CONVERGENT COMPUTATION BY ITERATIVE MAJORIZATION: THEORY AND APPLICATIONS IN MULTIDIMENSIONAL DATA ANALYSIS

Willem J. Heiser

Department of Data Theory
University of Leiden

RR-93-06

CONVERGENT COMPUTATION BY ITERATIVE MAJORIZATION: THEORY AND APPLICATIONS IN MULTIDIMENSIONAL DATA ANALYSIS

Willem J. Heiser

Department of Data Theory
University of Leiden

december 1993

Second draft for chapter in "*Recent Advances in Descriptive Multivariate Analysis*"
Edited by Wojtek J. Krzanowski, *published by* Oxford University Press.

Convergent Computation by Iterative Majorization: Theory and Applications in Multidimensional Data Analysis

1. Introduction

Many problems in multidimensional data analysis involve the optimization of quadratic functions, due to the common assumption of normally distributed errors, together with the prevalence of linear and bilinear models. By present standards, the resulting optimization problems are of moderate complexity, frequently involving the search for eigenvectors and eigenvalues, or projections of vectors on subspaces. Even in fairly complicated situations, such as for example generalized canonical correlation analysis with optimal scaling of the variables (Van der Burg, De Leeuw, and Verdegaal, 1988), it is often possible, by partitioning the parameter space into convenient regions, to split the problem into a connected series of simpler subproblems so that monotonic convergence to at least a local minimum remains guaranteed. This approach is called NIPALS (Wold, 1966), for *N*onlinear *I*terative *PA*rtial *L*east *S*quares, or ALS (De Leeuw, Young, and Takane, 1976), for *A*lternating *L*east *S*quares, and is strongly related to the *Gauss-Seidel* and *block decomposition* (or *relaxation*) methods, which are well-known in numerical analysis for iteratively solving linear systems (e.g., Burden and Faires, 1985).

There are at least two important situations in which another general strategy, called *iterative majorization* (IM), may be useful for constructing convergent algorithms, as this paper tries to demonstrate. Iterative majorization also involves solving a connected series of subproblems, but these are not defined by partitioning the parameter space into subspaces, but by covering the objective function surface with a series of simple (quadratic or linear) auxiliary functions. In the first kind of situation to be considered, the optimization function is of at most quadratic curvature, and its domain may be defined by some particular structure of constraints. The sole purpose of the IM strategy here is to generate a convergent sequence of simpler subproblems that can be solved by standard methods.

The other kind of situation consists of *nonsmooth optimization problems*, i.e., problems that involve functions with discontinuous gradients, or, practically speaking, very rapidly changing

gradients. Problems of this type arise in the areas of multidimensional scaling and robust estimation. The purpose of iterative majorization now is – in addition to obtaining simplicity or flexibility – to resolve two difficulties that one is sure to encounter when applying standard methods of smooth optimization to functions that have discontinuous gradients. First, one has to define an analogue of the gradient at the points at which it does not exist. This difficulty can mostly be overcome rather easily, by taking limits at the points of nondifferentiability. These limits can be non-unique, but then it is frequently possible to define *gradient sets*, which are called *subdifferentials* in the special case of convex functions (Rockafellar, 1970). The second difficulty, however, is much harder; it is how to define reasonable search directions and steplengths. If we take the direction of the negative gradient in the region where it is defined, we may converge to a nonstationary point (Wolfe, 1975). If we take the direction negative to that of an arbitrary subgradient, it may turn out not to be a direction of descent (Shor, 1985). And if we fail to select the right steplength, the algorithm may start oscillating or stop prematurely, i.e., again at a nonstationary point (see Heiser, 1991, for an example of this phenomenon in the present area of application).

Iterative majorization provides a reliable mechanism for generating search directions, and a natural adjustment of steplength in a wide variety of practical problems. Cornerstones of the IM approach as described here were laid in a series of papers by De Leeuw and Heiser (De Leeuw and Heiser, 1977; De Leeuw, 1977; De Leeuw and Heiser, 1980) and a dissertation (Heiser, 1981). At first, the strategy was primarily motivated by nonsmoothness, but later it was extended in various directions when its potential for simplifying other problems had been better recognized. Effective use of IM frequently requires a thorough analysis of the structure of the optimization problem, to bring it into one of a number of standard forms. The first part of this paper therefore gives a short introduction to the type of data analysis problems that forms the background of this theory, followed in the second part by an attempt to systematization, stating the three basic principles of IM, and offering an ordered classification of standard functions.

The classification is based on *type of curvature* of (parts of) the objective function, ranging from various forms of convexity to various forms of concavity. To avoid switching back and forth between majorization (covering from above) and minorization (covering from below), all tasks are

formulated as minimization problems, even when their natural formulation is the reverse. Next, a number of general considerations are presented for bringing problems into standard form, and this part is concluded by a discussion of the conditions that ensure proper convergence, and of some acceleration schemes.

The third part of the paper is devoted to applications of IM in multidimensional data analysis, particularly in multiple regression with restrictions on the parameters, robust fitting of principal components, global optimization in multidimensional scaling, and maximum likelihood estimation with missing data. Apart from presenting examples of each of the standard functions from the hierarchy of curves, this part also demonstrates how different majorizations can be combined, how IM can be integrated with (NIP)ALS, and in what sense the EM algorithm is an IM algorithm. In most cases, the IM algorithms were developed because no alternatives seemed to be feasible, or as a justification for already existing procedures. Not much can be said in regard to comparison with competing strategies. However, there is a dilemma between tailor-made majorization versus general-purpose algorithms for large classes of objective functions, and a few remarks on this issue are made in the Discussion section.

2. Type of Data Analysis Problems Considered

Expanding upon the basic schemes of Kruskal and Carroll (1969) and Gifi (1990), it may be said that most tasks in multidimensional data analysis require the minimization of some badness-of-fit function (or other *objective function*), which frequently takes one of the three following forms:

$$\phi_A(\mathbf{X},\mathbf{Y}) = \sum_k \psi[\mathbf{H}_k - \Xi_k(\mathbf{X},\mathbf{Y})] \,, \tag{1}$$

$$\phi_H(\mathbf{X},\mathbf{Y}) = \sum_k \psi[\Xi_k(\mathbf{H},\mathbf{Y}) - \mathbf{X}], \tag{2}$$

$$\phi_B(\mathbf{X},\mathbf{Y}) = c - \psi[\mathbf{H},\mathbf{X},\mathbf{Y}] \,/\, \nu[\mathbf{H},\mathbf{X},\mathbf{Y}]. \tag{3}$$

Here $\mathbf{H}$ denotes a matrix of observed values, partitioned into submatrices $\mathbf{H}_k$, of order $n \times m_k$, with k running from 1 to K, while $\Xi_k(\cdot)$ denotes a matrix-valued function representing the model to be fitted, with unknown parameters $\mathbf{X}$ and $\mathbf{Y}$, $\psi[\cdot]$ and $\nu[\cdot]$ both denote some *volume-* or *size-*

function, and c is a constant. When badness-of-fit is measured by (1), we aggregate in $\phi_A(\cdot)$ the size of the residuals from the model predictions (*approximations*, or *reconstructions*) with respect to the data, i.e., the size of the *approximation errors*. The function $\phi_H(\cdot)$ in (2) sums the size of the deviations from some (parametric) function of the data with respect to the model quantities $\mathbf{X}$, i.e., it determines the dispersion around $\mathbf{X}$, or the *homogeneity* of the functions $\Xi_k(\cdot)$. Badness-of-fit, or other data analytic objectives, can frequently also be formulated with $\phi_B(\cdot)$ in (3), through the ratio of two size functions, which establishes a *balance* between two opposing objectives.

There are cases where it is desired to replace $\mathbf{H}$ by a column-wise transformed data matrix $\mathbf{Q}$ – containing *quantifications*; a process also called *optimal scaling* of the observations – to further improve the fit of the model (see Heiser and Meulman, 1994a, 1994b). Then $\phi_A(\cdot)$, $\phi_H(\cdot)$, and $\phi_B(\cdot)$ become a function of these quantifications too. The three types of objective function certainly do not exhaust the possibilities, but they are general enough for the present discussion.

It cannot be stressed enough that – in contrast to specific, well-defined *tasks* – the major *methods* of multidimensional data analysis are not characterized by one single badness-of-fit or objective function, indeed not even by a single type of function. In the examples to be discussed below, *Principal components analysis* (PCA) will be used as the prime example of a method with multiple objectives that happen to coincide under the standard choices of $\Xi(\cdot)$, $\psi[\cdot]$, and $\nu[\cdot]$.

Approximation Problems

When stated in *Eckart-Young form*, a name acknowledging the contribution of Eckart and Young (1936), PCA is of type (1): an approximation problem with $K = 1$, with model function $\Xi(\mathbf{X},\mathbf{Y}) = \mathbf{XY}'$, called the *bilinear* model, and usually with $\psi[\cdot]$ specified as the *squared* Euclidean norm, defined on the residuals $\mathbf{R} = \mathbf{H} - \mathbf{XY}'$ as

$$\psi[\mathbf{R}] = \|\mathbf{R}\|^2 = \text{tr}\,(\mathbf{R}'\mathbf{R})\,. \tag{4}$$

The squared Euclidean norm is by far the most commonly used size function, so it will be tacitly assumed throughout the paper, except for the section on resistant fitting of principal components. With these specifications, the Eckart-Young form of PCA looks like

$$\phi_A(\mathbf{X},\mathbf{Y}) = \|\mathbf{H} - \mathbf{X}\mathbf{Y}'\|^2 , \qquad (5)$$

where $\mathbf{X}$ is an $n \times p$ matrix of unknowns, and $\mathbf{Y}$ an $m \times p$ matrix of unknowns. Note that the product $\mathbf{X}\mathbf{Y}'$ will have rank $p \leq \min(n,m)$, with p typically low. The low-rank approximation problem in (5) could be solved for each choice of the rank (or *dimensionality*) p separately, but principal components have an additional property, called *nestedness*, which implies that the first p columns of a $(p+1)$-dimensional solution are equal to the p-dimensional solution, for $1 \leq p \leq m-1$. Thus Eckart-Young PCA fits a bilinear model with nested components.

Multidimensional scaling (MDS), including *Individual Differences Scaling* (IDS), is a group of methods that is most frequently cast into form (1), with $\mathbf{H}_k = \Delta_k$, a matrix of size $n \times n$, containing measures of proximity or dissimilarity, and with the functions $\Xi_k(\cdot) = D(\mathbf{X}_k)$, or, alternatively, $\Xi_k(\cdot) = D^2(\mathbf{X}_k)$ or $\Xi_k(\cdot) = \log D(\mathbf{X}_k)$, where $D(\cdot)$ denotes the matrix of Euclidean distances among the rows of its argument, and $D^2(\cdot)$ the matrix of squared Euclidean distances. Thus the class of least squares Euclidean MDS problems can be characterized by the approximation criterion

$$\phi_A(\mathbf{X}_1,\ldots,\mathbf{X}_K) = \Sigma_k \|\Delta_k - D(\mathbf{X}_k)\|^2 , \qquad (6)$$

which is to be minimized with constraints on the $n \times p$ matrices $\mathbf{X}_k$ of the form $\mathbf{X}_k = \mathbf{X}$, or, for IDS, $\mathbf{X}_k = \mathbf{X}\mathbf{Y}_k$, with $\mathbf{X}$ and $\mathbf{Y}_k$ in turn restricted in various ways (Bloxom, 1978). A well-known case is obtained when $\mathbf{Y}_k$ has to be diagonal, leading to the INDSCAL model (Carroll and Chang, 1970), but one could also use low-rank restrictions on $\mathbf{Y}_k$ (Heiser and Stoop, 1986; Heiser and Meulman, 1989), and there is a great variety of interesting constraints on $\mathbf{X}$ (De Leeuw and Heiser, 1980; Heiser and Meulman, 1983; Meulman and Heiser, 1984). In Meulman's (1986, 1992) distance approach to multivariate analysis, the class of MDS/IDS problems is extended by letting Δ_k be of the form $D(\mathbf{Q}_k)$, where the columns of the $\mathbf{Q}_k$ are constrained to be transformations of given sets of variables.

As another example of an approximation task – although it is usually not phrased this way –, consider a p-group clustering method based on

$$\phi_A(\mathbf{G},\mathbf{Y}) = \|\mathbf{H} - \mathbf{G}\mathbf{Y}\|^2 , \qquad (7)$$

where $\mathbf{G}$ is an unknown $n \times p$ indicator matrix, i.e., an orthogonal binary matrix indicating to which cluster each of the row objects of $\mathbf{H}$ belongs, and $\mathbf{Y}$ a $p \times m$ matrix of unknowns, containing the cluster centers. Depending upon the allocation strategy chosen, one obtains one of the K-means clustering methods (*cf.* Selim and Ismael, 1984).

Homogeneity Problems

Problems of type (2) are called homogeneity problems, because they are aimed at minimizing the dispersion of a number of functions defined on the row objects of $\mathbf{H}$ around some unknown comparison function $\mathbf{X}$, defined on the same (discrete) domain. It is possible to formulate PCA as the minimum of the homogeneity criterion $\phi_H(\cdot)$ in (2), by letting k index single column vectors, rather than matrices, and by specifying $\Xi_k(\cdot) = y_k\mathbf{h}_k$ and $\mathbf{x}'\mathbf{x} = 1$, so that we obtain

$$\phi_H(\mathbf{x},\mathbf{y}) = \sum_k \|y_k\mathbf{h}_k - \mathbf{x}\|^2 , \tag{8}$$

for an analysis with one principal component $\mathbf{x}$. This formulation could be called PCA in *Pearson form*, and it is based on the idea of forming linear combinations of the given column vectors that are representative for the collection as a whole. There are various ways to extend (8) for an analysis with more than one principal component (Gifi, 1990, chapter 3).

The Gifi system of multivariate analysis (De Leeuw, 1984; Gifi, 1990; Heiser and Meulman, 1994b) can technically be viewed as an extension of PCA in Pearson form with the possibility to optimize over specified classes of transformations of the columns of $\mathbf{H}$, allowing for various constraints on the unknown $\mathbf{X}$ and $\mathbf{Y}$. It is based on the homogeneity loss function

$$\phi_H(\mathbf{X},\mathbf{Y}) = \sum_k \|\mathbf{G}_k\mathbf{Y}_k - \mathbf{X}\|^2 , \tag{9}$$

where the components are collected in the columns of $\mathbf{X}$, and where the $\mathbf{Y}_k$ matrices contain the coefficients of the transformations. The $n \times m_k$ matrices $\mathbf{G}_k$ are *known* indicator matrices, or, more generally, known matrices containing basis functions that span the space of transformations of $\mathbf{h}_k$. A least absolute deviations version of the homogeneity function in (9), with $\|\cdot\|$ instead of $\|\cdot\|^2$, has been studied in Heiser (1987a).

When the matrices $\mathbf{H}_k$ in (2) are $n \times n$ dissimilarity matrices Δ_k that are to be grouped in p groups

of unknown composition (called *points-of-view*), using coefficients $\mathbf{Y} = \{y_{ka}\}$ for the kth matrix in the ath group, and when the grouped dissimilarity matrices are approximated, as in (6), by low-dimensional Euclidean distance matrices $D(\mathbf{X}_a)$, then the homogeneity criterion $\phi_H(\cdot)$ becomes equal to the badness-of-fit function used in Meulman and Verboon's (1993) generalization of Tucker and Messick's (1965) points-of-view analysis, which is based upon

$$\phi_H(\mathbf{X},\mathbf{Y}) = \Sigma_k \Sigma_a \, \| y_{ka} \Delta_k - D(\mathbf{X}_a) \|^2 \, . \tag{10}$$

In Meulman and Verboon's (1993) method, each row of $\mathbf{Y}$ is constrained to have $p - 1$ elements equal to zero, ensuring that each Δ_k belongs to one and only one point-of-view, but other prescribed patterns of zeros (at known or unknown positions) could be imposed as well.

Balance Problems

Badness-of-fit function $\phi_B(\cdot)$ in (3) is the natural choice when a compromise between two desiderata is to be achieved, by balancing two objectives against each other, or when – more radically – there is one major objective, but certain classes of solutions are to be excluded from occurring. A prime example is again PCA, formulated as a procedure to find linear combinations of the columns of $\mathbf{H}$ with maximal dispersion. Assuming that the columns of $\mathbf{H}$ are centered, and that the measure of dispersion chosen is the variance, this goal implies maximization of $\mathbf{y}'\mathbf{H}'\mathbf{H}\mathbf{y}$. So $\psi[\cdot]$ in (3) is specified as the squared Euclidean norm; however, in contrast to (1) and (2), not on the residuals, but on the linear combinations $\mathbf{H}\mathbf{y}$ themselves. Because a uniform expansion of $\mathbf{y}$ would trivially increase the variance, some normalization function is needed that also changes quadratically if $\mathbf{y}$ is rescaled, for example $\nu[\cdot] = \mathbf{y}'\mathbf{y}$, and from these specifications, one obtains the balance function

$$\phi_B(\mathbf{y}) = c - \mathbf{y}'\mathbf{H}'\mathbf{H}\mathbf{y} \; / \; \mathbf{y}'\mathbf{y} \, . \tag{11}$$

It is possible to show that the $\mathbf{y}$ minimizing $\phi_B(\mathbf{y})$ is equal to the solution of (5) and (8), up to a rescaling, where it should be remarked that additional specifications are needed again to define the remaining principal components. There are no unknown scores $\mathbf{x}$ in (11), but once some optimal $\mathbf{y}^*$ is determined, it is always possible to proceed with $\mathbf{x}^* = \mathbf{H}\mathbf{y}^*$. The formulation of principal components as the solution to a balance problem will be called *PCA in Kruskal form*; the terminology

is not chosen here for historical reasons, but credits the fact that Kruskal (1969, 1972) initiated the more general use of balance functions $\phi_B(\cdot)$, an approach which later became known by the name *projection pursuit* (a term coined by Kruskal).

Successive linear combinations $\mathbf{Y}$ project the observations onto a subpace of $\mathbf{H}$, a feature from which projection pursuit methods derive their name. These exploratory methods (Friedman and Tukey, 1974; Friedman and Stuetzle, 1981; Huber, 1985; Jones and Sibson, 1987) seek subspaces that optimize a combination of characteristics, often starting from robustified measures of dispersion (*e.g.*, the *trimmed* variance). In the present framework, $\psi[\cdot]$ could be the trimmed variance, while the additional characteristics can be incorporated in the function $\nu[\cdot]$; they usually involve some type of *clustering* of the projected observations, or – more generally – other projections that best show some unusual distribution of points.

The central technique for *sets of matrices* is *canonical correlation analysis*. Here we have a balance function, in the simplest case, with $\psi[\mathbf{y}_1,\mathbf{y}_2] = \mathbf{y}_1'\mathbf{H}_1'\mathbf{H}_2\mathbf{y}_2$ and $\nu[\mathbf{y}_1,\mathbf{y}_2] = \|\mathbf{H}_1\mathbf{y}_1\|\,\|\mathbf{H}_2\mathbf{y}_2\|$, so that the ratio $\psi[\mathbf{y}_1,\mathbf{y}_2] / \nu[\mathbf{y}_1,\mathbf{y}_2]$ is a correlation between linear combinations of the columns of $\mathbf{H}_1$ and $\mathbf{H}_2$, respectively (Hotelling, 1936). There are various ways to generalize canonical correlation analysis. For a review of this rather large, but not very coherent area, the reader is referred to Gifi (1990, chapters 5–7). A number of interesting balance problems for sets of matrices, involving adjustments for keeping the linear combinations representative for the data vectors in their own space, are discussed in Nierop (1993).

A final example of a balance problem is the tunneling method for obtaining a series of local minima with decreasing badness-of-fit (Groenen and Heiser, 1991; Groenen, 1992, 1993). Since the balance function for tunneling takes a rather complicated form, description of this method is postponed until the Applications section. After this concise overview of the way in which many multidimensional data analysis methods lead to three classes of optimization problems, the stage has been set for introducing iterative majorization as a general optimization strategy.

3. Theory of Iterative Majorization (IM)

Due to the variety of specifications for $\Xi(\cdot)$, $\psi[\cdot]$, and $\nu[\cdot]$, the function surfaces of the badness-of-fit functions $\phi_A(\cdot)$, $\phi_H(\cdot)$, and $\phi_B(\cdot)$ are generally not of one single shape, nor even of uniform curvature. However, for easy and reliable computation it is often necessary to take the specifics of shape into account. A primary consideration in the iterative majorization (IM) strategy of algorithm construction is to *isolate the problematic aspect* of the function surface. Such an analysis generally leads to a decomposition of the objective function into parts that are either simple enough to be optimized by routine methods, or are to be majorized by such a simple function. Thus the first principle of IM is *to split the function – not the domain.* Splitting the domain of the objective function, the first principle of (NIP)ALS, can consequently be combined with any form of IM. The most common decomposition is additive, with a convex part and a concave part.

The second principle of IM is *to construct a family of auxiliary functions* that majorizes the objective function, in such a way that (1) for each feasible point in the domain there exists a member of the family, which coincides with the objective function at that point, and (2) each auxiliary function has a unique minimum. Next, an iterative computational scheme is defined as a sequence of minimization problems in terms of the members of the family of auxiliary functions. Then the third, and last, principle of IM is *to select the sequence of minimization problems so that convergence is guaranteed.* We will start our discussion with the standard IM sequence, including a basic convergence proof, next elaborate on the different types of functions to be majorized, and return to variations on the third principle in the section on convergence and acceleration.

Model Algorithm

What does it mean for a function to be majorized by a family, and how does one guarantee a convergent sequence of objective function values by repeated majorization? Starting, for simplicity of exposition, with a minimization problem of one variable, let $\phi(x)$ denote the objective function (or the non-simple part of it), let $\mu(x \mid \underline{x})$ denote what is called the *majorizing function*, and suppose that x and $\underline{x}$ both vary in the same domain Ω. Note that the majorizing function $\mu(x \mid \underline{x})$ is a function of *two* arguments, x and $\underline{x}$, which is how we deal with a *parametric family* of functions of x; here $\underline{x}$ is the

parameter indexing the members of the family. The second IM principle leads to the requirements

$$\phi(x) \leq \mu(x \mid \underline{x}), \tag{12}$$

$$\phi(\underline{x}) = \mu(\underline{x} \mid \underline{x}), \tag{13}$$

for all $x \in \Omega$ and $\underline{x} \in \Omega$. Thus the function $\mu(x \mid \underline{x})$ majorizes $\phi(x)$ if (12) holds, and equation (13) specifies the existence of a point where the two functions coincide. Depending upon the conditions under which equality in (12) holds, we may distinguish two cases: *strong* majorization, in which equality is reached if and only if $x = \underline{x}$, and *weak* majorization, in which equality is not necessarily reached in $\underline{x}$ only. The most common way to effectuate the third IM principle is to choose the majorizing function so that its minimum over the domain Ω exists and is unique, and to define, for any $\underline{x}$, the mapping

$$\overline{x} = \underset{x \in \Omega}{\operatorname{argmin}} \; \mu(x \mid \underline{x}). \tag{14}$$

Then it is possible to build up a standard sequence of mappings according to the following *IM model algorithm*:

(1) $\underline{x} \Leftarrow x_0$ for some initial guess $x_0 \in \Omega$;

(2) while $\overline{x} \neq \underline{x}$, do $\underline{x} \Leftarrow \overline{x}$;

(3) stop when $\overline{x} = \underline{x}$.

Of course, the equality checks must be understood as approximate equalities, in relation to the desired accuracy and the computer precision. As illustrated in Figure 1, $\overline{x}$, called the *successor point*,

Insert Figure 1 about here

The function $\mu(x \mid \underline{x})$ majorizes $\phi(x)$ at the supporting point $\underline{x}$.

minimizes the majorizing function over the domain Ω with respect to $\underline{x}$, called the *supporting point*, and $\overline{x}$ in turn serves to select the new auxiliary function $\mu(x \mid \underline{x} \Leftarrow \overline{x})$. This sequence leads to convergence, because it continually satisfies the basic IM property

$$\phi(\overline{x}) \leq \mu(\overline{x} \mid \underline{x}) = \min_{x \in \Omega} \mu(x \mid \underline{x}) \leq \mu(\underline{x} \mid \underline{x}) = \phi(\underline{x}). \tag{15}$$

Here the last equality follows from (13), the first one from (14). The first inequality in (15) follows from (12), and as long as the second inequality is strict, we continue the process. When the second inequality becomes an equality, we must have $\overline{x} = \underline{x}$, because of the uniqueness of the minimum of $\mu(x \mid \underline{x})$, and we stop the process. If the process is stopped, the vanishing derivatives of $\mu(x \mid \underline{x})$ at $\overline{x}$ imply that the derivatives of $\phi(x)$, if they exist at $\overline{x}$, vanish too. This property at the point of termination holds, because the second IM principle and continuity imply that the minorant and the majorizer have the same derivatives at any supporting point.

Illustrative Example: Iterative Calculation of the Median

Let us now have a look at a simple example of IM, applied to the *calculation of the median*. Given the n observations h_i, $i = 1, ..., n$, the median corresponds to the minimizing argument of

$$\phi(x) = \sum_i |h_i - x| \,. \tag{16}$$

This objective function is piecewise linear and convex (see Figure 2), and the (mildly) problematic

Insert Figure 2 about here

Piecewise linear function for determining the least total absolute deviaton from the observations {1, 3, 4, 7, 9, 13, 18}.

aspect is the discontinuity of its derivative. Heiser (1987a) proposed the family of quadratic functions

$$\mu(x \mid \underline{x}) = 1/2\ \phi(\underline{x}) + 1/2 \sum_i |h_i - \underline{x}|^{-1}\ (h_i - x)^2. \tag{17}$$

Using the inequality $(|h_i - x| - |h_i - \underline{x}|)^2 \geq 0$, it may be verified that the expression in (17) indeed satisfies requirements (12) and (13) for a proper majorizing function. Starting with any initial estimate $\underline{x} \Leftarrow x_0 \neq h_i$, we have to minimize the second part at the right-hand side of (17), i.e. we have to compute a weighted mean with weights proportional to $|h_i - \underline{x}|^{-1}$. According to (15), an improved value $\phi(\overline{x})$ is obtained, and computation of *iteratively reweighted means* may be continued until $\overline{x} = \underline{x}$. At this point, $\underline{x}$ satisfies the stationary equation for the least squares function (17)

$$\Sigma_i (h_i - \underline{x}) / |h_i - \underline{x}| = \Sigma_i \text{ sign } (h_i - \underline{x}) = 0, \tag{18}$$

which is identical to the stationary equation for (16); hence, the process stops at a desirable point. Note that (18) implies that when n is odd, $\overline{x}$ should be chosen as the element $h_{(n/2)}$, where $h_{(i)}$ denotes the ith of the ordered observations, and when n is even, that any value in the interval between $h_{(n/2)}$ and $h_{((n+2)/2)}$ is equivalent, because it always yields two residuals with opposite sign. If we would actually try to compute the median iteratively in practice, (17) shows that we would have to take care to avoid actual division by zero, which can easily be done through appropriate bounding of the weights.

Majorization of Standard Functions, Arranged by Curvature

According to the first IM principle, an objective function must be decomposed into components that are either simple enough to remain intact, or are candidates for majorization. Additive decomposition is a prime possibility, because majorization still holds under addition; for three functions f_1, f_2, and f_3, we have: if $f_1 \leq f_2$, then $f_1 + f_3 \leq f_2 + f_3$. As a general rule, linear and quadratic functions are to be regarded as simple, but there are circumstances where it is useful to majorize them. The situation can be structured by considering various standard functions in terms of their curvature, and arranging them in a hierarchy from *positively curved* (convex) functions, via linear functions, to *negatively curved* (concave) functions. This ordered classification in terms of curvature will be discussed first, with suggestions for majorization, and then the next section gives general combination rules to majorize functions built from components that have one of these standard forms. The scheme proposed here is not the only possible one, but it has proven to be rather valuable in actual work, and it has the virtue of simplicity.

(1) General quadratic functions with positive curvature. A quadratic function has positive curvature when its second derivative is positive. For a function of one variable, we have

$$\phi(x) = a\, x^2 + b\, x + c, \tag{19}$$

with $a \geq 0$. In this case, it is not difficult to verify that the family

$$\mu(x \mid \underline{x}) = \beta x^2 + [b + 2\,(a - \beta)\,\underline{x}]\,x + [c - (a - \beta)\,\underline{x}^2] \tag{20}$$

has the desired properties of a majorizing function, (12) and (13), provided that the coefficient β is chosen so that $\beta > a$. If $\beta = a$, then (20) becomes identical to (19). Thus, a quadratic function of one variable can be majorized by a family of steeper quadratics. Majorization becomes a lot more relevant when we have a function of m variables, of the form

$$\phi(\mathbf{x}) = \mathbf{x}'\mathbf{A}\mathbf{x} + \mathbf{b}'\mathbf{x} + c. \tag{21}$$

Now the function has positive curvature in *all* directions when its matrix of quadratic coefficients $\mathbf{A}$ is *positive definite*, i.e. as long as $\mathbf{x}'\mathbf{A}\mathbf{x} > 0$ for all $\mathbf{x} \in R^m$. As we shall see shortly, it may be desirable to work with a majorizing function in which $\mathbf{A}$ is replaced by a matrix of more simple form, *e.g.* the identity matrix, a (tri-)diagonal matrix, or a block-diagonal matrix. In all such cases, $\phi(\mathbf{x})$ can be majorized by

$$\mu(\mathbf{x} \mid \underline{\mathbf{x}}) = \mathbf{x}'\mathbf{P}\mathbf{x} + [\mathbf{b} + 2\,\mathbf{N}\underline{\mathbf{x}}]'\mathbf{x} + [c - \underline{\mathbf{x}}'\mathbf{N}\,\underline{\mathbf{x}}], \tag{22}$$

where $\mathbf{A}$ has been decomposed as $\mathbf{A} = \mathbf{P} + \mathbf{N}$, with $\mathbf{P}$ some positive definite matrix of simpler form than $\mathbf{A}$ itself, and $\mathbf{N}$ some negative definite matrix. The fact that $\mathbf{N}$ be negative definite is vital for the proof that $\phi(\mathbf{x}) \leq \mu(\mathbf{x} \mid \underline{\mathbf{x}})$, which is based upon the elementary inequality

$$(\mathbf{x} - \underline{\mathbf{x}})'\mathbf{N}(\mathbf{x} - \underline{\mathbf{x}}) \leq 0. \tag{23}$$

Substituting the decomposition $\mathbf{A} = \mathbf{P} + \mathbf{N}$ into (21) and using (23) yields the desired result.

Of course, there are many ways to decompose $\mathbf{A}$ into a positive definite and a negative definite component, so it is often useful to construct a parametric class of matrices in the role of $\mathbf{P}$, and then to derive an interval for the parameter which ensures negative definiteness of $\mathbf{N} = \mathbf{A} - \mathbf{P}$. For instance, suppose $\mathbf{P} = \beta\,\mathbf{I}$; then $\mathbf{A} - \mathbf{P}$ is negative definite whenever we have, for all $\mathbf{x} \in R^m$

$$\mathbf{x}'\mathbf{A}\mathbf{x} \leq \beta\,\mathbf{x}'\mathbf{x}, \tag{24}$$

which is true for all choices of β exceeding the largest eigenvalue of $\mathbf{A}$, because the latter quantity maximizes the Rayleigh quotient $\mathbf{x}'\mathbf{A}\mathbf{x} / \mathbf{x}'\mathbf{x}$. Condition (24) – derived from different arguments –

was first used in this context by De Leeuw and Bijleveld (1988) and Heiser (1987a). Note that we cannot use (24) directly for majorizing $\phi(\mathbf{x})$ in (21), because IM requires a function of $\mathbf{x}$ *and* $\underline{\mathbf{x}}$, where $\underline{\mathbf{x}}$ serves as the supporting point. The idea to derive the majorizing function in (22) from the decomposition $\mathbf{A} = \mathbf{P} + \mathbf{N}$ and inequality (23) is new; it provides a more general way to proceed in similar situations, and naturally follows from the first IM principle.

Summarizing, a positive definite quadratic in m variables can be majorized by simpler quadratics – often, but not necessarily, by functions with circular or (hyper)spherical contour lines.

(2) Piecewise linear convex functions. Convex functions frequently arise as the result of taking a positive sum of elementary convex functions, or taking the maximum over a set of simpler (e.g., linear) functions. The basic version is the absolute value function

$$\phi(x) = |ax| = \max_{\phi_1\ \phi_2} \{\phi_1(x) = ax,\ \phi_2(x) = -ax\}, \tag{25}$$

with $a \geq 0$, and the simplest way to majorize it, a result already used in the earlier illustrative example of the median, is by means of the quadratic family

$$\mu(x \mid \underline{x}) = 1/2\ a\ |\underline{x}| + 1/2\ a\ |\underline{x}|^{-1}\ x^2\ , \tag{26}$$

which can be derived from the inequality $(|x| - |\underline{x}|)^2 \geq 0$. Note that the sum of absolute value functions is piecewise linear, whereas the sum of quadratic functions is again quadratic. In (26), we have tacitly assumed that $|\underline{x}| \neq 0$; it depends on the context of application how to handle the situation when $|\underline{x}| = 0$ does occur. One possibility is to switch to the Huber function (47) with very small tuning constant, a function which will be discussed in the Applications section.

(3) Linear functions. In our classification the natural successor of a piecewise linear function arises when there is only one linear piece, in which case the function is both convex and concave, and has zero curvature. For $\phi(x) = b\ x + c$ we can use the majorizing quadratic family

$$\mu(x \mid \underline{x}) = \beta\ (x^2 - 2\underline{x}\ x + \underline{x}^2) + (b\ x + c)\ , \tag{27}$$

for any value of $\beta > 0$. Note that (27) is a special case of (20) with $a = 0$. It is not immediately obvious why IM with quadratics – without further modifications – would have much to offer in the plain linear programming case, since linear programming is such a well-developed area, but methods similar to IM can be useful for particular constrained problems (Shor, 1985, chapter 4). Also, (27) will reappear in (53), where Newton's method is reinterpreted in terms of majorization.

(4) *Concave functions of a positive argument: the rth root and the log.* In this section it is assumed that $x \geq 0$ and $\underline{x} \geq 0$. For the rth root, we may derive a majorization function from the inequality

$$x^q \leq q\, \underline{x}^{q-1}\, x + (1 - q)\, \underline{x}^q, \tag{28}$$

which is particular way to write the arithmetic-geometric mean inequality for merely two terms, with parametrization $q = 1/r$ for all $r > 1$ so that $0 < q < 1$ (Hardy, Littlewood and Pólya, 1952, chapter II). Equality in (28) is obtained when $x = \underline{x}$, as it should in a proper majorization function. Note that the rth root $x^{1/r}$ is a concave function of x for $r > 1$, because its second derivative is negative, so the simplest majorization family of the rth root is linear. Combining (28) with the inequality that was used to derive (26), another family of majorizing functions can be found that is positively quadratic in x. These results were first used for IM in Groenen and Heiser (1991).

Logarithmic functions are also concave in the positive reals, and it is a well-known fact that they can be majorized by linear functions; such a majorizing family may be derived by determining the coefficients of a general linear function $f(x) = a\, x + b$ according to basic IM principles. The minorant and the majorizer should have the same derivatives at any supporting point $\underline{x}$, so $a = 1 \,/\, \underline{x}$, and they should be equal at $x = \underline{x}$, so $b = \log \underline{x} - 1$. This way we obtain

$$\log x \leq \underline{x}^{-1}\, x + \log \underline{x} - 1, \tag{29}$$

with equality if and only if $x = \underline{x}$, a result valid for all positive x and $\underline{x}$. More generally, concave functions can always be majorized (Hardy, Littlewood and Pólya, 1952, p. 94-96) by a linear function of the form $\mu(x \mid \underline{x}) = \phi(\underline{x}) + \alpha(x - \underline{x})$, with α some suitably chosen constant, and we see that $\alpha = \underline{x}^{-1}$ in (29).

(5) Concave functions: the negative of the Euclidean norm. First consider the simplest case of a concave piecewise linear function of one variable, which is the negative of (25), $\phi(x) = -|ax|$. It has two linear pieces, $\phi_1(x) = ax$ for $x < 0$ and $\phi_2(x) = -ax$ for $x > 0$, which coincide at $x = 0$. Therefore, it can be weakly majorized by the linear family $\mu(x \mid \underline{x}) = -a \text{ sign}(\underline{x})\, x$, where $\text{sign}(\underline{x})$ keeps track of which side of the line we are. Now, the negative of Euclidean norm is a function of m variables, defined as

$$\phi(\mathbf{x}) = -\|\mathbf{x}\| \,, \tag{30}$$

which has the homogeneity property $\phi(a\mathbf{x}) = a\phi(\mathbf{x})$. Due to this property, its contour lines – which are concentric and circular in the case of two variables – are *equally spaced*, so that linear majorization is again possible. In particular, for $\phi(\cdot)$ in (30) we have the majorizer

$$\mu(\mathbf{x} \mid \underline{\mathbf{x}}) = -\underline{\mathbf{x}}'\mathbf{x} \,/\, \|\underline{\mathbf{x}}\| \,. \tag{31}$$

Thus $\mu(\mathbf{x} \mid \underline{\mathbf{x}})$ is linear in $\mathbf{x}$, where the coefficients are $-\underline{\mathbf{x}} \,/\, \|\underline{\mathbf{x}}\|$, the unit length vector that indicates to which side of the plane (in general, in which direction of the R^m) the supporting point extends. The fact that (31) majorizes (30) is an immediate consequence of the well-known *Cauchy-Schwarz* inequality $\|\mathbf{x}\|\,\|\underline{\mathbf{x}}\| \geq \underline{\mathbf{x}}'\mathbf{x}$ (for the latter, see Hardy, Littlewood and Pólya, 1952, p. 16).

(6) Concave functions: quadratics with negative curvature. A quadratic function $\phi(\mathbf{x}) = \mathbf{x}'\mathbf{N}\mathbf{x}$ has negative curvature (and is concave) when its matrix of coefficients $\mathbf{N}$ is *negative definite*, i.e. when it satisfies $\mathbf{x}'\mathbf{N}\mathbf{x} < 0$ for all $\mathbf{x}$. Majorization follows from the elementary inequality (23), which leads to

$$\mathbf{x}'\mathbf{N}\mathbf{x} = \phi(\mathbf{x}) \leq \mu(\mathbf{x} \mid \underline{\mathbf{x}}) = 2\,\underline{\mathbf{x}}'\mathbf{N}\mathbf{x} - \underline{\mathbf{x}}'\mathbf{N}\,\underline{\mathbf{x}}. \tag{32}$$

It is of some interest to compare the right-hand side of (32) with (31). The coefficients of the linear majorizer, $2\,\underline{\mathbf{x}}'\mathbf{N}$, are no longer unit normalized, but are expressed in the (negative) metric $\mathbf{N}$. This matrix specifies the exact shape of the function in various directions, while (the negative of) the Euclidean norm is *isotropic*: the same in all directions; i.e., the special case where the negative metric $\mathbf{N}$ equals $-\mathbf{I}$. There is also an additional intercept term $\underline{\mathbf{x}}'\mathbf{N}\,\underline{\mathbf{x}}$ in (32), which accounts for the fact that the contour lines of a quadratic function are not equally spaced; their spacing decelerates as a function

of the distance from the origin, here indicated by the position of the supporting point.

Functions with more strongly decelerating contour lines, such as fourth power concave functions, can in turn be majorized by negatively curved quadratics. Negatively curved functions with accelerating contour lines (or with changing acceleration, *e.g.*, $\phi(x) = \exp[-ax^2]$) are no longer concave, and cannot be majorized by a linear function, but convex quadratic majorizers can still be constructed in this case.

Combination and Decomposition Rules

Additive decomposition has already been mentioned as a prime strategy to isolate the special aspects of the objective function surface, and the reason is that majorization remains valid under additive combination. If $\mu_1(\cdot \mid \cdot)$ majorizes $\phi_1(\cdot)$, and $\mu_2(\cdot \mid \cdot)$ majorizes $\phi_2(\cdot)$, then

$$\alpha\phi_1(\cdot) + \beta\phi_2(\cdot) \leq \alpha\mu_1(\cdot \mid \cdot) + \beta\mu_2(\cdot \mid \cdot) , \tag{33}$$

for $\alpha \geq 0$ and $\beta \geq 0$. For nonnegative functions, we also have a multiplicative rule:

$$\phi_1(\cdot)\ \phi_2(\cdot) \leq \mu_1(\cdot \mid \cdot)\ \mu_2(\cdot \mid \cdot) . \tag{34}$$

where it is assumed that $\phi_1(\cdot) \geq 0$ and $\phi_2(\cdot) \geq 0$. Inequality (34) often leads to higher order functions on the majorizing side, and therefore it is very useful to also have a direct majorization rule for products (Groenen and Heiser, 1991):

$$\phi_1(x)\ \phi_2(x) \leq 1/2\ [\phi_2(\underline{x}) / \phi_1(\underline{x})]\ \phi_1^2(x) + 1/2\ [\phi_1(\underline{x}) / \phi_2(\underline{x})]\ \phi_2^2(x) , \tag{35}$$

and, for $\phi(x) > c > 0$, the division rule:

$$\phi^{-1}(x) \leq [c\phi^2(\underline{x})]^{-1}\ \{\phi^2(x) - [2\phi(\underline{x}) + c]\ \phi(x) + [\phi^2(\underline{x}) + 2c\phi(\underline{x})]\} . \tag{36}$$

Proofs of these assertions are left as an exercise for the reader.

Analysis of an objective function must lead to a decomposition into parts that remain unchanged and parts that can be majorized. Additive decomposition into a convex and a concave part is of particular interest, because concave functions can always be majorized linearly. The majority of IM applications use a decomposition of this form. Each term in the summation of MDS functions like (6)

and (10), for example, have a convex subterm $\phi_1(\mathbf{x}_i,\mathbf{x}_j) = \delta_{ij}^2 + \|\mathbf{x}_i - \mathbf{x}_j\|^2$, and a concave subterm $\phi_2(\mathbf{x}_i,\mathbf{x}_j) = -\delta_{ij}\|\mathbf{x}_i - \mathbf{x}_j\|$, and since convexity and concavity are preserved under addition, all least squares MDS functions can be decomposed into a convex part and a concave part. In this case, the term $\phi_2(\mathbf{x}_i,\mathbf{x}_j)$ also embodies the problematic aspect: a discontinuous gradient at $\mathbf{x}_i = \mathbf{x}_j$. Another example is the decomposition $\mathbf{x}'\mathbf{A}\mathbf{x} = \mathbf{x}'\mathbf{P}\mathbf{x} + \mathbf{x}'\mathbf{N}\mathbf{x}$ of a general quadratic form, with $\mathbf{P}$ positive definite and $\mathbf{N}$ negative definite, used in (22), which has many applications.

How to Ensure Convergence and How to Accelerate

The third IM principle forces us to make sure that convergence is guaranteed. This principle is extremely powerful in practice, and it has to be strictly maintained, because during the development of an algorithm lack of convergence, or even a single increase in the objective function, always infallibly diagnoses the presence of programming or implementation errors. Generally, there are three major points that may be checked at a theoretical level:

(a) Equality of $\phi(\cdot)$ and $\mu(\cdot \mid \cdot)$ at the supporting point (cf. equation (13));

(b) Boundedness (from below) of the majorizer $\mu(\cdot \mid \cdot)$ over its domain;

(c) Equivalence of the necessary and sufficient conditions for a minimum of $\phi(\cdot)$ and $\mu(\cdot \mid \cdot)$ *at the optimal point.*

Note that especially point (c) guards IM from convergence to a nonstationary point, a real danger for all methods that walk along the direction of steepest descent (the gradient). For example, minimization of the objective function $\phi(x) = x^2$ could be attempted with an iteration scheme based on the mapping $x_{t+1} \Leftarrow 1/2 + 1/2\, x_t$, where t indexes iterations. It may be verified that such a scheme is a gradient method, it converges, but it converges to 1, not to the optimal value of 0.

Whereas proper convergence is reassuring, slow convergence is a nuisance, and that is what we typically have in IM. Very often, however, there are good possibilities to speed the basic model algorithm up by staying slightly away from the minimal point of the majorizer. De Leeuw and Heiser (1980) showed that use of a double-step relaxed update for their MDS algorithm keeps the process convergent, but approximately *halves* the number of iterations without too much extra calculations, and it will be shown here how to construct such an update for *any* quadratic majorizing function.

Consider an adjustment of the second rule of the model algorithm, where instead of $\overline{x}$, we assign

the update $\mathring{x} = g(x_t, x_{t-1}, x_{t-2}, \ldots)$, in which $x_t = \underline{x}$, to the role of the next supporting point. From the proof of (15) it follows that any definition of the mapping $g(\cdot)$ is allowed, as long as we have

$$\mu(\mathring{x} \mid \underline{x}) \leq \mu(\underline{x} \mid \underline{x}) \,. \tag{37}$$

Various convergent acceleration schemes can be devised that satisfy this condition, which is independent of the objective function. Among the simple ones, the so-called *relaxed update* is a member of the class $\mathring{x} = (1 - \alpha)\underline{x} + \alpha\overline{x}$. For a general quadratic majorizing function of m variables,

$$\mu(\mathbf{x} \mid \underline{\mathbf{x}}) = \mathbf{x}'\mathbf{A}(\underline{\mathbf{x}})\mathbf{x} + \mathbf{b}(\underline{\mathbf{x}})'\mathbf{x} + c(\underline{\mathbf{x}}), \tag{38}$$

using the explicit notation $\mathbf{A}(\underline{\mathbf{x}})$, $\mathbf{b}(\underline{\mathbf{x}})$, $c(\underline{\mathbf{x}})$ to indicate the dependence of the coefficients on $\underline{\mathbf{x}}$, and assuming that $\mathbf{A}(\underline{\mathbf{x}})$ is positive semi-definite, the minimum is reached for $\overline{\mathbf{x}}$ satisfying the stationary equation $2\mathbf{A}(\underline{\mathbf{x}})\overline{\mathbf{x}} = -\mathbf{b}(\underline{\mathbf{x}})$. Inserting the latter equation and the definition of $\mathring{\mathbf{x}}$ into (38), and using the notation $\|\mathbf{x}\|^2_{\mathbf{M}} = \mathbf{x}'\mathbf{M}\mathbf{x}$ for the squared generalized Euclidean norm with metric $\mathbf{M}$, we obtain

$$\mu(\mathring{\mathbf{x}} \mid \underline{\mathbf{x}}) = \mu(\underline{\mathbf{x}} \mid \underline{\mathbf{x}}) + (\alpha^2 - 2\alpha)\, \|\overline{\mathbf{x}} - \underline{\mathbf{x}}\|^2_{\mathbf{A}(\underline{\mathbf{x}})} \,. \tag{39}$$

It follows from (39) that (37) holds for $0 \leq \alpha \leq 2$, but of course the choice $\alpha = 0$ should be avoided. When $\alpha = 1$, we find $\mathring{\mathbf{x}} = \overline{\mathbf{x}}$, and $\alpha = 2$ gives the double-step relaxed update $\mathring{\mathbf{x}} = \overline{\mathbf{x}} + (\overline{\mathbf{x}} - \underline{\mathbf{x}})$.

When $\mathbf{x}$ is constrained, we have to be a bit more careful, because the relaxed update may bring us outside the feasible region. In the quadratic case, it is useful to write condition (37) as

$$\|\mathring{\mathbf{x}} - \overline{\mathbf{x}}\|^2_{\mathbf{A}(\underline{\mathbf{x}})} \leq \|\overline{\mathbf{x}} - \underline{\mathbf{x}}\|^2_{\mathbf{A}(\underline{\mathbf{x}})} \,, \tag{40}$$

i.e., the update $\mathring{\mathbf{x}}$ should be located inside a hyperellipsoid, centered at the unconstrained minimizer $\overline{\mathbf{x}}$ and with maximal radius the distance between $\overline{\mathbf{x}}$ and the previous constrained update $\underline{\mathbf{x}}$ in the metric $\mathbf{A}(\underline{\mathbf{x}})$. How simple the consequences of (40) are, depends on the regularity of the domain Ω. Let the notation $P(\cdot)$ be used for the projection onto Ω; normally, we would use $P(\overline{\mathbf{x}})$, yielding the shortest distance of any $\mathbf{x} \in \Omega$ to $\overline{\mathbf{x}}$. Now, in the important special case where Ω is a subspace of R^m, it turns out that we could either project the unconstrained relaxed update onto the domain, obtaining $\mathring{\mathbf{x}} = P((1 - \alpha)\,\underline{\mathbf{x}} + \alpha\,\overline{\mathbf{x}})$, or stay in the domain and take $(1 - \alpha)\underline{\mathbf{x}} + \alpha P(\overline{\mathbf{x}})$, which equals $P(\overline{\mathbf{x}}) +$

$(\alpha - 1)[P(\overline{\mathbf{x}}) - \underline{\mathbf{x}}]$, and happens to coincide with $\mathring{\mathbf{x}}$. Regardless the route taken, the squared distance of $\mathring{\mathbf{x}}$ to $\overline{\mathbf{x}}$ in the metric $\mathbf{A}(\underline{\mathbf{x}})$ becomes

$$\|\mathring{\mathbf{x}} - \overline{\mathbf{x}}\|^2_{\mathbf{A}(\underline{\mathbf{x}})} = \|\overline{\mathbf{x}} - \underline{\mathbf{x}}\|^2_{\mathbf{A}(\underline{\mathbf{x}})} - \{1 - (\alpha - 1)^2\}\, \|P(\overline{\mathbf{x}}) - \underline{\mathbf{x}}\|^2_{\mathbf{A}(\underline{\mathbf{x}})}\,, \tag{41}$$

a result which shows that, for the interval $0 \leq \alpha \leq 2$, condition (40) is satisfied, and therefore the algorithm will remain convergent when using $\mathring{\mathbf{x}}$, and in particular also the double-step constrained update $P(\overline{\mathbf{x}}) + [P(\overline{\mathbf{x}}) - \underline{\mathbf{x}}]$. When Ω is not closed under addition and scalar multiplication, as it actually is in the case just discussed, it is safest to use $P(g(\cdot))$, where $g(\cdot)$ is a rule that brings us as far as possible away from previous points. The savings with respect to the basic model algorithm will almost always be worthwhile.

Sometimes it is not efficient to minimize the majorizing function completely in each iteration. Then another adjustment of the standard IM model algorithm is to modify (14) into a mapping that merely decreases $\mu(\mathbf{x} \mid \underline{\mathbf{x}})$, as in (37), for instance by a few ALS steps. It is important that the partial minimization is done by steps from a convergent algorithm, to ensure that the adjusted IM process does not stop prematurely.

4. Applications in Multidimensional Data Analysis

Elements of this theory of iterative majorization have been applied to quite a broad range of multidimensional data analysis problems, which cannot be reviewed exhaustively here. However, by taking examples from each of the three classes of the classification discussed earlier, hopefully a good impression is provided of the great flexibility and relative simplicity of the IM approach.

Some examples not discussed include: least squares fitting of three-way distance models (De Leeuw and Heiser, 1980; Heiser and Stoop, 1986; Meulman and Verboon, 1993), multidimensional unfolding (Heiser, 1981, 1987*b*; De Soete and Heiser, 1993), dissimilarity-driven nonlinear components analysis (Meulman and Heiser, 1984), optimal distance approximation in nonlinear multivariate analysis (Meulman, 1986, 1992), optimal rotation in factor analysis (Ten Berge, Knol, and Kiers, 1988), robust multidimensional scaling (Heiser, 1988), least squares fitting of longitudinal reduced rank regression models (Bijleveld and De Leeuw, 1989), optimal scaling of multivariate time series (Van Buuren, 1990), DEDICOM modelling of asymmetric matrices (Kiers, Ten Berge, Takane, and De Leeuw, 1990), maximum likelihood estimation of latent time budgets (De Leeuw, Van der Heijden and Verboon, 1990), weighted orthonormal Procrustes analysis (Koschat and Swayne, 1991), least squares fitting of additional points in a given configuration of points (Heiser, 1987*b*; Meulman and Heiser, 1993), non-Euclidean multidimensional scaling (De Leeuw, 1977; Heiser, 1989; Groenen, Mathar, and Heiser, 1992), low-rank approximation of optimally scaled scalar product matrices (Meulman, 1993; Kiers, 1993), and analysis of asymmetry by the slide-vector model (Zielman and Heiser, 1993).

The present discussion starts with a group of analysis problems that can be characterized as *regression with restrictions on the parameters*, which subsumes quite a number of different contexts (Meulman, 1986; Heiser, 1987*a*; Van der Lans, 1989, 1992; Verboon and Heiser, 1992; Verboon, 1993). A common feature of these situations is that IM can easily deal with squared distance functions in some non-isotropic metric, which may have additional structure, or could be arbitrary. Next, it is shown how the computational complications arising from the use of nonstandard size functions of the residuals, motivated by considerations of robustness, can be resolved by IM. The

analysis problem of interest here is *resistant fitting of principal components* (Verboon and Heiser, 1993), and the IM approach yields a rationale and convergence proof for Gabriel and Odoroff's (1984) iteratively reweighted least squares procedure. A third kind of application concerns *local minimum problems in multidimensional scaling* (Groenen, 1990, 1993; Groenen and Heiser, 1991). When searching for the global minimum, old local minima – and irrelevant new ones – are to be avoided, which can be done by introducing locally effective penalty terms in the loss function. Finally, it is shown how the EM algorithm (Dempster, Laird, and Rubin, 1977) can be understood as an IM algorithm.

Regression with Restrictions on the Parameters

Under the standard least squares assumptions, some applications of the linear regression model lead to the following approximation problem:

$$\min_{\mathbf{y}\in\Omega} \ \phi_A(\mathbf{y}) = \|\mathbf{h} - \mathbf{F}\mathbf{y}\|^2 , \tag{42}$$

where $\mathbf{h}$ is the (known) criterion variable, a vector of length n, $\mathbf{F}$ is the $n \times p$ matrix of (known) real-valued predictor or design variables (*factors*), and $\mathbf{y}$ the p-vector of unknown parameters. The restrictions are formulated in terms of a *feasible region* Ω, and could consist of the requirement (*a*) that some (or all) elements of $\mathbf{y}$ are non-negative or increasing, (*b*) that subsets of $\mathbf{y}$ are orthogonal, or (*c*) that $\mathbf{y}$ should satisfy row-wise and column-wise inequality restrictions if rearranged in matrix form (cf. Van der Lans, 1992, for more examples).

What makes (42) special is the fact that $\mathbf{y}$ is constrained as $\mathbf{y}\in\Omega$; yet it is still useful, theoretically, to refer to the *unconstrained minimizer* $\hat{\mathbf{y}}$, which is well-known to be a vector satisfying $\mathbf{F}'\mathbf{F}\hat{\mathbf{y}} = \mathbf{F}'\mathbf{h}$. The usual orthogonality of the residuals $\mathbf{h} - \mathbf{F}\hat{\mathbf{y}}$ with respect to any vector in the predictor space, such as $\mathbf{F}(\hat{\mathbf{y}} - \mathbf{y})$, gives the additive decomposition of $\phi_A(\mathbf{y})$ into two terms:

$$\phi_A(\mathbf{y}) = \|\mathbf{h} - \mathbf{F}\hat{\mathbf{y}}\|^2 + \|\hat{\mathbf{y}} - \mathbf{y}\|^2_{\mathbf{F}'\mathbf{F}} . \tag{43}$$

Since the first term of (43) does not involve $\mathbf{y}$, the restricted parameters can be found by minimizing the second term, i.e., by solving a *least distance problem* with respect to $\hat{\mathbf{y}}$ in the metric $\mathbf{F}'\mathbf{F}$.

Let $\phi_{F'F}(\mathbf{y})$ denote the second term of decomposition (43); this is a quadratic function with positive curvature of the form (21), with $\mathbf{A} = \mathbf{F'F}$, $\mathbf{b} = -2\,\mathbf{F'h}$, and $c = \mathbf{h'F}\hat{\mathbf{y}}$. So a majorization function of the form (22) can be used; if we split $\mathbf{F'F}$ into an isotropic positive definite and a complementary negative semidefinite part, i.e., $\mathbf{F'F} = \beta\mathbf{I} + (\mathbf{F'F} - \beta\mathbf{I})$, with β not smaller than the largest eigenvalue of $\mathbf{F'F}$, then it may be verified that the least distance function can be majorized as

$$\phi_{F'F}(\mathbf{y}) \leq \underline{c} + \beta \|\bar{\mathbf{y}} - \mathbf{y}\|^2 , \qquad \text{with} \quad \bar{\mathbf{y}} = \underline{\mathbf{y}} + \beta^{-1}\,\mathbf{F'}[\mathbf{h} - \mathbf{F}\underline{\mathbf{y}}] , \tag{44}$$

where $\underline{\mathbf{y}}$ is a previous estimate, and where several constants are collected in $\underline{c}$. The implication of (44) is, that we can solve a least distance problem in the metric $\mathbf{F'F}$ by repeatedly solving an unweighted problem in terms of $\bar{\mathbf{y}}$, by whatever standard method is available for doing the orthogonal projection onto Ω, i.e., for finding $P(\bar{\mathbf{y}})$. Notice that to calculate the successive improvements $\bar{\mathbf{y}}$, we only need to know the previous estimate $\underline{\mathbf{y}}$, and the fixed quantities $\mathbf{h}$ and $\mathbf{F}$, but not the unrestricted estimate $\hat{\mathbf{y}}$, as would seem to be necessary at first sight, in view of (43). In practice, direct calculation of $\hat{\mathbf{y}}$ is better avoided, because it involves the inversion of $\mathbf{F'F}$, which may be a large matrix, or an ill-conditioned one. For that matter, if $\Omega = R^p$, one might want to calculate $\hat{\mathbf{y}}$ with the above IM algorithm – just to avoid explicit inversion.

For the sake of completeness, it should be mentioned that a similar development is possible in least distance problems with some other positive (semi-)definite matrix $\mathbf{M}$ in the role of $\mathbf{F'F}$; for example, (*i*) a smoothing operator, or – conversely – a matrix that filters out trend, or (*ii*) an idempotent projection matrix that restricts all operations to a subspace (*cf* Meulman, 1986, section 4.7.3), or (*iii*) an estimate of $\mathbf{\Sigma}^{-1}$, when there are heterogeneous or correlated errors. The case of an idempotent matrix is particularly convenient, because idempotency implies that all eigenvalues are either zero or one, so the positive-definite part of the decomposition of $\mathbf{M}$ can be the identity.

Resistant Fitting in Principal Components Analysis

Robustness of a technique refers to its potential to give the same or similar results under substantial disturbances of the typical circumstances for which it had been designed. When some of the disturbances are even allowed to be excessive, the technique is called *resistant*. Early developments (Huber, 1981) almost exclusively dealt with the linear model, except for attempts to

robustify the exploration of multidimensional space with *projection pursuit* (Huber 1985). As we have seen in the typology of optimization problems, the objective function of projection pursuit qualifies as a balance function, and it provides only one way to define principal (or other) components. Here, components analysis will be discussed from the different perspectives of approximation and homogeneity.

The objective function of the Eckart-Young form of PCA is (5), a total sum of squared residuals, which can be split by variable as

$$\phi_A(\mathbf{X},\mathbf{Y}) = \sum_k \|\mathbf{r}_k\|^2 = \sum_k \|\mathbf{h}_k - \mathbf{X}\mathbf{y}_k\|^2 \ . \tag{45}$$

What do the variable-wise residuals $\mathbf{r}_k$ measure? They measure the prediction error that would occur when we would predict scores in $\mathbf{h}_k$ from the object scores in the rows $\mathbf{x}_i$ of $\mathbf{X}$, projected on direction $\mathbf{y}_k$, since in general we can write $\mathbf{x}_i'\mathbf{y}_k = a_k P_k(\mathbf{x}_i)$ with $P_k(\mathbf{x}_i)$ the projection of $\mathbf{x}_i$ on direction $\mathbf{y}_k$, and a_k some uniform scaling constant. Note that it is not the distance of a (multidimensional) score, or position, $\mathbf{x}_i$ towards the vector $\mathbf{y}_k$, but the displacement of $P_k(\mathbf{x}_i)$ with respect to $\mathbf{h}_k$ that is counted as error. So displacements are only measured in certain unknown directions, and an outlier in the residuals need *not* correspond to an outlying data point (the residual may even be small when an outlying data point is *also* an outlier in the whole of $\mathbf{X}\mathbf{y}_k$), but refers to a model point for which $P_k(\mathbf{x}_i)$ happens to *mismatch* with $\mathbf{h}_k$, and – by the same token – an outlier in the data need *not* obtain outlying object scores or outlying residuals.

The homogeneity version – or Pearson form – of principal components analysis is given in (8) for one component, and can be generalized to more components $a = 1,...,p$ as

$$\phi_H(\mathbf{X},\mathbf{Y}) = \sum_k \sum_a \|\mathbf{s}_{ka}\|^2 = \sum_k \sum_a \|y_{ka}\mathbf{h}_k - \mathbf{x}_a\|^2 \ . \tag{46}$$

Here, the residuals $\mathbf{s}_{ka}$ are defined with respect to each component $\mathbf{x}_a$ *separately*, and they measure the deviations from a *central value*: the estimated component scores for the individuals. Each row of $\mathbf{H}$ is reduced to p scores; the variables are differentially weighted by y_{ka} in this estimation process, and it is not the primary aim of the analysis to reproduce $\mathbf{h}_k$ – its role is to be taken by $\mathbf{x}_a$. In (46), an outlying data point will tend to be masked in the residuals through a small value of y_{ka}.

Although (45) and (46) can be shown to be equivalent under the standard assumptions, their robust versions will tend to become more clearly differentiated, because of these different roles of the parameter vectors and the residuals. Let us now look at two alternative size functions to replace $\|\cdot\|^2$.

(1) Huber function. To downweight the influence of the large residuals, Huber (1964) suggested to measure loss in a least squares fashion only when the residuals are small, and to use a least absolute deviations form when they are large. Thus, the Huber size function for the vector $\mathbf{r}$ with elements r_i is

$$\psi_{Hub}[\mathbf{r}] = 1/2 \sum_i [1/2\, r_i^2 + c\,|r_i| - 1/2\, c^2] - 1/2 \sum_i \text{sign}(|r_i| - c)\,[1/2\, r_i^2 - c\,|r_i| + 1/2\, c^2]\,, \qquad (47)$$

where c is a constant that marks the transition from the small to the large, called the *tuning constant.* It is clear from (47) that if $|r_i| < c$, then a fully quadratic term is added and the part that is linear in $|r_i|$ disappears, while if $|r_i| > c$, then the quadratic term cancels and the linear part remains. In the case of equality, the first term in the summation becomes equal to r_i^2. This construction will cause $\psi_{Hub}[\cdot]$ to become piecewise linear when c is chosen small enough (smaller than the smallest residual), while for c large enough, $\psi_{Hub}[\cdot]$ becomes equal to the least squares function.

A majorizer of the Huber function can be derived from the inequality $(|r_i| - |\underline{r}_i|)^2 \geq 0$, generalizing the way in which (17) majorizes the loss function of the median (Heiser, 1987a). Explicitly, we obtain

$$\mu_{Hub}(\mathbf{r} \mid \underline{\mathbf{r}}) = 1/2 \sum_i \underline{w}_i\, r_i^2 + 1/2 \sum_i [\underline{b}_i + \text{sign}(|\underline{r}_i| - c)\, \underline{b}_i]\,, \qquad (48)$$

where the intercept terms $\underline{b}_i = 1/2\, c\,|\underline{r}_i| - 1/2\, c^2$ are dependent on the previous residuals $\underline{r}_i$, as are the quantities $\underline{w}_i$, defined as $\underline{w}_i = 1$ for $|\underline{r}_i| < c$ and $\underline{w}_i = c \,/\, |\underline{r}_i|$ otherwise. Only the first term of (48) depends on r_i, and the $\underline{w}_i$ are weights in this weighted least squares loss function. These weights are never larger than one, and they (slowly) tend to zero when the previous residuals become large, relative to c.

(2) Tukey's biweight. More radical downweighting of the large residuals is possible by introducing an asymptote in the individual residual elements, i.e. a maximum value that bounds the influence of any single element. The classic example of such a size function, built up from locally supported terms with redescending derivatives, is Tukey's biweight function:

$$\psi_{Tuk}[\mathbf{r}] = c^2/12 \sum_i \{1 - \text{sign}(|r_i| - c)\}[1 - \{1 - (r_i/c)^2\}^3] + c^2/12 \sum_i \{1 + \text{sign}(|r_i| - c)\}\ , \qquad (49)$$

where residuals beyond c add a constant value of $c^2/6$ to the loss. By definition, the small residuals satisfy $\text{sign}(|r_i| - c) = -1$, and therefore contribute only to the first term of (49). To identify a basic characteristic of $\psi_{Tuk}[\cdot]$, note that this first term is built up from the quantities $t_i = \{1 - (r_i/c)^2\} \geq 0$. Therefore, $[1 - t_i^3]$ is a *concave* function of t_i, and, as we have seen, such functions can be majorized by a linear function of t_i. Equating the derivative of the general class of linear functions (the slope) with the derivative of $[1 - t_i^3]$ at a supporting point $\underline{t}_i$, and adjusting the intercept so that they coincide at that point, we obtain $1 - t_i^3 \leq 1 - 3\ \underline{t}_i^2\ t_i + 2\ \underline{t}_i^3$. Inserting the definition of t_i, it follows from this inequality and some rearrangements that the family

$$\mu_{Tuk}(\mathbf{r} \mid \underline{\mathbf{r}}) = 1/2 \sum_i \underline{w}_i r_i^2 + c^2/6\ [2 \sum_i\ \underline{t}_i^3 - 3 \sum_i\ \underline{t}_i^2 + n] \qquad (50)$$

majorizes $\psi_{Tuk}[\mathbf{r}]$, with weights $\underline{w}_i = \{1 - (\underline{r}_i/c)^2\}^2$ for $|\underline{r}_i| \leq c$, and $\underline{w}_i = 0$ elsewhere (for an alternative proof, see Verboon, 1990). Note that in this case it would not help much to write out $\psi_{Tuk}[\cdot]$ as a polynomial and majorize it term by term, since it would not be obvious how to majorize a positively curved sixth degree term. However, our interest in the function does not extend to the entire domain, but is limited to the range $|r_i| \leq c$, and in this restriction lies the key to understanding its local behavior and the possibility of majorization by a quadratic function.

Iteratively reweighted least squares. Since both Huber's and Tukey's biweight function can be majorized by a quadratic, as has been shown in (48) and (50), with different coefficients due to a different definition of the weights (the different additive constants are irrelevant for minimization), the IM model algorithm tells us to solve a series of weighted least squares problems, where the weights are redefined in each iteration as a function of the previous residuals. Thus IM theory provides a convergence proof for the use of *iteratively reweighted least squares* (IRLS) in resistant methods in general, and for Gabriel and Odoroff's (1984) resistant Eckart-Young approximation procedure, in particular (Verboon and Heiser, 1993). As is evident from the review by Green (1984), many different thoughts lead to IRLS, but the insight that it works via a *coherent* sequence of iteratively reshaped quadratic curves or surfaces provides a unifying justification. For the Eckart-Young form

(45), which sums m size functions across variables, we have to repeatly solve

$$\min_{\mathbf{X};\mathbf{Y}} \ \Sigma_k \, (\mathbf{h}_k - \mathbf{X}\mathbf{y}_k)' \underline{\mathbf{W}}_k (\mathbf{h}_k - \mathbf{X}\mathbf{y}_k) \ , \tag{51}$$

where the variable-wise weight matrices $\underline{\mathbf{W}}_k$ are diagonal, and depend on the previous residuals – according to Huber function, the Tukey function, or some other resistant alternative. Element-wise weighted bilinear approximation is an option, for example, in the computer program MULTIPALS (Verboon, Van der Lans, and Heiser, 1991), which also includes the possibility of optimal scaling of the $\mathbf{h}_k$, thus providing the basis of a method for resistant *nonlinear* principal components analysis.

The Pearson form could be handled analogously, but there is a caveat. Due to the fact that the residuals $\mathbf{s}_{ka}$ could be made arbitrary small by joint rescalings of $\mathbf{X}$ and $\mathbf{Y}$, normalization is not just a matter of identification, as it is in (51), but an influential consideration. Finding out what would be a correct or efficient normalization for resistant size functions is an open research question.

Towards Global Optimization: Tunneling (in Unconstrained Multidimensional Scaling)

Suppose that we have a convergent algorithm for multidimensional scaling, minimizing $\phi_A(\cdot)$ in (6), that always brings us to a local minimum (e.g., the SMACOF algorithm, Heiser and De Leeuw, 1977). A recurrent query is: could there be another stationary point with a lower value of the objective function, or is the current local minimum also the global minimum? Although evidence is accumulating that under normal conditions – and excepting the one-dimensional, or *seriation* case – the local minimum problem is not as severe as might be expected (Groenen, 1993), evidently the identification of the global minimum is an important computational objective.

Iterative methods generally use only local information to continue their trajectory of successive improvements, and therefore a major challenge is how to get away from a given stationary point, since – by definition – once we are there, all local information drives us back to the same point. One answer to this challenge is provided by an approach called *tunneling*, which employs a very effective tool, called the *movable pole*, to annihilate the attraction of globally suboptimal stationary points, and gives a descending series of local minima. Although tunneling does not *guarantee* the identification of the global minimum, in practice it often seems to come rather close to that goal. The development of

this approach for unconstrained global MDS has been discussed in more detail by Groenen and Heiser (1991); also see Groenen (1992, 1993).

Basic tools of tunneling. The tunneling approach to global optimization is due to Levy and Gomez (reviewed in Levy and Gomez, 1985; also see Gomez and Levy, 1982, for the constrained case), who suggested a procedure with two alternating phases. Each phase is characterized by the use of one basic tool: either (1) a local optimizer, or (2) a persistent zero finder. In one phase, called the *minimization phase*, the local optimizer searches for the closest local minimum, and in the other phase, called the *tunneling phase*, the persistent zero finder is used to locate another point that has the same objective function value as the current local minimum. These phases are shown in Figure 3,

Insert Figure 3 about here

Two alternating phases in the tunneling method for finding the global minimum.

based on Figure 1 of Levy and Gomez (1985), which displays an arbitrary function of one variable with multiple local minima. As suggested by the usual landscape metaphor, the procedure tunnels underneath the objective function surface, disregarding irrelevant local minima, to a new starting position from which a better function value can be obtained; hence its name. Since a local optimizer can always be constructed following IM theory, the present discussion focusses on the process to locate a new point with the same function value.

The *tunneling function* $\phi_B(\mathbf{X})$, which controls this process, has a numerator that adds an intercept to the objective function $\phi_A(\mathbf{X})$ so that the current local minimum would obtain a value zero, and a denominator that introduces a *pole* at (nearly) stationary points. The pole is a multiplicative penalty function that should act locally: it serves to make selected points unattractive by lifting the function surface in their immediate neighbourhood. If $\mathbf{X}^*$ denotes the previous local minimum, the numerator of the tunneling function is $\zeta(\mathbf{X}) = \phi_A(\mathbf{X}) - \phi_A(\mathbf{X}^*)$, and a basic candidate for the denominator is $\nu(\mathbf{X}) = \|\mathbf{X}^* - \mathbf{X}\|$; so as a first try, we use the balance function $\phi_B(\mathbf{X}) = \zeta(\mathbf{X}) / \nu(\mathbf{X})$. The objective of the tunneling phase is to find a *root* of the equation $\phi_B(\mathbf{X}) = 0$, or a *zero* of the function $\phi_B(\cdot)$. One of the classic ways to find a root of an equation is Newton's method, and we will make a slight

digression to show that this method has a particular IM interpretation. From this discussion it also follows that a quadratic IM root-finding procedure can be devised that provides an alternative with guaranteed convergence from *any* point of initialization.

Recall that Newton's root-finding method for a function of one variable $\phi(x)$ consists of choosing improved estimates of the root by the recursion $x^+ = \underline{x} - \phi(\underline{x}) \,/\, \phi'(\underline{x})$, where $\underline{x}$ is the old estimate. Thus the direction and the length of each step $(x^+ - \underline{x})$ satifies $\phi'(\underline{x})(x^+ - \underline{x}) = -\phi(\underline{x})$. Generalized to *one* function of *more* variables $\phi_B(\mathbf{X})$, where the matrix shape of $\mathbf{X}$ is disregarded so that it can be called a point or a vector, Newton's method works with a displacement vector $\Delta\mathbf{X} = \mathbf{X}^+ - \underline{\mathbf{X}}$ that is chosen concurrent with the direction of greatest change, indicated by $\nabla\phi_B(\underline{\mathbf{X}})$, the vector of partial derivatives evaluated in $\underline{\mathbf{X}}$. Analogously to the condition on the step-size in the one-variable case, the length of the displacement vector is determined by the matrix equation $\mathrm{tr}\,(\nabla\phi_B(\underline{\mathbf{X}}))'(\Delta\mathbf{X}) = -\phi_B(\underline{\mathbf{X}})$, which is satisfied if the recursion is defined as:

$$\mathbf{X}^+ = \underline{\mathbf{X}} - \frac{\phi_B(\underline{\mathbf{X}})}{\|\nabla\phi_B(\underline{\mathbf{X}})\|^2}\,\nabla\phi_B(\underline{\mathbf{X}})\ . \tag{52}$$

This recursion is given explicitly here, because most textbooks treat Newton's method only for finding either the root of one equation in one variable, or the set of roots of m equations in m variables. But it is not necessary, and in the present application even undesirable in view of computational costs, to execute the tunneling phase by finding the (matrix-valued) root of $\nabla\,|\phi_B(\mathbf{X})\,| = \mathbf{0}$, as appears to be the suggestion in Levy and Gomez (1985), rather than of $\phi_B(\mathbf{X}) = 0$ itself, as proposed here. In the derivation of (52) the function $\mathrm{tr}\,(\nabla\phi_B(\mathbf{X}))'(\Delta\mathbf{X})$ was used, called the *directional derivative* (Rudin, 1964), which measures the rate of change in direction $\Delta\mathbf{X}$. The rate of change is maximal if we choose $\Delta\mathbf{X}$ proportional to $\nabla\phi_B(\mathbf{X})$.

It is well-known that Newton's method works best if – near the root – the system is close to linear and the derivatives are bounded away from zero; also, its convergence is guaranteed, but only if it is started close enough to the root, or within an interval that satisfies certain conditions (Henrici, 1964). As can be deduced from (52), if the recursion is expressed in terms of the normalized partial derivatives, the step-size is $\phi_B(\underline{\mathbf{X}}) \,/\, \|\nabla\phi_B(\underline{\mathbf{X}})\|$, the 'residual' (the current distance to zero) divided by the size of the derivative. So excessively big steps are to expected only if the derivative would vanish

when approaching the root. For simplicity of presentation, the one-variable case is now used to give a majorization interpretation of Newton's method.

Suppose that we are actually looking for the zero of a linear function $\phi(x) = bx + c$ with $b > 0$ (the case of $b < 0$ runs analogously). If the process is started at some point where $\phi(\underline{x}) > 0$, then IM may be used to find an improved estimate $\overline{x}$ with $\phi(\overline{x})$ guaranteed to be lower. In (27), a majorization function was given for the linear case, depending upon some freely chosen constant β. Now, if we choose $\beta = {}^1/_2\, b^2 / (b\underline{x} + c)$, which is always defined provided that the supporting point $\underline{x}$ is not itself a root (in which case we are ready), the majorizing quadratic becomes

$$\mu_{New}(x \mid \underline{x}) = {}^1/_2\, [\{(bx + c)^2 / \phi(\underline{x})\} + \phi(\underline{x})] \,. \tag{53}$$

The derivative of (53), evaluated in $\underline{x}$, is equal to b, which is also equal to the (constant) derivative of $\phi(x)$. As an aside, note that if $\phi(\underline{x}) < 0$, the denominator of the first term in (53) is negative, and we actually are using *minorization* for finding the zero, but in the tunneling problem we would stop, because $\underline{x}$ would be a good position to enter a new minimization phase. Continuing with finding the best step in iteratively solving a linear equation, IM theory tells us to choose the argument minimizing $\mu_{New}(x \mid \underline{x})$, which is $\overline{x} = -\,c / b$, i.e. the zero of $\phi(x)$; wherever we start, the root is found in one step! Writing the successor point as

$$\overline{x} = -\,c / b = \underline{x} - (\underline{x} + c / b) = \underline{x} - (b\underline{x} + c) / b = \underline{x} - \phi(\underline{x}) / \phi'(\underline{x}) \,, \tag{54}$$

we obtain the Newton recursion. So Newton's method can be interpreted as an IM method, in which the quadratic majorizer is chosen in such a way that its minimum is the root, if the function is locally approximately linear. The important point is that, without assuming linearity, quadratic IM *always* would provide a successor point with $\phi(\overline{x})$ closer to zero than $\phi(\underline{x})$. An IM procedure will find a root regardless of the shape of the function or the initialization point, and therefore it is regarded here as the preferred way to proceed.

Difficulties that need to be taken care of in the tunneling function. In the development of the tunneling approach in the context of least squares multidimensional scaling, three difficulties were encountered that are likely to occur more generally.

First, there is the problem of *pole strength.* The idea of a pole is actually used for two purposes: to cancel out the previous local minimum, and to cancel out irrelevant local minima of the tunneling function, so that the *radius of convergence* of the zero finder is extended. If the algorithm approaches $\mathbf{X}^\circ$, a point where $\|\nabla\phi_B(\mathbf{X}^\circ)\|$ almost vanishes while $\phi_B(\mathbf{X}^\circ) > 0$, the denominator $\nu(\mathbf{X})$ is changed into the product of two norms $\|\mathbf{X}^* - \mathbf{X}\|\,\|\mathbf{X}^\circ - \mathbf{X}\|$, which eliminates the attraction of $\mathbf{X}^\circ$ without destroying the zeros of $\phi_B(\mathbf{X})$. This valuable idea is the reason for the qualification *persistent* zero finder; however, to actually achieve cancellation of these points of attraction, the pole has to be strong enough. The strength of the denominator is relative to the curvature of the numerator $\zeta(\mathbf{X})$, and therefore the influence of the pole can be increased either by raising $\nu(\mathbf{X})$ to the power r with $r > 1$, or by taking the rth root of $\zeta(\mathbf{X})$. The latter is preferred in the IM approach, because it is easier to majorize the rth root than to deal with a power in the denominator. Since the tunneling phase can be stopped as soon as $\zeta(\mathbf{X}) < 0$, it may be assumed that $\zeta(\mathbf{X}) \geq 0$, as is actually done in Groenen (1993, section 3.2.5), and therefore the problem of the pole strength can be handled by switching to $(\zeta(\mathbf{X}))^{1/r}$ in the numerator of the tunneling function.

Secondly, there is a need to eliminate rotations of $\mathbf{X}$, because we do not want to tunnel to merely a rotation of the previous local minimum, which is a real possibility, because the MDS objective function $\phi_A(\mathbf{X})$ is invariant under rotation of its argument: i.e., $\phi_A(\mathbf{XR}) = \phi_A(\mathbf{X})$ for any square $p \times p$ matrix $\mathbf{R}$ satisfying $\mathbf{R'R} = \mathbf{I}$. The invariance is due to the fact that the objective function deals with a Euclidean distance model, in which $D(\mathbf{X})$ is fitted to the data, and the invariants of the distance function are transferred to the objective function. This problem can be handled by extending the effect of the pole to rotations as well, i.e. by changing $\nu(\mathbf{X})$ into $\|D(\mathbf{X}^*) - D(\mathbf{X})\|$ or $\|\mathrm{D}(\mathbf{X}^*) - \mathrm{D}(\mathbf{X})\| \times \|\mathrm{D}(\mathbf{X}^\circ) - \mathrm{D}(\mathbf{X})\|$, functions which are also invariant under rotation.

Thirdly, it became obvious that if the pole is made strong enough to annihilate a local minimum, the effect of $\nu(\mathbf{X})$ is no longer local, as it was intended to be, but causes the problem of an *attractive horizon*, meaning that $\phi_B(\mathbf{X}) \rightarrow 0$ when $\mathbf{X}$ becomes uniformly large. The attactive horizon can be removed by switching from $\nu(\mathbf{X})$ to $\nu(\mathbf{X}) / (1 + \nu(\mathbf{X}))$, which approaches 1 in the limit.

Combining all adjustments, the tunneling function becomes

$$\phi_B(\mathbf{X}) = [\|\Delta - D(\mathbf{X})\| - \phi_A(\mathbf{X}^*)]^{1/r} \; \frac{1 + \|D(\mathbf{X}^*) - D(\mathbf{X})\|}{\|D(\mathbf{X}^*) - D(\mathbf{X})\|} \; . \tag{55}$$

This balance function involves a concave part (the rth root), several convex parts (the Euclidean norms), another concave part (the cross-product term – tr $\Delta' D(\mathbf{X})$), a convex quadratic part (the sum of squares tr $D(\mathbf{X})^2$), a product, and a division. So there is plenty of room for applying the present IM theory to obtain a convergent algorithm. Since the majorization is nested in several layers, there is also plenty of room for tuning, as computations become rather heavy (Groenen (1993) reports runs of several hours on a SUN-SPARC workstation for $n = 17$, or several days for larger problems); however, the important point is, that a decreasing series of local minima can be obtained.

The IM Aspect of the EM Algorithm

The EM algorithm (Dempster, Laird and Rubin, 1977) is a very general computational strategy for statistical calculations with missing information that converges monotonically. To clarify this feature within the present framework, some adjustments in notation have to be made. Let $f(\mathbf{z} \mid \mathbf{x})$ denote the sampling density of the complete data $\mathbf{z}$, given the unknown parameter vector $\mathbf{x}$. The vector $\mathbf{z} = (\mathbf{h},\mathbf{y})$ combines information on the actually observed data $\mathbf{h}$ and the 'missing data', or additional unknown quantities, $\mathbf{y}$. Let $g(\mathbf{h} \mid \mathbf{x})$ be the likelihood of the incomplete data, i.e. the actual likelihood to be maximized, and $k(\mathbf{z} \mid \mathbf{h},\mathbf{x})$ be the conditional density $f(\mathbf{z} \mid \mathbf{x}) / g(\mathbf{h} \mid \mathbf{x})$. The principle to split the objective function in this case consists of the decomposition of minus the loglikelihood of the incomplete data as

$$- \log g(\mathbf{h} \mid \mathbf{x}) = - E[\log f(\mathbf{z} \mid \mathbf{x}) \mid \mathbf{h},\underline{\mathbf{x}}] + E[\log k(\mathbf{z} \mid \mathbf{h},\mathbf{x}) \mid \mathbf{h},\underline{\mathbf{x}}] \,, \tag{56}$$

which follows from equations (2.4)-(2.6) in Dempster *et al.* (1977), i.e. from rearranging the definition of $k(\mathbf{z} \mid \mathbf{h},\mathbf{x})$, and taking the log and expectations. The first term in (56) is (minus) the current expectation of the complete data specification, given observed values $\mathbf{h}$ and current estimates $\underline{\mathbf{x}}$, and the second term is the expected conditional density of the complete data, also given observed values $\mathbf{h}$ and current estimates $\underline{\mathbf{x}}$. It is this second term that can be majorized as

$$E[\log k(\mathbf{z} \mid \mathbf{h},\mathbf{x}) \mid \mathbf{h},\underline{\mathbf{x}}] \leq E[\log k(\mathbf{z} \mid \mathbf{h},\underline{\mathbf{x}}) \mid \mathbf{h},\underline{\mathbf{x}}], \tag{57}$$

an inequality that is a consequence of Jensen's inequality for conditional expectations (Marshall and Olkin, 1979, Chap. 16C).

To better understand (57), consider the simple finite case, and suppose that we express the m elements of the conditional density $k(\mathbf{z} \mid \mathbf{h},\mathbf{x})$ as $\{p_1,\ldots,p_j,\ldots,p_m\}$ and the m elements of the currently estimated density $k(\mathbf{z} \mid \mathbf{h},\underline{\mathbf{x}})$ as $\{\underline{p}_1,\ldots,\underline{p}_j,\ldots,\underline{p}_m\}$. Since they are conditional probabilities, these quantities satisfy $\Sigma_j p_j = \Sigma_j \underline{p}_j = 1$. Because the logarithm is concave, we have for a convex combination of terms $p_j/\underline{p}_j$ the inequality

$$\log \Sigma_j \underline{p}_j (p_j/\underline{p}_j) \geq \Sigma_j \underline{p}_j \log (p_j/\underline{p}_j), \tag{58}$$

with equality if and only if $p_j = \underline{p}_j$ for all j. But the fact that the probabilities sum to one causes the left-hand side of (58) to be zero, from which it follows that

$$\Sigma_j \, p_j \, \log p_j \leq \Sigma_j \, p_j \, \log \underline{p}_j, \tag{59}$$

which is the discrete version of (57). It is also possible to derive (59) by substituting $x = p_j$ in (29) and taking a weighted sum with weights $\underline{p}_j$.

Equality is attained in (57) if and only if $\mathbf{x} = \underline{\mathbf{x}}$, so that the general case satisfies IM condition (13), too. Most importantly, note that the right-hand side of (57) does not depend on $\mathbf{x}$, and therefore (56) can be minimized by maximizing $E[\log f(\mathbf{z} \mid \mathbf{x}) \mid \mathbf{h}, \underline{\mathbf{x}}]$, the current expectation of the complete data. In the E-step of the EM algorithm one determines the expectation of $\mathbf{z}$, given the current estimates $\underline{\mathbf{x}}$, which corresponds to choosing a member from the family of majorizing functions, and then in the M-step one maximizes the expected complete data loglikelihood, which corresponds to finding the minimum of the majorizing function.

5. Discussion

Three principles, a number of inequalities, and several combination rules form the core of this theory of algorithm construction that insists on monotonic convergence. When algorithms are developed heuristically, it is often not so much the heuristic itself that is detrimental to success, but the corrections that are brought in when the method starts behaving unexpectedly. Without the guidance of a monotonic descent property, the cumulation of repairs tends to make the whole process incomprehensible rather quickly. In many cases, the IM procedure follows a trajectory along directions of steepest descent, as many other algorithms do. However, the essential thing is, that it always pursues a course of *monotonic* descent.

Most of the results presented here are of a qualitative nature, showing how optimal points can be reached, but not how fast. Quantitative indications of convergence rates for multidimensional scaling have been presented by De Leeuw (1988). Acceleration is a recurring need, since convergence actually is often slow, due to the fact that the algorithm tends to move along (sub)gradient directions that are almost perpendicular to the direction pointing towards the minimum. Indeed, slow

convergence is probably IM's major drawback, and therefore it is hoped that the acceleration device discussed in the present paper will be useful in other areas than multidimensional scaling, too.

The last remark does not extend immediately to the EM algorithm, because – even though it is an IM algorithm – the majorization function involved is not necessarily quadratic, but depends on the shape of the expected loglikelihood function. Here, it might be feasible to make progress if (57) could be replaced by a sharper bounding function, since majorization by a constant is the weakest possibility. Generally speaking, majorization with functions that are closer to the minorant is more efficient, a fact used by Kiers and Ten Berge (1992) to improve the general-purpose scheme proposed by Kiers (1990). This example illustrates the rule that tailor-made algorithms, which use as many special characteristics of the objective function as possible, are to be preferred to algorithms that work for a wider class of functions, since in the latter case, the special characteristics have to be recognized at execution time by the algorithmic process itself, which is always more costly.

Both in theory and in actual applications, the decomposition of an objective function into concave and convex parts turns out to be a very powerful idea. It enables a systematic search for majorizers of many objective functions in multidimensional data analysis. A lot of the majorization results given for concave functions can be obtained by a general application of the Taylor series expansion of $\phi(x)$ around $\underline{x}$, or, in the multivariable version, of $\phi(\mathbf{x})$ around $\underline{\mathbf{x}}$ (e.g., (28), (29), (31), (32)). We just use the first two terms and note that the remainder defined through the third term must be negative given the particular $\phi(x)$ we started with. This general strategy is recommended for all cases in which the objective function can be split into a convex and a concave part (or, equivalently, the difference between two convex functions). If the concave part is non-differentiable, the first derivatives in the Taylor expansion are replaced by a subdifferential, and the same reasoning applies – by the very definition of subdifferentiability. Functions with parts that are steeper than quadratic cannot be majorized immediately, but need a preliminary analysis, as shown in the case of Tukey's biweight function. Although the IM strategy is designed for local descent, it can also be used in procedures for global descent, such as the tunneling procedure, which is based upon an adaptation of Newton's method that radically extends its radius of convergence.

Acknowledgement

The author is indebted to Patrick Groenen, Ivo van der Lans, Larry Hubert, and Rudolf Mathar for their helpful comments on an earlier draft of this paper.

6. References

Bijleveld, C., and De Leeuw, J. (1991). Fitting longitudinal reduced rank regression models by alternating least squares. *Psychometrika, 56*, 433-447.

Bloxom, B. (1978). Constrained multidimensional scaling in *N* spaces. *Psychometrika, 43*, 397-408.

Burden, R.L., and Faires, J.D. (1985). *Numerical Analysis (3rd Edition)*. Boston, Mass.: PWS-Kent Publishers.

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

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

De Leeuw, J. (1984). The Gifi system of multivariate analysis. In E. Diday et al. (Eds.), *Data Analysis and Informatics, Vol III*. Amsterdam: North-Holland.

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

De Leeuw, J., and Bijleveld, C. (1988). *Fitting longitudinal reduced rank regression models by alternating least squares*. Research Report RR-88-03. Leiden: Department of Data Theory.

De Leeuw, J., and Heiser, W.J. (1977). Convergence of correction matrix algorithms for multidimensional scaling. In J.C. Lingoes *et al.* (Eds.), *Geometric Representations of Relational Data*, pp.735-752. Ann Arbor, Mich.: Mathesis Press.

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

De Leeuw, J., Van der Heijden, P.G.M., and Verboon, P. (1990). A latent time budget model. *Statistica Neerlandica, 44*, 1-22.

De Leeuw, J., Young, F.W., and Takane, Y. (1976). Additive structure in qualitative data: an alternating least squares method with optimal scaling features. *Psychometrika, 41*, 471-503.

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

De Soete, G., and Heiser, W.J. (1993). A latent class unfolding model for analyzing single stimulus

preference ratings. *Psychometrika, 58,* 545-565.

Eckart, C. and Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika, 1*, 211-218.

Friedman, J.H. and Tukey, J.W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers, C-23*, 881-890.

Friedman, J.H. and Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association, 76*, 817-823.

Gabriel, K.R. and Odoroff, C.L. (1984). Resistant lower rank approximation of matrices. In E. Diday, M. Jambu, L. Lebart, J. Pages, and R. Tomassone (Eds.), *Data Analysis and Informatics, III*, pp. 23-30. Amsterdam: North-Holland.

Gifi, A. (1990). *Nonlinear Multivariate Analysis*. New York: Wiley.

Gomez, S. and Levy, A.V. (1982). The tunneling method for solving the constrained global optimization problem with non-connected feasible regions. *Lecture Notes in Mathematics, 909*, 34-47. Berlin: Springer Verlag.

Green, P.J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society, B, 46*, 149-192.

Groenen, P.J.F. (1990). *The tunneling method applied to multidimensional scaling: Progress report 1*. Research Report RR-90-03. Leiden: Department of Data Theory.

Groenen, P.J.F. (1992). A comparison of two methods for global optimization in multidimensional scaling. In O. Opitz, B. Lausen, and R. Klahr (Eds.), *Information and Classification: Concepts, Methods and Applications*, pp. 145-155. Berlin: Springer Verlag.

Groenen, P.J.F. (1993). *The Majorization Approach to Multidimensional Scaling: Some Problems and Extensions*. Doctoral Dissertation. Leiden: DSWO Press.

Groenen, P.J.F. and Heiser, W.J. (1991). *An improved tunneling function for finding a decreasing series of local minima in MDS*. Research Report RR-91-06, Department of Data Theory, University of Leiden.

Groenen, P.J.F., Mathar, R., and Heiser, W.J. (1992). *The majorization approach to multi-dimensional scaling for Minkowski distances*. Research Report RR-92-11, Department of Data Theory, University of Leiden. To appear in: *Journal of Classification, 12* (1995), in press.

Hardy, G.H., Littlewood, J.E., and Pólya, G. (1952). *Inequalities* (2nd Edition). Cambridge: Cambridge University Press.

Henrici, P. (1964). *Elements of Numerical Analysis*. New York: Wiley.

Heiser, W.J. (1981). *Unfolding Analysis of Proximity Data*. Doctoral dissertation, University of Leiden, The Netherlands.

Heiser, W.J. (1987a). Correspondence analysis with least absolute residuals. *Computational Statistics and Data Analysis, 5*, 337-356.

Heiser, W.J. (1987b). Joint ordination of species and sites: the unfolding technique. In P. Legendre and L. Legendre (Eds.), *Developments in Numerical Ecology*, pp. 189-221. New York: Springer.

Heiser, W.J. (1988). Multidimensional scaling with least absolute residuals. In H.H. Bock (Ed.), *Classification and Related Methods of Data Analysis*, pp. 455-462. Amsterdam: North-Holland.

Heiser, W.J. (1989). The city-block model for three-way multidimensional scaling. In R. Coppi and S. Bolasco (Eds.), *Multiway Data Analysis*, pp. 395-404. Amsterdam: North-Holland.

Heiser, W.J. (1991). A generalized majorization method for least squares multidimensional scaling of pseudodistances that may be negative. *Psychometrika, 55*, 7-27.

Heiser, W.J. and De Leeuw, J. (1977). How to use SMACOF-1. Research Report, Leiden: Department of Data Theory.

Heiser, W.J. and Meulman, J.J. (1983). Constrained multidimensional scaling, including confirmation. *Applied Psychological Measurement, 7*, 381-404.

Heiser, W.J. and Meulman, J.J. (1989). The approximation of K subspaces by K other ones in a reduced common space. *Bulletin of the International Statistical Institute, Contributed Papers of the 47th Session*. Paris: ISI (1989), 430-431.

Heiser, W.J. and Meulman, J.J. (1994a). Homogeneity analysis: exploring the distribution of variables and their nonlinear relationships. In M. Greenacre, J. Blasius, and W. Kristof (Eds.), *Correspondence Analysis in the Social Sciences: Recent Developments and Applications*. New York, NY: Academic Press, in press.

Heiser, W.J. and Meulman, J.J. (1994b). Nonlinear methods for the analysis of homogeneity and heterogeneity. In W. Krzanowski (Ed.), *Recent Advances in Descriptive Multivariate Analysis*. Oxford: Oxford University Press, in press.

Heiser, W.J. and Stoop, I. (1986). *Explicit SMACOF algorithms for individual differences scaling*. Research Report RR-86-04. Leiden: Department of Data Theory.

Hotelling, H. (1936). Relations between two sets of variables. *Biometrika, 28*, 321-377.

Huber, P.J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics, 35*, 73-101.

Huber, P.J. (1981). *Robust Statistics*. New York: Wiley.

Huber, P.J. (1985). Projection pursuit. *The Annals of Statistics, 13*, 435-475.

Jones, M.C. and Sibson, R. (1987). What is projection pursuit? *Journal of the Royal Statistical Society, A, 150*, 1-36.

Kiers, H.A.L. (1990). Majorization as a tool for optimizing a class of matrix functions. *Psychometrika, 55*, 417-428.

Kiers, H.A.L. (1993). Handling ordinal variables in three-way analysis of quantification matrices for variables of mixed measurement levels. *British Journal of Mathematical and Statistical Psychology, 46*, 135-152.

Kiers, H.A.L. and Ten Berge, J.M.F. (1992). Minimization of a class of matrix trace functions by means of refined majorization. *Psychometrika, 57*, 371-382.

Kiers, H.A.L., Ten Berge, J.M.F., Takane, Y., and De Leeuw, J. (1990). A generalization of Takane's algorithm for DEDICOM. *Psychometrika, 55*, 151-158.

Koschat, M.A., and Swayne, D.F. (1991). A weighted Procrustes criterion. *Psychometrika, 56,* 229-239.

Kruskal, J.B. (1969). Toward a practical method which helps uncover the structure of a set of multivariate observations by finding the linear transformation which optimizes a new 'index of condensation'. In R.C. Milton and J.A. Nelder (Eds.), *Statistical Computation*. New York: Academic Press.

Kruskal, J.B. (1972). Linear transformation of multivariate data to reveal clustering. In R.N. Shepard *et al.* (Eds.), *Multidimensional Scaling: Theory and Applications in the Behavioral Sciences, vol. I, Theory*. New York: Seminar Press.

Kruskal, J.B. and Carroll, J.D. (1969). Geometric Models and badness-of-fit functions. In P.R. Krishnaiah (Ed.), *Multivariate Analysis, Vol. II*. New York: Academic Press.

Levy, A.V. and Gomez, S. (1985). The tunneling method applied to global optimization. In P.T. Boggs, R.H. Byrd, and R.B. Schnabel (Eds.), *Numerical Optimization 1984*. Philadelphia: SIAM.

Marshall, A.W. and Olkin, I. (1979). *Inequalities: Theory of Majorization and its Applications*. London: Academic Press.

Meulman, J.J. (1986). *A Distance Approach to Multivariate Analysis*. Doctoral Dissertation. Leiden: DSWO Press.

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

Meulman, J.J. (1993). Principal coordinates analysis with optimal transformations of the variables: minimizing the sum of squares of the smallest eigenvalues. *British Journal of Mathematical and Statistical Psychology, 46, in press.*

Meulman, J.J. and Heiser, W.J. (1984). Constrained multidimensional scaling: more directions than dimensions. In T. Havránek et al (Eds.), *COMPSTAT 1984*, pp.137-142 Wien: Physica Verlag.

Meulman, J.J. and Heiser, W.J. (1993). Nonlinear biplots for nonlinear mappings. In O. Opitz, B. Lausen, and R. Klar (Eds.), *Information and Classification: Concepts, Methods and Applications*, pp. 201-213. Berlin : Springer Verlag.

Meulman, J.J. and Verboon, P. (1993). Points of view analysis revisited: fitting multidimensional structures to optimal distance components with cluster restrictions on the variables. *Psychometrika, 58*, 7-35.

Nierop, A.F.M. (1993). *Multidimensional Analysis of Grouped Variables: An Integrated Approach*. Leiden: DSWO Press.

Rockafellar, R.T. (1970). *Convex Analysis*. Princeton, NJ: Princeton University Press.

Rudin, W. (1964). *Principles of Mathematical Analysis*. New York: McGraw-Hill.

Selim, S.Z., and Ismael, M.A. (1984). K-means type algorithms: a generalized convergence theorem and characterization of local optimality. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6*, 81-87.

Shor, N.Z. (1985). *Minimization Methods for Non-differentiable Functions*. Berlin: Springer Verlag.

Ten Berge, J.M.F., Knol, D.L. and Kiers, H.A.L. (1988). A treatment of the orthomax Rotation family in terms of diagonalization, and a re-examination of a singular value approach to varimax rotation. *Computational Statistics Quarterly, 3*, 207-217.

Tucker, L.R., and Messick, S. (1963). An individual differences model for multidimensional scaling. *Psychometrika, 28*, 333-367.

Van Buuren, S. (1990). *Optimal Scaling of Time Series*. Doctoral Dissertation. Leiden: DSWO Press.

Van der Burg, E., De Leeuw, J., and Verdegaal, R. (1988). Homogeneity analysis with *K* sets of variables: an alternating least squares method with optimal scaling features. *Psychometrika, 53*, 177-197.

Van der Lans, I.A. (1989). *Nonlinear reduced rank generalized canonical correlation analysis, including common scale quantifications and data weights*. Research Report RR-89-06, Department of Data Theory, University of Leiden.

Van der Lans, I.A. (1992). *Nonlinear Multivariate Analysis for Multiattribute Preference Data*. Doctoral dissertation, Leiden: DSWO Press.

Verboon, P. (1990). *Majorization with iteratively reweighted least squares: a general approach to optimize a class of resistant loss functions*. Research Report RR-90-07, Department of Data Theory, University of Leiden.

Verboon, P. (1993). Robust regression with optimal scaling. *British Journal of Mathematical and Statistical Psychology, 46, in press.*

Verboon, P. and Heiser, W.J. (1992). Resistant orthogonal Procrustes analysis. *Journal of Classification, 9*, 237-256.

Verboon, P. and Heiser, W.J. (1993). Resistant lower rank approximation of matrices by iterative majorization. *Computational Statistics and Data Analysis, 16, in press.*

Verboon, P., Van der Lans, I.A., and Heiser, W.J. (1991). *The Multipals algorithm*. Research Report RR-91-04, Department of Data Theory, University of Leiden.

Wold, H. (1966). Nonlinear estimation by iterative least squares procedures. In F.N. David (Ed.), *Research Papers in Statistics: Festschrift for J. Neyman*, p. 411-444. New York: Wiley.

Wolfe, P. (1975). A method of conjugate subgradients for minimizing nondifferentiable functions. In M.L. Balinski and P. Wolfe (Eds.), *Mathematical Programming Studies 3*, pp. 145-173. Amsterdam: North-Holland.

Zielman, B., and Heiser, W.J. (1993). Analysis of asymmetry by a slide vector. *Psychometrika, 58*, 101-114.

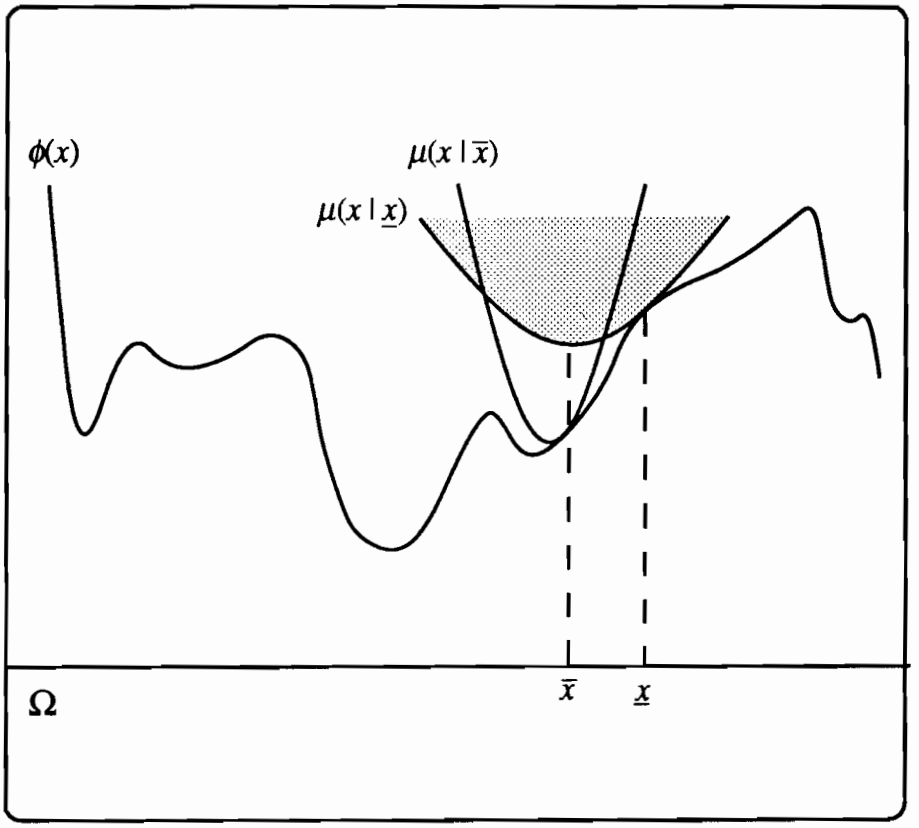

Figure 1.

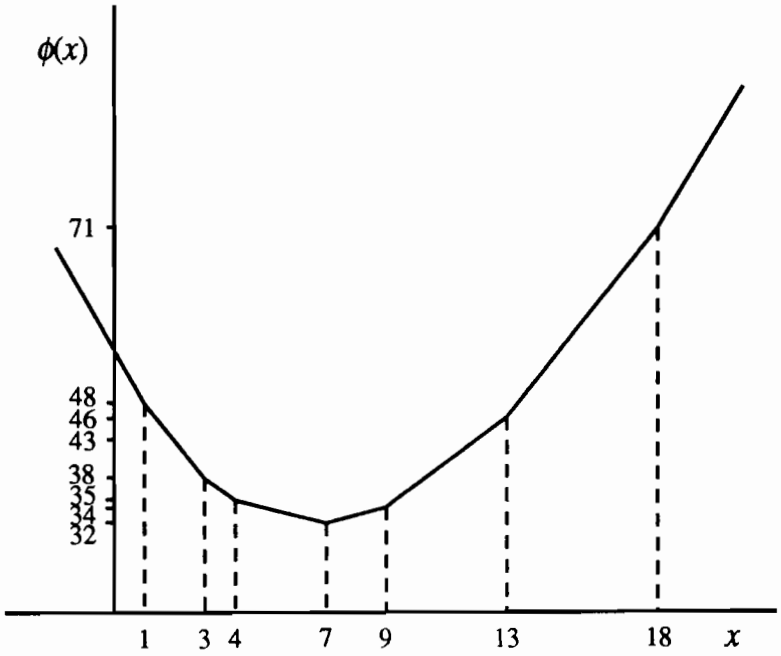

Figure 2.

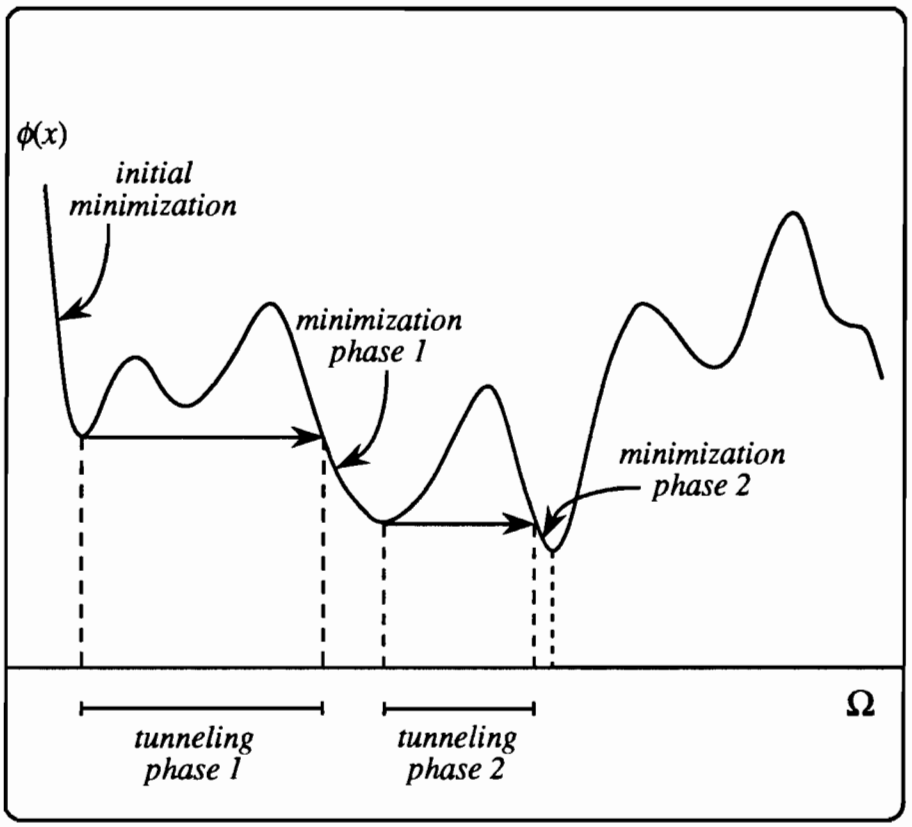
$\phi(x)$
initial
minimization
minimization
phase 1
minimization
phase 2
tunneling
phase 1
tunneling
phase 2
Ω

Figure 3.