RN 001 - 70

THE POSITIVE ORTHANT METHOD FOR NONMETRIC MULTIDIMENSIONAL SCALING

Jan de Leeuw

January 1970

Department of data theory for the social sciences / University of Leiden

SUMMARY

In this paper we describe our approach to nonmetric multidimensional scaling and compare it with some of the already existing approaches. On the basis of an analysis of what we mean by 'nonmetric' we develop an algorithm that can handle the most general type of nonmetric data, and that can choose from a continuum of possible error theories with the possibility of an arbitrary close approximation to a nonmetric error theory. Special attention is given to the properties of the loss-functions, to the precise nature of the nonmetric requirements, to the treatment of ties and inconsistencies, to the initial configuration, and to algorithmic aspects. We conclude with an enumeration of the problems that are still to be investigated.

NOTE

This report is a revised and expanded version of a previous mimeographed paper. About forty copies of that paper were sent away last year. If you happen to have one, throw it away and replace it by this report.

CONTENTS

1. Introduction

1.1. Definition of nonmetric multidimensional scaling

In nonmetric multidimensional scaling (NMS) the primitives are a (finite) stimulus set A and a function ϕ that maps a subset of A x A into the set Δ. Over the set Δ a binary relation $\underset{\delta}{\geq}$ (not necessary an order) is defined. The elements of Δ are called dissimilarities. We shall frequently use the notation

$$\delta_{ij} =_{\text{def}} \phi(a_i, a_j). \tag{1}$$

Let Re^p be the set of all ordered p-tuples of real numbers and d a metric on Re^p. If ω is a mapping of A into Re^p we shall write

$$d_{ij} =_{\text{def}} d(\omega\ (a_i),\ \omega(a_j)). \tag{2}$$

The image of A under ω(a finite subset of Re^p) is called a representation of A. The purpose of NMS-algorithms is to construct the mapping ω in such a way that the requirements

$$\delta_{ij} \underset{\delta}{\geq} \delta_{kl} \Rightarrow d_{ij} \geq d_{kl} \tag{3}$$

are optimally satisfied for all (δ_{ij}, δ_{kl}) ε $\underset{\delta}{\geq}$. (The expression A $\Rightarrow$B must be read as: if A then our representation must be found in such a way that B). The use op 'optimally' implies that all NMS-theories need a particular error theory. The define a loss-function that measures the departure from perfect fit to the requirements (3) and the algorithms minimize this coefficient.

1.2 Classification of existing NMS-algorithms

Nonmetric multidimensional scaling techniques use only the relational properties of the data, i.e. it suffices to input the indicator of $\underset{\delta}{\geq}$. If Δ is a subset of the reals and $\underset{\delta}{\geq} \equiv \geq$ then this is equivalent to the assertion that the results of the algorithms are invariant under a strictly monotonic increasing transformation of the δ_{ij}. The existing algorithms can be classified according to the assumptions they make about the nature of the relation $\underset{\delta}{\geq}$. In the original versions of the Kruskal-Roskam (KR) and Guttman-Lingoes(GL) series $\underset{\delta}{\geq}$ is supposed to be a weak order (Kruskal 1964 a,b, Guttman 1968). Both have been extended to the conditional case in which Δ can be partitioned into several weakly ordered subsets, and in which the partial order $\geq_\delta$ is the union of these weak orders (Gleason 1967, Roskam 1968, Lingoes 1966 a,b).

The term'nonmetric' is used in yet another sense. For a particular representation of A each violation of the requirements (3) is called an <u>error</u>. We shall say that an error theory is nonmetric if and only if the loss function is a strictly monotonic increasing funtion of the number of errors. If the size of the errors also influences the value of the coefficient the error theory is called metric.
In the first case the errors are counted, in the second case they are measured. Both the KR- and the GL-approach use metric error theories.
A third method for the multidimensional scaling of similarity judgments is due to Coombs and Hays (Coombs 1964,Chapter 21 and 22). According to the definition in the previous section the CH-method is not an NMS-algorithm because it does not arrive at a metric representation of the stimuli.
It only produces partial ordenings of the projections of A on the dimensions of Euclidean p-space. This has some obvious disadvantages (cf Shepard 1966). Another practical disadvantage is that the algorithm is not programmed for a computer as yet. The CH-approach is different in two other major respects. In the first place it is able to handle any binary relation over Δ. In its most complete form the algorithm includes, as a first stage, the possibility to convert the arbitrary binary relation $\geq_\delta$ into a partial order. This is a definite advantage in a considerable number of practical situations. It eliminates, for example, the need to 'blow up' partially ordered data structures into weakly ordered ones. Moreover it makes it easier to analyze the data of individual subjects separately, thus avoiding the problems connected with averaging (cf Shepard 1964). A rather serious limitation of the CH-method is that it only considers order relations on the conjoint distances ('within triples'). Of course this does not restrict its generality in the conditional case or in the case that the data are collected by the method of trials.
In the second place the error theory of the CH-approach is nonmetric. The chief reason for calling this an advantage is that it seems less arbitrary. There is only one way to count errors, but there is an infinite number of ways to weight errors. Moreover our optimal solution is invariant under strictly monotone transformations of the d_{ij}. This may be important from a computational point of view (taking d_{ij}^2 simplifies the treatment in the Euclidean case),but also from a theoretical point

of view. If ϕ is a strictly monotone, subadditive, real valued function with $\phi(0) = 0$, and d is a metric, then $\phi(d)$ also is a metric. An optimal solution for d is an optimal solution for the class of metrics $\phi(d)$ iff we use a nonmetric error theory.
In this paper we shall develop an algorithm for NMS that is also able to handle binary relations over Δ, but arrives at a metric representation of A. Moreover it concludes the possibility of an arbitrary close approximation to a nonmetric error theory. We are able to regulate the degree of metricity of the error theory, so to speak.

2. Data.

2.1 Nature of the data

Suppose the stimulus set A has n elements. Let N be the set of the first n positive integers. In NMS the data can be defined as a subset L of the Cartesian product N^4. L can be constructed in several ways. The most complete type of data arises if we show the subject all n^4 possible combinations of pairs of stimuli and ask him if stimuli a_i and a_j are more dissimilar than stimuli a_k and a_l. Let $(i,j,k,l) \varepsilon L$ iff he responds in the affirmative, and $(k,l,i,j) \varepsilon L$ otherwise. In other cases the raw data consist of a mapping ϕ of A x A into a set Δ, strictly ordered by a relation $\underset{o}{>}$. We may define L by the rule

$$(i,j,k,l)\varepsilon L \Longleftrightarrow \delta_{ij} \underset{o}{>} \delta_{kl}. \qquad (4)$$

Because the order relation is asymmetric and irreflexive, L will have $C(n^2,2) = \frac{1}{2}n^2(n+1)(n-1)$ elements (we write the binomal coefficients as C(n,m)). In most cases, however, there will be some symmetry in the data such that

$$\delta_{ij} = \delta_{ji} \underset{o}{>} \delta_{ii} =_o \delta_{jj} \qquad (5)$$

for all, $i,j,\varepsilon N, i \neq j$. In that case we only have to consider the mapping of a subset S of A x A into Δ, with $(a_i,a_j)\varepsilon$ S iff $j > i$. Because of the ≠ $\frac{1}{8}n(n-1)(n-2)(n+1)$ elements. In the sequel we shall write D_n for this number. In other cases there are mappings of the sets $\{a_i\}$ x A into the strictly ordered sets $<\Delta_i, >_i>$ for all $i \varepsilon N$. Evidently this produces

≠ asymmetry and irreflexivity it follows that L contains C(C(n,2),2) =

$nC(n,2) = \frac{1}{2}n^2(n-1)$ elements in L. In the method of k-ads all possible sets of k stimuli are presented and the $C(k,2)$ dissimilarities in each of these sets are strictly ordered. This procedure in a similar way $C(n,k)C(C(k,2),2) = C(n,k)D_k$ elements in L. If k=3 we have the method of triads with $n(L) = \frac{1}{2}n(n-1)(n-2)$. In the sequel the number of elements in L will be denoted by m. Moreover λ is a one-one mapping of L onto $M =_{\text{def}} \{1,2,\ldots,m\}$.

2.2 <u>Treatment of ties</u>

In the previous section we have assumed in most cases that the relevant subsets of A x A are strictly ordered. Or, in other words, that $\underset{\delta}{\geq}$ is a weak order and that the data structure does not contain ties. As in most non(para)metric problems (of Kendall 1962,ch 3) the appearance of ties is a nuisance for NMS-algorithms too. We may handle ties in two different ways(of Kruskal 1964a p 22, Roskam 1968 p 39-40, Guttman 1968 p 477).Consider the case in which ϕ maps a subset of A x A into a subset of the reals with their usual ordening, and let ε be a nonnegative number. The first rule we shall discuss is

$$(i,j,k,l) \in L \iff \delta_{ij} - \delta_{kl} \geqslant \varepsilon \ . \tag{6}$$

If $\varepsilon=0$ and if $\delta_{ij} = \delta_{kl}$ then both (i,j,k,l) and(k,l,i,j) $\in$ L. If $\varepsilon > 0$ and $\delta_{ij} = \delta_{kl}$ then both (i,j,k,l) and (k,l,i,j) $\notin$ L. The interpretation is clear: ε is an estimate of the precision with which we have measured our dissimilarities and $\varepsilon = 0$ is the special case of perfect precision. The interpretation is reasonable but the choice of ε is in most cases rather arbitrary. If we have information about the standard error of the δ_{ij} (either by knowledge of the sampling distributions or by replications) then we can choose ε in a more or less rational way. Another rule is

$$(i,j,k,l) \in L \iff \delta_{ij} - \delta_{kl} \geqslant -\varepsilon \ . \tag{7}$$

In this case, if $-\varepsilon \leqslant \delta_{ij} - \delta_{kl} \leqslant +\varepsilon$, then evidently both (i,j,k,l) and (k,l,i,j) $\in$ L. All values between the boundaries are defined as ties and (as we shall see) must be represented by ties in the representation. The use of rule (7) seems hard to justify in most practical situations.

The treatment of ties is especially important in the analysis of adjacency matrices of graphs (as in sociometry). The entries of the matrix $D=\{\delta_{ij}\}$ are either zero or one, i.e. $\delta_{ij}=0$ iff subject a_i has 'chosen' subject a_j, otherwise $\delta_{ij} = 1$. In this case the rule must be

$$(i,j,k,l) \in L \Longleftrightarrow \delta_{ij} = 1 \wedge \delta_{kl} = 0 \wedge i = k \wedge l > j \qquad (8)$$

We use tie-rule (6) with $1 \leqslant \varepsilon < 0$. In this case the number of elements in L equals $\sum_i \{n\sum_j \delta_j^2 - (\sum_j \delta_j)^2\}$.

3. The algorithmic problem.

3.1 The nonmetric requirements

The algorithmic problem can now be stated as follows: the mapping ω must be found in such a way that

$$(i,j,k,l) \in L \Rightarrow d_{ij} \geqslant d_{kl} \qquad (9)$$

A more severe set of requirements would be

$$(i,j,k,l) \in L \Rightarrow d_{ij} \geqslant d_{kl} \qquad (10a)$$

$$(i,j,k,l) \in L \wedge (k,l,i,j) \notin L \Rightarrow d_{ij} > d_{kl} \qquad (10b)$$

But we shall see that our algorithm allows for cases with

$$(i,j,k,l) \in L \wedge (k,l,i,j) \notin L \wedge d_{ij} = d_{kl} \qquad (11)$$

The same thing is true for the KR- and the GL-approaches (of Roskam 1968 p43-45). The requirements in the measurement theory of NMS (of Beals, Krantz and Tversky 1968) are

$$\delta_{ij} \underset{\delta}{\geqslant} \delta_{kl} \Longleftrightarrow d_{ij} \geqslant d_{kl} \qquad (12)$$

If $\underset{\delta}{\geqslant}$ is a weak order and $\varepsilon = 0$ this is equivalent to (10a) and (10b). The NMS-algorithms, however, do not try to find representations in such a way that (12) is optimally satisfied. Moreover, the discussion in the previous section implies that cases should be possible with

$$\delta_{ij} = \delta_{kl} \wedge d_{ij} \neq d_{kl} \qquad (13)$$

(for example in the case of graphs).

Observe that our algorithmic requirements (9) and (10) both imply

$$(i,j,k,l) \in L \wedge (k,l,i,j) \in L \Rightarrow d_{ij} = d_{kl} \qquad (14)$$

In combination with tie-rule (7)

$$- \varepsilon \leq \delta_{ij} - \delta_{kl} \leq + \varepsilon \Rightarrow d_{ij} = d_{kl} \tag{15}$$

If $\varepsilon = 0$ both tie-rules reduce to

$$\delta_{ij} = \delta_{kl} \Rightarrow d_{ij} = d_{kl} \tag{16}$$

There are more subtle ways in which (9) requires ties in the representation. For example

$$\{(i,j,k,l),(k,l,i',j'),(i',j',i,j)\} \subseteq L \Rightarrow$$
$$d_{ij} = d_{kl} = d_{i'j'}. \tag{17}$$

In general: all cycles in the graph of the binary relation corresponding with L are to be mapped into equivalence classes of distances.

3.2 Choice of metric

We write $\omega(a_i) =_{\text{def}} (x_{i1},\ldots,x_{is},\ldots,x_{ip})$, where the x_{is} are real numbers. Define

$$d_{ij} = \left[\sum_{s=1}^{p} | x_{is} - x_{js}|^r\right]^R \tag{18}$$

If $r \geq 1$ and $R = r^{-1}$ then d is a metric by Minkovski's inequality. If $0 < r \leq 1$ and $R = 1$ then d also is a metric (Hardy, Littlewood and Polya 1952, p 30-32).

3.3 Formulation as a system of homogeneous inequalities

In the sequel we shall frequently use the m-element vector t. Suppose we have already chosen the number of dimensions p and the power $r > 0$. The vector t is defined in the following way: if $\lambda(i,j,k,l) = b$ then

$$t_b = \sum_{s=1}^{p} | x_{is} - x_{js} |^r - \sum_{s=1}^{p} | x_{ks} - x_{ls} |^r. \tag{19}$$

Because the positive power R in (18) clearly does not affect the ordening of the distances, we may reformulate our algorithmic problem as finding a representation of the set A in such a way that the m homogeneous inequalities

$$t_b \geq 0 \tag{20}$$

are optimally satisfied.

The inequalities (20) have an obvious trivial solution :set $x_{is} = \underline{a}$ for all $i=1,\ldots,m$; $s=1,\ldots,p$. The value of $\underline{a}$ is immaterial. Another trivial solution with all $t_b = 0$ can always be found in n-1 dimensions by defining

$$x_{is} = \delta^{is} \qquad i,s = 1,\ldots,n-1 \qquad (21)$$

(superscripted δ always denotes the Kronecker operator), and by setting x_{ns} equal to one of the roots of the equation

$$(n-2) \mid y \mid^{r} + \mid 1-y \mid^{r} -2 = 0 \qquad (22)$$

for all s. It is not difficult to see that if $r > 0$ and $n \geqslant 3$ then this equation always has a real root y_0 in the open interval (0,2). In the Euclidean case (r=2) the n points are the vertices of a regular simplex in n-1 dimensions. We have proved: if there is a p-dimensional representation of A which satisfies all m inequalities (20) and the additional conditio that for at least one s there are i,j, ε N such that $x_{is} \neq x_{js}$ and if no solution is possible in less than p dimensions then $p \leqslant n-1$. The example in which $\varepsilon = 0$ and $\underset{0}{\geqslant}$ is an equivalence relation that connects the set S defined in section 2.1 shows that no sharper inequality is possible. For more specialized situations we have the already famous SSA-I theorems of Guttman, which read in an adapted version: if the C(n,2) upper-diagonal dissimilarities are weakly ordered and we use ties-approach (6) with $\varepsilon > 0$, then an (n-2)-dimensional Euclidean representation with $t_b > 0$ for all b can always be found. If we set $\varepsilon = 0$ or use ties-approach (7) with $\varepsilon \geqslant 0$ then such a solution always exists in n-1 dimensions (Guttman 1968, p 476-477

4. Loss Function

4.1 The class of coefficients f(q)

A well-known method (cf for example Wolfe 1967 p 105) to find a solution of the m homogeneous inequalities $t_b \geqslant 0$ is to minimize the function

$$F = \sum_{b=1}^{m} (\max(0,t_b) - t_b)^{q}. \qquad (23)$$

with q =0. Evidently $t_b \geqslant 0$ for all b iff F =0. A more convenient way to write F is

$$F = 2^{-q} \sum_{b=1}^{m} (\mid t_b \mid - t_b)^{q} \qquad (24)$$

Because F = 0 iff t lies in the positive orthant of the m-dimensional space defined by the t_b we call this algorithm the positive orthant method (POM). If q = 2 it is identical to Guttman's absolute value principle as used, for example, in MSA - I and SSA - IV (Guttman,1969).

In the NMS-case we know that there always is a (trivial) solution with F = 0 in one dimension and another trivial solution with F = 0 in n-1 dimensions. In order to exclude these solutions we define the scaled loss-function

$$f(q) = \frac{\Sigma\ (|t_b| - t_b)^q}{2^q \Sigma |t_b|^q} \qquad (25)$$

Another advantage of this scaling is that f(q) is invariant under uniform stretching and shrinking of t, while F is not . In other words:we confine our attention to the points on the m-dimensional closed surface in t-space whose equation is $\Sigma|t_b|^q = 1$. If $q \neq 1$ the function f(q) is differentiable vis-a-vis the t_b. If q = 1 the derivatives fail to exist at a number of points, but in practical computational work this does not create any difficulties. Throughout this section we shall assume in our discussion of the f(q) that there is at least one $t_b \neq 0$.
Two important special cases of f(q) are

$$\Gamma =_{\text{def}} f(2) = \frac{\Sigma(|t_b| - t_b)^2}{4\ \Sigma\, t_b^2} \qquad (26)$$

$$\delta =_{\text{def}} f(1) = \frac{\Sigma\ (|t_b| - t_b)}{2\Sigma\ |\ t_b|} \qquad (27)$$

It is easy to show that (if M_- and M_+ are the subsets of M which $t_b < 0$ and $t_b > 0$ respectively)

$$f(q) = \frac{\sum_{b \varepsilon M_-} |\ t_b|^q}{\sum_{b \varepsilon M} |\ t_b|^q}, \qquad (28)$$

which implies that

$$0 \leqslant f(q) \leqslant 1. \qquad (29)$$

Moreover $f(q) = 0$ iff $t_b \geqslant 0$ for all t_b, $f(q) = 1$ iff $t_b \leqslant 0$ for all t_b, and $f(q) = \frac{1}{2}$ iff $\Sigma_{\overline{\overline{b\ \varepsilon\ M}}} |t_b|^q = \Sigma_{\overline{\overline{b\ \varepsilon\ M_+}}} |t_b|^q$. If we change all signs of the t_b, then the sum of the two corresponding values of $f(q)$ equals unity.

There is yet another instructive way to write the $f(q)$. In fact

$$f(q) = \frac{\Sigma |t_b|^q (1 - \text{sign}\ (t_b))^q}{2^q\ \Sigma |t_b|^q} \qquad (30)$$

By checking the possible values of sign (t_b) we see that

$$\Sigma\ |t_b|^q\ (1 - \text{sign}\ (t_b))^q = 2^{q-1} \Sigma |\ t_b|^q\ (1\text{-sign}(t_b)). \qquad (31)$$

Define

$$y_b = \tfrac{1}{2}\ (1\text{-sign}\ (t_b)). \qquad (32)$$

Substitution of (31) and (32) into (30) yields

$$f(q) = \frac{\Sigma |t_b|^q y_b}{\Sigma |t_b|^q} \qquad (33)$$

implying that $f(q)$ is a weighted mean of the values y_b. Of course

$$0 \leqslant y_b \leqslant 1 \qquad (34)$$

for all b.

A simple geometric interpretation may make the meaning of $f(q)$ even clearer. If t is a vector in Re^m, and $P(t)$ is the projection of the vector t on the positive orthant of that space, then

$$P(t) = \tfrac{1}{2}(|t| + t). \qquad (35)$$

Let S be the L_q-distance between the endpoints of t and $P(t)$, and let T be the L_q-norm of t. Then

$$f(q) = (\tfrac{S}{T})^q. \qquad (36)$$

In particular

$$\Gamma = \sin^2 \measuredangle\ (t, P(t)) \qquad (37)$$

The 2^m possible choices of the signs in t correspond with the 2^m possible faces of the positive orthant. Local minima may arise because we find a solution with $P(t)$ on a particular face of the orthant, while there may exist solutions with lower values of $f(q)$ that project on other faces Sydow, 1968).

.2 Limiting cases of f(q)

It is clear from (33) that minimizing f(q) with $q > 0$ means using a metric error theory. If $y_b = 1$ we make a nonmetric error, but this error is weighted by $|t_b|^q$, i.e. according to its size. The smaller we take q, the smaller the influence of this differential weighting of errors will be. The limit of f(q) for $q \rightarrow 0$ exists, and

$$\phi =_{\text{def}} \lim_{q \rightarrow 0} f(q) = \frac{m_-}{m_- + m_+} \tag{38}$$

where m_- and m_+ denote the number of elements in M_- and M_+. An algorithm that minimizes ϕ clearly uses a nonmetric error theory. Because ϕ is a step function, minimization of ϕ is impossible with a gradient-type algorithm. Minimization procedures that do not use derivatives have been tried out (De Leeuw, 1968c, p 49) but with little success, because the algorithms almost invariantly get stuck on one of the nonoptimal steps. By choosing q very small in (25) we approximate the step function ϕ by the continuous and differentiable function f(q), and the gradient methods can again be used.

At the other extreme end of the q-scale we find that the limit of f(q) for $q \rightarrow \infty$ also exists. If

$$|t_s| =_{\text{def}} \max_{b=1}^{m} (|t_b|) \tag{39}$$

and s_- and s_+ are respectively the number of negative elements and the number of positive elements of t whose absolute values equal $|t_s|$, then

$$\rho =_{\text{def}} \lim_{q \rightarrow \infty} f(q) = \frac{s_-}{s_- + s_+} \tag{40}$$

In most cases $|t_b| < |t_s|$ for all $b \neq s$ and

$$\rho = \begin{cases} 1 & \text{if } t_s < 0 \\ 0 & \text{if } t_s > 0 \end{cases} \tag{41}$$

Approximating a minimum of ρ can be done by taking q very large, but evidently ρ is a useless coefficient. Even in this extreme case, however, the requirements

$$\max_{b \varepsilon M_-} (|t_b|) < \max_{b \varepsilon M_+} (|t_b|) \tag{42}$$

still have some nonmetric characteristics.

Both limiting cases have some remarkable properties in common. In the first place, by (38) and (40), the ranges of both ϕ and ρ are the rational numbers between zero and one. In the second place the solutions of the minimization problems for ϕ and ρ can be expected to be far from unique. If we choose our representation at random we minimize ρ, for example, in approximately half of the cases. For ϕ the situation is much less serious and the diameter of the m-dimensional region M which $\phi \equiv \phi_{min}$ can be expected to be quite small. If we generate random representations the sampling distribution of ϕ is nearly normal with small variance and mean $\frac{1}{2}$, the sampling distribution of ρ is, of course, a two-point distribution.

Another advantage of using nonmetric error theories is that the statistical aspects seem to become somewhat more tractable. Consider the case in which $t_b \neq 0$ for all b, $\geq_c$ is a partial order, and we use tiesapproach (6). Define

$$T =_{def} 1 - 2\phi = \frac{m_+ - m_-}{m_+ + m_-} \tag{43}$$

Then T is a coefficient of disarray in the sense of Kendall (1962, ch 2). In fact, if $\geq_o$ is a weak order, T is identical with Kendall's familiar rank correlation coefficient and we may use the distributional theory of T to test the hypothesis that a particular value of ϕ could also have been found by chance. This, however, is not the statistical hypothesis we are interested in. The distribution of ϕ under random permutations of the d_{ij} is much less interesting than the distribution of ϕ_{min}, the value of the coefficient found after application of our algorithm under random permutations of the δ_{ij}. In general (for all q) we have that $E(f_{min}) = E(f)$ iff f is constant on the space of all possible vectors of distances. In all other cases $E(f_{min}) < E(f)$. The unlikely special case occurs, for example, if $(k, l, i, j) \varepsilon L$ iff $(l, j, k, l) \varepsilon L$.

In some cases it may be important to know how much we have gained by using a metric error theory. Define

$$\gamma =_{def} \frac{\frac{1}{m_+} \sum_{b \varepsilon M_+} |t_b|^q}{\frac{1}{m_-} \sum_{b \varepsilon M_-} |t_b|^q} = \left(\frac{\phi}{1 - \phi}\right)\left(\frac{1 - f(q)}{f(q)}\right) \tag{44}$$

Evidently

$$\lim_{q \to 0} \gamma = 1. \tag{45}$$

If γ is very large this indicates that the value of f(q) reflects mainly the size of the errors and not so much their number.

4.3 Inequalities relating the f(q)

For all q the function f(q) is a weighted mean of the y_b. It follows that f(q) - f(q') is a contrast of the y_b. There is an interesting inequality for contrasts that can be used in this case. If $\alpha = \Sigma w_i x_i$ and $\beta = \Sigma v_i x_i$ are weighted arithmetical means of the x_i, then

$$|\alpha - \beta| \leqslant \tfrac{1}{2} \max |x_i - x_j| \; \Sigma \; |w_i - v_i| \tag{46}$$

Proof: Define the m-element vector c by $c_i =_{\text{def}} w_i - v_i$. Then $\Sigma c_i = 0$. Let the c_i be arranged in such a way that $c_i \geqslant 0$ for i=1,...,k and $c_i < 0$ for i=k+1,...,m.

Then

$$\sum_{i=1}^{m} c_i x_i = \sum_{i=1}^{m} |c_i| x_i - \sum_{i=k+1}^{m} |c_i| \; x_i \leqslant \max_{i=1}^{k} (x_i) \sum_{i=1}^{k} | \; c_i| -$$

$$\min_{i=k+1}^{m} \; (x_i) \sum_{i=k+1}^{m} |c_i| = \tfrac{1}{2}(\max_{i=1}^{k}(x_i) - \min_{i=k+1}^{m} \; (x_i)) \sum_{i=1}^{m} | \; c_i| \leqslant$$

$$\tfrac{1}{2}(\max_{i=1}^{m}(x_i) - \min_{i=1}^{m} \; (x_i)) \sum_{i=1}^{m} |c_i| = \tfrac{1}{2} \max_{i,j=1}^{m} | \; x_i - x_j \; | \sum_{i=1}^{m} | \; c_i|.$$

We obtain equality in (46) iff there are constants c and d such that

$$\left.\begin{array}{ll} x_i = c \text{ or } 0 & i = 1,\ldots,k \\ x_i = d & i = k+1,\ldots,m \\ c \geqslant d & \end{array}\right\} . \tag{47}$$

In our case

$$\max_{b,d=1}^{m} |y_b - y_d| \leqslant 1 \tag{48}$$

so

$$|f(c) - f(q')| \leqslant \tfrac{1}{2} \sum_{b=1}^{m} \left| \frac{|t_b|^q}{\Sigma |t_b|^q} - \frac{|t_b|^{q'}}{\Sigma |t_b|^{q'}} \right| . \tag{49}$$

A necessary and sufficient condition for equality in (49) can be found by using (32) and (47): $\text{sign}(c_b) = \text{sign}(t_b)$ for all b. This case can be realized so no sharper inequality is possible.

An important special case is (assuming for the moment that $t_b \neq 0$ for all b, and setting $u_b =_{\text{def}} |t_b|^q$)

$$|f(q) - \phi| \leq \tfrac{1}{2} \Sigma \left| \frac{u_b}{\Sigma u_b} - \frac{1}{m} \right| = \tfrac{1}{2} \frac{\frac{1}{m}\Sigma \left| u_b - \frac{1}{m}\Sigma u_b \right|}{\frac{1}{m}\Sigma u_b} \,. \tag{50}$$

Of course this is a coefficient of variation. The numerator is the mean absolute deviation of the u_b, the denominator is the mean of these quantities. The more variation in the u_b, the more difference between f(q) and ϕ is possible. If we drop the assumption that $t_b \neq 0$ for all b then the upper bound is identical to the same coefficient of variation, but in this case computed for the nonzero elements of t. Of course we may combine (50) with the well-known fact that the mean absolute deviation $s_1(u)$ is not greater than the standard deviation $s_2(u)$ and obtain (with m(u) for the mean)

$$|f(q) - \phi| \leq \tfrac{1}{2} \frac{s_1(u)}{m(u)} \leq \tfrac{1}{2} \frac{s_2(u)}{m(u)} \,. \tag{51}$$

It also follows from inequality (50) that a sufficient condition for the equality of f(q) for all q is that all non-zero $|t_b|$ are equal. Moreover, if f(q) is equal to zero or unity for one $q < \infty$, then it is equal to zero or unity for all q.

4.4 Introduction of weights

An obvious generalization of f(q) is

$$f(q) = \frac{\Sigma w_b (|t_b| - t_b)^q}{2^q \; \Sigma w_b |t_b|^q} \,, \tag{52}$$

where the w_b are m nonnegative weights. These weights can be constructed in several ways. Suppose the responses of d different subjects generate d sets of quadruples $L_c \subseteq M^4$, c=1,...,d. Let ϕ_c be the indicators of these sets, i.e.

$$\phi_c \colon M^4 \to \{0,1\} \,. \tag{53}$$

define

$$w_{\lambda(i,j,k,l)} = \sum_{c=1}^{d} \phi_c(i,j,k,l). \tag{54}$$

This is a simple way to analyze the responses of a set of subjects. If we have numerical values for the dissimilarities we may introduce a metric element by using the weights in the following way

$$w_{\lambda(i,j,k,l)} = \delta_{ij} - \delta_{kl} \qquad (55)$$

4.5 Generalization by partitioning L.

Suppose P is a partitioning of L into k subsets $L_1, \ldots, L_k$. Of course $1 \leq k \leq m$. A generalization of f(q) is given by

$$f_p(q) = \frac{1}{k} \sum_{j=1}^{k} f_j(q) = \frac{1}{k} \sum_{j=1}^{k} \frac{\sum_{b \in L_j} (|t_b| - t_b)^q}{2^q \sum_{b \in L_j} |t_b|^q} . \qquad (56)$$

Suppose that, for all j,

$$\sum_{b \in L_j} |t_b|^q \neq 0. \qquad (57)$$

There are two 'extreme' cases. If k = 1 then $f_p(q) \equiv f(q)$, if k = m then

$$f_p(q) = \frac{1}{m} \sum_{b=1}^{m} \frac{(|t_b| - t_b)^q}{2^q |t_b|^q} = \frac{1}{m} \sum_{b=1}^{m} y_b = \phi \qquad (58)$$

because by (57) $t_b \neq 0$ for all b.

This generalization of f(q) can be expected to be useful in two different ways. In the first place $f_p(q)$ with an a priori defined partitioning of L can be used to circumvent the problem of defenerate solutions. An important example is multidimensional unfolding in which we have a conditional order on each row of the off-diagonal submatrix. If we use f(q) a degenerate solution in one dimension is always possible (cf Kruskal and Carroll 1968). If we let the order $>_i$ for each subject correspond with a subset of L, and we use $f_p(q)$ then such degenerate solutions are not possible. In this respect the generalization $f_p(q)$ is analogous to Roskam's modification of Kruskal's loss-function (Roskam 1968 p 34-35).

Suppose that we find a partitioning P such that all $|t_b|$ within each subset L_j are either equal to a constant c or equal to zero (not all of them zero, of course). Then $f_p(q) = \phi$ for this partitioning. This suggests a different method to minimize ϕ in which we change the partitioning of L. After each minimization M is repartitioned into groups in which we change the values of $|t_b|$ are approximate $f_p(q)$ of formula (58), i.e. ϕ. This method has not been tried out so far. Theoretically it seems less elegant than the other method we discussed, i.e. approximating ϕ by taking q very small. A third method was already tried out in De Leeuw (1968c p 20-21). Iterations are carried out on the weights in formula (52) by taking w_b equal to the values of $|t_b|^{-q}$, at the minimum in the previous iteration. The results were quite unsuccessfull.

5. The algorithm

5.1. Initial configuration

The first thing we need to get our algorithm going is an initial configuration (IC). The choice of an IC is very important because the function f(q) can be expected to have local minima. Numerical experiments with the MINISSA-program (Roskam 1969) show that local minima are very frequent indeed if we start with a completely arbitrary IC such as the one proposed by Kruskal. Roskam's results also indicate that if we start close to the absolute minimum of f(q), we shall probably stay away from these local minima. Our IC is based on the rationale of the canonical discriminant analysis of relational data or CDARD -series (De Leeuw, 1968 b), a slight modification and considerable extension of some ideas of Guttman (1941, 1946, 1959). It provides solutions to nonmetric problems by computing principal components.

Define the m x n matrices S and T in the following way: if $\lambda(i,j,k,l) = b$ then

$$s_{bq} = \delta^{qi} - \delta^{qj} - \delta^{qk} + \delta^{ql}, \qquad (59)$$

and

$$t_{bq} = \delta^{qi} - \delta^{qj} + \delta^{qk} - \delta^{ql}. \qquad (60)$$

If x is an n-element vector, then the m- element vectors Sx and Tx contain, respectively, the elements $(x_i - x_j) - (x_k - x_l)$ and $(x_i - x_j) + (x_k - x_l)$ on the appropriate places. Define $g(x) = x'T'Sx$, then

$$g(x) = \sum_{\overline{(i,j,k,l)\epsilon L}} (x_i-x_j)^2 - (x_k-x_l)^2. \qquad (61)$$

The relation with (19) for $r = 2$ is obvious. Moreover

$$x'T'Sx = x' \{(T'S + S'T)/2 + (T'S - S'T)/2\} \; x =$$

$$= x' \{(T'S + S'T) /2 \}x \qquad (62)$$

because $T'S - S'T$ is skew symmetric, which implies that $x'(T'S - S'T)x = 0$ for all x. Define the Lagrangian function

$$\Phi_A(x) = x'T'Sx - \mu(x'x - 1). \qquad (63)$$

Symbolic differentation with respect to all elements of x simultaneously gives

$$\frac{\partial \Phi \; (x)}{\partial x'} = (S'T + T'S)x - 2 \; \mu x \qquad (64)$$

Finding the extreme values of $x'T'Sx$ means solving the eigenproblem

$$\{(S'T + T'S)/2 \}x = \; \mu x \qquad (65)$$

Of course, by (62), we could also have defined

$$\Phi_B(x) = x' \{(S'T + T'S)/2 \}x \; - \mu(x'x - 1) \qquad (66)$$

with exactly the same result.

It is quite easy to see that S and T are composed of two m x n matrices, say S_1 and S_2, with $S = S_1-S_2$ and $T = S_1+ S_2$. Evidently

$$S'T = S'_1S_1- S'_2S_1 + S'_1 S_2 - S'_2S_2 \; , \qquad (67a)$$

$$T'S = S'_1S_1 - S'_1S_2 + S'_2S_1 - S'_2S_2 \; , \qquad (67b)$$

so

$$Q =_{def} \tfrac{1}{2}(S'T+T'S) = S'_1S_1 - S'_2S_2. \qquad (68)$$

Because the row sums of both S_1 and S_2 disappear, S'_1S_1 and S'_2S_2 are doubly centered and consequently, so is Q. It follows that Q is singular, so at least one of the eigenvalues equals zero. We may order the eigenvalues:

$$\mu_1 \geq \mu_2 \geq \ldots \geq \mu_k = 0 \geq \cdots \geq \mu_n. \qquad (69)$$

The eigenvector associated with μ_k has constant elements. Define

$$g_p(x) = \sum_{s=1}^{p} \mu_s =$$

$$= \sum_{(i,j,k,l)\varepsilon L} \left[\sum_{s=1}^{p} (x_{is}-x_{js})^2 - \sum_{s=1}^{p} (x_{ks}-x_{ls})^2 \right] \qquad (70)$$

which makes the relation with (19) even more obvious.

If it is true for all $(i,j,k,l)\varepsilon L$ that $i \neq j$ and $k \neq l$, then both S_1 and S_2 have two nonzero elements in each row, one of them equal to +1, the other to -1. It follows that

$$\text{trace } (S'_1S_1) = \text{trace } (S'_2S_2) = 2m, \qquad (71)$$

so trace (Q) = 0, and the sum of the eigenvalues equals zero.

For each eigenvector with nonvanishing eigenvalue we obtain for its contribution to the squared Euclidean distances

$$\sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} (x_i - x_j)^2 = 2n \sum_{i=1}^{n} x_i^2 \qquad (72)$$

This means that maximization of $x'S'Tx$ under the condition that $\Sigma\Sigma\, d_{ij}^2$ is some constant value would give us the same results. This fact can be used to relate the coefficients μ and $f(1) = \delta$ in the Euclidean case. Suppose we scale each eigenvector (corresponding with a positive eigenvalue) in such a way that the sum of squares of its elements equals the eigenvalue. In that case, by (72),

$$\sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij}^2 = 2n \sum_{s=1}^{n} \mu_s . \qquad (73)$$

By (70)

$$\sum_{b=1}^{m} t_b = \sum_{s=1}^{p} \mu_s^2 \qquad (74)$$

From (27) it follows that

$$1-2\delta = \frac{\Sigma\, t_b}{\Sigma |t_b|} . \qquad (75)$$

We consider three different cases. In the first case all upper-diagonal dissimilarities are strictly ordered, $m = D_n$. Each squared distance d_{ij}^2 with $j > i$ appears $\frac{1}{2}\, n(n-1)-1 = \frac{1}{2}(n-2)(n+1)$ times in the m instances of formula (19). By applying Minkovski's inequality

$$\sum_{b=1}^{m} |t_b| \leq \tfrac{1}{2}(n-2)(n+1) \sum_{i} \sum_{<j} d_{ij}^2 = \tfrac{1}{4}(n-2)(n+1) \Sigma\Sigma\, d_{ij}^2 . \tag{76}$$

Combination of (73)-(76) gives uw the result

$$\delta \leq \frac{n(n-2)(n+1)\Sigma\,\mu_s - 2\Sigma\,\mu_s^2}{2n(n-2)(n+1)\Sigma\,\mu_s} \tag{77}$$

In the conditional case, in which $m = \frac{1}{2}n^2(n-1)$, Minkovski's inequality gives us

$$\Sigma|t_b| \leq (n-1)\, \Sigma\Sigma\, d_{ij}^2 \tag{78}$$

and we obtain in a similar way

$$\delta \leq \frac{2n(n-1)\ \Sigma\mu_s - \Sigma\mu_s^2}{4n(n-1)\ \ \Sigma\mu_s} \tag{79}$$

For the method of triads

$$\Sigma|t_b| \leq (\ n-2)\Sigma\Sigma d_{ij}^2\ , \tag{80}$$

and

$$\delta \leq \frac{2n(n-2)\ \Sigma\mu_s - \Sigma\mu_s^2}{4n\ (n-2)\ \Sigma\mu_s} . \tag{81}$$

Observe that the inequalities (77) and (79) remain valid if there are ties and we use tie-rule (6) with $\varepsilon > 0$, because this simply implies that we delete some of the elements of L. The left side of the inequalities (76) and (78) decreases, the right side remains constant. Moreover the discussion in this section implies that (if $Q \neq 0$) there always exists a one-dimensional Euclidean solution of our NMS-problem for which $\delta < \frac{1}{2}$.

Our IC is a generalization of the IC used in the GL-SSA-I and MINISSA-I programs (Guttman 1968 p 499-502, Roskam and Lingoes 1969). To prove this we apply our procedure to the complete case in which all n^2 dissimilarities are ordered. Because of the definition of Q it is immaterial whether we use

$$\delta_{ij} = \delta_{kl} \Rightarrow (i,j,k,l) \in L \wedge (k,l,i,j) \in L, \tag{82}$$

or

$$\delta_{ij} = \delta_{kl} \Rightarrow (i,j,k,l) \notin L \wedge (k,l,i,j) \notin L \tag{83}$$

for the tied elements. Let a_{ij} denote the number of elements of Δ less than δ_{ij}, and b_{ij} the number of larger elements. Clearly $a_{ij} + b_{ij} \leqslant n^2-1$. If e_i is the n-element unit vector, the vector e_i-e_j appears a_{ij} times as a row of S_1 and b_{ij} times as a row of S_2. The vector e_j-e_i appears a_{ji} times in S_1 and b_{ji} times in S_2. It follows that the elements of $C^{(1)} =_{\text{def}} S_1'S_1$ are given by

$$c_{ij}^{(1)} = \delta^{ij} \sum_{k=1}^{n} (a_{ik} + a_{kj}) - (a_{ij} + a_{ji}), \tag{84}$$

and the elements of $C^{(2)} =_{\text{def}} S_2'S_2$ by

$$c_{ij}^{(2)} = \delta^{ij} \sum_{k=1}^{n} (b_{ik} + b_{kj}) - (b_{ij} + b_{ji}). \tag{85}$$

So

$$q_{ij} = \delta^{ij} \sum_{k=1}^{n} \{ (a_{ik}-b_{ik})+(a_{kj}-b_{kj})\} - \{(a_{ij}-b_{ij})+(a_{ji}-b_{ji})\} . \tag{86}$$

In Guttman's technique the matrix

$$\tilde{C} = C - 2\bar{D}\,(nI-J) \tag{87}$$

is factored. In (87)

$$c_{ij} = \delta^{ij} \sum_{k=1}^{n} (D_{ik} + D_{kj}) - (D_{ij} + D_{ji}), \tag{88}$$

J is an n x n matrix with all elements equal to unity, the D_{ij} are arbitrary numbers with the same rank order properties as the δ_{ij}, and $\bar{D}$ is the mean of the D_{ij}. Formula (86) can be simplified by letting r_{ij} denote the rank number of δ_{ij} (as usual, for tied elements the relevant rank numbers are averaged), and by using the identities

$$a_{ij}-b_{ij} = \sum_{l=1}^{n} \sum_{v=1}^{n} \text{sign}\,(\delta_{ij} - \delta_{lv}) = 2r_{ij} - (n^2+1). \tag{89}$$

Substitution of (89) into (86) and simplification yields

$$\tfrac{1}{2}q_{ij} = \delta^{ij} \sum_{k} (r_{ik}+r_{kj})-(r_{ij}+r_{ji}) -$$

$$\delta^{ij} n(n^2+1) + (n^2+1). \tag{90}$$

Substituting (86) with $D_{ij} = r_{ij}$ in (90) gives

$$\tfrac{1}{2}Q = C - (n^2 + 1)(nI - J). \tag{91}$$

Because $2\bar{r} = n^2+1$ we finally have

$$\tfrac{1}{2}Q = \tilde{C} \tag{92}$$

if we use $D_{ij} = r_{ij}$. This proves that the two techniques are essentially identical if applied to an important special case. In the case of missing data we obtain again formula (92), where $\tilde{C}$ is now defined by Guttman's formula (106). In the case of (complete) conditional matrices (86) is valid again but r_{ij} of(89) becomes the rank number of δ_{ij} in the relevant row (or column). The formulas for the rectangular case can be easily obtained by using Guttman's notation for missing data.

We have obtained two significant extensions of Guttman's results. In the first place a generalization of the IC to an arbitrary binary relation over Δ, in the second place an optimal property (within the framework of maximizing g_p) of the r_{ij}, compared with other possible choices of D_{ij}. Both Guttman's technique and ours give us a prelimary upper bound for p_{opt}, the optimal dimensionality. Because g_p will increase as long as we add new dimensions with positive eigenvalues, a reasonable upper bound is the number of positive eigenvalues of Q.

Our IC can be modified to permit the introduction of weights. They must be collected in a diagonal matrix W of order m, and

$$Q = S_1'WS_1 - S_2'WS_2. \tag{93}$$

This generalization is used in the CDARD-7 program that constructs an optimal representation of the cognitive space by using the dissimilarity estimates of a number of subjects at the same time, with values of g(x) computed for each dimension and each subject separately. The NMSEMS-program (De Leeuw 1968c,p 11-14) is a more simple version, where the weights must be part of the input (EMS stands for Euclidean maximum sum, because in (19) r=2 and the program maximizes the sum of the elements of t). The NMSPOM-algorithm, described in this paper, uses the EMS-solution as an initial configuration, the NMSEMS-program gives it as an independent solution of the multidimensional scaling problem. In general the EMS-rationale provides us with a very good IC (even if r≠2 and q≠1), and in some cases the results are even more satisfactory than those obtained with NMSPOM. The reasons are clear: NMSEMS computes the absolute maximum of the function $g_p(x)$

for all p. In fact, because the eigenvectors of $\tilde{C}$ are the same as those of C (Guttman 1968 p 502) and because C is Gramian (Guttman 1968 p. 493), maximizing $g_p(x)$ is equivalent to maximizing a concave function. Moreover, in cases where a so-called degenerate solution (Roskam 1968 p 43-45, p 61-64, p 122-126) is possible, NMSEMS may be much better than NMSPOM. Nevertheless at the moment I do think it wise to include NMSEMS in the class of NMS-algorithms, because there seems to be at least two necessary conditions for calling an algorithm nonmetric. In the first place the results must be invariant under a strictly monotonic increasing transformation of the data. This is true for NMSEMS, as it is true for all CDARD-type techniques. But in the second place the algorithm must minimize a loss-function with an a priori known lower bound, and the coefficient must equal its lower bound if and only if all nonmetric restraints are nontrivially satisfied. This second condition is not met by the CDARD-techniques (and of course this is exactly the reason why they are more robust against inherent degeneracy in the data). It is true of CDARD that the loss-functions have a lower bound, and that a necessary (but not sufficient) condition for attaining this lower bound is that all nonmetric contraints are satisfied. The same thing is true, however, for maximizing the PM-correlation between d_{ij} and δ_{ij}: if $r(\delta,d) = 1$ then $\delta_{ij} \geq \delta_{kl}$ iff $d_{ij} \geq d_{kl}$. Techniques for which only the first criterion is satisfied will be called semi-metric. The most important fact that follows from both theoretical and computational developments in psychological measurement and scaling theory since the book of Coombs, is that the additional requirement that the representation must be nonmetric as well is unnecessary restrictive (Shepard 1966).

5.2 Variance algorithm

Both the KR- and GL-methods use a simple gradient (SG) algorithm to minimize their DPF-coefficients. An alternative is to use revised gradient (RG) algorithms such as those of Fletcher and Powell (1963) and Fletcher and Reeves (1964). These methods have a faster eventual convergence because they converge for a quadratic function in a finite number of steps(not greater than the number of variables), and because each twice-differentiable function is dominated in the neighbourhood of the minimum by the second order terms in its

Taylor expansion. Moreover the RG-methods provide additional useful information about the curvature of the function at the minimum. The prize we must pay for these advantages is an increased amount of computational work in each iteration and an increase of the memory space required for the storage of the relevant arrays. Comparative studies of the RG- and SG-methods have been carried out by Box (1966), and, for a more specific psychometric problem, by Jöreskog(1967)(cf also Jöreskog 1966). The conclusion from both studies is that in general RG-methods were superior, where superiority is defined in terms of quite a number of different criteria. In the Jöreskog study the only serious rival of the RG-methods was a Gauss-Seidel-type technique, but algorithms of this kind can only be applied in very specific cases and even in these cases the algorithm must be rederived for each new problem (Harman and Jones 1966 is also relevant in this context).

In our NMSPOM-program we use a recent RG-algorithm due to Davidon (1968). It is called the variance algorithm and it is comparable to the older variable metric algorithm devised by Davidon (1959) and perfected by Fletcher and Powell (1963). Both algorithms build up an estimate of the inverse matrix of second partial derivatives of the function (because of asymtotic maximum likelihood theory this matrix is called the variance matrix by Davidon). Explicit calculation of the second order derivatives is extremely time-consuming in the NMS-case, so we have preferred this approach to the classical Newton-Raphson method. We have selected the variance algorithm rather than the variable metric algorithm because the former is, at least theoretically, about twice as efficient as the latter.

The reasons to use an RG-method (and not an SG-method) at all in the NMSPOM-case are following. The relative efficiency of RG compared with SG increases if the main burden of the computational work in each iteration is the calculation of the function values and the gradient. In other words, if updating the estimate of the variance matrix uses only a small portion of the time required for each iteration. In NMSPOM this definitely is the case because of the large number of elements in L. A second reason is that our IC enables us to start close to the minimum we are looking for. It is well-known that SG-methods are quite efficient if the function is still far from its minimum, but that convergence in the neighbourhood of the minimum is very slow. It can be improved by modifications such as subrelaxation, diagonal-step, and partan (cf Wolfe 1967), but it cannot be remedied completely. RG-methods are definitely superior in the neighbourhood of the minimum.

For a detailed development and theoretical analysis of the variance algorithm we refer the reader to Davidon's paper. For our purposes it suffices to know that the algorithm produces a nonincreasing sequence of function values, that it computes a matrix that converges to the variance matrix, and that it provides us with an estimate of the excess of the function value in a particular iteration above its minimum value.

5.3. Derivatives

Of course the POM-approach can be used for all nonmetric measurement models based on differentiable functions. Therefor we develop the derivatives of f(q) in two steps. The first formula can be used for any measurement model. Let

$$T =_{\text{def}} \Sigma |t_b|^q \qquad (92)$$

and suppose for the moment that $q \neq 1$, $r \neq 1$ (again, in actual computation these assumptions cause no loss of generality). Then

$$\frac{\partial\ f(q)}{\partial\ t_b} = \frac{q}{2T} ((1 - 2f(q))\ \text{sign}(t_b) - 1)\ |t_b|^{q-1}. \qquad (93)$$

The second formula is for the specific NMS-case: if $\lambda(i,j,k,l)=b$ then

$$\frac{\partial t_b}{\partial x_{gh}} = r \left[\text{sign}\ (x_{ih} - x_{jh})|\ x_{ih} - x_{jh}|^{r-1}\ (\delta^{ig} - \delta^{jg}) - \right.$$
$$\left. \text{sign}\ (x_{kh} - x_{lh})|\ x_{kh} - x_{lh}|^{r-1}\ (\delta^{kg} - \delta^{lg}) \right]. \qquad (96)$$

The formula's must be combined by using

$$\frac{\partial f(q)}{\partial x_{gh}} = \sum_{b=1}^{m} \frac{\partial f(q)}{\partial t_b} \cdot \frac{\partial t_b}{\partial x_{gh}} \cdot \qquad (97a)$$

It is interestingto observe that a sufficient condition for the numerator of f(q)to be a convex function in the variables x is that t_b is a concave function of x for all b. Because t_b is the difference of values obtained by applying the same functional form to two different subsets of the variables, this functional form must be both concave and convex, i.e. linear. This may be an explanation of the fact that the algorithms for additive conjoint measurement using the POM-approach always seem to converge to the absolute minimum (cf also Roskam 1968 p 41). The same thing should be true for scalogram analysis, nonmetric discriminant analysis, and nonmetric multiple regression. Derivatives for the generalizations of f(q) in sections 44 and 45 are easily found. Suppose there is a partitioning of L into d subsets, then

$$\frac{\partial f_p(q)}{\partial x_{gh}} = \frac{1}{d} \sum_{c=1}^{d} \sum_{b \varepsilon L_c} \frac{\partial f_c(q)}{\partial t_b} \cdot \frac{\partial t_b}{\partial x_{gh}} \qquad (97b)$$

5.4. A matrix interative process

In the Euclidean case Guttman (1968) managed to rewrite the stationary equations that give necessary conditions for an extreme value in the form AX = X. This was done, of course, for the Kruskal-Guttman type of approach only. In this section we show that the same thing can be done for the POM-method. We use Guttman's definition of the signature of the data (Guttman 1968,1969). The signature is a function δ_{ijkl} defined on x with

$$\delta_{ijkl} = \begin{cases} -1 & \text{if } \delta_{ij} < \delta_{kl} \\ +1 & \text{if } \delta_{ij} > \delta_{kl} \\ 0 & \text{otherwise} \end{cases} \qquad (98)$$

Using this function t can be redefined as

$$t_{ijkl} = \delta_{ijkl} \, (d^2_{ij} - d^2_{kl}), \qquad (99)$$

and f(q) as

$$f(q) = \frac{\sum_i \sum_j \sum_k \sum_l (|t_{ijkl}| - t_{ijkl})^q}{2^q \sum_i \sum_j \sum_k \sum_l |t_{ijkl}|^q} . \tag{100}$$

This formulation is more elegant than the one we used in the previous sections, but special provisions must be made if δ_{ij} and δ_{kl} are compared more than once (as in the method of triads). For all these cases Guttman has developed a notation that fully deserves to become standard. Now

$$\frac{\partial f}{\partial x_{gh}} = \frac{q}{2T} \sum_i \sum_j \sum_k \sum_l ((1-2f)\mathrm{sign}(tijkl)-1).$$

$$|t_{ijkl}|^{q-1} \frac{\partial t_{ijkl}}{\partial x_{gh}}. \tag{101}$$

In the Euclidean case

$$\frac{\partial t_{ijkl}}{\partial x_{gh}} = 2\sigma_{ijkl}(x_{ih}-x_{jh})(\delta^{ig}-\delta^{jg}) -$$

$$2\sigma_{ijkl}(x_{kh}-x_{lh})(\delta^{kg}-\delta^{lg}). \tag{102}$$

Substituting (102) into (101) and a considerable amount of algebraic manupulation gives

$$\frac{\partial f}{\partial x_{gh}} = \sum_{j=1}^{n} (x_{gh}-x_{jh})c_{gj}, \tag{103}$$

with

$$c_{ij} = \frac{4q}{T} \sum_k \sum_l \sigma_{ijkl}|t_{ijkl}|^{q-1} ((1-2f)\mathrm{sign}(tijkl)-1). \tag{104}$$

A necessary condition for an extreme value is, according to (103)

$$\sum_j c_{gj}x_{jh} = x_{gh} \sum_j c_{gj}. \tag{105}$$

If we collect the row sums of C in the diagonal matrix D, this suggests the iterative process

$$X_{(s+1)} = D^{-1}_{(s)} C_{(s)}X_{(s)}. \tag{106}$$

In this paper the process (106) is introduced as a curiosity. We do not investigate the deeper properties of C and D any further.

6. Comparison with other algorithms

6.1 Data

The NMSPOM-program can handle a more general type of data than both the KR- and GL-series. From the program description of Lingoes(1968) it can be seen that NMSPOM can be used to analyze the data for SSA-I, SSA-I , SSAR-I tot SSAR-V, and MSA-II. The prize we must pay for this generality is that the data must be coded in the form of a large number of quadruples, i.e. (n-2) (n+1) times as many numbers as for the comparable GL-or KR-program. This made it necessary to write a special program, SMPUNCH, that punches all quadruples to be included in L.The program allows for various options: the numerical matrices can be interpreted as conditionally or weakly ordered, off diagonal or complete, symmetric or asymmetric. The diagonal elements can be included in or excluded from Δ, the ties approach used can be either (6) or (7). The difference between our tie-rules (6) and (7) can be interpreted as a difference in the definition of a tie in the data structure. A tie is defined as a member of the binary relation $=_1$ over Δ. If we use ties-approach (6) with $\varepsilon > 0$ then $=_1$ is empty. If we use ties approach (7) with $\varepsilon > 0$ then

$$\delta_{ij} =_1 \delta_{kl} \Longleftrightarrow -\varepsilon \leqslant \delta_{ij} - \delta_{kl} \leqslant +\varepsilon . \qquad (107)$$

Observe that in this case $=_1$ is not necessary an equivalence relation (it is reflexive and symmetric, but it may be intransitive). If $\varepsilon = 0$ both approaches define

$$\delta_{ij} =_1 \delta_{kl} \Longleftrightarrow \delta_{ij} = \delta_{kl} . \qquad (108)$$

6.2 Algorithmic problem.

Both Kruskal and Guttman define δ_{ij} and δ_{kl} as tied if $\delta_{ij} = \delta_{kl}$. Kruskal considers two different sets of algorithmic requirements (1964 a,p22). The primary approach is defined by

$$\delta_{ij} > \delta_{kl} \Rightarrow d_{ij} \geq d_{kl} \tag{109}$$

The secondary approach requires (98) in conjunction with

$$\delta_{ij} = \delta_{kl} \Rightarrow d_{ij} = d_{kl}. \tag{110}$$

Using the definition of $=_1$ in the previous section, we may formulate our algorithmic requirements (for numerical dissimilarities) as

$$\begin{cases} \delta_{ij} > \delta_{kl} \Rightarrow d_{ij} \geq d_{kl}, & \text{(111a)} \\ \delta_{ij} =_1 \delta_{kl} \Rightarrow d_{ij} = d_{kl}. & \text{(111b)} \end{cases}$$

The primary approach corresponds with (6) and $\varepsilon > 0$ (which implies that $=_1$ is empty). The secondary approach with (6) with $\varepsilon = 0$, or with (7) with

$$0 < \varepsilon < \min_{i,j,k,l} |\delta_{ij} - \delta_{kl}| \cdot \tag{112}$$

The main practical difference between our treatment of ties and that of Kruskal is that, once we have defined what elements are to be considered as tied, no further preprocessing is required in each iteration. We use

$$\delta_{ij} =_1 \delta_{kl} \Rightarrow (i,j,k,l) \in L \wedge (k,l,i,j) \in L \tag{113}$$

only once: in the construction of L.

A very important aspect of both the KR-and the POM-algorithms is that they may tie distances even if the corresponding pair of dissimilarities is not an element of $=_1$. They do not require the reverse implication in (99) or (100b). Guttman tries to circumvent this possibility, but we shall try to show that the devices he uses are theoretically unsound and based on a nonrigourous formulation of the algorithmic problem. The requirements (100) are called the weak monotonicity requirements by Guttman.

He also considers semi-strong monotonicity. In the complete case this requires

$$\delta_{ij} > \delta_{kl} \Longleftrightarrow d_{ij} > d_{kl} \tag{114}$$

whenever $i \neq j$ and $k \neq l$. For tied elements of Δ there are no requirements. Strong monotonicity corresponds with

$$\delta_{ij} \geqslant \delta_{kl} \Longleftrightarrow d_{ij} \geqslant d_{kl} \tag{115}$$

whenever $i \neq j$ and $k \neq l$. It is easy to see that (104) implies both (103) and

$$\delta_{ij} = \delta_{kl} \Longleftrightarrow d_{ij} = d_{kl} \tag{116}$$

whenever $i \neq j$ and $k \neq l$. Verbally strong monotonicity requires that tied distances must only occur on the places where we want them to occur.

Now suppose that there are v dissimilarities in Δ. We reformulate the algorithmic problem by introducing v additional variables e_l $(l=1,\ldots,v)$. The problem is to find a representation of A (for given o and r) and a set of numbers e_l such that

$$S = \frac{\Sigma(d_l - e_l)^2}{\Sigma d_l^2} \tag{117}$$

is minimized, subject to the conditions

$$\begin{cases} \delta_l > \delta_k \Rightarrow e_l \geqslant e_k, & (118a) \\ \delta_l =_1 \delta_k \Rightarrow e_l = e_k, & (118b) \end{cases}$$

where $=_1$ may, of course,be empty. If there is a configuration $\omega(A)$ with distances $\tilde{d}_l$ and a set of numbers $\tilde{e}_l$ such that this problem is solved, then $\tilde{e}_l$ also solves the problem: find numbers e_l such that

$$S_{\tilde{d}} = \frac{\Sigma\ (\tilde{d}_l - e_l)^2}{\Sigma\ \tilde{d}_l^2} \tag{119}$$

is minimized, subject to the conditions (107). Because $S_{\tilde{d}}$ is a convex function of the e_l only two cases are possible: either $\tilde{e}_l = \tilde{d}_l$for all l=1,...,ν, or $\tilde{e}_l$ lies on one of the boundaries of the convex defined by (107a). If we consider the conditions (103) or (104) then a necessary condition for these conditions to hold is that min(S) = 0. In general, of course, min(S) > 0

In the POM-case we may force strict monotonicity by requiring

$$\delta_{ij} > \delta_{kl} \Rightarrow t_{\lambda(i,j,k,l)} \geq \varepsilon_2, \quad (120a)$$

$$\delta_{ij} =_1 \delta_{kl} \Rightarrow t_{\lambda(i,j,k,l)} = 0, \quad (120b)$$

where ε_2 is a positive constant. A similar device is used by Guttman (1968 p 496). The choice of ε_2 is of course extremely arbitrary. If we consider ε_2 as an additional variable, then it follows from our discussion that min(f(q)) > 0 implies that $\varepsilon_2 = 0$ at the minimum.

The e_l can also be interpreted as a function of the d_l (and, thus, as a function of the x_{is}). In fact, if $\hat{e}_l$is the set of real numbers that minimizes

$$S^* = \Sigma(d_l - e_l)^2, \quad (121)$$

under the conditions (107), then each of the $\hat{e}_l$ is related to the d_l by a continuous function (Van Eeden, 1958, ch I). Guttman's rank-image transformational principles use the additional requirement that e_l must be a permutation of the elements of d_l. The set of values e_l^* that results is not continuously related to the d_l. Moreover

$$S^*(\hat{e}) \leq S^*(e^*). \quad (122)$$

The interrelations between the two approaches were thoroughly analyzed by Roskam(1969a,b). For this purpose he uses Guttman's distinction between soft-squeeze and hard-squeeze approaches. The soft approach concentrates on the minimization of S^* given by formula (121), the hard approach on the minimization of S. Moreover a distinction can be made between split-step and joint-step algorithms. They both start with an initial configuration $X^{(0)}$, and corresponding distances $d_l^{(0)}$ and values of $\tilde{e}_l$ (either $\hat{e}_l$ or e_l^*). In the split-step algorithms the first step is minimization with respect to the x_{is}, for given (fixed) values of $\tilde{e}_l$. The minimization problem does not have to be solved completely, it is sufficient that the value of S is decreased. If the first split is finished, we compute new values of d and $\tilde{e}$ (second split), and repeat the first split using the new $\tilde{e}$. Split-step algorithms correspond with the well known relaxation methods in which a subset of the variables is held constant during minimization. If we have reached the minimum another subset is held constant, and so on. A well known example is the Gauss-Seidel method. The joint-step algorithms also minimize S (or S^*), but now the $\tilde{e}_l$ are considered as functions of the x_{is}. In the almagamation case the derivatives of the $\hat{e}_l$ vis-à-vis the x_{is} vanish (essentially because the $\hat{e}$ are constructed by averaging the d), in the rank image case the derivatives do not vanish in general (at some places they do not exist, and they most certainly are not continuous). This has unpleasant consequences for the convergence properties of the rank-image process.

6.3 Loss Functions

We continue our comparative analysis with a discussion of the different loss-functions. Of course all coefficients have the property that they attain their lower bound (zero) if and only if all nonmetric constraints are (nontrivially) satisfied. Another desirable property would be: the loss-functions have a priori known upper bound that is attained if and only if all nonmetric constraints are violated. Our coefficients f(q) have this property for all $q < \infty$. They share it with all members of the class of generalized correlation coefficients (Kendall 1962, ch 2). The KR- and GL- coefficients do not have this property. For the GL-approach we have (for normalized phi)

$$S_\phi = \frac{\Sigma(d_l - e_l^*)^2}{2\ \Sigma d_l^2} = 1 - \frac{d'Pd}{d'd}, \qquad (123)$$

where P is a permutation matrix. For the upper bound we have (for a given set of weakly ordered distances with $d_k \leqslant d_l$ if $k < l$)

$$S_\phi \leqslant 1 - \frac{\sum_{l=1}^{v} d_l d_{v-l+1}}{d_l^2} \leqslant 1. \qquad (124)$$

(cf Hardy et al, 1952, p 267).

This means that the upper bound depends on the values of d, which is not very elegant.

Kruskal's original loss-function (stress-one) is

$$S_1 = \surd \left[\frac{\Sigma \, (d_l - \hat{e}_l)^2}{\Sigma d_l^2} \right] \qquad (125)$$

For a given set of d_l the maximum value of S_1 is given by

$$S_1^{(max)} = \surd \left[\frac{\Sigma \, (d_l - \bar{d}\,)^2}{\Sigma \, d_l^2} \right] \qquad (126)$$

The maximum value of $S_1^{(max)}$ is reached for an arbitrary partial order if $d_k = a$ for one k, and d vanishes for the other v-1 elements. This situation can be realized, and we have

$$S_1^{(max)} \leqslant \surd \left(\frac{v-1}{v} \right) . \qquad (127)$$

For the complete case the maximum is attained if $x_k = a$ for one k, and x_i vanishes in the other n-1 cases. Then

$$S_1^{(max)} \leqslant \surd \left[\frac{n-2}{n} \right] \qquad (128)$$

The upper bound given by (128) is sharper than that given by (127), as it should be (substitute v = C(n,2) in (127)). The fact that the upper bound depends on n may explain one of the results of Young's Monte Carlo studies (Young 1968). He found that the expected value of S_1 increased with the number of points for fixed p and level of error. Observe moreover that the upper bound given by (126) is a coefficient of variation, because it is the quotient of the standard deviation and the root mean square of the elements of d.

In the complete case the minimum value of $S_1^{(max)}$ in the one dimensional case is attained for $x_i = i(i=1,\ldots,n)$. In that case

$$S_1^{(max)} = \surd \left[\frac{n^2 + 2}{3n^2} \right] \qquad (129)$$

It follows that, if $S_1^{(min)}$ is the value obtained by applying the KR-algorithm to the complete case, then

$$S_1^{(min)} \leqslant \frac{1}{3} \surd 3 \qquad (130)$$

This may be an explanation of a result of Wagenaar and Padmos (1968) who invariantly found $S_1^{(\min)} \leqslant \frac{1}{2}$ in extensive Monte Carlo studies (of also Klahr 1969). The approach used in finding inequality (129) can be generalized In the same way we can prove that the minimum value of S_1 in two dimensions (no matter what the order-relations are) is given by

$$S_1^{(\min)} \leqslant \surd\left(\frac{\Pi^2 -8}{\Pi^2}\right) \qquad (131)$$

Kruskal's modified loss-function (stress two) is

$$S_2 = \surd\left[\frac{(d_l - \hat{e}_l)^2}{(d_l - \bar{d}\ )^2}\right]. \qquad (132)$$

For any set of distances $S_2^{(\max)} = 1$. Observe that

$$S_2 = \frac{S_1}{S_1^{(\max)}} \qquad (133)$$

where $S_1(\max)$ is given by (126). A sufficient (but not necessary) condition is that all nonmetric constraints are violated. In general S_2 seems more satisfactory than S_1. Because

$$\Sigma\hat{e}_l = \Sigma\ d_l, \qquad (134)$$

S_2^2 is the ratio of the residual variance after fitting the model to the total variance of the d, S_2 is a coefficient of alienation, and $\surd(1-S_2^2)$ can be interpreted as a measure of (monotonic) correlation. Observe that S_2 excludes the regular simplex solution discussed in section 3.3.

6.4 Algorithm

For a discussion of the convergence of the procedures we must distinguish the various special cases. It is, of course, always true that the loss-functions S and S^* are bounded below by zero. Both the hard joint and the hard split almagamation processes produce (with suitable choices of step-size) a non-increasing sequence of S-values, which means that they converge to a value $\tilde{S}$. Because the derivatives exist **and are** continuous everywhere (Kruskal 1969) this value corresponds with at least a local minimum.The same thing is true for the POM-approach. The rank image processes do not always produce a nonincreasing sequence, and the derivatives are not continuous, which means that the familiar theorems on the convergence of gradient methods cannot be used. In fact Roskam(1969a) did indeed find that the GL-SSA programs do not converge.

In the soft squeeze almagation processes convergence to a local minimum of S^* is also assured, but this value may very well correspond to a trivial solution. In general soft- squeeze processes do not minimize the complete loss-funtion S (cf De Leeuw 1968a).

Some theoretical considerations of Guttman (1968) and some numerical results of Roskam (1969b) indicate that split step-algorithms are less vulnerable to local minima than joint-step algorithms. The same considerations and results indicate, however, that the nature of the IC has a more profound influence on the freuqency of local minima. The one used by Guttman is very good indeed, and it can even be improved upon a good deal by using some soft-split procedure. The discussion in this section implies that we must always end our iterative process with hard joint almagamation. The soft split rank image options can be looked upon as devices that can be employed to perfect the IC. And this is exactly the way they are used in Roskam's MINISSA programs.

In the case that perfect solutions are possible the rank image principle is more likely to produce strictly monotonic solutions (Roskam 1969a,c, Lingoes & Roskam 1970), but this is not very important. If such a perfect solution is possible we are sure that there actually is a set of perfect solutions and we need additional criteria to make a selection from this set (such as the maximin approach of Abelson and Tukey, or the pr nciple of least squares).

A degenerate solution arises, for example, if A can be partitioned into two subsets A_1 and A_s and the following condition is also satisfied. Let $\psi_{ij} = 1$ iff a_i and a_j are both elements of the same subset and $\psi_{ij} = 0$ otherwise. If

$$\psi_{ij} = 1 \wedge \psi_{kl} = 0 \Rightarrow (i,j,k,l) \notin L \tag{135}$$

for all i,j,k,l ε N then a degenerate solution in one dimension is possible. If a and b are two real numbers, $a \neq b$, and

$$a_i \in A_1 \Rightarrow x_i = a \tag{136a}$$

$$a_i \in A_2 \Rightarrow x_i = b \tag{136b}$$

then f(q) vanishes for all q (and so do the loss- funtions for the KR- and GL- approaches). In the NMSPOM-case with q >>0 we may expect to reach the degenerate solution quite fast, because of the weighting of errors. The negative elements in t will soon be made very small. In minimizing f(q) with q very small the successive iterations will probably use a route with more acceptable solutions. Using NMSEMS may be a satisfactory way out. The same thing is true for a reanalysis of A_1 and A_2 separately. These separate analysis may be carried out by partitioning the elements of L into those that correspond with within-group-distances and those that correspond with between-group-distances. The weights for the second set are set equal to zero. A much more elegant solution is using $f_p(q)$ on this partitioned set L. In NMSEMS the first dimensions invariantly contrasts the groups, and the additional dimensions can be expected to give information comparable to the separate analysis of A_1 and A_2. It is of interest to note that Roskam(1968 p 45, also 1969a) found that the GL-programs produce trivial solutions whenever the KR-programs do so. This means that the GL-algorithms essentially produce weakly monotonic solutions too (at least in the critical cases).

In order to compare the POM-approach with the other algorithms we rewrite the loss-function as follows

$$f(q) = \frac{\Sigma|\tilde{t}_b - t_b|^q}{\Sigma|t_b|^q} \qquad (137)$$

where $\tilde{t}_b$ must be nonnegative for all b. If we substitute

$$\tilde{t}_b = \tfrac{1}{2}(|t_b| + t_b) \qquad (138)$$

into (137) we obtain our original formula (25). This formulation of the problem makes it possible to distinguish between hard-soft and joint-split in the POM-method too. The algorithm outlined in De Leeuw (1968a) for nonmetric discriminant analysis is a soft-squeeze POM-process, the algorithm in this paper a hard-squeeze joint-step POM-process. An important observation in this context is that in the linear case (De Leeuw 1969) we do not need the distinction between split and joint, both minimization problems can be solved in one step, i.e. without using gradient methods at all

There is a remarkable conceptual correspondence between the KR- and the POM-approaches. The main difference is that the two algorithms work in a different space. In the KR-case each distance defined a dimension, the nonmetric restraints define a polyhedral convex cone in this space, $\hat{d}$ is the projection of d on the cone, and

$$S_1 = \sin \sphericalangle \; (d,\hat{d}). \qquad (139)$$

(In the S_2 case the vectors d and $\hat{d}$ must be replaced by $d - \bar{d}$ and $\hat{d} - \bar{d}$). In the POM-case each difference of distances defined a dimension, the nonmetric constraints again define a polyhedral convex cone in this space (the positive orthant), $\tilde{t}$ given by formula (138) is the projection of t on the cone, and

$$\Gamma = \sin^2 \sphericalangle \; (t,\tilde{t}). \qquad (140)$$

In which of the two spaces we prefer to work depends on the nature of the data, more precisely on what we consider as 'observed'. If we observe differences of distances we may prefer the POM-space, if we observe numerical similarities we may prefer the KR-space (cf also Roskam 1969 c,p 21).

The KR-method can be easily generalized to partial orders. We have done this by solving the quadratic programming problem (107),(110) by the Hildreth-d'Esope method. By using the familiar Birkhoff-theorem on the decomposition of doubly stochastic matrices into a weighted mean of permutation matrices we may show that finding the optimal rank-image transformation for a partial order reduces to a linear programming problem. In this sense the rank-image transformation is easier to find, even in the general case. Both the KR- rnd the GL- approach are faster than the POM-approach, but they may be called less subtle because they do not include the possibility of weighting each inequality separately and approximating a nonmetric error theory. The KR-approach is, however, a theoretically sound and completely acceptable alternative. In my opinion the GL-approach has no (theoretical) advantages whatsoever over the KR-approach, except for the initial configuration, which can, of course, also be used in the KR-algorithm.

The CH-method cannot be considered a serious alternative any more, because of two reasons. In the first place it takes too much time, because it must be done by hand. In the second place the representation is far from satisfactory. In NMS, even more so than in factor analysis, we must pay attention to the structure of the configuration and not to the projections on arbitrary axis (Guttman, 1966,1967).

The conclusion from this comparative analysis is clear: for each problem separately we have to make a choice between the KR-approach with loss-function S_2 and NMSPOM. Which one of the two we choose depends on the number of elements in L, on the nature of $\underset{0}{\geq}$, on the value we attach to a nonmetric error theory, and possibly also on other considerations. If the programs produce a degenerate solution we may even prefer NMSEMS.

7. Remaining problems

7.1. Algorithmic problems

There are a number of problems that must be solved before we can make an optimal use of the variance algorithm. In the first place we want to find an optimal way to choose the initial estimate of the variance matrix, $V^{(0)}$. This may have a considerable effect on the speed of convergence (Davidon 1968 p 409, cf also Jöreskog 1967). The variance algorithm has two additional parameters α and β , $0 < \alpha < 1 < \beta$, that regulate the rate of change in the variance estimate within one iteration. In the current version of NMSPOM we have taken as an initial estimate of the variance matrix a scalar matrix: $V^{(0)} = \theta I$, so the first iteration is a steepest descend step. Moreover we have chosen $\alpha = 10^{-3}$, and $\beta = 10$. The problem is to find the optimal choice of α, β and θ.
Another problem is the influence of the value of q on the convergence of the procedure. We may of course expect that taking q very small means that convergence will be less smooth and more difficult (because f(q) behaves very much like a step function). A related problem is to find the optimal way to minimize ϕ. We may choose q very small at the outset, but a safer procedure seems to be to start with a relatively large value of q (because of the initial configuration it may be wise to choose q around unity), obtain the optimal representation for this value of q and use this configuration (and possibly also its variance matrix) as an initial estimate for iterations with a smaller value of q. This possibility is build into the NMSPOM-program (as is the possibility of keeping any part of the configuration constant throughout the iterations). Of course the question whether the variance algorithm itself is an optimal choice from the class of minimization algorithms is also an algorithmic problem. In earlier versions of the NMSPOM-program we used SG-methods.

Some of the examples in the appendix are computed with these earlier versions. It is very hard to compare the computational efficiency of SG and RG in this case, because of two reasons. In the SG-case we used q=2 throughout with fixed binary exponentiation, in the RG-programs q was made a (short) float decimal variable. Moreover we made no attempt to optimize step-size procedures in the SG-case, nor did we experiment with the RG-parameters α and β.

7.2 Comparative problems.

Another class of problems compares the results om NMSPOM with those of other approaches (this may include approaches to NMS we did not discuss in this paper such as those of Shepard, McGee, Sydow, and Young-Torgerson). We may use the output of each of the programs as initial configuration for the other programs and see how there loss-functions are related, how well the output of approach A approcimates a minimum in approach B, etcetera. In particular we may compute the values of ϕ at the minimum of each of the different loss-functions and compare them. This was already done by Coombs (1966) for one single example. One of our relevant results is, for example, that if we use the KR-approach in additive conjoint measurement, start with metric (least squares) estimates of the parameters, and compute the calue τ of (43) in each iteration, then the optimal solution in KR-terms often has a τ- value that is slightly less than that of initial configuration. Other examples (from nonmetric discriminant analysis) also indicate that minimizing Γ can easily result in an increase of ϕ (a number of small errors is created to eliminate one large error).

7.3 Determinateness problems.

We start with known underlying configurations with varying number of points and dimensions, compute the distances and construct L. In addition we may add various degrees of random error before we compute the distances used to construct L. Then apply the algorithm and compute a measure of correspondance between the error-free distances and the recovered distances (the cosine between the two vectors seems an appropriate measure, more so than the correlation). Studies of this kind have already been carried out by Young (1968), and Sherman and Young (1968) for the complete case. More restricted studies (error-free data, two dimensions) had already been carried out by Shepard (1966), while Green and Maheshwari (1969) investigated the conditional case. All studies used stress one, which is unfortunate.

An important measure in the context is the completeness index η defined as

$$\eta = \frac{m}{D_n} . \tag{141}$$

Obviously for small values of η, our solution will be less determinate. For the complete case without ties $\eta = 1$ for the method of triads

$$\eta = \frac{4}{n+1} , \tag{142}$$

and for the conditional case (without ties)

$$\eta = \frac{4n}{(n+1)(n-2)} . \tag{143}$$

In the Green and Maheshwari paper the effect of the number of ties elements in each row of a conditional matrix was studied (using the TORSCA-program with the primary approach to ties). In NMSPOM the ways to vary the value of η are of course more subtle than in other approaches. Moreover we may study the effect of q on the determinateness of our solutions.

A related problem is the effect of upgrading the data on the solutions. In the case of an arbitrary binary relation we may input the corresponding L, but we may also find the best fitting partial order first. Partial orders may be converted into weak orders by a technique such as $CDARD_2$ (De Leeuw 1968b), by using Goode's Δ-method (Coombs 1964), or Abelson and Tukey's maximin approach (Abelson and Tukey 1963). Conditional matrices can be made symmetric by averaging. Responses in the method of triads can be put in numerical form by summing over subjects. In the Green and Maheshwari paper the effect of doing things like this was studied on a small scale.

Another very important problem is more or less specific for NMSPOM.

In the case of a consistent partial order some of the elements of L are 'inessential', because the fact that they are included follows by transitivity from the inclusion of other 'essential' quadruples. If the essential elements of t are nonnegative, then it follows that the inessential elements are nonnegative too. Perfect solutions for the complete L are also perfect solutions for the reduced L. In the compete case without ties the number of essential quadruples is C (n,2) - 1, which is very much less than D_n (this suggests an alternative measure of completeness: the ratio of the number of essential elements in L to C(n,2)-1). We are interested in the effect of deleting inessential quadruples from L on the solution, because this reduction of the number of elements in L considerably decreases the time needed for each iteration. Moreover the effect can be expected to be influenced by the value of q, although for all values of q it is true that the values of f(q) for the reduced set are not a monotonic function of those for the complete set.

There is yet another way to study the degree in which nonmetric constraints determine the configuration. Suppose we start with an arbitrary configuration X and a set of requirements which are satisfied by this configuration. There exists a set of configurations (in Re^{np}) that satisfies these requirements. How large is this set? Of course we limit our discussion to configurations whose centroid is the origin, and whose root mean square distance to the origin is some constant value. In the case that r = 2 we may restrict the set even further by requiring that the dimensions coincide with the principal components. Te answer this question we do not need any NMS-algorithm. Some relevant theoretical work has been done by Abelson and Tukey (1963), and Benzécri (1964,1965). The complete answer can be given by the complete description method of De Leeuw (1970).

7.4 Statistical problems.

The most interesting statistical problem for us is the distribution of the minimum value of f(q) for fixed values of η, n, p, and q. In the complete case (η = 1) this was already investigated (for stress one) by Wagenaar and Padmos (1968) for n = 7 (1) 11 and p = 1 (1) 5. Stenson and Knoll (1969) investigated the expected value of S_{min} for n = 10(10)60 and p = 1(1)10. Klahr (1969) repeated the Wagenaar - Padmos study over a somewhere wider range of n and p. The principal result from these studies is that it is useless to devide the range of the loss-functions into intervals and to label these as 'fair-to-good','excellent','poor', etcetera. The signigicance levels were strongly dependent upon the values op p and n.

7.5 Conclusion.

All four classes of problems have one remarkable aspect in common: they must be solved by extensive sampling experiments. The reason is, of course, that a purely mathematical approach to their solution (although theoretically possible) very soon becomes extremely complicated, and, somewhat later, even prohibitive. Of course ultimately a rigorous theoretical solution must always be preferred to a set of Monte Carlo results, but in some cases one of the alternatives is simply unfeasible. Using the Monte Carlo method costs an enormous amount of machine time, but at least it can be done.

REFERENCES

Abelson, R.P. and Tukey, J.W.
Efficient utilization of nonnumerical. information in quantitative analysis : general theory and the case of simple order.

Beals,R.W., Krantz,D.H., and Tversky,A.
The foundations of multidimensional scaling
Psych. Rev., 1968,75,127-142.

Benzécri, J.P.
Analyse factorielle des proximités.(I and II)
Publ.de l'institute de statistique de l' université de Paris.
I : 1964,13, 235-282.
II: 1965,14, 65- 80.

Box, M.J.
A comparison of several current optimization methods.
The Computer Journal, 1966, 9, 67 -

Coombs, C.H.
A theory of data.
New York, Wiley, 1964.

Coombs, C.H.
General mathematical psychology, chapter scaling and data theory.
Unpublished manuscript, Univ. Michigan, May 1966.

Davidon, W.C.
Variable metric method for minimization.
Argonne Natl. Lab. Report 5990,1959.

Davidon, W.C.
Variance algorithm for minimization.
The Computer Journal, 1968, 11, 406-410.

De Leeuw, J.
Nonmetric discriminant analysis
Department of data theory for the social sciences, University of Leiden, Research Note RN 006-68 (a).

De Leeuw, J.
Canonical discriminant analysis of relational data.
Department of data theory for the social sciences, University of Leiden, Research Note RN 007-68 (b).

De Leeuw, J. Nonmetric multidimensional scaling.
Department of data theory for the social sciences, University of Leiden, Research Note RN 010-68 (c)

De Leeuw, J. The linear nonmetric model.
Department of data theory for the social sciences, University of Leiden, Research Note RN 003 - 69.

De Leeuw, J. The complete description method for nonmetric Euclidean scaling.
Department of data theory for the social sciences, University of Leiden, Research Note RN 003-70.

Fletcher, R., and Powell, M.J. A rapidly convergent descend method for minimization.
The Computer Journal, 1963, 6, 163-168.

Fletcher, R., and Reeves, C.M. Function minimization by conjugate gradients.
The Computer Journal, 1964, 7, 149-154.

Gleason, T.C. A general model for nonmetric multidimensional scaling.
Michigan Mathematical Psychology Program, MMPP 67 - 3.

Green, P.E., and Maheshwari,A. A note on the multidimensional scaling of conditional proximity data.
Unpublished manuscript, Univ. Pennsylvania,1969.

Guttman, L. The quantification of a class of attributes: a theory and method of scale construction.
In: Horst,P.(ed): The prediction of personal adjustment.
New York, Social Science Research Council, 1941, 253 - 312.

Guttman, L. An approach for quantifying paired comparisons and rank order.
Ann. Math. Statist., 1946, 17, 144-163.

Guttman,L.
An introduction to facet design and analysis.
In: Proceedings of the 15th international
congress of psychology.
Brussels, 1959, 130-132 (abstract).

Guttman, L.
Order analysis of correlation matrices.
In: R.B. Cattell(ed) : Handbook of multivariate
experimental psychology.
New York, Rand McNally, 1966, 438-458.

Guttman,L.
The development of nonmetric space analysis:
letter to professor John Ross.
Multivariate Behavioral Research, 1967,3, 71-83.

Guttman, L.
A general nonmetric technique for finding the
smallest coordinate space for a configuration
of points.
Psychometrika, 1968, 33, 469 - 506.

Guttman, L.
Smallest space analysis by the absolute
value principle.
Paper presented at the symposium on 'Theory
and practice of measurement' at the XIXth
international Congress of Psychology, London,
August 1, 1969.

Hardy, G.H., Littlewood, J.E.,
and Polya, G.
Inequalities.
Cambridge University Press, 1952.

Harman,H.H. and Jones, W.H.
Factor analysis by minimizing residuals.
Psychometrika, 1966, 31, 443-482.

Jöreskog, K.G.
Testing a simple structure hypotheses in
factor analysis.
Psychometrika, 1966, 31, 165-178.

Jöreskog, K.G.
Some contributions to maximum likelihood
factor analysis.
Psychometrika, 1967, 32, 443-482.

Kendall, M.G.
Rank correlation methods.
London, Griffin, 1962.

Klahr,D.
A Monte Carlo investigation of the statistical significance of Kruskal's nonmetric scaling procedure.
Psychometrika, 1969, 34, 319-330.

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

Kruskal, J.B.
Nonmetric multidimensional scaling : a numerical method.
Psychometrika, 1964, 29, 115-129 (b).

Kruskal, J.B.
Monotone regression: continuity and differentiability properties.
Unpublished mimeographed report, Bell Telephone Labs, Murray Hill, N.J., 1969.

Kruskal, J.B., and Carroll, J.D.
Geometric models and Badness-of-Fit Functions.
In: P.R. Krishnaiah(ed). Multivariate Analysis II
New York, Academic Press, 1969, 639-671.

Lingoes, J.C.
New computer developments in pattern analysis and nonmetric techniques.
In: Uses of computers in Psychological research-The 1964 IBM symposium of Statistics.
Paris: Gauthier-Villars, 1966, 1-22 (a).

Lingoes, J.C.
Recent computational advances in nonmetric methodology for the behavioral sciences.
In: Proceedings of the International Symposium: Mathematical and Computational Methods in Social Sciences.
Rome, International Computation Centre, 1966, 1-38 (b).

Lingoes, J.C.
The rationale of the Guttman-Lingoes nonmetric series: a letter to doctor Philip Runkel
Multivariate Behavioral Research, 1968, 3, 495-509.

Roskam, E.E.C.I. Metric analysis of ordinal data in psychology.
Doctoral dissertation, University of Leiden,1968.

Roskam, E.E.C.I. A comparison of principles for algorithm construction in nonmetric scaling.
Michigan Mathematical Psychology Program,
MMPP 69 - 2(a).

Roskam, E.E.C.I. Results of MINISSA-I with random data.
Unpublished mimeographed paper, University of Nijmegen, 1969 (b).

Roskam, E.E.C.I. Data theory and algorithms for nonmetric scaling(Parts I and II).
Unpublished mimeographed paper, University of Nijmegen. 1969 (c).

Roskam, E.E.C.I., and Lingoes,J.C.MINISSA-I: A FORTRAN IV (G) program for the smallest space analysis of square symmetric matrices.
Behavioral Science, 1969, 14, (in press).

Shepard, R.N. Attention and the metric structure of the stimulus space.
J. Math. Psychol., 1964, 1, 54-87.

Shepard, R.N. Metric structures in ordinal data.
J. Math. Psychol., 1966, 3, 287-315.

Sherman, C.R., and Young,F.W. Nonmetric multidimensional scaling: A Monte Carlo study.
Proceedings 76th annual convention, APA, 1968, 207-208.

Stenson, H.H., and Knoll, R.L. Goodness of fit for random rankings in Kruskal's nonmetric scaling procedure.
Psychol. Bull., 1969, 71, 122-127.

Sydow, H. Ähnlichkeitsskalierung mittels allgemeiner Minkovskiräume.
Paper read at the 2nd annual meeting of the Gruppe Mathematische Psychologie, Marburg/Lahn, April 1968.

Van Eeden, C.
Testing and estimating ordered parameters
of probability distributions.
Unpublished Doctoral Dissertation, University
of Amsterdam, 1958.

Wagenaar, W.A. and Padmos, P.
The significance of a stress percentage
obtained with Kruskal's multidimensional
scaling technique.
Institutefor perception RVO/TNO, Soesterberg
The Netherlands, Report IZF 1968-22.

Wolfe, P.
Methods of nonlinear programming.
In: Abadie,J. (ed): Nonlinear programming.
Amsterdam, North Holland Publishing Company,1967.

Young, F.W.
Nonmetric multidimensional scaling: development
of an index of metric determinacy.
The L.L. Thurstone Psychometric Laboratory,
Univ. N. Carolina, report 1968-68.

Appendix I: Corrections

p. 7, line 12	For '(of Kendall...' read '(cf Kendall ...'.
p. 7, line 14	For '(of Kruskal...' read '(cf Kruskal...' .
p. 8, line 6	For '$1 \leq \xi < 0$' read '$0 \leq \xi < 1$'
p. 8, line 18	For '(of Roskam...' read '(cf Roskam...'.
p. 8, line 19	For '(of Beals...' read '(cf Beals...'.
p.10, line 28	For '$q=0$' read '$q > 0$'
p.12, line 4	For '$\sum_{b\in M}\lvert t_b\rvert^q$' read '$\sum_{b\in M-}\lvert t_b\rvert^q$'
p.15, line 23	For '$\lvert f(c)-f(q')\rvert$' read '$\lvert f(q)-f(q')\rvert$'.
p.18, line 24	The sentence beginning with 'After each...' must be replaced by:' After each minimization M is repartitioned into groups in which the values of $\lvert t_b\rvert$ are approximately equal. By making the partitionings finer and finer we approximate $f_F(q)$ of formula (58), i.e. ϕ.'
p.22, line 25	For '(86)' read '(88)'
p.24, line 28	For 'DPF coefficients' read 'loss functions'
p.29, line 7	For 'SSA-I,SSA-I,' read 'SSA-I, SSA-II'
p.30,line 5	For'(98)' read '(109)'
p.30, line 21	For '(99)' read '(110)0
p.30, line 22	For '(100b)' read (111b)'
p.30, line 24	For '(100)' read '(109)'
p.31, line 7	For'(104)'read '(115)'
	For '(103)' read'(114)'
p.31, line 14	For '(for given o and r)' read '(for given n and r)
p.32, line 1	For '(107)' read '(118)'
p.32, line 3	For 'convex defined' read 'convex region defined'
	For '(107a)' read '(118a)'
p.32, line 4	For '(103)' read '(114)'
	For '(104)' read '(115)'
p.32, line 16	For '(107)' read '(118)'
p.33, line 27	In the numerator of formula (124) 'd_e^2' should read 'Σd_e^2'
p.35,line 4	For 'S_1' read 'S_1^{max}'
p.36, line 28	For 'A_s' read 'A_2'
p.38, line 15	For '(107)' read '(108)'
	For '(110)' read '(121)'

Appendix I (vervolg)

p.38, line 21	For'rnd' read 'and'
p.40, line 1	There is no appendix with examples. Some of the **SG** examples can be found in RN010-68.
p.40, line 20	For'calue' read 'value'.

Appendix II : Addenda:

II.1:

The main parts of this report were written about two years ago. This implies that it is not representative any more for the way in which I write on NMS. One of the things that must be changed is the notation Guttman's signature notation is much more compact and much easier to manipulate as my $(i,j,k,l)\varepsilon L$. And in the second place the terminology has been improved in my more recent report 'The Euclidean distance model' and in an unpublished paper 'The abstract structure of scaling theories'. Those reports will be referred to in this appendix as EDM and ASST.

II.2:

EDM has a new derivation for the EMS-rationale, which uses signature notation and is, consequently, more elegant.

II.3:

In EDM we have derived very sharp versions of the SSA-I theorems. The method of proof is similar to the method used in Lingoes' recent MMPP-report, though more compact and elegant. A more general and more elegant version is possible, using the geometrical language of EDM. A semimetric is the samething as a metric, only it is not necessary that the triangle inequality is satisfied for all triples of points. In the space of all real symmetric n x n matrices the semimetrics are a polyhedral convex cone P.

The matrix D is defined by $d_{ij} = 1-\delta^{ij}$, the ray T_d by $T_d=\{X|X = \lambda D \wedge \lambda \geq 0\}$ Moreover the ε ts E_p^q are those real symmetric matrices whose elements can be interpreted as q^{th} powers of the Euclidean distances of a configuration in p dimensions. Our representation theorem covers the case in which there are p square matrices A_i, q square matrices B_j and vectors c and d with, respectively, p and q elements. The system of linear inequalities and equations

$$\mathrm{Tr}(A_iX) \geq c_i \qquad i=1,\ldots, p$$

$$\mathrm{Tr}(B_jX) = d_j \qquad j=1,\ldots,q$$

has solution set $S = S_X \cap S_B$ with S_X a convex polyhedron and S_B a subspace. Moreover $\bar{S}_A$ is the convex closure of S_A, and $Q = \{X \mid \mathrm{Tr}(B_jX) = 0;\ j=1,\ldots,q\}$.

Vervolg II.3

<u>Theorem</u>: If $S \cap P \neq \phi$, $D \in \bar{S}_A \cap Q$, then $E'_{n-1} \cap S \neq \phi$. If in addition $S \neq T_d$ then $E^2_{n-2} \cap S \neq \phi$

II.4:

Formula (131) is the limit for $n \to \infty$ of

$$S_1^{(\min)} \leqslant \sqrt{\left(1 - \frac{2\cot^2\frac{\Pi}{2n}}{n(n-1)}\right)},$$

which is the value of $S_1^{(\max)}$ for n points equally spaced on a circle.

II.5.

If we apply partitioning to Kruskal's S_2 for the case of paired comparisons of pairs of distances, then

$$S_2 = \frac{1}{N}\sum \frac{(d_{ij} - \hat{d}_{ij})^2 + (d_{kl} - \hat{d}_{kl})^2}{\frac{1}{2}(d_{ij}-d_{kl})2} = \frac{M-}{M_+ + M_-} = \phi$$

if $d_{ij} \neq d_{kl}$ for all i,j,k,l.

II.6

A better definition of f(q) would be

$$f(q) = \left[\frac{\Sigma\Sigma\Sigma\ (|t_{ijkl}| - t_{ijkl})^q}{2^q\ \Sigma\Sigma\Sigma|t_{ijkl})^q}\right]^Q$$

with $Q = 1$ if $q < 1$ and $Q = 1/q$ otherwise. Of course $t_{ijkl} = \sigma_{ijkl}(d_{ij}^2 - d_{kl}^2)$

This improves the limiting behavior for $q \to \infty$

$$\lim_{q \to \infty} f(q) = \frac{\max\ (|t_{ijkl}| - t_{ijkl})}{2\ \max\ |t_{ijkl}|}$$

Consequently $f(\infty) = 0$ iff $t_{ijkl} \geqslant 0$ for all i,j,k,l and > 0 for at least one, $f(\infty) = 1$ iff there is a $t_{ijkl} < 0$.

Thus $f(\infty) = 0$ iff we have a perfect nontrivial solution, and $f(\infty) = 1$ iff we have a nonperfect,nontrivial solution.

In the terminilogy of ASST $f(\infty)$ is a binary loss function

II.7:

Let

$$S_{\phi}^{q} = \left[\frac{\Sigma\Sigma e_{ij}(\quad \phi\, (d_{ij}) - \widehat{\phi(d_{ij})})^{q}}{\Sigma\Sigma\; e_{ij}(\phi(d_{ij}) - \overline{\phi(d_{ij})})^{2}} \right]^{Q}$$

with ϕ strictly monotone and $\phi(0) = 0$, and let F_{ϕ}^{q} be defined as $f(q)$ with $t_{ijkl} = t_{ijkl}(\phi(d_{ij}) - \phi\,(d_{kl}))$.

Theorem: $F_{\phi}^{q} \leqslant S_{\phi}^{q}$