Latent Features in Similarity Judgments: A Nonparametric Bayesian Approach

Daniel J. Navarro
School of Psychology
University of Adelaide

Thomas L. Griffiths
Department of Psychology
University of California, Berkeley

Abstract

One of the central problems in cognitive science is determining the mental representations that underlie human inferences. Solutions to this problem often rely on the analysis of subjective similarity judgments, on the assumption that recognizing "likenesses" between people, objects and events is crucial to everyday inference. One such solution is provide by the additive clustering model, which is widely used to infer the features of a set of stimuli from their similarities, on the assumption that similarity is a weighted linear function of common features. Existing approaches for implementing additive clustering often lack a complete framework for statistical inference, particularly with respect to choosing the number of features. To address these problems, this paper develops a fully Bayesian formulation of the additive clustering model, using methods from nonparametric Bayesian statistics to allow the number of features to vary. We use this to explore several approaches to parameter estimation, showing that the nonparametric Bayesian approach provides a straightforward way to obtain estimates of both the number of features and their importance.

Introduction

One of the central problems in cognitive science is determining the mental representations that underlie human inferences. A variety of solutions to this problem are based on the analysis of subjective similarity judgments, on the assumption that recognizing "likenesses" between people, objects and events is crucial to everyday inference. However, since subjective similarity cannot be derived from a straightforward analysis of objective stimulus characteristics (Goodman, 1972), it is important that mental representations be constrained by empirical data (Komatsu, 1992; Lee, 1998). By defining a probabilistic model that accounts for the similarity between stimuli based on their representation, statistical methods can be used to infer these underlying representations from human judgments. The particular methods used to infer represention from similarity judgments depend on the nature of the underlying representations. For stimuli that are assumed to be represented as points in some psychological space, multidimensional scaling algorithms (Torgerson, 1958) can be used to translate similarity judgments into stimulus locations. For stimuli that are assumed to be represented in terms of a set of latent features (Tversky, 1977), *additive clustering* is the method of choice (Shepard & Arabie, 1979).

Additive clustering provides a method of assigning a set of latent features to a collection of objects, based on the observable similarities between those items. The model is related to factor analysis, multidimensional scaling and latent class models, and shares a number of important issues. When extracting a set of latent features we need to infer the dimension of the model (i.e., number of features), determine the best feature allocations, and estimate the saliency (or importance) weights associated with each feature. Motivated in part by these issues, this paper develops a fully Bayesian formulation of the additive clustering model, using methods from nonparametric Bayesian statistics to allow the number of features to vary. We use this to explore several approaches to parameter estimation, showing that the nonparametric Bayesian approach provides a straightforward way to obtain estimates of both the number of features and their importance.

In what follows, we assume that the data take the form of an $n \times n$ similarity matrix $\mathbf{S} = [s_{ij}]$, where s_{ij} is the judged similarity between the ith and jth of n objects. The various similarities are assumed to be symmetric (with $s_{ij} = s_{ji}$) and non-negative, often constrained to lie on the interval $[0, 1]$. The source of such data can vary considerably: in psychology alone, similarity data have been collected using a number of experimental methodologies, including rating scales (e.g., Kruschke, 1993), confusion probabilities (e.g., Shepard, 1972), sorting tasks (e.g., Rosenberg & Kim, 1975), or forced-choice tasks (e.g., Navarro & Lee, 2002). Additionally, in applications outside psychology similarity matrices are often calculated using aspects of the objective structure of the stimulus items (e.g., Dayhoff, Schwartz, & Orcutt, 1978; Henikoff & Henikoff, 1992).

Latent Variable Models for Similarity Judgment

The analysis of similarities is perhaps best treated as a question of inferring a *latent structure* from the observed similarity data, for which a variety of methods have been

proposed. For instance, besides the latent features approach, latent metric spaces have been found using multidimensional scaling, latent classes found by partitioning, and latent trees constructed using hierarchical clustering and related methods. Even the factor analysis model for the analysis of covariances has been used for this purpose. Since additive clustering has close ties to these methods, we provide a brief overview.

Multidimensional scaling. The first method to be developed explicitly for the extraction of latent structure in similarity data was multidimensional scaling (Torgerson, 1958; Attneave, 1950; Shepard, 1962; Kruskal, 1964a, 1964b; Young & Householder, 1938), in which items are assumed to be represented as points in a low dimensional space, usually equipped with one of the Minkowski distance metrics (Minkowski, 1891). The motivation behind this approach comes from measurement theory, with particular reference to psychophysical measurement (e.g., Stevens, 1946, 1951). In psychophysical scaling, the goal is to construct a latent scale that translates a physical measurement (e.g., frequency) into a subjective state (e.g., pitch). Typically, however, stimuli may vary simultaneously in multiple respects, so the single scale generalizes to a latent metric space in which observed stimuli are located. Although multidimensional scaling is not always formalized as a statistical model, it is common to use squared error as a loss function, which agrees with the Gaussian error model adopted by some authors (e.g., Lee, 2001).

Factor analysis. The well-known factor analysis model (Thurstone, 1947; Spearman, 1904, 1927) and the closely-related principal component analysis technique (Pearson, 1901; Hotelling, 1933; in effect the same model, minus the error theory - see Lawley & Maxwell, 1963) both predate multidimensional scaling by some years. In these approaches stimulus items are assumed to "load" on a set of latent variables or "factors". Since these variables are continuous-valued, factor analysis is closely related to multidimensional scaling. However, the "common factors" model makes different assumptions about similarity to the Minkowski distance metrics, so the two are not equivalent. Historically, neither factor analysis nor principal components were widely used for modeling similarity (but see Ekman, 1954, 1963). In recent years this has changed somewhat, with principal components analysis becoming a standard method for making predictions about document similarities, under the name of "latent semantic analysis" (Landauer & Dumais, 1997). On occasions, those predictions have been compared to human similarity judgments (Lee, Pincombe, & Welsh, 2005).

Partitions. A discrete alternative to the continuous methods provided by multidimensional scaling and factor analysis is clustering. The aim behind clustering is the unsupervised extraction of a classification system for different items, such that similar items tend to be assigned to the same class (Sokal, 1974). Clustering methods vary extensively (A. K. Jain, Murty, & Flynn, 1999), with different models imposing different structural constraints on how objects can be grouped, as illustrated in Figure 1. The *partitioning approach*, very commonly used as a general data analysis technique, forces each object to be assigned to exactly one cluster. This approach can be interpreted as grouping the objects into equivalence classes without specifying how the clusters relate to each other. For example, if the objects A though H in Figure 1(a) correspond to people, the partition might indicate which of four different companies employs each person. Commonly-used

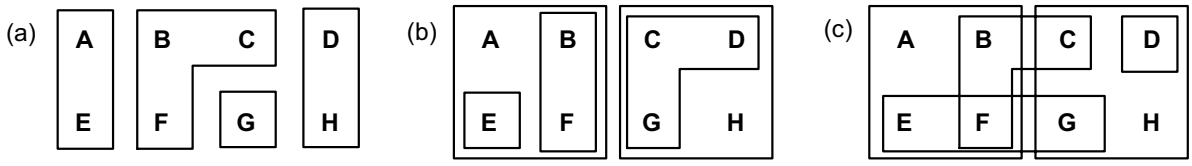


Figure 1: Three different representational assumptions for clustering models, showing (a) partitioning, (b) hierarchical, and (c) overlapping structures.

methods for extracting partitions include heuristic methods such as k-means (McQueen, 1967; Hartigan & Wong, 1979), as well as more statistically-motivated approaches based on mixture models (Wolfe, 1970; McLachlan & Basford, 1988; Kontkanen, Myllymäki, Buntine, Rissanen, & Tirri, 2005). Partitioning models are rarely used for the representation of stimulus similarities, though they are quite common for representing the similarities between people, sometimes in conjunction with the use of other models for representing stimulus similarities (e.g., McAdams, Winsberg, Donnadieu, de Soete, & Krimphoff, 1995).

Hierarchies. The major problem with partitioning models is that, in the example above, 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. Other clustering schemes allow objects to belong to multiple classes. The hierarchical approach (Sneath, 1957; Sokal & Sneath, 1963; Johnson, 1967; D'Andrade, 1978) allows for nested clusters, for instance. Thus the arrangement in Figure 1(b) could show not just the company employing each person, but also the division they work in within that company, and further subdivisions in the organizational structure. Useful extensions to this approach are provided by the additive tree (Buneman, 1971; Sattath & Tversky, 1977), extended tree (Corter & Tversky, 1986) and bidirectional tree (Cunningham, 1978) models.

Additive Clustering: A Latent Feature Model

The additive clustering (ADCLUS) model (Shepard & Arabie, 1979) was developed to provide a discrete alternative to multidimensional scaling, allowing similarity models to encompass a range of data sets for which spatial models seem inappropriate (Tversky, 1977). It provides a natural extension of the partitioning and hierarchical clustering models, and has an interpretation as a form of binary factor analysis. Viewed as a clustering technique, additive clustering is an example of *overlapping clustering* (Jardine & Sibson, 1968; Cole & Wishart, 1970), which imposes no representational restrictions on the clusters, allowing any cluster to include any object and any object to belong to any cluster (e.g., Hutchinson & Mungale, 1997, p. 88). By removing these restrictions, overlapping clustering models can be interpreted as assigning features to objects. For example, in Figure 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 shared features. This representational flexibility allows overlapping clustering to be applied far more broadly than hierarchical clustering or partitioning methods.

Additive clustering relies on the *common features* measure for item similarities (Tversky, 1977; Navarro & Lee, 2004), in which the empirically observed similarity s_{ij} between items i and j is assumed to be well-approximated by a weighted linear function μ_{ij} of the features shared by the two items,

$$\mu_{ij} = \sum_{k=1}^{m} w_k f_{ik} f_{jk}. \tag{1}$$

In this expression, $f_{ik} = 1$ if the ith object possesses the kth feature, and $f_{ik} = 0$ if it does not, and w_k is the non-negative saliency weight applied to that feature. Under these assumptions, a representation that uses m common features to describe n objects is defined by the $n \times m$ feature matrix $\mathbf{F} = [f_{ik}]$, and the saliency vector $\mathbf{w} = (w_1, \ldots, w_m)$. Accordingly, additive clustering techniques aim to uncover a feature matrix and saliency vector that provide a good approximation to the empirical similarities. In most applications it is assumed that there is a fixed "additive constant", a required feature possessed by all objects.

To formalize additive clustering as a statistical model, it has become standard practice (Tenenbaum, 1996; Lee, 2002a) to assume that the empirically observed similarities are drawn from a normal distribution with common variance σ^2, and means described by the common features model. Given the latent featural model $(\mathbf{F}, \mathbf{w})$, we may write

$$s_{ij} \quad | \quad \mathbf{F}, \mathbf{w}, \sigma \quad \sim \quad \text{Gaussian}(\mu_{ij}, \sigma^2). \tag{2}$$

Note that σ is a nuisance parameter in this model, denoting the amount of noise in the data, but carrying no psychological significance. The statistical formulation of the model allows us to obtain the additive clustering decomposition of the similarity matrix,

$$\mathbf{S} = \mathbf{F}\mathbf{W}\mathbf{F}' + \mathbf{E}, \tag{3}$$

where $\mathbf{W} = \text{diag}(\mathbf{w})$ is a diagonal matrix with nonzero elements corresponding to the saliency weights, and $\mathbf{E} = [\epsilon_{ij}]$ is an $n \times n$ matrix with entries drawn from a Gaussian$(0, \sigma^2)$ distribution. This is illustrated in Figure 2, which decomposes a continuously varying similarity matrix $\mathbf{S}$ into the binary feature matrix $\mathbf{F}$, non-negative weights $\mathbf{W}$, and error terms $\mathbf{E}$.

Additive clustering also has a factor analytic interpretation (Shepard & Arabie, 1979; Mirkin, 1987), since Equation 3 has the same form as the factor analysis model, with the "feature loadings" f_{ik} constrained to 0 or 1. By imposing this constraint, additive clustering enforces a variant of the "simple structure" concept (Thurstone, 1947, ch. 14) that provides the theoretical basis for many factor rotation methods currently in use (see Browne, 2001). To see this, it suffices to note that the most important criterion for simple structure is *sparsity*. In the extreme case, illustrated in the top row of Figure 3,

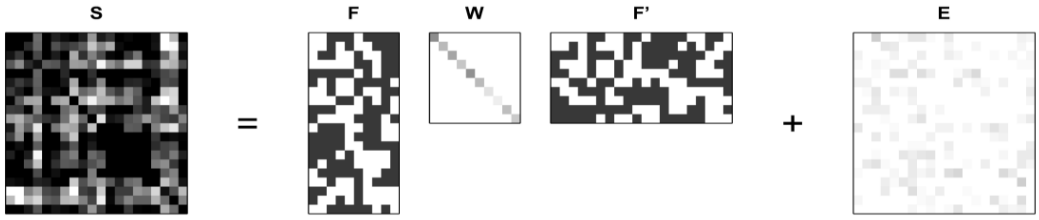


Figure 2: The additive clustering decomposition of a similarity matrix. A continuously varying similarity matrix **S** may be decomposed into a binary feature matrix **F**, a diagonal matrix of non-negative weights **W**, and a matrix of error terms **E**.

each item might load only on a single factor, yielding a partition-like representation (panel a) of the item vectors (shown in panel b). As a consequence, the factor-loading vectors project onto a very constrained part of the unit sphere (panel c). Although most factor rotation methods seek to approximate this partition-like structure (Browne, 2001, p. 116), Thurstone himself allowed more general patterns of sparse factor loadings. Figure 3(d) provides an illustration, corresponding to a somewhat different configuration of items in the factor space (panel e) and on the unit sphere (panel f). The additive clustering model is similarly general in terms of the pattern of zeros it allows, as illustrated in Figure 4(a). However, by forcing all "loadings" to be 0 or 1, every feature vector is constrained to lie at one of the vertices of the unit cube, as shown in Figure 4(b). When these vectors are projected down onto the unit sphere, they show a different, though clearly constrained pattern. It is in this sense that the additive clustering model implements the simple structure concept, and is the motivation behind the "qualitative factor analysis" view of additive clustering (Mirkin, 1987).

Existing Approaches to Additive Clustering

Since the introduction of the additive clustering model, a range of algorithms have been used to infer features, including "subset selection" (Shepard & Arabie, 1979), expectation maximization (Tenenbaum, 1996), continuous approximations (Arabie & Carroll, 1980) and stochastic hillclimbing (Lee, 2002b) among others. A review, as well as an effective combinatorial search algorithm, is given by Ruml (2001). However, in order to provide a context, we present a brief discussion of some of the existing approaches.

The original additive clustering technique (Shepard & Arabie, 1979) was a combinatorial optimization algorithm that employed a heuristic method to reduce the space of possible cluster structures to be searched. Shepard and Arabie observed that a subset of the stimuli in the domain is most likely to constitute a feature if the pairwise similarities of the stimuli in the subset are high. They define the s-level of a set of items c, to be the lowest pairwise similarity rating for two stimuli within the subset. Further, the subset c is *elevated* if and only if every larger subset that contains c has a lower s-level than

c. Having done so, they constructed the algorithm in two stages. In the first step, all elevated subsets are found. In the second step, the saliency weights are found and the set of included features is reduced. The weight initially assigned to each potential cluster is proportional to its *rise*, defined as the difference between the s-level of the subset and the minimum s-level of any subset containing the original subset. The weights are then iteratively adjusted by a gradient descent procedure.

The next major development in inference algorithms for the ADCLUS model was the introduction of a mathematical programming approach (Arabie & Carroll, 1980). In this technique, the discrete optimization problem is recast as a continuous one. The cluster membership matrix $\mathbf{F}$ is initially allowed to assume continuously varying values, rather than the binary membership values required in the final solution. An error function is defined as the weighted sum of two parts, the first being the sum squared error and the second being a penalty function designed to push the elements of $\mathbf{F}$ towards 0 or 1.

A statistically motivated approach proposed by Tenenbaum (1996) uses the expectation maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977). As with the mathematical programming formulation, the number of features needs to be specified in advance, and the discrete problem is (in effect) converted to a continuous one. The EM algorithm for additive clustering consists of an alternating two-step procedure. In the E-step, the saliency weights are held constant, the expected sum squared error is estimated, and (conditional on these saliency weights), the expected values for the elements of $\mathbf{F}$ are calculated. Then, using the expected values for the feature matrix calculated during the E-step, the M-step finds a new set of saliency weights that minimize the expected sum squared error. As the EM algorithm iterates, the value of σ is reduced, and the expected assignment values converge to 0 or 1, yielding a final feature matrix $\mathbf{F}$ and saliency weights $\mathbf{w}$.

Note that the EM approach treates σ as something more akin to a "temperature" parameter rather than a genuine element of the data-generating process. Moreover, it still requires the number of features to be fixed in advance. To redress some of these problems, Lee (2002b) proposed a simple stochastic hillclimbing algorithm that "grows" an additive clustering model. The algorithm initially specifies a single-feature representation, which is optimized by "flipping" the elements of $\mathbf{F}$ (i.e., $f_{ik} \rightarrow 1 - f_{ik}$) one at a time, in a random order. Every time a new feature matrix is generated, best-fitting saliency weights $\mathbf{w}^*$ are found by solving the corresponding non-negative least squares problem (see Lawson & Hanson, 1974), and the solution is evaluated. Whenever a better solution is found, the flipping process restarts. If flipping f_{ik} results in an inferior solution, it is flipped back. If no element of $\mathbf{F}$ can be flipped to provide a better solution, a local minimum has been reached. Since, as Tenenbaum (1996) observed, additive clustering tends to be plagued with local minima problems, the algorithm allows the locally optimal solution to be "shaken", by randomly flipping several elements of $\mathbf{F}$ and restarting, in order to find a globally optimal solution. Once this process terminates, a new (randomly generated) cluster is added, and this solution is used as the starting point for a new optimization procedure. Importantly, potential solutions are evaluated using the stochastic complexity measure (Rissanen, 1996), which provides a statistically-principled method for determin-

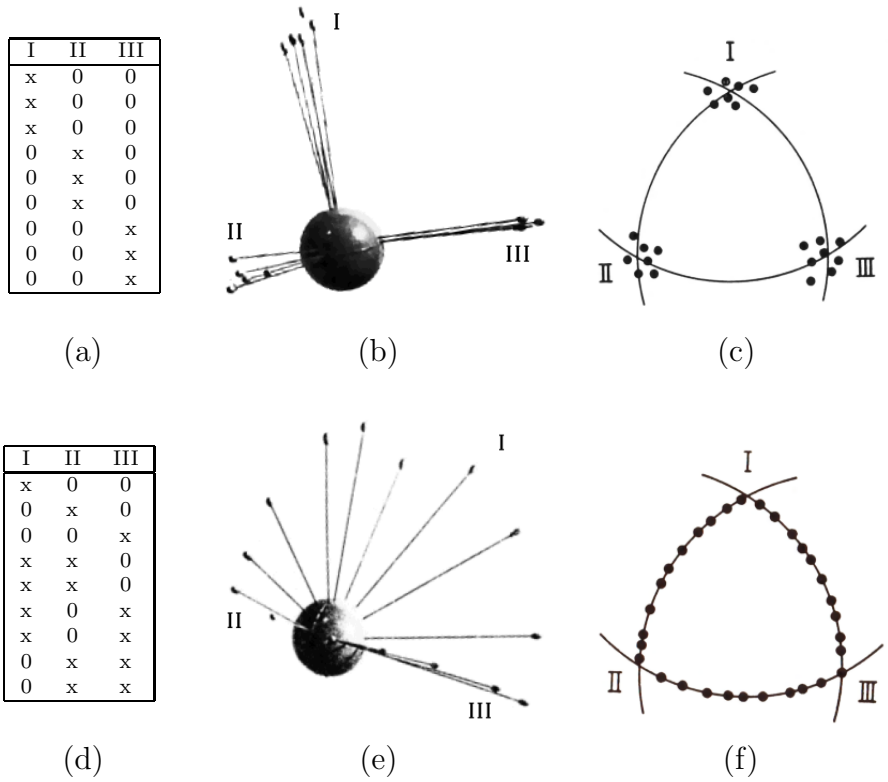


Figure 3: Simple structures in a three-factor solution, adapted from Thurstone's (1947) original examples (p.126-127, 183-186). In the tables, crosses denote non-zero factor loadings. The middle panels illustrate possible item vectors in the solutions, and the right panels show corresponding projections onto the unit sphere. The partition-style solution shown in the top row (panels a-c) is the classic example of a simple structure, but more general sparse structures of the kind illustrated in the lower row (panels d-f) are allowed.

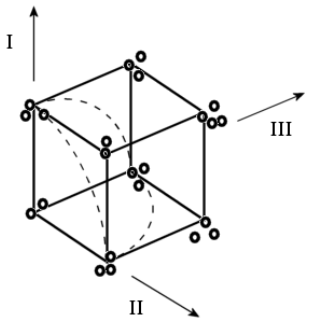
ing the number of features to include in the representation (and under some situations has a Bayesian interpretation; see Myung, Balasubramanian, & Pitt, 2000).

A Nonparametric Bayesian ADCLUS Model

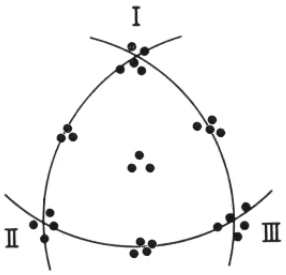
The additive clustering model provides a method for relating a latent feature set to an observed similarity matrix. In order to complete the statistical framework, we need to specify a method for learning a feature set and saliency vector from data. In contrast to the approaches discussed in the previous section, our solution is to cast the additive clustering model in an explicitly Bayesian framework, placing priors over both $\mathbf{F}$ and $\mathbf{w}$, and then basing subsequent inferences on the full joint posterior distribution $p(\mathbf{F}, \mathbf{w}|\mathbf{S})$ over possible representations in light of the observed similarities. However, since we wish

I	II	III
1	0	0
0	1	0
0	0	0
0	0	1
1	1	0
1	0	1
1	0	1
0	1	1
1	1	1

(a)



(b)



(c)

Figure 4: The variant of simple structure enforced by the ADCLUS model. Any sparse pattern of binary loadings is allowable (panel a), and the natural way to interpret item vectors is in terms of the vertices of the unit cube (panel b) on which all feature vectors lie, rather than project the vectors onto the unit sphere (panel c).

to allow the additive clustering model the flexibility to extract a range of structures from the empirical similarities $\mathbf{S}$, we want the implied marginal prior $p(\mathbf{S})$ to have broad support. In short, we have a nonparametric problem, in which the goal is to learn from data without making any strong prior assumptions about the family of distributions that might best describe those data.

The rationale for adopting a nonparametric approach is that the generative process for a particular data set is unlikely to belong to any finite-dimensional parametric family, so it would be preferable to avoid making this false assumption at the outset. From a Bayesian perspective, nonparametric assumptions require us to place a prior distribution that has broad support across the space of probability distributions. In general, this is a hard problem: thus, to motivate a nonparametric prior for a latent feature model, it is useful to consider the simpler case of latent class models. In these models, a common choice relies on the Dirichlet process (Ferguson, 1973). The Dirichlet process is by far the most widely-used distribution in Bayesian nonparametrics, and specifies a distribution that has broad support across the discrete probability distributions. The distributions indexed by the Dirichlet process can be expressed as countably infinite mixtures of point masses (Sethuraman, 1994), making them ideally suited to act as priors in infinite mixture models (Escobar & West, 1995; Rasmussen, 2000). For the current paper, however, it is more important to note that Dirichlet process also implies a distribution over latent class assignments: any two observations in the sample that were generated from the same mixture component may be treated as members of the same class, allowing us to specify priors over infinite partitions. This implied prior can be useful for data clustering purposes (e.g., Navarro, Griffiths, Steyvers, & Lee, 2006), particularly since samples from this prior can be generated using a simple stochastic process known as the *Chinese restaurant*

process[1] (Blackwell & MacQueen, 1973; Aldous, 1985; Pitman, 1996). In a similar manner, it is possible to generate infinite latent hierarchies using other priors, such as the Pólya tree (Ferguson, 1974; Kraft, 1964) and Dirichlet diffusion tree (Neal, 2003) distributions. The key insight in all cases is to separate the prior over the *structure* (e.g., partition, tree, etc) from the prior over the other parameters associated with that structure. For instance, most Dirichlet process priors for mixture models are explicitly constructed by placing a Chinese restaurant process prior over the infinite latent partition, and using a simple parametric prior for the parameters associated with each element of that partition.

This approach is well-suited for application to the additive clustering model. For simplicity, we assume that the priors for $\mathbf{F}$ and $\mathbf{w}$ are independent of one another. Moreover, we assume that feature saliencies are independently generated, and employ a fixed Gamma distribution as the prior over these weights. This yields the simple model

$$\begin{array}{rcl} s_{ij} \mid \mathbf{F}, \mathbf{w}, \sigma & \sim & \text{Gaussian}(\mu_{ij}, \sigma^2) \\ w_k \mid \lambda_1, \lambda_2 & \sim & \text{Gamma}(\lambda_1, \lambda_2) \end{array} \tag{4}$$

The choice of Gamma priors is primarily one of convenience, and it would be straightforward to extend this to more flexible distributions. As with Dirichlet process models, the key element is the prior distribution over model structure: specifically, we need a prior over infinite latent feature matrices. By specifying such a prior, we obtain the desired nonparametric additive clustering. Moreover, infinite models have some inherent psychological plausibility here, since it is commonly assumed that there are an infinite number of features that may be validly assigned to an object (Goodman, 1972). As a result, we might expect the number of features required to grow arbitrarily large, providing that a sufficiently large number of stimuli were observed to elicit the appropriate contrasts.

Our approach to this problem employs the *Indian buffet process* (IBP; Griffiths & Ghahramani, 2005), a simple stochastic process that generates samples from a distribution over sparse binary matrices with a fixed number of rows and an unbounded number of columns. This is particularly useful as a method for placing a prior over $\mathbf{F}$, since there is generally no good reason to assume an upper bound on the number of features that might be relevant to a particular similarity matrix. The IBP can be understood by imagining an Indian restaurant, in which there is a buffet table containing an infinite number of dishes. Each customer entering the restaurant samples a number of dishes from the buffet, with a preference for those dishes that other diners have tried. For the kth dish sampled by at least one of the first $i-1$ customers, the probability that the ith customer will also try that dish is

$$p(f_{ik} = 1 | \mathbf{F}_{i-1}) = \frac{n_k}{i}, \tag{5}$$

where $\mathbf{F}_{i-1}$ records the choices of the previous customers, and n_k denotes the number of previous customers that have sampled that dish. Being adventurous, the new customer may try some hitherto untasted meals from the infinite buffet on offer. The number

[1] The origin of the term is due to Jim Pitman and Lester Dubner, and refers to the Chinese restaurants in San Francisco that appear to have infinite seating capacity. The term "Indian buffet process" introduced later is named by analogy to the Chinese restaurant process.

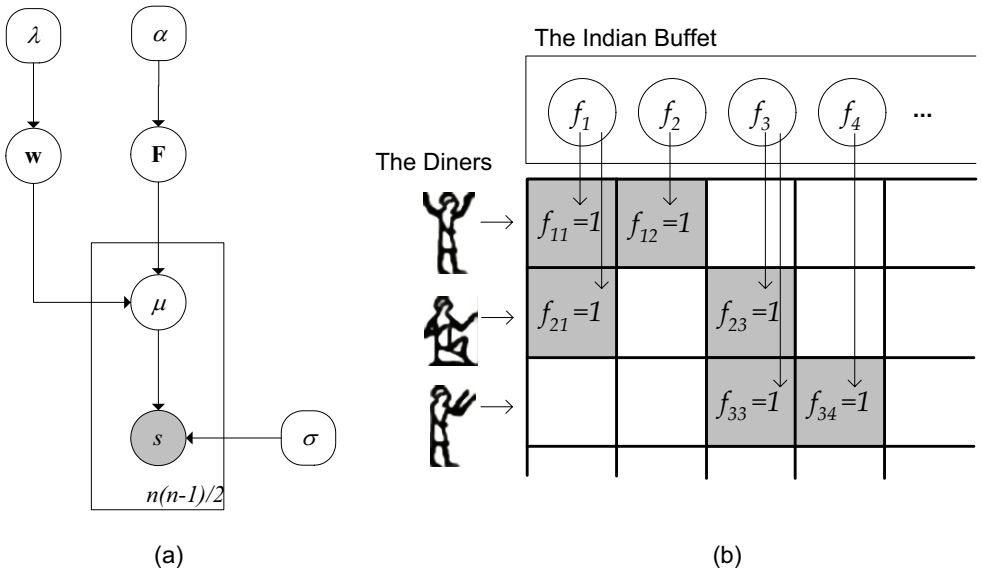


Figure 5: Graphical model representation of the IBP-ADCLUS model. Panel (a) shows the hierarchical structure of the ADCLUS model, and panel (b) illustrates the method by which a feature matrix is generated using the Indian buffet process.

of new dishes taken by customer n follows a Poisson(α/n) distribution. Importantly, this sequential process generates *exchangeable* observations (see Griffiths & Ghahramani, 2005, for a precise treatment). In other words, the probability of a binary feature matrix $\mathbf{F}$ does not depend on the order in which the customers appear (and is thus invariant under permutation of the rows). As a consequence, it is always possible to treat a particular observation as if it were the last one seen: much of the subsequent development in the paper relies on this property. As a consequence, we will often be interested in the special case of this equation corresponding to the nth customer, in which

$$p(f_{nk} = 1|\mathbf{F}_{n-1}) = \frac{n_k}{n}. \tag{6}$$

If we assume that the ordering of the columns in the feature matrix is irrelevant to the model (which is true for additive clustering), then every binary feature matrix $\mathbf{F}$ is a member of a particular equivalence class $[\mathbf{F}]$, and it is at the level of these equivalence classes that we wish to specify our priors. Accordingly, we are interested in the distribution over equivalence classes from which the IBP generates samples (Griffiths & Ghahramani, 2005), which assigns probability to $[\mathbf{F}]$ as follows:

$$p([\mathbf{F}]|\alpha) = \frac{\exp(-\alpha H_n)\,\alpha^m}{\prod_{h=1}^{2^n-1} m_h!} \prod_{k=1}^{m} \frac{(n-n_k)!(n_k-1)!}{n!}, \tag{7}$$

where H_n denotes the nth harmonic number, $H_n = \sum_{j=1}^{n} 1/j$. In this expression, h is an index variable that refers to one of the $2^n - 1$ possible assignments of items to a

particular feature, excluding the case where all elements are zero, and m_h counts the number of features in $\mathbf{F}$ that have that particular pattern of assignments. To summarize, the nonparametric Bayesian additive clustering model may be written,

$$\begin{array}{rcll} s_{ij} & | \; \mathbf{F}, \mathbf{w}, \sigma & \sim & \text{Normal}(\mu_{ij}, \sigma^2) \\ w_k & | \; \lambda_1, \lambda_2 & \sim & \text{Gamma}(\lambda_1, \lambda_2) \\ \mathbf{F} & | \; \alpha & \sim & \text{IBP}(\alpha). \end{array} \tag{8}$$

With a fully specified statistical model in place, we are now in a position to discuss methods for performing the required inferences. Specifically, given the observed similarities $\mathbf{S}$, our goal is to infer the features $\mathbf{F}$ and saliencies $\mathbf{w}$ that underlie those data. To do so, we fix the hyperparameters α, σ, λ_1 and λ_2, and then infer the posterior distribution over possible latent features. In the next two sections, we first discuss a numerical method for approximating this posterior distribution, and then move on to a discussion of the kinds of estimators that can be constructed from this distribution.

A Gibbs-Metropolis Sampling Scheme

Having now provided a complete specification of the model, we now turn to the question of how we can perform inference. As a Bayesian formulation of additive clustering, statistical inference in Equation 8 is based on the posterior distribution over feature matrices and saliency vectors, $p(\mathbf{F}, \mathbf{w}|\mathbf{S})$. Since the priors over $\mathbf{F}$ and $\mathbf{w}$ are independent, the application of Bayes' rule yields,

$$p(\mathbf{F}, \mathbf{w}|\mathbf{S}) = \frac{p(\mathbf{S}|\mathbf{F}, \mathbf{w})p(\mathbf{F})p(\mathbf{w})}{p(\mathbf{S})} \tag{9}$$

where, for ease of exposition, we omit the dependence on the hyperparameters. Naturally, in any Bayesian model the ideal approach is to calculate posterior quantities using exact methods. Unfortunately, this is quite difficult in this case, particularly since the number of features is unbounded *a priori*. In view of this difficulty, a natural alternative is to use Markov chain Monte Carlo (MCMC) methods to repeatedly sample from the posterior distribution: estimates of posterior quantities can be made using these samples as proxies for the full distribution. Accordingly, we now describe a simple MCMC scheme for the Bayesian ADCLUS model, which uses a combination of Gibbs sampling (Geman & Geman, 1984) and more general Metropolis proposals (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953).

The idea behind our procedure, as with all basic MCMC methods, is to start with randomly chosen values for $\mathbf{F}$ and $\mathbf{w}$, and update these values using a simple resampling procedure, discussed below. In our approach, at each step only a single f_{ik} value or w_k value is resampled; all other variables are held fixed at their pre-existing values. An iteration consists of a single sweep through all feature assignments and saliency weights. The logic behind the approach is that, irrespective of the starting point, a properly chosen sampling scheme will converge on samples from the "target" distribution, in this case the posterior distribution $p(\mathbf{F}, \mathbf{w}|\mathbf{S})$. A good introduction to the general approach, along with

the conditions upon which MCMC convergence relies, is given by Gilks, Richardson, and Spiegelhalter (1995). In the remainder of the section, we describe the sampling procedure that we used. The sampler is broken into three parts: (1) resampling a saliency w_k for a feature k that is currently non-empty (i.e., possessed by at least one item), (2) resampling a feature assignment f_{ik}, where k again refers to a non-empty feature, and (3) resampling *all* the assignments and saliencies for the (infinite) set features that are currently empty. We discuss each part in turn.

Saliency Weights (Non-Empty Features). When resampling the saliency of a non-empty feature, we use a Metropolis scheme with a Gaussian proposal distribution. In other words, if the current state saliency is w_k, a candidate w_k^* is first generated by drawing a sample from a Gaussian$(w_k, 0.05)$ distribution. The value of w_k is then reassigned using the Metropolis update rule. If $\mathbf{w}_{-k}$ denotes the set of all saliencies except w_k, then this update rule can be written as

$$w_k \leftarrow \begin{cases} w_k^* & \text{if } u < a \\ w_k & \text{otherwise} \end{cases}, \quad \text{where } a = \frac{p(\mathbf{S}|\mathbf{F}, \mathbf{w}_{-k}, w_k^*)p(w_k^*)}{p(\mathbf{S}|\mathbf{F}, \mathbf{w}_{-k}, w_k)p(w_k)}. \tag{10}$$

In this expression, u is a uniform random variate on the interval $[0, 1)$. Note that, since the candidates are drawn from a Gaussian distribution, it is quite possible for the Metropolis sampler to propose replacement values w_k^* that are less than zero. However, the Metropolis update rule will never accept one of these proposals, since $p(w_k^*) = 0$ for all $w_k^* < 0$.

Feature Assignments (Non-Empty Features). For features that are currently non-empty (i.e., $f_{ik} = 1$ for at least one value of i), the feature assignments are updated using a standard Gibbs sampler: the value of f_{ik} is drawn from the conditional posterior distribution over $f_{ik}|\mathbf{S}, \mathbf{F}_{-ik}, \mathbf{w}$. Since feature assignments are discrete, it is easy to find this conditional probability by noting that

$$p(f_{ik}|\mathbf{S}, \mathbf{F}_{-ik}, \mathbf{w}) \propto p(\mathbf{S}|\mathbf{F}, \mathbf{w})p(f_{ik}|\mathbf{F}_{-ik}), \tag{11}$$

where $\mathbf{F}_{-ik}$ denotes the set of all feature assignments except f_{ik}. The first term in this expression is just the likelihood function for the ADCLUS model, and is simple to calculate. Moreover, since feature assignments in the IBP are exchangeable (see Griffiths & Ghahramani, 2005, for details), we can treat the kth assignment as if it were the last. Given this, Equation 6 indicates that $p(f_{ik}|\mathbf{F}_{-ik}) = n_{-ik}/n$, where n_{-ik} counts the number of stimuli (besides the ith) that currently possess the kth feature. Notice that this Gibbs sampler *does* apply to "singleton" features (i.e., those for which $f_{ik} = 1$ for *exactly* one value of i). However, when resampling the value of the feature assignment that is currently 1, that value will always be set to 0, since in this case $n_{-ik} = 0$. In effect, any time the Gibbs sampler runs into a singleton feature, it automatically converts it to an empty feature, adding it to the infinite collection of empty features.

Empty Features. We now turn to how the sampler deals with the infinite set of currently-empty features. Since the IBP describes a prior over infinite feature matrices, the resampling procedure really does need to accommodate the remaining (infinite) set of features that are not currently represented among the manifest (i.e., non-empty) features.

At a conceptual level it is helpful to note that if we were to redraw all these values simultaneously, some finite number of those currently-latent features would become manifest. In our MCMC procedure, we sample from the conditional posterior over feature assignments for the ith stimulus, holding all assignments fixed for the non-empty features, and also keeping the infinite set of (empty) feature assignments fixed for the other items. Now, in this case, if there were no data observed, this would be exactly equivalent to drawing the "new" features from the IBP. That is, we would introduce some number of "singleton" features possessed only by the ith object, where (due to exchangeability) this number is always drawn from a Poisson(α/n) distribution as noted previously. Fortunately, singleton features do not affect the probability of the data $\mathbf{S}$ in the additive clustering model, so the conditional posterior is exactly equivalent to the prior. In short, a very simple way to redraw for the infinite set of empty features is to sample Poisson(α/n) new singleton features for every stimulus.

When working with this algorithm, we typically run several chains. For each chain, we initialize the Gibbs-Metropolis sampler more or less arbitrarily. After a "burn-in" period is allowed for the sampler to converge to a sensible location (i.e., for the state to represent a sample from the posterior), we make a "draw" by recording the state of the sampler, leaving a "lag" of several iterations between successive draws to reduce the autocorrelation between samples. When doing so, it is important to ensure that the Markov chains converge on the target distribution $p(\mathbf{F}, \mathbf{w}|\mathbf{S})$. We did so by inspecting the time series plot formed by graphing probability of successive samples. To illustrate this, one of the chains used in our simulations (see Section 5) is displayed in Figure 6, with nine parallel chains used for comparison: the time series plot shows no long-term trends, and that different chains are visually indistinguishable from one another. Although elaborations and refinements are possible for both the sampler (Chen, Shao, & Ibrahim, 2000) and the convergence check (Cowles & Carlin, 1996), we have found this approach to be reasonably effective for the moderate-sized problems considered in our applications.

Four Estimators for the ADCLUS Model

One advantage to the IBP-ADCLUS approach is that it allows us to discuss a range of different estimators within a single framework. This is particularly useful since, in the absence of any explicit discussion of estimators, authors have adopted several different methods, evidently with little recognition of the fact. The difficulty stems from the fact that there are several slightly different questions that one might wish to answer, and authors vary in the importance they attach to each. For instance,

1. How should features be assigned to stimuli?
2. How much importance attaches to each feature?
3. How many features should be included in the model?
4. What predictions should be made about the similarities?
5. What is the probability that a particular feature is represented?

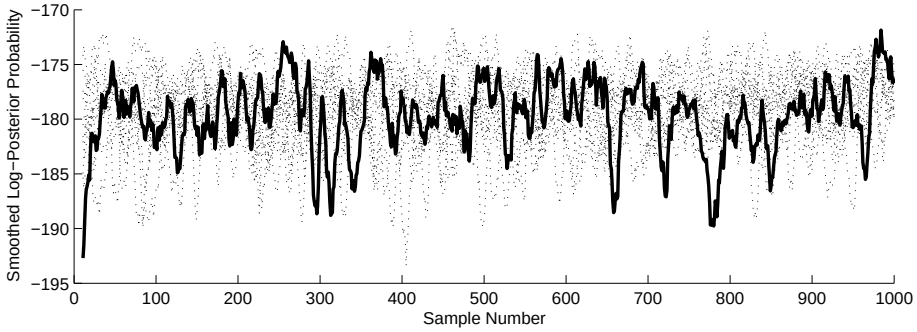


Figure 6: Smoothed time series showing log-posterior probabilities for successive draws from the Gibbs-Metropolis sampler, for simulated similarity data with $n = 16$. The bold line shows a single chain, while the dotted lines show the remaining nine chains. Note that, strictly speaking, since the normalizing contant for the posterior probability is unknown, the quantity being plotted differs from the actual log-probabilities by a constant.

These are by no means the only questions that one could ask. For instance, question 5 is essentially about obtaining measures of uncertainty associated with the feature set $\mathbf{F}$. Accordingly, we could also ask for uncertainty estimates for the feature weights and the dimensionality of the model. However, we omit these questions for now, since all existing estimators can accommodate those in a natural fashion.

In this section we outline four different estimators that one might employ when using the additive clustering model, and show how each of the four relates to these different questions. As indicated, the key difference between the estimators is not the underlying model, nor the choice of utility function (though there are disagreements regarding these). The main issue here regards which parameters are "nuisance parameters" and which are "parameters of interest". In the Bayesian framework, if the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\phi}, \boldsymbol{\lambda})$ can be partitioned into a set of interesting parameters $\boldsymbol{\phi}$ and a set of nuisance parameters $\boldsymbol{\lambda}$, it is typical to draw inferences about the interesting parameters using the marginal posterior:

$$p(\boldsymbol{\phi}|x) = \int p(\boldsymbol{\theta}|x) d\boldsymbol{\lambda} = \int p(\boldsymbol{\phi}, \boldsymbol{\lambda}|x) d\boldsymbol{\lambda} \tag{12}$$

(see Bernardo & Smith, 2000, p. 245). In the event that an estimate $\hat{\boldsymbol{\phi}}$ has already been obtained for the interesting parameters, it can be convenient to report estimates of the nuisance parameters conditional on the estimated values of the interesting parameters, using the conditional posterior distribution $p(\boldsymbol{\lambda}|\hat{\boldsymbol{\phi}}, x)$. This is often the case for the additive clustering model.

Selecting the Posterior Mode

The simplest approach to estimation in the additive clustering model is to report the posterior mode. This *maximum a posteriori* (MAP) estimator gives the single most likely

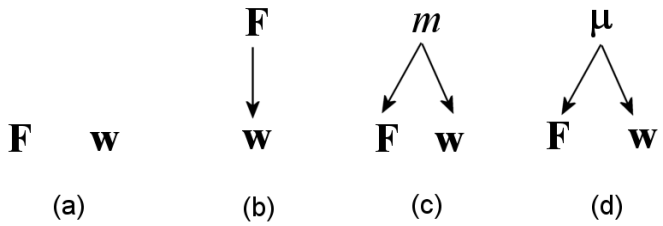


Figure 7: The implied "hierarchies" used to constrain the manner in which different estimators are used. In all cases one (often implicitly) marginalizes over the variables listed in at the bottom to estimate the variable listed at the top. Having done so, one reports the best values of the lower-level variables conditional on the top-level variable. In the simplest case, one can directly estimate $\mathbf{F}$ and $\mathbf{w}$, as illustrated in panel a. In MAP feature estimation (panel b), the feature set $\mathbf{F}$ is estimated first (marginalizing over $\mathbf{w}$), and then saliencies estimated conditional on those features. Alternatively, in conditional MAP estimation (panel c), the model order m is estimated first, and then the features $\mathbf{F}$ and saliencies $\mathbf{w}$ are reported conditional on model order. Finally, one could infer denoised (i.e., posterior predictive) similarity matrix $\boldsymbol{\mu}$ by marginalizing over $\mathbf{F}$ and $\mathbf{w}$, and then report features and saliencies that contribute to this prediction (panel d).

set of parameter values, and is hence optimal under a 0-1 loss function. The MAP estimate is obtained by selecting the the feature matrix $\hat{\mathbf{F}}_1$ and saliency vector $\hat{\mathbf{w}}_1$ such that,

$$\hat{\mathbf{F}}_1, \hat{\mathbf{w}}_1 = \arg\max_{\mathbf{F},\mathbf{w}} p(\mathbf{F}, \mathbf{w}|\mathbf{S}). \tag{13}$$

However, the MAP estimator is not generally used in additive clustering literature. The main reason for this is that the saturated model that contains a single two-stimulus feature for every cell in the similarity matrix can obtain perfect data fit by a judicious choice of weights. Accordingly, the only way to avoid this representation being chosen every time is to emphasize parsimony in either the prior or the loss function. In view of this obvious difficulty, the simple MAP estimator is almost never used. However, since many theoretically-motivated priors (including the IBP) allow the researcher to emphasize parsimony, it is possible to use this approach so long as one chooses the prior with care. Interestingly, this approach is paralleled in the frequentist literature: in order to produce parsimonious representations, some frequentist methods used in ADCLUS-like models apply penalty functions when obtaining a maximum likelihood feature model (Frank & Heiser, in press).

When assessed in terms of the questions listed earlier, the simple MAP estimator gives priority to questions 1 and 2, since it is explicitly constructed to provide the best possible estimates (under 0-1 loss) for the features and weights. In contrast, it provides answers to the remaining questions only indirectly. The number of features $\hat{m}_1$ is implied by the feature matrix, and similarity predictions $\hat{\mathbf{S}}_1$ could be constructed using the reported features and saliencies. However, while some authors report measures of the uncertainty

associated with the modal saliencies and number of features (e.g., Frank & Heiser, in press), it is not usual to report uncertainty measures for the specific features.

Selecting the Most Likely Feature Set

A rather different approach has been adopted by other authors (Lee, 2002b; Navarro, 2003), based on the typical application of the additive clustering model. The idea is that the saliency weights are of less psychological importance than the features to which they are applied, since saliencies often reflect context-specific levels of attention and are subject to frequent change. In contrast, the feature assignments themselves are usually treated as fixed properties of the stimulus representation (e.g., Lee & Navarro, 2002). Accordingly, the feature matrix $\mathbf{F}$ is taken to define a model with parameters $\mathbf{w}$, as illustrated by the implied hierarchy in Figure 7b. Under this view, the saliencies are treated as nuisance parameters, on the assumption that the psychologically important question is the identification of the feature matrix. Again based on a 0-1 loss function, the estimator $\hat{\mathbf{F}}_2$ for the feature matrix is then taken to be the marginal MAP feature set:

$$\hat{\mathbf{F}}_2 = \arg\max_{\mathbf{F}} p(\mathbf{F}|\mathbf{S}) = \arg\max_{\mathbf{F}} \left[\int p(\mathbf{F}, \mathbf{w}|\mathbf{S}) d\mathbf{w} \right]. \tag{14}$$

In general, the integral in Equation 14 is not tractable, so previous applications (e.g., Lee, 2002b; Navarro, 2003) have employed asymptotic approximations to $p(\mathbf{F}|\mathbf{S})$ for the sake of expediency, based on the Bayesian information criterion (Schwarz, 1978), Laplacian approximations (de Bruijn, 1958) or other related methods such as geometric complexity (Balasubramanian, 1997) and minimum description length (Rissanen, 1996). Given the primacy of the feature matrix, it is then sensible to estimate the saliencies conditional on this estimate. The corresponding MAP estimator is thus,

$$\hat{\mathbf{w}}_2 = \arg\max_{\mathbf{w}} p(\mathbf{w}|\hat{\mathbf{F}}_2, \mathbf{S}). \tag{15}$$

In this approach, the number of features $\hat{m}_2$ emerges as a consequence of the feature selection process, and similarity predictions $\hat{\mathbf{S}}_2$ would be made using the estimated features. Thus, like the simpler MAP estimator outlined above, this approach considers questions 3 (dimensionality), 4 (prediction) and 5 (model uncertainty) to be of secondary importance. However, it differs in that it treats quesion 1 (feature selection) as being considerably more important than question 2 (feature weighting).

Dimensionality Estimation and Conditional Posterior Modes

The most common approach in the additive clustering literature is to divide the estimation problem into a model selection problem and a parameter estimation problem (Tenenbaum, 1996; Arabie & Carroll, 1980; Ruml, 2001). In this approach, the number of features m is taken to define a particular model, with parameters corresponding to the $n \times m$ binary matrix $\mathbf{F}$ and the $m + 1$ length vector $\mathbf{w}$. For the purposes of parameter estimation, neither $\mathbf{F}$ nor $\mathbf{w}$ are considered to be nuisance parameters, and so are estimated jointly.

Again, assuming that we have some fixed value for m, we select the conditonally *maximum a posterior* (MAP) parameter values, given by

$$\hat{\mathbf{F}}_3, \hat{\mathbf{w}}_3 = \arg\max_{\mathbf{F},\mathbf{w}} p(\mathbf{F}, \mathbf{w}|\mathbf{S}, m) \tag{16}$$

In practice, m is rarely if ever known in advance. As a result, we need to solve a model selection problem in order to arrive at some estimate $\hat{m}_3$. When solving this problem, the psychologically-interesting variables $\mathbf{F}$ and $\mathbf{w}$ actually become the nuisance parameters. Accordingly, under 0-1 loss the (generally implicit) formulation of the model selection problem becomes,

$$\hat{m}_3 = \arg\max_m p(m|\mathbf{S}) = \arg\max_m \left[\sum_{\mathbf{F}\in\mathcal{F}_m} \int p(\mathbf{F}, \mathbf{w}|\mathbf{S}) \; d\mathbf{w} \right]. \tag{17}$$

In this expression, $\mathcal{F}_m$ denotes the set of feature matrices containing m unique features. The logic is that if we are to treat this as a model order selection problem, then the dimension m defines a model that had parameters $\mathbf{F}$ and $\mathbf{w}$. Ideally, then, we would choose the most likely (or maximum utility) model $\hat{m}_3$ by integrating over the parameters of that model. Then, having estimated this model, we would report parameter values $\hat{\mathbf{F}}_3, \hat{\mathbf{w}}_3$ conditional on this model. In practice, given the difficulty of working with Equation 17, it is typical to fix m on the basis of intuition, or via some heuristic method. Either way, the key point is that there is an implied "estimation hierarchy" among the different variables, since we implicitly marginalize $\mathbf{F}$ and $\mathbf{w}$ when estimating m, but condition on m when estimating $\mathbf{F}$ and $\mathbf{w}$. This is illustrated in Figure 7(c). In terms of the four questions listed earlier, the model selection problem treats dimensionality (question 3) as the only question of interest, while the parameter estimation problem treats feature selection (question 1) and saliency (question 2) as equally important, given that the model selection problem has been solved. Again, the prediction (question 4) and feature uncertainty (question 5) are considered ancillary.

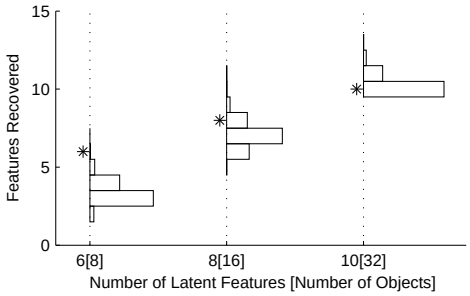
Approximating the Posterior Predictive Similarity Matrix

The three methods discussed above have all been applied to some extent in the existing literature. However, given that none are designed explicitly to address the questions of prediction and feature uncertainty, we now suggest a fourth possibility that complements the existing three. In this approach, we seek to uncover a small set of features that best approximates the posterior predictive similarity matrix. Letting $\hat{r}_k = p(\mathbf{f}_k|\mathbf{S})$ denote the posterior probability that feature $\mathbf{f}_k$ is manifest, we obtain

$$\hat{r}_k = p(\mathbf{f}_k|\mathbf{S}) = \sum_{\mathbf{F}:\mathbf{f}_k\in\mathbf{F}} p(\mathbf{F}|\mathbf{S}). \tag{18}$$

This allows us to construct a vector $\hat{\mathbf{r}} = [\hat{r}_k]$ that contains these probabilities for all 2^n possible features. Although this vector discards the covariation between features across

(a)



(b)

n		$\hat{\mathbf{S}}_1$	$\hat{\mathbf{S}}_2$	$\hat{\mathbf{S}}_3$	$\hat{\mathbf{S}}_4$
8	$\mathbf{S}_o$	79	81	79	84
	$\mathbf{S}_n$	78	81	78	84
	$\mathbf{S}_t$	87	88	87	92
16	$\mathbf{S}_o$	89	88	89	90
	$\mathbf{S}_n$	90	88	90	90
	$\mathbf{S}_t$	96	95	96	97
32	$\mathbf{S}_o$	91	91	91	91
	$\mathbf{S}_n$	91	91	91	91
	$\mathbf{S}_t$	100	100	100	100

Figure 8: Posterior distributions (a) over the number of features $p(m|\mathbf{S}_o)$ in simulations containing $m_t = 6$, 8 and 10 features respectively. Variance accounted for (b) by the four similarity estimators $\hat{\mathbf{S}}$, where the target is either the observed training data $\mathbf{S}_o$, a new test data set $\mathbf{S}_n$, or the true similarity matrix $\mathbf{S}_t$.

the posterior distribution, it is very useful, since the expected posterior similarities can be written as follows:

$$\hat{s}_{ij} = E[{s_{ij}}^*|\mathbf{S}] \quad = \quad \sum_{\mathbf{f}_k} f_{ik} f_{jk} \hat{r}_k \hat{w}_k, \tag{19}$$

where $\hat{w}_k = E\left[w_k|\mathbf{f}_k, \mathbf{S}\right]$ denotes the expected saliency for feature $\mathbf{f}_k$ on those occasions when it is represented. Equation 19 relies on the fact that features combine linearly in the ADCLUS model, and is straightforward to derive.

In practice, it is impossible to report all 2^n features, so although Equation 19 provides the estimate $\hat{\mathbf{S}}_4$ for the predicted similarities, one would typically report only those features that make the most substantial contributions to this estimate. While there are obviously several ways that we can formalize the notion of contribution, for the current purposes it will suffice to select those features for which $\hat{r}_k \hat{w}_k$ is largest. Similarly, there are several ways to determine the number of features to report, but for now we will simply ensure that the number of features reported for this method are in line with those suggested by the three existing methods.

Obviously, this approach treats prediction (question 4) as primary, but it is worth noting that since it involves the calculation of $\hat{r}_k$, the probability that feature $\mathbf{f}_k$ should be included in the representation, it gives some fairly explicit consideration to the question of feature uncertainty (question 5). This can be theoretically useful, since the concept of feature uncertainty is implicit in more general discussions of mental representation (Medin & Ortony, 1989) that ask whether or not a specific predicate is likely to be represented. However, unlike the other three estimators, it treats feature selection, feature saliency and dimensionality estimation (questions 1–3) as secondary.

Recovering Noisy Feature Matrices

By using the IBP-ADCLUS framework, we can compare the performance of the four estimators in a reasonable fashion. Loosely following Ruml (2001), we generated noisy similarity matrices with $n = 8$, 16 and 32 stimuli, based on "true" feature matrices $\mathbf{F}_t$ in which $m_t = 2\log_2(n)$, where each object possessed each feature with probability 0.5. Saliency weights $\mathbf{w}_t$ were generated uniformly from the interval $[1, 3]$, but were subsequently rescaled to ensure that the "true" similarities $\mathbf{S}_t$ had variance 1. Two sets of Gaussian noise were injected into the similarities with fixed $\sigma = 0.3$, ensuring that the noise accounted for approximately 9% of the variance in the "observed" data matrix $\mathbf{S}_o$ and the "new" matrix $\mathbf{S}_n$. We fixed $\alpha = 2$ for all simulations: since the number of manifest features in an IBP model follows a Poisson(αH_n) distribution (Griffiths & Ghahramani, 2005), the prior has a strong bias toward parsimony, since the prior expected number of features is approximately 5.4, 6.8 and 8.1 (as compared to the true values of 6, 8 and 10).

We approximated the posterior distribution $p(\mathbf{F}, \mathbf{w}|\mathbf{S}_1)$, by drawing samples in the following manner. For a given similarity matrix, 10 Gibbs-Metropolis chains were run from different start points, and 1000 samples were drawn from each. The chains were burnt in for 1000 iterations, and a lag of 10 iterations was used between successive samples. Visual inspection suggested that five chains in the $n = 32$ condition did not converge: log-posteriors were low, differed substantially from one another, and had noticable positive slope. In this case, the estimators were constructed from the five remaining chains.

Figure 8(a) shows the posterior distributions over the number of features m for each of the three simulation conditions. There is a tendency to underestimate the number of features when provided with small similarity matrices, with the modal number being 3, 7 and 10. However, since the posterior estimate of m is below the prior estimate when $n = 8$, it seems this effect is data-driven, as 79% of the variance in the data matrix $\mathbf{S}_o$ can be accounted for using only three features.

Since each approach allows the construction of an estimated similarity matrix $\hat{\mathbf{S}}$, a natural comparison is to look at the proportion of variance this estimate accounts for in the observed data $\mathbf{S}_o$, the novel data set $\mathbf{S}_n$, and the true matrix $\mathbf{S}_t$. In view of the noise model used to construct these matrices, the "ideal" answer for these three should be around 91%, 91% and 100% respectively. When $n = 32$, this profile is observed for all four estimators, suggesting that in this case all four estimators have converged appropriately. For the smaller matrices, the joint MAP and conditional MAP estimators ($\hat{\mathbf{S}}_1$ and $\hat{\mathbf{S}}_3$) behave similarly. The MAP feature approach $\hat{\mathbf{S}}_2$ appears to perform slightly better, though the difference is very small. The expectation method $\hat{\mathbf{S}}_4$ provides the best estimate.

Modeling Empirical Similarities

We now turn to the analysis of empirical data. To keep the presentation as brief as possible, we limit the discussion to the most novel IBP-ADCLUS estimators, namely the direct estimates of dimensionality provided through Equation 17, and the features

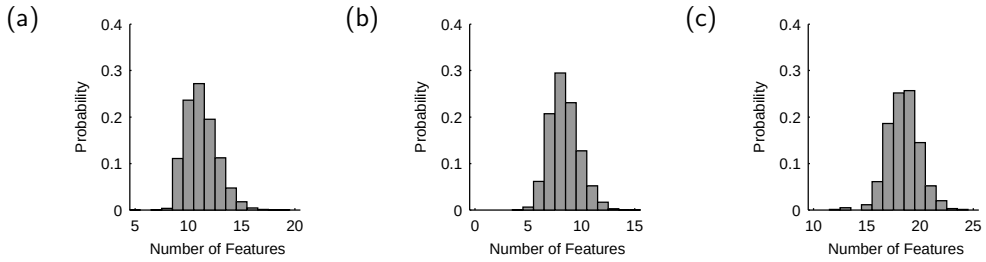


Figure 9: Posterior distributions over the number of features when the Bayesian ADCLUS model is applied to (a) the numbers data, (b) the countries data and (c) the letters data.

extracted via "approximate expectation".

Featural representations of numbers.

A standard data set used in evaluating additive clustering models measures the conceptual similarity of the numbers 0 through 9 (Shepard, Kilpatric, & Cunningham, 1975). This data set is often used as a benchmark due to the complex interrelationships between the numbers. Table 1(a) shows an eight-feature representation of these data, taken from Tenenbaum (1996) who applied a maximum likelihood approach. This representation explains 90.9% of the variance, with features corresponding to arithmetic concepts and to numerical magnitude. Fixing $\sigma = 0.05$, and $\alpha = 0.5$, we drew 10,000 lagged samples to construct estimates. Although the posterior probability is spread over a large number of feature matrices, 92.6% of sampled matrices had between 9 and 13 features. The modal number of represented features was $\hat{m}_3$=11, with 27.2% of the posterior mass. The posterior distribution over the number of features is shown in Figure 9(a). Since none of the existing literature has used the "approximate expectation" approach to find highly probable features, it is useful to note the strong similarities between Table 1(a) and Table 1(b), which reports the ten highest-probability features across the entire posterior distribution. Applying this approach to obtain an estimate of the posterior predictive similarities $\hat{\mathbf{S}}_4$ revealed that this matrix accounts for 97.4% of the variance in the data. Moreover, unlike the feature set reported by Tenenbaum (1996), even and odd numbers both appear as distinct features.

Featural representations of countries

A second application is to human forced-choice judgments of the similarities between 16 countries (Navarro & Lee, 2002). In this task, participants were shown lists of four countries and asked to pick out the two countries most similar to each other. Applying the Bayesian model to these data with the empirically-estimated value of $\sigma = 0.1$ reveals that only eight features appear in the representation more than 25% of the time. Given this, it is not surprising that the posterior distribution over the number of features, shown in Figure 9(b), indicates that the modal number of features is eight. The eight most probable

Table 1: Two representations of the numbers data. (a) The representation reported in (Tenenbaum, 1996), extracted using an EM algorithm with the number of features fixed at eight. (b) The 10 most probable features extracted using the Bayesian ADCLUS model. The first column gives the posterior probability that a particular feature belongs in the representation. The second column displays the average saliency of a feature in the event that it is included.

(a)

FEATURE	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
additive constant	0.148

(b)

FEATURE	PROB.	WEIGHT
3 6 9	0.79	0.326
2 4 8	0.70	0.385
0 1 2	0.69	0.266
2 3 4 5 6	0.59	0.240
6 7 8 9	0.57	0.262
0 1 2 3 4	0.42	0.173
2 4 6 8	0.41	0.387
1 3 5 7 9	0.40	0.223
4 5 6 7 8	0.34	0.181
7 8 9	0.26	0.293
additive constant	1.00	0.075

features are listed in Table 2. The "approximate expectation" method explains 85.4% of the variance, as compared to the 78.1% found by a MAP feature approach (Navarro & Lee, 2002). The features are interpretable, corresponding to a range of geographical, historical, and economic regularities. Moreover, while the features recovered are very similar to those found by Navarro and Lee (2002), a comparison of the saliency weights reported in that paper (the "NL-weight" row in Table 2) to the saliencies and inclusion probabilities found here (the "prob." and "weight" rows) reveals that there is considerable uncertainty associated with some features but not others, a fact that was not evident in the original paper. The first three features have $\hat{r}_k$ values that are very close to 1, whereas the other features may or may not form part of the underlying mental representation.

Featural Representations of Letters

To provide a contrast with the last data set, in which there is considerably uncertainty associated with several of the features, we analyzed a somewhat larger data set, consisting of kindergarten children's assessment of the perceptual similarity of the 26 capital letters (Gibson, Osser, Schiff, & Smith, 1963). In this case, we used $\sigma = 0.05$, and the Bayesian model accounted for 89.2% of the variance in the children's similarity judgments. The posterior distribution over the number of represented features is shown in Figure 9(c). Table 3 shows the ten features that appeared in more than 90% of samples from the posterior. The model recovers an extremely intuitive set of overlapping features. For

Table 2: Featural representation of the similarity between 16 countries. The table shows the eight highest-probability features extracted by the Bayesian ADCLUS model. Each column corresponds to a single feature, with the associated probabilities and saliencies shown below. The average weight associated with the additive constant is 0.035. The last line (NL-weight) lists the saliencies for the various features in the representation reported by Navarro and Lee (2002).

	FEATURE							
	Italy Germany Spain	Vietnam China Japan Philippines Indonesia	Germany Russia USA China Japan	Zimbabwe Nigeria	Zimbabwe Nigeria Cuba Jamaica Iraq Libya	Iraq Libya	Zimbabwe Nigeria Iraq Libya	Philippines Indonesia
PROB.	1.00	1.00	0.99	0.62	0.52	0.36	0.33	0.25
WEIGHT	0.593	0.421	0.267	0.467	0.209	0.373	0.299	0.311
NL-WEIGHT	0.641	0.371	0.262	0.742	-	0.613	-	0.414

example, it picks out the long strokes in I, L, and T, and the elliptical forms of D, O, and Q. Moreover, since we the estimation method is sensitive to the full variation in the posterior distribution, we are able to say with a very high degree of certainty that all 10 features should be included as part of the inferred representation.

Discussion

Learning how similarity relations are represented is a difficult modeling problem. Additive clustering provides a framework for learning featural representations of stimulus similarity, but remains underused due to the difficulties associated with the inference. By adopting a Bayesian approach to additive clustering, we are able to obtain a richer characterization of the structure behind human similarity judgments. Moreover, by using nonparametric Bayesian techniques to place a prior distribution over infinite binary feature matrices via the Indian buffet process, we can allow the data to determine the number of features that the algorithm recovers. This is theoretically important as well as pragmatically useful. As noted by Medin and Ortony (1989), people are capable of recognizing that individual stimuli possess an arbitrarily large number of characteristics, but in any particular context will make judgments using only a finite, usually small number of properties that form part of our current mental representation. In other words, by moving to a Bayesian nonparametric form, we are able to bring the ADCLUS model closer to the kinds of assumptions that are made by psychological theories.

Table 3: Featural representation of the perceptual similarity between 26 capital letters. The table shows the ten highest-probability features extracted by the Bayesian ADCLUS model. Each column corresponds to a single feature, with the associated probabilities and saliencies shown below. The average weight associated with the additive constant is 0.003.

	FEATURE									
	M	I	C	D	P	E	E	K	B	C
	N	L	G	O	R	F	H	X	G	J
	W	T		Q					R	U
PROB.	1.00	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.98	0.92
WEIGHT	0.686	0.341	0.623	0.321	0.465	0.653	0.322	0.427	0.226	0.225

A number of avenues for extending this work present themselves. At a statistical level, placing hyperpriors over α, σ and λ would seem to be a good idea, in order to infer their values from data. It would also be useful to switch to an extended version of the IBP, in which the implied prior over the number of features is decoupled from the implied prior over the number of objects that possess a particular feature. In order to scale the approach to larger domains, the MCMC algorithm could be augmented by using more advanced proposals (e.g., split-merge steps; S. Jain & Neal, 2004). Psychologically, the ability to perform reliable inference in the ADCLUS model suggests that the model itself be extended to cover some of the more recent theoretical ideas in similarity. For instance, it would allow explicit testing of the "size principle" (Tenenbaum & Griffiths, 2001), which proposes on theoretical grounds that small features should be more salient on average. The underlying model itself could be extended, allowing more complex, structured representations to be learned from data, in keeping with more recent theories for similarity and analogy (e.g., Gentner, 1983; Goldstone, 1994). In the meantime, however, the adoption of a nonparametric Bayesian approach goes a long way towards making additive clustering a more reliable technique.

Acknowledgments

DJN was supported by an Australian Research Fellowship (ARC grant DP-0773794). TLG was supported by a Junior Faculty Research Grant from the University of California, Berkeley. We thank Nancy Briggs, Simon Dennis and Michael Lee for helpful comments.

References

Aldous, D. (1985). Exchangeability and related topics. In *École d'été de probabilités de Saint-Flour, XIII—1983* (pp. 1–198). Berlin: Springer.

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

Attneave, F. (1950). Dimensions of similarity. *American Journal of Psychology*, *63*, 546-554.

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

Bernardo, J. M., & Smith, A. F. M. (2000). *Bayesian theory (2nd ed)*. New York: Wiley.

Blackwell, D., & MacQueen, J. (1973). Ferguson distributions via Polya urn schemes. *The Annals of Statistics*, *1*, 353-355.

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, *36*, 111-150.

Buneman, P. (1971). The recovery of trees from measures of dissimilarity. In F. R. Hodson, D. G. Kendall, & P. Tautu (Eds.), *Mathematics in the archaelogical and historical sciences* (p. 387-395). Edinburgh, UK: Edinburgh University Press.

Chen, M., Shao, Q., & Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Compu tation*. New York, NY: Springer.

Cole, A. J., & Wishart, D. (1970). An improved algorithm for the jardine-sibson method of generating overlapping clusters. *The Computer Journal*, *13*, 156-163.

Corter, J., & Tversky, A. (1986). Extended similarity trees. *Psychometrika*, *51*, 429-451.

Cowles, M. K., & Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, *91*, 833-904.

Cunningham, J. P. (1978). Free trees and bidirectional trees as representations of psychological distance. *Journal of Mathematical Psychology*, *17*, 165-188.

D'Andrade, R. (1978). U-statistic hierarchical clustering. *Psychometrika*, *4*, 58-67.

Dayhoff, M., Schwartz, R., & Orcutt, B. (1978). A model of evolutionary change in proteins. In M. O. Dayhoff (Ed.), *Atlas of protein sequence and structure 5(3)* (p. 345-352). Washington: National Biomedical Research Foundation.

de Bruijn, N. G. (1958). *Asymptotic methods in analysis*. Amsterdam: North-Holland.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B*, *39*.

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

Ekman, G. (1963). A direct method for multidimensional ratio scaling. *Psychometrika*, *28*, 33-41.

Escobar, M. D., & West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, *90*, 577-588.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, *1*, 209-230.

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *Annals of Statistics*, *2*, 615-629.

Frank, L. E., & Heiser, W. J. (in press). Feature selection in Feature Network Models: Finding predictive subsets of features with the Positive Lasso. *British Journal of Mathematical and Statistical Psychology*.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721-741.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*, 155-170.

Gibson, E. J., Osser, H., Schiff, W., & Smith, J. (1963). *An analysis of crictical features of letters, tested by a confusion matrix* (Tech. Rep. No. Cooperative Research Project No. 639).

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1995). *Markov chain monte carlo in practice*. London, UK: Chapman and Hall.

Goldstone, R. L. (1994). Similarity, interactive activation, and mapping. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 3-28.

Goodman, N. (1972). Seven strictures on similarity. In N. Goodman (Ed.), *Problems and projects*. New York: The Bobbs-Merrill Co.

Griffiths, T. L., & Ghahramani, Z. (2005). *Infinite latent feature models and the Indian buffet process* (Tech. Rep. No. 2005-001). Gatsby Computational Neuroscience Unit.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Applied Statistics*, *28*, 100-108.

Henikoff, S., & Henikoff, J. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences USA*, *89*, 10915-10919.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*, 417-441, 498-520.

Hutchinson, J. W., & Mungale, A. (1997). Pairwise partitioning: A nonmetric algorith, for identifying feature-based similarity structures. *Psychometrika*, *62*, 85-117.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys*, *31*, 264-323.

Jain, S., & Neal, R. M. (2004). A split-merge Markov chain Monte Carlo procedure for the Dirichlet Process mixture model. *Journal of Computational and Graphical Statistics*, *13*, 158-182.

Jardine, N., & Sibson, R. (1968). The construction of hierarchic and non-hierarchic classifications. *The Computer Journal*, *11*, 177-184.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, *32*, 241-254.

Komatsu, L. K. (1992). Recent views of conceptual structure. *Psychological Bulletin*, *112*(3), 500-526.

Kontkanen, P., Myllymäki, P., Buntine, W., Rissanen, J., & Tirri, H. (2005). An MDL framework for data clustering. In *Advances in minimum description length: Theory and applications*. Cambridge, MA: MIT Press.

Kraft, C. H. (1964). A class of distribution function processes which have derivatives.

Journal of Applied Probability, *1*, 385-388.

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

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

Kruskal, J. B. (1964b). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, *29*, 115-129.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211-240.

Lawley, D. N., & Maxwell, A. E. (1963). *Factor analysis as a statistical model*. London: Butterworths.

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

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

Lee, M. D. (2001). Determining the dimensionality of multidimensional scaling models for cognitive modeling. *Journal of Mathematical Psychology*, *45*, 149-166.

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

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

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

Lee, M. D., Pincombe, B. M., & Welsh, M. B. (2005). An empirical evaluation of models of text document similarity. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society* (p. 1254-1259). Mahwah, NJ: Lawrence Erlbaum.

McAdams, S., Winsberg, S., Donnadieu, S., de Soete, G., & Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. *Psychological Research*, *58*, 177-192.

McLachlan, G. J., & Basford, K. E. (1988). *Mixture models. inference and applications to clustering*. New York, NY: Dekker.

McQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth berkeley symposium on mathematical statistics and probability* (p. 281-297).

Medin, D. L., & Ortony, A. (1989). Psychological essentialism. In *Similarity and analogical reasoning*. New York: Cambridge University Press.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, *21*, 1087-1092.

Minkowski, H. (1891). [on positive quadratic forms and on algorithms suggesting continued fractions]. *Journal für die reine und angewandte Mathematik*, *107*, 278-297.

Mirkin, B. G. (1987). Additive clustering and qualitative factor analysis methods for

similarity matrices. *Journal of Classification*, *4*, 7-31.

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

Navarro, D. J. (2003). *Representing stimulus similarity.* Ph.D. Thesis, University of Adelaide.

Navarro, D. J., Griffiths, T. L., Steyvers, M., & Lee, M. D. (2006). Modeling individual differences using dirichlet processes. *Journal of Mathematical Psychology*, *50*, 101-122.

Navarro, D. J., & Lee, M. D. (2002). Commonalities and distinctions in featural stimulus representations. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society* (p. 685-690). Mahwah, NJ: Lawrence Erlbaum.

Navarro, D. J., & Lee, M. D. (2004). Common and distinctive features in stimulus representation: A modified version of the contrast model. *Psychonomic Bulletin and Review*, *11*, 961-974.

Neal, R. M. (2003). Density modeling and clustering using dirichlet diffusion trees. *Bayesian Statistics*, *7*, 619-629.

Pearson, K. (1901). On lines and planes of closest fit to a system of points in space. *Philosophical Magazine*, *2*, 559-572.

Pitman, J. (1996). Some developments of the blackwell-macqueen urn scheme. In T. F. et al (Ed.), *Statistics, probability and game theory: Papers in honor of david blackwell* (p. 245-267). Hayward, CA: Institute of Mathematical Studies.

Rasmussen, C. (2000). The infinite gaussian mixture model. In *Advances in Neural Information Processing Systems 12.* Cambridge, MA: MIT Press.

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

Rosenberg, S., & Kim, M. P. (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 *Advances in Neural Information Processing Systems 14.* Cambridge, MA: MIT Press.

Sattath, S., & Tversky, A. (1977). Additive similarity trees. *Psychometrika*, *42*, 319-345.

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

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, *4*, 639-650.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with and unknown distance function. i and ii. *Psychometrika*, *27*, 125-140, 219-246.

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

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

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

Sneath, P. H. A. (1957). The application of computers to taxonomy. *The Journal of General Microbiology*, *17*, 201-206.

Sokal, R. R. (1974). Classification: Purposes, principles, progress, prospects. *Science*, *185*, 1115-1123.

Sokal, R. R., & Sneath, P. H. A. (1963). *Principles of numerical taxonomy*. San Francisco: Freeman.

Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology*, *15*, 201-293.

Spearman, C. (1927). *The abilities of man*. New York: Macmillan.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*, 677-680.

Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. In *Handbook of experimental psychology*. New York: Wiley.

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

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*, 629-641.

Thurstone, L. L. (1947). *Multiple-factor analysis*. Chicago: University of Chicago Press.

Torgerson, W. S. (1958). *Theory and methods of scaling*. New York: Wiley.

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

Wolfe, J. H. (1970). Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research*, *5*, 329-350.

Young, G., & Householder, A. S. (1938). Discussion of a set of points in terms of their mutual distances. *Psychometrika*, *3*, 19-22.