Journal of Mathematical Psychology **45**, 131–148 (2001)
doi:10.1006/jmps.1999.1299, available online at http://www.idealibrary.com on IDEAL®

On the Complexity of Additive Clustering Models

Michael D. Lee

Defence Science and Technology Organisation

Additive clustering provides a conceptually simple and potentially powerful approach to modeling the similarity relationships between stimuli. The ability of additive clustering models to accommodate similarity data, however, typically arises through the incorporation of large numbers of parameterized clusters. Accordingly, for the purposes of both model generation and model comparison, it is necessary to develop quantitative evaluative measures of additive clustering models that take into account both data-fit and complexity. Using a previously developed probabilistic formulation of additive clustering, the Bayesian Information Criterion is proposed for this role, and its application demonstrated. Limitations inherent in this approach, including the assumption that model complexity is equivalent to cluster cardinality, are discussed. These limitations are addressed by applying the Laplacian approximation of a marginal probability density, from which a measure of cluster structure complexity is derived. Using this measure, a preliminary investigation is made of the various properties of cluster structures that affect additive clustering model complexity. Among other things, these investigations show that, for a fixed number of clusters, a model with a strictly nested cluster structure is the least complicated, while a model with a partitioning cluster structure is the most complicated.

INTRODUCTION

Additive clustering models (e.g., Arabie & Carroll, 1980; Chaturvedi & Carroll, 1994; Mirkin, 1987; Shepard & Arabie, 1979; Tenenbaum, 1996) provide powerful yet conceptually simple accounts of the observed similarities between sets of stimuli. Given a matrix of pairwise similarities $\mathbf{S} = [s_{ij}]$, additive clustering derives a set of weighted stimulus clusters, which may, in various contexts, also be interpreted as domain classes or features. What distinguishes additive clustering from other clustering approaches is that the relationship between the given set of stimuli and the derived clusters is entirely unconstrained. As Shepard and Arabie (1979, p. 91) argue "generally, the discrete psychological properties of objects overlap in arbitrary

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

ways." For this reason, unlike standard partitioning clustering approaches that place each stimulus in only one cluster, additive clustering allows each stimulus to belong to any number of clusters. Furthermore, unlike hierarchical clustering approaches that allow stimuli to lie in more than one cluster, additive clustering places no constraints upon the set of stimuli that may be encompassed by a cluster.

The similarity model underpinning additive clustering, introduced by Arabie and Shepard (1973; see also Shepard, 1974; Shepard & Arabie, 1979), assumes that the similarity between any given pair of stimuli is determined by the number of clusters to which both stimuli belong. Formally, if the m derived clusters for n stimuli are defined by an $n \times m$ matrix of binary membership variables $\mathbf{F} = [f_{ik}]$, where

$$f_{ik} = \begin{cases} 1 & \text{if stimulus } i \text{ is in cluster } k \\ 0 & \text{otherwise,} \end{cases}$$

and the kth cluster is assigned a weight w_k, denoting its importance or salience, then the estimated similarity of the ith and jth stimuli is

$$\hat{s}_{ij} = \sum_k w_k f_{ik} f_{jk} . \tag{1}$$

Typically, the patterns of cluster membership and the weights extracted from a given similarity matrix are determined by minimizing an error measure of the form

$$E = \sum_{i<j} (s_{ij} - \hat{s}_{ij})^2.$$

It is generally recognized that the binary nature of the cluster membership variables makes this a difficult optimization problem and, accordingly, a wide variety of extraction techniques have been proposed, including mathematical programming (Arabie & Carroll, 1980), qualitative factor analytic (Mirkin, 1987), and probabilistic expectation-maximization (Tenenbaum, 1996) approaches. While all of these techniques have shortcomings, it is probably fair to suggest that they generally achieve sufficiently good minima to derive models of some theoretical and practical utility.

The focus of this paper, however, is on another difficulty in deriving additive clustering models, relating to the need to control model complexity, which is less satisfactorily addressed by established techniques. As noted by Shepard and Arabie (1979, p. 98), the ability to specify an arbitrarily overlapping cluster structure, when coupled with the ability to manipulate cluster weightings, enables any similarity structure to be accommodated perfectly by an additive clustering model. This means that E can always be reduced to zero or, equivalently, that the variance of the similarity data accounted for by the model, which is measured by

$$v = 1 - \frac{E}{\sum_{i<j} (s_{ij} - \bar{s})^2}, \tag{2}$$

where $\bar{s}$ is the arithmetic mean of the similarity values, can always assume unity.

While the modeling flexibility afforded by additive clustering is clearly desirable in terms of providing an ability to accommodate similarity data, the introduction of unconstrained and parameterized cluster structures potentially detracts from other fundamental modeling goals such as the achievement of interpretability, explanatory insight, and the ability to generalize accurately beyond given information.

This familiar conflict between maximizing data-fit and minimizing model complexity is often acknowledged in the development of previous techniques and has typically been tackled through the general strategy of attempting to use a minimal number of clusters to account for a maximal degree of similarity structure variance. Some techniques (e.g., Tenenbaum, 1996) accomplish this by setting the number of clusters to be derived at a fixed value and then seeking the best data-fit possible, while other techniques (e.g., Lee, in press b) set a target data-fit level and then seek a minimal number of clusters that achieve this fit.

No established technique, however, explicitly quantifies the trade-off between accuracy and complexity during the process of model generation. Consequently, the criteria by which the data-fit benefits of, for example, including an extra cluster are weighed against the concurrent complexity drawbacks are difficult to elucidate in any precise or practically useful form. Furthermore, without a quantitative measure, it is difficult to compare different additive clustering models when one model provides better data-fit than another, but relies upon a greater number of clusters to achieve this fit. Clearly, the articulation of an evaluative measure of a model that considers both the data-fit and the number of clusters would be useful in the generation and comparison of additive clustering models. The first goal of this paper is to suggest that, when the complexity of additive clustering models is equated with cluster cardinality, the application of the Bayesian Information Criterion (BIC) provides such a measure.

APPLICATION OF THE BAYESIAN INFORMATION CRITERION

The BIC is an established and well understood measure which incorporates both data-fit and model complexity (Schwarz, 1978; see Kass & Raftery, 1995; Myung & Pitt, 1997 for overviews). For a particular model A, the BIC takes the general form

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

where $p(\mathrm{ML}_A)$ 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, in terms of both model development and comparison, the candidate model with the minimal BIC value is to be preferred.

Cast in terms of additive clustering models, the maximum likelihood estimate is the probability of a similarity matrix $\mathbf{S}$, given the derived cluster matrix $\mathbf{F}$, and associated weight values $\mathbf{w} = (w_1, ..., w_m)$. An appropriate formulation of this

probability is provided by Tenenbaum (1996), in which it is assumed that $p(\mathbf{S}\,|\,\mathbf{F}, \mathbf{w})$ has a Gaussian distribution with common variance σ^2, as follows:

$$p(\mathbf{S}\,|\,\mathbf{F}, \mathbf{w}) = \prod_{i<j} \frac{1}{(\sigma\sqrt{2\pi})} \exp\left(-\frac{(s_{ij}-\hat{s}_{ij})^2}{2\sigma^2}\right)$$
$$= \frac{1}{(\sigma\sqrt{2\pi})^{n(n-1)/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i<j}(s_{ij}-\hat{s}_{ij})^2\right). \quad (3)$$

An additive clustering instantiation of the BIC may be generated using this framework, by equating the number of parameters in an additive clustering model with the number of clusters and observing that a similarity matrix for n stimuli incorporates $n(n-1)/2$ measures, to give

$$\mathrm{BIC}_A = -2\log\left\{\frac{1}{(\sigma\sqrt{2\pi})^{n(n-1)/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i<j}(s_{ij}-\hat{s}_{ij})^2\right)\right\} + m\log\left(\frac{n(n-1)}{2}\right)$$
$$= -2\log\left\{\frac{1}{(\sigma\sqrt{2\pi})^{n(n-1)/2}}\right\} - 2\log\left\{\exp\left(-\frac{1}{2\sigma^2}\sum_{i<j}(s_{ij}-\hat{s}_{ij})^2\right)\right\}$$
$$+ m\log\left(\frac{n(n-1)}{2}\right)$$
$$= n(n-1)\log(\sigma\sqrt{2\pi}) + \frac{1}{\sigma^2}\sum_{i<j}(s_{ij}-\hat{s}_{ij})^2 + m\log\left(\frac{n(n-1)}{2}\right)$$
$$= n(n-1)\log(\sigma\sqrt{2\pi}) + \frac{E}{\sigma^2} + m\log\left(\frac{n(n-1)}{2}\right). \quad (4)$$

In effect, σ quantifies the inherent precision of a matrix of similarity data and provides an indication of the level of data-fit an additive clustering representation should seek to model the stated relationships between stimuli. If the given similarity values are known to be very accurate, for example, the inclusion of additional clusters may well be warranted to capture all of the detail provided. If, however, the constraining data are imprecise, then it may be appropriate to model only the major representational clusters using a relatively simple model. An important property of this conception is that precision is a property of a similarity matrix itself and is independent of its use within any representational framework. This means that σ should be derived from an understanding of the process by which the similarity values themselves were generated and not estimated as a parameter within the process of fitting a particular representational model.

One means of determining data precision, particularly applicable within the common experimental situation where the final similarity matrix is derived by averaging across measures provided by a number of subjects, is to calculate σ as the average of the standard deviations for each of the pooled cells in the matrix. In other words, σ is a sample estimate of the precision obtained by averaging across the standard deviation of each of the similarity values. Effectively, these values quantify the agreement across different subjects, or across the same subject on different occasions, in

their ratings of the stimulus pairs. Close agreement, which provides precise data, is heralded by small values of σ, whereas noisy or imprecise data are indicated by large values of σ.

Unfortunately, when the raw data needed to form estimates of precision in this or any other rigorous sort of way are not available, the choice of σ must be made on more subjective and heuristic grounds. In this case, results obtained on a collection of various similarity matrices (Lee, 1999a) might provide some broad guidance. For normalised similarity or proximity matrices, a σ value of about 0.10 seems to correspond to reasonably precise data, while values over 0.20 seem to correspond to particularly imprecise data. Despite the generality of these sorts of guidelines, the role of σ, in forcing an explicit and quantitative assumption to be made about data precision before fitting an additive clustering model, is an important one. It is possible for two similarity 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 are likely to demand additive clustering models with different levels of complexity. This allows precise data collected, say, from domain experts exhibiting close agreement in their judgments, to be fit by a detailed model with many parameters, while ensuring that less precise data are not over-fit by a similarly complex model.

An Illustrative Application

As an illustrative application of the BIC approach, consider four competing additive clustering models of the domain of $n = 16$ consonant phonemes. Each of these models used a similarity matrix given by Shepard (1972, Table 4.1), originally derived from Tables I through VI of the auditory confusion probabilities reported by Miller and Nicely (1995). The first two additive clustering models are provided by Tenenbaum (1996, Table 2) and Shepard and Arabie (1979, Table 2), accounting, respectively, for 90.2% of the variance using 9 clusters and 94.5% of the variance using 17 clusters. These patterns of data-fit and cluster cardinality clearly raise issues of model comparison, in that it is not obvious to what extent the additional clusters incorporated within Shepard and Arabie's (1979) model are sufficiently justified by the evident advantages in similarity structure accommodation.[1]

The third and fourth models of the phoneme domain, detailed in Table 1, were generated using the additive clustering technique developed by Lee (1999b) and raise the closely related issue of model generation. The 10 cluster model augments the set of middle voiceless fricatives $\{\theta, \mathrm{s}\}$ to the clusters of the 9 cluster model, thus increasing, after an adjustment of the cluster weights, the variance accounted for from 85.5 to 87.4%. Once again, in terms of a process of incremental model construction, it is not obvious whether or not this additional cluster should, on the basis of the improvement in data-fit, be included in the final model.

[1] Note that this comparison focuses upon the relative merits of the two specific models and not the general utility of the two techniques by which they were generated. Indeed, on the basis of the available evidence, it seems likely that a 17 cluster model derived using Tenenbaum's (1996) method would be able to account for more than 94.5% of the variance.

TABLE 1

The 10 and 9 Cluster Phoneme Models

Phonemes in cluster	Weight in 10 cluster model	Weight in 9 cluster model
{f, θ}	0.386	0.391
{d, g}	0.220	0.226
{p, k}	0.202	0.197
{p, t, k}	0.200	0.202
{v, ∂}	0.171	0.159
{b, v, ∂}	0.139	0.144
{m, n}	0.113	0.107
{d, g, z, ž}	0.084	0.082
Additive constant	0.028	0.029
{θ, s}	0.112	—

The phoneme similarity matrix **S** from which the 10 and 9 cluster models were derived has variance $\sum_{i<j}(s_{ij}-\bar{s})^2 = 0.694$, which allows E measures for each of the models to be evaluated using (2). These measures, together with the cluster cardinalities, were used to generate BIC indices for each of the four models across the interval $\sigma = [0.02, ..., 0.20]$. The results of this analysis are shown in Fig. 1 and provide a range of useful information relating to questions of model comparison and construction.

In terms of comparing the models of Tenenbaum (1996) and Shepard and Arabie (1979), the BIC measure suggests that the additional clusters of the latter model are justifiable if the target similarity data are believed to be exceptionally precise. Once

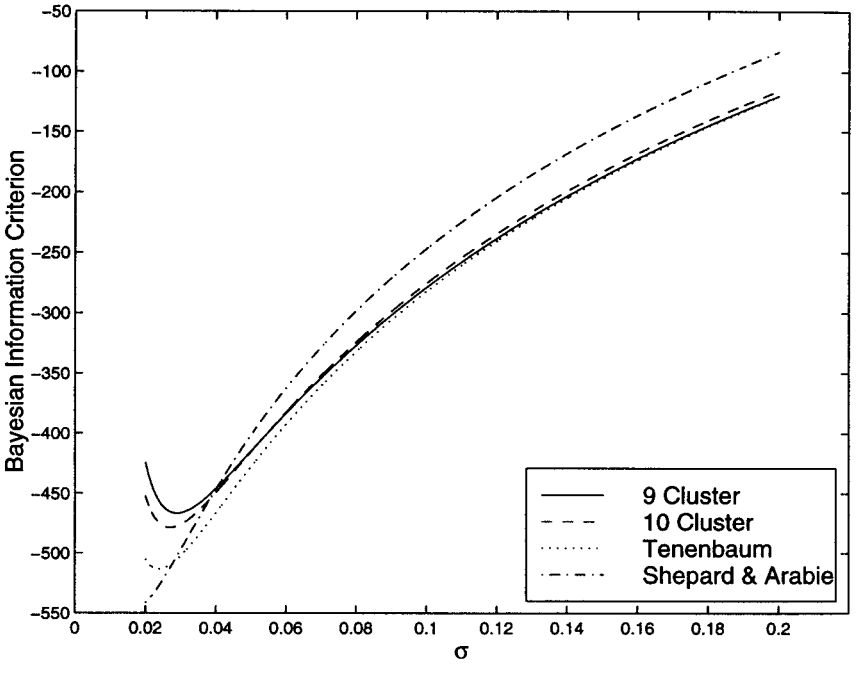

FIG. 1. Bayesian Information Criterion values across different levels of data precision for the four phoneme models.

(approximately) $\sigma > 0.03$, corresponding to an assumed variance in (3) of about 0.001, the pattern of change of BIC measures indicates that the data are no longer capable of supporting the additional data-fit provided by the more complicated model. A similar state of affairs is evident in relation to the 10 and 9 cluster models, where the inclusion of the extra cluster constitutes a modeling improvement only when (approximately) $\sigma < 0.04$. It is also worth observing that Tenenbaum's (1996) model is superior to both the 10 and 9 cluster models across the entire interval of σ values considered. This is to be expected, since the 10 and 9 the cluster models have the same or greater number of clusters, yet explain less of the similarity structure variance than Tenenbaum's (1996) model.

While the preceding form of analysis assists decisions to be made regarding which model is better at a given level of precision, it is not immediately obvious as to how the significance of magnitude differences in BIC measures should be gauged. Fortunately, as its name suggests, the BIC measure is grounded in Bayesian probabilistic reasoning (again, see Kass & Raftery, 1995; Myung & Pitt, 1997). As such, it is possible to interpret differences in BIC measures across models or levels of data precision in terms of a meaningful scale of probabilities, rather than resorting to some form of consensual subjective callibration. One way in which this objective evaluative comparison may be accomplished is through considering the ratio of the posterior odds of the cluster structures of model A, in relation to a second model, B, as revised according to Bayes' theorem

$$\frac{p(\mathbf{F}_A \mid \mathbf{S})}{p(\mathbf{F}_B \mid \mathbf{S})} = \frac{p(\mathbf{F}_A)}{p(\mathbf{F}_B)} \frac{p(\mathbf{S} \mid F_A)}{p(\mathbf{S} \mid F_B)},$$

where $p(\mathbf{F}_A)$ and $p(\mathbf{F}_B)$ are the prior probabilities of the two cluster structures, which are reasonably assumed to be equal in most practical applications.[2] In this case, the so called Bayes factor

$$B_{AB} = \frac{p(\mathbf{S} \mid F_A)}{p(\mathbf{S} \mid F_B)}$$

determines the posterior odds. However, as Kass and Raftery (1995, p. 778) demonstrate, the logarithm of the Bayes factor is approximated by the difference between the BIC measures of the two models, as follows

$$2 \log B_{AB} \approx \mathrm{BIC}_B - \mathrm{BIC}_A . \tag{5}$$

Through this relationship, it is possible to interpet differences in BIC measures. In particular, as Kass and Raftery (1995) observe, $2 \log B_{AB}$ exists on the same scale as likelihood ratio test statistics. In this context, Table 2 reproduces the standards of evidence suggested by Kass and Raftery (1995, p. 777), which serve as a useful

[2] A worthwhile topic for future research, however, is to determine the way in which computational complexity theory might be employed to quantify these prior probabilities (see Myung & Pitt, 1997, note 6).

TABLE 2

Kass and Raftery's (1995) Suggested Interpretative Scale for $2 \log B_{AB}$

$2 \log B_{AB}$	Evidence
0 to 2	Not worth more than a bare mention
2 to 6	Positive
6 to 10	Strong
>10	Very strong

initial framework within which to evaluate differences in the BIC measures of additive clustering models.

Figure 2 depicts $2 \log B_{AB}$ for the two pairs of phoneme models considered earlier, with the various demarcation points of the interpretative scale superimposed as horizontal lines. The comparison of Tenenbaum's (1996) model with that of Shepard and Arabie (1979) reveals that the difference in BIC values in Fig. 1 provides strong evidence in favor of the latter model at low σ values, but strong evidence in favor of the former when (approximately) $\sigma > 0.03$. As before, any decision as to which model is to be preferred hinges upon what is known or assumed regarding the precision of the similarity data. In contrast, the developmental comparison of the 9 and 10 cluster models shows strong evidence in favor of the inclusion of the additional cluster for low σ values, whereas the impetus for maintaining the 9 cluster model ranges only to "positive" as σ increases. As such, unless there are

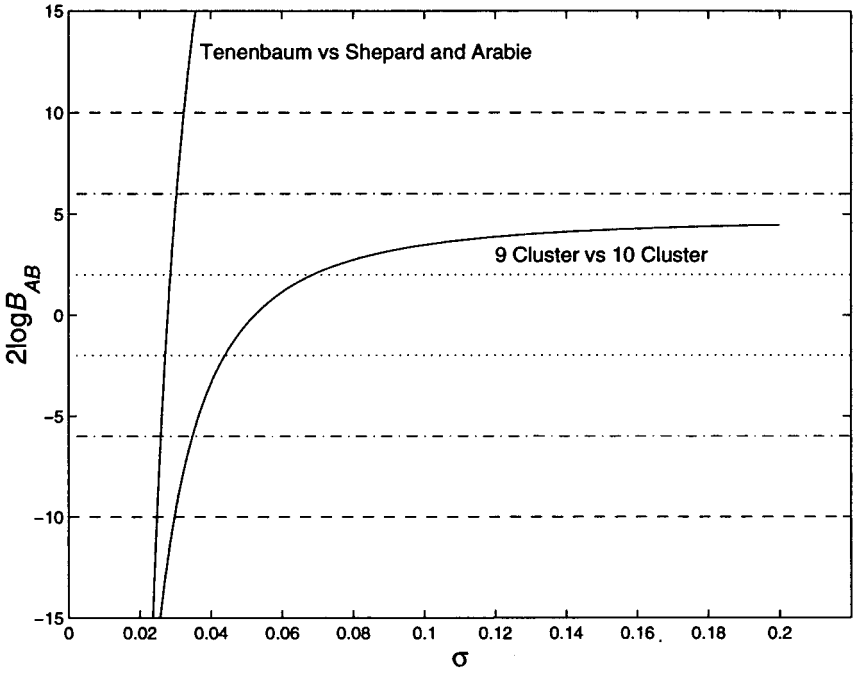

FIG. 2. Approximation to twice the logarithm of the Bayes factor comparing the phoneme models of Tenenbaum (1996) and Shepard and Arabie (1979), and the 9 and 10 cluster phoneme models.

strong grounds for doubting the precision of the similarity data, it seems reasonable to conclude that the inclusion of the additional cluster is justified.

The application of BIC measures to comparing and generating additive clustering models in the ways described above has the advantage of being both conceptually and computationally simple. In terms of practical operation, the primary domain-specific issue is one of determining, through whatever means are available, a reasonable prior assumption regarding the precision of the similarity data. Once such an assumption is made, the generation of BIC measures and Bayes factor approximations provides valuable quantitative insight regarding the trade-off between data-fit and complexity operating within additive clustering models.

There are, however, at least two deficiencies of the BIC measure approach that warrant further consideration. First, it is important to emphasize that (5) only approximates (twice the logarithm of) the Bayes factor and that the BIC measure in general makes simplifying assumptions regarding model complexity which may not always be justified (see Kass & Raftery 1995, p. 790). Because of these limitations, at the very least, the application of the BIC measure to target domains containing a relatively small number of stimuli requires caution. The second deficiency of the proposed BIC measure involves the implicit assumption that the complexity of an additive clustering model is equivalent to its cluster cardinality. Although this assumption pervades previous discussion of additive clustering model complexity, it seems clear that, in fact, model complexity is determined by the patterns of cluster encompassment, cluster overlap, and general cluster structure of a model and not simply the number of clusters employed. The second goal of this paper is to attempt to develop a quantitative understanding of the way in which these types of features of cluster structures relate to additive clustering model complexity.

BEYOND CARDINALITY AS A MEASURE OF COMPLEXITY

The progression from cluster cardinality as a measure of complexity to one which allows for general cluster structure relates to the distinction drawn by Myung and Pitt (1997, p. 81) between the number of parameters and the functional form components of model complexity. Traditionally, when **F** and **w** values are derived from a given similarity matrix, the clusters are treated as the model *per se*, while the associated weights are viewed as cluster parameters tuned to minimize the error measure.[3] It is in this sense that the number of clusters constitutes the number of parameters complexity component, as was assumed to develop the BIC measure in (4). The functional form component, in contrast, relates to the way in which the parameters interact within the model which, in the additive clustering context, is controlled by the patterns of encompassment and overlap within a cluster structure.

[3] Indeeed, additive clustering models are often (e.g., Arabie & Carroll, 1980; Hojo, 1982; Shepard & Arabie, 1979) conveyed visually by overlaying the derived clusters upon spatial characterizations of the domain, of the type generated by techniques such as multidimensional scaling. In such model presentations, the derived values of the weights are usually not included, although sometimes the weight ranked order of each cluster is indicated. In other words, it is the clusters contained in the matrix **F** which are typically interpreted as the model extracted from a given similarity structure, with the weights being regarded merely as a suitable parameterization of this model.

Application of Laplacian Approximation

One way in which a quantitative understanding of the relationship between cluster structure and model complexity may be developed is through the so-called Laplacian approximation of the marginal probability density (see Kass & Raftery, 1995, p. 777), given by

$$p(\mathbf{S}\,|\,\mathbf{F}) = \int_{[0,\,1]^m} p(\mathbf{S}\,|\,\mathbf{F},\,\mathbf{w})\; p(\mathbf{w}\,|\,\mathbf{F}).d\mathbf{w}$$

$$\approx \frac{(2\pi)^{m/2}\, p(\mathbf{S}\,|\,\mathbf{F},\,\mathbf{w}^*)\; p(\mathbf{w}^*\,|\,\mathbf{F})}{|-\nabla^2 \log(p(\mathbf{S}\,|\,\mathbf{F},\,\mathbf{w}^*)\; p(\mathbf{w}^*\,|\,\mathbf{F}))|^{1/2}}, \tag{6}$$

where $\mathbf{w}^*$ is the maximum likelihood estimate of the weights, $p(\mathbf{w}^*\,|\,\mathbf{F})$ is the prior probability distribution of the weights over the given cluster structure, and $|\cdot|$ denotes the matrix determinant function. Kass and Raftery (1995, p. 778) suggest that this approximation is problematic when $n(n-1)/2 < 5m$. Given the usual additive clustering practice, initially advanced by Shepard and Arabie (1979, p. 102), of seeking models with $m \leqslant n$, this restriction seems likely to be satisfied. In particular, taking the worst case when $m = n$, the lower bound is exceeded for domains containing only 11 or more stimuli. In addition, the explicit assumption of normality in (3) should be expected to assist the accuracy of the approximation.

Another potential difficulty with the Laplacian approximation, more particular to the additive clustering context, relates to the limited domain of integration. Additive clustering models require nonnegative weights, but the approximation to the definite integral in (6) is made over the entire space $\mathbf{R}^m$. Once again, however, this may be justified on the grounds of the limited numbers of clusters derived in additive clustering models. Since only clusters associated with significantly nonzero weights are typically included in a final model, the maximum likelihood weight vector $\mathbf{w}^*$ should lie well within the bounds of valid integration. This means, in turn, that the Laplacian approximation should tend not to accumulate significant volumes of integrand mass beyond the applicable domain of nonnegative weights.

Determining a universally appropriate prior distribution expressing the relationship between the clusters and their associated weights is more difficult, because it seems likely to be domain specific.[4] For example, in some domains of application, there may be grounds for believing that clusters encompassing larger numbers of stimuli have larger associated weightings. Alternatively, it may be the case that nonoverlapping isolated clusters typically have smaller weights. Previously, Tenenbaum (1996) has assumed, in generating artificial data, that the operation of a uniform distribution across a restricted interval is "grossly typical" (p. 6) of observed weight distributions. This does not seem unreasonable and has the attraction of being the maximum entropy distribution if plausible maximal and minimal values can be specified (Kagan, Linnik, & Rao 1973), although other choices might appear to be equally well motivated (see Kass & Wasserman, 1996, for a general overview).

[4] This is to be distinguished from the requirement of the Laplacian approximation that the prior be uniform in the immediate region of the best-fitting weights, which seems less problematic.

Perhaps the most reasonable way to proceed is by consolidating those components of the Laplacian approximation dependent upon the determination of a prior distribution with those components relating to data precision and cluster cardinality into one general function $Q(m, \sigma)$. Whatever prior assumption relating particular weight values to cluster structures is made, it is reasonable to expect $Q(m, \sigma)$ to decrease as m increases, since an arbitrary higher-dimensional weight vector will generally be less likely than a lower-dimensional one. With this function Q in place, the Laplacian approximation of (6) simplifies to

$$p(\mathbf{S}\,|\,\mathbf{F}) \approx \frac{p(\mathbf{S}\,|\,\mathbf{F}, \mathbf{w}^*)}{|-\nabla^2 \log p(\mathbf{S}\,|\,\mathbf{F}, \mathbf{w}^*)|^{1/2}} \times Q(m, \sigma). \tag{7}$$

The denominator of (7) is a Hessian matrix, given by

$$-\nabla^2 \log p(\mathbf{S}\,|\,\mathbf{F}, \mathbf{w}^*) = \nabla^2 \left(\frac{1}{2\sigma^2} \sum_{i<j} \left(s_{ij} - \sum_k w_k f_{ik} f_{jk} \right)^2 \right),$$

which may be found by noting that

$$\begin{aligned} \frac{\partial^2}{\partial w_x \partial w_y} \left[\frac{1}{2\sigma^2} \sum_{i<j} \left(s_{ij} - \sum_k w_k f_{ik} f_{jk} \right)^2 \right] &= \frac{\partial}{\partial w_y} \left[-\frac{1}{\sigma^2} \sum_{i<j} \left(s_{ij} - \sum_k w_k f_{ik} f_{jk} \right) f_{ix} f_{jx} \right] \\ &= \frac{1}{\sigma^2} \sum_{i<j} f_{ix} f_{jx} f_{iy} f_{jy}, \end{aligned}$$

implying that the complexity depends on the generic cluster structure, independent of specific values of the best-fitting weights and takes the form

$$\mathbf{H} = \frac{1}{\sigma^2} \mathbf{G},$$

where

$$\mathbf{G} = \begin{bmatrix} \sum_{i<j} f_{i1} f_{j1} & \sum_{i<j} f_{i1} f_{j1} f_{i2} f_{j2} & \cdots & \sum_{i<j} f_{i1} f_{j1} f_{im} f_{jm} \\ \sum_{i<j} f_{i2} f_{j2} f_{i1} f_{j1} & \sum_{i<j} f_{i2} f_{j2} & \cdots & \sum_{i<j} f_{i2} f_{j2} f_{im} f_{jm} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i<j} f_{im} f_{jm} f_{i1} f_{j1} & \sum_{i<j} f_{im} f_{jm} f_{i2} f_{j2} & \cdots & \sum_{i<j} f_{im} f_{jm} \end{bmatrix}.$$

The $1/\sigma^2$ may be incorporated into the general function $Q(m, \sigma)$ in (7). This means that the Laplacian approximation can finally be written as

$$\begin{aligned} p(\mathbf{S}\,|\,\mathbf{F}) &\approx \frac{p(\mathbf{S}\,|\,\mathbf{F}, \mathbf{w}^*)}{|\mathbf{G}|^{1/2}} \times Q(m, \sigma) \\ &\propto \frac{p(\mathbf{S}\,|\,\mathbf{F}, \mathbf{w}^*)}{|\mathbf{G}|^{1/2}} \qquad \text{if} \quad m, \sigma \text{ are fixed.} \end{aligned} \tag{8}$$

The first evaluative measure of an additive clustering model given in (8) is readily interpretable in terms of its three components. The maximum likelihood $p(\mathbf{S}|\mathbf{F}, \mathbf{w}^*)$ is the measure of data-fit, which needs to be maximized, the value $|\mathbf{G}|^{1/2}$ is a measure of cluster structure complexity, which needs to be minimized, and the function Q quantifies the complexity effect of cluster cardinality, and the general effect of changes in data precision, in the ways described earlier. When considering a model with a fixed number of clusters, relating to data of a given precision, the value of $Q(m, \sigma)$ is a constant. This leads to the second evaluative measure given in (8), which incorporates only the data-fit and cluster complexity components. Clearly, it is through the analysis of the complexity matrix $\mathbf{G}$ that an understanding of the complexity effects of various cluster structures in additive clustering models can be developed.

Interpretation and Nature of the Complexity Matrix

The interpretation of the elements of $\mathbf{G}$ in terms of the cluster variables f_{ik} is relatively straightforward. The kth diagonal element, $\sum_{i<j} f_{ik} f_{jk}$, constitutes a count of the number of pairs of domain stimuli lying in the kth cluster, whereas each off-diagonal element, of the form $\sum_{i<j} f_{ix} f_{jx} f_{iy} f_{jy}$, counts the number of pairs of stimuli lying in both the xth and yth clusters. Accordingly, the diagonal elements give an indication of cluster size within a model, while the off-diagonal elements relate to the patterns of cluster overlap. This interpretation makes it clear that $\mathbf{G}$ is constrained by the relationship $g_{ij} \leqslant \min(g_{ii}, g_{jj})$, since the number of stimulus pairs in the overlap of two clusters cannot be greater than the number of pairs in the whole of either cluster.

Importantly, it will generally be the case that $\mathbf{G}$ is positive definite, which ensures that the complexity measure $|\mathbf{G}|^{1/2}$ is strictly positive, as is required for meaningful interpretation. To see this, note that $\mathbf{G}$ may be rewritten as $\mathbf{C}^{\mathbf{T}}\mathbf{C}$, where

$$\mathbf{C} = \begin{bmatrix} f_{11}f_{21} & f_{11}f_{31} & \cdots & f_{21}f_{31} & \cdots & f_{(n-1)1}f_{n1} \\ f_{12}f_{22} & f_{12}f_{32} & \cdots & f_{22}f_{32} & \cdots & f_{(n-1)2}f_{n2} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ f_{1m}f_{2m} & f_{1m}f_{3m} & \cdots & f_{2m}f_{3m} & \cdots & f_{(n-1)m}f_{nm} \end{bmatrix},$$

and therefore is positive definite when $\mathbf{C}^{\mathbf{T}}$ has full rank. Since the best-fitting weights of an additive clustering models satisfy the relationship $\mathbf{C}^{\mathbf{T}}\mathbf{w} = \mathbf{s}$, where $\mathbf{s} = (s_{12}, s_{13}, ..., s_{23}, ..., s_{(n-1)n})^{\mathbf{T}}$, the full rank of $\mathbf{C}^{\mathbf{T}}$ allows the matrix inversion which, together with the non-negativity constraint on $\mathbf{w}$, determines the values of the weights. When $\mathbf{C}^{\mathbf{T}}$ does not have full rank, however, it is not clear whether potential solutions for the weights will be sufficiently constrained to allow the derivation of a meaningful model. In any case, in general, it seems very unlikely that well-constructed additive clustering models, given their emphasis on cluster parsimony, will exhibit the degeneracies associated with $\mathbf{C}^{\mathbf{T}}$ not having full rank.

It is also worth noting that if, somehow, a method of model construction does create a degenerate cluster structure, it may often be able to be remedied using the following observation. When $\mathbf{C}^{\mathbf{T}}$ does not have full column rank, each of its

columns can be written as a linear combination of the others. Providing one of these columns, say the xth, can be rewritten as $f_{ix} f_{jx} = \sum_{k \neq x} \alpha_k f_{ik} f_{jk}$ $\forall i < j$, with $(w_x \alpha_k + w_k) \geqslant 0$ $\forall k$; this cluster can be removed from the cluster structure, without altering the estimated similarity values, since

$$\begin{aligned} \hat{s}_{ij} &= \sum_k w_k f_{ik} f_{jk} \\ &= w_x f_{ix} f_{jx} + \sum_{k \neq x} w_k f_{ik} f_{jk} \\ &= w_x \sum_{k \neq x} \alpha_k f_{ik} f_{jk} + \sum_{k \neq x} w_k f_{ik} f_{jk} \\ &= \sum_{k \neq x} (w_x \alpha_k + w_k) f_{ik} f_{jk} . \end{aligned} \tag{9}$$

This constructive approach to remediation would apply, for example, in the case where four stimuli are partitioned by two clusters into two pairs of stimuli, and a third cluster which encompassed all four stimuli is also included. The relationship expressed in (9) suggests that this least-weighted partitional cluster should be removed, with its weight added to the outer cluster, and the weight of the remaining partitioning cluster reduced by the same amount.

Implications for Cluster Structure Complexity

Given a cluster structure for which $\mathbf{G}$ is positive definite, the application of Hadamard's inequality (see, for example, Bellman, 1970, pp. 129–130) shows that

$$|\mathbf{G}| \leqslant \prod_k \sum_{i<j} f_{ik} f_{jk} , \tag{10}$$

which implies, for a fixed number of clusters, that a partitioning cluster structure is the most complicated. This is because equality is achieved in (10) iff $\mathbf{G}$ is a diagonal matrix, which requires that there be no overlap between clusters. Intuitively, the complexity of partitional models arises from the fact that their weight parameterization is constrained by a relatively limited subset of the available similarity measures. Each cluster involves only the pairwise relations between stimuli within that cluster, to the exclusion of the large number of pairwise relationships between stimuli in different clusters. This means that the best-fitting weights are determined in relation to only part of the data available in the similarity matrix. Consequently, for weight parameterizations which differ from the maximum likelihood estimates, the data-fit of partitional models will generally be poor. Cluster structures which allow some degree of overlap, in contrast, have the potential to consider all of the available similarity data. Models with overlapping structures will generally, therefore, have best-fitting

weights which represent a compromise between various competing fine-tunings and exhibit data-fit which is more robust across a range of weight parameterizations.

Of the different types of partitioning cluster structures, it follows from (10) that complexity is minimized when all but one of the clusters encompasses only one pair of stimuli, and the remaining cluster covers the rest. To see this, note that for any two clusters in a partition, containing $\alpha \geqslant \beta \geqslant 2$ stimuli respectively, the product of the complexity matrix determinant associated with these clusters, $\alpha(\alpha-1)/2 \times \beta(\beta-1)/2$, is reduced by transferring one stimulus from the smaller cluster to the larger one. That is, $\alpha(\alpha+1) \times (\beta-1)(\beta-2) < \alpha(\alpha-1) \times \beta(\beta-1)$. The minimal complexity matrix determinant, therefore, is achieved by transferring stimuli from smaller clusters to larger ones, until one cluster is as large as possible. Once again, this may be understood in terms of the extent to which the similarity measures constrain the weight parameterization of a cluster structure. Since the number of stimulus pairs considered by a cluster increases quadratically with the number of stimuli encompassed, the proportion of a similarity matrix that constrains weight estimation in a partitioning model is maximized by including the largest cluster possible. Adopting a minimum description length perspective on model complexity (Rissanen, 1989; Zemel, 1995) provides another means of understanding this result. As noted by Li and Vitányi (1993, p. 71), the so-called noiseless coding theorem (Shannon, 1948) indicates that the minimal average message length needed to convey a structure is approximately given by the entropy of that structure. Therefore, a partition in which each cluster encompasses the same number of stimuli is more complicated because each of the clusters becomes equally likely, maximizing the entropy of the cluster structure and, in turn, maximizing the message description length necessary to communicate the structure.

Typically, however, additive clustering models avoid partitioning cluster structures through the introduction of a universal cluster. This cluster encompasses all stimuli and corresponds to the inclusion of an additive constant in the basic similarity model (1). As noted by Arabie and Carroll (1980, p. 212), the incorporation of the universal cluster is necessary if the data-fit of a model is to be assessed by the variance explained measure (2). The impact of including this cluster on model complexity may be assessed as a special case of the following general result. If a cluster structure with complexity matrix $\mathbf{G}$ has an additional cluster encompassing z pairs of stimuli with overlaps defined by the vector $\mathbf{y}$ appended, the augmented matrix

$$\mathbf{G}^{+} = \begin{bmatrix} \mathbf{G} & \mathbf{y} \\ \mathbf{y}^{\mathbf{T}} & z \end{bmatrix}$$

bears the relationship

$$|\mathbf{G}^{+}| = |\mathbf{G}|\,(z - \mathbf{y}^{\mathbf{T}}\mathbf{G}^{-\mathbf{1}}\mathbf{y}) \leqslant z\,|\mathbf{G}|,$$

with equality iff $\mathbf{y} = \mathbf{0}$ (Magnus & Neudecker, 1988, p. 23).

In the special case of a partitioning cluster structure, with clusters encompassing $a, b, \ldots, x$ pairs of stimuli, the addition of the universal cluster increases the complexity measure $ab \cdots x$ by a factor of

$$\frac{n(n-1)}{2} - [a \;\; b \;\; \cdots \;\; x]^T \begin{bmatrix} \frac{1}{a} & 0 & \cdots & 0 \\ 0 & \frac{1}{b} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \frac{1}{x} \end{bmatrix} [a \;\; b \;\; \cdots \;\; x]$$

$$= \frac{n(n-1)}{2} - (a + b + \cdots + x),$$

so that the increase is minimized when the sum $a + b + \cdots + x$ is as large as possible. Encouragingly, this occurs when the original partitioning cluster structure assumes exactly the same minimal complexity structure as before, with one cluster encompassing all available stimuli, under the restriction that each of the other clusters must contain at least two stimuli. To see this, note that for any two clusters in a partition, containing $\alpha \geqslant \beta \geqslant 2$ stimuli respectively, the part of the sum associated with these clusters, $\alpha(\alpha-1)/2 + \beta(\beta-1)/2$, is increased by transferring one stimulus from the smaller cluster to the larger one. That is, $\alpha(\alpha+1) + (\beta-1)(\beta-2) > \alpha(\alpha-1) + \beta(\beta-1)$. The maximal sum, therefore, is achieved by continuing to transfer stimuli from smaller clusters to larger ones until one cluster is as large as possible. When coupled with the earlier result, this means that the complexity of a partitioning model that includes the universal cluster is also minimized by having one large cluster encompassing most of the stimuli.

To examine ways in which the complexity of more general overlapping cluster structures may be minimized, consider first a two cluster model in which the first cluster encompasses a pairs of stimuli and the second encompasses c pairs. Without loss of generality it may be assumed that $a \geqslant c$, so the cluster structure complexity of this matrix is given by

$$\mathbf{G} = \begin{bmatrix} a & b \\ b & c \end{bmatrix},$$

where $b \leqslant c$. The cluster complexity of this model is minimized when $|\mathbf{G}| = ac - b^2$ is minimized, which requires that $b = c$. If both of the clusters encompass the same number of stimuli, this choice is degenerate, since it amounts to including exactly the same cluster in a model twice. Otherwise, however, the choice of overlap b which minimizes cluster structure complexity is a strictly nested one in which the smaller second cluster encompasses a subset of the larger first cluster.

An intuitive analysis of cluster structures containing arbitrary numbers of clusters suggests that this result may be generalized. Suppose the first n clusters of the model, encompassing $a \geqslant b \geqslant c \geqslant \cdots \geqslant x$ stimulus pairs, are arranged in a strictly nested fashion, and a decision is to be made regarding the pattern of overlap of the next

smallest cluster, which contains $z \leqslant x$ stimuli. The complexity matrix then takes the form

$$\mathbf{G} = \begin{bmatrix} a & b & c & \cdots & x & y_1 \\ b & b & c & \cdots & x & y_2 \\ c & c & c & \cdots & x & y_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ x & x & x & \cdots & x & y_n \\ y_1 & y_2 & y_3 & \cdots & y_n & z \end{bmatrix},$$

and it is required that $y_1 \geqslant y_2 \geqslant y_3 \geqslant \cdots \geqslant z$ be chosen to minimize $|\mathbf{G}|^{1/2}$. This is equivalent to minimizing $|\mathbf{G}|$ which, geometrically, corresponds to minimizing the volume of the $(n+1)$-dimensional parallepiped with edges defined by the rows (or columns) of $\mathbf{G}$ (Horn & Johnson, 1985, p. 477). The two ways of accomplishing this are through minimizing the lengths of the edges, which suggests choosing y_1, y_2, y_3, ..., $y_n = 0$, or through decreasing the interior angle between the edges, which suggests choosing y_1, y_2, y_3, ..., $y_n = z$. However, a little thought, and some concrete numerical investigation, suggests that the second of these approaches is to be preferred. It is easy to see that minimizing the length of the edge associated with the new cluster forces it to be orthogonal with all of the established edges of the parallepiped. Minimizing the interior angle of the new edge with all others, in contrast, comes only at the cost of lengthening the one new edge, which, in any case, is constrained to be shorter than all of the other edges. Accordingly, the volume minimization benefits of edge reduction seem to be outweighed by the benefits of interior angle reduction. This implies that the new cluster should be assigned patterns of overlap which maintain the strictly nested cluster structure. When formalized, this intuitive argument provides the inductive step which, together with the case for $n = 2$ given earlier, demonstrates that the complexity of additive clustering models is minimized by a strictly nested cluster structure. Clearly, a formal proof of this claim should be a priority for future research.

In the meantime, it may be observed that for a strictly nested cluster structure, the following elementary row operation

$$\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ -1 & 1 & 0 & \cdots & 0 \\ -1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -1 & 1 & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} a & b & c & \cdots & x \\ b & b & c & \cdots & x \\ c & c & c & \cdots & x \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x & x & x & \cdots & x \end{bmatrix} = \begin{bmatrix} a & b & c & \cdots & x \\ b-a & 0 & 0 & \cdots & 0 \\ c-a & c-b & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x-a & x-b & x-c & \cdots & 0 \end{bmatrix}$$

reveals that $|\mathbf{G}| = (b-a)(c-b)\cdots x$. Since a strictly nested model is restricted to having $a > b > c > \cdots > x$, this means that the complexity of such a structure is minimized by having each of the successively decreasing clusters encompass almost as many stimulus pairs as its predecessor.

CONCLUSION

Because of its flexibility and ease of interpretation, additive clustering seems well suited to modeling the similarity relationships observed to exist between a set of stimuli. A crucial determinant of the success of such modeling, however, is the degree to which the complexity of derived models may be controlled. The BIC measure developed and demonstrated in this paper provides a quantitive measure, incorporating a cluster cardinality conceptualization of additive clustering model complexity, from which decisions in model comparison and construction may be made. This measure is readily calculated, allows for the introduction of prior knowledge or assumptions regarding the precision of the similarity data, and is amenable to interpretation on a meaningful probabilistic scale through its relationship to Bayes factors. The BIC measure, however, does rely on the simplifying assumption that model complexity is equivalent to cluster cardinality, which may not be appropriate in all practical situations.

Accordingly, a second measure of additive clustering model complexity, which considers the complexity effects of different cluster structures, was developed using the Laplacian approximation of a marginal probability density. A preliminary investigation of this complexity measure revealed that when a model has the same number of clusters and the exhibits the same level of data-fit, those with partitioning cluster structures are the most complicated, while those with strictly nested clusters are the least complicated. There remains, however, considerable scope for further analysis of this source of additive clustering model complexity, providing a fertile ground for further theoretical research with potential practical application.

ACKNOWLEDGMENTS

I thank Geoff Latham, Kenneth Pope, Simon Williams, and Chris Woodruff for their assistance and In Jae Myung and an anonymous reviewer for their helpful comments on earlier versions of this article.

REFERENCES

Arabie, P., & Carroll, J. D. (1980). MAPCLUS: A mathematical programming approach to fitting the ADCLUS model. *Psychometrika*, **45**, 211–235.

Arabie, P., & Shepard, R. N. (1973). *Representation of similarities as combinations of discrete, overlapping properties*. Paper presented at Mathematical Psychological Meeting, Montréal.

Bellman, R. (1970). *Introduction to matrix analysis* (second ed.). New York: McGraw–Hill.

Chaturvedi, A., & Carroll, J. D. (1994), An alternating combinatorial optimization approach to fitting the INDCLUS and generalized INDCLUS models. *Journal of Classification*, **11**, 155–170.

Hojo, H. (1982). A maximum likelihood method for additive clustering and its applications. *Japanese Psychological Research*, **25**(4), 191–201.

Horn, J. A., & Johnson, C. R. (1985). *Matrix analysis*. Cambridge, MA: Cambridge University Press.

Kagan, A. M., Linnik, Y. V., & Rao, C. R. (1973). *Characterization problems in mathematical statistics*. New York: Wiley.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**(430), 773–795.

Kass, R. E., & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, **91**(435), 1343–1370.

Lee, M. D. (1999a). *Algorithms for representing similarity data.* Technical Report DSTO–RR–0152, Defence Science and Technology Organisation.

Lee, M. D. (1999b). An extraction and regularization approach to additive clustering. *Journal of Classification*, **16**, 255–281.

Li, M., & Vitányi, P. (1993). *An introduction to Kolmogorov complexity and its applications*. Reading, MA: Addison–Wesley.

Magnus, J. R., & Neudecker, H. (1988). *Matrix differential calculus with applications in statistics and econometrics.* New York: Wiley.

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America*, **27**, 338–352.

Mirkin, B. G. (1987). Additive clustering and qualitative factor analysis methods for similarity matrices. *Journal of Classification*, **4**, 7–31.

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

Rissanen, J. (1989). *Stochastic complexity in statistical inquiry*. Singapore: World Scientific.

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

Shannon, C. E. (1948). The mathematical theory of communication. *Bell Systems Technical Journal*, **27**, 379–243, 623–656.

Shepard, R. N. (1972). Psychological representation of speech sounds. In E. E. David & P. B. Denes (Eds.), *Human communication*: *A unified view* (pp. 67–113). New York: McGraw–Hill.

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

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

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* (Vol. 8, pp. 3–9). Cambridge, MA: MIT Press.

Zemel, R. S. (1995). Minimum description length analysis. In M. A. Arbib (Ed.), *Handbook of brain theory and neural networks* (pp. 572–575). Cambridge, MA: MIT Press.

Received: January 27, 1998; revised August 4, 1998