Journal of Classification 19:69-85 (2002)
DOI: 10.1007/s00357-001-0033-y

Generating Additive Clustering Models with Minimal Stochastic Complexity

Michael D. Lee

University of Adelaide

Abstract: Additive clustering models provide a conceptually simple and representationally powerful approach to extracting features from similarity data. Objects are represented according to the presence or absence of a set of weighted features, and their observed similarities are modeled using the recovered features that objects have in common. Unlike partitioning or hierarchical clustering approaches, most approaches to additive clustering place no constraints on the way features may be assigned to objects. This representational freedom demands, however, that the issue of additive clustering model complexity is addressed. It is important that additive clustering models are generated so as to balance the competing demands of goodness-of-fit and complexity for substantive interpretation. This paper uses previous analytic results to derive a stochastic complexity criterion measure for additive clustering models. This measure simultaneously takes into account the goodness-of-fit, the number of clusters, and the complexity associated with the patterns of cluster inclusion and overlap within the model. A new algorithm for fitting additive clustering models to similarity data is then developed, using the stochastic complexity measure to control the balance between goodness-of-fit and complexity. The ability of the algorithm to recover known features and weights is assessed using Monte Carlo techniques, and its application to empirical data is demonstrated using a previously examined data set that measures the similarities among kinship terms.

Keywords: Additive clustering; Overlapping clustering; Stochastic complexity.

I thank Phipps Arabie and three anonymous reviewers for helpful comments on an earlier version of this article.

Author's Address: Michael D. Lee, Department of Psychology, University of Adelaide, Adelaide SA 5005, Australia; telephone:+61 8 8303 6096; fax: +61 8 8303 3770; e-mail: michael.lee@psychology.adelaide.edu.au. URL: http://www.psychology.adelaide.edu.au/ members/staff/michaellee/homepage .

1. Introduction

Additive clustering models (e.g., Arabie and Carroll 1980; Chaturvedi and Carroll 1994; Mirkin 1987, 1996; Shepard and Arabie 1979; Tenenbaum 1996) provide representationally powerful—but conceptually simple—accounts of the observed similarities between sets of objects. 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 nearly always entirely unconstrained[1]. Unlike standard partitioning clustering approaches that place each stimulus in only one cluster, additive clustering allows each object to belong to any number of clusters. Unlike hierarchical clustering approaches, additive clustering places no nesting constraints upon the sets of objects that may be encompassed by a cluster.

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

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

and the weights fitted for each of the clusters, denoting their importance or salience, are defined by a derived vector $\mathbf{w} = (w_1, w_2, \ldots, w_m)$, then the estimated similarity of the ith and jth objects is

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

which includes a positive constant, in the form of the weight associated with a universal cluster that includes all of the objects.

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 matrix to be accommodated perfectly by an additive clustering model. While the flexibility afforded

The exception comes in the form of additive clustering techniques (e.g., Carroll and Chaturvedi 1995; Chaturvedi, Green, and Carroll 2001; DeSarbo 1982) that allow users to place constraints on overlapping clustering solutions.

by additive clustering is clearly desirable in providing an ability to model 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 goodness-of-fit and minimizing model complexity is often acknowledged in the development of techniques to generate additive clustering models, and has typically been tackled through the general strategy of attempting to use a minimal number of clusters to provide a maximal level of goodness-of-fit. Some techniques (e.g., Tenenbaum 1996) accomplish this task by setting the number of clusters to be derived at a fixed value, and then seeking the best goodness-of-fit possible, while other techniques (e.g., Lee 1999a) set a target goodness-of-fit level, and then seek a minimal number of clusters that achieve this fit. Only one technique (Lee, in press) explicitly quantifies the trade-off between accuracy and complexity during the process of model generation. This technique uses a formulation of the Bayesian Information Criterion (BIC: Schwarz 1978; see Kass and Raftery 1995; Myung and Pitt 1997 for overviews) developed by Lee (2001b).

A weakness of this approach, however, is that the BIC, like Akaike's (1974) Information Criterion (AIC), is sensitive only to the parametric complexity of additive clustering models, which corresponds simply to the number of clusters they use. As argued by Lee (2001b), additive clustering model complexity is also influenced by the patterns of cluster encompassment, cluster overlap, and general cluster structure of a model, and not simply the number of clusters. Lee (2001b) proceeds to derive a measure of additive clustering model complexity, based on the so-called Laplacian approximation to a marginal probability density (see Kass and Raftery 1995, p. 777), that is sensitive to variations in cluster structure.

As it turns out, much of the analysis presented in Lee (2001b) may be used to derive a measure of the 'stochastic complexity' of additive clustering models, as defined by Rissanen (1996). The stochastic complexity measure has the advantage of combining goodness-of-fit, the number of clusters, and the structural complexity of an additive clustering model. This paper presents a new algorithm for fitting additive clustering models to similarity data, using the stochastic complexity measure to constrain the derived representation so that it balances the competing demands of goodness-of-fit and model complexity.

2. Stochastic Complexity

Rissanen (1996) presents a reparameterization-invariant form of the Stochastic Complexity Criterion (SCC), as:

$$\mathrm{SCC} = -\ln\, p\left(D \mid \mathbf{p}^*\right) + \frac{P}{2}\ln\left(\frac{N}{2\pi}\right) + \ln\int\sqrt{\det\mathbf{I}\left(\mathbf{p}\right)}.d\mathbf{p}, \qquad (2)$$

where D is a data sample of size N that constrains the model, $\mathbf{p}$ is a vector containing P model parameters, $p\left(D \mid \mathbf{p}^*\right)$ is the maximum probability density of the model, found by evaluating at the best-fitting parameter values $\mathbf{p}^*$, and $\mathbf{I}\left(\mathbf{p}\right)$ is the Fisher information matrix, defined using the expectation:

$$\mathbf{I}_{xy}\left(\mathbf{p}\right) = -E_{\mathbf{p}}\left[\frac{\partial^2 \ln\, p\left(D \mid \mathbf{p}\right)}{\partial p_x \partial p_y}\right]. \qquad (3)$$

The first term in the stochastic complexity criterion may be regarded as a 'maximum-likelihood' term, measuring the goodness-of-fit of the model under the best-fitting parameterization. The remaining terms may be regarded as 'complexity' terms, relevant to the complexity effects of including extra parameters in a model, and the 'functional form' complexities (Myung and Pitt 1997) associated with the degree to which values of parameters constrain each other.

In the context of measuring the complexity of a given additive clustering model, Lee (2001b) argues that it is reasonable to treat the cluster structure defined by the membership variables in the matrix $\mathbf{F}$ as the model *per se*, and the weights in the vector $\mathbf{w}$ as parameters of the model. In addition, Lee (2001b) follows Tenenbaum (1996), in using a probabilistic formulation that characterizes each of the similarities as a Gaussian distribution with common variance. This approach means that the probability of a similarity matrix $\mathbf{S}$ arising under a particular additive clustering structure $\mathbf{F}$, using a particular weight parameterization $\mathbf{w}$, is given by

$$\begin{aligned} p\left(\mathbf{S} \mid \mathbf{F}, \mathbf{w}\right) &= \prod_{i<j} \frac{1}{\left(\sigma\sqrt{2\pi}\right)} \exp\left(-\frac{\left(s_{ij} - \hat{s}_{ij}\right)^2}{2\sigma^2}\right) \\ &= \frac{1}{\left(\sigma\sqrt{2\pi}\right)^{n(n-1)/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i<j}\left(s_{ij} - \hat{s}_{ij}\right)^2\right), \qquad (4) \end{aligned}$$

where σ^2 is the common variance. As argued by Lee (2001a; see also Lee submitted, 1999b), this variance quantifies the *inherent* precision of the data and can be estimated based on an understanding of the process by which the data was generated.

Given a sample estimate s of σ, the first 'maximum likelihood' term of the stochastic complexity criterion may be given the following additive clustering formulation:

$$\begin{aligned} -\ln p\left(D \mid \mathbf{p}^{*}\right) &= -\ln p\left(\mathbf{S} \mid \mathbf{F}, \mathbf{w}^{*}\right) \\ &= \frac{1}{2s^{2}} \sum_{i<j} \left(s_{ij} - \hat{s}^{*}_{ij}\right)^{2} + \frac{n\left(n-1\right)}{2} \ln\left(s\sqrt{2\pi}\right). \end{aligned} \quad (5)$$

A symmetric similarity matrix for n objects, where self-similarity often remains undefined, contains $n\left(n-1\right)/2$ similarity measures. Accordingly, for additive clustering models, the first of the 'complexity' terms in the stochastic complexity criterion is given by

$$\frac{P}{2}\ln\left(\frac{N}{2\pi}\right) = \frac{m}{2}\ln\left(\frac{n\left(n-1\right)}{4\pi}\right). \quad (6)$$

The additive clustering formulation of the second 'complexity' term is more involved, and relies on the derivation given in Lee (2001b), whereby

$$\begin{aligned} -\frac{\partial^{2}\ln p\left(\mathbf{S} \mid \mathbf{F}, \mathbf{w}\right)}{\partial w_{x}\partial w_{y}} &= \frac{\partial^{2}}{\partial w_{x}\partial w_{y}}\left[\frac{1}{2s^{2}}\sum_{i<j}\left(s_{ij}-\hat{s}_{ij}\right)^{2}\right] \\ &= \frac{\partial^{2}}{\partial w_{x}\partial w_{y}}\left[\frac{1}{2s^{2}}\sum_{i<j}\left(s_{ij}-\sum_{k} w_{k} f_{ik} f_{jk}\right)^{2}\right] \\ &= \frac{\partial}{\partial w_{y}}\left[-\frac{1}{s^{2}}\sum_{i<j}\left(s_{ij}-\sum_{k} w_{k} f_{ik} f_{jk}\right) f_{ix} f_{jx}\right] \\ &= \frac{1}{s^{2}}\sum_{i<j} f_{ix} f_{jx} f_{iy} f_{jy}. \end{aligned}$$

This result allows the required Fisher information matrix to be written as

$$\begin{aligned} \mathbf{I}_{xy}\left(\mathbf{w}\right) &= -E_{\mathbf{w}}\left[\frac{\partial^{2}\ln p\left(\mathbf{S} \mid \mathbf{F}, \mathbf{w}\right)}{\partial w_{x}\partial w_{y}}\right] \\ &= \frac{1}{s^{2}}\mathbf{G}, \end{aligned} \quad (7)$$

where

$$\mathbf{G} =$$

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

As noted in Lee (2001b, p.142), the interpretation of the elements of **G** using 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 objects 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 objects 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.

For normalized similarity data, where a linear rescaling has been used to limit the similarity data to the interval $[0, 1]$, it is reasonable to restrict each of the weight parameters to the same interval. Accordingly, the second complexity term may finally be given as

$$\begin{aligned} \ln \int \sqrt{\det \mathbf{I}\left(\mathbf{w}\right)}.d\mathbf{w} &= \ln \int_0^1 \int_0^1 \cdots \int_0^1 \sqrt{\det \mathbf{I}\left(\mathbf{w}\right)}.dw_1.dw_2.dw_m \\ &= \ln \int_0^1 \int_0^1 \cdots \int_0^1 \sqrt{\det\left(\frac{1}{s^2}\mathbf{G}\right)}.dw_1.dw_2.dw_m \\ &= \ln \sqrt{\det\left(\frac{1}{s^2}\mathbf{G}\right)} \\ &= \ln \sqrt{\det \mathbf{G}} - \ln s. \end{aligned} \tag{8}$$

Combining these three terms gives a measure of the stochastic complexity of an additive clustering model, as follows:

$$\begin{aligned} \mathrm{SCC}_{\mathrm{adclus}} &= \frac{1}{2s^2}\sum_{i<j}\left(s_{ij} - \hat{s}^*_{ij}\right)^2 + \frac{n\left(n-1\right)}{2}\ln\left(s\sqrt{2\pi}\right) \\ &+ \frac{m}{2}\ln\left(\frac{n\left(n-1\right)}{4\pi}\right) + \ln\sqrt{\det \mathbf{G}} - \ln s \\ &= \frac{1}{2s^2}\sum_{i<j}\left(s_{ij} - \hat{s}^*_{ij}\right)^2 + \frac{m}{2}\ln\left(\frac{n\left(n-1\right)}{4\pi}\right) \\ &+ \ln\sqrt{\det \mathbf{G}} + \text{constant}. \end{aligned} \tag{9}$$

3. A Concrete Example

As a concrete example of the benefits of using the stochastic complexity criterion to constrain additive clustering models, consider a four-object domain with the normalized symmetric similarity matrix

$$\mathbf{S} = \begin{bmatrix} 1.0000 & 0.4981 & 0.4700 & 0.5402 \\ 0.4981 & 1.0000 & 0.4325 & 0.4044 \\ 0.4700 & 0.4325 & 1.0000 & 0.9789 \\ 0.5402 & 0.4044 & 0.9789 & 1.0000 \end{bmatrix}.$$

These similarity values were chosen because, for the case of a two-cluster model, there are two different cluster structures that provide the same maximal goodness-of-fit. These two models are shown in Figure 1. The model on the left includes a universal cluster with weight 0.4618, a cluster containing the first and second objects with weight 0.0363, and a cluster containing the third and fourth objects with weight 0.5171. The model on the right includes a universal cluster with weight 0.4512, a cluster containing the first, second, and fourth objects with weight 0.0296, and a cluster containing the third and fourth objects with weight 0.5277.

Both these models explain 92.92% of the variance in the similarity data, using the same number of clusters. Because their goodness-of-fit and parametric complexity are the same, these models would be equivalent under a complexity measure such as the AIC or BIC. The SCC, however, through its sensitivity to functional form complexity, is able to distinguish the models. The second complexity term in the SCC for the model on the left is

$$\ln \left(\det \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 6 \end{bmatrix} \right)^{\frac{1}{2}} = \ln 2,$$

whereas for the model on the right it is

$$\ln \left(\det \begin{bmatrix} 3 & 0 & 3 \\ 0 & 1 & 1 \\ 3 & 1 & 6 \end{bmatrix} \right)^{\frac{1}{2}} = \ln \sqrt{6},$$

which means that the model on the left is to be preferred. This result accords with intuition, since it prefers an account of the data that is a straightforward

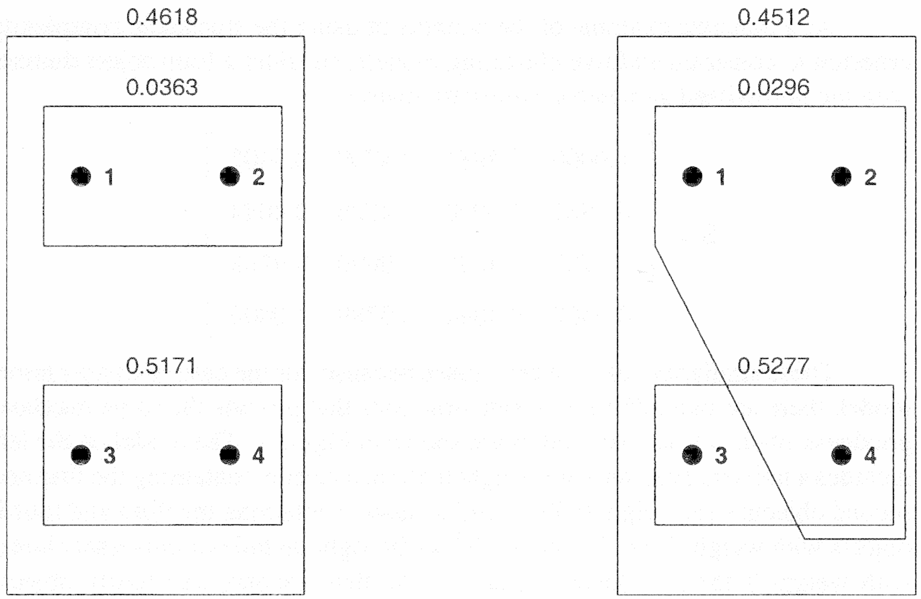


Figure 1: Two additive clustering models that provide equally good goodness-of-fit using the same number of clusters, but have different stochastic complexity. The stochastic complexity criterion measures the model on the left as being simpler than the one on the right.

hierarchical decomposition, rather than a less easily interpreted overlapping cluster structure. A more detailed analysis of the nature of the matrix **G**, and its implications for cluster structure complexity, may be found in Lee (2001b).

4. Fitting Algorithm

The stochastic complexity measure provides an opportunity to generate additive clustering models that consider both goodness-of-fit and complexity, without requiring the number of clusters to be prespecified. The fitting algorithm developed here 'grows' a series of additive clustering models by successively adding clusters, similarly to Mirkin (1987), and produces the model with the minimal stochastic complexity. The algorithm may be described according to the following five steps.

Step 1: The best-fitting '0-cluster' model is found and its stochastic complexity calculated. The '0-cluster' model is the degenerate additive clustering model that uses only the constant arising from a universal cluster (cf. Equation 1). The stochastic complexity measure (Equation 9) may be simplified for this one-parameter model to

$$\frac{1}{2s^2}\sum_{i<j}(s_{ij}-\bar{s})^2+\frac{1}{2}\ln\left(\frac{n\,(n-1)}{4\pi}\right)+\ln\sqrt{\frac{n\,(n-1)}{2}}, \tag{10}$$

where $\bar{s}$ is the arithmetic mean of the similarity values. By construction, this model always explains 0% of the variance in the similarity data.

Step 2: A new cluster is created, and the objects it initially includes are determined by a 'seeding' heuristic. The heuristic is based on ideas used in the ADDI-S algorithm for additive clustering developed by Mirkin (1987). When a new cluster is added, the similarity not already modeled by the preceding clusters is calculated. For the ith and jth objects, this residual similarity is given by

$$s_{ij}^{r}=\max\left(s_{ij}-\sum_{k}w_k f_{ik} f_{jk}, 0\right). \tag{11}$$

The two objects with the greatest residual similarity are included in the new cluster. The seeding heuristic then continues to add the object not in the new cluster that has the greatest average residual similarity to those objects already in the cluster, providing this average residual similarity measure is more than half the average within-cluster residual similarity.

Step 3: The algorithm then attempts to optimize the seeded model, searching for a cluster structure and set of weights that have minimal stochastic complexity. The combinatorial part of this optimization, which involves adjusting the patterns of discrete cluster assignment, is based on stochastic hill-climbing, so that the algorithm starts by constructing a random ordering of all the binary membership variables in the seeded representation. By following this ordering, the membership variables are changed either to remove an object from a cluster, or to include an object in a cluster. For the new cluster structure, best-fitting weights are found as the solution to a non-negative least-squares problem:

$$\mathbf{w}^{*}=\arg\min_{\mathbf{w}}\sum_{i<j}\left(s_{ij}-\sum_{k}w_k f_{ik} f_{jk}\right)^{2} \quad \text{subject to } \mathbf{w}\geq\mathbf{0}, \tag{12}$$

using the method described by Lawson and Hanson (1974, pp. 160-165). The stochastic complexity of the new model is then calculated according to Equation 9. If the stochastic complexity decreases as a result of the change of cluster membership, the change is accepted, and a new random ordering for all of the binary membership variables is constructed. If the stochastic complexity does not decrease, however, the change is rejected, and the effects of changing the next variable in the random sequence are considered. Step 3 continues iteratively until an entire random sequence of variable changes has been exhausted, meaning that the model has (locally) minimal stochastic complexity, before proceeding to Step 4.

Step 4: If the stochastic complexity of the current model is within an *evidence* parameter (Lee 2001b, pp. 137-138) of the lowest stochastic complexity recorded for any model the algorithm returns to Step 2 to continue adding clusters; otherwise the algorithm terminates by continuing to Step 5.

Step 5: The additive clustering model with the lowest stochastic complexity is returned, using the matrix of cluster membership variables and a corresponding vector giving the best-fitting weights for this cluster structure.

5. Monte Carlo Evaluation

To evaluate the effectiveness of the proposed algorithm, Monte Carlo simulations were undertaken examining its ability to recover known cluster structures and weights in the presence of noise. Rather than generating random cluster structures, a number of manually specified cluster structures were used. This approach was taken partly because it is important that the recovery properties of the algorithm are tested for cluster structures that simultaneously involve partitioning, hierarchical and overlapping clusters, and this goal is most easily achieved by specifying appropriately challenging structures. More fundamentally, it is possible for similarity data derived from randomly generated additive clustering models to be equally well accommodated by different clusters and weights, making problematic the assessment of whether an algorithm recovers the 'correct' model.

An example of the cluster structures and weights used in the Monte Carlo evaluation is shown in Figure 2, This four-cluster structure involves seven stimuli and incorporates partitioning, hierarchical, and overlapping clusters with different weights. The similarity data generated from this model were corrupted by the addition of zero-mean Gaussian noise with standard deviation of 0.10.

The proposed additive clustering algorithm was then applied to the data, assuming the value $s = 0.10$, and using Akaike's Information Criterion, the Bayesian Information Criterion, and the Stochastic Complexity Criterion to control the complexity of the derived cluster structures. The results of this analysis, across 50 runs of each algorithm, are summarized in Figure 3. The mean pattern of change in the percentage of variance explained and various complexity measures are shown as the number of clusters increases, together with best- and worst-case performance bounds.

Figure 3 shows that the AIC tends to recover the wrong cluster structure,

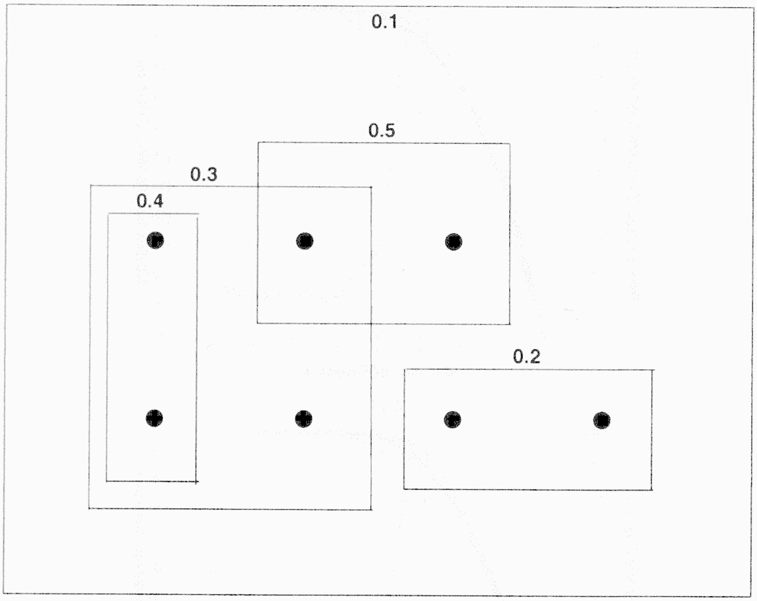
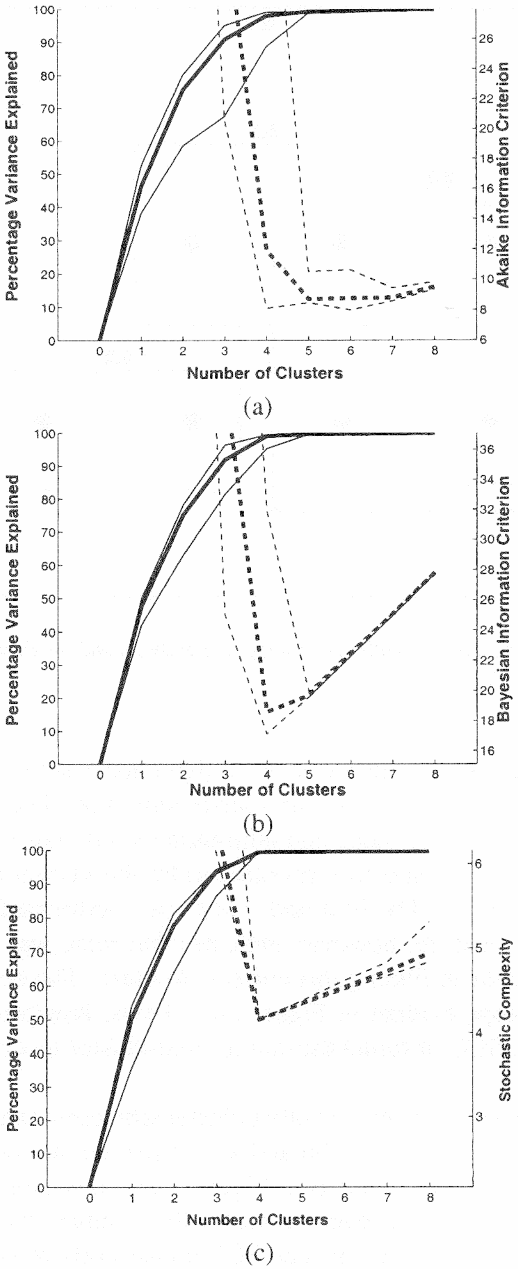


Figure 2: The 5-cluster structure, with partitioning, hierarchical and overlapping clusters used in the reported Monte Carlo study.

with its mean value being minimized for a five-cluster structure, and the lowest value found corresponding to a six-cluster structure. The BIC, in contrast, has both its minimum mean value and minimum overall value for a four-cluster structure, which visual inspection revealed to be the known generating structure shown in Figure 2. The best-and worst-case performance envelope surrounding the BIC measure, however, indicates that there was some significant variability across runs in finding this correct structure. The convergence of the performance envelope evident in Figure 3(c) shows, however, that, when the algorithm used the SCC, it found the correct four-cluster model on all of the 50 runs.

In the interests of brevity, the other cluster structures examined, and other levels of noise considered ($s = 0.05$ and $s = 0.15$), are not presented in detail here. The pattern of results shown in Figure 3 are, however, typical of those obtained. Generally, the AIC over-estimated the number of clusters in the recovered model, while both the BIC and SCC consistently recovered the correct models. One the basis of these results, it may well be the case that the BIC can usefully be used with the proposed algorithm in many cases. This would have some practical advantages in terms of computational efficiency, since the BIC is easier to calculate than the SCC. The temptation to adopt this short-cut should, however, be tempered with an understanding of the theoretical advantages of



(a)

(b)

(c)

Figure 3: Pattern of change in the percentage of variance explained (solid lines, left axis), and complexity measure (broken lines, right axis), when applying the additive clustering algorithm to the artificially generated similarity data, using (a) the AIC, (b) the BIC, and (c) the SCC. For both the variance explained and complexity measures, the mean performance across 50 runs is shown in bold, and the envelope of best- and worse-case performance is shown by the surrounding lines.

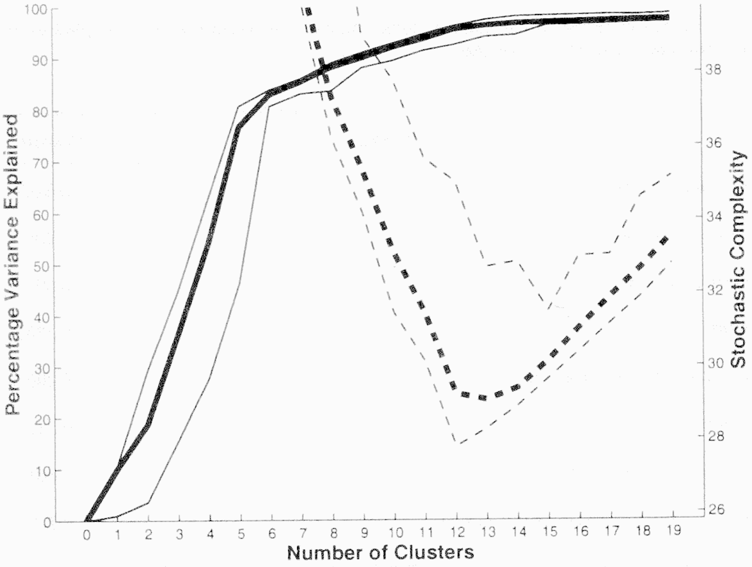


Figure 4: Pattern of change in the percentage of variance explained (solid lines, left axis), and stochastic complexity measure (broken lines, right axis), when applying the additive clustering algorithm to the kinship data. For both measures, the mean performance across 50 runs is shown in bold, and the envelope of best and worse case performance is shown by the surrounding lines.

the SCC in terms of measuring functional form complexity, as demonstrated by the concrete example presented earlier.

6. Illustrative Application

As an illustrative application of the algorithm to real data, consider the similarities collected by Rosenberg and Kim (1975), as published in Arabie, Carroll and DeSarbo (1987, pp. 62–63) involving 15 common kinship terms, such as 'father', 'cousin', and 'grandmother'. These data were collected using a sorting procedure performed by six groups of 85 subjects, where each kinship term was placed into one of a number of groups, under various conditions of instructions to the subjects.

For similarity data obtained by averaging across a number of data sources in this way, Lee (2001a) proposes a straightforward means of estimating data precision. In general, given a set of similarity matrices $\mathbf{S}^k = [s^k_{ij}]$ provided by $k = 1, 2, \ldots, K$ data sources, the precision of the averaged similarity matrix $\mathbf{S} = \frac{1}{K}[\sum_k s^k_{ij}] = [s_{ij}]$ may be estimated as the average of the sample standard deviations for each of the pooled cells in the final matrix, as follows:

Table 1: The 12-cluster model of the kinship data, showing the kinship terms included in each of the clusters, and the associated cluster weights

Stimuli In Cluster	Weight
brother sister	0.320
father mother	0.305
daughter son	0.304
aunt uncle	0.270
grandfather grandmother	0.269
nephew niece	0.266
granddaughter grandson	0.263
aunt cousin nephew niece uncle	0.225
aunt daughter granddaughter grandmother mother niece sister	0.225
brother father grandfather grandson nephew son uncle	0.224
granddaughter grandfather grandmother grandson	0.209
brother daughter father mother sister son	0.168
additive constant	0.237
Variance Explained	96.2%

$$s = \frac{1}{n(n-1)/2} \sum_{i<j} \sqrt{\frac{\sum_k \left(s_{ij}^k - s_{ij}\right)^2}{K-1}}. \tag{13}$$

This approach was applied to the six data sources available for the kinship data by arithmetically averaging the six individual similarity matrices to form a single averaged matrix, and using the sample standard deviations of the averaged cells to estimate the precision of this averaged matrix as $s = 0.091$.

Figure 4 summarizes the results of 50 independent applications the algorithm to the kinship data, at this level of data precision, using an evidence parameter value of 6. As before, the increase in the percentage of variance accounted for by successive models as clusters are added is shown by the solid lines, against the left axis. The pattern of change in the stochastic complexity measure is also shown, using a broken line against the right axis. For both measures, the performance 'envelope' across the 50 runs is given by the surrounding lines, which correspond to best and worse case performances.

Looking at the best-case performance in Figure 4, the stochastic complexity measure is minimized for a 12-cluster model. This model is detailed in Table 1, listing the kinship terms included in each cluster, the weights for each of the clusters, and the value of the additive constant. Each of the clusters is amenable to substantive interpretation, capturing classes within the domain such as 'nuclear family', 'extended family', 'grandparents', 'siblings', 'parents', 'females' and 'males'. This example also highlights the representational flexibility of additive clustering models. The use of arbitrarily overlapping clusters is necessary, for example, to place the kinship term 'brother' within the clusters corresponding to the classes 'siblings', 'nuclear family' and 'male'. By avoiding the constraints of partitioning or hierarchical clustering models, additive clustering allows several contrasting contexts (e.g., 'generation' and 'gender') to be captured simultaneously.

It is interesting to compare this particular solution with that produced by the SINDCLUS algorithm described by Chaturvedi and Carroll (1994), and the SEFIT algorithm described by Mirkin (1996). Chaturvedi and Carroll (1994) report the same five-cluster solution that was found on 41 out of the 50 runs of the current algorithm. This earlier solution contains the following subset of the clusters listed in Table 1: {granddaughter, grandfather, grandmother, grandson}, {aunt, cousin, nephew, niece, uncle }, {brother, daughter, father, mother, sister, son}, {aunt, daughter, granddaughter, grandmother, mother, niece, sister}, and {brother, father, grandfather, grandson, nephew, son, uncle}. These are basically the five 'large' clusters that are subsequently decomposed by the 12-cluster model. The solution reported by Mirkin (1996, Table 4.9) also contains these clusters, but appends the clusters {brother, sister}, {aunt, uncle}, and {nephew, niece}. The stochastic complexity measure on which the current algorithm is based suggests that, at the estimated level of data precision, the extra clusters present in the 12-cluster solution are warranted. The fact that these clusters are readily amenable to substantive interpretation, capturing important concepts within the kinship domain such as 'parent' and 'sibling', is encouraging.

7. Conclusion

A measure of the stochastic complexity of additive clustering models has been developed. Unlike such measures as the AIC or BIC, this measure takes into account goodness-of-fit, the number of parameters, and the way in which these parameters interact within the model. This measure is thus sensitive not only to the number of parameters used by an additive clustering model to achieve a level of goodness-of-fit, but also to the patterns of cluster inclusion and overlap that are used.

An algorithm for generating additive clustering models with minimal stochastic complexity was also developed. This algorithm starts with a one-cluster model, and successively 'grows' by adding new clusters, until the stochastic complexity measure starts to increase. Monte Carlo simulations showed that the algorithm is able to recover known cluster structures and weights in the presence of Gaussian noise. The algorithm was also applied to a standard similarity data set involving kinship terms, where it was shown to generate a meaningful domain representation that extends upon those achieved by other additive clustering approaches. More generally, through its use of the stochastic complexity measure, the algorithm has the ability to generate features from similarity data that balance the competing demands of accuracy and simplicity.

References

AKAIKE, H. (1974), "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control 19*, 716–723.

ARABIE, P., and CARROLL, J.D. (1980), "MAPCLUS: A Mathematical Programming Approach to Fitting the ADCLUS Model," *Psychometrika 45*(2), 211–235.

ARABIE, P., CARROLL, J.D., and DESARBO, W.S. (1987), *Three-Way Scaling and Clustering*, Newbury Park, CA: Sage.

ARABIE, P., and SHEPARD, R.N. (1973), "Representation of Similarities as Combinations of Discrete, Overlapping Properties," Paper presented at Mathematical Psychological Meeting, Montréal.

CARROLL, J.D., and CHATURVEDI, A. (1995), "A General Approach to Clustering and Multidimensional Scaling of Two-Way, Three-Way, or Higher-Way Data," In *Geometric Representations of Perceptual Phenomena,* Eds., R.D. Luce, M. D'zmura, D.D. Hoffman, G. Iverson, and A.K. Romney, Mahwah, NJ: Erlbaum, pp. 295-318.

CHATURVEDI, A., and CARROLL, J.D. (1994), "An Alternating Combinatorial Optimization Approach to Fitting the INDCLUS and Generalized INDCLUS Models," *Journal of Classification 11*, 155-170.

CHATURVEDI, A., GREEN, P.E., and CARROLL, J.D. (2001), "K-modes Clustering," *Journal of Classification 18*, 35-56.

DESARBO, W.S. (1982), "GENNCLUS: New Models for General Nonhierarchical Cluster Analysis", *Psychometrika 47*, 449-475.

KASS, R.E., and RAFTERY, A.E. (1995), "Bayes Factors," *Journal of the American Statistical Association 90*(430), 773-795.

LAWSON, C.L., and HANSON, R.J. (1974), *Solving Least Squares Problems*, Englewood Cliffs, NJ: Prentice-Hall.

LEE, M.D. (1999a), "An Extraction and Regularization Approach to Additive Clustering," *Journal of Classification 16*(2), 255-281.

LEE, M.D. (1999b). "On the Complexity of Multidimensional Scaling, Additive Tree, and Additive Clustering Representations," Paper presented at the Symposium on Model Complexity held at the 32nd Annual Meeting of the Society for Mathematical Psychology, Santa Cruz, CA, July/August 1999.

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.

LEE, M.D. (submitted). "Avoiding the Dangers of Averaging Across Subjects when using Multidimensional Scaling," Manuscript submitted for publication.

LEE, M.D. (in press), "A Simple Method for Generating Additive Clustering Models with Limited Complexity," *Machine Learning*.

MIRKIN, B.G. (1987), "Additive Clustering and Qualitative Factor Analysis Methods for Similarity Matrices," *Journal of Classification 4,* 7-31.

MIRKIN, B.G. (1996), *Mathematical Classification and Clustering*, Boston, MA: Kluwer.

MYUNG, I.J., and PITT, M.A. (1997), "Applying Occam's Razor in Modeling Cognition: A Bayesian Approach," *Psychonomic Bulletin & Review 4*(1), 79-95.

RISSANEN, J.J. (1996), "Fisher Information and Stochastic Complexity," *IEEE Transactions on Information Theory 42*(1), 40-47.

ROSENBERG, S., and 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., and ARABIE, P. (1979), "Additive Clustering Representations of Similarities as Combinations of Discrete Overlapping Properties," *Psychological Review 86*(2), 87-123.

TENENBAUM, J.B. (1996), "Learning the Structure of Similarity," In *Advances in Neural Information Processing Systems, Volume 8*, Eds., D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, Cambridge, MA: MIT Press, 3-9.