Journal of Mathematical Psychology **45**, 149–166 (2001)
doi:10.1006/jmps.1999.1300, available online at http://www.idealibrary.com on IDEAL®

# Determining the Dimensionality of Multidimensional Scaling Representations for Cognitive Modeling

Michael D. Lee

*Defence Science and Technology Organisation*

Multidimensional scaling models of stimulus domains are widely used as a representational basis for cognitive modeling. These representations associate stimuli with points in a coordinate space that has some predetermined number of dimensions. Although the choice of dimensionality can significantly influence cognitive modeling, it is often made on the basis of unsatisfactory heuristics. To address this problem, a Bayesian approach to dimensionality determination, based on the Bayesian Information Criterion (BIC), is developed using a probabilistic formulation of multidimensional scaling. The BIC approach formalizes the trade-off between data-fit and model complexity implicit in the problem of dimensionality determination and allows for the explicit introduction of information regarding data precision. Monte Carlo simulations are presented that indicate, by using this approach, the determined dimensionality is likely to be accurate if either a significant number of stimuli are considered or a reasonable estimate of precision is available. The approach is demonstrated using an established data set involving the judged pairwise similarities between a set of geometric stimuli. © 2001 Academic Press

## COGNITIVE MODELING AND MULTIDIMENSIONAL SCALING

Multidimensional scaling techniques (Shepard, 1962; Kruskal, 1964; see Cox & Cox, 1994, for a recent overview) generate spatial representations of stimulus sets based on information regarding the similarity relationships existing between the stimuli. Typically, each stimulus is identified with a point in a coordinate space such that the distance between representative points decreases as the similarity of the corresponding stimuli increases. The spatial nature of these representations means that they are well suited to practical application in the context of data visualization (e.g., Lowe & Tipping, 1996; Mao & Jain, 1995).

Address correspondence and reprint requests to Michael D. Lee, Communications Division, Defence Science and Technology Organisation, PO Box 1500, Salisbury SA, 5108 Australia. Fax: +61 8 8259 7110. E-mail: michael.d.lee@dsto.defence.gov.au.



0022-2496/01 $35.00
Copyright © 2001 by Academic Press
All rights of reproduction in any form reserved.

More fundamentally, however, multidimensional scaling representations have their origins in (Shepard, 1957), and some considerable status as, plausible models of human conceptual structure, particularly in relation to low-level, continuous sensory stimulus domains. Shepard (1987; see also Myung & Shepard, 1996; Shepard, 1994) has provided compelling empirical and theoretical evidence that the spatial representation of a stimulus domain generated by multidimensional scaling affords a simple and elegant characterization of the way in which the fundamental cognitive process of generalization operates within that domain. In particular, he argues that the generalization gradient that relates psychological similarity to distance in the representational space is invariant and closely approximated by the exponential decay functional form. Accordingly, multidimensional scaling representations have been employed as the underpinnings of a number of successful models, generically described as *cognitive process models* (Nosofsky, 1992), such as the Generalized Context Model (Nosofsky 1984, 1986), ALCOVE (Kruschke, 1992), and others (e.g., Getty, Swets, Swets, & Green 1979). In each of these models, the fixed spatial stimulus representations generated by multidimensional scaling—which are sometimes called *psychological spaces* in this context—are manipulated by processes that model cognitive phenomena such as identification and categorization.

The existence of an exponentially decaying relationship between psychological similarity and distance in a psychological space has the further computational advantage of allowing cognitive modeling to proceed on the basis of metric multidimensional scaling. Metric, as opposed to nonmetric, multidimensional scaling techniques require an additional assumption to be made regarding the form of the monotonically decreasing function that relates similarity measures to proximities and hence allows the problem of deriving a similarity-preserving spatial representation to be conceived in terms of deriving a proximity-preserving spatial representation. On the basis of Shepard's (1987) results, it may be appropriate to convert an empirically observed similarity measure between the $i$th and the $j$th stimuli, $s_{ij}$, into a target proximity measure $d_{ij}$, using the relationship $d_{ij} \propto -\log(s_{ij})$. This transformation does, theoretically, assume a ratio level of data scaling which may not always be appropriate, given the common practice of collecting similarity data on interval and ordinal scales. It is less clear, however, to what extent this theoretical deficiency manifests itself as a practical deficiency in terms of the appropriateness of the distance-like proximity[1] values it generates. In any case, from this perspective, the generation of multidimensional scaling representations for cognitive modeling involves the application of some form of optimization method to minimize an error measure of the form

$$E \propto \sum_{i<j} (d_{ij} - \hat{d}_{ij})^2, \tag{1}$$

where $\hat{d}_{ij}$ is the current distance between representative points $\mathbf{p}_i = (p_{i1}, ..., p_{im})$ and $\mathbf{p}_j = (p_{j1}, ..., p_{jm})$ in an $m$-dimensional representative space, as measured according

[1] The intended difference between the terms "proximity" and "distance" is that the former does not imply adherence to the triangle inequality.

to some distance metric. The distance metrics of interest are commonly restricted to the family of Minkowskian $r$-metrics[2], given by

$$\hat{d}_{ij} = \left[ \sum_{k=1}^{m} |p_{ik} - p_{jk}|^r \right]^{1/r}, \tag{2}$$

with a particular emphasis having been placed on the $r = 1$ (city-block) and $r = 2$ (Euclidean) cases because of their relationship, respectively, to so-called separable and integral stimulus domains (Garner, 1974). Recently, Shepard (1987, 1991) provided a compelling theoretical basis for this association, based on the correlation of the extension of different dimensions of abstract geometric structures termed *consequential regions*. Specifically, it is shown that perfect correlation of these structures, to be expected in the case of stimulus integrality, gives rise to circular equisimilarity contours indicating the operation of the Euclidean metric, while a complete lack of correlation, corresponding to separable component stimulus dimensions, produces diamond-shaped contours indicative of the city-block metric[3]. More generally, it has been argued (see Shepard, 1991, p. 61 for a list of references) that the distinction between separable and integral stimuli may represent endpoints of a continuum rather than a dichotomy and that, therefore, some stimulus domains may be modeled appropriately using an $r$ value between 1 and 2. Although Minkowski $r$-metrics with $r > 2$ are sometimes considered (e.g., Kruskal, 1964; Shepard, 1991), it is difficult to provide a psychological interpretation, in terms of the component structure, for stimuli modeled in this way. Pure integrality at $r = 2$ would seem to constitute a psychological upper limit on the degree to which underlying stimulus dimensions may be combined. In contrast, the adoption of metrics with $r < 1$ has been given a psychological justification (Gati & Tversky, 1982; see also Shepard 1987, 1991) in terms of modeling stimuli with component dimensions that compete for attention. These assertions are also consistent with Shepard's (1987, 1991) theory, since, as noted above, the limit of complete correlation corresponds to the Euclidean metric, but the use of consequential regions with negatively correlated degrees of extension correspond to Minkowski $r$-metrics with $r < 1$. It seems reasonable, therefore, to conclude that there is some psychological impetus for restricting the family of Minkowki $r$-metrics considered in cognitive modeling to the range $0 < r \leqslant 2$.

## DIMENSIONALITY DETERMINATION

While it has been argued that the minimization of the error measure given in Eq. (1) is a nontrivial optimization problem, particularly when the representational space is one-dimensional (see Shepard, 1974), a number of gradient descent based

[2] Although multidimensional scaling techniques have been developed (e.g., Cox & Cox, 1991; Lindman & Caelli, 1978) that operate in spaces not accommodated by the Minkowskian family of metrics.

[3] Despite the theoretical elegance of this association between stimulus structure and metric structure, it should be acknowledged that there are empirical findings (e.g., Maddox & Ashby, 1998; Potts, Melara, & Marks 1998) that question the adequacy of the simple form of the relationship often assumed.

(e.g., Demartines & Hérault, 1997; Kruskal, 1964) and other (e.g., Cohen, 1997; Klock & Buhmann, 1997) techniques have been proposed that appear to generate useful solutions. However, a second fundamental problem, that of determining the appropriate dimensionality of the coordinate space in which the representative points are to be embedded, has been less satisfactorily addressed.

The commonly advocated practice for dimensionality determination (e.g., Davison, 1983, pp. 91–92; Schiffman, Reynolds, & Young, 1981, pp. 10–13) is a heuristic one of seeking a relatively sharp drop or elbow in the pattern of change of the error measure across representational spaces of different dimensionalities[4]. Usually, this is accomplished through some form of visual inspection or by comparing obtained error values to baselines generated by Monte Carlo studies. Unfortunately, it has often been noted (e.g., Borg & Lingoes, 1987, p. 68; Grau & Nelson, 1988) that the pattern of change of data-fit across dimensionality, rather than containing an elbow, is often better characterized as one of smooth and gradual decline. Accordingly, additional prior information regarding the stimulus domain or similarity data, including notions of subjective substantive interpretation or beliefs regarding the precision of the data, are also often introduced into the dimensionality decision-making process in an informal and ad hoc manner.

Clearly, these practices can act to exaggerate the well-documented problems (Shepard, 1974) of deciding upon representations of spuriously low dimensionality to facilitate visualization or representations of inappropriately high dimensionality to accommodate almost all of the variance in empirical data. Similarly, in terms of cognitive modeling, it is imperative that a representation of the stimulus domain is chosen that captures only those relationships between stimuli that are important for the cognitive process being modeled. A failure to include all relevant dimensions of the underlying psychological representation will often mean that necessary information is simply not available, whereas the inclusion of spurious dimensions may interfere with the representational structure required to model a cognitive process.

From a more general modeling perspective, the issue of determining the appropriate dimensionality of a multidimensional scaling representation is a familiar and pervasive one, addressed by what is variously referred to as Ockham's Razor or the Principle of Parsimony. In essence, it concerns the trade-off between providing the ability for a model to accommodate the data that it must seek to explain while simultaneously ensuring the complexity of the model is minimized so that it is capable of generalization and prediction. One way of tackling this general dilemma, previously adopted in both psychological (e.g., Myung & Pitt, 1997) and other modeling, is through the adoption of Bayesian approaches to model selection (see Kass & Raftery, 1995, for an overview). The basic notion is one of considering the evidence provided for various models by a particular set of data, taking into account both the descriptive adequacy and the inherent complexity of those models. A measure of model goodness, known as the Bayesian Information Criterion (BIC)

[4] While there are at least two multidimensional scaling techniques (Lee, 1997; Shepard, 1962) that attempt to determine automatically the dimensionality of the representations they derive, it is fair to observe that neither could claim a rigorous and principled basis for this determination. Certainly, neither technique is widely employed to generate multidimensional scaling solutions.

(Schwarz, 1978), is perhaps the simplest of these approaches and seems well suited to addressing the problem of dimensionality determination in multidimensional scaling.

## A BAYESIAN INFORMATION CRITERION FORMULATION

The BIC takes the general form

$$\mathrm{BIC} = -2\log(p(\mathrm{ML})) + P\log N, \tag{3}$$

where $p(\mathrm{ML})$ is the maximum likelihood estimate of the model, $P$ is the number of parameters in the model, and $N$ is the sample size. Qualitatively, it can be seen that this measure increases whenever either model complexity, as measured by the number of model parameters, increases or when the model's accommodation of the data worsens. Accordingly, the candidate model with the minimal BIC value is to be preferred.

Clearly, however, to apply the BIC to multidimensional scaling, it is necessary to provide a probabilistic formulation of the data-fit exhibited by different spatial configurations. Following Tenenbaum's (1996) treatment of additive cluster modeling (Shepard & Arabie, 1979), this seems naturally achieved by assuming that the probability of a set of target proximities, given a particular spatial configuration of dimensionality $m$, has a Gaussian distribution with common variance $\sigma^2$,

$$p(\mathbf{D} \mid m, \hat{\mathbf{D}}) \propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i<j}(d_{ij}-\hat{d}_{ij})^2\right), \tag{4}$$

where $\mathbf{D} = [d_{ij}]$ and $\hat{\mathbf{D}} = [\hat{d}_{ij}]$.

It is important to note that this formulation, while not unrelated, differs significantly from previous probabilistic characterizations of multidimensional scaling developed by Ramsay (1977) and Takane (1978a), in that measures of similarity, rather than *dis*similiarity, are assumed to be log-normally distributed. The current proposal follows from the acceptance of Shepard's (1987) empirical and theoretical evidence in favor of an exponentially decaying relationship between similarity and distance, which allows similarity measures to be converted into proximities in a psychological space. In contrast, the previously proposed dissimilarity-based formulations seem to have been less rigorously motivated, as evidenced by statements such as "the author tends to favor lognormal distribution as a first guess at the behavior of dissimilarity data" (Ramsay, 1977, p. 246).

While, on this basis, it seems clear that the current probalistic formulation has a more principled grounding in the context of cognitive modeling it is, nevertheless, fair to observe that both Ramsay's (1977, 1978, 1980, 1982) and Takane's (1978a, 1978b) maximum likelihood multidimensional scaling techniques fare relatively well with regard to the issue of dimensionality determination. Within their probabilistic frameworks, a number of useful but underutilized approaches to comparing representations of different dimensionality are developed, including chi-squared tests and comparisons based on the Akaike Information Criterion (AIC) (Akaike,

1974). Indeed, it should be acknowledged that the AIC measure is closely related to the BIC measure currently being proposed (see, for example, Myung & Pitt, 1997). There are, however, considerable grounds for asserting the superiority of the current choice. First, it has been shown (Kashyap, 1980), in the context of autoregressive models, that the AIC has a finite probability of incorrect model selection that is never less than approximately 16%, even when infinite data are available. Second, the results of a very thorough series of comparative tests undertaken by Luetkepohl (1985) indicate that the BIC estimates the correct dimensionality of a model most often and generally outperforms the AIC on other performance measures. Importantly, in the context of the cognitive modeling application of multidimensional scaling, these differences are most pronounced when dealing with small sample sizes. Most fundamentally, as argued by Kass and Raftery (1995, p. 790) the AIC is appropriate "only if the precision of the prior is comparable to that of the likelihood." In terms of the construction of psychological spaces by multidimensional scaling, this is certainly not the case, since the prior assumption made explicit by Shepard (1987) is that each point in a multidimensional space is equally likely to be a representative point, effectively ensuring that derived solutions are almost entirely constrained by the available similarity data.

In any case, under the probabilistic interpretation given in Eq. (4), the logarithm of the maximum likelihood estimate is, ignoring an additive constant, proportional to the minimum error measure attained. In an $m$ dimensional space, a multidimensional scaling model effectively employs $mn$ parameters, since it contains $m$ coordinate values for each of the $n$ objects that have been derived. Furthermore, in an $n \times n$ symmetric dissimilarity matrix, these parameters are constrained by a total of $n(n-1)/2$ data values. Accordingly, a multidimensional scaling formulation of the BIC measure takes the form

$$\mathrm{BIC} = \frac{1}{s^2} \sum_{i<j} (d_{ij} - \hat{d}_{ij})^2 + mn \log\left(\frac{n(n-1)}{2}\right), \tag{5}$$

where $s$ is a sample estimate[5] of the data precision population parameter $\sigma$.

The intended role of $s$ is one of quantifying the *inherent* precision of the data, independent of its subsequent application to multidimensional scaling or any other type of cognitive representational analysis. As Zinnes and Mackay (1992, p. 36) summarize, in a slightly different context, "for individual analyses, the variances... reflect the degree of unfamiliarity or uncertainty that the individual has concerning the nature of the stimulus [proximities]... for group analyses, the variances instead indicate how heterogenous the people in the group are with respect to their perception of stimulus [proximities]."

In terms of the empirical collection of similarity and dissimilarity data, the second of these conceptions is the most important, since the established and prevailing practice (e.g., Ekman, 1954; Gati & Tversky, 1982; Gregson, 1976;

[5] It is important to distinguish between the use of the symbol $s$ as a sample estimate of the precision of the data, and the use of $s_{ij}$ to denote the similarity between a pair of stimuli. Fortunately, there is no ambiguity, since the presence or absence of subscripts distinguishes the two cases.

Johnson & Tversky, 1984; Kruschke, 1993) is to form similarity or dissimilarity matrices by averaging the individual ratings or confusion probabilities of a large number of subjects. That is, given a set of individual proximity matrices $\mathbf{D}^k = [d_{ij}^k]$ derived from the data collected from each of $k = 1, 2, ..., K$ subjects, it is the averaged proximity matrix $\mathbf{D} = \frac{1}{K}[\sum_k d_{ij}^k] = [d_{ij}]$ that is used to generate a multidimensional scaling representation. In this case, the natural approach to determining $s$ is to calculate the average of the standard deviations for each of the pooled cells in the final proximity matrix, as follows:

$$s = \frac{1}{n(n-1)/2} \sum_{i<j} \sqrt{\frac{\sum_k (d_{ij}^k - d_{ij})^2}{K-1}}. \tag{6}$$

This estimate of data precision is entirely determined by the raw data and may be calculated before fitting multidimensional scaling representations with different dimensionalities to the averaged proximity matrix. The evaluation of BIC measures for each of these candidate representations is then straightforward, requiring the substitution of $s$ into Eq. (5) and using the known parametric complexities and residual errors. The representation with the minimal BIC value may then be taken as constituting an appropriate compromise between the need to accommodate the original data and the requirement to minimize the dimensionality of the representational model.

It is worth emphasizing the role of $s$ in forcing an explicit and quantitative estimate of data precision to be made as part of the complexity analysis. Through the averaging process, it is possible for two proximity matrices to be identical in terms of their individual entries, but to have different associated levels of precision. Under the approach being advocated here, these two matrices may demand multidimensional scaling representations with different levels of complexity. This allows precise data collected, say, from domain experts exhibiting close agreement in their judgments to be represented using many dimensions, while ensuring that less precise data are not over-fit by a similarly complicated representation.

The data precision estimate $s$ also offers some promise in addressing concerns raised by Ashby, Maddox, and Lee (1994) regarding the interpretation of multidimensional scaling analyses that use pooled data. These authors note that multidimensional scaling models are often observed to provide less convincing accounts of single-subject data than corresponding averaged data and present an insightful theoretical account of this phenomenon in terms of the geometrical effects of the averaging process. In general, there would seem to be two important potential sources of disagreement between subjects in observed similarity or dissimilarity judgments of stimuli. Besides fundamental differences in the perceived psychological similarity or dissimilarity of presented stimuli, there is also the possibility of decisional noise being introduced by crude measurement instruments, such as low-resolution ratings scales. To the extent that averaging helps smooth the noise arising from this decision-making process, it may well be justified, although a better solution would be to use measurement instruments that are not arbitrarily quantized. If it is the more fundamental perceptual source of disagreement that is being

alleviated, however, it is clear that "averaging across subjects changes the underlying psychological structure of the data" (Ashby *et al.*, 1994, p. 147), and the entire procedure is invalidated. The estimated value of data precision given by $s$ should be of some utility in this regard, since it measures the extent to which the similarity or dissimilarity judgments of individual subjects differ and provides a basis on which to decide whether a pooled proximity matrix is sufficiently precise to justify a multidimensional scaling representation.

## MONTE CARLO EVALUATION

As a preliminary examination of the validity, robustness, and general behavior of the BIC measure for determining the dimensionality of multidimensional scaling representations, a Monte Carlo study was undertaken. For 10, 20, 40, and 80 point domains, a total of 25 artificial configurations was constructed in spaces of each dimensionality between 1 and 7 by independently and uniformly selecting random points within the unit hypercube of the required dimensionality[6]. The matrix of interpoint distances $\mathbf{D}$ was then generated according to a given Minkowski $r$-metric.

For each of these configurations, the so-called pinning variant of gradient descent (Demartines & Hérault, 1997) was employed to generate 10 multidimensional scaling solutions, using different initial locations, in spaces of each of the seven possible dimensionalities. The iterative learning rule by which this was accomplished took the form

$$p_{jk}^{new} = p_{jk}^{old} + \lambda(d_{ij} - \hat{d}_{ij})\, \hat{d}_{ij}^{1-r}\, |p_{ik} - p_{jk}|^{r-1}\, \mathrm{sgn}(p_{ik} - p_{jk}) \qquad \forall j \neq i, \tag{7}$$

where $\lambda$ is a learning rate parameter fixed at 0.05, and $\mathrm{sgn}(\cdot)$ is the signum function.

The effectiveness of this optimization procedure was verified by examining its ability to recover noise-free configurations generated using both city-block and Euclidean metrics. For each of the $7 \times 25 = 175$ configurations, one of the 10 alternative solutions of the correct dimensionality was found that accounted for at least 99.999% of the variance of the data. Visual inspection of several of the low dimensional solutions confirmed that the original relationship between representative points had been recovered. Although Procrustes analysis (Sibson, 1978) could have been employed to provide similar confirmation for higher dimensional Euclidean cases, this seemed unnecessary given the accuracy of the recovery. The evident ability of the optimization approach to recover non-Euclidean configurations is particularly important, given doubts raised (Arabie, 1991; Hubert, Arabie, & Hesson-Mcinnis, 1992) about the capability of many multidimensional scaling techniques in this regard.

The ability of BIC measures to determine known spatial dimensionality in the more realistic case of noise-peturbed target proximities was examined using configurations that incorporated the independent cell-wise addition of zero mean

[6] The selection of an upper limit of 7 is somewhat arbitrary, but seems to constitute an appropriate overestimate of the dimensionality of useful multidimensional scaling models.

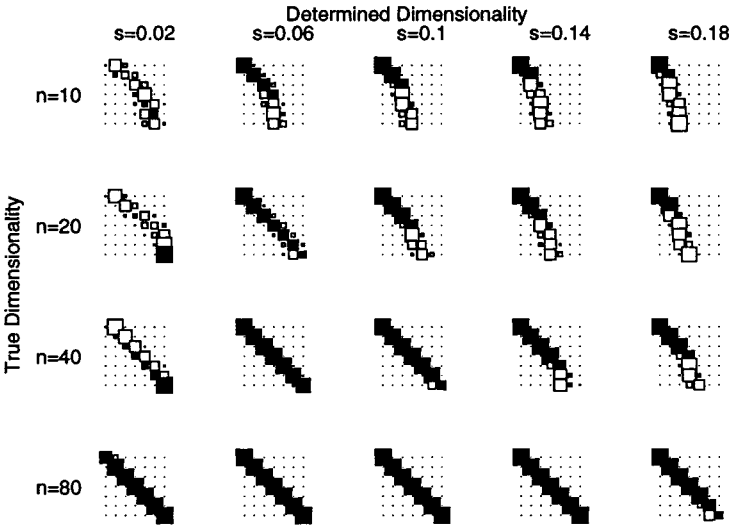


**FIG. 1.** Confusion matrices, using different values of $s$, for the noise peturbed 10, 20, 40, and 80 object city-block configurations.

Gaussian noise with a standard deviation[7] of 0.10. The minimal error value across the 10 alternatives for each of the seven dimensions was used to find BIC measures from which, upon specifying a particular estimate $s$ of $\sigma$, the minimum value was selected, indicating the determined dimensionality of the original space.

Figures 1 and 2 provide graphical summaries of the confusion matrices for this process of dimensionality determination in the city-block and Euclidean cases, respectively, as they vary across domain cardinality and accuracy of the estimate of data precision. As indicated, the rows in Figs. 1 and 2 correspond to the 10, 20, 40, and 80 object configurations, while the columns correspond to the specified choices of $s$. Each confusion matrix takes the form of a $7 \times 7$ grid, oriented in the conventional fashion, with the value of each cell being indicated by the size of the square located in the appropriate position on the grid. Squares corresponding to cells on the main diagonal, indicating correct dimensionality determination, are depicted in black, while the remaining squares are shown in white. To assist in the interpretation of this form of graphical representation, the upper-left confusion matrix in Fig. 1, resulting from dimensionality determination for the 10 object configuration using $s = 0.02$, is detailed in Table 1.

Figures 1 and 2 are very similar and indicate that the natural tendency for the BIC measure is to overestimate dimensionality when precision is overestimated and underestimate dimensionality when precision is underestimated. This pattern of results is to be expected, since the incorporation of additional dimensions in precise data is warranted, but corresponds to the fitting of noise when the belief in precision is mistaken. In contrast, too few dimensions are included when the mistaken

[7] This value of data precision was chosen as typical following the analysis of several real data sets, one of which is presented later. While preliminary analyses of the recovery properties of the BIC suggest that different levels of data precision do not result in significant changes, there is clearly scope for a more detailed investigation in this regard.

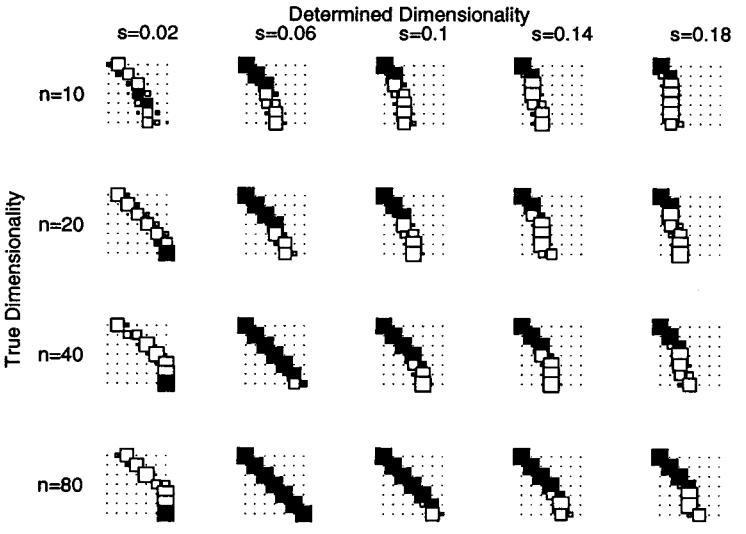


**FIG. 2.** Confusion matrices, using different values of $s$, for the noise peturbed 10, 20, 40, and 80 object Euclidean configurations.

assumption is made that the data do not have the level of precision required to justify the fitting of additional dimensions.

The practical implication of these results is to caution against the use of estimates of data precision that are not calculated directly from available raw data. There are clearly many factors that have the potential to influence the precision of empirically collected similarity or dissimilarity data. The various experimental methodologies commonly used to gather this data—ratings scales, sorting tasks, identification tasks, and so on—are likely to have different impacts upon precision. The nature of the stimulus domain under consideration, and the particular set of stimuli selected for an experimental task from this domain, will also influence the final precision of similarity or dissimilarity data. In addition, it seems possible that there may be significant second-order interactions between stimuli and methodology, so that similarity or dissimilarity data are more precisely gathered using different

**TABLE 1**

**Confusion Matrix, Using $s = 0.02$, in the 10 Object City-Block Domain**

| | | Determined dimensionality | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| True dimensionality | 1 | **1** | 18 | 6 | 0 | 0 | 0 | 0 |
| | 2 | 0 | **6** | 11 | 8 | 0 | 0 | 0 |
| | 3 | 0 | 0 | **1** | 15 | 9 | 0 | 0 |
| | 4 | 0 | 0 | 0 | **5** | 20 | 0 | 0 |
| | 5 | 0 | 0 | 0 | 3 | **8** | 13 | 1 |
| | 6 | 0 | 0 | 0 | 1 | 14 | **10** | 0 |
| | 7 | 0 | 0 | 0 | 0 | 7 | 15 | **3** |

experimental tasks for different stimulus sets. All of these possibilities place a caveat on the generalizability of an estimate $s$ calculated from one data set to another stimulus domain or to the same domain evaluated in a different way.[8]

Figures 1 and 2 also reveal that the characteristic pattern of dimensional under- and overestimation is most pronounced for configurations containing relatively few objects. Indeed, it is clear that the dimensionality of the 10 object domain is only determined accurately if it is, at most, two dimensional, while the 20 and 40 object domains seem to support dimensional upper limit of about 3 and 5 respectively. This conclusion concurs with established heuristic suggestions (see Schiffman, Reynolds, & Young, 1981, p. 24 for details), and better grounded empirical guidance (Rodgers, 1991), regarding the number of stimuli required in multidimensional scaling studies to generate spatial representations of various dimensionalities. In other words, it is accepted practice to limit, *a priori*, the dimensionality of derived multidimensional scaling representations on the basis of the cardinality of the stimulus set, and, within these bounds, the BIC measure is capable of accurate dimensionality determination when a moderately accurate estimate of data precision is supplied. In any case, the dependence on the accuracy of this estimate weakens as more objects are included in the configuration to the extent that the dimensionality of an 80 object configuration is determined correctly across something approaching an order of magnitude in the quantitative estimate made regarding data precision.

## DEMONSTRATION OF THE CRITERION

Kruschke (1993) examined the ability of the ALCOVE connectionist model to emulate human performance on various categorization tasks, using a geometric stimulus set consisting of rectangles that assumed one of four possible height values, with an interior segment that varied across four possible lateral positions. Of the 16 possible stimuli that could be generated by the exhaustive combination of these possibilities, only eight were selected as appropriate for examining the categorization phenomena of interest. The nature of the stimuli strongly suggests psychological separability, as confirmed by Kruschke's (1993, pp. 35–36) choice of the city-block metric to underpin the psychological representation used in modeling.

The representation used by Kruschke (1993) was constructed on the basis of empirical similarity ratings formed by averaging across a total of 400 observations for each possible stimulus pair. After converting each of the individual similarity

[8] More problematically, G. Ashby has noted that it is possible to construct an argument whereby, in extreme cases, the true, but unknown dimensionality of a stimulus domain influences data precision. A geometric property of coordinate spaces is that as dimensionality increases, the variance of interpoint distance measures tends to decrease, although this tendency remains sensitive to the actual spatial configuration under consideration. To the extent that stimulus similarity or dissimilarity does vary less for instrinsically high-dimensional stimulus domains, however, it is reasonable to expect the judgments of subjects to be less precise. Obviously, this possibility introduces a self-defeating circularity into the proposed method of dimensionality determination. In practice, however, it seems reasonable to expect that the influence of the stimulus set and experimental methodology will dominate any effect of dimensionality to the extent that it may safely be regarded as negligible.

**TABLE 2**

**Standard Deviations of Each of the Normalized Target Proximities for the Rectangle Domain, Based on the Raw Similarity Ratings from Kruschke's (1993) Experiment**

| | | Second stimulus | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| First stimulus | 1 | — | 0.080 | 0.084 | 0.126 | 0.122 | 0.160 | 0.140 | 0.137 |
| | 2 | 0.061 | — | 0.114 | 0.087 | 0.141 | 0.129 | 0.153 | 0.1389 |
| | 3 | 0.099 | 0.105 | — | 0.145 | 0.091 | 0.156 | 0.144 | 0.158 |
| | 4 | 0.105 | 0.106 | 0.136 | — | 0.163 | 0.097 | 0.160 | 0.131 |
| | 5 | 0141 | 0.153 | 0.080 | 0.169 | — | 0.150 | 0.111 | 0.131 |
| | 6 | 0.145 | 0.123 | 0.167 | 0.082 | 0.137 | — | 0.121 | 0.124 |
| | 7 | 0.124 | 0.154 | 0.113 | 0.166 | 0.094 | 0.152 | — | 0.052 |
| | 8 | 0.146 | 0.144 | 0.135 | 0.126 | 0.111 | 0.112 | 0.078 | — |

ratings to a proximity rating using an exponential decay transformation, the standard deviations for the normalized proximity between each pair of stimuli were calculated and are shown in Table 2. It can be seen that the standard deviations are both reasonably similar and do not seem to exhibit any systematic or marked deviation from symmetry. On this basis, following the practice of Kruschke (1993) with regards to the similarity measures, the matrix shown in Table 2 was made symmetric by pairwise averaging, resulting in a range of standard deviations between 0.065 and 0.166. The average value of these standard errors was calculated according to Eq. (6), giving an estimate of data precision $s = 0.125$ for this data set.

The same multidimensional scaling algorithm developed for the Monte Carlo analysis was used to find representations in spaces having between one and seven dimensions, all operating under the city-block metric. The residual errors obtained from this process, together with the known parametric complexities and estimate of data precision, were then substituted into Eq. (5) to give BIC values for each of the candidate representations. In addition, BIC values for data precision estimates of $s = 0.10$ and $s = 0.15$ were obtained, so that the sensitivity of the pattern of change in the BIC to the $s$ value estimated directly from the data could be examined. While, as argued earlier, it is important that the estimate of data precision be constrained by the raw data available from the empirical process of collection, it is also prudent to examine the extent to which dimensionality determination relies on this estimate. Even using the same collection methodology with the same stimulus set, it is likely that estimates of data precision will vary from experiment to experiment. The peturbation of the estimated $s$ value provides a simple mechanism for examining the robustness of the conclusions of a complexity analysis to these effects.

The results of this analysis for the Kruschke (1993) data are shown in Fig. 3, which shows the pattern of change of the BIC as the dimensionality of the underlying representational space increases, for each of the three estimates of data precision. Also shown, against the right hand scale, is the percentage of variance in the target proximity data explained by the best fitting configuration of each dimensionality. It can be seen that the minimal BIC measure for each of the assumptions regarding data precision occurs when the dimensionality is two, corresponding to

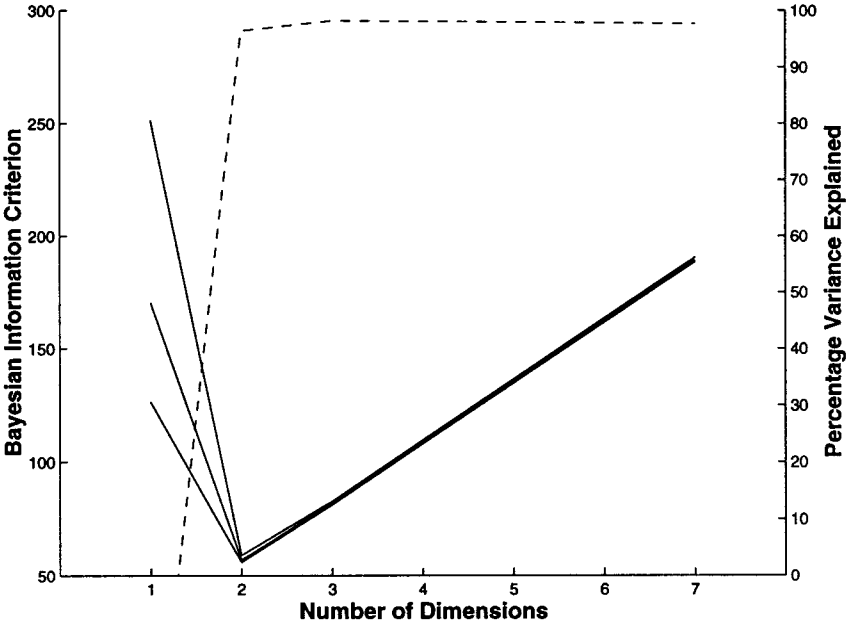


**FIG. 3.** BIC values (left hand scale) for the rectangle data, using $s$ values of 0.10, 0.125, and 0.15 (top to bottom). The variance explained (right hand scale) for the best fitting configuration of each dimensionality is shown by the broken line.

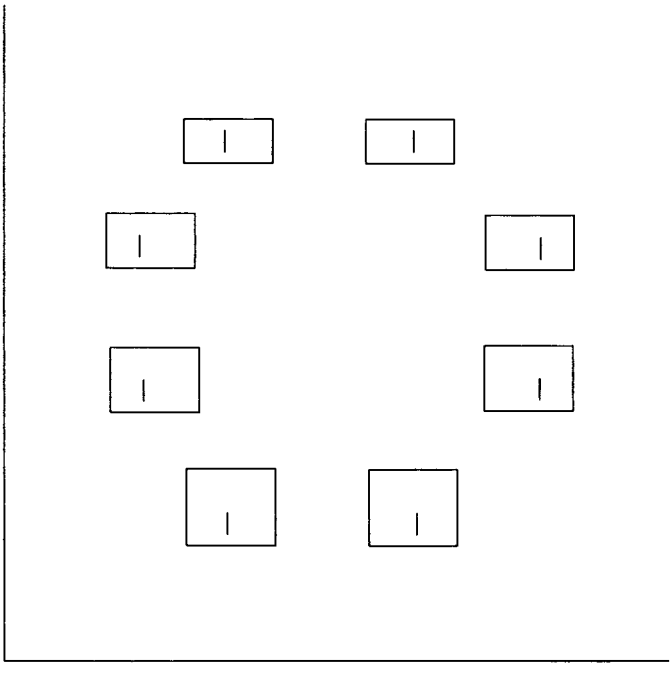

**FIG. 4.** Best fitting two dimensional configuration for the rectangle domain (cf. Kruschke 1993, Fig. 8).

a marked elbow in the variance explained index of data accommodation. It is reasonable to note that, for this data set, the appropriate dimensionality could have been heuristically determined solely on the basis of a measure of data-fit with considerable confidence. In general, however, the BIC approach provides a principled and objective means to make this decision for data that are more problematic and, unlike the detection of an elbow by a human observer, are readily amenable to automation. The choice of the Kruschke (1993) data to demonstrate the application of the BIC was largely based on the availability of the raw data needed to estimate $s$ directly. Unfortunately, the great majority of published similiarity or proximity matrices obtained by averaging across subjects do not provide information regarding the variances of the individual pairwise measures.

In any case, the best fitting two dimensional configuration is shown in Fig. 4, with depictions of each of the stimuli being placed at their derived locations. The spatial model is entirely consistent with the explicit procedure by which the stimuli were constructed, in the sense that the $x$-axis corresponds to the lateral position of the interior segments while the $y$-axis corresponds to the height of the rectangles. It is also worth noting that the representation shown in Fig. 4 is visually identical to that presented by Kruschke (1993, Fig. 8), upon which the reported cognitive process modeling was based.

## DISCUSSION

The Monte Carlo evaluation and practical application presented above suggest that the BIC has considerable utility in terms of evaluating and developing multidimensional scaling representations for cognitive modeling. When a spatial configuration appropriately represents the similarity structures observed within a domain, the BIC measure provides a simple and principled means of evaluating the relative evidence in favor of configurations with different dimensionalities. In this way, a model may be formulated with sufficient dimensionality to accommodate the similarity relationships without introducing the unnecessary complexity of superfluous dimensions. Furthermore, the evaluation of the BIC allows for—indeed requires—information regarding the precision of the available data to be explicitly and rationally introduced into the decision process of dimensionality determination. For these reasons, the BIC approach to dimensionality determination seems to constitute a significant advance on the heuristic search for error 'elbows' that characterizes current practice.

There are, however, several shortcomings of the suggested approach that should be acknowledged. First, it is important to understand that the BIC measure represents only an approximation to the true Bayesian measure of model goodness (see Kass & Raftery, 1995, p. 778). In particular, the BIC considers only what Myung and Pitt (1997) term the *number of parameters* component of model complexity and, as a consequence, measures model complexity solely in terms of increases in dimensionality. A consequence of this is that the BIC is insensitive to the component of model complexity Myung and Pitt (1997) term the *functional form* component, which relates to the different levels of complexity that are

involved in different forms of parametric interaction within a model. In the context of multidimensional scaling representations, this is a significant shortcoming, since there are both theoretical grounds and empirical evidence (Shepard, 1974) suggesting that non-Euclidean distance metrics are generally capable of achieving more error-free representations than the Euclidean metric. Presumably, these differences arise from variations in the inherent complexity of the way in which the coordinate locations of a spatial representation interact, as dictated by different distance metrics. Accordingly, the use of the BIC to compare spatial representations generated within coordinate spaces using different distance metrics is not justified. Clearly, given the importance of being able to use multidimensional scaling representations of integral, separable, and other stimulus domains in cognitive modeling, the extension of the general Bayesian framework for model comparison to consider this issue constitutes a worthwhile topic for future research.

A second, more general, shortcoming concerns the need to retain the criterion of substantive interpretability in model generation noted by Shepard (1974). While, as argued in the Introduction, it is inappropriate to base dimensionality decisions solely upon grounds of interpretability, the BIC measure is only sensitive to evidence related to the data-fit and parametric complexity of the derived model. Any exercise in model building should be constrained by both theory and data, which means that both substantive interpretation (theory) and a measure such as the BIC (data) should be taken into account when considering the relative merits of various multidimensional scaling representations.

## ACKNOWLEDGMENTS

I thank Kenneth Pope and Chris Woodruff for helpful discussions and Greg Ashby, In Jae Myung, and Joe Rodgers for their comments on earlier versions of this article.

## REFERENCES

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

Arabie, P. (1991). Was Euclid an unnecessarily sophisticated psychologist? *Psychometrika*, **56**, 567–587.

Ashby, F. G., Maddox, W. T., & Lee, W. W. (1994). On the dangers of averaging across subjects when using multidimensional scaling or the similarity-choice model. *Psychological Science*, **5**(3), 144–151.

Ashby, F. G. & Perrin, N. A. (1988). Toward a unified theory of similarity and recognition. *Psychological Review*, **95**(1), 124–150.

Borg, I., and Lingoes, J. (1987). *Multidimensional similarity structure analysis*. New York: Springer-Verlag.

Cohen, J. D. (1997). Drawing graphs to convey proximity: An incremental arrangement method. *ACM Transactions on Computer-Human Interaction*, **4**(3), 197–229.

Cox, T. F., & Cox, M. A. A. (1991). Multidimensional scaling on a sphere. *Communications in Statistics: Theory and Methods*, **20**, 2943–2953.

Cox, T. F., & Cox, M. A. A. (1994). *Multidimensional scaling*. London: Chapman and Hall.

Davison, M. L. (1983). *Multidimensional scaling*. New York: Wiley.

Demartines, P., & Hérault, J. (1997). Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets. *IEEE Transactions on Neural Networks*, **8**, 148–154.

Ekman, G. (1954). Dimensions of color vision. *The Journal of Psychology*, **38**, 467–474.

Ennis, D. M. (1988a). Confusable and discriminable stimuli: Comment on Nosofsky (1986) and Shepard (1986). *Journal of Experimental Psychology*: *General*, **117**, 408–411.

Ennis, D. M. (1988b). Toward a universal law of generalization. *Science*, **242**, 944.

Ennis, D. M. (1992). Modeling similarity and identification when there are momentary fluctuations in psychological magnitudes, In F. G. Ashby (Ed.). *Multidimensional models of perception and cognition*, pp. 279–298. Hillsdale, NJ: Erlbaum.

Garner, W. R. (1974). *The processing of information and structure*. Potomac, MD: Erlbaum.

Gati, I., & Tversky, A. (1982). Representations of qualitative and quantitative dimensions. *Journal of Experimental Psychology*: *Human Perception and Performance*, **8**, 325–340.

Getty, D. J., Swets, J. A., Swets, J. B., & Green, D. M. (1979). On the prediction of confusion matrices from similarity judgements. *Perception & Psychophysics*, **26**, 1–19.

Grau, J. W., & Nelson, D. K. (1988). The distinction between integral and separable dimensions: Evidence for the integrality of pitch and loudness. *Journal of Experimental Psychology*: *General*, **117**, 347–370.

Gregson, R. A. M. (1976). A comparative evaluation of seven similarity models. *British Journal of Mathematical and Statistical Psychology*, **29**, 139–156.

Hubert, L., Arabie, P., & Hesson-Mcinnis, M. (1992). Multidimensional scaling in the city-block metric: A combinatorial approach. *Journal of Classification*, **9**, 211–236.

Johnson, E. J., & Tversky, A. (1984). Representations of perceptions of risks. *Journal of Experimental Psychology*: *General*, **113**, 55–70.

Kashyap, R. L. (1980). Inconsistency of the AIC rule for estimating the order of autoregressive models. *IEEE Transactions on Automatic Control*, 25, 996–998.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.

Klock, H. J., & Buhmann, J. (1997). Multidimensional scaling by deterministic annealing. In *Proceedings of the International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, *Venice*, 1997.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning, *Psychological Review*, **99**, 22–44.

Kruschke, J. K. (1993). Human category learning: Implications for backpropagation models. *Connection Science*, **5**, 3–36.

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

Lee, M. D. (1997). The connectionist construction of psychological spaces. *Connection Science*, **9**, 323–351.

Lindman, H., & Caelli, T. (1978). Constant curvature Riemannian scaling. *Journal of Mathematical Psychology*, **17**, 89–109.

Lowe, D., & Tipping, M. E. (1996). Feed-forward neural networks and topographic mappings for exploratory data analysis. *Neural Computing and Applications*, **4**, 83–95.

Luetkepohl, U. (1985). Comparison of criteria for estimating the order of a vector autoregressive process. *Journal of Time Series Analysis*, **6**, 35–52.

Maddox, W., & Ashby, F. (1998). Selective attention and the formation of linear decision boundaries: Comment on McKinley and Nosofsky (1996). *Journal of Experimental Psychology*, **24**, 301–321.

Mao, J., & Jain, A. K. (1995). Artificial neural networks for feature extraction and multivariate data projection. *IEEE Transactions on Neural Networks*, **6**, 296–317.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, **4**, 79–95.

Myung, I. J., & Shepard, R. N. (1996). Maximum entropy inference and stimulus generalization. *Journal of Mathematical Psychology*, **40**, 342–347.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology*: *Learning*, *Memory*, *and Cognition*, **10**, 104–114.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology*: *General*, **115**, 39–57.

Nosofsky, R. M. (1988). On exemplar-based representations: Reply to Ennis (1988). *Journal of Experimental Psychology*: *General*, **117**, 412–414.

Nosofsky, R. M. (1992). Similarity scaling and cognitive process models. *Annual Review of Psychology*, **43**, 25–53.

Potts, B., Melara, R. D., & Marks, L. (1998). Circle size and diameter tilt: A new look at integrality and separability. *Perception & Psychophysics*, **60**, 101–112.

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

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

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

Ramsay, J. O. (1982). Some statistical approaches to multidimensional scaling data. *Journal of the Royal Statistical Society*: *Series A*, **145**, 285–312.

Rodgers, J. L. (1991). Matrix and stimulus sample sizes in the weighted MDS model: Empirical metric recovery functions. *Applied Psychological Measurement*, **15**, 71–77.

Schiffman, S. S., Reynolds, M. L., & Young, F. W. (1981). *Introduction to multidimensional scaling*. New York: Academic Press.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 461–464.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, **22**, 325–345.

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

Shepard, R. N. (1974). Representation of structure in similarity data: Problems and prospects. *Psychometrika*, **39**, 373–422.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, **237**, 1317–1323.

Shepard, R. N. (1988). Time and distance in generalization and discrimination: Reply to Ennis (1988). *Journal of Experimental Psychology*: *General*, **117**, 415–416.

Shepard, R. N. (1991). Integrality versus separability of stimulus dimensions: From an early convergence of evidence to a proposed theoretical basis, In J. R. Pomerantz & G. L. Lockhead (Eds.). *The perception of structure*: *Essays in honor of Wendell R Garner* (pp. 53–71). Washington, DC: American Psychological Association.

Shepard, R. N. (1994). Perceptual-cognitive universals as reflections of the world. *Psychonomic Bulletin & Review*, **1**, 2–28.

Shepard, R. N., & Arabie, P. (1979). Additive clustering representations of similarities as combinations of discrete overlapping properties. *Psychological Review*, **86**, 87–123.

Sibson, R. (1978). Studies in the robustness of multidimensional scaling: Procrustes statistics. *Journal of the Royal Statistical Society*, *Series B*, **40**, 234–238.

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

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

Tenenbaum, J. B. (1996). Learning the structure of similarity, In D. S. Touretzky, M. C. Mozer, & M. E. Hasselmo (Eds.). *Advances in neural information processing systems*, Volume 8, pp. 3–9. Cambridge, MA: MIT Press.

Zinnes, J. L., & MacKay, D. B. (1992). A probabilistic multidimensional scaling approach: Properties and procedures, In F. G. Ashby (Ed.). *Multidimensional Models of Perception and Cognition* (pp. 35–88). Hillsdale, NJ: Erlbaum.

Received: April 16, 1998