PSYCHOMETRIKA—VOL. 46, NO. 1.
MARCH, 1981

MULTIDIMENSIONAL SUCCESSIVE CATEGORIES SCALING: A MAXIMUM LIKELIHOOD METHOD

YOSHIO TAKANE

MCGILL UNIVERSITY

A single-step maximum likelihood estimation procedure is developed for multidimensional scaling of dissimilarity data measured on rating scales. The procedure can fit the euclidian distance model to the data under various assumptions about category widths and under two distributional assumptions. The scoring algorithm for parameter estimation has been developed and implemented in the form of a computer program. Practical uses of the method are demonstrated with an emphasis on various advantages of the method as a statistical procedure.

Key words: similarity ratings, maximum likelihood multidimensional scaling (MDS), method of successive categories.

Introduction

In applications of multidimensional scaling (MDS) dissimilarity judgments are often measured on rating scales with a relatively small number of observation categories. For example, an investigator may ask the subject to rate the degree of similarity between a pair of objects on a 7-point rating scale. Descriptive names of those categories might be: very similar, moderately similar, somewhat similar, neutral (neither similar nor dissimilar), somewhat dissimilar, moderately dissimilar, and very dissimilar. The same-different judgment is another example of category judgment in which the number of observation categories is restricted to two. Thus, the confusion data arising from the same-different judgments (Type I confusion data in Kruskal and Wish's terminology, 1978; see Rothkopf, 1957, for an example) can be analyzed as if they were two-category judgments. Rating judgments with a relatively few observation categories (say, up to seven or nine) are not only easy to make on the part of subjects, but also offer an important advantage regarding the statistical model evaluation. Since observed proportions of a certain dissimilarity falling into certain categories provide the minimum sufficient statistics for true proportions, it is possible to test the appropriateness of a distance representation against observed proportions.

In this paper we develop a maximum likelihood (ML) multidimensional scaling procedure specifically designed to analyze this type of similarity data. Our procedure is similar in spirit to Zinnes and Wolff's [1977] procedure for the same-different judgments, but is much more widely applicable. A maximum likelihood estimation procedure has been proposed for the method of successive categories scaling by Schönemann and Tucker [1967]. The major distinction between their method and ours is that the former scales whatever is measured on rating scales without any concern about what is measured. The current method, on the other hand, explicitly takes into account the content of measurement (that the data pertain to similarity), and not only *scales* the data, but also *represents* them by a distance model.

The research reported here was partly supported by Grant A6394 to the author by Natural Sciences and Engineering Research Council of Canada. Portions of this research were presented at the Psychometric Society meeting in Uppsala, Sweden, in June, 1978. MAXSCAL-2.1, a program to perform the computations discussed in this paper may be obtained from the author. Thanks are due to Jim Ramsay for his helpful comments.

Requests for reprints should be sent to Yoshio Takane, Department of Psychology, McGill University, 1205 Docteur Penfield Avenue, Montreal, Quebec H3A 1B1, Canada.

0033-3123/81/0300-2947$00.75/0

Categorical similarity ratings have been analyzed in at least two different ways. One approach, proposed by Messick [1956], is in the tradition of classical MDS [Torgerson, 1952], in which original observations are first transformed into "observed" distances based on some Thurstonian rationale, and those "observed" distances are subjected to Young and Householder's [1938] method to find a spatial representation. This approach, however, has a theoretical weakness in that it involves two distinct steps which are not compatible with each other. Perhaps for this reason it faded away with the advent of more powerful analysis tools.

The Shepard-Kruskal type of multidimensional scaling [Shepard, 1962; Kruskal, 1964a, b] has successfully been applied to many exploratory types of studies in which similarity data are collected by the rating method having a few observation categories. With their procedure the rating data can be analyzed either under the equal interval assumption (metric MDS) or under the unequal interval assumption (nonmetric MDS) between adjacent categories. Unlike classical MDS, the Shepard-Kruskal type of MDS is a single-step procedure in which both the optimal data transformation and the optimal data representation are obtained through the optimization of a single criterion. However, as a data analysis technique it has one definite drawback, namely the lack of statistical inference capability.

The maximum likelihood estimation procedure to be described in this paper, on the other hand, is not only capable of performing both metric and nonmetric types of analyses (like the Shepard-Kruskal type of procedures), but also offers the important advantages of permitting hypothesis testing and providing estimators which are best asymptotically normal (BAN) (see Wilks, 1962, for example). We refer to Ramsay [1977, 1978], Takane [1978a, b], and Takane and Carroll [Note 1] for maximum likelihood MDS procedures designed for other types of similarity data. Those procedures, though designed for different types of similarity data, also enjoy the same statistically nice properties.

The Method

Preliminary Considerations

In developing a scaling procedure it is important to consider at least three aspects of data: its empirical content, the nature of error processes, and the method by which the data are collected. The empirical content of the data refers to the kind of conceptual relationship that the data are pertaining to. For example, the data we typically deal with in MDS represent some sort of similarity relationship between objects. The scaling (or more precisely, the transformation) of the original similarity data in this case should reflect reasonable assumptions about our concept of similarity. That is, the scaling of the similarity data presupposes a model of similarity. This model is called the representation model of the data.

Observed data are typically subject to a sizable amount of measurement error. How we treat the error has important consequences upon the results of scaling, and thus the scaling procedure should reflect reasonable assumptions about the error processes (i.e., where and how randomness enters into a data generation process). The scaling of the original similarity data, for example, should in some sense be consistent with a specific error model for similarity.

Data, whether similarity or not, arise in a variety of experimental settings, and there are as many kinds of similarity data as there are methods to collect them. Similarity judgments are sometimes obtained by pair comparisons or rank orderings of similarities; sometimes they are obtained by ratings of similarities; yet at other times confusion probabilities are taken as similarity measures. How do people compare similarities? How

do they rank them? How do they categorize similarities and respond to rating scales? What processes underlie a confusion process, and what is its relationship to similarity? Presumably, different types of judgments require different mental operations on the part of subjects. The scaling procedure should reflect reasonable assumptions about the psychological processes involved in a specific task situation which generates a specific type of data. The likelihood function, or any other sensible criterion for parameter estimation, must be constructed in such a way that it captures essential features of the data generating processes. The model of these processes is called the response model for the data.

Any scaling procedure, explicitly or implicitly, presupposes the three types of models discussed above. Our contention is that they should always be made explicit, that they should be empirically sound, and that the data transformation as well as the data representation must be optimal with respect to all of them.

Likelihood Function

In line with the discussion in the previous section we present our method centering around its major constituents, namely the representation model, the error model and the response model. The discussion on these models should naturally lead to a specification of the likelihood function.

The representation model: A variety of models other than the distance model have been proposed for similarity data [see, for example, Johnson, 1967; Carroll, 1976; Tversky, 1977]. For the purpose of this paper, however, we deliberately restrict our attention to the euclidian model defined by

$$d_{ij} = \left\{ \sum_{a=1}^{A} (x_{ia} - x_{ja})^2 \right\}^{1/2}, \tag{1}$$

where d_{ij} is the euclidian distance between objects i and j, x_{ia} is the coordinate of object i on dimension a, and A is the dimensionality of the space. This by no means implies that our approach cannot be generalized to other representation models. In fact it is fairly straightforward to extend our approach to models other than euclidian distance.

The error model: The distance defined by (1) is assumed to be error-perturbed by some process, and the error model specifies distributional properties of this perturbation process.

The following two error models will be considered in this paper:

$$\begin{cases} \lambda_{ijk}^{(t)} = d_{ij} + e_{ijk}^{(t)} \\ e_{ijk}^{(t)} \sim N(0, \sigma_k^2), \end{cases} \quad \text{(Additive error model)} \tag{2}$$

and

$$\begin{cases} \lambda_{ijk}^{(t)} = d_{ij} e_{ijk}^{(t)} \\ \ln e_{ijk}^{(t)} \sim N(0, \sigma_k^2). \end{cases} \quad \text{(Multiplicative error model)} \tag{3}$$

The parenthesized superscript (t) here indicates an occasion of a particular judgment. (This occasion index may be omitted in the discussion to follow.) The k is an index of subject (or individual). Note that σ_k^2 has subscript k, implying that we may allow for individual differences in dispersion. However, when subjects are taken as mere replications, index k is null, and σ_k^2 reduces to a single common dispersion (σ^2).

Model (2) postulates that the error is additive, and is normally distributed over replications of judgments with constant variance σ_k^2 (within each subject). This error model was implicit in the initial scaling phase of classical MDS [Torgerson, 1952]. More recently this model was successfully employed in the confusion-choice model for multidimensional psychophysics by Nakatani [1974].

Model (3), on the other hand, assumes that the error is multiplicative, and that the log transform of the error follows the normal distribution with a constant variance. This

error model is equivalent to the log-normal distributional assumption made by Ramsay [1977] in his maximum likelihood MDS for dissimilarities measured on relatively "continuous" scales, and he also discussed some desirable properties of the log-normal assumption.

Our approach to the problem of specifying the error model is purely empirical. Rather than deciding which error model is correct on a priori grounds, we incorporate both error models in the estimation procedure. We can then analyze the same set of data under the two different distributional assumptions, compare the goodness of fit of the two competing error models, and decide which model fits better in a specific situation. If one model is found consistently better than the other, we can subsequently employ this better model in similar situations without carrying out the comparison each time. (We will demonstrate how this comparison may be done with our procedure in the later section.)

The response model: Based on the representation and error models given above we now specify the model which captures the transformation mechanism from $\lambda_{ijk}^{(t)}$ to a category judgment. When rating scales have a small number of observation categories (as has been assumed in the present case), it is more natural to assume that each category represents an interval rather than a single point. An observation that a certain similarity falls in a certain category is interpreted as implying that the similarity takes some unknown value within the interval corresponding to that category. An appropriate response model in this case would be the successive categories scaling model [see Torgerson, 1958].

In the method of successive categories, observation categories are represented by a set of ordered intervals, mutually exclusive and exhaustive, defined on the set of all possible values of similarity. These intervals are delimited by the upper and lower boundaries. Let b_{km} denote the upper boundary of the mth category. (The upper boundary of the mth category coincides with the lower boundary of the $(m+1)$st category.) Without loss of generality we may assume $-\infty = b_{k0} \leq \cdots \leq b_{km} \leq \cdots \leq b_{kM} = \infty$ where M is the total number of categories on a scale. (Note that the category boundaries have subscript k to allow for possible individual difference.) Then the probability (p_{ijkm}) that the similarity between objects i and j for subject k falls in category m is given by

$$p_{ijkm} = Pr(b_{k(m-1)} < \lambda_{ijk} \leq b_{km}). \tag{4}$$

Under the distributional assumption made in (2) or (3), this equation becomes

$$p_{ijkm} = \int_{b_{k(m-1)}}^{b_{km}} g(\lambda_{ijk})\, d\lambda_{ijk}, \tag{5}$$

where g is the density function of λ_{ijk}. By standardizing λ_{ijk} in (5) we obtain for the additive error model

$$p_{ijkm} = \int_{a_{ijk(m-1)}}^{a_{ijkm}} f(z)\, dz, \tag{6}$$

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

$$\begin{cases} z = \dfrac{\lambda_{ijk} - d_{ij}}{\sigma_k} \\ a_{ijk(m-1)} = \dfrac{b_{k(m-1)} - d_{ij}}{\sigma_k} \\ a_{ijkm} = \dfrac{b_{km} - d_{ij}}{\sigma_k}. \end{cases} \tag{7}$$

For the multiplicative model (7) should be replaced by

$$\begin{cases} z = \dfrac{\ln \lambda_{ijk} - \ln d_{ij}}{\sigma_k} \\ a_{ijk(m-1)} = \dfrac{\ln b_{k(m-1)} - \ln d_{ij}}{\sigma_k} \\ a_{ijkm} = \dfrac{\ln b_{km} - \ln d_{ij}}{\sigma_k}. \end{cases} \tag{8}$$

If we let

$$F_{ijkm} = \int_{-\infty}^{a_{ijkm}} f(z)\, dz, \tag{9}$$

then

$$p_{ijkm} = F_{ijkm} - F_{ijk(m-1)}. \tag{10}$$

Define an indicator variable Z_{ijkmr} which takes the value of one when the similarity between objects i and j for subject k falls in category m at replication r, and zero otherwise. That is,

$$Z_{ijkmr} = \begin{cases} 1 \text{ when } o_{ijkr} \in C_m \\ 0 \text{ otherwise,} \end{cases} \tag{11}$$

where $o_{ijkr} \in C_m$ means that o_{ijkr} (the similarity between objects i and j for subject k at replication r) falls in category m (C_m). Since each o_{ijkr} belongs to one and only one category, we have $\sum_{m=1}^{M} Z_{ijkmr} = 1$ and $Z_{ijkmr} Z_{ijkm'r} = 0$ for $m \neq m'$. The probability of a particular rating judgment can then be written as

$$p_{ijkr} = \prod_{m=1}^{M} p_{ijkm}^{Z_{ijkmr}}. \tag{12}$$

The joint likelihood of the total observations is now stated as

$$L = \prod_{k} \prod_{i,j} \prod_{r} p_{ijkr}, \tag{13}$$

where the products are to be taken over the terms for which Z_{ijkmr}'s ($m = 1, \ldots, M$) are actually observed. Substituting (12) in (13) gives

$$L = \prod_{k} \prod_{i,j} \prod_{r} \prod_{m} p_{ijkm}^{Z_{ijkmr}} = \prod_{k} \prod_{i,j} \prod_{m} p_{ijkm}^{Y_{ijkm}}, \tag{14}$$

where

$$Y_{ijkm} = \sum_{r} Z_{ijkmr}. \tag{15}$$

The Y_{ijkm} in the above equation is the frequency with which the similarity between i and j falls in category m. It is well-known that it is a minimum sufficient statistic for p_{ijkm} [Wilks, 1962]. The expression L in terms of this statistic (the second equation in (14)) is very convenient when we have a large number of replicated observations.

The log of L in (14),

$$\ln L = \sum_{k} \sum_{i,j} \sum_{m} Y_{ijkm} \ln p_{ijkm}, \tag{16}$$

is maximized over the stimulus coordinates $\{x_{ia}\}$ in the representation model, over the dispersions $\{\sigma_k\}$ in the error model, and over the category boundaries $\{b_{km}\}$ in the response model.

Constraints on Category Boundaries

There are $N \times (M-1)$ [number of subjects times number of category boundaries per scale] category boundaries to be estimated in the response model if they are fully realized. (The only restriction so far is that they are ordered within each subject.) Thus, it seems worthwhile to try to reduce this number by making various structural assumptions.

One plausible assumption is that the category boundaries are equally spaced. (This allows us to perform a metric type of analysis with the current method.) Many alternative parametrizations are possible in this case, among which we choose the following expression:

$$b_{km} = am + b, \qquad \text{(Linear constraint)} \tag{17}$$

where $a(>0)$ is a scale factor, and b is an additive constant. Both a and b are to be estimated from the data. Equation (17) states that the category boundaries are linearly related to each other, and that there are no individual differences in subjects' response style. To put it differently category boundaries are assumed to constitute an interval scale common to all subjects. (If, further, b is assumed to be zero, it would be a ratio scale.) Under (17) there are only two response parameters (a and b) to be estimated.

The restriction (17) should be applicable, at least in principle, to the case of multiplicative error model (3) as well. In this case, however, it is more natural to assume that

$$b_{km} = bm^a, \qquad \text{(Log-linear constraint)} \tag{18}$$

instead of (17). Again, $a(>0)$ and $b(>0)$ are the parameters to be estimated from the data. The above restriction is equivalent to saying that category boundaries constitute a log-interval scale. (If a is further restricted to be unity, it reduces to a ratio scale.) Note that (18) does not imply equal spacing among original category boundaries or among log-transformed category boundaries, unless a is fixed to unity.

A much more flexible restriction than (17) or (18) is

$$b_{km} = b_m. \qquad \text{(Unrestricted, no individual differences)} \tag{19}$$

There are no individual differences, but no particular functional relationships are assumed among category boundaries. Thus, they are free to take any values except that they are common to all subjects and that they have a prescribed order. The above restriction permits a nonmetric type of analysis (without individual differences in subjects' response style) with the current method.

Finally, we may impose no additional restrictions on b_{km} (unrestricted, individual differences). With this option we can perform a nonmetric MDS with individual differences in subjects' response style.

The chief advantage of allowing different structural assumptions on category boundaries in the estimation procedure is again that we can empirically choose the best structural model using the hypothesis testing capability of the present method.

Algorithmic Considerations and Derivatives

The log likelihood defined in (16) can be maximized by various numerical optimization methods. In this section we briefly discuss one of those methods, which is implemented in the current procedure.

While good approximation methods for the normal integral in (6) exist [Hastings, 1951], it is more convenient to substitute the logistic distribution for the normal distribution. The logistic distribution is known to give a close approximation to the normal distribution [Bock & Jones, 1968; Lord & Novick, 1968]. One of the desirable features of

the logistic distribution is that its distribution function has an explicit analytical expression, which is given by

$$F_{ijkm} = \frac{1}{1 + \exp(-s_k a_{ijkm})}, \tag{20}$$

where s_k is approximately $\pi/(3^{1/2}\sigma_k)$.

In order to maximize the log likelihood we must seek a point where the gradients of the log-likelihood function vanish. We use Fisher's scoring algorithm (see, for example, Rao, 1952) to locate a desired point. It is an iterative procedure in which parameter values are updated by the following formula:

$$\theta^{(q+1)} = \theta^{(q)} + \varepsilon^{(q)} I(\theta^{(q)})^{-1} u(\theta^{(q)}), \tag{21}$$

where (q) is the index of iteration number, θ is a vector of unknown parameters, ε is a step-size parameter, and $u(\theta)$ and $I(\theta)$ are defined by

$$u(\theta) = \left(\frac{\partial \ln L}{\partial \theta}\right), \tag{22}$$

and

$$I(\theta) = \operatorname{cov}[u(\theta)] = E\left[\left(\frac{\partial \ln L}{\partial \theta}\right)\left(\frac{\partial \ln L}{\partial \theta}\right)'\right] = -E\left[\frac{\partial^2 \ln L}{\partial \theta \partial \theta'}\right], \tag{23}$$

respectively. When the information matrix $I(\theta)$ defined above is singular due to nonuniqueness of parameters (as in the present case), we may simply replace the regular inverse in (21) by the Moore-Penrose inverse of $I(\theta)$ [Ramsay, 1978]. Of course, in the course of iterations the inverse of $I(\theta)$ does not have to be explicitly evaluated. We may simply solve $\varepsilon^{(q)} I(\theta^{(q)})(\theta^{(q+1)} - \theta^{(q)}) = u(\theta)^{(q)}$ for $\theta^{(q+1)}$.

Fisher's scoring algorithm is almost as quick as the Newton-Raphson method which has the quadratic convergence, and is known to be generally very effective for locating a precise optimum (see Takane, 1978a, b, for example).

Fisher's scoring algorithm is also known to be equivalent to the Gauss-Newton method for minimizing a weighted least squares criterion, when the assumed population distribution belongs to the regular exponential family [Jennrich & Moore, Note 2]. This is the case here, since as we can see in (12), we have a kernel of the multinomial distribution. The equivalent weighted least squares criterion in this case is given by

$$\sum_k \sum_{i,j} \sum_m \frac{1}{n_{ijk} p_{ijkm}} (Y_{ijkm} - n_{ijk} p_{ijkm})^2, \tag{24}$$

where $n_{ijk} = \sum_{m=1}^{M} Y_{ijkm}$. The expected value of the Hessian (the matrix of second order derivatives) of (24) is given by

$$H(\theta) = 2 \sum_k \sum_{i,j} \sum_m \frac{n_{ijk}}{p_{ijkm}} \left(\frac{\partial p_{ijkm}}{\partial \theta}\right)\left(\frac{\partial p_{ijkm}}{\partial \theta}\right)' \tag{25}$$

which is known to be proportional to $I(\theta)$. [$H(\theta) = 2I(\theta)$.] Thus, the explicit form of the information matrix can be readily obtained without taking the second order derivatives of the log likelihood.

The gradient $u(\theta)$ is given by

$$u(\theta) = \sum_k \sum_{i,j} \sum_m \frac{Y_{ijkm}}{p_{ijkm}} \left(\frac{\partial p_{ijkm}}{\partial \theta}\right) \tag{26}$$

Thus, we need to obtain $\partial p_{ijkm}/\partial\theta$ to evaluate both $u(\theta)$ and $I(\theta)$. We have

$$\frac{\partial p_{ijkm}}{\partial\theta} = \frac{\partial F_{ijkm}}{\partial\theta} - \frac{\partial F_{ijk(m-1)}}{\partial\theta}. \tag{27}$$

By breaking down the θ into its components, we have

$$\frac{\partial F_{ijkm}}{\partial x_{qa}} = \left(\frac{\partial F_{ijkm}}{\partial d_{ij}}\right)\left(\frac{\partial d_{ij}}{\partial x_{qa}}\right), \tag{28}$$

where

$$\frac{\partial F_{ijkm}}{\partial d_{ij}} = -F_{ijkm}(1 - F_{ijkm})s_k, \qquad \text{(Additive error model)} \tag{29}$$

or

$$\frac{\partial F_{ijkm}}{\partial d_{ij}} = -\frac{F_{ijkm}(1 - F_{ijkm})s_k}{d_{ij}}, \qquad \text{(Multiplicative error model)} \tag{29'}$$

and where

$$\frac{\partial d_{ij}}{\partial x_{qa}} = \frac{(\delta_{iq} - \delta_{jq})(x_{ia} - x_{ja})}{d_{ij}} \tag{30}$$

($\delta_{..}$ is a Kronecker delta). For derivatives of F_{ijkm} with respect to the dispersion parameters we have

$$\frac{\partial F_{ijkm}}{\partial s_k} = -F_{ijkm}(1 - F_{ijkm})h_{ijkm} \tag{31}$$

where $h_{ijkm} = \sigma_k a_{ijkm}$. Derivatives of F_{ijkm} with respect to the parameters of the response model have to be given separately for each different type of restriction. Under the linear or the log-linear assumption [(17) or (18)] we have

$$\begin{aligned} \frac{\partial F_{ijkm}}{\partial a} &= F_{ijkm}(1 - F_{ijkm})s_k m \\ \frac{\partial F_{ijkm}}{\partial b} &= F_{ijkm}(1 - F_{ijkm})s_k \end{aligned} \qquad \text{(Additive error model)} \tag{32}$$

or

$$\begin{aligned} \frac{\partial F_{ijkm}}{\partial a} &= F_{ijkm}(1 - F_{ijkm})s_k \ln m \\ \frac{\partial F_{ijkm}}{\partial b} &= \frac{F_{ijkm}(1 - F_{ijkm})s_k}{d_{ij}}. \end{aligned} \qquad \text{(Multiplicative error model)} \tag{32'}$$

For the monotonic case without individual differences [(19)], we have

$$\frac{\partial F_{ijkm}}{\partial b_m} = F_{ijkm}(1 - F_{ijkm})s_k \qquad \text{(Additive error model)} \tag{33}$$

or

$$\frac{\partial F_{ijkm}}{\partial b_m} = \frac{F_{ijkm}(1 - F_{ijkm})s_k}{b_m}. \qquad \text{(Multiplicative error model)} \tag{33'}$$

Finally, for the completely unrestricted case we have

$$\frac{\partial F_{ijkm}}{\partial b_{km}} \equiv F_{ijkm}(1 - F_{ijkm})s_k \qquad \text{(Additive error model)} \tag{34}$$

$$\frac{\partial F_{ijkm}}{\partial b_{km}} \equiv \frac{F_{ijkm}(1 - F_{ijkm})s_k}{b_{km}}. \qquad \text{(Multiplicative model)} \tag{34'}$$

Results and Discussion

As noted earlier, the statistical inference capability of maximum likelihood estimation offers several important advantages over other fitting procedures. In this section we would like to demonstrate some of these in the context of multidimensional scaling.

One of the advantages of maximum likelihood estimation derives from the fact that the inverse of the information matrix at the maximum provides asymptotic variance and covariance estimates of estimated parameters. Using this property along with the asymptotic normality of maximum likelihood estimators, we can obtain asymptotic confidence intervals or regions (of a prescribed size) for any subsets of parameters.

With maximum likelihood estimation a goodness of fit criterion can be readily derived from the general principle of likelihood ratio [Wilks, 1962]. Specifically, minus twice the log likelihood ratio has asymptotically a chi-square distribution with degrees of freedom equal to the difference in the number of parameters. This statistic can be used to determine whether one model fits to the data significantly better than the other.

The use of the asymptotic chi square for a model comparison is, however, limited to the case in which one model is a restricted version of the other. It cannot be used, for example, to compare the additive error model (2) with the multiplicative error model (3), since neither is a proper subset of the other. The AIC statistic, introduced by Akaike [1973, 1977], can be used instead, where the asymptotic chi-square test is not feasible. The AIC statistic associated with a model is defined by

$$\text{AIC} = -2 \ln L + 2 \text{ d.f.}, \tag{35}$$

where L is the maximum likelihood of the model and d.f. is the effective number of parameters in the model.

Note that adding twice the degrees of freedom compensates for the "loss" of error d.f. as a consequence of increasing the number of parameters. If more parameters are used to describe the data, it is natural to get a larger likelihood, possibly without improving a true goodness of fit, and the AIC avoids this spurious improvement of fit by penalizing the use of additional parameters. The model which gives the minimum AIC value is considered the best fitting model. We may, of course, divide the AIC by -2 and choose the model which maximizes $\ln L -$ d.f.

The addition of 2 d.f. to $-2 \ln L$ in (35) is by no means arbitrary. The AIC has been derived as a measure which approximately maximizes the entropy defined by

$$B(f, \hat{f}) = E\left[\int \ln\left\{\frac{\hat{f}(y; \mathbf{x})}{f(y)}\right\} f(y)\, dy\right] \tag{36}$$

$$= E \int f(y) \ln \hat{f}(y; \mathbf{x})\, dy - \int f(y) \ln f(y)\, dy,$$

where $\mathbf{x}$ is the vector of observations, $f(y) \equiv f(y|\boldsymbol{\theta}_0)$ and $\hat{f}(y; \mathbf{x}) \equiv f(y; \mathbf{x}|\hat{\boldsymbol{\theta}})$ (where $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}$ are the true and estimated parameter values, respectively) are the probability density functions of the true and fitted models, respectively, and where E is the expectation of $\hat{\boldsymbol{\theta}}$ with respect to $f(\mathbf{x})$. Since the second term in (36) is a function of only $f(y)$, the

$\ln \hat{f}(y; \mathbf{x})$ which maximizes the first term is deemed desirable. Also, two competing models, $\hat{f}$ and $\hat{\hat{f}}$, can be compared entirely on the basis of the first term, since $B(f, \hat{f}) - B(f, \hat{\hat{f}})$ does not involve the second term.

The crucial thing is then to estimate the first term of $B(f, \hat{f})$. Unfortunately the formal derivation of this estimate is beyond the scope of this paper, and the interested reader should consult Akaike's original articles [1973, 1974, 1977]. It suffices here to point out that $\ln L -$ d.f. turns out to be the estimate.

The AIC cannot be used to compare two fitted models for which the density functions of the true models are not the same. (This is because the second term in $B(f, \hat{f})$ would not cancel out.) This effectively excludes the possibility of comparing two models applied to two different samples. On the other hand, AIC may be used to compare any two models, so far as $f(y)$ is common. No distributional properties of AIC are yet known, so that it is a descriptive index of the goodness of fit of a model (as are the maximum likelihood criterion and the least squares criterion). We choose the model with the minimum AIC value, just as we choose the maximum likelihood or the least squares solution over nonoptimal solutions. The AIC has been derived based on the asymptotic properties of maximum likelihood estimators. Like the use of the asymptotic chi square, its unmodified use should be limited to large sample problems. (We do not have a definite guideline as to how large a sample should be, but a rule of thumb indicates ten times the number of parameters estimated are more than enough, at least in the context of maximum likelihood MDS of the sort presented in this paper. A study is being conducted [Takane & Carroll, Note 3] to find a rule to modify AIC so that it is applicable to small samples as well).

The AIC has successfully been applied to a variety of settings related to the identification of the best fitting models. Those include the detection of outliers [Kitagawa, 1979], the identification of the best regression model in the analysis of cross-classification tables [Sakamoto & Akaike, 1978], and the selection of the order in an auto-regressive process [Shimizu, 1978], just to mention a few. More recently Steiger and Lind [Note 4] compared the performance of AIC with other statistics (sequential likelihood ratio and BIC, Schwarz, 1978) in determining the number of factors in factor analysis. We refer to those articles as well as many others referenced in those articles for examples of useful applications of AIC.

Krantz and Tversky's Example

With the understanding of the general characteristics of maximum likelihood estimation as described above, we now turn to a specific example of analyses with MAXSCAL-2.1, a Fortran program incorporating the theoretical developments in the previous section.

In Krantz and Tversky's [1975] study on the psychological dimensions of similarity between rectangles, a set of 17 rectangles were used to test the additivity of stimulus dimensions in subjective judgments of dissimilarities. They considered two alternatives, the height and width combination and the area (height $\times$ width) and shape (height/width) combination, as possible candidates of psychological dimensions of rectangles. (Rectangles are uniquely identifiable in terms of either one of these combinations.)

They approached the problem from a measurement theoretic viewpoint. A set of axioms characterizing the model were postulated, from which testable conditions were derived. Those conditions were tested against empirical data. It was found that neither combinations of stimulus dimensions were satisfactory. (The area and shape combination was found to fit slightly better than the height and width combination.)

Our first analysis is concerned with the tests of Krantz and Tversky's two hypotheses about the psychological dimensions of rectangles. Using the constrained optimization feature of MAXSCAL-2.1 it is possible to test their hypotheses based on a rigorous statistical inference, which was not possible in their original analysis. However, our test is of a limited nature in the sense that the euclidian model, which is but a special case of the additive difference model that Krantz and Tversky were dealing with, is assumed for the representation of the similarity data.

A small experiment was conducted to collect the relevant data. Stimuli were cut from Krantz and Tversky's original report [Note 5], and were each pasted on a 6 by 3 inch blank index card. Stimuli were placed in pairs on the desk in front of the subject, side by side (approximately two inches apart) at approximately $1\frac{1}{2}$ feet from the subject. The subject rated the similarity between all possible pairs (136 pairs) of rectangles on 7-point rating scales (very similar to very dissimilar). A single subject (male adult) replicated the experiment six times (either one or two days apart between sessions). Each session lasted about 50 minutes. The order of stimulus presentation was randomized over replications of the experiment. Seventeen perceptual stimuli were large enough in number to prevent systematic memory effects on the judgments of dissimilarities. Replications were made within a single subject rather than over subjects in this experiment, since Krantz and Tversky's analysis indicated clear individual differences in the evaluation of stimulus dimensions.

In order to determine the most appropriate model the data were analyzed under various sets of assumptions. The results are reported in the upper half of Table 1 (the portion labeled "Unconstrained Solution"). Three figures are reported in each cell; the top figure is minus twice the log likelihood, the middle is the value of the AIC statistic and the bottom is the d.f. associated with the model. The d.f. for a model is evaluated by $nA - A(A + 1)/2 + n_c$ where n is the number of stimuli, A is the dimensionality of the space and n_c is the number of independent parameters needed to specify category boundaries. For example, $n_c = M - 1$ for the unrestricted, no individual differences case and $n_c = 2$ for the linear or log-linear constraint. The MAXSCAL-2.1 program is capable of performing MDS under two different distributional assumptions about the error process. Which error model is more descriptive of the error process is as empirical a question as the identification of the best representation model. Columns 1 and 3 of Table 1 compare the goodness of fit of the multiplicative or log-normal error model with that of the additive error model. In both two- and three-dimensional solutions the multiplicative model shows a better fit. From our limited experience the multiplicative error model seems to fit to the rating data generally better than the additive error model. However, in other types of judgments (e.g., pair comparisons and rankings) the additive error model has been found consistently superior [Takane, 1978b; Takane & Carroll, Note 1].

Given that the error model is multiplicative, we now examine the restriction on category boundaries. The first and second columns of Table 1 compare the goodness of fit of two assumptions: the unrestricted, no individual differences case (19) and the log-linear case (18). In both two- and three-dimensional solutions, the former gives smaller values of AIC, indicating a better fit. The asymptotic chi squares representing the differences between (18) and (19) were 104.92 with 4 d.f. ($p < .001$) for three-dimensional solution, and 117.95 with 4 d.f. ($p < .001$) for two-dimensional solutions. Thus, in both dimensionalities the unrestricted case is significantly better than the log-linear constraint. The unrestricted category boundaries estimated under the multiplicative error model and with a two-dimensional stimulus configuration are displayed in Figure 1 along with the 95% asymptotic confidence intervals. In the figure, tick marks are placed at the middle of vertical line segments (corresponding to the 95% asymptotic confidence intervals) to indicate point estimates of category boundaries. Circular dots indicate estimates of

TABLE 1

Summary of MAXSCAL-2.1 Analyses of Rectangle Data

Error Model		Multiplicative		Additive
Category Boundaries		unconstrained no individual differences	log linear	unconstrained no individual differences
Unconstrained Solutions				
3 dimensions	$-2\ln(L)$	1543.10	1648.02	1604.18
	AIC	1645.10	1742.02	1706.18
	(d.f.)	(51)	(47)	(51)
2 dimensions	$-2\ln(L)$	1571.66	1689.61	1654.34
	AIC	1645.66	1755.61	1728.34
	(d.f.)	(37)	(33)	(37)
Constrained Solutions				
Shape and area factorial hypothesis	$-2\ln(L)$	2037.24		
	AIC	2065.24		
	(d.f.)	(14)		
Height and Width factorial hypothesis	$-2\ln(L)$	1910.83		
	AIC	1946.83		
	(d.f.)	(18)		
Schönemann's hypothesis 1	$-2\ln(L)$	1707.22	1847.06	1793.63
	AIC	1737.22	1869.06	1823.63
	(d.f.)	(15)	(11)	(15)
Schönemann's hypothesis 2	$-2\ln(L)$	1830.93		
	AIC	1848.93		
	(d.f.)	(9)		
Schönemann's hypothesis 3	$-2\ln(L)$	1933.97		
	AIC	1949.97		
	(d.f.)	(8)		

category boundaries under the log-linear assumption. The function relating these dots is virtually linear, though slightly convex downward ($\hat{a} = 1.025$, $\hat{b} = .375$). However, it indicates a clear departure from the unrestricted case. It may be seen from the graph that the fit of the log-linear constraint would have been much better, if an additive constant term were included (i.e., $b_{km} = bm^a + c$). The power transformation has to be more convex in order to more closely approximate the unrestricted category boundaries, but this seems to be hindered by the fact that the power transformation passes through the origin.

Assuming that the unrestricted, no individual differences restriction is appropriate, we are now in the position to choose appropriate dimensionality of the representation space. We have the AIC values of 1645.10 and 1645.66 for three- and two-dimensional solutions (Column 1 of Table 1) respectively, implying that the three-dimensional solution is slightly better. The difference is small. Nevertheless, it seems significant ($\chi^2 = 28.56$ with 14 d.f.; $p < .05$). Furthermore, the third dimension is clearly interpretable; it contrasts inside stimuli with those located at the periphery of the configuration. Thus, we are inclined to conclude that three dimensions are needed to describe the data adequately. (An alternative interpretation of the third dimension will be given later.) For illustrative convenience the two-dimensional configuration is displayed instead in Figure 2. The reader may exert a little imagination here to obtain the desired

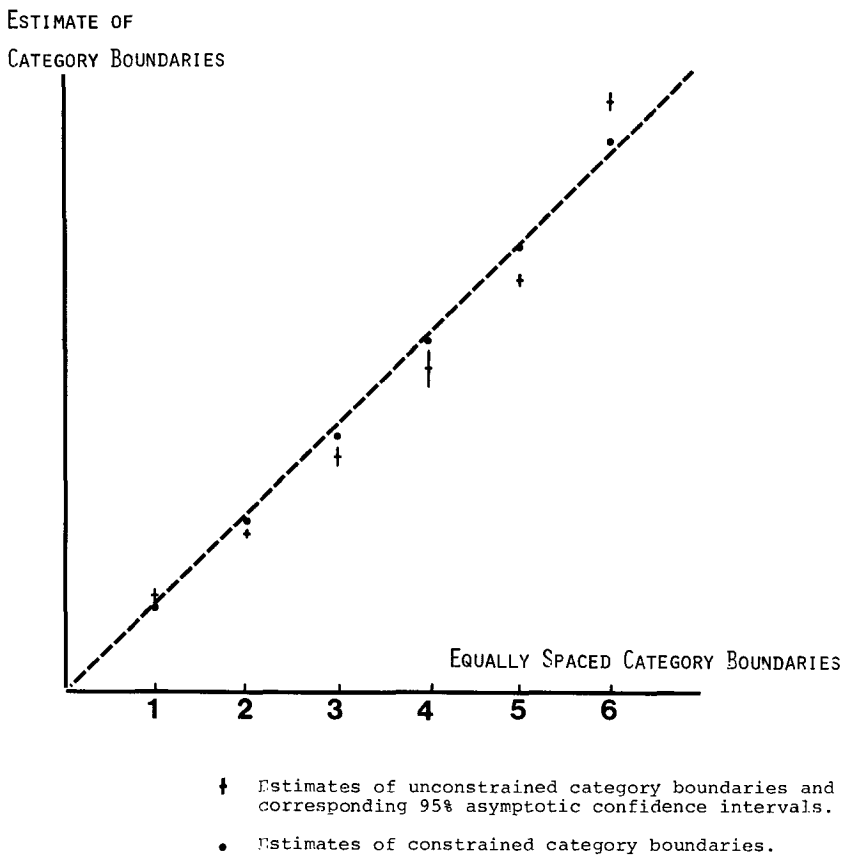

FIGURE 1
Estimates of category boundaries and 95% asymptotic confidence intervals obtained from unconstrained two-dimensional solution.

three-dimensional configuration; all outside stimuli (1 through 8) are slightly above the surface, while all inside stimuli (9 through 17) are slightly below the surface. Otherwise, this two-dimensional configuration is very much like the one obtained by Krantz and Tversky [1975]; the horizontal axis roughly corresponds with the area dimension and the vertical axis with the shape dimension (with some qualification to be noted later on).

We may now go on to test Krantz and Tversky's hypotheses. Their original hypotheses are, strictly speaking, inconsistent with a three-dimensional solution. However, it would still be informative to see how they are formally tested using MAXSCAL-2.1 in two dimensions.

A set of rectangles employed in Krantz and Tversky's study was constructed by factorially combining several levels of the area and shape dimensions of rectangles. Thus, under the additivity assumption of these dimensions, some of the rectangles ought to take equal coordinate values along these dimensions. In Figure 2 those stimuli whose coordinate values must be equal are connected by line segments for each dimension. For example, Rectangles 1, 2 and 3 have the same area, and Rectangles 3, 4 and 5 have the same shape, etc. The additivity hypothesis can be tested by explicitly obtaining an MDS solution which satisfies the condition implied by the hypothesis, and then by comparing the goodness of fit of this constrained solution with that of the unconstrained solution obtained earlier. Equality constraints can be imposed by treating those parameters

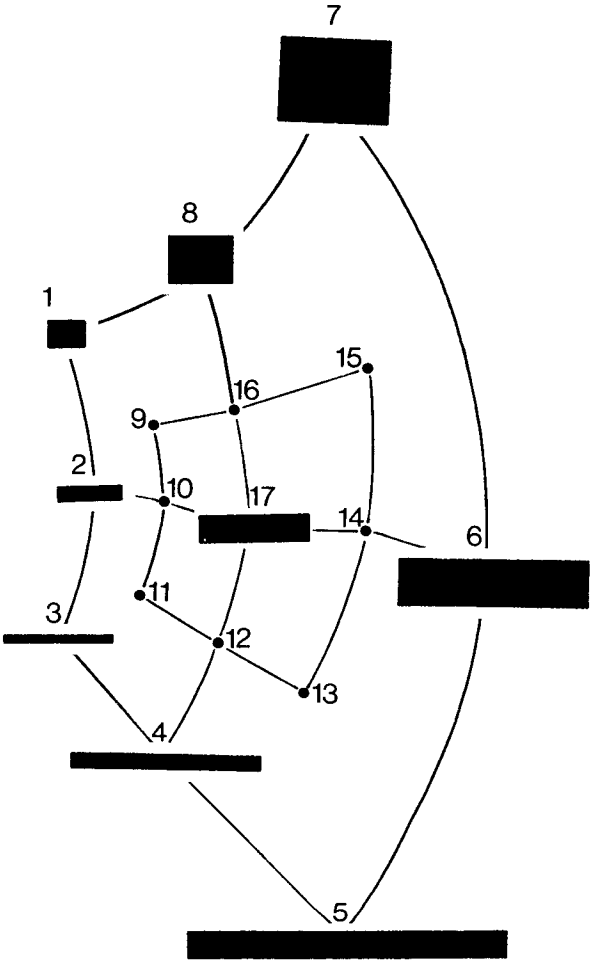

FIGURE 2
Unconstrained two-dimensional solution.

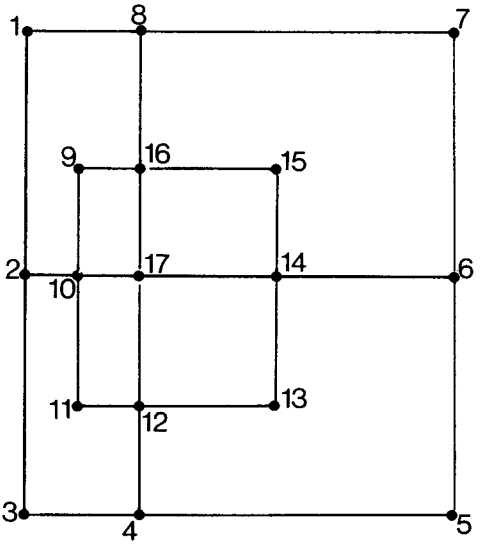

FIGURE 3
Constrained solution obtained under the area (horizontal) and shape (vertical) factorial hypothesis.

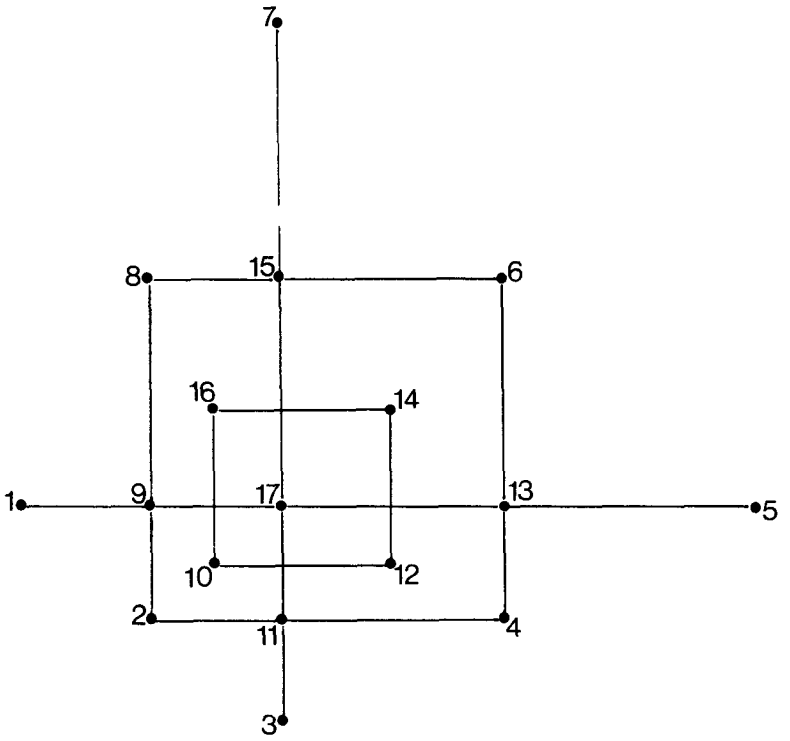

FIGURE 4
Constrained solution obtained under the width (horizontal) and height (vertical) additivity hypothesis.

assumed equal as a single parameter [Takane, 1978a]. The constrained solution is presented in Figure 3. The value of the AIC statistic associated with this solution is 2065.34, (the eighth row, first column) which is a way higher than 1645.66 for the two-dimensional unconstrained solution ($\chi^2 = 465.58$ with 23 d.f.; $p < .001$) under the equivalent condition. This is, of course, not unexpected after having looked at the unconstrained stimulus configuration. In Figure 2 it can be observed that the shape difference tends to be judged larger at larger area levels, indicating the failure of the area and shape additivity hypothesis.

The set of rectangles employed in the experiment was also constructed in such a way that some rectangles have the same height or width levels. Thus, we may test the height and width additivity hypothesis by imposing another set of equality constraints. Figure 4 displays the constrained solution obtained under this hypothesis. Note that coordinate values are equated for equal height (vertical direction) and width (horizontal direction) levels. The AIC value of this solution is 1946.83, (eleventh row, first column) which is much smaller than that of the previous solution, but is still substantially larger than that of the unconstrained solution ($\chi^2 = 339.17$ with 19 d.f.; $p < .001$). Thus, the height and width additivity hypothesis is also rejected. The fact that this case is still better than the area and shape additivity hypothesis, though it contradicts Krantz and Tversky's finding, may be clearly seen in Figure 4; if the configuration is rotated about 90° clockwise, it becomes evident that it better captures an important feature of the unconstrained solution (Figure 2) that the judged shape difference increases as the area increases.

Schönemann [1977] observed the same general pattern in Krantz and Tversky's configuration, and proposed a hypothesis that the perceived difference in shape increases in a specific way with the area of rectangles. A hypothesis slightly more general than

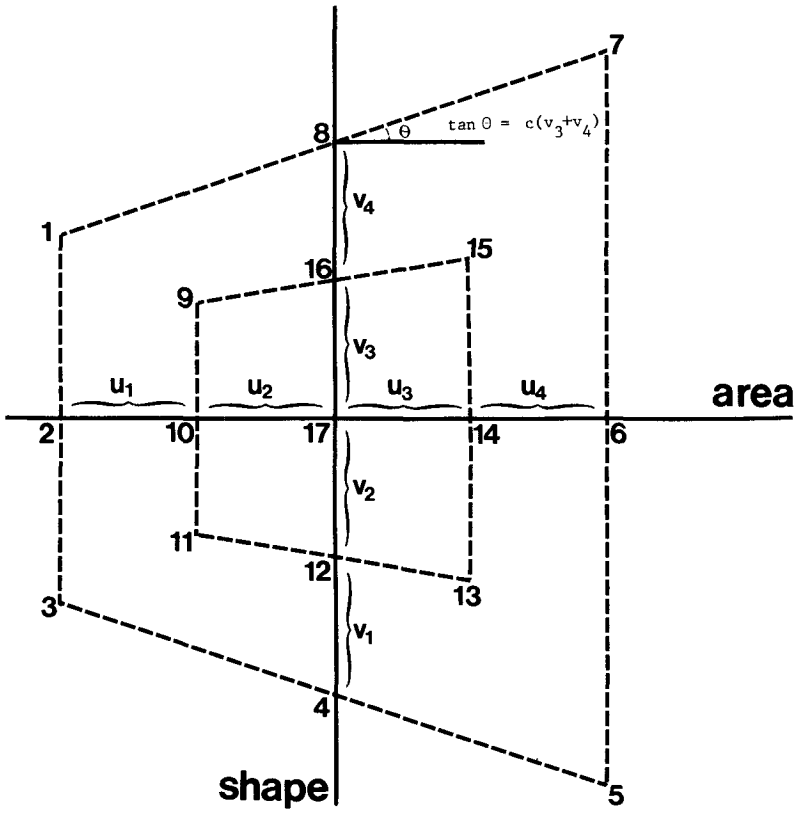

FIGURE 5
Illustration of Schönemann's hypothesis that the perceived shape difference increases as the area increases.

Schönemann's original hypothesis is depicted in Figure 5. While area levels are assumed to contribute to overall dissimilarities independently from the shape dimensions (as indicated by the parallel vertical lines connecting rectangles with physically identical areas), shape levels are assumed to diverge, so to speak, as area levels go up.

Without loss of generality Rectangle 17 may be placed at the center (origin) of the configuration. If we denote each successive interval between adjacent area levels by u_j ($j = 1, \ldots, 4$) from left to right and each successive interval between adjacent shape levels at area level = 0 by v_j ($j = 1, \ldots, 4$) from bottom to top, and if we assume that the divergence rate of the shape levels is $\tan \theta = cy$ where c is some constant and y is the vertical coordinate of a particular shape level at area level = 0, then the stimulus coordinates can be reparametrized as shown in Table 2. Let us call the structural assumption represented in Table 2 Schönemann's Hypothesis 1. We may also assume that all intervals along the area dimension are equal ($u_1 = u_2 = u_3 = u_4$), and that all intervals along the shape dimension are equal ($v_1 = v_2 = v_3 = v_4$). This is in fact Schönemann's original hypothesis (we call this Schönemann's Hypothesis 2). Finally, all intervals ($u_1, \ldots, u_4, v_1, \ldots, v_4$) may be assumed equal (Schönemann's Hypothesis 3).

The values of the AIC statistic corresponding to these solutions are 1737.22 (Hypothesis 1), 1848.93 (Hypothesis 2), and 1949.97 (Hypothesis 3) (see Table 1). Hypothesis 1 has the minimum AIC value among these. There is a clear tendency that the area intervals get larger for higher area levels, though this tendency is less clear for the shape intervals. (Note that rectangles are equally spaced in terms of the log transformed physical area and shape dimensions.) However, the AIC value of 1737.22 from

TABLE 2

Stimulus Coordinates as Functions of Reduced Model Parameters given in Figure 6

Stimulus	Dimension: Area	Dimension: Shape
1	$-(u_1 + u_2)$	$(v_3 + v_4)\{1 - c(u_1 + u_2)\}$
2	$-(u_1 + u_2)$	0
3	$-(u_1 + u_2)$	$-(v_1 + v_2)\{1 - c(u_1 + u_2)\}$
4	0	$-(v_1 + v_2) - (v_1 + v_2)$
5	$u_3 + u_4$	$-(v_1 + v_2)\{1 + c(u_3 + u_4)\}$
6	$u_3 + u_4$	0
7	$u_3 + u_4$	$(v_3 + v_4)\{1 + c(u_3 + u_4)\}$
8	0	$v_3 + v_4$
9	$-u_2$	$v_3(1 - cu_2)$
10	$-u_2$	0
11	$-u_2$	$-v_2(1 - cu_2)$
12	0	$-v_2$
13	u_3	$-v_2(1 + cu_3)$
14	u_3	0
15	u_3	$v_3(1 + cu_3)$
16	0	u_3
17	0	0

Hypothesis 1 is still substantially larger than 1645.66 from the two-dimensional unconstrained solution ($\chi^2 = 135.56$ with 22 d.f.; $p < .001$). It seems that the unconstrained solution is a better account of the data. From Figure 2 we can observe that the configuration not only diverges along the area dimension, but also is curved in an interesting way; curves connecting the rectangles with the same area levels shape like arcs drawn from a common focal point, so that the whole configuration looks like an open fan. This feature is not captured in any of Schönemann's hypotheses.

The alternative hypothesis to Schönemann's is proposed in Figure 6, which more closely simulates the above feature of the unconstrained solution. In terms of the polar coordinate system whose origin is placed at P, this hypothesis can be completely characterized by four angle parameters ($\theta_1 \sim \theta_4$) and five interval parameters ($r_1 \sim r_5$). Of course, we may further assume $\theta_1 = \theta_2 = \theta_3 = \theta_4$ and/or $r_2 = r_3 = r_4 = r_5$. Unfortunately MAXSCAL-2.1 is not yet up to fitting this hypothesis.

This hypothesis, however, is interesting in another respect; it gives important insight about the nature of the third dimension we discussed previously. We interpreted the third dimension as contrasting inside stimuli with outside stimuli, but an alternative interpretation is also possible; all stimuli are seen equally distant from an origin, just like stimuli with a same shape are at a constant distance from P in Figure 6. Note that all the stimuli employed were rectangles with constant number of sides, constant curvature of sides, etc. Since the plane at a constant distance from a point is necessarily curved in the euclidian space, the third dimension emerged. If more heterogeneous stimuli had been included in the experiment, the existence of the third dimension might have been more clearly substantiated.

Concluding Remarks

We have seen the kind of analyses which can take place with MAXSCAL-2.1. First of all we can perform either metric or nonmetric types of MDS. Alternatively, we may

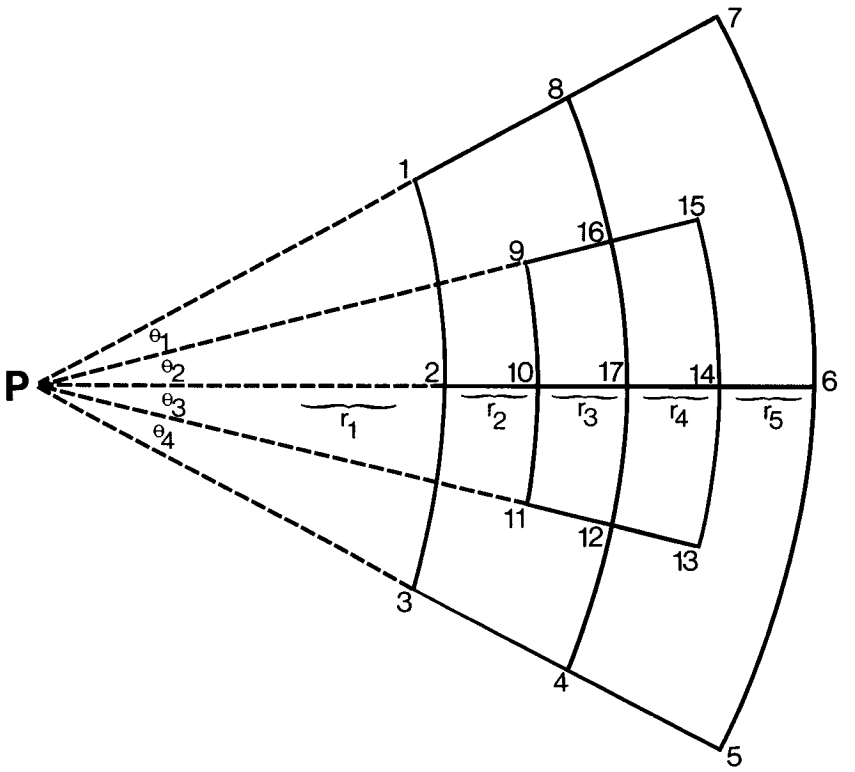

FIGURE 6
The proposed alternative hypothesis to Schönemann's.

perform both types of analyses to choose a more appropriate measurement assumption. The important point is that this can be done strictly on an empirical basis.

Secondly, we can perform MDS under two representative distributional assumptions about the error process. The minimum AIC criterion can be used to compare the goodness of fit of the two error models. The conventional significance testing procedure has little to say in a situation like this, where either one of the two models is not subsumed under the other.

The MAXSCAL-2.1 program has a constrained optimization feature. Thus, in addition to the test of dimensionality of the representation space and the test of appropriate restrictions on category boundaries we may perform the tests of various structural hypotheses on the stimulus configuration. To the best of the present author's knowledge MAXSCAL-2.1 is the first MDS program which is capable of imposing the kind of constraints discussed in this paper. The importance of this is in the fact that MAXSCAL-2.1 can provide suitable statistical criteria for identification of the best fitting model. Although the range of constraints that can be handled by MAXSCAL-2.1 is still limited, it should not be very difficult to extend its scope to the kinds of constraints discussed by Bentler and Weeks [1978] and Bloxom [1978]. (See also Ramsay, 1980a.)

The chief advantage of collecting dissimilarity data by rating scales with a relatively small number of observation categories lies in the existence of the minimum sufficient statistic for p_{ijkm}. Specifically, Y_{ijkm}/n_{ijkm} is the maximum likelihood estimate of p_{ijkm}, when no further structural assumption is made about p_{ijkm}. Thus, Y_{ijkm}/n_{ijkm}, as a model of p_{ijkm}, may be called the null model. We may then ask a question about the adequacy of the (euclidian) distance representation (5) of the data against the null model. This can be done by comparing the goodness of fit of the two models. The log likelihood of the

null model is naturally very large ($-2 \ln L = 1127.90$), but it uses 816 parameters to achieve this. The best solution under model (5) (i.e., the three-dimensional unconstrained solution), on the other hand, has 1543.10 for $-2 \ln L$ with 51 d.f. The difference between the two models is not significant ($\chi^2 = 415.20$ with 765 d.f.). Thus it seems fair to conclude that the distance representation is adequate in this instance. Note also that the AIC value of 2759.90 for the null model is the worst among all solutions obtained. (See Table 1.)

There are a number of things left to be done in the future, of which the following two are perhaps worth mentioning here. First, the small sample behavior of maximum likelihood estimates should be systematically investigated, as has been done for MULTISCALE [Ramsay, 1980b]. Second, the relative efficiency of collecting dissimilarity data using rating scales with a small number of response categories should be examined in relation to other data collection methods. The effect of the number of response categories has been studied [Ramsay, 1973; Green & Rao, 1970; Okada, Note 6] in somewhat different contexts. Now that MAXSCAL-2.1 is available, this can be done in the specific context of maximum likelihood multidimensional scaling.

REFERENCE NOTES

1. Takane, Y., & Carroll, J. D. *Maximum likelihood multidimensional scaling from directional rankings of similarities.* Paper submitted for publication, 1980.
2. Jennrich, R. I., & Moore, R. H. *Maximum likelihood estimation by means of nonlinear least squares* (RB-75-7). Princeton, N.J.: Educational Testing Service, 1975.
3. Takane, Y., & Carroll, J. D. *On the robustness of AIC in the context of maximum likelihood multidimensional scaling.* Manuscript in preparation.
4. Steiger, J. H., & Lind, J. M. *Statistically based tests for the number of common factors.* Handout for a paper presented at the Psychometric Society meeting, Iowa, 1980.
5. Krantz, D. H., & Tversky, A. *Similarities of rectangles: An analysis of subjective dimensions.* Ann Arbor, Mich.: Michigan Mathematical Psychology Program 73-8, 1973.
6. Okada, A. *Nonmetric multidimensional scaling and rating scales.* San Diego, California: Proceedings of the U.S.-Japan Seminar on Theory, Methods and Applications of Multidimensional Scaling and Related Techniques, 1975, 141-150.

REFERENCES

Akaike, H. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *The second international symposium on information theory.* Budapest: Akadémiai Kiado, 1973.

Akaike, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 1974, *19*, 716-723.

Akaike, H. On entropy maximization principle. P. R. Krishnaiah (Ed.). *Applications of Statistics.* Holland: North-Holland Publishing Co., 1977.

Beck, J. V., & Arnold, K. J. *Parameter estimation in engineering and science.* New York: Wiley, 1977.

Bentler, P. M., & Weeks, D. G. Restricted multidimensional scaling models. *Journal of Mathematical Psychology*, 1978, *17*, 138-151.

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

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

Carroll, J. D. Spatial, non-spatial and hybrid models of scaling. *Psychometrika*, 1976, *41*, 439-463.

Green, P. E., & Rao, V. R. Rating scales and information recovery—how many scales and response categories to use. *Journal of Marketing*, 1970, *34*, 33-39.

Hastings, C. *Approximation for digital computers.* Princeton, NJ: Princeton University Press, 1955.

Johnson, S. C. Hierarchical clustering schemes. *Psychometrika*, 1967, *32*, 241-254.

Kitagawa, G. On the use of AIC for the detection of outliers. *Technometrics*, 1979, *21*, 193-199.

Krantz, D. H., & Tversky, A. Similarity of rectangles; An analysis of subjective dimensions. *Journal of Mathematical Psychology*, 1975, *12*, 4-34.

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

Kruskal, J. B. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 1964, *29*, 115-129.

Kruskal, J. B. & Wish, M. *Multidimensional scaling.* Beverly Hills, Calif.: Sage Publications, 1978.

Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores.* Reading, Massachusetts: Addison-Wesley, 1968.

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

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

Ramsay, J. O. The effect of number of categories in rating scales on precision of estimation of scale values. *Psychometrika*, 1973, *38*, 513–532.

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

Ramsay, J. O. Confidence regions for multidimensional scaling analysis. *Psychometrika*, 1978, *43*, 145–160.

Ramsay, J. O. Joint analysis of direct ratings, pairwise preferences and dissimilarities. *Psychometrika*, 1980a, *45*, 149–165.

Ramsay, J. O. Some small sample results for maximum likelihood estimation in multidimensional scaling. *Psychometrika*, 1980b, *45*, 139–144.

Rao, C. R. *Advanced statistical methods in biometric research.* New York: Wiley, 1952.

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

Sakamoto, Y. & Akaike, H. Analysis of cross classified data by AIC. *Annals of the Institute of Statistical Mathematics*, 1978, *B30*, 185–197.

Schönemann, P. H. Similarity of rectangles. *Journal of Mathematical Psychology*, 1977, *16*, 161–165.

Schönemann, P. H., & Tucker, L. R. A maximum likelihood solution for the method of successive intervals allowing for unequal stimulus dispersions. *Psychometrika*, 1967, *32*, 403–418.

Schwartz, G. Estimating the dimensions of a model. *The Annals of Statistics*, 1978, *6*, 461–464.

Shepard, R. N. The analysis of proximities: Multidimensional scaling with an unknown distance function, I & II. *Psychometrika*, 1962, *27*, 125–140, & 219–246.

Shimizu, R. Entropy maximization principle and selection of the order of an autoregressive Gaussian process. *Annals of the Institute of Statistical Mathematics*, 1978, *A30*, 263–270.

Takane, Y. A maximum likelihood method for nonmetric multidimensional scaling: I. The case in which all empirical pairwise orderings are independent—theory. *Japanese Psychological Research*, 1978a, *20*, 7–17.

Takane, Y. A maximum likelihood method for nonmetric multidimensional scaling: II. The case in which all empirical pairwise orderings are independent—evaluations. *Japanese Psychological Research*, 1978b, *20*, 105–114.

Thurstone, L. L. *The measurement of values.* Chicago, Ill.: University of Chicago Press, 1959.

Torgerson, W. S. Multidimensional scaling: I. Theory and method. *Psychometrika*, 1952, *17*, 401–419.

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

Tversky, A. Features of similarity. *Psychological Review*, 1977, *84*, 327–352.

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

Young, G., & Householder, A. S. Discussion of a set of points in terms of their mutual distances. *Psychometrika*, 1938, *3*, 19–22.

Zinnes, J. L., & Wolff, R. P. Single and multidimensional same-different judgments. *Journal of Mathematical Psychology*, 1977, *16*, 30–50.

Manuscript received 3/19/80
Final version received 10/21/80