2 Minimum Description Length and Psychological Clustering Models

Michael D. Lee
Department of Psychology
University of Adelaide
South Australia 5005
Australia
michael.lee@adelaide.edu.au
http://www.psychology.adelaide.edu.au/members/staff/michaellee/homepage

Daniel J. Navarro
Department of Psychology
Ohio State University
1885 Neil Avenue
Columbus, Ohio 43210
USA
navarro.20@osu.edu
http://quantrm2.psy.ohio-state.edu/navarro/

Clustering is one of the most basic and useful methods of data analysis. This chapter describes a number of powerful clustering models, developed in psychology, for representing objects using data that measure the similarities between pairs of objects. These models place few restrictions on how objects are assigned to clusters, and allow for very general measures of the similarities between objects and clusters. Geometric Complexity Criteria (GCC) are derived for these models, and are used to fit the models to similarity data in a way that balances goodness-of-fit with complexity. Complexity analyses, based on the GCC, are presented for the two most widely used psychological clustering models, known as "additive clustering" and "additive trees".

2.1 Introduction

Clustering is one of the most basic and useful methods of data analysis. It involves treating groups of objects as if they were the same, and describing how the groups relate to one another. Clustering summarizes and organizes data, provides a framework for understanding and interpreting the relationships between objects, and proposes a simple description of these relationships that has the potential to generalize to new or different situations. For these reasons, many different clustering models have been developed and used in fields ranging from computer science and statistics to marketing and psychology [see Arabie, Hubert, and De Soete 1996; Everitt 1993; Gordon 1999 for overviews].

Different clustering models can be characterized in terms of the different assumptions they make about the *representational structure* used to define clusters, and the *similarity measures* that describe the relationships between objects and clusters.

2.1.1 Representational Assumptions

Representationally, it is possible for different types of constraints to be imposed on how objects can be grouped into clusters. Three different assumptions are shown in Figure 2.1:

(a) The partitioning approach forces each object to be assigned to exactly one cluster. This approach can be interpreted as grouping the objects into equivalence classes, and essentially just summarizes the objects, without specifying how the clusters relate to each other. For example, if the objects 'A' though 'H' in Figure 2.1(a) correspond to people, the partitioning could be showing which of four different companies employs each person. The representation does not allow a person to work for more than one company, and does not convey information about how the companies themselves are related to each other.

(b) The hierarchical approach allows for nested clusters. This can be interpreted as defining a tree structure, where the objects correspond to terminal nodes. For example, the hierarchical clustering in Figure 2.1(b) could be showing not just the company employing each person, but also the division they work for within that company, and further subdivisions in the organizational structure. Each of these subdivisions corresponds to a branch in the tree, and the overall topology of the tree relates objects and clusters to one another.

(c) The overlapping approach imposes no representational restrictions, allowing any cluster to include any object and any object to belong to any cluster. Overlapping clustering models can be interpreted as assigning features to objects. For example, in Figure 2.1(c), the five clusters could correspond to features like the company a person works for, the division they work in, the football team they support, their nationality, and so on. It is possible for two people in different companies to support the same football team, or have the same nationality, or have any other pattern of

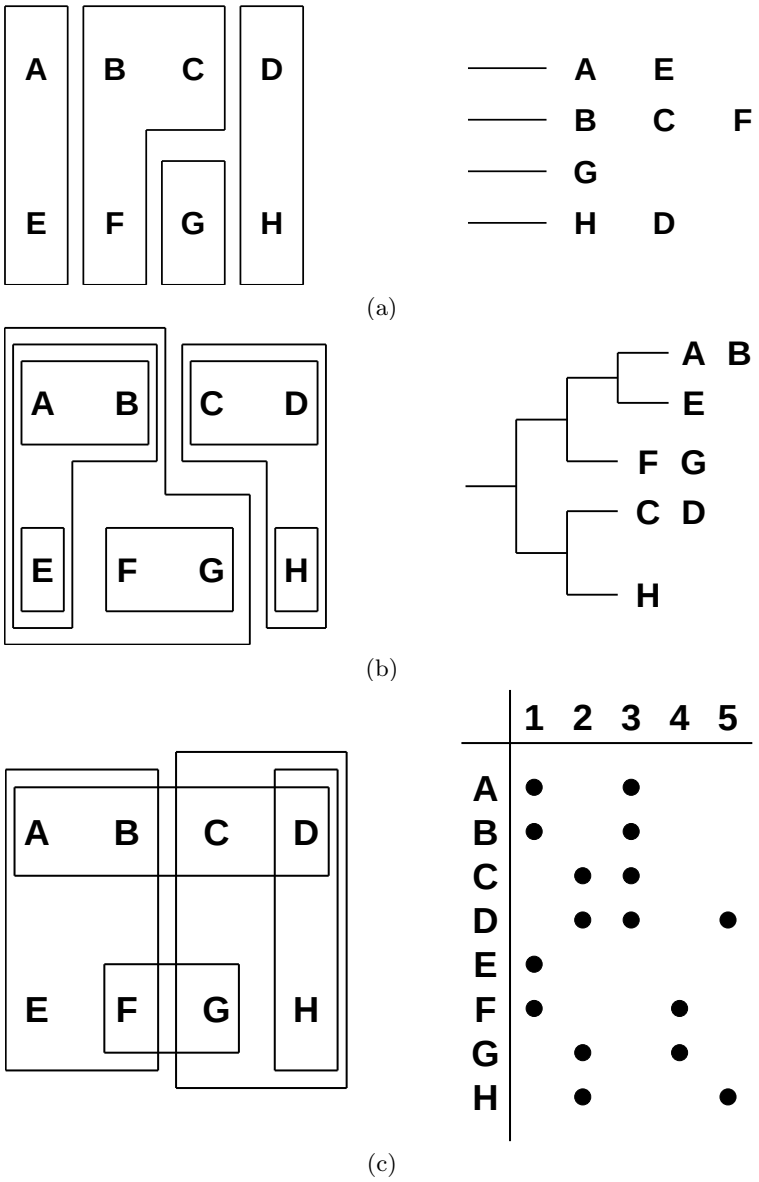


Figure 2.1 Three different representational assumptions for clustering models, showing (a) partitioning, (b) hierarchical, and (c) overlapping structures, and their interpretation as (a) equivalence classes, (b) tree structures, and (c) feature assignments.

shared features.

2.1.2 Similarity Assumptions

A clustering model also makes assumptions about how the similarity between objects is measured. One possibility, most compatible with partitioning representations, is to treat all objects in the same cluster as being equally similar to one another, and entirely different from objects not in that cluster. In hierarchical and overlapping representations, more detailed measures of similarity are possible. Because objects may belong to more than one cluster, various similarity measures can be constructed by considering the clusters objects have in common, and those that distinguish them, and combining these sources of similarity and dissimilarity in different ways.

2.1.3 Psychological Clustering Models

In many fields that use clustering models, most applications have relied on a relatively small range of the possible representational and similarity assumptions. Great emphasis is given to partitioning approaches like k-means clustering, and various tree-fitting approaches using hierarchical representations. Sometimes (although not always) this emphasis comes at the expense of overlapping representations, which have hierarchical and partitioning representations as special cases.

One field, perhaps surprisingly, that has a long tradition of using overlapping clustering models is psychology. In cognitive psychology, a major use of clustering models has been to develop accounts of human mental representations. This is usually done by applying a clustering model to data that describes the empirically observed similarities between objects, and then interpreting the derived clusters as the cognitive features used by people to represent the object. At least as early as Shepard and Arabie [1979, p. 91], it has been understood that "generally, the discrete psychological properties of objects overlap in arbitrary ways", and so representations more general than partitions or hierarchies needed to be used.

Psychological clustering models have also considered a variety of possible similarity processes. In particular, they have drawn a useful distinction between common and distinctive features [Tversky 1977]. Common features are those that make two objects with the feature more similar, but do not affect the similarities of objects that do not have the feature. For example, think of two people with an unusual characteristic like blue hair. Having this feature in common makes these two people much more similar to each other than they otherwise would be, but does not affect the similarities between other people being considered who have 'normal' hair colors. Distinctive features, on the other hand, are those that make objects both having and not having the feature more similar to each other. For example, whether a person is male or female is a distinctive feature. Knowing two people are male makes them more similar to each other, knowing two people are female makes them more similar to each other, and knowing one person is male while the other is fe-

male makes them less similar to each other. Using common and distinctive features allows clustering models to deal with two different kind of regularities: common features capture the idea of 'similarity within', whereas distinctive features captures the notion of 'difference between'. In addition, psychological clustering models usually associate a weight with every cluster, which can be interpreted as measuring its 'importance' or 'salience'. By combining the weights of common and distinctive features in various ways, a wide range of similarity assumptions is possible.

A consequence of considering clustering models with great flexibility in both their representations and similarity measures, however, is that it becomes critical to control for model complexity. As noted by Shepard and Arabie [1979, p. 98], an overlapping clustering model that is also able to manipulate the similarity measures it uses may be able to fit any similarity data perfectly. The possibility of developing overly-complicated clustering representations, of course, conflicts with the basic goals of modeling: the achievement of interpretability, explanatory insight, and the ability to generalize accurately beyond given information. In psychology, it is particularly important to control the complexity of cluster representations when they are used in models of cognitive processes like learning, categorization, and decision-making. Because the world is inherently dynamic, representations of the environment that are too detailed will become inaccurate over time, and provide a poor basis for decision-making and action. Rather, to cope with change, cognitive models need to have the robustness that comes from simplicitly. It is this need for simple representations that makes psychological clustering models ideal candidates for Minimum Description Length (MDL) methods.

2.1.4 Overview

This chapter describes the application of modern MDL techniques to a number of psychological clustering models. The next section provides a formal description of the clustering models considered, the common and distinctive models of similarity, and the form of the similarity data from which models are learned. Geometric Complexity Criteria [GCC: Balasubramanian 1997; Myung, Balasubramanian, and Pitt 2000] are then derived for the clustering models. As it turns out, these are equivalent to Rissanen's [1996] Fisher Information approximation to the Normalized Maximum Likelihood. With the GCC measures in place, two established psychological clustering models, known as "additive clustering" and "additive trees", are considered in some detail. Illustrative examples are given, together with analysis and simulation results that assess the complexity of these models. Finally, two new psychological clustering models are described that raise different challenges in measuring and understanding model complexity.

2.2 Formal Description of Clustering Models

2.2.1 Similarity Data

Psychological clustering models are learned from similarity data, in the form of a $n \times n$ similarity matrix $\mathbf{S} = [s_{ij}]$, where s_{ij} is the similarity between the ith and jth of n objects. Usually these data are normalized to lie in the interval $[0, 1]$, and often the assumption of symmetry is made so that $s_{ij} = s_{ji}$ for all i and j pairs. Similarities are usually based on empirical measures of human performance, including ratings scales, identification tasks, sorting or grouping procedures, and a range of other experimental methodologies. It is also possible to generate psychological similarity data theoretically, using quantitative descriptions of objects. There are, for example, many methods for measuring the semantic similarity of text documents [e.g., Damashek 1995; Griffiths and Steyvers 2002; Landauer and Dumais 1997; Lund and Burgess 1996], based on the words (or sequences of characters or words) they contain. The pairwise similarities between all of the documents in a corpus could be used as the data for learning a clustering representation.

However similarity data are generated, a standard assumption [e.g., Lee 2001; Tenenbaum 1996] is that the similarity between the ith and jth objects comes from a Gaussian distribution with mean s_{ij}, and that the Gaussian distribution for each pair has common variance σ^2. The variance quantifies the inherent precision of the data, and can be estimated based on an understanding of the process by which the data were generated. For example, most empirical methods of collecting similarity data generate repeated measures for the similarity between each pair of objects, by having more than one person do a task, or having the same person do a task more than once. Given a set of similarity matrices $\mathbf{S}^k = [s^k_{ij}]$ provided by $k = 1, 2, \ldots, K$ data sources, the variance of the arithmetically averaged similarity matrix $\mathbf{S} = \frac{1}{K}[\sum_k s^k_{ij}] = [s_{ij}]$ can be estimated as the average of the sample variances for each of the pooled cells in the final matrix.

2.2.2 Cluster Structures

A clustering model that uses m clusters for n objects is described by a $n \times m$ matrix $\mathbf{F} = [f_{ik}]$, where $f_{ik} = 1$ if the ith object is in the kth cluster, and $f_{ik} = 0$ if it is not. When the clusters are interpreted as features, the vector $\mathbf{f}_i = (f_{i1}, \ldots, f_{im})$ gives the featural representation of the ith object. Each cluster has an associated weight, w_k for the kth cluster, which is a positive number. Generally, the cluster structure $\mathbf{F}$ is treated as the model, and the clusters weights $\mathbf{w} = (w_1, \ldots, w_m)$ are treated as model parameters.

	A	B	C	D	E
A	--				
B	17	--			
C	17	17	--		
D	05	05	14	--	
E	05	05	05	05	--

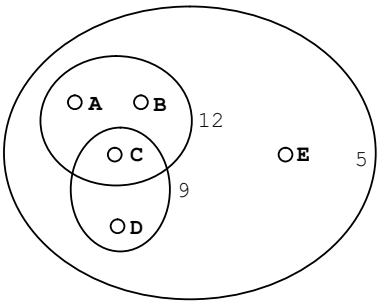


Figure 2.2 An example of an additive clustering representation and its associated similarity matrix.

2.2.3 Common Features Similarity

The common features similarity model assumes that two objects become more similar as they share more features in common, and that the extent to which similarity increases is determined by the weight of each common feature. This means that the modeled similarity between the ith and jth objects, denoted as $\hat{s}_{ij}$, is simply the sum of the weights of the common features:

$$\hat{s}_{ij} = c + \sum_k w_k f_{ik} f_{jk}. \tag{2.1}$$

The "additive constant" c in Eq. (2.1) increases the similarity of each pair of objects by the same amount, and so measures the degree to which all of the objects are similar to each other. It can be interpreted as the saliency weight of a 'universal' cluster containing all objects.

Combining overlapping clusters with common features similarity corresponds to what is known as the "additive clustering" model in psychology [e.g., Arabie and Carroll 1980; Chaturvedi and Carroll 1994; Lee 2002a; Mirkin 1987, 1996; Ruml 2001; Shepard 1980; Shepard and Arabie 1979; Tenenbaum 1996]. A simple example of an additive clustering model, and the similarity matrix on which it is based, is shown in Figure 2.2. Notice that the sums of the weights of the clusters shared by each pair of objects corresponds to their similarity in the matrix.

2.2.4 Distinctive Features Similarity

The distinctive features similarity model assumes that two stimuli become more dissimilar to the extent that one stimulus has a feature that the other does not. As with the common features approach, the extent to which similarity is decreased by a distinctive feature is determined by the weight of that feature. This model can be expressed as:

$$\hat{s}_{ij} = c - \sum_k w_k \left| f_{ik} - f_{jk} \right|. \tag{2.2}$$

	A	B	C	D	E
A	--				
B	15	--			
C	20	25	--		
D	18	23	06	--	
E	20	25	20	18	--

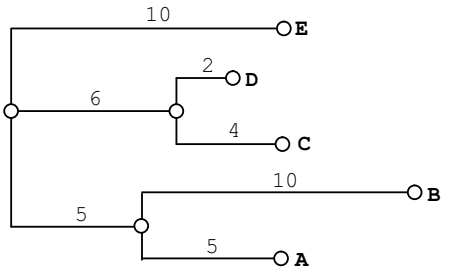


Figure 2.3 An example of an additive tree representation and its associated dissimilarity matrix.

For hierarchical representations, distinctive features similarity corresponds to what is known as the "additive tree" model in psychology [Corter 1996; Johnson and Tversky 1984; Sattath and Tversky 1977; Shepard 1980; Tversky and Hutchinson 1986]. These models are usually applied to *dis*similarity data, generated by reversing the scale of similarity measures. A simple example of an additive tree model, and the dissimilarity matrix on which it is based, is shown in Figure 2.3. The model has an additive constant of 30 and seven clusters: one for each of the objects 'A' to 'E', with weights 5, 10, 4, 2, and 10 respectively; one for the pair of objects 'A' and 'B', with weight 5; and one for the pair of objects 'C' and 'D', with weight 6. Each of these clusters corresponds to a node in the tree, and represents a feature that distinguishes between all of the objects that lie under the different branches coming from that node. Accordingly, the weights of the clusters can be interpreted as the length of the edges between nodes. This means that, in Figure 2.3 the length of the unique path between each pair of objects corresponds to their dissimilarity in the matrix.

For overlapping representations, distinctive features similarity corresponds to a discrete version of what is known as the "multidimensional scaling" model in psychology. Multidimensional scaling models [e.g., Cox and Cox 1994; Shepard 1962; Kruskal 1964] represent objects as points in a multidimensional space, so that the distance between the points corresponds to the dissimilarity between the objects. Discrete multidimensional scaling [e.g., Clouse and Cottrell 1996; Lee 1998; Rohde 2002] restricts the points to binary values, and so most of the distance metrics commonly used in the continuous version (i.e., Minkowskian metrics) reduce to the distinctive features model.

2.3 Geometric Complexity of Clustering Models

Traditionally, the complexity of clustering models in psychology has been dealt with in incomplete or heuristic ways. Most often [e.g., Arabie and Carroll 1980; Chaturvedi and Carroll 1994; DeSarbo 1982; Shepard and Arabie 1979; Tenenbaum 1996], the approach has been to find cluster structures that maximize a goodness-

of-fit measure using a fixed number of clusters. More recently [Lee 2001, 2002b], the Bayesian Information Criterion [Schwarz 1978] has been applied, so that the number of clusters does not need to be pre-determined, but the appropriate number can be found according to the goodness-of-fit achieved and the precision of the data. Both of these approaches, however, have the weakness of equating model complexity with only the number of clusters.

In general, both the representational and similarity assumptions made by a clustering model contribute to its complexity. Moving from partitions to hierarchies to overlapping clusters leads to progressively more complicated models, able to explain a progressively larger range of data. Controlling for this complexity requires more than counting the number of clusters, and needs to be sensitive to measures like the number of objects in the clusters, and the patterns of overlap or nesting between clusters. Different similarity assumptions control how the weight parameters interact, and so also affect model complexity. In addition, the complexities associated with representational and similarity assumptions will generally not be independent of one another, but will interact to create the overall complexity of the clustering model. For these reasons, it is important that psychological clustering models are evaluated against data using criteria that are sensitive to the full range of influences on model complexity.

The goal of psychological clustering is to find the best representation of empirical similarity data. The defining part of a representation is the cluster structure $\mathbf{F}$, which encodes fixed assumptions about the representational regularities in a stimulus environment. Unlike these core assumptions, the saliency weights $\mathbf{w}$ and constant c are parameters of a particular representation, which are allowed to vary freely so that the representational model can be tuned to the data. In general, finding the best parameter values for a given set of clusters is straightforward. The difficulty is finding the best set of clusters. This involves the theoretical challenge of developing criteria for comparing different cluster representations, and the practical challenge of developing combinatorial optimization algorithms for finding the best cluster representations using these criteria .

This chapter relies on the Geometric Complexity Criterion [GCC: Myung, Balasubramanian, and Pitt 2000; see also Pitt, Myung, and Zhang 2002] for model evaluation. In the GCC goodness-of-fit is measured by the maximum log likelihood of the model, $\ln p\left(D \mid \theta^*\right)$, where $p\left(\cdot\right)$ is the likelihood function, D is a data sample of size N, and θ is a vector of the k model parameters which take their maximum likelihood values at θ^*. The complexity of the model is measured in terms of the number of distinguishable data distributions that the model indexes through parametric variation. The geometric approach developed by Myung, Balasubramanian and Pitt [2000] leads to the following four term expression:

$$\mathrm{GCC} = -\ln p\left(D \mid \theta^*\right) + \frac{k}{2}\ln\left(\frac{N}{2\pi}\right) + \ln\int d\theta\sqrt{\det \mathbf{I}\left(\theta\right)} + \frac{1}{2}\ln\left(\frac{\det \mathbf{J}\left(\theta^*\right)}{\det \mathbf{I}\left(\theta^*\right)}\right),$$

where

$$\mathbf{I}_{ij}(\theta) = -E_\theta\left[\frac{\partial^2 \ln p(D \mid \theta)}{\partial\theta_i \partial\theta_j}\right]$$

is the Fisher Information Matrix of the model parameters, and

$$\mathbf{J}_{ij}(\theta^*) = -\left[\frac{\partial^2 \ln p(D \mid \theta)}{\partial\theta_i \partial\theta_j}\right]_{\theta=\theta^*}$$

is the covariance matrix of the model parameters at their maximum likelihood values.

Under the assumption that the similarities follow Gaussian distributions, with common variance estimated by $\hat{\sigma}^2$, the probability of similarity data $\mathbf{S}$ arising for a particular featural representation $\mathbf{F}$, using a particular weight parameterization $\mathbf{w}$, is given by

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

and so the log-likelihood takes is the sum of squared difference between the empirical data and model predictions, as scaled by the estimated precision of the data. The first term of the GCC, which measures data-fit, is simply the maximum of this log-likelihood, corresponding to the maximum likelihood modeled similarities $\hat{s}^*_{ij}$, as follows:

$$-\ln p(\mathbf{S} \mid \mathbf{F}, \mathbf{w}^*) = \frac{1}{2\hat{\sigma}^2}\sum_{i<j}(s_{ij} - \hat{s}^*_{ij})^2 + \text{constant}. \tag{2.3}$$

The second term of the GCC for a model with m clusters is found by noting that it uses $m+1$ parameters (including the additive constant), and that an $n \times n$ similarity matrix contains $n(n-1)/2$ observations, giving

$$\frac{m+1}{2}\ln\left(\frac{n(n-1)}{4\pi}\right). \tag{2.4}$$

For the common and distinctive similarity models given in Eqs. (2.1) and (2.2), the calculation of the second-order partial derivatives

$$\frac{\partial^2 \ln p(\mathbf{S} \mid \mathbf{F}, \mathbf{w})}{\partial w_x \partial w_y}$$

is straightforward, and allows the Fisher Information Matrix $\mathbf{I}(\mathbf{w})$ and the covari-

ance matrix $\mathbf{J}(\mathbf{w})$ to be specified. As it turns out, these two matrices are identical for all of the clustering models considered here, and so the fourth term of the GCC vanishes. This makes the GCC identical to Rissanen's [1996] asymptotic approximation to the Normalized Maximum Likelihood [see Grünwald this volume].

In fact, the two matrices $\mathbf{I}(\mathbf{w})$ and $\mathbf{J}(\mathbf{w})$ assume a constant value that is independent of the weight parameters, and is determined entirely by $\mathbf{F}$, which also simplifies the third term of the GCC. This constant value is conveniently written as the determinant of an $(m+1)\times(m+1)$ "complexity matrix", $\mathbf{G} = [g_{xy}]$, defined as

$$g_{xy} = \sum_{i<j} e_{ijx} e_{ijy},$$

where

$$e_{ijk} = \begin{cases} f_{ik} f_{jk} & \text{for common features,} \\ |f_{ik} - f_{jk}| & \text{for distinctive features.} \end{cases}$$

Using the complexity matrix, and assuming that, since the similarity values are normalized, the weight parameters range over the interval $[0, 1]$, the third term of the GCC is given by:

$$\begin{aligned} \ln \int d\mathbf{w} \sqrt{\det \mathbf{I}(\mathbf{w})} &= \ln \int_0^1 \int_0^1 \ldots \int_0^1 \sqrt{\det\left(\frac{1}{\hat{\sigma}^2}\mathbf{G}\right)}.dw_1.dw_2 \ldots dw_{m+1} \\ &= \frac{1}{2} \ln \det \mathbf{G} - \frac{m+1}{2} \ln \hat{\sigma}^2. \end{aligned} \tag{2.5}$$

Putting together the results in Eqs. (2.3), (2.4) and (2.5), the GCC for the clustering models is given as

$$\text{GCC} = \frac{1}{2\hat{\sigma}^2} \sum_{i<j} (s_{ij} - \hat{s}^*_{ij})^2 + \frac{m+1}{2} \ln\left(\frac{n(n-1)}{4\pi\hat{\sigma}^2}\right) + \frac{1}{2} \ln \det \mathbf{G} + \text{constant}.$$

Strictly speaking, the GCC requires a number of regularity conditions are met. However, Takeuchi [this volume] shows that the asymptotic GCC approximation of the Normalized Maximum Likelihood holds under a wide variety of conditions, and for a wide variety of models. While we have not checked all of the conditions, the most important ones (including positive definiteness of the Fisher Information Matrix) certainly hold.

2.4 Established Psychological Clustering Models

Additive clustering and additive trees are by far the most commonly used clustering models in psychology. In this section, illustrative examples of these models are provided demonstrating them being fit to similarity data using the GCC, together with analysis and simulation results based on their complexity matrices.

2.4.1 Additive Clustering

2.4.1.1 Illustrative Example

Lee and Navarro [2002] considered the similarities between nine colored shapes that combined the colors red, green and blue with the shapes circle, square and triangle. Twenty subjects rated the similarity of all 36 possible object pairs, presented in a random order, on a five point scale. The final similarity matrix was arithmetically averaged across subjects, and made symmetric by transpose averaging.

Figure 2.4 shows the additive clustering representation of these data corresponding to the minimum GCC value, as found using a stochastic hill-climbing optimization algorithm [Lee 2002a]. This model explains 99.3% of the variance in the data, and each of the clusters is readily interpreted as a color or shape. Interestingly, the weights of the clusters suggest that people assigned relatively greater emphasis to common color than common shape when judging similarity. The representation also highlights the need for overlapping clusters, so that the orthogonal color and shape characteristics of the objects can both be accommodated.

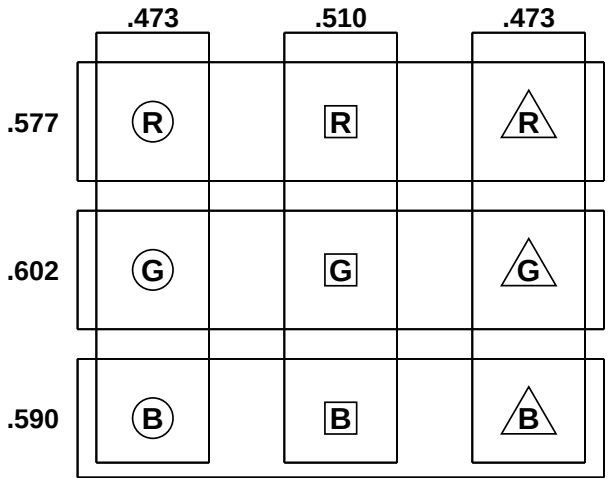


Figure 2.4 Overlapping common features representation, including cluster weights, of the colored shapes.

2.4.1.2 Interpretation of Complexity Matrix

The complexity matrix for additive clustering models is

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

The diagonal elements, $\sum_{i<j} f_{ik}f_{jk}$, count the number of object pairs in the kth cluster, and so measure cluster size. The off-diagonal elements, $\sum_{i<j} f_{ix}f_{jx}f_{iy}f_{jy}$, count the number of object pairs that are in both the xth and yth clusters, and so measure the overlap between clusters.

To make these ideas concrete, observe that the complexity matrix for the representation of the colored shapes in Figure 2.4 is

$$\mathbf{G} = \begin{bmatrix} 3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3 \end{bmatrix}.$$

Because each cluster has three objects, and hence three pairs of objects, all of the diagonal elements are three. Because each pair of clusters either has no overlap, or has one object in common, no pair of clusters share a pair of objects, and so all of the off-diagonal elements are zero.

2.4.1.3 Partitions

It is possible to show that, in general, $\mathbf{G}$ will be positive definite [Lee 2001, pp. 142-143]. This allows Hadamard's inequality [e.g., Bellman 1970, pp. 129-130] to be applied, so that the determinant is less than or equal to the product of the main diagonal,

$$\det \mathbf{G} \leq \prod_k g_{kk} = \prod_k \sum_{i<j} f_{ik}f_{jk},$$

with equality occurring when all off-diagonal elements are zero. This suggests that partitions, which have diagonal complexity matrices, are complicated cluster structures. There are, however, two important caveats to be placed on the generality of this result [Navarro 2003]. First, while being a partition is sufficient for a diagonal complexity matrix, it is not necessary. Since the counts in $\mathbf{G}$ are of object pairs,

clusters that have only one object in common also produce zero off-diagonal entries. The complexity matrix for the colored shapes in Figure 2.4 is a good example of this. Secondly, Hadamard's inequality requires that the product of the main diagonal elements remains constant, and so can only be used to compare cluster structures where the number of object pairs, and hence the number of objects, in each cluster is the same.

For partitions, or other cluster structures with diagonal complexity matrices, the determinant is simply the product of the diagonal elements, and so the number of objects in clusters determines model complexity. In particular, complexity is decreased by removing an object from a cluster, or by moving an object from a smaller cluster to a larger cluster.

Both of these results still hold when the universal cluster corresponding to the additive constant is included. This can be demonstrated by considering the complexity matrix $\mathbf{G}^+$ obtained when incorporating the universal cluster, which is

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

where $z = n\,(n-1)\,/2$ is the total number of object pairs, and $\mathbf{y}$ is a vector of the diagonal elements in $\mathbf{G}$. A standard result [e.g., Magnus and Neudecker 1988, p. 23] is that the determinant of this augmented complexity matrix can be written as

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

and it turns out [Lee 2001, pp. 144-145] that removing objects from clusters, or moving them from smaller to larger clusters, continues to increase complexity.

Interestingly, the reduction in complexity achieved by making clusters different sizes has a natural interpretation in terms of Shannon's [1948] Noiseless Coding Theorem. This theorem shows that the minimum average message length needed to convey a structure is approximately given by the entropy of that structure [Li and Vitányi 1993, p. 71]. From this perspective, a partition where each cluster has the same number of objects is more complicated because each cluster is equally likely, maximizing the entropy of the representation and its message length.

2.4.1.4 Nested Clusters

A two cluster model has complexity matrix

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

where $a \geq c$ and $b \leq c$. Since $\det \mathbf{G} = ac - b^2$ is minimized when $b = c$, the simplest possible two cluster model is a strictly nested one. Lee [2001] follows this

observation with an intuitive argument that, given a strictly nested cluster structure with i clusters, the increase in complexity from adding the $(i+1)$th cluster is minimized by making it also strictly nested. Together, these two arguments lead to the induction that strictly nested cluster structures are maximally simple additive clustering models.

Given a strictly nested cluster structure, the elementary row operation

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

shows that $\det \mathbf{G} = (-1)^{m+1}(b-a)(c-b)\ldots x$. Since a strictly nested model is restricted to having $a > b > c > \ldots > x$, this means that the complexity of nested representation is minimized by having each successive cluster encompass one fewer object pairs than its predecessor.

2.4.1.5 General Cluster Structures

For general cluster structures, Hadamard's inequality suggests two ways of reducing model complexity. The first is to minimize the number of objects in clusters, since this minimizes the diagonal elements whose product determines the upper bound on complexity. The second is to introduce overlap between the clusters, since this creates non-zero off-diagonal elements. In general, these two strategies conflict with one another, since increasing the overlap between clusters is often best achieved by increasing their size, and reducing cluster size will often come at the expense of reducing overlap.

Navarro [2002] reported the results of a simulation study designed to explore of how cluster size and overlap interact to determine complexity. This study used sample of 10^5 randomly generated cluster structures with ten objects and six clusters, and measured their complexity, average cluster size, and average overlap. The size of a cluster containing a objects out of the total $n = 10$ was measured as the proportion of object pairs that were included, $a(a-1)/(n(n-1))$. Similarly, the overlap between two clusters containing $a \geq b$ objects, of which c were included in both, was measured as $c(c-1)/(b(b-1))$. Figure 2.5 shows the relationship between size, overlap and complexity for a representative subsample of 10^3 of the cluster structures. Figure 2.6 shows the relationship between overlap and complexity for the 823 cluster structures with a constant average cluster size of approximately 41.6%. The basic results are that increasing size increases complexity, increasing cluster overlap decreases complexity, but that the increase due to size outweighs the decrease due to overlap.

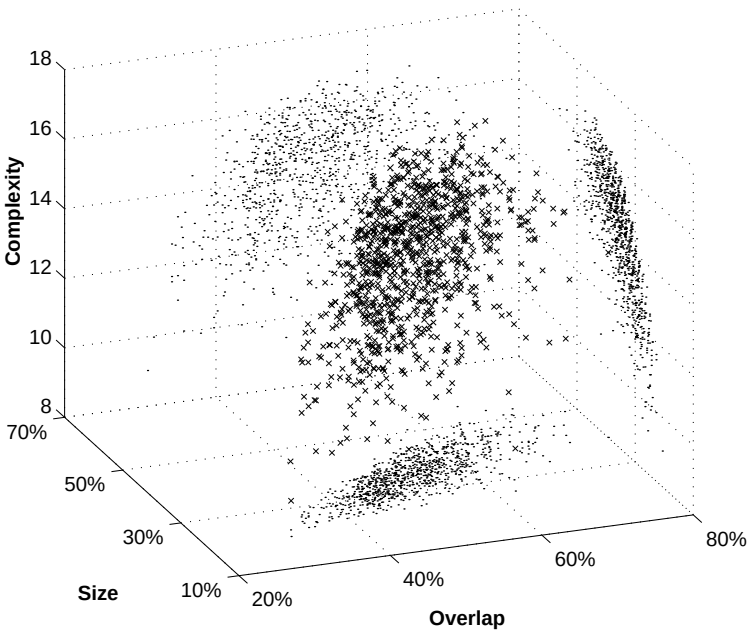


Figure 2.5 The complexity of a sample of 10^3 cluster structures with ten objects and six clusters, shown by crosses as a function of average size and overlap. The projection of each pair of measures is also shown.

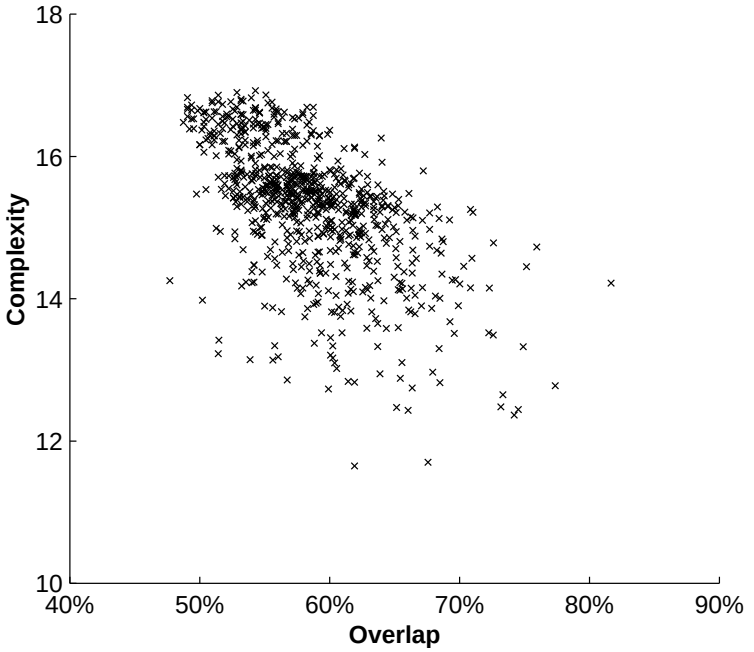


Figure 2.6 The complexity of a sample of 823 cluster structures with constant average size and variable overlap.

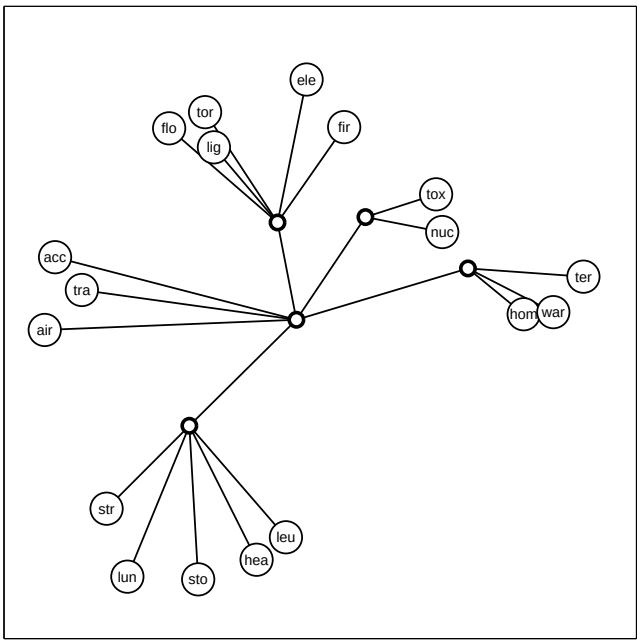


Figure 2.7 Additive tree representation of the risk similarity data

2.4.2 Additive Trees

2.4.2.1 Illustrative Example

Johnson and Tversky [1984, Table A1, lower triangular half] collected similarity data for 18 different 'risks', obtained by pooling the ratings made by subjects for each pair on a nine-point scale. Figure 2.7 shows the additive tree representation of these data, found using a stochastic search algorithm to minimize the GCC. The internal nodes correspond to clusters of risks that can be interpreted as (clockwise from top) 'natural disasters', 'technological disasters', 'violent acts', 'illnesses' and 'accidents'.

It is interesting to compare this representation, which explains about 70% of the variance in the data, with previous additive tree analyses of the same data [Johnson and Tversky 1984; Corter 1996]. These previous analyses did not explicitly consider model complexity, but instead fitted 'full' trees with $(n-3) = 15$ internal nodes, explaining about 75% of the variance. Interpretation of these more complicated trees, however, is only offered for nodes near to the top of tree, and basically corresponds to those concepts shown in Figure 2.7. This lack of extra interpretability suggests that the superior goodness-of-fit achieved by the more complicated trees does not come from finding additional meaningful regularities in the data.

2.4.2.2 Interpretation of Complexity Matrix

The complexity matrix for additive tree models is

$$\mathbf{G} = \begin{bmatrix} \sum_{i<j} e_{ij1} & \sum_{i<j} e_{ij1}e_{ij2} & \cdots & \sum_{i<j} e_{ij1}e_{ijm} \\ \sum_{i<j} e_{ij2}e_{ij1} & \sum_{i<j} e_{ij2} & \cdots & \sum_{i<j} e_{ij2}e_{ijm} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i<j} e_{ijm}e_{ij1} & \sum_{i<j} e_{ijm}e_{ij2} & \cdots & \sum_{i<j} e_{ijm} \end{bmatrix}.$$

where $e_{ijk} = 1$ if the kth edge is on the unique path between objects i and j, and $e_{ijk} = 0$ if it is not. The diagonal elements count the number paths connecting objects that include each edge. The off-diagonal elements count the number of paths connecting objects that use each possible pairing of edges.

2.4.2.3 Extending Star Trees

Additive trees with a single internal (non-terminal) node are called star trees, and have complexity matrix

$$\mathbf{G}_{\text{star}} = \begin{bmatrix} n-1 & 1 & 1 & \cdots & 1 \\ 1 & n-1 & 1 & \cdots & 1 \\ 1 & 1 & n-1 & \cdots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \cdots & n-1 \end{bmatrix}.$$

If a star tree is extended to have two internal nodes, its complexity matrix becomes

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_{\text{star}} & \mathbf{y} \\ \mathbf{y^T} & z \end{bmatrix},$$

where z counts the number of paths that pass through the edge between the internal nodes, and $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ is a column vector where y_i counts the number of paths that pass through both the internal edge and the edge from the terminal node representing the ith object. The determinant of this complexity matrix can be written as

$$\det \mathbf{G} = \det \mathbf{G}_{\text{star}}(z - \mathbf{y^T}\mathbf{G}_{\text{star}}^{-1}\mathbf{y}),$$

where

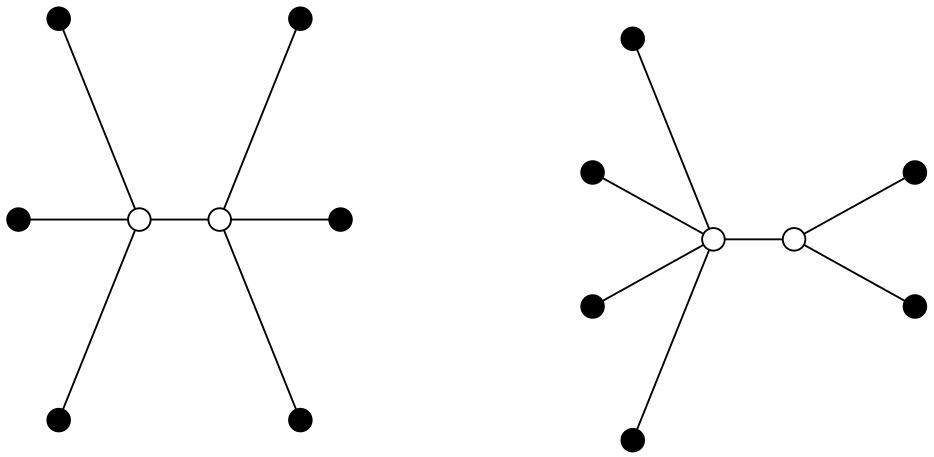


Figure 2.8 The two possible ways of adding a second internal node to a star tree representing six objects.

$$\mathbf{G}_{\text{star}}^{-1} = \frac{1}{2(n-1)(n-2)} \begin{bmatrix} 2n-3 & -1 & -1 & \cdots & -1 \\ -1 & 2n-3 & -1 & \cdots & -1 \\ -1 & -1 & 2n-3 & \cdots & -1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & -1 & \cdots & 2n-3 \end{bmatrix}.$$

In the simplest interesting case, the additional internal node is added to a star tree representing six objects. There are two possibilities, shown in Figure 2.8. The tree on the left divides the objects into two clusters of three. Here $z = 9$, $\mathbf{y}^{\mathbf{T}} = (3, 3, 3, 3, 3, 3)$, and so $\det \mathbf{G} = 3.6$. The tree on the right divides the objects into a cluster of four and a cluster of two. Here $z = 8$, $\mathbf{y}^{\mathbf{T}} = (2, 2, 2, 2, 4, 4)$, and so $\det \mathbf{G} = 2.2$. The tree on the left, with an equal number of objects in each cluster, is more complicated.

More generally, adding an internal node to a star tree representing n objects creates one cluster with r objects, and another cluster with the remaining $(n - r)$. Here $z = r\,(n - r)$, the first r elements of $\mathbf{y}^{\mathbf{T}}$ are $(n - r)$ and the remaining $(n - r)$ elements are r. This results in

$$\det \mathbf{G} = r\,(n-r)\left(1 + \frac{2r\,(n-r)}{(n-1)\,(n-2)} - \frac{n}{n-2}\right),$$

which increases monotonically with $r\,(n - r)$. This generalizes the six object result, showing that dividing any number of objects evenly between the two clusters leads to the greatest complexity.

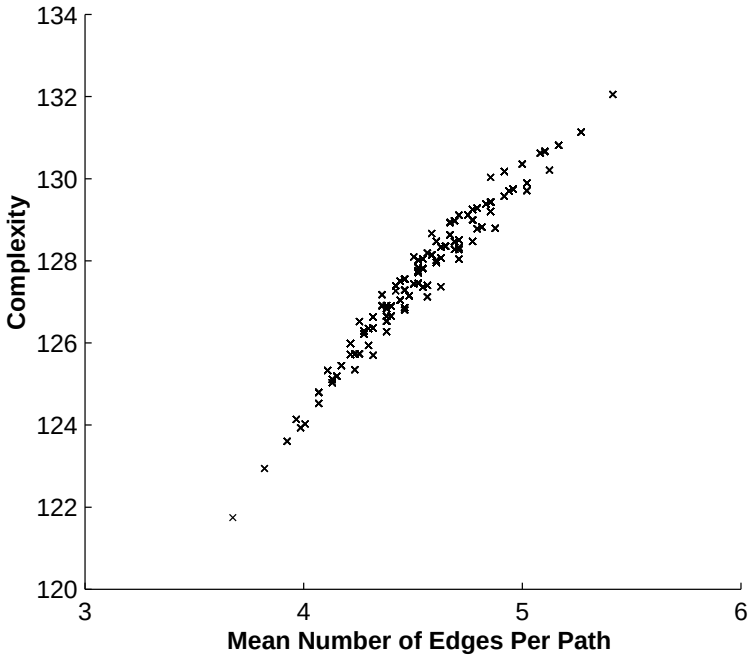


Figure 2.9 The relationship between complexity and mean edges per path for all possible additive trees with ten internal nodes, where each has three terminal nodes.

2.4.2.4 General Tree Structures

The complexity matrix of an additive tree with m clusters can be represented as the result of adding $(m-1)$ clusters to a star tree, so that

$$\mathbf{G} = \begin{bmatrix} \mathbf{G}_{\text{star}} & \mathbf{Y} \\ \mathbf{Y^T} & \mathbf{Z} \end{bmatrix}.$$

The $(m-1) \times (m-1)$ matrix $\mathbf{Z}$ has both rows and columns corresponding to edges between internal nodes, counting the number of paths between objects that include each possible pairing of these edges. The $n \times (m-1)$ matrix $\mathbf{Y}$ has rows corresponding to edges connecting terminal nodes, columns corresponding to edges between internal nodes, and elements counting the number of paths between objects that include each possible combination of these internal and terminal edges. This decomposition allows the determinant to be given as

$$\det \mathbf{G} = \det \mathbf{G}_{\text{star}} \det \left(\mathbf{Z} - \mathbf{Y^T} \mathbf{G}_{\text{star}}^{-1} \mathbf{Y} \right),$$

which depends only on $\mathbf{Y}$ and $\mathbf{Z}$ for a fixed number of objects.

To explore the relationship between the topology of additive trees and their

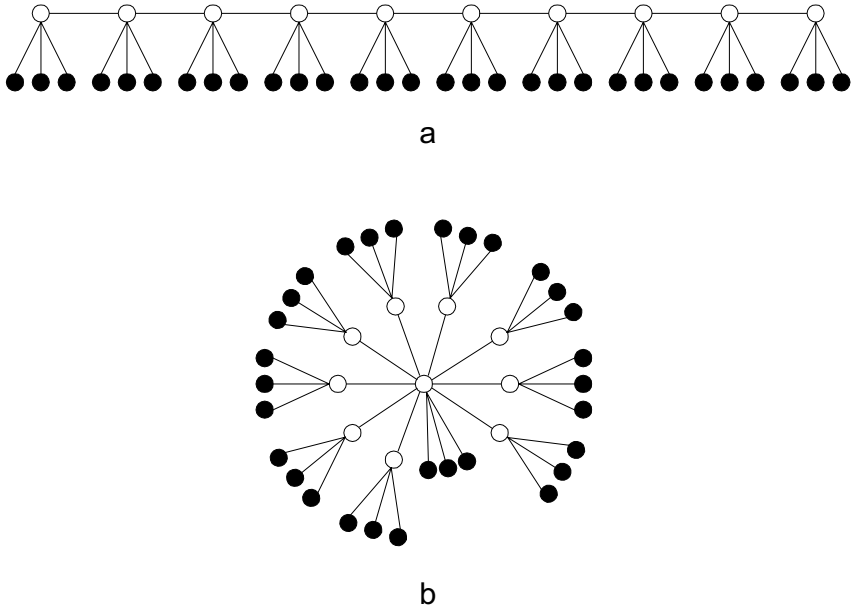


Figure 2.10 The (a) most complicated and (b) least complicated additive trees with ten internal nodes, where each has three terminal nodes.

complexity, Navarro [2002] generated all possible trees with between five and ten internal nodes, under the restriction that all internal nodes were connected to two, three or four terminal nodes. For a given number of internal nodes, complexity was observed to increase roughly linearly with the average number of edges in the paths connecting objects, regardless of the number of terminal nodes. Figure 2.9 shows the relationship for trees of ten internal nodes with three terminal nodes each. Figure 2.10 shows the most and least complicated of these trees. The basic result is that broad trees, which have longer average path lengths, are more complicated than deep trees, which have shorter average path lengths.

2.5 New Psychological Clustering Models

This section presents two new psychological clustering models that extend the representational possibilities of additive clustering and additive trees. The first model uses a similarity measure that considers both common and distinctive features, while the second moves beyond clustering to incorporate continuous dimensions in its representations. Both models are demonstrated by applying them to similarity data under complexity constraints, but analyses of the complexity of these models have not been made. The study of the complexity of these models is

an important area for future research.

2.5.1 Overlapping Common and Distinctive Features

Tversky [1977] proposed two similarity models combining common and distinctive features, known as the Contrast Model and the Ratio Model. Under the Contrast Model, similarity is measured as an additive mixture of common and distinctive features. Under the Ratio Model, similarity is measured as the proportion of common to distinctive features. The Ratio Model has a natural interpretation in terms of a Bayesian theory of generalization [Tenenbaum and Griffiths 2001], but the Contrast Model is more difficult to interpret, because it treats each cluster as being part common feature and part distinctive feature. To overcome this difficulty, Navarro and Lee [2002] proposed a modified version of the Contrast Model that designates each cluster as being either a completely common or completely distinctive feature, but allows both types of cluster in the same model.

Under this "Modified Contrast Model" approach, similarity is measured as

$$\hat{s}_{ij} = c + \sum_{k \in CF} w_k f_{ik} f_{jk} - \sum_{k \in DF} w_k \left| f_{ik} - f_{jk} \right|, \tag{2.6}$$

where $k \in CF$ means that the sum is taken over the common features, and $k \in DF$ means that the sum is taken over the distinctive features. The complexity matrix $\mathbf{G}$ and GCC for this similarity model can be derived in exactly the same way as the purely common and distinctive cases, by making the appropriate choice in Eq. (2.6) for each cluster.

2.5.1.1 Illustrative Example

Rosenberg and Kim [1975] collected data, later published by Arabie, Carroll and DeSarbo [1987, pp. 62–63], measuring the similarities between 15 common kinship terms, such as 'father', 'daughter', and 'grandmother'. The similarities were based on a sorting task undertaken by six groups of 85 subjects, where each kinship term was placed into one of a number of groups, under various instructions to the subjects. A slightly modified version of this data set that excludes the term 'cousin' is considered, because it is interesting to examine how the model deals with the concept of gender, and 'cousin' is the only ambiguous term in this regard.

Table 2.1 describes the overlapping common and distinctive features clustering found by applying stochastic hill-climbing optimization to minimize the GCC. The clusters correspond to easily interpreted common and distinctive features. It has four distinctive features, dividing males from females, once removed terms (aunt, nephew, niece uncle) from those not once removed, extreme generations (granddaughter, grandfather, grandmother, grandson) from middle generations, and the nuclear family (brother, daughter, father, mother, sister, son) from the extended family. It also has six common features, which correspond to meaningful

Table 2.1 Overlapping common and distinctive features representation of the kinship terms.

Type	Objects in Cluster	Weight	Interpretation
DF	Brother, Father, Grandfather, Grandson, Nephew, Son, Uncle	0.452	Gender
CF	Aunt, Uncle	0.298	Adult extended family
CF	Nephew, Niece	0.294	Child extended family
CF	Brother, Sister	0.291	Siblings
CF	Grandfather, Grandmother	0.281	Grandparents
CF	Father, Mother	0.276	Parents
CF	Granddaughter, Grandson	0.274	Grandchildren
DF	Aunt, Nephew, Niece, Uncle	0.230	Once-Removed
DF	Granddaughter, Grandfather, Grandmother, Grandson	0.190	Extreme Generation
DF	Brother, Daughter, Father, Mother, Sister, Son	0.187	Nuclear Family
	universal cluster	0.660	

subsets within the broad distinctions, such as parents, siblings, grandparents and grandchildren. These concepts are common features since, for example, a brother and sister have the similarity of being siblings, but this does not make those who are not siblings, like an aunt and a grandson, more similar.

The kinship data provide a good example of the need to consider both common and distinctive features in the same clustering model. Common features models, such as additive clustering, are inefficient in representing concepts like 'gender', because they need to include separate equally-weighted clusters for 'male' and 'female'. Distinctive feature models, on the other hand, generally cannot represent concepts like 'siblings', where the objects outside the cluster do not belong together.

2.5.1.2 Complexity Issues

The Modified Contrast Model uses both the common and distinctive similarity measures in Eqs. (2.1) and (2.2) to model similarity. This means that, in a way unlike additive clustering or additive tree models, the weight parameters of the model have different 'functional forms' [Myung and Pitt 1997] of interaction, depending on whether they are associated with a common or distinctive feature. An interesting model complexity issue raised by combining common and distinctive features, therefore, relates to the relative complexity of the two different similarity models. Some preliminary evidence [Navarro 2002, pp. 122-124], based on simulation studies, suggests that common features increase the complexity of a model more than distinctive features. Analysis of the complexity matrix for the Modified Contrast Model provides an opportunity to understand the basis and generality of this finding, and is a worthwhile area for further research.

2.5.2 Combining Features with Dimensions

Whatever representational assumptions are made, and whatever similarity measure is used, clustering models are inefficient when dealing with the inherently continuous aspects of the variation between objects. Most psychological modeling in these cases uses the "multidimensional scaling" model described earlier, where objects are represented by values along one or more continuous dimensions, so that they correspond to points in a multidimensional space. The dissimilarity between objects is then measured by the distance between their points. While dimensional representation naturally captures continuous variation, it is constrained by the metric axioms, such as the triangle inequality, that are violated by some empirical data.

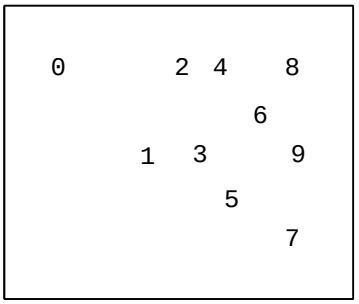
It has been argued [e.g., Carroll 1976; Tenenbaum 1996; Tversky 1977] that spatial representations are most appropriate for low-level perceptual stimuli, whereas cluster representations are better suited to high-level conceptual domains. In general, though, stimuli convey both perceptual and conceptual information, and so both dimensional and clustering representations need to be combined. As Carroll [1976, p. 462] concludes: "Since what is going on inside the head is likely to be complex, and is equally likely to have both discrete and continuous aspects, I believe the models we pursue must also be complex, and have both discrete and continuous components".

In this spirit, Navarro and Lee [2003] developed a representational model that combines continuous dimensions with discrete features. Objects take values on a number of dimensions, as well as potentially belonging to a number of clusters. If there are v dimensions and m features, this means the ith object is defined by a point $\mathbf{p}_i$, a vector $\mathbf{f}_i$, and the cluster weights $\mathbf{w} = (w_1, \ldots, w_m)$. The similarity between the ith and jth objects is then modeled as the sum of the similarity arising from their common features, minus the dissimilarity arising from their dimensional differences under the Minkowskian r-metric, so that:

$$\hat{s}_{ij} = \left(\sum_{k=1}^{m} w_k f_{ik} f_{jk} \right) - \left(\sum_{k=1}^{v} |p_{ik} - p_{jk}|^r \right)^{\frac{1}{r}} + c.$$

2.5.2.1 Illustrative Example

Shepard, Kilpatric and Cunningham [1975] collected data measuring the "abstract conceptual similarity" of the numbers 0 through 9. Figure 2.11(a) displays a two-dimensional representation of the numbers, using the City-Block metric, found by multidimensional scaling. This representation explains only 78.6% of the variance, and fails to capture important regularities in the raw data, such as the fact that the number 7 is more similar to 8 than it is to 9, and that 3 is much more similar to 0 than it is to 8. Figure 2.11(b) shows an eight-cluster representation of the numbers using the same data, found by Tenenbaum [1996] using additive clustering. This representation explains 90.9% of the variance, with clusters corresponding to arithmetic concepts (e.g., $\{2, 4, 8\}$ and $\{3, 6, 9\}$) and to numerical magnitude

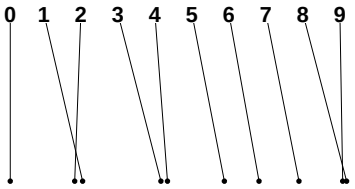


(a)

Objects in Cluster	Weight
2 4 8	0.444
0 1 2	0.345
3 6 9	0.331
6 7 8 9	0.291
2 3 4 5 6	0.255
1 3 5 7 9	0.216
1 2 3 4	0.214
4 5 6 7 8	0.172
universal cluster	0.148

(b)

Figure 2.11 Representations of the numbers similarity data using the (a) dimensional and (b) clustering models.



Objects in Cluster	Weight
2 4 8	0.286
3 6 9	0.282
2 4 6 8	0.224
1 3 9	0.157
universal cluster	0.568

Figure 2.12 Representation of the numbers similarity data using the combined model with one dimension (shown on the left) and four clusters (shown on the right).

(e.g., $\{1, 2, 3, 4\}$ and $\{6, 7, 8, 9\}$). While the clusters are appropriate for representing the arithmetic concepts, a 'magnitude' dimension seems to offer a more efficient and meaningful representation of this regularity than the five clusters used in Figure 2.11(b).

Navarro and Lee [2003] fitted combined models with between one and three dimensions and one and eight clusters to the similarity data. Because analytic results for the complexity of the combined model are not available, the Bayesian approach of selecting the most likely model given the data was used [e.g., Kass and Raftery 1995], based on an approximation to the log posterior found by importance sampling [e.g., Oh and Berger 1993]. The best representation under this measure contains one dimension and four clusters, explains 90.0% of the variance, and is shown in Figure 2.12. The one dimension almost orders the numbers according to their magnitude, with the violations being very small. The four clusters all capture meaningful arithmetic concepts, corresponding to "powers of two", "multiples of three", "multiples of two" (or "even numbers") and "powers of three".

2.5.2.2 Complexity Issues

The combined model also raises interesting complexity issues related to the functional form of parameter interaction. The coordinate locations of the points interact according to the Minkowskian distance metric that is used to model similarity. In psychological applications of multidimensional scaling, particular emphasis has been placed on the $r = 1$ (City-Block) and $r = 2$ (Euclidean) cases because of their relationship, respectively, to so-called 'separable' and 'integral' dimensions [Garner 1974]. Pairs of separable dimensions are those, like shape and size, that can be attended to separately. Integral dimensions, in contrast, are those rarer cases like hue and saturation that are not easily separated. Metrics with $r < 1$ have also been given a psychological justification [Gati and Tversky 1982; Shepard 1991] in terms of modeling dimensions that 'compete' for attention. Little is known about the relative complexities of these different metrics, although there is some simulation study evidence [Lee and Pope 2003] that the City-Block metric is complicated, because it allows multidimensional scaling models to achieve high levels of goodness-of-fit, even for data generated using another metric. There is a need, however, for much more detailed analysis of the complexity of the combined model.

2.6 Conclusion

Clustering aims to find meaningful and predictive representations of data, and so is a fundamental tool for data analysis. One of the strengths of clustering models is that they potentially allow for great representational flexibility, and can accommodate sophisticated measures for assessing the relationships between objects. The price of these freedoms, however, is the need to control their complexity, so that they capture the regularities underlying data that are important for explanation and prediction.

This chapter has attempted to meet the challenge by treating clustering models as statistical models, and using the Geometric Complexity Criterion for the statistical inference of model selection. Theoretically, this statistical approach offers interpretable measures of the complexity of clustering models. The results for additive clustering and additive tree models are good examples of this. Practically, the statistical approach offers a useful way of generating models from data. It compares favorably with the collections of heuristics that must otherwise be used to determine basic properties of a model, such as how many clusters it uses. The illustrative applications of the additive clustering and additive tree models are good examples of the sorts of representations that can be learned from data under complexity constraints. Finally, this chapter has also attempted to demonstrate the potential for new clustering models, and the new complexity issues they raise. Clustering models, like all good scientific models, should be developed and extended boldly, seeking general and powerful accounts of data, but also need to be evaluated and differentiated carefully, taking account of all of the complexities bound up in their

generality and power.

Acknowledgments

This research was supported by Australian Research Council Grant DP0211406, and by the Australian Defence Science and Technology Organisation. We wish to thank Helen Braithwaite, Peter Grünwald, Geoff Latham, In Jae Myung, Kenneth Pope, Chris Woodruff and the reviewers for helpful comments.

Bibliography

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

Arabie, P., J. D. Carroll, and W. S. DeSarbo (1987). *Three-Way Scaling and Clustering*. Newbury Park, CA: Sage.

Arabie, P., L. J. Hubert, and G. De Soete (1996). *Clustering and Classification*. Singapore: World Scientific.

Balasubramanian, V. (1997). Statistical inference, Occam's razor and statistical mechanics on the space of probability distributions. *Neural Computation 9*, 349–368.

Bellman, R. (1970). *Introduction to Matrix Analysis* (Second ed.). New York, NY: McGraw-Hill.

Carroll, J. D. (1976). Spatial, non-spatial and hybrid models for scaling. *Psychometrika 41*, 439–463.

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

Clouse, D. S. and G. W. Cottrell (1996). Discrete multidimensional scaling. In *Proceedings of the Eighteenth Cognitive Science Conference*, San Diego, CA, pp. 290–294. Mahwah, NJ: Erlbaum.

Corter, J. E. (1996). *Tree Models of Similarity and Association*. Thousand Oaks, CA: Sage.

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

Damashek, M. (1995). Gauging similarity with n-grams: Language-independent categorization of text. *Science 267*, 843–848.

DeSarbo, W. S. (1982). GENNCLUS: New models for general nonhierarchical cluster analysis. *Psychometrika 47*, 449–475.

Everitt, B. S. (1993). *Cluster Analysis* (Third ed.). London: Edward Arnold.

Garner, W. R. (1974). *The Processing of Information and Structure*. Potomac, MD: Erlbaum.

Gati, I. and A. Tversky (1982). Representations of qualitative and quantitative

dimensions. *Journal of Experimental Psychology: Human Perception and Performance 8*(2), 325–340.

Gordon, A. D. (1999). *Classification* (Second ed.). London: Chapman & Hall/CRC Press.

Griffiths, T. L. and M. Steyvers (2002). A probabilistic approach to semantic representation. In W. G. Gray and C. D. Schunn (Eds.), *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, pp. 381–386. Mahwah, NJ: Erlbaum.

Grünwald, P. D. (this volume). Universal modeling: Introduction to modern MDL. In P. D. Grünwald, I. J. Myung, and M. A. Pitt (Eds.), *Advances in Minimum Description Length: Theory and Applications*, Cambridge, MA: MIT Press.

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

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

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

Landauer, T. K. and S. T. Dumais (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review 104*(2), 211–240.

Lee, M. D. (1998). Neural feature abstraction from judgments of similarity. *Neural Computation 10*(7), 1815–1830.

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

Lee, M. D. (2002a). Generating additive clustering models with limited stochastic complexity. *Journal of Classification 19*(1), 69–85.

Lee, M. D. (2002b). A simple method for generating additive clustering models with limited complexity. *Machine Learning 49*, 39–58.

Lee, M. D. and D. J. Navarro (2002). Extending the ALCOVE model of category learning to featural stimulus domains. *Psychonomic Bulletin & Review 9*(1), 43–58.

Lee, M. D. and K. J. Pope (2003). Avoiding the dangers of averaging across subjects when using multidimensional scaling. *Journal of Mathematical Psychology 47*, 32–46.

Li, M. and P. Vitányi (1993). *An Introduction to Kolmogorov Complexity and its Applications*. Reading, MA: Addison-Wesley.

Lund, K. and C. Burgess (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behaviour Research Methods, Instruments, & Computers 28*(2), 203–208.

Magnus, J. R. and H. Neudecker (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. New York, NY: Wiley.

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

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

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

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

Navarro, D. J. (2002). *Representing Stimulus Similarity*. Ph. D. thesis, University of Adelaide.

Navarro, D. J. (2003). Regarding the complexity of additive clustering models: Comment on Lee (2001). *Journal of Mathematical Psychology 47*, 241–243.

Navarro, D. J. and M. D. Lee (2002). Commonalities and distinctions in featural stimulus representations. In W. G. Gray and C. D. Schunn (Eds.), *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, pp. 685–690. Mahwah, NJ: Erlbaum.

Navarro, D. J. and M. D. Lee (2003). Combining dimensions and features in similarity-based representations. In S. Becker, S. Thrun, and K. Obermayer (Eds.), *Advances in Neural Information Processing Systems 15*, Cambridge, MA. MIT Press.

Oh, M. and J. O. Berger (1993). Integration of multimodal functions by Monte Carlo importance sampling. *Journal of the American Statistical Association 88*, 450–456.

Pitt, M. A., I. J. Myung, and S. Zhang (2002). Toward a method of selecting among computational models of cognition. *Psychological Review 109*(3), 472–491.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory 42*(1), 40–47.

Rohde, D. L. T. (2002). Methods for binary multidimensional scaling. *Neural Computation 14*(5), 1195–1232.

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

Ruml, W. (2001). Constructing distributed representations using additive clustering. In T. G. Dietterich, S. Becker, and Z. Ghahramani (Eds.), *Advances in Neural Information Processing 14*, Cambridge, MA. MIT Press.

Sattath, S. and A. Tversky (1977). Additive similarity trees. *Psychometrika 42*,

319–345.

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

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

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika 27*(2), 125–140.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science 210*, 390–398.

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

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

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

Takeuchi, J. (this volume). Minimax regret for stochastic processes. In P. D. Grünwald, I. J. Myung, and M. A. Pitt (Eds.), *Advances in Minimum Description Length: Theory and Applications*, Cambridge, MA: MIT Press.

Tenenbaum, J. B. (1996). Learning the structure of similarity. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo (Eds.), *Advances in Neural Information Processing Systems*, Volume 8, pp. 3–9. Cambridge, MA: MIT Press.

Tenenbaum, J. B. and T. L. Griffiths (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences 24*(4), 629–640.

Tversky, A. (1977). Features of similarity. *Psychological Review 84*(4), 327–352.

Tversky, A. and J. W. Hutchinson (1986). Nearest neighbor analysis of psychological spaces. *Psychological Review 93*(1), 3–22.

Index