Japanese Psychological Research
1978, Vol. 20, No. 1, 7-17

A MAXIMUM LIKELIHOOD METHOD FOR NONMETRIC MULTIDIMENSIONAL SCALING: I. THE CASE IN WHICH ALL EMPIRICAL PAIRWISE ORDERINGS ARE INDEPENDENT—THEORY

YOSHIO TAKANE[1]

Department of Psychology, McGill University

A maximum likelihood estimation procedure is developed for nonmetric multidimensional scaling (MDS) which applies to the situation in which all empirical pairwise orderings of dissimilarities are assumed to be independent. The proposed method, while formulated within Thurstonian framework, does not presuppose initial unidimensional scaling of " observed " distances. Rather, the original nonmetric data (which are the set of empirical ordinal relations on the dissimilarities between stimuli) are directly related to the representation model (which is a distance function of some form) through a single optimization criterion based on the maximum likelihood principle.

There are various experimental procedures in which ordinal dissimilarity measures arise. Three broad categories of experimental operations will be distinguished for the present discussion.

One is the pair comparison type methods in which subjects are required to make a set of pairwise ordinal judgments of dissimilarities between pairs of stimuli. The methods of tetrads and the method of triads (Torgerson, 1952, 1958) are in this category. The two methods are distinguished in terms of the sets of pairs of dissimilarities on which empirical orderings are defined.

Another class of methods is the rating type methods, such as the method of successive categories (Messick, 1956). Rating scales could be combined with the pair comparison type judgments providing a method which might be appropriately called the successive categories method of tetrads following Sjöberg's (1967) successive categories method of pair comparisons.

The rank-order type methods constitute yet a third class of empirical operations. This includes the simple rank-order method, and the conditional rank-order method (Klingberg, 1941; Young, 1975) and the method of triadic combinations (Richardson, 1938) as two special cases of the (partial) rank-order method.

The above distinctions are made, despite the presence of certain intriguing relationships among these methods, in view of the ease with which the statistical independence among observations may be obtained. For example, in the method of tetrads it is possible, at least in principle, to obtain statistically independent observations (i.e., independent pairwise orderings), while in the rank-order type methods order statistics are, totally or partially, dependent with each other. Particular dependence structures assumed of observations (or of statistics derived from observations) play a crucial role in

[1] The author wishes to express his appreciation to Drs. Forrest W. Young and Elliot M. Cramer of the University of North Carolina for their helpful comments. Portions of the work were done while the author was at the University of Tokyo and at the L. L. Thurstone Psychometric Laboratory, the University of North Carolina. Requests for reprints should be addressed to Yoshio Takane, Department of Psychology, McGill University, 1205 McGregor Avenue, Montreal, Quebec, Canada, H3A 1B1.

the specification of the likelihood function.

In this paper we are primarily concerned with the development and evaluation of a maximum likelihood (ML) estimation procedure for nonmetric multidimensional scaling (MDS) specifically designed for the first type of empirical operations for obtaining ordinal dissimilarities. In subsequent papers we consider some plausible extensions of the current approach to other experimental situations in which data are collected.

The now classical approach to ordinal empirical dissimilarities in MDS, as most notably exemplified by Torgerson (1952), is to scale the original observations into " observed " distances. Those " observed " distances are then subjected to a metric MDS method to find a spatial representation. A possible conceptual difficulty with this approach is that the initial transformation of the observed ordinal relations is performed under an assumption which has nothing to do with the representation model of the data. A least squares (LS) criterion is set up for the estimation of initial " observed " distances, and another for the spatial representation. There are no logical connections bridging the two LS criteria.

The advent of nonmetric MDS (Shepard, 1962; Kruskal, 1964) partially resolved the problem. As is well known, Shepard-Kruskal type nonmetric MDS utilizes only ordinal information in finding a spatial representation, and yet does not require preprocessings of ordinal information into " observed " distances. In fact, the gist of the Shepard-Kruskal type of nonmetric procedures is that they consider both the optimal transformation of data (under certain measurement restrictions) and the optimal estimation of model parameters based on a single optimization criterion.

However, inferential problems still remain unsolved. There does not seem, at least to the present author, to be any prospect of developing reasonable parametric tests of statistical hypotheses within the conventional LS framework (i.e., with Kruskal's transformational approach). One may be tempted to suppose that the theory of isotonic regression (Barlow, Bartholomew, Bremner, & Brunk, 1972) may directly apply to the present case. However, viewed as an isotonic regression problem, distances serve as observations on the dependent variable whereas observed dissimilarities are measures on the independent variable (defining a convex cone of observations to which distances are regressed), and disparties are the parameters of the model to be estimated. The difficulty is obviously that distances calculated from a smaller number of parameters (stimulus coordinates) cannot be statistically independent. We maintain that the results of isotonic regression, particularly those pertaining to the distribution theory, do not carry over to the nonmetric MDS situation (despite the equivalence of the algebraic operations involved in the two problems). It seems necessary to reformulate the estimation procedure based on a criterion with statistically better properties.

In this paper we develop a single step ML estimation procedure for nonmetric MDS, along with the associated tests of the goodness of fit. One of the major conceptual differences between Kruskal's formulation and the present one is that we view distances (not disparities) as parameters which are further related to stimulus coordinates by some specific distance function. Distances are assumed to be error-perturbed, and to give rise to a particular set of observations (which are the empirical ordinal relations among dissimilarities) at a particular time. Due to the error perturbation the observed ordinal relations may sometimes violate the " true " orderings of underlying distances. Nonetheless they should convey some information concerning the likely state of underlying distances (i.e., large distances tend to be judged larger more often). We

attempt to recover the " true " distances in such a way that the likelihood of the observed orderings of dissimilarities is a maximum.

An ML estimation procedure has been proposed for metric MDS by Ramsay (1976, 1977) under the distributional assumptions similar to the present case. However, it should be emphasized that the present method is the first ML procedure which directly applies to nonmetric data.

Theoretical Developments

The Representation Model

It is the usual practice in statistical analyses of data to isolate systematic variations in data from random components. Systematic variations are identified with representation models of data and random components with stochastic error models. We construct estimation procedures of parameters both in representation and error models based on some plausible distributional assumptions on random components.

For representation models we employ a particular class of distance functions called Minkowski power metric models

$$d_{ij} = \left\{ \sum_{a=1}^{r} | x_{ia} - x_{ja} |^{p} \right\}^{1/p} \quad (i, j = 1, \ldots, n) \qquad (1)$$

for $p \geq 1$, where d_{ij} is the distance between stimuli i and j, x_{ia} is the coordinate of stimulus i on dimension a, r is the dimensionality of the space, p is the power of the metric space, and n is the number of stimuli. Without loss of generality, we assume that the stimulus coordinates are dimensionwise centered, and that the overall size of the stimulus configuration is constant.

The Error Model

We assume that d_{ij} in (1) is error-perturbed by processes of indeterminate nature. We consider two different manners in which errors are exerted on distances. We then discuss the most critical assumption; i.e., conditions for independence of observations.

The additive error model. Errors are assumed to operate on distances in an additive fashion. That is,

$$\lambda_{ij}^{(t)} = d_{ij} + e_{ij}^{(t)}, \qquad (2)$$

where $e_{ij}^{(t)}$ is the error random variable. The parenthesized superscript indicates occasion. (Unless necessary to avoid confusions, the occasion index will not be explicitly specified in the following discussion.)

[Normal errors] We assume that e_{ij} is normally distributed with zero mean and a finite, but unknown, variance σ_{ij}^2. That is,

$$e_{ij} \sim N(0, \sigma_{ij}^2), \qquad (3)$$

where σ_{ij} is the discriminal dispersion.

It might be noted that the normality assumption here is by no means the most " natural " distributional assumption in the present context. Suppes and Zinnes (1963), for example, derived a noncentral chi-square model for squared Euclidian distances based on the assumption that stimulus coordinates are independently normally distributed with a constant variance across stimuli and dimensions. The noncentral chi-square model is intuitively more appealing particularly in light of the fact that distances, by definition, can never be negative. In contrast, λ_{ij} as defined in (2) can be negative under the normality assumption on e_{ij}. However, the ratio of two independent noncentral chi-square variables (with the same degrees of freedom) has a doubly noncentral F distribution whose integral should be evaluated in order to obtain $\Pr(\lambda_{ij} > \lambda_{kl})$. This evaluation involves double summations of infinite series. One may consequently have to resort to some approximation methods, either by taking a finite sum of the infinite terms (Saito, 1974), or by taking an appropriate normal integral (Zinnes & Griggs, 1974).

The numerical complications and limitations which arise from the noncentral chi-square model have been the major obstacle toward a further progress along this line. Perhaps for this reason Ramsay (1976) has discarded this intuitively more attractive assumption, and has proposed a maximum likelihood estimation procedure for metric MDS based on a distributional assumption which is essentially equivalent to (3). Nakatani (1972) has also employed the normality assumption for his multidimensional confusion-choice model for pure recognition experiments.

The negative λ_{ij} may still be a problem, particularly in the metric case where $o_{ij}=\lambda_{ij}$, which tempts one to interpret λ_{ij} as an observed distance. We avoid this interpretation following the lead of Nakatani, who simply regards a set of λ_{ij} as "quantities which are positively and linearly correlated with interpoint distances." We use a noncommittal term, discriminal process, for λ_{ij}.

Note that this neutral interpretation is afforded by the fact that λ_{ij} is completely a hypothetical construct in the present case (as well as in Nakatani's case). Furthermore, we are only concerned with the ordinal property of the λ_{ij}. The possibility of negative λ_{ij} will not be a problem since there is no difficulty in establishing ordinal relations between two negative values or between positive and negative values.

[Structures on discriminal dispersion] The discriminal dispersion λ_{ij} may be different from one pair of stimuli to another. However, the principle of parsimony dictates a preference for the smallest possible number of parameters in the model. The effective number of independent parameters may be reduced by assuming various structures on the discriminal dispersion. The structure must be a plausible one reflecting reasonable assumptions about the data. We have chosen to impose

$$\sigma_{ij}^2 = \sigma^2 d_{ij}^s , \tag{4}$$

and to restrict our attention to the following three cases (Ramsay, 1976).

When $s=0$, Eq. (4) reduces to the constant variance assumption. When $s=2$, which postulates that the discriminal dispersion is proportional to the corresponding mean (distance), Eq. (4) roughly simulates the situation implied by Weber's law. When $s=1$ and $\sigma=2$, Eq. (4) approximates the relation (variance proportional to mean) which holds between mean and variance of the noncentral chi-square distribution.

The multiplicative error model. We may alternately assume, instead of (2), a multiplicative model,

$$\lambda_{ij}^{(t)} = d_{ij} e_{ij}^{(t)} \tag{5}$$

which reduces to a linear (additive) model like the one in (2) by the logarithmic transformation,

$$\lambda_{ij}' = d_{ij}' + e_{ij}' , \tag{6}$$

where $\lambda_{ij}'=\ln \lambda_{ij}$, $d_{ij}'=\ln d_{ij}$ and $e_{ij}'=\ln e_{ij}$. We assume that e_{ij}' is normally distributed; i.e.,

$$e_{ij}' \sim N(0, \sigma_{ij}^2) . \tag{7}$$

Then λ_{ij} follows the log-normal distribution (Aitchson & Brown, 1963) with the median of d_{ij}. The log-normal assumption on λ_{ij} has been discussed by Ramsay (1977).

One of the desirable consequences of the log-normal distribution, besides asymmetry (positive skewness) of the distribution, is that λ_{ij} assumes only positive values. The conceptual difficulty associated with the possible negativity of λ_{ij} we have encountered in the additive model does not exist in this case. Furthermore, the dispersion of λ_{ij} is proportional to its median (d_{ij}) with the approximate proportionality constant of $\exp(\sigma_{ij}^2)$. This proportionality holds over different σ_{ij} if σ_{ij} is constant for all i and j. Note that this is equivalent

to assuming $s=0$ in (4). Yet the constant variance assumption on e_{ij}' leads to the dispersion being proportional to the median for e_{ij}, the relation we have observed by setting $s=2$ in the additive model.

Independence. The independence consideration between distinct observations is very crucial in the current developments, since the definition of the likelihood function (Eq. (18)) is critically dependent on this assumption. The simplest case is when discriminal processes at different occasions take place in different individuals. This requires single-judgment sampling (Bock & Jones, 1968) in which a subject makes one and only one judgment. Then the independence assumption can reasonably be made. However, it would be unrealistic to employ single-judgment sampling in all situations for which the current procedure is designed. Most often we have to employ multiple-judgment sampling in which a subject makes more than a single judgment. Then we should take the covariance structure of observations into account in constructing the likelihood function. However, it can be shown that the independence assumption is still valid for multiple-judgment sampling, if certain conditions (much weaker than the complete independence) hold on the covariances between discriminal processes. The importance of the following development, which to the best of the author's knowledge has not be explicitly noted before, should not be overlooked. Bock and Jones (1968), for example, develop various procedures for pair comparisons almost entirely based on Thurstone's case V, which implies an equal covariance between two discriminal processes involved in a single judgment (Mosteller, 1951), across different judgments, and yet they fail to observe that multiple-judgment sampling reduces to single-judgment sampling, *if* the equal covariance assumption holds between discriminal processes involved in two distinct judgmental processes.

Let $\mathbf{e}^{(t)}$ and $\mathbf{e}^{(t')}$ be two-component vectors of random errors associated with the discriminal processes involved in judgments t and t' ($t \neq t'$). If the matrix of covariances between $\mathbf{e}^{(t)}$ and $\mathbf{e}^{(t')}$, namely $C^{(tt')}$, is expressible in the form of

$$C^{(tt')} = \mathbf{lg}' + \mathbf{hl}' \tag{8}$$

for *arbitrary* two-component vectors $\mathbf{g}$ and $\mathbf{h}$, and a two-component vector of ones $\mathbf{1}$, two difference processes at two distinct occasions t and t' will be statistically independent. Define

$$A = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} = \begin{bmatrix} \mathbf{a}_1' \\ \mathbf{a}_2' \end{bmatrix},$$

and

$$\mathbf{e}' = (\mathbf{e}^{(t)'}, \mathbf{e}^{(t')'}).$$

We have

$$A\mathbf{e} = \begin{pmatrix} e_{ij}{}^{(t)} - e_{kl}{}^{(t)} \\ e_{i'j'}{}^{(t')} - e_{k'l'}{}^{(t')} \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}. \tag{9}$$

The covariance between b_1 and b_2 is given by

$$\mathrm{Cov}(b_1, b_2) = \mathbf{a}_1' V(\mathbf{e}) \mathbf{a}_2 = \mathbf{q}' C^{(tt')} \mathbf{q}, \tag{10}$$

where $V(\mathbf{e})$ is the variance-covariance matrix of $\mathbf{e}$, and $\mathbf{q}'=(1, -1)$. Since $\mathbf{q}'\mathbf{1} = \mathbf{1}'\mathbf{q} = 0$, we have

$$\mathbf{q}' C^{(tt')} \mathbf{q} = 0. \tag{11}$$

Similar conditions for the variance-covariance matrix of repeated measures have been discussed by Huynh and Feldt (1970). Note that equal covariance cases follow as special cases by setting both $\mathbf{g}$ and $\mathbf{h}$ to be constant vectors (all four covariances are equal), only $\mathbf{g}$ to be a constant vector (covariances are rowwise equal), and only $\mathbf{h}$ to be a constant vector (covariances are columnwise equal).

It is important to realize that the equal covariance assumptions are much weaker than the strict independence assumption, and that, so far as the serial effects persist over only a few trials, it should not be too difficult to arrange the judgment sequence

(i.e., the order of stimulus presentation) so that either one of the equal covariance conditions will be satisfied.

Likelihood Function

Unless otherwise specified we assume the additive model for the discriminal process throughout this section.

We have as observations a set of empirical orderings on a well defined set of pairs of stimuli. Define

$$Y_{ijkl} = \begin{cases} 1, & \text{if } o_{ij} \succ o_{kl} \\ 0, & \text{if } o_{ki} \prec o_{kl} \end{cases} \tag{12}$$

where o_{ij} is the (possibly unobserved) empirical dissimilarity measure between stimuli i and j, $\succ$ is the empirical ordering defined on a subset Ω of Ξ^2 where Ξ is the set of pairs of stimuli such that $i>j$. We assume that $Y_{ijkl}=1$ whenever λ_{ij} defined in (2) is greater than λ_{kl}, and consequently that

$$\begin{aligned} \Pr(Y_{ijkl} = 1) &= \Pr(\lambda_i >_j \lambda_{kl}) \\ &= \Pr(\lambda_{ij} - \lambda_{kl} > 0) \end{aligned} \tag{13}$$

which is an important assumption implied by Thurstone's law of comparative judgment. Note that for the multiplicative model we have correspondingly that

$$\begin{aligned} \Pr(Y_{ijkl} = 1) &= \Pr(\ln \lambda_{ij} - \ln \lambda_{kl} > 0) \\ &= \Pr(\lambda_{ij}' - \lambda_{kl}' > 0) . \end{aligned} \tag{14}$$

By the normality assumption we obtain

$$\Pr(\lambda_{ij} > \lambda_{kl}) = \int_{-\infty}^{h_{ijkl}} f(z)dz = F_{ijkl}(h_{ijkl}) , \tag{15}$$

where f is the density function of the standard normal distribution and where

$$h_{ijkl} = (d_{ij} - d_{kl})/\sigma_{ijkl} . \tag{16}$$

The σ_{ijkl} in the present case assumes the form

$$\sigma_{ijkl} = (\sigma_{ij}^2 + \sigma_{kl}^2)^{1/2} = \sigma(d_{ij}^s + d_{kl}^s)^{1/2} \tag{17}$$

by virtue of (4).

The likelihood function L can now be stated as

$$L = \prod_{\Omega} T_{ijkl} , \tag{18}$$

where

$$T_{ijkl} = F_{ijkl}^{Y_{ijkl}}(1 - F_{ijkl})^{1-Y_{ijkl}} , \tag{19}$$

which is the simple Bernoulli distribution. The F_{ijkl} in this case is, of course, related to stimulus coordinates x_{ia} through (15), (16), (17) and (1).

The definition of the likelihood function as the product of T_{ijkl} is justified by the independence assumption of the observations, which, in turn, is justified by the particular experimental operations for obtaining empirical orderings of dissimilarities for which the present method is specifically designed.

We are to determine stimulus coordinates x_{ia}, and dispersion multiplier σ, so that L defined in (18) is maximal, or equivalently the log of L,

$$\ln L = \sum_{\Omega} \ln T_{ijkl} \tag{20}$$

is a maximum. (It might be pointed out at this point that in either the additive model with $s=2$ or the multiplicative model the h_{ijkl} defined in (16) is invariant over the choice of σ. In either case we can arbitrarity choose the value of σ and there exists no estimation problem thereof.)

Replications. When there are more than a single observation per tetrad, we may incorporate the number of replications in the estimation procedure. This can be done by defining the likelihood function as

$$L = \prod_{\alpha=1}^{N} \prod_{\Omega\alpha} T_{ijkl(\alpha)} , \tag{21}$$

where α is the index of replications (note that we have introduced a parenthesized subscript (α) to indicate T_{ijkl} for a particular replication). N is the total number of replications. The $\Omega\alpha$ is the set of pairs of dissimilarities for which an empirical ordering is obtained for replication α.

By rearranging (21) we have

$$L = \prod_{\Omega} \prod_{R_{ijkl}} T_{ijkl(\alpha)} = \prod_{\Omega} T_{ijkl}^*, \quad (22)$$

where $\Omega = \bigcup_{\alpha=1}^{N} \Omega_\alpha$, R_{ijkl} is the set of replications for which $T_{ijkl(\alpha)}$ is defined, and where $T_{ijkl}^* = \prod_{R_{ijkl}} T_{ijkl(\alpha)}$. We thus obtain, by defining N_{ijkl} to be the number of replications for a tetrad involving (ij) and (kl) pair (i.e., $N_{ijkl} = \sum_{R_{ijkl}} (1)$), and $Z_{ijkl} = \sum_{R_{ijkl}} Y_{ijkl(\alpha)}$ ($Y_{ijkl(\alpha)}$ is the Y_{ijkl} defined in (13) for replication α), that

$$L = \prod_{\Omega} T_{ijkl}^* = \prod_{\Omega} F_{ijkl}^{Z_{ijkl}} (1-F_{ijkl})^{N_{ijkl}-Z_{ijkl}}. \quad (23)$$

This familiar looking expression is the product of the portion of the binomial probability distribution related to parameters of the distribution.

Goodness of Fit Tests

Reasonable test statistics for the goodness of fit of a model can be readily constructed based on the general principle of the likelihood ratio criterion (Wilks, 1962). Define

$$\theta = \text{MAX } L(\pi_1)/\text{MAX } L(\pi_2) \quad (24)$$

where MAX $L(\pi)$ is the likelihood of model π maximized over its associated parameters. Model π_1 is subsumed under model π_2. Then

$$\chi^2 = -2 \ln \theta = -2\{\ln \text{MAX } L(\pi_1) - \ln \text{MAX } L(\pi_2)\} \quad (25)$$

is distributed asymptotically according to chi-square with degrees of freedom equal to the difference in the number of parameters in the two models.

For the general test of the goodness of fit we find that

$$\text{MAX } L(\pi_1) = \prod_{\Omega} \tilde{F}_{ijkl}^{Z_{ijkl}} (1-\tilde{F}_{ijkl})^{N_{ijkl}-Z_{ijkl}}, \quad (26)$$

where $\tilde{F}_{ijkl}$ (defined in (15)) is evaluated at the ML estimates of model parameters, and

$$\text{MAX } L(\pi_2) = \prod_{\Omega} \hat{F}_{ijkl}^{Z_{ijkl}} (1-\hat{F}_{ijkl})^{N_{ijkl}-Z_{ijkl}}, \quad (27)$$

where $\hat{F}_{ijkl} = Z_{ijkl}/N_{ijkl}$ (we define $0^0 = 1$) which is an ML estimate of a sample proportion. The degree of freedom is $n_2 - n_1$ where n_2 is the number of elements in Ω and n_1 is the number of prameters in model π_1.

An important class of tests of the goodness of fit in the present context is the test of the number of significant dimensions. In this case π_1 is associated with the solution with one dimension less than the number of dimensions in π_2. Under the representation model (1), the number of independent parameters is determined by the following rule. As has been alluded to earlier, we need only one degree of freedom for σ which is counterbalanced by a degree of freedom for the uniform dilation of stimulus coordinates in the additive error model with $s=0$ or 1. The number of parameters in the stimulus configuration is nr for n stimuli and r dimensions. However, distances are invariant over the translations (shifts) of origin of the stimulus coordinates (for any Minkowski power metric models) so that r of nr parameters may be chosen arbitrarily, leaving $r(n-1)$ parameters to be estimated. In case of the Euclidian distance we have another kind of indeterminacy; i.e., the Euclidian distance is also invariant over the orthogonal transformations of stimulus coordinates. An additional $r(r-1)/2$ parameters are arbitrarily chosen in order to make the solution unique. Thus, we have for the Euclidian model $r(n-1)-r(r-1)/2$ free parameters in the model. In the additive model with $s=2$ or the multiplicative error model, the h_{ijkl} defined in (16) and consequently the likelihood L, are invariant over the choice of σ. We need to fix one

more parameter in these cases to obtain a unique solution, so that the numbers of free parameters are one less than that given by the above formulae.

A third class of tests of the goodness of fit involves specific hypotheses on parameters. In the course of data analyses one may have more or less specific hypotheses about the parameters being estimated. In the next section we discuss an important class of such hypotheses. Statistical tests of specific hypotheses consist, again, of comparisons of the likelihood of two models, one (π_1) with the restrictions which follow from the hypotheses and the other (π_2) without restrictions. Then the χ^2 defined in (25) is with degree of freedom equal to the number of (linear) constraints (the difference in the number of independent parameters between the two models), which is the reduction in the number of free parameters as a consequence of imposed constraints.

An interesting statistical criterion which serves as a guide for choosing appropriate models (appropriate specifications of a model) is the Akaike Information Criterion (the AIC; Akaike, 1976) defined by

$$\mathrm{AIC}(\pi) = -2 \ln \mathrm{MAX}\, \mathrm{L}(\pi) + 2n_\pi$$

where n_π is the number of free parameters in model π. The usefulness of the statistic has been amply demonstrated in similar situations; e.g., in the determination of the appropriate numbers of common factors in factor analysis and in the appropriate choice of a model in regression analysis. We choose the model which minimizes the AIC (MAICE: Minimum AIC Estimator).

Linear Constraints

We discuss a special type of constraints (among general linear hypotheses) that may be made about the parameters, and that are particularly interesting in the present context, namely the equality constraints.

The equality constraints specify subsets of parameters which assume equal, but unknown, values. Notice that a set of equality constraints can generally be represented in matrix notation as

$$\boldsymbol{\theta} = \mathrm{A}\boldsymbol{\theta}^* + \mathbf{b} \tag{28}$$

where $\boldsymbol{\theta}$ is a vector of parameters in the original set, and $\boldsymbol{\theta}^*$ is a vector of parameters in the constrained set, $\mathbf{b}$ is a vector of fixed values, which is identically $\mathbf{0}$ (a zero vector) for equality constraints, and A is a matrix defining some linear dependence (relations) among parameters. Matrix A also has a specialized pattern for equality constraints, consisting of only ones and zeros. For example, if $\theta_2 = \theta_4 = \theta_3^*$, then the second and fourth rows of A contain a one in the third column with all other columns filled with zeros.

The necessary modification in the estimation procedure in the presence of constraints characterized in the form of (28) is given in the derivatives section. An interesting application of equality constraints will be demonstrated in the companion paper (Takane, 1978).

Numerical Method

To optimize L defined in (23), the derivatives of ln L with respect to unknown parameters (X and σ) are set to zero to obtain likelihood equations, which may be solved numerically by various methods. Among the most promising alternatives we illustrate the Gauss-Newton method, which has been incorporated in the current MAXSCAL-1 (Fortran IV program written along the theoretical developments presented in this paper), and which has a certain convenient characteristic in the present context. It has been shown that for the regular exponential type of distributions, the maximum likelihood method yields estimators which are asymptotically equivalent to those derived from the weighted LS criterion (Bradley, 1973), and that Fisher's scoring algorithm for maximum likelihood equations is equivalent to the Gauss-Newton algorithm for

weighted LS problems (Jennrich & Moore, 1975), if weighted and reweighted properly in each iteration. We exploit this fact to construct the Gauss-Newton algorithm for our ML estimation problem.

The updating equation for the Gauss-Newton algorithm for minimizing the weighted LS criterion,

$$\phi = (\hat{\mathbf{f}}-\mathbf{f})'\mathrm{W}(\hat{\mathbf{f}}-\mathbf{f}),$$

where $\hat{\mathbf{f}}$ is the vector of $\hat{\mathrm{F}}_{ijkl}$, $\mathbf{f}$ is the vector of F_{ijkl} and W is the diagonal matrix of $\mathrm{N}_{ijkl}/\mathrm{F}_{ijkl}(1-\mathrm{F}_{ijkl})$, is given by

$$\boldsymbol{\theta}^{(q+1)} = \boldsymbol{\theta}^{(q)}+\varepsilon\mathrm{H}(\boldsymbol{\theta}^{(q)})^{-1}\mathbf{g}(\boldsymbol{\theta}^{(q)}), \quad (29)$$

where

$$\mathrm{H}(\boldsymbol{\theta}) = (\boldsymbol{\partial}\mathbf{f}/\boldsymbol{\partial}\boldsymbol{\theta})'\mathrm{W}(\boldsymbol{\partial}\mathbf{f}/\boldsymbol{\partial}\boldsymbol{\theta}), \quad (30)$$

and

$$\mathbf{g}(\boldsymbol{\theta}) = (\boldsymbol{\partial}\mathbf{f}/\boldsymbol{\partial}\boldsymbol{\theta})'\mathrm{W}(\hat{\mathbf{f}}-\mathbf{f}), \quad (31)$$

and where $\boldsymbol{\theta}$ is the general representation of the vector of unknown parameters. The ε is the step size and the parenthesized superscript indicates the iteration number. It has been proved that $\mathrm{H}(\boldsymbol{\theta})$ is equivalent to Fisher's information matrix $\mathrm{I}(\boldsymbol{\theta})$ which is defined by

$$\mathrm{I}(\boldsymbol{\theta}) = \mathrm{E}(\mathbf{s}(\boldsymbol{\theta})\mathbf{s}(\boldsymbol{\theta})'),$$

where $\mathbf{s}(\boldsymbol{\theta})=(\boldsymbol{\partial}\ln\mathrm{L}/\boldsymbol{\partial}\boldsymbol{\theta})$ is Fisher's scoring vector, which also coincides with the negative gradient $-\mathbf{g}(\boldsymbol{\theta})$ derived from ϕ. Equation (29) is iteratively applied until convergence is reached. Computational details of the Gauss-Newton implementation of the algorithm can be found in Takane (1977).

Derivatives

In this section we collect expressions for the derivatives necessary for the optimization procedures discussed in the previous section.

For the general result we have

$$\begin{aligned}\boldsymbol{\partial}\ln\mathrm{L}/\boldsymbol{\partial}\theta_s &= \sum\boldsymbol{\partial}\ln\mathrm{T}_{ijkl}^{*}/\boldsymbol{\partial}\theta_s \\ &= \sum(\boldsymbol{\partial}\ln\mathrm{T}_{ijkl}^{*}/\boldsymbol{\partial}\mathrm{F}_{ijkl})(\boldsymbol{\partial}\mathrm{F}_{ijkl}/\boldsymbol{\partial}\theta_s),\end{aligned} \quad (32)$$

where θ_s is any parameter.

For specific results we have

$$\frac{\boldsymbol{\partial}\ln\mathrm{T}_{ijkl}^{*}}{\boldsymbol{\partial}\mathrm{F}_{ijkl}} = \frac{(\mathrm{Z}_{ijkl}-\mathrm{N}_{ijkl}\mathrm{F}_{ijkl})}{\mathrm{F}_{ijkl}(1-\mathrm{F}_{ijkl})}, \quad (33)$$

$$\frac{\boldsymbol{\partial}\mathrm{F}_{ijkl}}{\boldsymbol{\partial}\theta_s} = (\boldsymbol{\partial}\mathrm{F}_{ijkl}/\boldsymbol{\partial}\mathrm{q}_{ijkl})(\boldsymbol{\partial}\mathrm{q}_{ijkl}/\boldsymbol{\partial}\theta_s), \quad (34)$$

and

$$\frac{\boldsymbol{\partial}\mathrm{F}_{ijkl}}{\boldsymbol{\partial}\mathrm{q}_{ijkl}} = \mathrm{f}(\mathrm{q}_{ijkl}), \quad (35)$$

where

$$\mathrm{q}_{ijkl} = (\mathrm{d}_{ij}-\mathrm{d}_{kl})/\sigma(\mathrm{d}_{ij}{}^{\mathrm{s}}+\mathrm{d}_{kl}{}^{\mathrm{s}})^{1/2} \quad (36)$$

By rearranging (33) into

$$\frac{\mathrm{N}_{ijkl}}{\mathrm{F}_{ijkl}(1-\mathrm{F}_{ijkl})}(\hat{\mathrm{F}}_{ijkl}-\mathrm{F}_{ijkl}),$$

it may be seen that it is equal to $\mathrm{W}(\hat{\mathbf{f}}-\mathbf{f})$ part of (31). What we need is a complete expression for $\boldsymbol{\partial}\mathrm{q}_{ijkl}/\boldsymbol{\partial}\mathrm{x}_{ma}$, which is equivalent to $\boldsymbol{\partial}\mathbf{f}/\boldsymbol{\partial}\boldsymbol{\theta}$ part of (31).

We have

$$\begin{aligned}&\boldsymbol{\partial}\mathrm{q}_{ijkl}/\boldsymbol{\partial}\mathrm{x}_{ma} \\ &\quad = \boldsymbol{\partial}(\mathrm{d}_{ij}/\sigma_{ijkl})\boldsymbol{\partial}\mathrm{x}_{ma}-\boldsymbol{\partial}(_{kl}/\sigma_{ijkl})\boldsymbol{\partial}\mathrm{x}_{ma},\end{aligned} \quad (37)$$

$$\begin{aligned}&\frac{\boldsymbol{\partial}(\mathrm{d}_{ij}/\sigma_{ijkl})}{\boldsymbol{\partial}\mathrm{x}_{ma}} \\ &\quad = \frac{\sigma_{ijkl}(\boldsymbol{\partial}\mathrm{d}_{ij}/\boldsymbol{\partial}\mathrm{x}_{ma})-\mathrm{d}_{ij}(\boldsymbol{\partial}\sigma_{ijkl}/\boldsymbol{\partial}\mathrm{x}_{ma})}{\sigma_{ijkl}{}^{2}}\end{aligned} \quad (38)$$

and

$$\begin{aligned}&\boldsymbol{\partial}\mathrm{d}_{ij}/\boldsymbol{\partial}\mathrm{x}_{ma} \\ &= \frac{(\delta_{im}-\delta_{jm})|\mathrm{x}_{ia}-\mathrm{x}_{ja}|^{\mathrm{p}-1}\,\mathrm{signum}\,(\mathrm{x}_{ia}-\mathrm{x}_{ja})}{\mathrm{d}_{ij}{}^{\mathrm{p}-1}}\end{aligned} \quad (39)$$

where we tacitly assumed that $\mathrm{d}_{ij}\neq 0$ for all i and j such that $i\neq j$. Finally we have

$$\begin{aligned}\frac{\boldsymbol{\partial}\sigma_{ijkl}}{\boldsymbol{\partial}\mathrm{x}_{ma}} &= \frac{1}{2}\sigma(\mathrm{d}_{ij}{}^{\mathrm{s}}+\mathrm{d}_{kl}{}^{\mathrm{s}})^{-1/2} \\ &\quad \cdot\mathrm{s}\left\{\mathrm{d}_{ij}{}^{\mathrm{s}-1}\frac{\boldsymbol{\partial}\mathrm{d}_{ij}}{\boldsymbol{\partial}\mathrm{x}_{ma}}+\mathrm{d}_{kl}{}^{\mathrm{s}-1}\frac{\boldsymbol{\partial}\mathrm{d}_{kl}}{\boldsymbol{\partial}\mathrm{x}_{ma}}\right\},\end{aligned} \quad (40)$$

$$\boldsymbol{\partial}\mathrm{q}_{ijkl}/\boldsymbol{\partial}\sigma = -\mathrm{q}_{ijkl}/\sigma. \quad (41)$$

For the multiplicative model (5) and (6) we obtain

$$\partial d_{ij}'/\partial\theta_s = \partial \ln d_{ij}/\partial\theta_s$$
$$= (1/d_{ij})(\partial d_{ij}/\partial\theta_s) . \quad (42)$$

For the constrained case in which the original parameters $\boldsymbol{\theta}$ are further related to a smaller set of parameters $\boldsymbol{\theta}^*$ by a linear function as in (28), we have

$$\partial \ln L/\partial\boldsymbol{\theta}^* = (\partial\boldsymbol{\theta}/\partial\boldsymbol{\theta}^*)'(\partial \ln L/\partial\boldsymbol{\theta}) , \quad (43)$$

and

$$\partial\boldsymbol{\theta}/\partial\boldsymbol{\theta}^* = A .$$

This means, for the special case of equality constraints, that the gradients of θ_s^* are just the sums of the gradients for θ_t's which are assumed to be equal to θ_s^*. The θ_s^* and θ_t are the elements of $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}$, respectively.

SUMMARY

In this paper a maximum likelihood estimation procedure is described for nonmetric multidimensional scaling focussing on its conceptual and mathematical foundations. The likelihood function is derived making strong parametric assumptions on the process generating a set of ordinal judgments. Note that " nonmetric " does not mean " nonparametric ", nor does it necessarily imply " transformational ". Rather, the essence of a nonmetric procedure is in its exclusive use of ordinal information in finding a representation.

Empirical evaluations of the procedure will be reported in the companion paper (Takane, 1978).

REFERENCES

AITCHISON, J., & BROWN, J.A.C. 1963 *The log-normal distribution.* Cambridge: The Cambridge University Press.

AKAIKE, H. 1976 On entropy maximization principle. Paper presented at the Symposium on Applications of Statistics, Dayton, Ohio.

BARLOW, R. E., BARTHOLOMEW, D. J., BREMNER, J. M., & BRUNK, H. D. 1970 *Statistical inference under order restrictions.* London: Wiley.

BOCK, R. D., & JONES, L. V. 1968 *The measurement and prediction of judgment and choice.* San Francisco: Holden-Day.

BRADLEY, E. L. 1973 The equivalence of maximum likelihood and weighted least squares estimates in the exponential family. *Journal of the American Statistical Association,* **68**, 199–200.

HUYNH, H., & FELDT, L. S. 1970 Conditions under which mean square ratios in repeated measurements designs have exact *F*-distributions. *Journal of the American Statistical Association,* **65**, 1582–1589.

JENNRICH, R. I., & MOORE, R. H. 1975 Maximum likelihood estimation by means of nonlinear least squares. Research Bulletin RB-75-7, Educational Testing Service, Princeton, N.J.

KLINGBERG, F. L. 1941 Studies in measurement of the relations among sovereign states. *Psychometrika,* **6**, 335–352.

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

MESSICK, S. J. 1956 An empirical evaluation of multidimensional successive intervals. *Psychometrika,* **21**, 367–375.

MOSTELLER, F. 1951 Remarks on the method of paired comparisons. I. The least squares solution assuming equal standard deviations and correlations. *Psychometrika,* **16**, 203–206.

NAKATANI, L. H. 1972 Confusion-choice model for multidimensional psychophysics. *Journal of Mathematical Psychology,* **9**, 104–127.

RAMSAY, J. O. 1976 Two algorithms and various statistical models for multidimensional scaling by maximum likelihood. Unpublished paper, McGill University.

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

RICHARDSON, M. W. 1938 Multidimensional psychophysics. *Psychological Bulletin,* **35**, 659.

SAITO, T. 1974 Multidimensional Thurstonian scaling and its applications (I). *Nippon Univac Soken Kiyo,* **4**, 87–112.

SHEPARD, R. N. 1962 The analysis of proximities: Multidimensional scaling with unknown distance functions, I and II. *Psychometrika,* **27**, 125–140; 219–246.

SJÖBERG, L. 1967 Successive intervals scaling of paired comparisons. *Psychometrika,* **32**, 297–308.

SUPPES, P., & ZINNES, J. L. 1963 Basic measurement theory. In R.. D. Luce et al. (Eds.),

Handbook of mathematical psychology, Vol. I. New York: Wiley.

TAKANE, Y. 1977 Statistical procedures for nonmetric multidimensional scaling. Unpublished doctoral dissertation. The University of North Carolina.

TAKANE, Y. 1978 A maximum likelihood method for nonmetric multidimensional scaling: I. The case in which all empirical pairwise orderings are independent—evaluations. *Japanese Psychological Research*, **20**, (in press).

TORGERSON, W. S. 1952 Multidimesional scaling: I. Theory and method. *Psychometrika*, **17**, 273–286.

TORGERSON, W. S. 1958 *Theory and methods of scaling*. New York: Wiley.

WILKS, S. S. 1962 *Mathematical statistics*. New York: Wiley.

YOUNG, F. W. 1975 Scaling replicated conditional rank-order data. In D. R. Heise (Ed.), *Sociological methodology*. San Francisco: Jossey-Bass Publishers.

ZINNES, J. L., & GRIGGS, R. A. 1974 Probabilistic multidimensional unfolding analysis. *Psychometrika*, **39**, 327–350.

(Received July 28, 1977)