Machine Learning, 49, 39–58, 2002

A Simple Method for Generating Additive Clustering Models with Limited Complexity

MICHAEL D. LEE michael.lee@psychology.adelaide.edu.au
(http://complex.psych.adelaide.edu.au/members/staff/michaellee/homepage/)
Department of Psychology, University of Adelaide, SA 5005, Australia

Editor: Padhraic Smyth

Abstract. Additive clustering was originally developed within cognitive psychology to enable the development of featural models of human mental representation. The representational flexibility of additive clustering, however, suggests its more general application to modeling complicated relationships between objects in non-psychological domains of interest. This paper describes, demonstrates, and evaluates a simple method for learning additive clustering models, based on the combinatorial optimization approach known as Population-Based Incremental Learning. The performance of this new method is shown to be comparable with previously developed methods over a set of 'benchmark' data sets. In addition, the method developed here has the potential, by using a Bayesian analysis of model complexity that relies on an estimate of data precision, to determine the appropriate number of clusters to include in a model.

Keywords: additive clustering, population-based incremental learning, PBIL, Bayesian information criterion, BIC, cognitive modeling

Introduction

Additive clustering

Additive clustering models, developed by Arabie and Shepard (1973; see also Shepard, 1974; Shepard & Arabie, 1979) in the context of cognitive modeling, provide simple yet powerful accounts of the observed similarities between sets of stimuli. Central to additive clustering is the fundamental premise it shares with all clustering models: that groups of objects in a domain may, in some context, be treated as equivalent. As Shepard and Arabie (1979, p. 91) argue, however, "generally, the discrete psychological properties of objects overlap in arbitrary ways". Accordingly, additive clustering does not enforce either of the constraints that limit partitioning and hierarchical clustering approaches. Unlike partitioning clustering approaches, additive clustering allows each stimulus to belong to any number of clusters and, unlike hierarchical clustering approaches, additive clustering places no 'nesting' constraints upon the sets of stimuli that may be encompassed by successive clusters. Rather, additive clustering characterizes stimuli in terms of a series of discrete, potentially overlapping properties.

Formally, an additive clustering representation involving m clusters and n stimuli is defined by an $n \times m$ matrix of binary membership variables $\mathbf{F} = [f_{ik}]$,

where:

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

Using this representation, the observed similarity between each pair of stimuli $\mathbf{S} = [s_{ij}]$ is considered to have arisen from the clusters to which both belong. In particular, if 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 the sum of the weights of the common clusters, as follows:

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

Under this simple similarity model, the representational flexibility of additive clustering has allowed the development of convincing accounts, in several stimulus domains, of the complicated patterns of interaction between different sources of stimulus similarity evident in human judgments (Hojo, 1982; Lee, 1999b; Shepard & Arabie, 1979; Tenenbaum, 1996). For example, using data measuring the 'abstract conceptual similarity' of the numbers 0–9, Tenenbaum (1996) developed an 8 cluster model, explaining in excess of 90% of the variance in the data, that incorporated clusters relating to both the numerical magnitude (e.g., $\{1, 2, 3, 4\}$ and $\{6, 7, 8, 9\}$) of the numbers, and various arithmetic concepts (e.g., $\{2, 4, 8\}$ and $\{3, 6, 9\}$). As Tenenbaum (1996) argues, to capture the human judgments of similarity in this way "an overlapping clustering model is necessary ... to accommodate the multiple causes of similarity".

Despite reported successes predominantly relating to psychological measures of similarity, there is no fundamental reason for restricting additive clustering analyses to cognitive modeling. In many non-psychological modeling tasks, the target domain of interest is naturally characterized in terms of a set of objects, related to each other through measures of pairwise similarity (Lee, 1999a). A host of measures, such as ratings of similarity, indices of correlation or proximity, counts or probabilities of co-occurrence, confusion, or substitution, or any other measures which might broadly be termed 'associative', are all amenable to interpretation as similarities. In addition, sets of objects that are best characterized in terms of quantitative lists of properties may be subjected to any distance metric or various other comparative techniques to provide measures of the similarity between each pair of objects (Cox & Cox, 1994).

Clearly, the notion of representing domains of interest in terms of their constituent objects, and the strength of the relationships between them, is a very general one. Given the success of additive clustering models in developing meaningful psychological representations from this sort of information, it seems likely that additive clustering techniques may be amenable to more general application.

Limiting additive clustering complexity

Previous techniques

Typically, the cluster membership variables $\mathbf{F}$ and weights $\mathbf{w} = (w_1, \ldots, w_m)$ extracted from a given similarity structure $\mathbf{S}$ 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. These include mathematical programming (Arabie & Carroll, 1980; Chaturvedi & Carroll, 1994), qualitative factor analytic (Mirkin, 1987), extraction and regularization (Lee, 1999b; Shepard & Arabie, 1979), 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.

As noted by Shepard and Arabie (1979, p. 98), however, 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:

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

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 an ability to generalize accurately beyond given information.

Previously developed techniques for learning additive clustering models have most often addressed this potential difficulty by limiting the number of clusters extracted to some pre-determined number. A rationale for making this choice is rarely articulated, although Shepard and Arabie (1979, p. 102) suggest that for an n stimulus domain it is appropriate to seek a model accounting "for at least 80% of the variance ... with no more than about n weights in most cases". In general, however, the development of additive clustering models does not appear to have been explicitly or systematically constrained by an analysis which balances the need to improve a model's fit to the data, with the competing need to limit model complexity.

Bayesian complexity measure

Lee (2001b) presented a simple means of quantitatively addressing this balance using the Bayesian Information Criterion (BIC), an established and well understood measure that incorporates both data-fit and model complexity (Schwarz, 1978; see also Kass & Raftery,

1995; Myung & Pitt, 1997). The BIC takes the general form:

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

where p (ML) is the maximum probability density of the model, P is the number of parameters in the model, and N is the sample size. 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 $\mathbf{w}$. An appropriate formulation of this probability is provided by Tenenbaum (1996), in which it is assumed that $p\,(\mathbf{S} \mid \mathbf{F}, \mathbf{w})$ has a Gaussian distribution with common variance σ^2, as follows:

$$p\,(\mathbf{S} \mid \mathbf{F}, \mathbf{w}) \propto \exp\left(-\frac{1}{2\sigma^2} \sum_{i>j} (s_{ij} - \hat{s}_{ij})^2\right) = \exp\left(-\frac{E}{2\sigma^2}\right). \tag{2}$$

By equating the number of parameters in an additive clustering model with the number of cluster weights[1], ignoring an additive constant arising from the coefficient of proportionality, and observing that a similarity matrix for n objects incorporates $n(n-1)/2$ samples, an additive clustering formulation of the BIC measure is given by:

$$\text{BIC} = \frac{E}{s^2} + m \log\left(\frac{n(n-1)}{2}\right), \tag{3}$$

where s is a sample estimate[2] of the data precision population parameter σ.

It should be noted that the BIC has limitations as a measure of model 'goodness', since it is derived as an asymptotic approximation to posterior probability of a model being true, and relies on a number of simplifying assumptions (Schwarz, 1978; see also Kass & Raftery, 1995). There are good grounds for preferring the BIC over the AIC (Akaike, 1974), because the BIC places relatively greater importance on likelihoods than priors (Kass & Raftery, 1995, p. 790). This is appropriate, since additive cluster modeling usually aims to reveal the latent structure in similarity data, which requires that model selection primarily be constrained by the data at hand. More serious challenges to the BIC arise from the fact that it counts only the number of parameters as a measure of model complexity, and is insensitive to the 'functional form' component of model complexity caused by differences in parametric interaction (Myung & Pitt, 1997). There are a number of alternative measures that do account for parametric interaction (see Kass & Raftery, 1995), including new results based on differential geometry (Myung, Balasubramanian, & Pitt, 2000), and at least one of these measures has been applied specifically to additive clustering models (Lee, 2001b). While many of these alternatives are theoretically preferable, they are also considerably more difficult to incorporate into an algorithm for model generation, and require significantly greater computational resources. In this sense, the BIC could be argued to represent a practical balance between the accuracy of results and simplicity of implementation. Perhaps the most significant practical limitation of the BIC is that it is a conservative measure (Raftery, 1999), in the sense that it has a tendency to favor additive clustering models with too few clusters. It could also be argued, however, this conservatism is the lesser of two

evils. As Grünwald (2000, p. 148) concludes: "If you overfit, you think you know more than you really know. If you underfit, you do not know much, but you know that you do not know much. In this sense, underfitting is relatively harmless, while overfitting dangerous".

Measuring data precision

The intended role of s is one of quantifying the *inherent* precision of the data, independent of its subsequent application to any particular type of cognitive or other representational analysis. In terms of the empirical collection of similarity data, the established and prevailing practice (e.g., Ekman, 1954; Gati & Tversky, 1982; Gregson, 1976; Johnson & Tversky, 1984; Kruschke, 1993) is to form similarity matrices by averaging the individual ratings or confusion probabilities of a large number of subjects. More formally, given a set of individual similarity matrices $\mathbf{S}^k = [s_{ij}^k]$ derived from the data collected from each of $k = 1, 2, \ldots, K$ subjects, it is the averaged similarity matrix $\mathbf{S} = \frac{1}{K}[\sum_k s_{ij}^k] = [s_{ij}]$ that is used. In this case, the natural approach to determining s is to calculate the average of the standard errors for each of the pooled cells in the final matrix, as follows:

$$s = \frac{1}{n\,(n-1)\,/2} \sum_{i<j} \sqrt{\frac{\sum_k \left(s_{ij}^k - s_{ij}\right)^2}{K-1}}. \tag{4}$$

This estimate of data precision is entirely determined by the raw data, and may be calculated before fitting any additive clustering model. The evaluation of BIC measures for candidate models is then straightforward, and the model with the minimal BIC value may be taken as constituting an appropriate compromise between the need to accommodate the original data, and the requirement to minimize the parametric complexity of the representational model.

Unfortunately, it is sometimes the case that the raw ratings data needed to calculate s is not available. In this case, it is necessary to rely on experience with other averaged similarity matrices collected in much the same way, as guidance regarding appropriate assumptions for s. For example, Lee (2001a) examined ratings of the similarity of rectangles with interior line segments used by Kruschke (1993), and found an s value of 0.125. Meanwhile, unpublished analyses of similarity ratings for simple geometric shapes (Lee, 1998), embedded textures (Woodruff, 1998), and various other stimulus sets (Mark Steyvers, personal communication) revealed s values between approximately 0.05 and 0.15. Heuristically, it seems that noise standard deviations under a Gaussian distribution of 5%, 10% and 15% correspond to what might loosely be termed 'precise', 'average' and 'imprecise' data sets. Thus, s values of 0.05, 0.10, and 0.15 may reasonably be employed with normalized similarity matrices, when the raw data required to generate a first-principles estimate of data precision is not available.

It is worth emphasizing, however, the importance 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 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 may demand additive clustering 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 clusters, while ensuring that less precise data are not over-fit by a similarly complicated representation. Using this argument, it is clear that previous additive clustering algorithms that have pre-determined the number of clusters to be derived have, effectively, been making implicit assumptions regarding the level of data precision. Even when the raw data needed to estimate s in a principled way is not available, it seems appropriate to make this assumption explicitly, by specifying s heuristically, rather than implicitly through constraining the number of clusters.

A new additive clustering method

A two-stage approach is adopted by the new method for generating additive clustering representations developed here. In the first stage, for an n object domain, an n cluster model is learned, and a record is made of the κ clusters with the greatest associated weights considered during this process. In the second stage, an assumption is made regarding the precision of the similarity data, and a model is built from these 'candidate' clusters on the basis of minimizing the BIC measure. Although it would be equally possible to apply the BIC in the case of an algorithm that simply generated candidate additive clustering representations, there are two reasons for using a two-stage approach. The primary motivation is that, as with extraction and regularization approaches Lee (1999b), domain specific knowledge, in the form of collateral information or substantive hypothesis, can potentially be used to augment or modify the set of candidate clusters. Secondly, in practice, the separation of the two stages is also useful when the data precision estimate s must be specified heuristically, since it is relatively computationally efficient to generate models corresponding to a range of different s values, using the same set of candidate clusters.

Algorithmically, both stages are tackled by adapting the 'black-box' combinational optimization technique known as Population-Based Incremental Learning (PBIL) (Baluja, 1994). Accordingly, before describing the operation of the two stages, it seems worth giving a generic overview of the PBIL optimization method.

Population-based incremental learning

The basic approach of PBIL is to encode potential solutions to an optimization problem in terms of a bit string, and maintain and update an explicit measure of the (unconditional) probabilities describing the state of each of these bits in good solutions. For a problem that requires α bits to specify a potential solution, this explicit measure takes the form of a probability vector $\mathbf{p} = (p_1, p_2, \ldots, p_\alpha)$.

On each iteration, a set of β potential solutions $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_\beta\}$ are stochastically generated according to the current state of $\mathbf{p}$. This means that each $\mathbf{v}_i$ is a bit string, where the probability of the j-th bit being 1 is given by p_j. Each of these candidate solutions is then evaluated against the optimization problem at hand, giving a set of measures $\{\Phi(\mathbf{v}_1), \Phi(\mathbf{v}_2), \ldots, \Phi(\mathbf{v}_\beta)\}$.

After ordering this set, the solution that corresponds to the best evaluative measure, $\mathbf{v}_{\text{best}}$, is then used to update $\mathbf{p}$ with the standard competitive learning rule (see Hertz, Krogh, & Palmer, 1991):

$$\mathbf{p}^{\text{new}} = (1 - \lambda)\,\mathbf{p}^{\text{old}} + \lambda \mathbf{v}_{\text{best}}, \tag{5}$$

where $0 < \lambda \leq 1$ is a learning rate parameter. Each element of the probability vector $\mathbf{p}$ then has a small probability μ_{prob} of being subjected to 'mutation', which involves an additive shift μ_{shift} towards the maximum entropy value of 0.5. The next iteration then commences, generating another set of candidate solutions according to the updated probability vector. Across all iterations a record is maintained of the best solution vector found, and the algorithm terminates once a fixed number of evaluations τ have proceeded without improvement to this best solution.

In both of the applications of the basic PBIL approach employed here, the following parameter values were used: a total of $\beta = 50$ solutions were generated in each potential solution set, a learning rate of $\lambda = 0.1$ was used, the mutation probability was $\mu_{\text{prob}} = 0.02$ with an associated shift of $\mu_{\text{shift}} = 0.05$, and the optimization process was terminated after $\tau = 5,000$ potential solutions were tried without improvement. These values were determined, reasonably uncritically, on the basis of values previously employed in tackling various other optimization problems using PBIL (see Baluja, 1994; Baluja & Davies, 1998), and proved to work well.

Candidate cluster generation

To generate an n cluster model, the PBIL probability vector $\mathbf{p}$ was constructed with $\alpha = n^2$ elements, each initialized to 0.5, encoding the presence or absence of each of the n objects in the n clusters. Effectively, $\mathbf{p}$ serves as a vectorial representation of the probabilities of the elements in the matrix $\mathbf{F}$ being 0 or 1. Each cluster structure represented in the potential solution set stochastically generated from $\mathbf{p}$ was appended with a cluster containing all stimuli, to accommodate an additive constant in the similarity model. For these augmented cluster structures, the best-fitting weight vector, in terms of the error measure E, was found using a standard non-negative least-squares algorithm (Lawson & Hanson, 1974). The actual minimal values of E were then used to compare the set of potential solutions, choose the best, and update the probability vector.

During the process of generating and evaluating potential solutions, a record was maintained of the $\kappa = 100$ clusters within these solutions with the greatest weights. It was these clusters which served as the candidates for inclusion in the final model generated during the second stage. For the 10-object numbers domain reported later, this process took about 315 seconds when implemented in Matlab, and run on an 1100 megahertz PC. For the 16-object phoneme domain reported later, this process took about 1,280 seconds.

Model generation

In the model generation stage, a probability vector $\mathbf{p}$ was constructed with $\beta = \kappa = 100$ elements, all initialized to 0.1, encoding the presence or absence of each of the candidate

clusters from a potential model. After stochastically generating a set of potential models, the all-encompassing cluster was again added, and the best-fitting weight vector, in terms of E, found as before. In this stage, however, the assumption regarding data precision embodied by the estimate s allowed each potential model to be evaluated in terms of the BIC measure, considering both the error E and the number of clusters included in the model. Thus, for each set of potential models, the one that constituted a trade-off between data-fit and model complexity was used to update the probability vector.

The initial value for the elements of $\mathbf{p}$ was set to 0.1 to reflect the fact that, in general, most of the candidate clusters were excluded from models with relatively low BIC values. Experience suggests that, unless the similarity data constraining model development is particularly precise, only about 10% of the candidate clusters, under the specified parameter regime, are likely to be included in the final model. While there is no fundamental reason why the theoretically preferable maximum entropy value of 0.5 could not be employed, using an initial value of 0.1 proved to generate equally good final models with greater computational efficiency. For example, in relation to the 10-object numbers domain reported later, the model generation process using the value 0.1 took about 54 seconds on average when implemented in Matlab, and run on an 1100 megahertz PC. Using the value 0.5 increased this average running time to about 251 seconds. In terms of scaling to larger problem sizes, the 16-object phoneme domain reported later took an average of 368 seconds when using the initial value 0.1.

Evaluation using artificial data

As a basic test of the proposed method for generating additive clustering models, its ability to recover a known, but noise corrupted, set of clusters with associated weights was examined. Largely following the methodology adopted by Tenenbaum (1996), a set of 12 stimuli were assigned to 8 clusters with probability 0.5, and cluster weights were chosen by sampling independently from the uniform distribution on the interval [0.1, 0.6]. The resultant similarity matrix was then calculated and normalized, and zero-mean Gaussian noise with a standard deviation of 0.10 was independently added to each similarity value.

In attempting to recover the known additive clustering representation, all three of the heuristically determined broad assumptions regarding data precision, corresponding to the s values 0.05, 0.10 and 0.15, were considered. Having derived a set of candidate clusters by recording the most highly weighted clusters found in learning an 8 cluster model, the model generating stage was repeated 50 times under each of these precision assumptions.

Figure 1 graphically depicts the results of the $3 \times 50 = 150$ additive clustering models generated in this way. Each model is represented in terms of the number of clusters it contained (along the x-axis), and the percentage of variance in the data it explained (along the y-axis). Models generated under different assumptions regarding data precision are shown by different markers. To allow some visual indication of the relative frequency with which various models were derived, the x location of each of the markers was subjected to a small random displacement about the appropriate integral cluster cardinality value.

Figure 2 shows the expected complexity controlling behavior of the new method, in that the assumption of greater data precision leads to models being generated that have

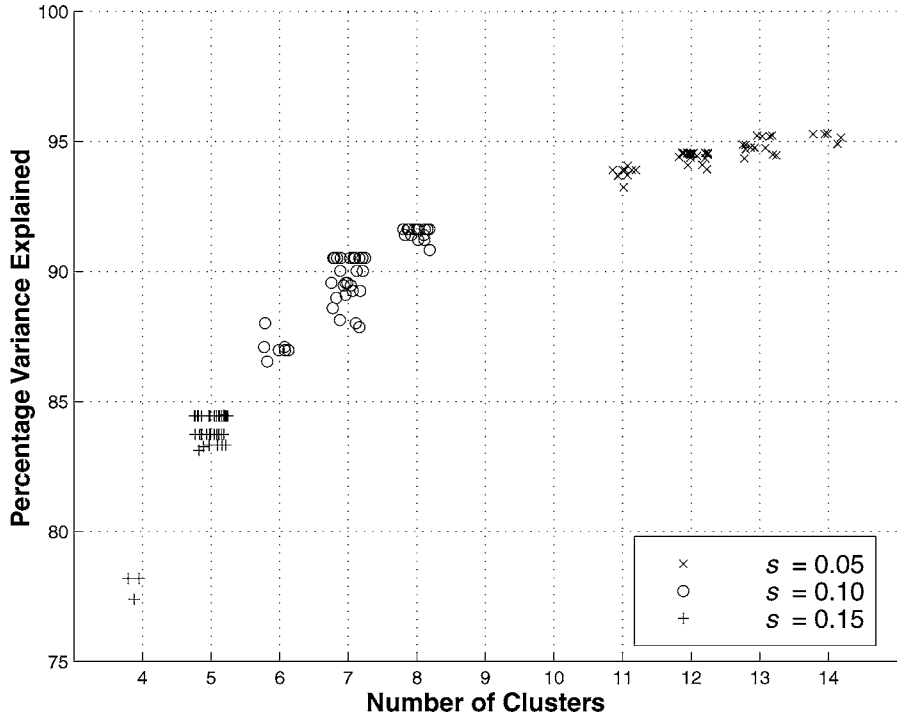

Figure 1. Performance on 8-cluster artifical data with noise standard deviation $\sigma = 0.10$.

greater numbers of clusters, and explain more of the variance in the data. It can be seen that, when the correct assumption regarding data precision is made by setting $s = 0.10$, there are two models, with 7 and 8 clusters respectively, that are frequently generated. An examination of the actual recovered cluster structures revealed that the best-fitting 8 cluster representation was the one used to generate the data, while the best-fitting 7 cluster representation simply omitted the least weighted cluster from the generating model. The frequent recovery of the 7 cluster model provides a concrete demonstrations of the tendency of the BIC to underfit the data by favoring an additive clustering model with too few clusters.

Nevertheless, the performance of the new method on this recovery problem, and on several others examined but not reported in detail here, suggests that it is capable of recovering the bulk of a noise perturbed additive clustering structure when an accurate estimate of data precision is available. The main caveat to place on the application of the new method is that great care should be taken in accepting additive clustering models that assume high levels of data precision on a heuristic basis. If the true precision of the data does not meet the same level as the assumed precision, the derived additive clustering model is likely to have overfit the data, and the less weighted clusters should be interpreted with caution.

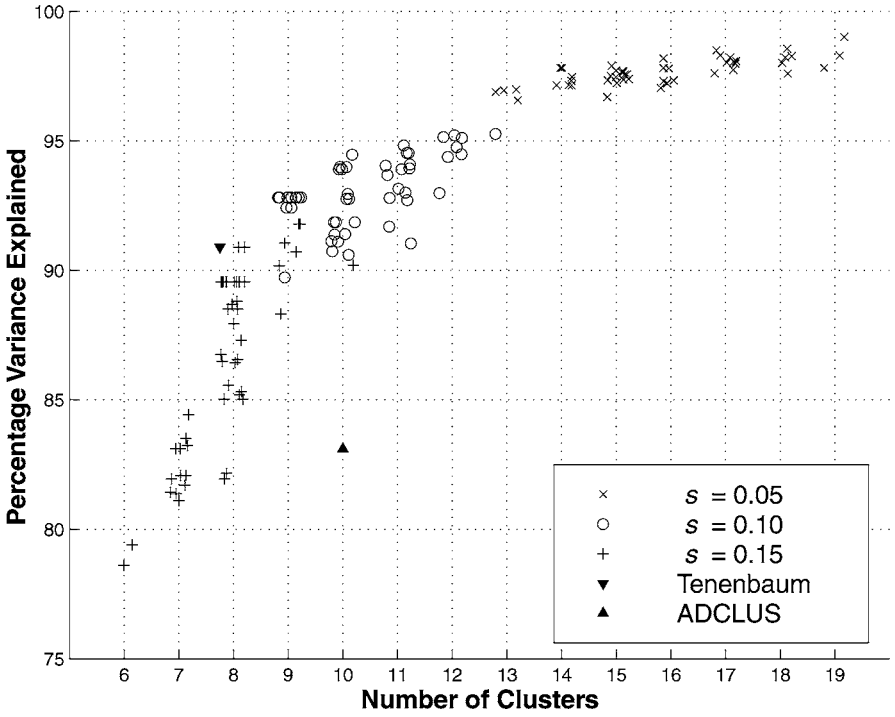

Figure 2. Performance on the numbers domain.

Comparative applications

Although a claim of this sort is necessarily subjective and open to debate, it is probably reasonable to argue that there are four previously reported additive clustering models that are sufficiently impressive to serve as de-facto performance benchmarks. Two of these, relating to number similarity and consonant phonemes confusion data, were generated using Tenenbaum's (1996) technique. The third relates to the social relationships between industrial workers, and was developed using Mirkin's (1987) approach, while the fourth relates to kinship data, and was derived using SINDCLUS. Accordingly, the application of the proposed new method to these four data sets provides a means for its comparative evaluation.

Numbers 0–9

As described earlier, Tenenbaum's (1996) number model relates to human judgments of the abstract conceptual similarities between the numbers 0–9 collected by Shepard, Kilpatrick and Cunningham (1975). The actual similarity matrix was generated by averaging ratings both across subjects, and across three conditions—verbal, written-numeral, and written-dots—of stimulus presentation. Unfortunately, as anticipated earlier, the raw data

needed to estimate the precision of this averaged data are not available, so the heuristically determined broad assumptions regarding data precision, corresponding to the s values 0.05, 0.10 and 0.15, were again considered.

Figure 2 shows the results of the $3 \times 50 = 150$ additive clustering models generated in this way. As before, each model is represented in terms of its number of clusters and data-fit, with models generated under different data precision assumption shown by different markers. Also shown are several models previously derived using the same data, represented in the same way, but indicated by filled markers.

Figure 2 again graphically depicts the general complexity controlling behavior of the new method, with greater precision giving more complicated models with better data-fit. More specifically, however, figure 2 reveals that, using $s = 0.15$, an 8 cluster model with data-fit equivalent to that presented by Tenenbaum (1996) was twice derived. Furthermore, the performance of the new method is clearly better than the original ADCLUS method (Shepard & Arabie, 1979), since models with the same number of clusters typically explain 10% more of the variance, and there are models with fewer clusters which explain the same amount of variance.

On the basis of figure 2, two of the more impressive models appear to be those which exhibit the best data-fit using 8 and 9 clusters. These models are detailed in Table 1, which lists the clusters, their associated weights, the additive constant, and the percentage of variance in the data explained. This information confirms that, as suspected on the basis of their equivalent data-fit, the 8 cluster models generated by the new method and by Tenenbaum (1996) are identical. The 9 cluster model simply appends a lowly weighted cluster, almost comprised of the even numbers, which contributes to the explanation of about another 2% of the variance.

Perhaps the most disappointing aspect of the new method's performance in the number domain is that the best-fitting 8 cluster solution was derived relatively infrequently. Tenenbaum (1996) indicates that this model was the best found in 5 runs of his algorithm,

Table 1. The 8 and 9 cluster number models.

Stimuli in cluster	Weight in 8 cluster model	Weight in 9 cluster model
2 4 8	0.444	0.353
0 1 2	0.345	0.358
3 6 9	0.331	0.326
6 7 8 9	0.291	0.249
2 3 4 5 6	0.255	0.241
1 3 5 7 9	0.216	0.239
1 2 3 4	0.214	0.232
4 5 6 7 8	0.172	0.183
2 4 6 7 8	–	0.105
Additive constant	0.148	0.129
Variance explained	90.9%	92.8%

whereas a ratio of closer to 1 in 25 is suggested by figure 2. However, in 50 separate trials of the new method using an s value of 0.125, the same 8 cluster model was found 7 times. Indeed, if attention is restricted to only those trials which derived 8 cluster models, the best-fitting model was found with a success rate of 1 in 2. On this basis, it seems reasonable to claim that the performance of the new method in the number domain is at least comparable to that reported by Tenenbaum (1996).

Consonant phonemes

The second benchmark additive clustering model reported by Tenenbaum (1996) relates to a similarity matrix for 16 consonant phonemes, derived from confusion probabilities originally collected by Miller and Nicely (1955). Tenenbaum (1996) generated an 8 cluster model accounting for 90.2% of the variance, and noted its superiority to an 8 cluster solution generated by the MAPCLUS algorithm (Arabie & Carroll, 1980) that explained only 88.3% of the variance. Even though the quantitative form of the relationship between cluster numbers and data-fit is unknown, both of these models would seem likely to compare favorably to the best reported ADCLUS model (Shepard & Arabie, 1979), which required 16 clusters to explain 94.5% of the variance. Therefore, as well as benchmarking the performance of the new method against Tenenbaum's (1996) 8 cluster model, it is also of some interest to examine the way in which this complexity issue is resolved by the application of the BIC evaluative measure.

To these ends, the phoneme similarity data was used to generate 150 models, once again employing the three levels of data precision corresponding to s values of 0.05, 0.10 and 0.15. Figure 3 summarizes the results of these trials, and indicates that, when $s = 0.05$, the new method frequently generated 8 cluster models with better data-fit than the model reported by Tenenbaum (1996). In addition, the range of cluster cardinalities in the derived models suggest that the 16 cluster ADCLUS model corresponds to an implicit assumption of extremely precise data that seems highly unlikely to be justified, particularly given the methodological origins of the data in measures of confusion. In any case, Figure 3 indicates that an 11 cluster model derived by the new method explained more of the variance in the data than the ADCLUS model.

Table 2 details the best of the 8 cluster models that exhibited greater data-fit than Tenenbaum's (1996) model, as well as the frequently derived best fitting 5 cluster model that explained more than 80% of the variance. The only difference in cluster structure between the new 8 cluster model and Tenenbaum's is the inclusion of s in the previously incomplete cluster of unvoiced consonants, which contributed to the explanation of another 1.6% of the variance. Meanwhile, the impressive 5 cluster model is seen simply to consist of a subset of the larger model, selecting the front unvoiced fricatives, unvoiced stops, back voiced stops, front voiced, and (almost) voiced consonant clusters.

Industrial workers

The third comparative evaluation of the new method involves correlational measures of similarity derived by Breiger, Boorman and Arabie (1975, Table 3)[3] from an observational

Table 2. The 5 and 8 cluster consonant phoneme models.

Stimuli in cluster	Weight in 5 cluster model	Weight in 8 cluster model
f θ	0.350	0.399
p t k	0.267	0.162
d g	0.243	0.243
b v ð	0.176	0.182
d g v ð z ž	0.066	0.075
p k	–	0.197
m n	–	0.127
p t k f θ s š	–	0.049
Additive constant	0.034	0.024
Variance explained	81.3%	91.8%

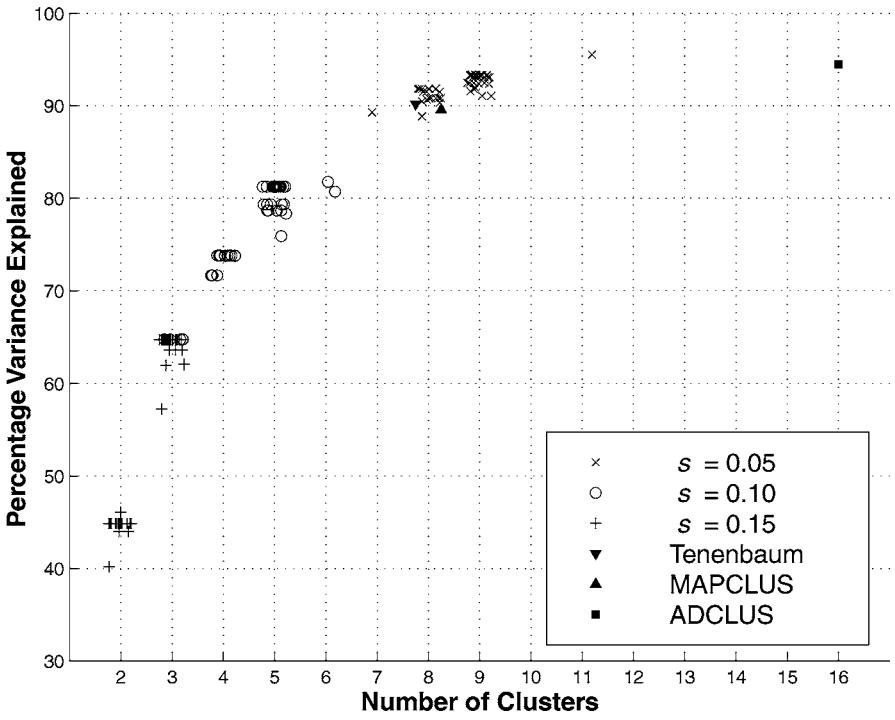

Figure 3. Performance on the consonant phonemes domain.

Table 3. The 12 cluster industrial worker model.

Stimuli in cluster	Weight
W2 W5 I3	0.356
W6 S2 I3	0.223
W5 W6 W7 W8 W9 S4 I1	0.194
W1 W3 W4 S1 I1	0.178
W4 W6 W7 W8 W9 S1 S2 S4	0.176
W7 W8 W9 S2 I3	0.155
W1 W2 I1	0.136
W1 W2 W3 W4 W5 S1 S2 I1	0.127
W1 W2 W3 W4	0.099
W6 W8 W9 S4	0.096
W1 W2 W3 W4 W5 S1 I1 I3	0.067
W1 W3 W4 W7 S1	0.065
Additive constant	0.041
Variance explained	94.8%

study of 14 industrial workers conducted by Roethlisberger and Dickson (1939). As before, 50 trials were conducted at each of three levels of data precision, and the results are presented in figure 4.

It can be seen, in relation to 7 cluster models, that the benchmark developed by Mirkin (1987) is not achieved by the new method, with the best fitting model explaining 87.7%, rather than 89.7%, of the variance in the data. Only once 8 clusters are incorporated is more than 90% of the variance able to be explained, although this performance remains superior to that reported for ADCLUS, which required 10 clusters to explain 89.0%. Nevertheless, under assumptions of data precision corresponding to s values greater than 0.1, which seems most reasonable, Mirkin's (1987) solution is likely to be preferable. Perhaps, therefore, the main contribution of the trials summarized in figure 4 relates to the 12 cluster model generated under the assumption of precise data, explaining almost 95% of the variance, which is detailed in Table 3.

Kinship

The final comparative evaluation involves data collected by Rosenberg and Kim (1975) measuring the similarity of 15 common kinship terms, such as 'father', 'cousin', and 'grandmother'. These data were calculated from the results a sorting procedure performed by 6 groups of 85 subjects, where each kinship term was placed into one of a number of groups, under various conditions of subject instruction.

Figure 5 displays the results achieved by the new method on this data, again using 50 trials, and assuming the same three heuristic levels of data precision. Also shown in the 'benchmark' performance of the SINDCLUS algorithm (Chaturvedi & Carroll, 1994; see

Table 4. The 5 and 9 cluster kinship models.

Stimuli in cluster	Weight	Weight
granddaughter grandfather grandmother grandson	0.295	0.296
aunt cousin nephew niece uncle	0.273	0.223
brother daughter father mother sister son	0.229	0.192
aunt daughter granddaughter grandmother mother niece sister	0.199	0.212
brother father grandfather grandson nephew son uncle	0.198	0.211
brother sister	–	0.291
father mother	–	0.276
aunt uncle	–	0.266
nephew niece	–	0.263
Additive constant	0.248	0.242
Variance explained	80.6%	89.8%

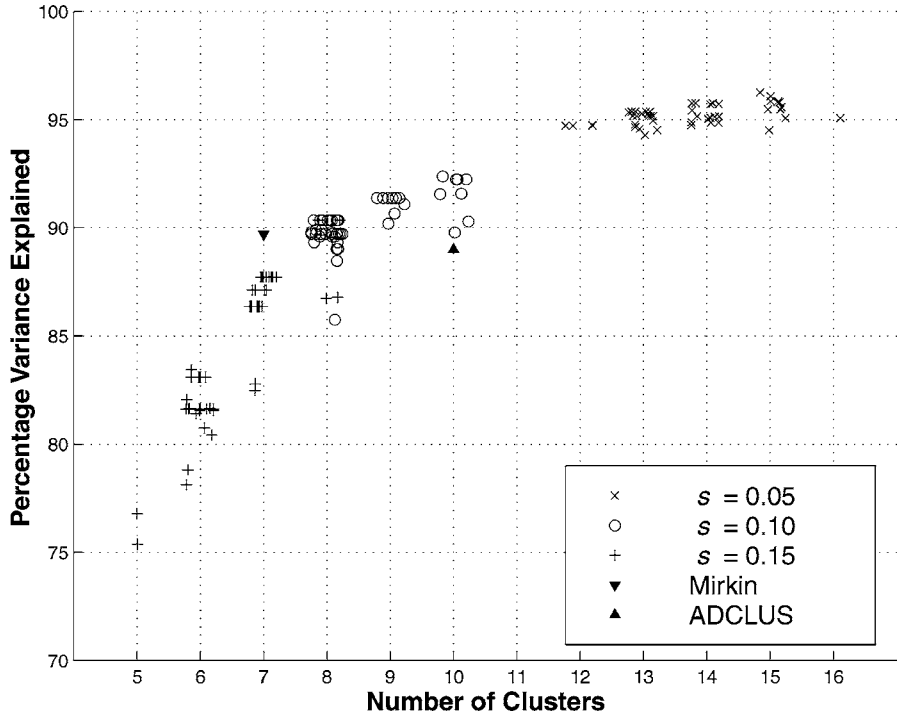

Figure 4. Performance on the industrial workers domain.

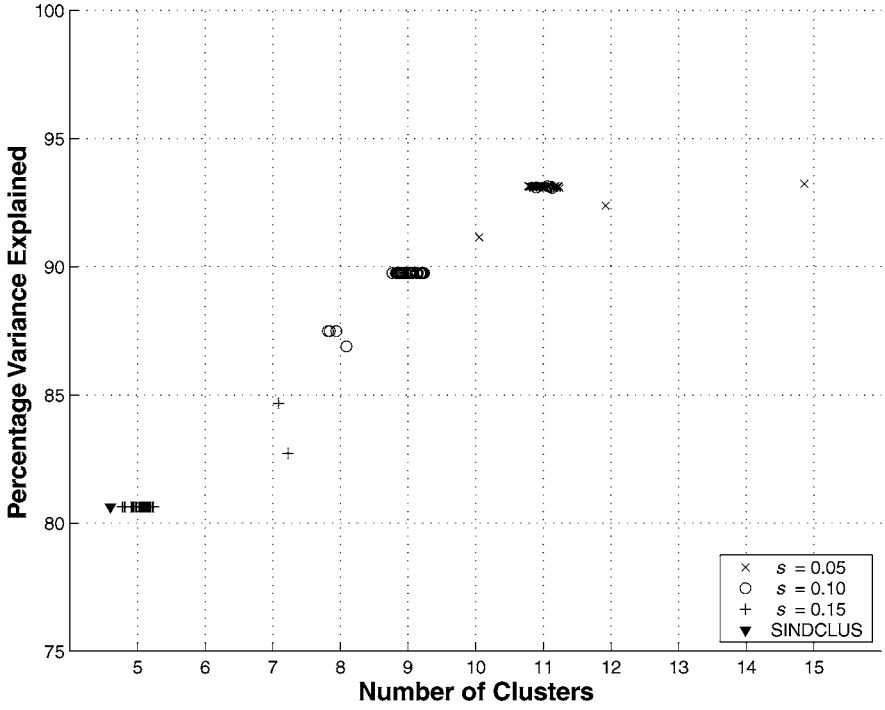

Figure 5. Performance on the kinship domain.

also Arabie, Carroll, & DeSarbo, 1987), which found a 5 cluster solution explaining 80.6% of the variance in the data[4].

Figure 5 shows that the new method derives the benchmark SINDCLUS model on 48 out of the 50 trials where $s = 0.15$. Table 4 details this model, and the 9 cluster model explaining 89.8% of the variance that is frequently generated when $s = 0.10$. The 5 cluster model is readily interpretable in terms of the familial relationships it represents (e.g., the nuclear family kinships 'brother', 'daughter', 'father', 'mother', 'sister', 'son'), and the 9 cluster model simply appends more detailed breakdowns of several of these relationships (e.g., augmenting the nuclear family cluster with the sibling 'brother' and 'sister', and parental 'father' and 'mother' clusters).

Discussion

On the basis of these four applications of the new additive clustering method, in which performance twice met a benchmark (numbers, kinship), once improved upon a benchmark (phonemes), and once fell short of a benchmark (workers), it seems reasonable to claim at least comparable performance with the best previously developed methods. Equally importantly, the potential of the new method to balance the number of clusters included in a

model against improvements in data-fit could extend the capabilities of previous methods. Without the explicit quantitative Bayesian basis provided by the BIC measure, there is no guarantee that models generated by any other method correspond to reasonable assumptions regarding the precision of the data from which they are derived.

One issue not directly addressed by the trials summarized by figures 2–5 is whether the evident consistency of performance happened to be the result of particularly good sets of underlying candidate clusters. An investigation of this possibility showed that, in fact, those clusters ultimately included in the best-fitting models of any cardinality, for all four domains, were consistently derived by repeated applications of the candidate cluster generation algorithm. For this reason, it seems unlikely that either using more candidate clusters, or combining repeated runs of cluster generation would significantly improve the new method's performance.

Similarly, a number of other explorations failed to produce a clearly superior method. Following Baluja and Davies (1997, 1998), the possibility of incorporating knowledge of the conditional dependencies between the bit string states stored in **p** was investigated. In practical terms, the additional computational complexity required to model these dependencies did not prove to have sufficiently large compensating performance benefits. It seems that available computational resources were better spent generating candidate solutions, rather than attempting to constrain further the region of the state space being explored. In particular, attempts to capture the dependencies between stimuli in terms of a single cluster, using optimal mutual information trees, actually resulted in a worsening of performance. Whether this observation remains true for domains containing larger numbers of objects than those examined here remains an open question, although it is worth noting that the performance of the new method on the largest domain, the phoneme domain, was probably the most impressive.

There is, however, some scope for further investigation of the candidate cluster generation approach. In particular, the decision to extract clusters from the process of generating an n cluster model is somewhat arbitrary. Clearly, if a model with too few clusters is sought, there is unlikely to be sufficient variety in the candidate cluster set, while the derivation of a model with too many clusters may prevent a sufficient focus being placed on good clusters, and also increases computational overheads. The decision to derive an n cluster model represents a heuristic compromise between these two scenarios, but it may be possible to choose an appropriate model size for candidate cluster generation on a more principled basis. This capability would be particularly useful when dealing with larger stimulus domains, since the main computational burden of the current method lies in generating the candidate clusters. The possibility of learning simple evaluation functions, rather than sophisticated probabilistic models, to constrain the search for good clusters or models (Boyan & Moore, 1998) remains to be explored. In this regard, an attempt to develop useful evaluation functions which identify key features of good additive clustering models seems likely to be a worthwhile avenue for future research.

In the meantime, however, the new method for additive clustering developed here has the advantages of being transparently simple, easily implemented, and reasonably computationally efficient. It allows for the introduction of collateral information into the process of model construction, and its progress is amenable to meaningful diagnostic interpretation

at all stages. On the four benchmark data-sets examined, it has proven to generate models that rival or better the best previously reported models and, finally, it has introduced a mechanism that has the potential to address the fundamental issue of additive clustering model complexity in a practical way.

Acknowledgments

The majority of this work was completed while the author was employed at the Australian Defence Science and Technology Organisation. The helpful comments of two anonymous referees are acknowledged.

Notes

1. The rationale for treating only the weights as parameters is that, given a particular cluster structure, it is the weights that are free to vary to minimize an error measure. This conceptualization is consistent with that originally advocated by Shepard and Arabie (1979, p. 102), but it should be noted that an alternative conceptualization is possible, in which the cluster membership variables are also treated as parameters.
2. 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.
3. Modified according to the corrections noted by Shepard and Arabie (1979).
4. Although Chaturvedi and Carroll (1994, p. 305) report that the representation explains 81.3% of the variance, a measure of 80.6% was obtained by finding the best-fitting weights in terms of the version of the data set used here. Presumably, the discrepancy lies in the data, or in the exact measure of variance accounted for that was used (e.g., whether self-similarity measures were considered). The important point is that the new method recovers the same cluster structure as SINDCLUS.

References

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

Arabie, P., & Carroll, J. D. (1980). MAPCLUS: A mathematical programming approach to fitting the ADCLUS model. *Psychometrika, 45:2*, 211–235.

Arabie, P., Carroll, J. D., & DeSarbo, W. S. (1987). *Three-way scaling and clustering*. Newbury Park, CA: Sage.

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

Baluja, S. (1994). Population-based incremental learning: A method for integrating genetic search based function optimization and competitive learning. Technical Report CMU-CS-94-163, Carnegie-Mellon University.

Baluja, S., & Davies, S. (1997). Using optimal dependency-trees for combinatorial optimization: Learning the structure of the search space. Technical Report CS-CMU-97-107, Carnegie-Mellon University.

Baluja, S., & Davies, S. (1998). Fast probabilistic modeling for combinatorial optimization. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence* (pp. 469–476). Madison, WI: AAAI Press/MIT Press.

Boyan, J. A., & Moore, A. W. (1998). Learning evaluation function for global optimization and boolean satisfiability. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence* (pp. 3–10). Madison, WI: AAAI Press/MIT Press.

Breiger, R. L., Boorman, S. A., & Arabie, P. (1975). An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling. *Journal of Mathematical Psychology, 12*, 328–383.

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.

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

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

Gati, I., & Tversky, A. (1982). Representations of qualitative and quantitative dimensions. *Journal of Experimental Psychology: Human Perception and Performance, 8:2*, 325–340.

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

Grünwald, P. (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology, 44:1*, 133–152.

Hertz, J., Krogh, A., & Palmer, R. G. (1991). *Introduction to the theory of neural computing*. Redwood City, CA: Addison-Wesley.

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

Johnson, E. J., & Tversky, A. (1984). Representations of perceptions of risks. *Journal of Experimental Psychology: General, 113:1*, 55–70.

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

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

Lawson, C. L., & Hanson, R. J. (1974). *Solving least squares problems*. Englewood Cliffs, NJ: Prentice-Hall.

Lee, M. D. (1998). *Active cognitive representation and learned categorical perception*. In *25th Australasian Experimental Psychology Conference*, Hobart.

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

Lee, M. D. (1999b). An extraction and regularization approach to additive clustering. *Journal of Classification, 16:2*, 255–281.

Lee, M. D. (2001a). Determining the dimensionality of multidimensional scaling representations for cognitive modeling. *Journal of Mathematical Psychology, 45:1*, 149–166.

Lee, M. D. (2001b). On the complexity of additive clustering models. *Journal of Mathematical Psychology, 45:1*, 131–148.

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., Balasubramanian, V., & Pitt, M. A. (2000). Counting probability distributions: Differential geometry and model selection. In *Proceedings of the National Academy of Sciences, 97*, 11170–11175.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review, 4:1*, 79–95.

Raftery, A. E. (1999). Bayes factors and BIC: Comment on Weakliem. *Sociological Methods and Research, 27*, 411–427.

Roethlisberger, F. J., & Dickson, W. J. (1939). *Management and the worker*. Cambridge, MA: Harvard University Press.

Rosenberg, S., & Kim, M. P. (1975). The method of sorting as a data-generating procedure in multivariate research. *Multivariate Behavioral Research, 10*, 489–502.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6:2*, 461–464.

Shepard, R. N. (1974). Representation of structure in similarity data: Problems and prospects. *Psychometrika, 39:4*, 373–422.

Shepard, R. N., & Arabie, P. (1979). Additive clustering representations of similarities as combinations of discrete overlapping properties. *Psychological Review, 86:2*, 87–123.

Shepard, R. N., Kilpatrick, D. W., & Cunningham, J. P. (1975). The internal representation of numbers. *Cognitive Psychology, 7*, 82–138.

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. 2). Cambridge, MA: MIT Press.
Woodruff, C. J. (1998). *Establishing a psychophysics of texture*. In *25th Australasian Experimental Psychology Conference*, Hobart.

Received June 26, 1998
Revised June 22, 2001
Accepted June 22, 2001
Final manuscript June 22, 2001