symbols that are not standard mathematical usage, when they occur for the first time in this book. Through the course of the book it will probably become clear why the mysterious factor $\frac{1}{4}$ is there. Clearly it has no influence on the actual minimization of the loss function.

In definition @ref(eq:stressall) we use the following objects and symbols.

1. $W = \{w_{ij}\}$ is a symmetric, non-negative, and hollow matrix of *weights*, where *hollow* means zero diagonal.

2. $\Delta = \{\delta_{ij}\}$ is a symmetric, non-negative, and hollow matrix of *dissimilarities*.
3. X is an $n \times p$ *configuration*, containing coordinates of n *points* in p dimensions.
4. $D(X) = \{d_{ij}(X)\}$ is a symmetric, non-negative, and hollow matrix of *Euclidean distances* between the n points in X. Thus $d_{ij}(X) := \sqrt{\sum_{s=1}^p (x_{is} - x_{js})^2}$.

Note that symmetry and hollowness of the basic objects W, Δ, and D allows us carry out the summation of the weighted squared residuals in formula @ref(eq:stressall) over the upper (or lower) diagonal elements only. Thus we can also write

$$\sigma(X) := \frac{1}{2}\sum\sum_{1\leq i<j\leq n} w_{ij}(\delta_{ij} - d_{ij}(X))^2 .(\#eq:stresshalf) \tag{2}$$

We use the notation $\sum\sum_{1\leq i<j\leq n}$ for summation over the lower-diagonal elements of a matrix.

The function D, which computes the distance matrix $D(X)$ from a configuration X, is matrix-valued. It maps the $n \times p$-dimensional *configuration space* $\mathbb{R}^{n\times p}$ into the set $D(\mathbb{R}^{n\times p})$ of Euclidean distance matrices between n points in $\mathbb{R}^p$, which is a subset of the convex cone of hollow, symmetric, non-negative matrices in the linear space $\mathbb{R}^{n\times n}$ (Datorro (2018)).

In basic MDS the weights and dissimilarities are given numbers, and we minimize stress over all $n \times p$ configurations X. Note that the *dimensionality* p is also supposed to be known beforehand, and that MDS in p dimensions is different from MDS in $q \neq p$ dimensions. We sometimes emphasize this by writing $pMDS$, which indicates that we will map the points into p-dimensional space.

Two boundary cases that will interest us are *Unidimensional Scaling* or *UDS*, where $p = 1$, and *Full-dimensional Scaling* or *FDS*, where $p = n$. Thus UDS is 1MDS and FDS is nMDS. Most actual MDS applications in the sciences use 1MDS, 2MDS or 3MDS, because configurations in one, two, or three dimensions can easily be plotted with standard graphics tools. Note that MDS is not primarily a tool to tests hypotheses about dimensionality and to find meaningful dimensions. It is a mostly a mapping tool for data reduction, to graphically find interesting aspects of dissimilarity matrices.

The projections on the dimensions are usually ignored, it is the configuration of points that is the interesting outcome. This distinguishes MDS from, for example, factor analysis. There is no Varimax, Oblimax, Quartimax, and so on. Exceptions are confirmatory applications of MDS in genetic mapping along the chromosome, in archeological seriation, in testing psychological theories of cognition and representation, in the conformation of molecules, and in geographic and geological applications. In these areas the dimensionality and general structure of the configuration are given by prior knowledge, we just do not know the precise location and distances of the points. For more discussion of the different uses of MDS we refer to De Leeuw and Heiser (1982).

2.1 Kruskal's stress

Definition @ref(eq:stressall) differs from Kruskal's original stress in at least three ways: in Kruskal's use of the square root, in our use of weights, and in our different approach to normalization.

We have paid so much attention to Kruskal's original definition, because the choices made there will play a role in the normalization discussion in the ordinal scaling chapter (section @ref(nmdsnorm)), in the comparison of Kruskal's and Guttman's approach to ordinal MDS (sections @ref(nmdskruskal) and @ref(nmdsguttman)), and in our discussions about the differences between Kruskal's stress @ref(eq:kruskalstressfinal) and smacof's stress @ref(eq:stressall) in the next three sections of this chapter.

2.1.0.1 Square root

Let's discuss the square root first. Using it or not using it does not make a difference for the minimization problem. Using the square root, however, does give a more sensible root-mean-square scale, in which stress is homogeneous of degree one, instead of degree two. But I do not want to compute all those unnecessary square roots in my algorithms, and I do not want to drag them along through my derivations. Moreover the square root potentially causes problems with differentiability at those X where $\sigma(X)$ is zero. Thus, througout the book, we do not use the square root in our formulas and derivations. In fact, we do not even use it in our computer programs, except at the very last moment when we return the final stress after the algorithm has completed.

2.1.0.2 Weights

There were no weights $W = \{w_{ij}\}$ in the original definition of stress by Kruskal (1964a), and neither are they there in most of the basic later contributions to MDS by Guttman, Lingoes, Roskam, Ramsay, or Young. We will use weights throughout the book, because they have various interesting applications within basic MDS, without unduly complicating the derivations and computations. In Groenen and Van de Velden (2016), section 6, the various uses of weights in the stress loss function are enumerated. They generously, and correctly, attribute the consistent use of weights in MDS to me. I quote from their paper:

> 1. Handling missing data is done by specifying $w_{ij} = 0$ for missings and 1 otherwise thereby ignoring the error corresponding to the missing dissimilarities.
> 2. Correcting for nonuniform distributions of the dissimilarities to avoid dominance of the most frequently occurring dissimilarities.

3. Mimicking alternative fit functions for MDS by minimizing Stress with w_{ij} being a function of the dissimilarities.
4. Using a power of the dissimilarities to emphasize the fitting of either large or small dissimilarities.
5. Special patterns of weights for specific models.
6. Using a specific choice of weights to avoid nonuniqueness.

In some situations, for example for huge data sets, it is computationally convenient, or even necessary, to minimize the influence of the weights on the computations. We can use *majorization* to turn the problem from a weighted least squares problem to an iterative unweighted least squares problem. The technique, which we call *unweighting*, is discussed in detail in section @ref(minunweight).

2.1.0.3 Normalization

This section deals with a rather trivial problem, which has however caused problems in various stages of smacof's 50-year development history. Because the problem is trivial, and the choices that must be made are to a large extent arbitrary, it has been overlooked and somewhat neglected.

In basic MDS we scale the weights and dissimilarities. It is clear that if we multiply the weights or dissimilarities by a constant, then the optimal approximating distances $D(X)$ and the optimal configuration X will be multiplied by the same constant. That is exactly why Kruskal's raw stress had to be normalized. Consequently we in basic MDS we always scale weights and dissimilarities by

$$\sum\sum_{1\leq i<j\leq n} w_{ij}=1, (\#eq:scaldiss1) \tag{3}$$
$$\sum\sum_{1\leq i<j\leq n} w_{ij}\delta_{ij}^2=1.(\#eq:scaldiss2) \tag{4}$$

This simplifies our formulas and makes them look better (see, for example, section @ref(prop-expand) and section @ref(secrhostress)). It presupposes, of course, that $w_{ij}\delta_{ij}\neq 0$ for at least one $i\neq j$, which we will happily assume in the sequel, because otherwise the MDS problem is trivial. Note that if all weights are equal (which we call the *unweighted case*) then they are equal to $1/\binom{n}{2}$ and thus we require $\sum\sum_{1\leq i<j\leq n}\delta_{ij}^2=\frac12 n(n-1)$.

Using normalized dissimilarities amounts to the same defining stress as

$$\sigma(X)=\frac12\frac{\sum\sum_{1\leq i<j\leq n} w_{ij}(\delta_{ij}^2-d_{ij}(X))^2}{\sum\sum_{1\leq i<j\leq n} w_{ij}\delta_{ij}^2}.(\#eq:stressrat) \tag{5}$$

This is useful to remember when we discuss the various normalizations for non-metric MDS in section @ref(nmdsnorm).

2.2 Local and Global Minima

In basic MDS our goal is to compute both $\min_X \sigma(X)$ and $\mathop{\mathrm{Argmin}}_X \sigma(X)$, where $\sigma(X)$ is defined as @ref(eq:stressall), and where we minimize over all configurations in $\mathbb{R}^{n\times p}$.

In this book we study both the properties of the stress loss function and a some of its generalizations, and the various ways to minimize these loss functions over configurations (and sometimes over transformations of the dissimilarities as well).

Emphasis local minima

Compute stationary points

2.3 Partitioning Loss

3 Generalizations

The most important generalizations of basic MDS we will study in later chapters of this book are discussed briefly in the following sections.

3.1 Non-linear MDS

3.2 Non-metric MDS

Basic MDS is a form of *Metric Multidimensional Scaling* or *MMDS*, in which dissimilarities are either known or missing. In chapter @ref(nonmtrmds) we relax this assumption. Dissimilarities may be partly known, for example we may know they are in some interval, we may only know their order, or we may know them up to some smooth transformation. MDS with partly known dissimilarities is *Non-metric Multidimensional Scaling* or *NMDS*. Completely unknown (missing) dissimilarities are an exception, because we can just handle this in basic MDS by setting the corresponding weights equal to zero.

In NMDS we minimize stress over all configurations, but also over the unknown dissimilarities. What we know about them (the interval they are in, the transformations that are allowed, the order they are in) defines a subset of the space of non-negative, hollow, and symmetric matrices.

Any matrix in that subset is a matrix of what Takane, Young, and De Leeuw (1977) call *disparities*, i.e. imputed dissimilarities. The imputation provides the missing information and transforms the non-numerical information we have about the dissimilarities into a numerical matrix of disparities. Clearly this is an *optimistic imputation*, in the sense that it chooses from the set of admissible disparities to minimize stress (for a given configuration).

One more terminological point. Often *non-metric* is reserved for ordinal MDS, in which we only know a (partial or complete) order of the dissimilarities. Allowing linear or polynomial transformations of the dissimilarities, or estimating an additive constant, is then supposed to be a form of metric MDS. There is something to be said for that. Maybe it makes sense to distinguish non-metric *in the wide sense* (in which stress must be minimized over both X and Δ) and *non-metric in the narrow sense* in which the set of admissible disparities is defined by linear inequalities. Nonmetric in the narrow sense will also be called *ordinal MDS* or *OMDS*.

It is perhaps useful to remember that Kruskal (1964a) introduced explicit loss functions in MDS to put the somewhat heuristic NMDS techniques of Shepard (1962a) onto a firm mathematical and computational foundation. Thus, more or less from the beginning of iterative least squares MDS, there was a focus on non-metric rather than metric MDS, and this actually contributed a great deal to the magic and success of the technique. In this book most of the results are derived for basic MDS, which is metric MDS, with non-metric MDS as a relatively straightforward extension not discussed until chapter @ref(nonmtrmds). So, at least initially, we take the numerical values of the dissimilarities seriously, as do W. S. Torgerson (1958) and Shepard (1962a), Shepard (1962b).

It may be the case that in the social and behavioural sciences only the ordinal information in the dissimilarities is reliable and useful. But, since 1964, MDS has also been applied in molecular conformation, chemometrics, genetic sequencing, archelogical seriation, and in network design and location analysis. In these areas the numerical information in the dissimilarities is usually meaningful and should not be thrown out right away. Also, the use of the Shepard plot, with dissimilarities on the horizontal axis and fitted distances on the vertical axis, suggests there is more to dissimilarities than just their rank order.

3.3 Fstress and Friends

Instead of defining the residuals in the least squares loss function as $\delta_{ij} - d_{ij}(X)$ chapter @ref(chrstress) discusses the more general cases where the residuals are $f(\delta_{ij}) - g(d_{ij}(X))$ for some known non-negative increasing function f. This defines the *fstress* loss function.

If $f(x) = x^r$ with $r > 0$ then fstress is called *rstress*. Thus stress is rstress with $r = 1$, also written as *1stress* or σ_1. In more detail we will also look at $r = 2$, which is called *sstress* by Takane, Young, and De Leeuw (1977). In chapter @ref(chsstressstrain) we look at the problem of minimizing sstress and weighted version *strain*. The case of rstress with $r \rightarrow 0$ is also of interest, because it leads to the loss function in Ramsay (1977).

3.4 Constraints

Instead of minimizing stress over all X in $\mathbb{R}^{n\times p}$ we will look in chapter @ref(cmds) at various generalizations where minimization is over a subset Ω of $\mathbb{R}^{n\times p}$. This is often called *Constrained Multidimensional Scaling* or *CMDS*.

The distinction may be familiar from factor analysis, where we distinguish between exploratory and confirmatory factor analysis. If we have prior information about the parameters then incorporating that prior information in the analysis will generally lead to more precise and more interpretable estimates. The risk is, of course that if our prior information is wrong, if it is just prejudice, then we will have a solution which is precise but incorrect. We have the famous trade-off between bias and variance. In MDS this trade-off does not seem to apply directly, because the necessary replication frameworks are missing.

and we do not attach much value to locating the true configuration.

Primal and Dual

$$\min_{X\in\Omega}\sigma(X)$$

$$\min_{X}\sigma(X)+\lambda\kappa(X,\Omega)$$

where $\kappa(X,\Omega)\geq 0$ and $\kappa(X,\Omega)=0$ if and only if $X\in\Omega$.

3.5 Individual Differences

Now consider the situation in which we have m different dissimilarity matrices Δ_k and m different weight matrices W_k. We generalize basic MDS by defining

$$\sigma(X_1,\cdots,X_m):=\frac{1}{2}\sum_{k=1}^{m}\sum_{1\leq i<j\leq n}\sum w_{ijk}(\delta_{ijk}-d_{ij}(X_k))^2,(\#eq:replistress)\tag{6}$$

and minimize this over the X_k.

There are two simple ways to deal with this generalization. The first is to put no further constraints on the X_k. This means solving m separate basic MDS problems, one for each k. The second way is to require that all X_k are equal. As shown in more detail in section

@ref(indifrepl) this reduced to a single basic MDS problem with dissimilarities that are a weighted sum of the Δ_k. So both these approaches do not really bring anything new.

Minimizing @ref(eq:replistress) becomes more interesting if we constrain the X_k in various ways. Usually this is done by making sure they have a component that is common to all k and a component that is specific or unique to each k. This approach, which generalizes constrained MDS, is discussed in detail in chapter @ref(chindif).

3.6 Asymmetry

We have seen in section @ref(datasym) of this chapter that in basic MDS the assumption that W and Δ are symmetric and hollow can be made without loss of generality. The simple partitioning which proved this was based on the fact that $D(X)$ is always symmetric and hollow. By the way, the assumption that W and D are non-negative cannot be made without loss of generality, as we will see below.

In chapter @ref(asymmds) we relax the assumption that $D(X)$ is symmetric (still requiring it to be non-negative and hollow). This could be called *Asymmetric MDS*, or *AMDS*. I was reluctant at first to include this chapter, because asymmetric distances do not exist. And certainly are not Euclidean distances, so they are not covered by the title of this book. But as long as we stay close to Euclidean distances, least squares, and the smacof approach, I now feel reasonably confident the chapter is not too much of a foreign body.

3.7 Non-Euclidean Distances

When Kruskal introduced gradient-based methods to minimize stress he also discussed the possibility to use Minkovski metrics other than the Euclidean metric. This certainly was part of the appeal of the new methods, in fact it seemed as if the gradient methods made it possible to use any distance function whatsoever. This initial feeling of empowerment was somewhat naive, because it ignored the seriousness of the local minimum problem, the combinatorial nature of one-dimensional and city block scaling, the problems with nonmetric unfolding, and the problematic nature of gradient methods if the distances are not everywhere differentiable. All these complications will be discussed in this book. But it made me decide to ignore Minkovski distances (and hyperbolic and elliptic non-Euclidean distances), because life with stress is complicated and challenging enough as it is.

Groenen, Mathar, and Heiser (1995), Mathar and Meyer (1994)

4 Principles of Algorithm Constuction

4.1 Alternating Least Squares

4.2 Majorization

4.3 Introduction to Majorization

Majorization, these days better known as MM (Lange (2016)), is a general approach for the construction of minimization algorithms. There is also minorization, which leads to maximization algorithms, which explains the MM acronym: minorization for maximization and majorization for minimization.

Before the MM principle was formulated as a general approach to algorithm construction there were some important predecessors. Major classes of MM algorithms avant la lettre were the EM Algorithm for maximum likelihood estimation of Dempster, Laird, and Rubin (1977), the *Smacof Algorithm* for MDS of De Leeuw (1977), the Generalized Weiszfeldt Method* of Vosz and Eckhardt (1980), and the Quadratic Approximation Method of Böhning and Lindsay (1988). The first formulation of the general majorization principle seems to be De Leeuw (1994).

Let's start with a brief introduction to majorization. Minimize a real valued function σ over $x \in \mathbb{S}$, where $\mathbb{S}$ is some subset of $\mathbb{R}^n$. There are obvious extensions of majorization to functions defined on more general spaces, with values in any partially ordered set, but we do not need that level of generality in this manual. Also majorization applied to σ is minorization applied to $-\sigma$, so concentrating on majorization-minimization and ignoring minorization-maximization causes no loss of generality

Suppose there is a real-valued function ω on $\mathbb{S} \otimes \mathbb{S}$ such that

$$\sigma(x) \leq \omega(x,y) \qquad \forall x,y \in \mathbb{S}, \tag{7}$$

$$\sigma(x) = \omega(x,x) \qquad \forall x \in \mathbb{S}. \tag{8}$$

The function ω is called a *majorization scheme* for σ on S. A majorization scheme is *strict* if $\sigma(x) < \omega(x,y)$ for all $x,y \in S$ withj $x \neq y$.

Define

$$x^{(k+1)} \in \mathop{\mathrm{argmin}}_{x \in \mathbb{S}} \omega(x, x^{(k)}), \tag{9}$$

assuming that ω attains its (not necessarily unique) minimum over $x \in \mathbb{S}$ for each y. If $x^{(k)} \in \mathrm{argmin}_{x\in\mathbb{S}}\, \omega(x, x^{(k)})$ then we stop.

By majorization property (7) $\sigma(x^{(k+1)})\leq\omega(x^{(k+1)},x^{(k)})$. Because we did not stop update rule Equation 9 implies $\omega(x^{(k+1)},x^{(k)})<\omega(x^{(k)},x^{(k)})$. and finally by majorization property (8) $\omega(x^{(k)},x^{(k)})=\sigma(x^{(k)})$.

If the minimum in Equation 9 is attained for a unique x then $\omega(x^{(k+1)},x^{(k)})<\omega(x^{(k)},x^{(k)})$. If the majorization scheme is strict then $\sigma(x^{(k+1)})<\omega(x^{(k+1)},x^{(k)})$. Under either of these two additional conditions $\sigma(x^{(k+1)})<\sigma(x^{(k)})$, which means that the majorization algorithm is a monotone descent algorithm, and if σ is bounded below on $\mathbb{S}$ the sequence $\sigma(x^{(k)})$ converges.

Note that we only use the order relation to prove convergence of the sequence of function values. To prove convergence of the $x^{(k)}$ we need stronger compactness and continuity assumptions to apply the general theory of Zangwill (1969). For such a proof the argmin in update formula Equation 9 can be generalized to $x^{(k+1)}=\phi(x^{(k)})$, where ϕ maps $\mathbb{S}$ into $\mathbb{S}$ such that $\omega(\phi(x),x)\leq\sigma(x)$ for all x.

We give a small illustration in which we minimize σ with $\sigma(x)=\sqrt{x}-\log x$ over $x>0$. Obviously we do not need majorization here, because solving $\mathcal{D}\sigma(x)=0$ immediately gives $x=4$ as the solution we are looking for.

To arrive at this solution using majorization we start with

$$\sqrt{x}\leq\sqrt{y}+\frac{1}{2}\frac{x-y}{\sqrt{y}},(\#eq:sqrtmaj) \tag{10}$$

which is true because a differentiable concave function such as the square root is majorized by its tangent everywhere. Inequality @ref(eq:sqrtmaj) implies

$$\sigma(x)\leq\eta(x,y):=\sqrt{y}+\frac{1}{2}\frac{x-y}{\sqrt{y}}-\log x.(\#eq:examplemaj) \tag{11}$$

Note that $\eta(\bullet,y)$ is convex in its first argument for each y. We have $\mathcal{D}_1\eta(x,y)=0$ if and only if $x=2\sqrt{y}$ and thus the majorization algorithm is

$$x^{(k+1)}=2\sqrt{x^{(k)}}(\#eq:examplealg) \tag{12}$$

The sequence $x^{(k)}$ converges monotonically to the fixed point $x=2\sqrt{x}$, i.e. to $x=4$. If $x^{(0)}<4$ the sequence is increasing, if $x^{(0)}<4$ it is decreasing. Also, by l'Hôpital,

$$\lim_{x\rightarrow 4}\frac{2\sqrt{x}-4}{x-4}=\frac{1}{2}(\#eq:hopi1) \tag{13}$$

and thus convergence to the minimizer is linear with asymptotic convergence rate $\frac{1}{2}$. By another application of l'Hôpital

$$\lim_{x\rightarrow 4}\frac{\sigma(2\sqrt{x}))-\sigma(4)}{\sigma(x)-\sigma(4)}=\frac{1}{4},(\#eq:hopi2) \tag{14}$$

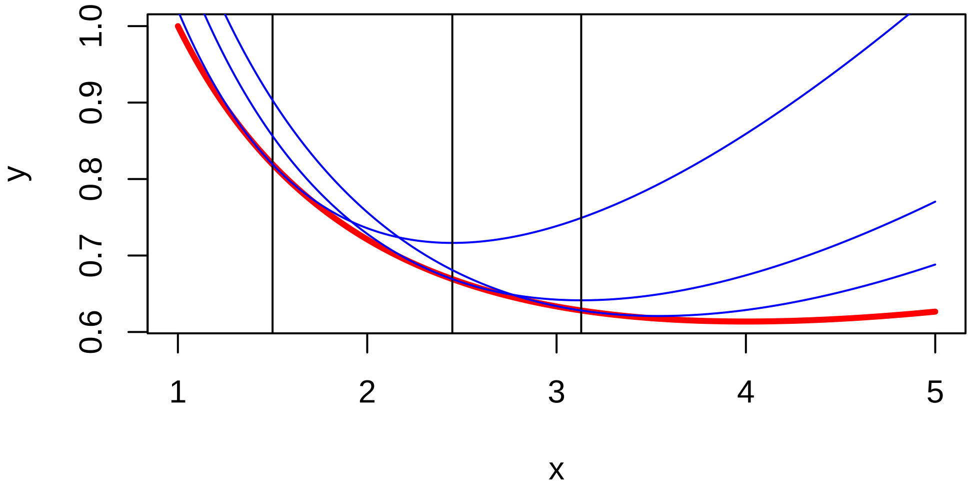
and convergence to the minimum is linear with asymptotic convergence rate $\frac{1}{4}$. Linear convergence to the minimizer is typical for majorization algorithms, as is the twice-as-fast linear convergence to the minimum value.

This small example is also of interest, because we minimize a *DC function*, the difference of two convex functions. In our example the convex functions are minus the square root and minus the logarithm. Algorithms for minimizing DC functions define other important subclasses of MM algorithms, the *DC Algorithm* of Tao Pham Dinh (see Le Thi and Tao (2018) for a recent overview), the *Concave-Convex Procedure* of Yuille and Rangarajan (2003), and the *Half-Quadratic Method* of Donald Geman (see Niikolova and Ng (2005) for a recent overview). For each of these methods there is a huge literature, with surprisingly little non-overlapping literatures. The first phase of the smacof algorithm, in which we improve the configuration for given disparities, is DC, concave-convex, and half-quadratic.

In the table below we show convergence of @ref(eq:examplealg) starting at $x = 1.5$. The first column show how far $x^{(k)}$ deviates from the minimizer (i.e. from 4), the second shows how far$\sigma(x^{(k)})$ deviates from the minimum (i.e. from $2 - \log 4$). We clearly see the convergence rates $\frac{1}{2}$ and $\frac{1}{4}$ in action.

```
itel   1 2.5000000000 0.2055741244
itel   2 1.5505102572 0.0554992066
itel   3 0.8698308399 0.0144357214
itel   4 0.4615431837 0.0036822877
itel   5 0.2378427379 0.0009299530
itel   6 0.1207437506 0.0002336744
itel   7 0.0608344795 0.0000585677
itel   8 0.0305337787 0.0000146606
itel   9 0.0152961358 0.0000036675
itel  10 0.0076553935 0.0000009172
itel  11 0.0038295299 0.0000002293
itel  12 0.0019152235 0.0000000573
itel  13 0.0009577264 0.0000000143
itel  14 0.0004788919 0.0000000036
itel  15 0.0002394531 0.0000000009
```

The first three iterations are shown in the figure below. The vertical lines indicate the value of x, function is in red, and the first three majorizations are in blue.


y
1.0
0.9
0.8
0.7
0.6
1
2
3
4
5
x

References

Böhning, D., and B. G. Lindsay. 1988. "Monotonicity of Quadratic-approximation Algorithms." *Annals of the Institute of Statistical Mathematics* 40 (4): 641–63.

Borg, I., and P. J. F. Groenen. 2005. *Modern Multidimensional Scaling*. Second Edition. Springer.

Carroll, J. D., and J. J. Chang. 1970. "Analysis of Individual Differences in Multidimensional scaling via an N-way generalization of "Eckart-Young" Decomposition." *Psychometrika* 35: 283–319.

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

Datorro, J. 2018. *Convex Optimization and Euclidean Distance Geometry*. Second Edition. Palo Alto, CA: Meebo Publishing. https://ccrma.stanford.edu/~dattorro/0976401304.pdf.

De Leeuw, J. 1968. "Nonmetric Multidimensional Scaling." Research Note 010-68. Department of Data Theory FSW/RUL. https://jansweb.netlify.app/publication/deleeuw-r-68-g/deleeuw-r-68-g.pdf.

———. 1973. "Canonical Analysis of Categorical Data." PhD thesis, University of Leiden, The Netherlands. https://jansweb.netlify.app/publication/deleeuw-b-73/deleeuw-b-73.pdf.

———. 1977. "Applications of Convex Analysis to Multidimensional Scaling." In *Recent Developments in Statistics*, edited by J. R. Barra, F. Brodeau, G. Romier, and B. Van Cutsem, 133–45. Amsterdam, The Netherlands: North Holland Publishing Company.

———. 1984. *Canonical Analysis of Categorical Data*. Leiden, The Netherlands: DSWO Press. https://jansweb.netlify.app/publication/deleeuw-b-84/deleeuw-b-84.pdf.

———. 1988. "Convergence of the Majorization Method for Multidimensional Scaling." *Journal of Classification* 5: 163–80.

———. 1994. "Block Relaxation Algorithms in Statistics." In *Information Systems and Data Analysis*, edited by H. H. Bock, W. Lenski, and M. M. Richter, 308–24. Berlin: Springer Verlag. https://jansweb.netlify.app/publication/deleeuw-c-94-c/deleeuw-c-94-c.pdf.

———. 2017a. "Shepard Non-metric Multidimensional Scaling." 2017. https://jansweb.netlify.app/publication/deleeuw-e-17-e/deleeuw-e-17-e.pdf.

———. 2017b. "Tweaking the SMACOF Engine." 2017. https://jansweb.netlify.app/publication/deleeuw-e-17-p/deleeuw-e-17-p.pdf.

De Leeuw, J., and W. J. Heiser. 1977. "Convergence of Correction Matrix Algorithms for Multidimensional Scaling." In *Geometric Representations of Relational Data*, edited by J. C. Lingoes, 735–53. Ann Arbor, Michigan: Mathesis Press.

———. 1980. "Multidimensional Scaling with Restrictions on the Configuration." In *Multivariate Analysis, Volume V*, edited by P. R. Krishnaiah, 501–22. Amsterdam, The Netherlands: North Holland Publishing Company.

———. 1982. "Theory of Multidimensional Scaling." In *Handbook of Statistics, Volume II*, edited by P. R. Krishnaiah and L. Kanal. Amsterdam, The Netherlands: North Holland Publishing Company.

De Leeuw, J., and P. Mair. 2009. "Multidimensional Scaling Using Majorization: SMACOF in R." *Journal of Statistical Software* 31 (3): 1–30. https://www.jstatsoft.org/article/view/v031i03.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. "Maximum Likelihood for Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society* B39: 1–38.

Gifi, A. 1990. *Nonlinear Multivariate Analysis*. New York, N.Y.: Wiley.

Groenen, P. J. F., R. Mathar, and W. J. Heiser. 1995. "The Majorization Approach to Multidimensional Scaling for Minkowski Distances." *Journal of Classification* 12: 3–19.

Groenen, P. J. F., and M. Van de Velden. 2016. "Multidimensional Scaling by Majorization: A Review." *Journal of Statistical Software* 73 (8): 1–26. https://www.jstatsoft.org/index.php/jss/article/view/v073i08.

Guttman, L. 1968. "A General Nonmetric Technique for Fitting the Smallest Coordinate Space for a Configuration of Points." *Psychometrika* 33: 469–506.

Harshman, R. A. 1970. "Foundations of the PARAFAC Procedure." Working Papers in Phonetics 16. UCLA.

Heiser, W. J., and J. De Leeuw. 1977. "How to Use SMACOF-I." Department of Data Theory FSW/RUL.

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

———. 1964b. "Nonmetric Multidimensional Scaling: a Numerical Method." *Psychometrika* 29: 115–29.

Lange, K. 2016. *MM Optimization Algorithms*. SIAM.

Le Thi, H. A., and P. D. Tao. 2018. "DC Programming and DCA: Thirty Years of Developments." *Mathematical Programming, Series B*.

Lingoes, J. C., and E. E. Roskam. 1973. "A Mathematical and Empirical Analysis of Two Multidimensional Scaling Algorithms." *Psychometrika* 38: Monograph Supplement.

Mair, P., P. J. F. Groenen, and J. De Leeuw. 2022. "More on Multidimensional Scaling in R: smacof Version 2." *Journal of Statistical Software* 102 (10): 1–47. https://www.jstatsoft.org/article/view/v102i10.

Mathar, R., and R. Meyer. 1994. "Algorithms in Convex Analysis to Fit l_p -Distance Matrices." *Journal of Multivariate Analysis* 51: 102–20.

Meulman, J. J., and W. J. Heiser. 2012. *IBM SPSS Categories 21*. IBM Corporation.

Niikolova, M., and M. Ng. 2005. "Analysis of Half-Quadratic Minimization Methods for Signal and Image Recovery." *SIAM Journal Scientific Computing* 27 (3): 937–66.

Ramsay, J. O. 1977. "Maximum Likelihood Estimation in Multidimensional Scaling." *Psychometrika* 42: 241–66.

Roskam, E. E. 1968. "Metric Analysis of Ordinal Data in Psychology." PhD thesis, University of Leiden.

Schoenberg, I. J. 1935. "Remarks to Maurice Frechet's article: Sur la Definition Axiomatique d'une Classe d'Espaces Vectoriels Distancies Applicables Vectoriellement sur l'Espace de Hllbert." *Annals of Mathematics* 36: 724–32.

Shepard, R. N. 1962a. “The Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function. I.” *Psychometrika* 27: 125–40.

———. 1962b. “The Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function. II.” *Psychometrika* 27: 219–46.

Takane, Y., F. W. Young, and J. De Leeuw. 1977. “Nonmetric Individual Differences in Multidimensional Scaling: An Alternating Least Squares Method with Optimal Scaling Features.” *Psychometrika* 42: 7–67.

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

Torgerson, W. S. 1952. “Multidimensional Scaling: I. Theory and Method.” *Psychometrika* 17 (4): 401–19.

———. 1965. “Multidimensional Scaling of Similarity.” *Psychometrika* 30 (4): 379–93.

Van de Geer, J. P. 1967. *Inleiding in de Multivariate Analyse*. Van Loghum Slaterus.

———. 1971. *Introduction to Multivariate Analysis for the Social Sciences*. San Francisco, CA: Freeman.

Vosz, H., and U. Eckhardt. 1980. “Linear Convergence of Generalized Weiszfeld’s Method.” *Computing* 25: 243–51.

Young, G., and A. S. Householder. 1938. “Discussion of a Set of Points in Terms of Their Mutual Distances.” *Psychometrika* 3 (19-22).

Yuille, A. L., and A. Rangarajan. 2003. “The Concave-Convex Procedure.” *Neural Computation* 15: 915–36.

Zangwill, W. I. 1969. *Nonlinear Programming: a Unified Approach*. Englewood-Cliffs, N.J.: Prentice-Hall.