Running head: CATEGORIZATION AS NONPARAMETRIC BAYESIAN DENSITY ESTIMATION

Categorization as nonparametric Bayesian density estimation

Thomas L. Griffiths

Department of Psychology

University of California, Berkeley

Adam N. Sanborn

Department of Psychological and Brain Sciences

Indiana University

Kevin R. Canini

Department of Computer Science

University of California, Berkeley

Daniel J. Navarro

School of Psychology

University of Adelaide

Categorization as nonparametric Bayesian density estimation

Rational models of cognition aim to explain the structure of human thought and behavior as an optimal solution to the computational problems that are posed by our environment (Anderson, 1990; Chater & Oaksford, 1999; Marr, 1982; Oaksford & Chater, 1998). Rational models have been developed for several aspects of cognition, including memory (Anderson, 1990; Shiffrin & Steyvers, 1997), reasoning (Oaksford & Chater, 1994), generalization (Shepard, 1987; Tenenbaum & Griffiths, 2001), and causal induction (Anderson, 1990; Griffiths & Tenenbaum, 2005). By examining the computational problems that underlie our cognitive capacities, it is often possible to gain a deeper understanding of the assumptions behind successful models of human cognition, and to discover new classes of models that might otherwise have been overlooked.

In this chapter, we pursue a rational analysis of *category learning*: inferring the structure of categories from a set of stimuli labeled as belonging to those categories. The knowledge acquired through this process can ultimately be used to make decisions about how to categorize new stimuli. Several rational analyses of category learning have been proposed (Anderson, 1990; Nosofsky, 1998; Ashby & Alfonso-Reese, 1995). These analyses essentially agree on the nature of the computational problem involved, casting category learning as a problem of *density estimation*: determining the probability distributions associated with different category labels. Viewing category learning in this way helps to clarify the assumptions behind the two main classes of psychological models: exemplar models and prototype models. Exemplar models assume that a category is represented by a set of stored exemplars, and categorizing new stimuli involves comparing these stimuli to the set of exemplars in each category (e.g., Medin & Schaffer, 1978; Nosofsky, 1986). Prototype models assume that a category is associated with a single prototype and categorization involves comparing new stimuli to these prototypes (e.g., Reed, 1972).

These approaches to category learning correspond to different strategies for density estimation used in statistics, being nonparametric and parametric density estimation respectively (Ashby & Alfonso-Reese, 1995).

Despite providing insight into the assumptions behind models of categorization, existing rational analyses of category learning leave a number of questions open. One particularly important question is whether rational learners should use an exemplar or prototype representation. The greater flexibility of nonparametric density estimation has motivated the claim that exemplar models are to be preferred as rational models of category learning (Nosofsky, 1998). However, nonparametric and parametric methods have different advantages and disadvantages: the greater flexibility of nonparametric methods comes at the cost of requiring more data to estimate a distribution. The choice of representation scheme should ultimately be determined by the stimuli presented to the learner, and existing rational analyses do not indicate how this decision should be made (although see Briscoe & Feldman, 2006). This question is complicated by the fact that prototype and exemplar models are not the only options. A number of models have recently explored possibilities between these extremes, representing categories using clusters of several exemplars (Anderson, 1990; Kruschke, 1990; Love, Medin, & Gureckis, 2004; Rosseel, 2002; Vanpaemel, Storms, & Ons, 2005). The range of representations possible in these models emphasizes the significance of being able to identify an appropriate category representation from the stimuli themselves: with many representational options available, it is even more important to be able to say which option a learner should choose.

Anderson's (1990, 1991) rational analysis of categorization presents a partial solution to this question, automatically selecting the number of clusters to be used in representing a set of objects, but has its own limitations. Anderson's approach uses a flexible representation in which new clusters are added as required. When a new stimulus

is observed, it can either be assigned to one of the pre-existing clusters, or to a new cluster of its own. As a result, the representation becomes more complex as new data are observed, with the number of clusters growing as needed to accommodate the rich structures that emerge as we learn more about our environment. Accordingly, a crucial aspect of the model is the method by which stimuli are assigned to clusters. Anderson (1990, 1991) proposed an algorithm in which stimuli are sequentially assigned to clusters, and assignments of stimuli are fixed once they are made. However, this algorithm does not provide any asymptotic guarantees for the quality of the resulting assignments, and is extremely sensitive to the order in which stimuli are observed, a property which is not intrinsic to the underlying statistical model.

In this chapter, we identify connections between existing rational models of categorization and work on density estimation in nonparametric Bayesian statistics. These connections have two consequences. First, we present two new algorithms that can be used in evaluating the predictions of Anderson's (1990, 1991) rational model of categorization. These two algorithms both asymptotically approximate the Bayesian posterior distribution over assignments of objects to clusters, and help to separate the predictions that arise from the underlying statistical model from those that are due to the inference algorithm. These algorithms also provide a source of hypotheses about the processes by which people could solve the challenging problem of performing probabilistic inference. Second, we develop a unifying model of categorization, of which existing rational models are special cases. This model goes beyond previous unifying models of category learning (e.g., Rosseel, 2002; Vanpaemel et al., 2005) by providing a rational solution to the question of which representation should be chosen, and when the representation should change, based purely on the information provided by the stimuli themselves.

Identifying the connection between models of human category learning and nonparametric Bayesian density estimation extends the scope of the rational analysis of

category learning. It also provides a different perspective on human category learning. Rather than suggesting that people use one form of representation or another, our approach indicates how it might be possible (and, in fact, desirable) for people to switch between representations based upon the structure of the stimuli they observe. This basic idea is similar to that underlying recent process models of category learning, such as SUSTAIN (Love et al., 2004). Our contribution is a rational account of when a given representation is justified by the data given a set of assumptions about the processes by which those data are produced, providing a way to explore the assumptions that underlie human category learning. We illustrate this approach by modeling data from Smith and Minda (1998), in which people seem to shift from using a prototype representation early in training to using an exemplar representation late in training, showing that such a shift can be understood as a rational statistical inference.

The plan of the chapter is as follows. The next section summarizes exemplar and prototype models, and the idea of interpolating between the two. We then discuss existing rational models of categorization, before going on to highlight the connection between the rational model proposed by Anderson (1990, 1991) and the Dirichlet process mixture model (Antoniak, 1974; Ferguson, 1983; Neal, 1998), a statistical model that is commonly used in nonparametric Bayesian statistics. This allows us to identify two new algorithms for use with Anderson's model, which we describe and evaluate, and to use generalizations of this statistical model as the basis for a more complete account of human categorization. We summarize the ideas behind the hierarchical Dirichlet process (Teh, Jordan, Beal, & Blei, 2004), and use it as the foundation for a unifying rational model of categorization. Finally, we show that this model can capture the shift from prototypes to exemplars in the data of Smith and Minda (1998).

Similarity-based models of categorization

While early work assumed that people use explicit classification rules in order to assign stimuli to categories (e.g., Bruner, Goodnow, & Austin, 1956), most categorization models developed in the last 30 years have assumed that categories are defined by a kind of "family resemblance" (e.g., Rosch, 1978). The two most influential approaches have been prototype models and exemplar models, which both assume that people assign stimuli to categories based on similarity, formalized in the following manner. Given a set of $N-1$ stimuli with features $\mathbf{x}_{N-1} = (x_1, x_2, \ldots, x_{N-1})$ and category labels $\mathbf{y}_{N-1} = (y_1, y_2, \ldots, y_{N-1})$, the probability that stimulus N with features x_N is assigned to category j is given by

$$P(y_N = j | x_N, \mathbf{x}_{N-1}, \mathbf{y}_{N-1}) = \frac{\eta_{N,j}\beta_j}{\sum_y \eta_{N,y}\beta_y} \tag{1}$$

where $\eta_{N,y}$ is the similarity of the stimulus x_N to category y and β_y is the response bias for category y. Thus, the decision is a function of the various category similarities, and involves a straightforward application of the standard choice rule (Luce, 1959). The key difference between the models is in how $\eta_{N,j}$, the similarity of a stimulus to a category, is computed.

Exemplars and prototypes

In an exemplar model (e.g., Medin & Schaffer, 1978; Nosofsky, 1986), all of the instances of that category are stored. The similarity of stimulus N to category j is calculated by summing the similarity of the stimulus to all these stored instances. That is,

$$\eta_{N,j} = \sum_{i|y_i=j} s_{N,i} \tag{2}$$

where $s_{N,i}$ is a symmetric measure of the similarity between the two stimuli x_N and x_i. The similarity measure is typically defined as a decaying exponential function of the distance between the two stimuli, following Shepard (1987). An example of the overall similarity function is shown in the rightmost panel of Figure 1. In contrast, prototype models (e.g., Reed, 1972), represent a category j in terms of a single prototypical instance. In this formulation, the similarity of stimulus N to category j is defined to be,

$$\eta_{N,j} = s_{N,p_j} \tag{3}$$

where p_j is the prototypical instance of the category and s_{N,p_j} is a measure of the similarity between stimulus N and the prototype p_j. One common way of defining the prototype is as the centroid of all instances of the category in some psychological space, i.e.,

$$p_j = \frac{1}{N_j} \sum_{i|y_i=j} x_i \tag{4}$$

where N_j is the number of instances of the category (i.e., the number of stimuli for which $y_i = j$). The panel on the left of Figure 1 illustrates the kind of category similarity functions employed by a prototype model.

Broader classes of representation

Although exemplars and prototypes have dominated the modern literature, a number of authors (e.g., Kruschke, 1990; Love et al., 2004; Vanpaemel et al., 2005) have proposed more general classes of category representation that interpolate between prototype and exemplar models. For example, Vanpaemel et al. (2005) formalized a set of interpolating models by partitioning instances of each category into clusters, where the number of clusters K_j ranges from 1 to N_j. Then each cluster is represented by a prototype, and the

similarity of stimulus N to category j is defined to be,

$$\eta_{N,j} = \sum_{k=1}^{K_j} s_{N,p_{j,k}} \tag{5}$$

where $p_{j,k}$ is the prototype of cluster k in category j. This is equivalent to the prototype model when $K_j = 1$, and the exemplar model when $K_j = N_j$. Thus, this generalized model, the Varying Abstraction Model (VAM), is more flexible than both the exemplar and prototype models (as illustrated by the middle panel of Figure 1), although it raises the problem of estimating which clustering people use in any particular categorization task (for details, see Vanpaemel et al., 2005).

The idea of representing a category using a set of clusters is reasonably intuitive, since explicitly labeled categories are not the only level at which homogeneity can be found in the world (Rosch, 1978). For example, while no two chairs are exactly the same, many chairs are of similar types, differing only in superficial properties like color. By clustering the instances of these similar types of chairs and storing a single prototype, we can avoid having to remember a large number of redundant instances. A similar property holds for natural categories, where, for example, species of animals might be composed of subspecies. This underlying structure supports a finer-grained representation than a single prototype, while not requiring the comprehensiveness of a full exemplar model.

Rational accounts of categorization

The models discussed in the previous section all explain categorization behavior in terms of cognitive processes, in particular similarity and choice. An alternative approach is to seek an explanation based on the form of the computational problem that underlies categorization. Following the methodology outlined by Anderson (1990), rational models of categorization explain human behavior as an adaptive solution to a computational

problem posed by the environment, rather than focusing on the cognitive processes involved. Existing analyses tend to agree that the basic problem is one of *prediction* – identifying the category label or some other unobserved property of an object using its observed properties (Anderson, 1990; Ashby & Alfonso-Reese, 1995; Rosseel, 2002). This prediction problem has a natural interpretation as a form of Bayesian inference. In a standard classification task, for instance, Bayes' rule allows us to compute the probability that object N belongs to category j given the features and category labels of $N-1$ objects:

$$P(y_N = j|x_N, \mathbf{x}_{N-1}, \mathbf{y}_{N-1}) = \frac{P(x_N|y_N = j, \mathbf{x}_{N-1}, \mathbf{y}_{N-1})P(y_N = j|\mathbf{y}_{N-1})}{\sum_y P(x_N|y_N = y, \mathbf{x}_{N-1}, \mathbf{y}_{N-1})P(y_N = y|\mathbf{y}_{N-1})}. \quad (6)$$

where we assume that the prior probability of an object coming from a particular category is independent of the features of the previous objects. In this expression, the posterior probability of category j is related to both the probability of sampling an object with features x_N from that category, and the prior probability of choosing that category. Category learning, then, becomes a matter of determining these probabilities – a problem known as *density estimation*. Since different rational models vary in how they approach this problem, we provide a brief overview of the various accounts.

The rational basis of exemplar and prototype models

Ashby and Alfonso-Reese (1995) observed that both prototype and exemplar models can be recast as rational solutions to the problem of categorization, highlighting the connection between the Bayesian solution presented in Equation 6 and the choice probabilities in the exemplar and prototype models (i.e., Equation 1). Specifically, the category similarity $\eta_{N,j}$ can be identified with the probability of generating an item, $P(x_N|y_N = j, \mathbf{x}_{N-1}, \mathbf{y}_{N-1})$, while the category bias β_j corresponds naturally to the prior probability of category j, $P(y_N = j|\mathbf{y}_{N-1})$. The difference between exemplar and

prototype models is thus the different ways of estimating $P(x_N|y_N = j, \mathbf{x}_{N-1}, \mathbf{y}_{N-1})$. The definition of $\eta_{N,j}$ used in an exemplar model (Equation 2) corresponds to estimating $P(x_N|y_n = j, \mathbf{x}_{N-1}, \mathbf{y}_{N-1})$ as the sum of a set of functions (known as "kernels") centered on the x_i already labeled as belonging to category j, with

$$P(x_N|y_N = j, \mathbf{x}_{N-1}, \mathbf{y}_{N-1}) \propto \sum_{i|y_i=j} f(x_N, x_i) \tag{7}$$

where $f(x, x_i)$ is a probability distribution centered on x_i.[1] This method is widely used for approximating distributions in statistics, being a simple form of nonparametric density estimation called kernel density estimation (e.g., Silverman, 1986). In contrast, the definition of $\eta_{N,j}$ used in a prototype model (Equation 3) corresponds to estimating $P(x_N|y_n = j, \mathbf{x}_{N-1}, \mathbf{y}_{N-1})$ by assuming that each category distribution comes from an underlying parametric family and then finding the parameters that best characterize the instances labeled as belonging to that category. The prototype is specified by these parameters, with the centroid being an appropriate estimate for distributions whose parameters characterize their mean. Again, this is a common method for estimating a probability distribution, known as parametric density estimation, in which the distribution is assumed to be of a known form but with unknown parameters (e.g., Rice, 1995).

The Mixture Model of Categorization

Casting exemplar and prototype models as different schemes for density estimation suggests that a similar interpretation might be found for interpolating models. Rosseel (2002) proposed one such model – the Mixture Model of Categorization (MMC) – assuming that $P(x_N|y_N = j, \mathbf{x}_{N-1}, \mathbf{y}_{N-1})$ is a mixture distribution. Specifically, each object x_i comes from a cluster z_i, and each cluster is associated with a probability distribution over the features of the objects generated from that cluster. When evaluating

the probability of a new object x_N, it is necessary to sum over all of the clusters from which that object might have been drawn. Accordingly,

$$P(x_N|y_N = j, \mathbf{x}_{N-1}, \mathbf{y}_{N-1}) = \sum_{k=1}^{K_j} P(x_N|z_N = k, \mathbf{x}_{N-1}, \mathbf{z}_{N-1}) P(z_N = k|\mathbf{z}_{N-1}, y_N = j, \mathbf{y}_{N-1}) \quad (8)$$

where K_j is the total number of clusters for category j, $P(x_N|z_N = k, \mathbf{x}_{N-1}, \mathbf{z}_{N-1})$ is the probability of x_N under cluster k, and $P(z_N = k|\mathbf{z}_{N-1}, y_N = j, \mathbf{y}_{N-1})$ is the probability of generating a new object from cluster k in category j. The clusters can either be shared between categories, or be specific to a single category (in which case $P(z_N = k|\mathbf{z}_{N-1}, y_N = j, \mathbf{y}_{N-1})$ is 0 for all clusters not belonging to category j). This model reduces to kernel density estimation when each object has its own cluster and the clusters are equally weighted, and parametric density estimation when each category is represented by a single cluster. By a similar argument to that used for the exemplar model above, we can connect Equation 8 with the definition of $\eta_{N,j}$ in the VAM (Equation 5), providing a rational justification for this method of interpolating between exemplars and prototypes.[2]

Anderson's Rational Model of Categorization

The MMC elegantly defines a rational model between exemplars and prototypes, but does not determine how many clusters are appropriate for representing each category, based on the available data. Anderson (1990) introduced the Rational Model of Categorization (RMC), which presents a partial solution to this problem. The RMC differs from the other models discussed in this section by treating category labels like features. Thus, the RMC specifies a joint distribution on features and category labels, rather than assuming that the distribution on category labels is estimated separately and then combined with a

distribution on features for each category. As in the MMC, this distribution is a mixture, with

$$P(\mathbf{x}_N, \mathbf{y}_N) = \sum_{\mathbf{z}_N} P(\mathbf{x}_N, \mathbf{y}_N|\mathbf{z}_N)P(\mathbf{z}_N) \tag{9}$$

where $P(\mathbf{z}_N)$ is a distribution over clusterings of the N objects. The key difference from the MMC is that the RMC provides an explicit prior distribution over possible partitions. Importantly, this distribution allows the number of clusters to be unbounded, with

$$P(\mathbf{z}_N) = \frac{(1-c)^K c^{N-K}}{\prod_{i=0}^{N-1}[(1-c)+ci]} \prod_{k=1}^{K}(M_k - 1)! \tag{10}$$

where c is a parameter called the *coupling probability*, and M_k is the number of objects assigned to cluster k. This is the distribution that results from sequentially assigning objects to clusters with probability

$$P(z_i = k|\mathbf{z}_{i-1}) = \begin{cases} \frac{cM_k}{(1-c)+c(i-1)} & \text{if } M_k > 0 \text{ (i.e., } k \text{ is old)} \\ \frac{(1-c)}{(1-c)+c(i-1)} & \text{if } M_k = 0 \text{ (i.e., } k \text{ is new)} \end{cases} \tag{11}$$

where the counts M_k are accumulated over $\mathbf{z}_{i-1}$. Thus, each object can be assigned to an existing cluster with probability proportional to the number of objects already assigned to that cluster, or to a new cluster with probability determined by c.

Despite having been defined in terms of the joint distribution of $\mathbf{x}_N$ and $\mathbf{y}_N$, the assumption that features and category labels are independent given the cluster assignments makes it possible to write $P(x_N|y_N = j, \mathbf{x}_{N-1}, \mathbf{y}_{N-1})$ in the same form as Equation 8. To do so, note that the probability that the Nth observation belongs to the kth cluster is given by,

$$P(z_N = k|\mathbf{z}_{N-1}, y_N = j, \mathbf{y}_{N-1}) \propto P(y_N = j|z_N = k, \mathbf{z}_{N-1}, \mathbf{y}_{N-1})P(z_N = k|\mathbf{z}_{N-1}) \tag{12}$$

where we take into account the fact that this observation belongs to category y_N. The second term on the right hand side is given by Equation 11. This defines a distribution over the same K clusters regardless of j, but the value of K depends on the number of clusters in $\mathbf{z}_{N-1}$. Substituting this expression into Equation 8 provides the relevant mixture model for the RMC. In general, the probabilities in Equation 12 will never be precisely zero, so all clusters contribute to all categories. The RMC can therefore be viewed as a form of the mixture model in which all clusters are shared between categories but the number of clusters is inferred from the data. However, the two models are not directly equivalent because the RMC assumes that both features and category labels are generated from the clusters. This assumption induces a dependency between labels and features, such that the prior over y_N depends on $\mathbf{x}_{N-1}$ as well as $\mathbf{y}_{N-1}$, violating the (arguably sensible) independence assumption made by the other models and embodied in Equation 6.

The RMC comes close to specifying a unifying rational model of categorization, capturing many of the ideas embodied in other models and allowing the representation to be inferred from the data. It can also be shown to mimic the behavior of other models of categorization under certain conditions (Nosofsky, 1991). However, the model is still significantly limited. First, the RMC assumes a single set of clusters for all categories, an assumption that is inconsistent with many models that interpolate between prototypes and exemplars (e.g., Vanpaemel et al., 2005). Second, the idea that category labels should be treated like other features has odd implications, such as the dependency between features and category labels mentioned above. Third, as we will discuss shortly, the approximate algorithm used for assigning objects to clusters in the RMC has serious drawbacks. In order to address these issues, we now discuss the connections between the RMC and nonparametric Bayesian statistics.

Nonparametric Bayes and categorization

One of the most interesting properties of the RMC is that it has a direct connection to nonparametric Bayesian statistics (Neal, 1998). The rationale for using nonparametric methods is that real data are not generally sampled from some neat, finite-dimensional family of distributions, so it is best to avoid this assumption at the outset. From a Bayesian perspective, the nonparametric approach requires us to use priors that include as broad a range of densities of possible, thereby allowing us to infer very complex densities if they are warranted by data. The most commonly used method for placing broad priors over probability distributions is the *Dirichlet process* (DP; Ferguson, 1973). 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). When used in this fashion, the resulting model is referred to as a *Dirichlet process mixture model* (DPMM; Antoniak, 1974; Ferguson, 1983; Neal, 1998).

Although a complete description of the Dirichlet process is beyond the scope of this chapter (for more details, see Navarro, Griffiths, Steyvers, & Lee, 2006), what matters for our purposes is that the Dirichlet process implies a distribution over partitions: any two observations in the sample that were generated from the same mixture component may be treated as members of the same cluster, allowing us to specify priors over an unbounded number of clusters. In the case where N observations have been made, the prior probability that a Dirichlet process will partition those observations into the clusters $\mathbf{z}_N$ is

$$P(\mathbf{z}_N) = \frac{\alpha^K}{\prod_{i=0}^{N-1}[\alpha + i]} \prod_{k=1}^{K} (M_k - 1)! \tag{13}$$

where α is the dispersion parameter of the Dirichlet process. This distribution over partitions can be produced by a simple sequential stochastic process (Blackwell &

MacQueen, 1973), known as the Chinese restaurant process (Aldous, 1985; Pitman, 2002). If observations are assigned to clusters one after another and the probability that observation $i+1$ is assigned to cluster k is

$$P(z_i = k|\mathbf{z}_{i-1}) = \begin{cases} \frac{M_k}{i-1+\alpha} & \text{if } M_k > 0 \text{ (i.e., } k \text{ is old)} \\ \frac{\alpha}{i-1+\alpha} & \text{if } M_k = 0 \text{ (i.e., } k \text{ is new)} \end{cases} \tag{14}$$

we obtain Equation 13 for the probability of the resulting partition. This distribution has a number of nice properties, with one of the most important being *exchangeability*: the prior probability of a partition is unaffected by the order in which the observations are received (Aldous, 1985). To make some of these ideas more concrete, Figure 2 presents a visual depiction of the relationship between the partitioning implied by the DP, the distribution over parameters that is sampled from the DP, and the mixture distribution over stimuli that results in the DPMM.

It should be apparent from our description of the DPMM that it is similar in spirit to the probabilistic model underlying the RMC. In fact, the two are directly equivalent, a point that was first made in the statistics literature by Neal (1998). If we let $\alpha = (1-c)/c$, Equations 10 and 13 are equivalent, as are Equations 11 and 14. Thus the prior over cluster assignments used in the RMC is exactly the same as that used in the DPMM. Anderson (1990, 1991) thus independently discovered one of the most celebrated models in nonparametric Bayesian statistics, deriving this distribution from first principles. This connection provides us with the opportunity to draw on work related to the DPMM in statistics to develop new rational models of categorization. In the remainder of the chapter, we use this approach to explore two new algorithms for approximate Bayesian inference in the RMC and a way to significantly extend the scope of the model.

Approximate inference algorithms

When considering richer representations than prototypes and exemplars it is necessary to have a method for learning the appropriate representation from data. Using Equation 9 to make predictions about category labels and features requires summing over all possible partitions $\mathbf{z}_N$. This sum rapidly becomes intractable for large N, since the number of partitions grows rapidly with the number of stimuli.[3] Consequently, an approximate inference algorithm is needed. The RMC does provide an algorithm, but it has some significant drawbacks. In this section, we first discuss the algorithm that Anderson (1990, 1991) originally proposed for the RMC, and then use the connections with the DPMM to motivate two alternative inference algorithms, which we will compare with exact Bayesian inference and human judgments in the next section.

The existence of alternative inference algorithms for the RMC is valuable for two reasons. The first is that these algorithms provide us with a way to separate the assumptions behind the underlying statistical model – the DPMM – and the scheme used for approximate inference when evaluating the predictions of the model. This is important, because different algorithms can have properties that significantly affect the predictions of the model, such as violating the exchangeability assumption. The second is that each inference algorithm provides us with a hypothesis about how people might go about solving the challenging problem of performing the probabilistic computations involved in Bayesian inference. Rational models are useful for testing assumptions learners make about the environment, but do not generally aim to describe the psychological processes used in solving the computational problems posed by the environment. The computations involved in solving these problems are often intractable, with the overwhelming number of partitions of a set of objects being just one example of a seemingly simple problem that rapidly exceeds the capacities of most computers. Computer science and statistics have developed useful algorithms for approximating intractable probability distributions.

Cognitive scientists can appropriate these algorithms for modeling categorization – assuming that people have rational goals and perhaps approximate the solutions using these same algorithms. Incorporating these algorithms into categorization models provides a way to convert principled rational models into practical process models, as well as tightening the link between these two levels of analysis.

The local MAP algorithm

Anderson (1990, 1991) identified two desiderata for an approximate inference algorithm: that it be incremental, assigning a stimulus to each cluster as it is seen, and that these assignments, once made, be fixed. These desiderata were based on beliefs about the nature of human category learning: that "people need to be able to make predictions all the time not just at particular junctures after seeing many objects and much deliberation" (Anderson, 1991, p. 412), and that "people tend to perceive objects as coming from specific categories" (Anderson, 1991, p. 411). He developed a simple inference algorithm that satisfies these desiderata. We will refer to this algorithm as the *local MAP* algorithm, as it involves assigning each stimulus to the cluster that has the highest posterior probability given the previous assignments (i.e., the maximum a posteriori or MAP cluster).

The local MAP algorithm approximates the sum in Equation 9 with just a single clustering of the N objects, $\mathbf{z}_N$. This clustering is selected by assigning each object to a cluster as it is observed. The posterior probability that stimulus i was generated from cluster k given the features and labels of all stimuli, along with the cluster assignments $\mathbf{z}_{i-1}$ for the previous $i-1$ stimuli is given by

$$P(z_i = k|\mathbf{z}_{i-1}, x_i, \mathbf{x}_{i-1}, y_i, \mathbf{y}_{i-1}) \propto \tag{15}$$
$$P(x_i|z_i = k, \mathbf{z}_{i-1}, \mathbf{x}_{i-1})P(y_i|z_i = k, \mathbf{z}_{i-1}, \mathbf{y}_{i-1})P(z_i = k|\mathbf{z}_{i-1})$$

where $P(z_i = k|\mathbf{z}_{i-1})$ is given by Equation 11. Under the local MAP algorithm, x_i is assigned to the cluster k that maximizes Equation 15. Iterating this process results in a single partition of a set of N objects. The local MAP algorithm approximates the complete joint distribution using only this partition. In effect, it assumes that

$$P(\mathbf{x}_N, \mathbf{y}_N) \approx P(\mathbf{x}_N, \mathbf{y}_N|\mathbf{z}_N) \tag{16}$$

where $\mathbf{z}_N$ is produced via the procedure outlined above. The probability that a particular object receives a particular category label would likewise be computed using a single partition. Unfortunately, although this approach is fast and simple, the local MAP algorithm has some odd characteristics. In particular, the quality of the approximation is often poor, and the algorithm violates the principle of exchangeability. In fact, the local MAP algorithm is *extremely* sensitive to the order in which stimuli are observed, perhaps more than human participants are (see Sanborn, Griffiths, & Navarro, 2006).

Monte Carlo methods

The connection between the RMC and the DPMM suggests a solution to the shortcomings of the local MAP algorithm. In the remainder of this section, we draw on the extensive literature on approximate inference for DPMMs to offer two alternative algorithms for the RMC: Gibbs sampling and particle filtering. These algorithms are less sensitive to order and are asymptotically guaranteed to produce accurate predictions. Both are Monte Carlo methods, in which the intractable sum over partitions is approximated numerically using a collection of samples. Specifically, to compute the probability that a particular object

receives a particular category label, a Monte Carlo approximation gives

$$\begin{aligned} P(y_N = j|\mathbf{x}_N, \mathbf{y}_{N-1}) &= \sum_{\mathbf{z}_N} P(y_N = j|\mathbf{x}_N, \mathbf{y}_{N-1}, \mathbf{z}_N) P(\mathbf{z}_N|\mathbf{x}_N, \mathbf{y}_{N-1}) \\ &\approx \frac{1}{m} \sum_{\ell=1}^{m} P(y_N = j|\mathbf{x}_N, \mathbf{y}_{N-1}, \mathbf{z}_N^{(\ell)}) \end{aligned} \tag{17}$$

where $\mathbf{z}_N^{(1)}, \ldots, \mathbf{z}_N^{(m)}$ are m samples from $P(\mathbf{z}_N|\mathbf{x}_N, \mathbf{y}_{N-1})$, and the approximation becomes exact as $m \to \infty$. This is the principle behind the two algorithms we outline in this section. However, since sampling from $P(\mathbf{z}_N|\mathbf{x}_N, \mathbf{y}_{N-1})$ is not straightforward, the two algorithms use more sophisticated Monte Carlo methods to generate a set of samples.

Gibbs sampling

The approximate inference algorithm most commonly used with the DPMM is Gibbs sampling, a Markov chain Monte Carlo (MCMC) method (see Gilks, Richardson, & Spiegelhalter, 1996). This algorithm involves constructing a Markov chain that will converge to the distribution from which we want to sample, in this case the posterior distribution over partitions. The state space of the Markov chain is the set of partitions, and transitions between states are produced by sampling the cluster assignment of each stimulus from its conditional distribution, given the current assignments of all other stimuli. The algorithm thus moves from state to state by sequentially sampling each z_i from the distribution

$$\begin{aligned} &P(z_i = k|\mathbf{z}_{-i}, x_i, \mathbf{x}_{-i}, y_i, \mathbf{y}_{-i}) \propto \\ &\qquad P(x_i|z_i = k, \mathbf{z}_{-i}, \mathbf{x}_{-i}) P(y_i|z_i = k, \mathbf{z}_{-i}, \mathbf{y}_{-i}) P(z_i = k|\mathbf{z}_{-i}) \end{aligned} \tag{18}$$

where $\mathbf{z}_{-i}$ refers to all cluster assignments except for the ith.

Equation 18 is extremely similar to Equation 15, although it gives the probability of

a cluster based on the all of the trials in the entire experiment except for the current trial, instead of just the previous trials. Exchangeability means that these probabilities are actually computed in exactly the same way: the order of the observations can be rearranged so that any particular observation is considered the last observation. Hence, we can use Equation 14 to compute $P(z_i|\mathbf{z}_{-i})$, with old clusters receiving probability in proportion to their popularity, and a new cluster being chosen with probability determined by α (or, equivalently, c). The other terms reflect the probability of the features and category label of stimulus i under the partition that results from this choice of z_i, and depend on the nature of the features.

The Gibbs sampling algorithm for the DPMM is straightforward (Neal, 1998). First, an initial assignment of stimuli to clusters is chosen. Next, we cycle through all stimuli, sampling a cluster assignment from the distribution specified by Equation 18. This step is repeated, with each iteration potentially producing a new partition of the stimuli. This process is illustrated in Figure 3. Since the probability of obtaining a particular partition after each iteration depends only on the partition produced on the previous iteration, this is a Markov chain. After enough iterations for the Markov chain to converge, we begin to save the partitions it produces. The partition produced on one iteration is not independent of the next, so the results of some iterations are discarded to approximate independence. The partitions generated by the Gibbs sampler can be used in the same way as samples $\mathbf{z}_N^{(\ell)}$ in Equation 17.

The Gibbs sampler differs from the local MAP algorithm in two ways. First, it involves sequentially revisiting the cluster assignments of all objects many times, while the local MAP algorithm assigns each object to a cluster exactly once. Second, the cluster assignment is sampled from the posterior distribution instead of always going to the cluster with the highest posterior probability. As a consequence, different partitions are produced on different iterations, and approximate probabilities can be computed using a collection

of partitions rather than just one. As with all Monte Carlo approximations, the quality of the approximation increases as the number of partitions in that collection increases.

The Gibbs sampler provides an effective means of constructing the approximation in Equation 17, and thus of making accurate predictions about the unobserved features of stimuli. However, it does not satisfy the desiderata Anderson (1990, 1991) used to motivate his algorithm. In particular, it is not an incremental algorithm: it assumes that all data are available at the time of inference. Depending on the experimental task, this assumption may be inappropriate. The Gibbs sampler is an excellent algorithm to model experiments where people are shown the full set of stimuli simultaneously. However, when the stimuli are shown sequentially, it needs to be run again each time new data are added, making it inefficient when predictions need to be made on each trial. In such situations, we need to use a different algorithm.

Particle filtering

Particle filtering is a sequential Monte Carlo technique that can be used to provide a discrete approximation to a posterior distribution that can be updated with new data (Doucet, de Freitas, & Gordon, 2001). Each "particle" is a partition $\mathbf{z}_i^{(\ell)}$ of the stimuli from the first i trials. Unlike the local MAP algorithm, in which the posterior distribution is approximated with a single partition, the particle filter uses m partitions. Summing over these particles gives us an approximation to the posterior distribution over partitions

$$P(\mathbf{z}_i|\mathbf{x}_i,\mathbf{y}_i) \approx \frac{1}{m}\sum_{\ell=1}^{m}\delta(\mathbf{z}_i,\mathbf{z}_i^{(\ell)}) \tag{19}$$

where $\delta(\mathbf{z},\mathbf{z}')$ is 1 when $\mathbf{z}=\mathbf{z}'$, and 0 otherwise. If Equation 19 is used as an approximation to the posterior distribution over partitions $\mathbf{z}_i$ after the first i trials, then we can approximate the distribution of $\mathbf{z}_{i+1}$ given the observations $\mathbf{x}_i,\mathbf{y}_i$ in the following

manner:

$$\begin{aligned} P(\mathbf{z}_{i+1}|\mathbf{x}_i, \mathbf{y}_i) &= \sum_{\mathbf{z}_i} P(\mathbf{z}_{i+1}|\mathbf{z}_i) P(\mathbf{z}_i|\mathbf{x}_i, \mathbf{y}_i) \\ &\approx \sum_{\mathbf{z}_i} P(\mathbf{z}_{i+1}|\mathbf{z}_i) \frac{1}{m} \sum_{\ell=1}^{m} \delta(\mathbf{z}_i, \mathbf{z}_i^{(\ell)}) \\ &= \frac{1}{m} \sum_{\ell=1}^{m} P(\mathbf{z}_{i+1}|\mathbf{z}_i^{(\ell)}) \end{aligned} \tag{20}$$

where $P(\mathbf{z}_{i+1}|\mathbf{z}_i)$ is given by Equation 14. We can then incorporate the information conveyed by the features and label of stimulus $i+1$, arriving at the approximate posterior probability

$$\begin{aligned} P(\mathbf{z}_{i+1}|\mathbf{x}_{i+1}, \mathbf{y}_{i+1}) &\propto P(x_{i+1}|\mathbf{z}_{i+1}, \mathbf{x}_i) P(y_{i+1}|\mathbf{z}_{i+1}, \mathbf{y}_i) P(\mathbf{z}_{i+1}|\mathbf{x}_i, \mathbf{y}_i) \\ &\approx \frac{1}{m} \sum_{\ell=1}^{m} P(x_{i+1}|\mathbf{z}_{i+1}, \mathbf{x}_i) P(y_{i+1}|\mathbf{z}_{i+1}, \mathbf{y}_i) P(\mathbf{z}_{i+1}|\mathbf{z}_i^{(\ell)}) \end{aligned} \tag{21}$$

The result is a discrete distribution over all the previous particle assignments and all possible assignments for the current stimulus. Drawing m samples from this distribution provides us with our new set of particles, as illustrated in Figure 4.

The particle filter for the RMC is initialized with the first stimulus assigned to the first cluster for all m particles. On each following trial, the distribution in Equation 21 is calculated, based on the particles sampled in the last trial. On any trial, these particles provide an approximation to the posterior distribution over partitions. The stimuli are integrated into the representation incrementally, satisfying one of Anderson's desiderata. The degree to which Anderson's fixed assignment criterion is satisfied depends on the number of particles. The assignments in the particles themselves are fixed: once a stimulus has been assigned to a cluster in a particle, it cannot be reassigned. However, the probability of a previous assignment across particles can change when a new stimulus is

introduced. When a new set of particles is sampled, the number of particles that carry a particular assignment of a stimulus to a cluster is likely to change. For large m, the assignments will not appear fixed. However, when $m = 1$, previous assignments cannot be changed, and Anderson's criterion is unambiguously satisfied. In fact, the single-particle particle filter is very similar to the local MAP algorithm: each assignment of a stimulus becomes fixed on the trial the stimulus is introduced. The key difference from the local MAP algorithm is that each stimulus is stochastically assigned a cluster by sampling from the posterior distribution, rather than being deterministically assigned to the cluster with highest posterior probability.

Comparing the algorithms to data

In this section we use data from Medin and Schaffer's (1978) Experiment 1 to compare how effective the algorithms are in approximating the full Bayesian solution, and how closely they match human performance. In order to do so, we need to specify a measure of the probability of a set of features given a particular partition. The RMC assumes that the features (and category label) of a stimulus are independent once the cluster it belongs to is known. Using this idea, we can write the probability of the features of a stimulus as

$$P(x_N|z_N = k, \mathbf{x}_{N-1}, \mathbf{z}_{N-1}) = \prod_d P(x_{N,d}|z_N = k, \mathbf{x}_{N-1}, \mathbf{z}_{N-1})$$

where $x_{N,d}$ is the value of the dth feature of object N. In this section, we collapse the distinction between category labels and features, treating category labels simply as a special kind of discrete feature. Anderson (1991) presents the likelihood for both discrete and continuous features, but we need only consider binary features for our applications. Given the cluster, the value on each feature is assumed to have a Bernoulli distribution. Integrating out the parameter of this distribution with respect to a Beta(β_0, β_1) prior, we

obtain

$$P(x_{N,d} = v | z_N = k, \mathbf{x}_{N-1}, \mathbf{z}_{N-1}) = \frac{B_v + \beta_v}{B_\cdot + \beta_0 + \beta_1} \tag{22}$$

where B_v is the number of stimuli with value v on the dth feature that $\mathbf{z}_N$ identifies as belonging to the same cluster as x_N. $B_\cdot$ denotes the number of other stimuli in the same cluster. We use $\beta_0 = \beta_1 = 1$ in all simulations.

Medin and Schaffer's (1978) experiment used six training items, each consisting of five binary features (including the category label, listed last): 11111, 10101, 01011, 00000, 01000, and 10110. In an experiment with only six training examples, the exact posterior probabilities can be computed, as can the partition with the highest posterior probability (the global MAP solution). The algorithms were trained on the six examples, and the category label of a set of test stimuli (shown in Table 1) was then predicted. Three coupling probabilities were compared: $c = 0.25$, $c = 0.45$, and $c = 0.75$. The local MAP algorithm was run on all 720 possible orders of the training stimuli. The Gibbs sampler was run for $1,100$ iterations on a single training order. The first 100 iterations were discarded and only every 10th iteration was kept for a total of 100 samples. The particle filter was run with 100 particles on a single training order. Linear correlations with the human confidence ratings reported by Medin and Schaffer (1978) were computed for all algorithms.

The results shown in the top row of Figure 5 show that the coupling parameter does not have a large effect on the exact solution, the particle filter, or the Gibbs sampler. Moreover, the particle filter and Gibbs sampler provide good approximations to the full posterior solution.[4] In contrast, the local MAP algorithm depends heavily on the value of the coupling parameter. Furthermore, the global MAP solution, which the local MAP algorithm attempts to discover, is not a very good approximation to the full posterior, and provides a worse fit to the human data than the local MAP solution.

The fits to the human data for the two Monte Carlo algorithms are not particularly good when shown one instance of each stimulus (i.e. one block of training), but improve when they are trained on ten blocks of the six stimuli, as shown in the lower panels of Figure 5. This is more relevant for the different algorithms to human data, as participants in the experiment received ten blocks of training data. The full posterior is not tractable for sixty trials, but we can still compare the three approximation algorithms. Again, all of the predictions across algorithms and values of the coupling parameter are similar except for the local MAP algorithm with a high coupling parameter. Overall, the local MAP algorithm does not predict the human data any better than the other algorithms, and is in fact substantially worse for some values of the coupling parameter.

Unifying rational models using hierarchical Dirichlet processes

In the previous sections, interpreting the RMC as a DPMM allowed us to propose approximate inference algorithms that improve the fit to empirical data and better approximate the ideal Bayesian solution to the categorization problem. In this section we extend the approach, showing how Bayesian nonparametric models can unify all of the rational models discussed so far, subsuming prototypes, exemplars, the MMC, and RMC into a single model that learns the most appropriate representational structure. The tool that we will use to do this is the *hierarchical Dirichlet process* (HDP).

The HDP, introduced by Teh, Jordan, Blei, and Beal (2004), is a straightforward generalization of the basic Dirichlet process. Observations are divided into groups, and each group is modeled using a Dirichlet process (with parameter α). A new observation is first compared to all of the clusters in its group, with the prior probability of each cluster determined by Equation 14. If the observation is to be assigned to a new cluster, the new cluster is drawn from a second Dirichlet process that compares the stimulus to all of the clusters that have been created across groups. This higher-level Dirichlet process is

governed by parameter γ, analogous to α, and the prior probability of each cluster is proportional to the number of times that cluster has been selected by any group, instead of the number of observations in each cluster. The new observation is only assigned to a completely new cluster if both Dirichlet processes select a new cluster. In this manner, stimuli in different categories can end up belonging to the same mixture component, simply by being drawn from the same partition in the higher level. An illustration of this is shown in Figure 6.

The HDP provides a way to model probability distributions across groups of observations. Each distribution is a mixture of an unbounded number of clusters, but the clusters can be shared between groups. Shared clusters allow the model to leverage examples from across categories to better estimate cluster parameters. A priori expectations about the number of clusters in a group and the extent to which clusters are shared between groups are determined by the parameters α and γ. When α is small, each group will have few clusters, but when α is large, the number of clusters will be closer to the number of observations. When γ is small, groups are likely to share clusters, but when γ is large, the clusters in each group are likely to be unique.

We can now define a unifying rational model of categorization, based on the HDP. If we identify each category with a "group" for which we want to estimate a distribution, the HDP becomes a model of category learning, subsuming all previous rational models through different settings of α and γ. Figure 7 identifies six models we can obtain by considering limiting values of α and γ.[5] We will refer to the different models using the notation $\text{HDP}_{\alpha,\gamma}$, where α and γ take on values corresponding to the values of the two parameters of the model (with + denoting a value in the interval $(0, \infty)$).Three of the models shown in Figure 7 are exactly isomorphic to existing models.[6] $\text{HDP}_{\infty,\infty}$ is an exemplar model, with one cluster per object and no sharing of clusters. $\text{HDP}_{0,\infty}$ is a prototype model, with one cluster per category and no sharing of clusters. $\text{HDP}_{\infty,+}$ is the

RMC, provided that category labels are treated as features. In $\text{HDP}_{\infty,+}$, every object has its own cluster, but those clusters are generated from the higher-level Dirichlet process. Consequently, group membership is ignored and the model reduces to a Dirichlet process.

Figure 7 also includes some models that have not previously been explored in the literature on categorization. $\text{HDP}_{0,+}$ makes the same basic assumptions as the prototype model, with a single cluster per category, but makes it possible for different categories to share the same prototype – something that might be appropriate in an environment where the same category can have different labels. However, the most interesting models are $\text{HDP}_{+,+}$ and $\text{HDP}_{+,\infty}$. These models are essentially the MMC, with clusters shared between categories or unique to different categories respectively, but the number of clusters in each category can differ and can be learned from the data. Consequently, these models make it possible to answer the question of whether a particular category is best represented using prototypes, exemplars, or something in between, simply based on the objects belonging to that category. In the remainder of the chapter, we show that one of these models – $\text{HDP}_{+,\infty}$ – can capture the shift that occurs from prototypes to a more exemplar-based representation in a recent categorization experiment.

Modeling the prototype-to-exemplar transition

Smith and Minda (1998) argued that people seem to produce responses that are more consistent with a prototype model early in learning, later shifting to exemplar-based representations. The models discussed in the previous section potentially provide a rational explanation for this effect: the prior specified in Equation 13 prefers fewer clusters and is unlikely to be overwhelmed by small amounts of data to the contrary, but as the number of stimuli consistent with multiple clusters increases, the representation should shift. These results thus provide an opportunity to compare the HDP to human data.

We focused on the non-linearly separable structure explored in Experiment 2 of

Smith and Minda (1998). In this experiment, 16 participants were presented with six-letter nonsense words labeled as belonging to different categories. Each letter could take one of two values, producing the binary feature representation shown in Table 2. Each category contains one prototypical stimulus (000000 or 111111), five stimuli with five features in common with the prototype, and one stimulus with only one feature in common with the prototype, which we will refer to as an "exception". No linear function of the features can correctly classify every stimulus, meaning that a prototype model cannot distinguish between the categories exactly. Participants were presented with a random permutation of the 14 stimuli and asked to identify each as belonging to either Category A or Category B, receiving feedback after each stimulus. This block of 14 stimuli was repeated 40 times for each participant, and the responses were aggregated into 10 segments of 4 blocks each. The results are shown in Figure 8 (a). The exceptions were initially identified as belonging to the wrong category, with performance improving later in training.

We tested three models: the exemplar model $\text{HDP}_{\infty,\infty}$, the prototype model $\text{HDP}_{0,\infty}$, and $\text{HDP}_{+,\infty}$. All three models were exposed to the same training stimuli as the human participants and used to categorize each stimulus after each segment of 4 blocks. The cluster structures for the prototype and exemplar models are fixed, so the probability of each category is straightforward to compute. However, since $\text{HDP}_{+,\infty}$ allows arbitrary clusterings, the possible clusterings need to be summed over when computing the probabilities used in categorization (as in Equation 8). We approximated this sum by sampling from the posterior distribution on clusterings using the MCMC algorithm described by Teh et al. (2004), which is a variant on the Gibbs sampling algorithm for the DPMM introduced above. Each set of predictions is based on an MCMC simulation with a burn-in of 1000 steps, followed by 100 samples separated by 10 steps each. The parameter α, equivalent to the coupling probability c, was also estimated by sampling.

As in Smith and Minda's original modeling of this data, a guessing parameter was

incorporated to allow for the possibility that participants were randomly responding for some proportion of the stimuli. In practice, rational models – which have perfect memory for the stimuli and access to their features – can outperform human learners, so introducing a guessing parameter to handicap the models is a necessary part of comparing them to human data. If a model originally assigned probability $P(y_N = j)$ to categorizing a stimulus to some category, and the guessing parameter for the participant in question was ϕ, this probability would be updated to $(1 - \phi)P(y_N = j) + \phi 0.5$. The guessing parameter was allowed to vary between 0 and 1 across individual participants, but was fixed per participant across every instance of every stimulus. Furthermore, the values of β_0 and β_1 in Equation 22 were fit to each participant, with the restriction that $\beta_0 = \beta_1$. Intuitively, this captures variation in the tendency to create new clusters, since the stronger bias towards feature probabilities near 0.5 resulting from high values of β_0 and β_1 makes it less likely that a new cluster will provide a better match to the particular features of a given object.

The predictions of the three models are shown in Figure 8. As might be expected, the prototype model does poorly in predicting the categories of the exceptions, while the exemplar model is more capable of handling these stimuli. We thus replicated the results of Smith and Minda (1998), finding that the prototype model fit better early in training, and the exemplar model better later in training. More interestingly, we also found that $\text{HDP}_{+,\infty}$ provided an equivalent or better account of human performance than the other two models after the first four segments. In particular, only this model captured the shift in the treatment of the exceptions over training. This shift occurred because the number of clusters in the HDP changes around the fourth segment: categories are initially represented with one cluster, but then become two clusters, one for the stimuli close to the prototype and one for the exception.

The HDP model produces the shift from performance similar to a prototype model

to performance similar to an exemplar model because this shift is justified by the data. The underlying structure – five stimuli that form a natural cluster and one exception in each category – supports a representation with more than a single cluster, and once evidence for this being the true structure accumulates, through the provision of enough instances of these stimuli, this is the structure favored by the posterior distribution. The model is able to capture similar predictions for other experiments reported by Smith and Minda (1998), as well as other standard datasets (e.g., Nosofsky, Gluck, Palmeri, McKinley, & Glauthier, 1994), but perhaps its greatest strength is in being able to explain how learning about one category can inform learning about another. In the general case, the HDP model allows clusters to be shared between categories, suggesting that we might be able to understand the great ease with which adults learn new categories of familiar objects (or new words) in terms of having acquired an accurate understanding of the clusters from which these categories could be composed through their previous experiences in category learning.

Conclusion

One of the most valuable aspects of rational models of cognition is their ability to establish connections across different fields. Here, we were able to exploit the correspondence between Anderson's (1990) Rational Model of Categorization and the Dirichlet process to draw on recent work in nonparametric Bayesian statistics. Using this correspondence, we identified more accurate approximation algorithms for use with Anderson's model and to define a more general rational model, based on the hierarchical Dirichlet process. The algorithms provide a source of hypotheses as to how people can solve the difficult problem of performing Bayesian inference, and the new model subsumes previous rational analyses of human category learning, indicating how learners should select the number of clusters to represent a category. The result is a picture of human categorization in which people

do not use a fixed representation of categories across all contexts, but instead select a representation whose complexity is warranted by the available data, using simple and efficient approximation algorithms to perform these computations.

While our focus in this paper has been on applying ideas from statistics to cognitive science, the connection between human category learning and methods used in nonparametric Bayesian density estimation also has the potential to lead to new kinds of models that might be useful in statistics. The ways in which people use different sources of data in forming categories, combine category learning with language learning, and exploit structured knowledge as well as statistical information when categorizing objects all provide challenging computational problems that are beyond the scope of existing statistical models. Understanding how people solve these problems is likely to require thinking about categorization in terms that are more sophisticated than the schemes for density estimation summarized in this chapter, although we anticipate that similar issues of determining the complexity of the underlying representations are likely to arise, and that solutions to these problems can be found in the methods of nonparametric Bayesian statistics.

References

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

Anderson, J. R. (1990). *The adaptive character of thought.* Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429.

Antoniak, C. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, *2*, 1152-1174.

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, *39*, 216-233.

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

Briscoe, E., & Feldman, J. (2006). Conceptual complexity and the bias-variance tradeoff. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society.* Mahwah, NJ: Erlbaum.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking.* New York, NY: Wiley.

Chater, N., & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Science*, *3*, 57-65.

Doucet, A., Freitas, N. de, & Gordon, N. (2001). *Sequential Monte Carlo methods in practice.* New York: Springer.

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

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

Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. In M. Rizvi, J. Rustagi, & D. Siegmund (Eds.), *Recent advances in statistics* (p. 287-302). New York: Academic Press.

Gilks, W., Richardson, S., & Spiegelhalter, D. J. (Eds.). (1996). *Markov chain Monte Carlo in practice.* Suffolk, UK: Chapman and Hall.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*, 354-384.

Kruschke, J. K. (1990). *A connectionist model of category learning.* Unpublished doctoral dissertation, University of California, Berkeley, Berkeley, CA.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, *111*, 309-332.

Luce, R. D. (1959). *Individual choice behavior.* New York: John Wiley.

Marr, D. (1982). *Vision.* San Francisco, CA: W. H. Freeman.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207-238.

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.

Neal, R. M. (1998). *Markov chain sampling methods for Dirichlet process mixture models* (Tech. Rep. No. 9815). Department of Statistics, University of Toronto.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39-57.

Nosofsky, R. M. (1991). Relation between the rational model and the context model of categorization. *Psychological Science*, *2*, 416-421.

Nosofsky, R. M. (1998). Optimal performance and exemplar models of classification. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (p. 218-247). Oxford: Oxford University Press.

Nosofsky, R. M., Gluck, M., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994). Comparing models of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition*, *22*, 352-369.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*, 608-631.

Oaksford, M., & Chater, N. (Eds.). (1998). *Rational models of cognition.* Oxford: Oxford University Press.

Pitman, J. (2002). *Combinatorial stochastic processes.* (Notes for Saint Flour Summer School)

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

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*, 393-407.

Rice, J. A. (1995). *Mathematical statistics and data analysis* (2nd ed.). Belmont, CA: Duxbury.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. Lloyd (Eds.), *Cognition and categorization* (p. 27-48). Hillsdale, New Jersey: Erlbaum.

Rosseel, Y. (2002). Mixture models of categorization. *Journal of Mathematical Psychology*, *46*, 178-210.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2006). A more rational model of categorization. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society.* Mahwah, NJ: Erlbaum.

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

Shepard, R. N. (1987). Towards a universal law of generalization for psychological science. *Science*, *237*, 1317-1323.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM: Retrieving Effectively from Memory. *Psychonomic Bulletin & Review*, *4*, 145-166.

Silverman, B. W. (1986). *Density estimation.* London: Chapman and Hall.

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 1411-1436.

Teh, Y., Jordan, M., Beal, M., & Blei, D. (2004). Hierarchical Dirichlet processes. In *Advances in Neural Information Processing Systems 17.* Cambridge, MA: MIT Press.

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

Vanpaemel, W., Storms, G., & Ons, B. (2005). A varying abstraction model for categorization. In

Proceedings of the 27th Annual Conference of the Cognitive Science Society. Mahwah, NJ: Erlbaum.

Author Note

TLG was supported by a Junior Faculty Research Grant from the University of California, Berkeley, and grant number FA9550-07-1-0351 from the Air Force Office of Scientific Reesarch. ANS was supported by a NSF Graduate Research Fellowship. DJN was supported by an Australian Research Fellowship (ARC grant DP-0773794). We thank Nancy Briggs for helpful comments, and J. Paul Minda for providing data.

Footnotes

[1]The constant of proportionality is determined by $\int f(x, x_i)dx$, being $\frac{1}{N_j}$ if $\int f(x, x_i)\ dx = 1$ for all i, and is absorbed into β_j to produce direct equivalence to Equation 2.

[2]Note, however, that the MMC is more general than the VAM, since the VAM does not allow clusters to be shared across categories.

[3]The number of partitions of a set of N stimuli is given by the Nth Bell number, with the first ten values being 1, 2, 5, 15, 52, 203, 877, 4140, 21147, and 115975.

[4]Though not shown, a particle filter with fewer particles produced correlations to human data that were similar to those produced with 100 particles.

[5]The case of $\gamma \to 0$ is omitted, since it simply corresponds to a model in which all observations belong to the same cluster across both categories, for all values of α.

[6]In stating these equivalence results, we focus just on the kind of representation acquired by the model. In order to produce the same predictions for new observations, we need to assume that different values of the α and γ parameters are used in acquiring a representation and applying it. Specifically, we need to assume that $\alpha = 0$ in $\text{HDP}_{\infty,\infty}$ when making categorization decisions, guaranteeing that the new object is compared to old exemplars. A similar assumption was made by Nosofsky (1991) in showing equivalence between the RMC and exemplar models.

Table 1
Test Stimuli Ordered by Category 1 Subject Ratings from Medin and Schaffer (1978)

1111	0101	1010	1101	0111	0001	1110	1000	0010	1011	0100	0000

Table 2
Categories A and B from Smith and Minda (1998)

	Stimuli
A	000000, 100000, 010000, 001000, 000010, 000001, 111101
B	111111, 011111, 101111, 110111, 111011, 111110, 000100

Figure Captions

Figure 1. Category similarity functions for a simple one-dimensional category. The panel on the left shows the similarity function for a prototype model, with a single prototype summarizing the structure of the category. The panel on the right shows the similarity function for an exemplar model, with the overall similarity resulting from summing a set of similarity functions centered on each exemplar. The similarity function shown in the middle panel comes from an intermediate model that groups the three stimuli on the left and the two stimuli on the right.

Figure 2. The relationship between (a) the clustering implied by the DP, (b) the distribution over parameters that is sampled from the DP, and (c) the mixture distribution over stimuli that results in the DPMM. The clustering assignments in (a) were produced by drawing sequentially from the stochastic process defined in Equation 14, and each cluster is associated with a parameter value θ. After an arbitrarily large number of cluster assignments have been made, we can estimate the probability of each cluster, and hence of the corresponding parameter value. The resulting probability distribution is shown in (b). If each value of θ is treated as the mean of a simple normal distribution (with fixed variance) over the value of some continuous stimulus dimension, then the resulting mixture distribution drawn from the DPMM is the one illustrated in (c). While the applications considered in this chapter use stimuli that have discrete features, not a single continuous dimension, the notion of a mixture distribution is more intuitive in the continuous setting.

Figure 3. Example of Gibbs sampling with three objects (circles, differentiated by numbers). A partition of the objects is expressed using boxes, where all objects within a box belong to the same element of the partition. At any point in time, a single partition is maintained. Stochastic transitions between partitions are produced by sequentially sampling the element of the partition to which each object is assigned from its conditional

distribution given the data and all other assignments. The partition produced by a full iteration of sampling (i.e. reassignment of all three objects) is shown by the solid boxes, with the intermediate steps being illustrated by dotted boxes. After many iterations, the probability of producing a particular partition corresponds to the posterior probability of that partition given the observed data (features and category labels).

Figure 4. Example of particle filtering, involving three particles and three sequentially observed objects (circles, differentiated by numbers). On any given trial, we take the sampled distribution over partitions (boxes) from previous trial, and treat it as an approximation to the full posterior over partitions for that trial (Equation 19). We then update to an approximate posterior for the current trial using Equation 21 and redraw a collection of particles. Note that since we are sampling with replacement, it is possible for particles to "exchange histories", as is illustrated by the states of particles 2 and 3 in this figure.

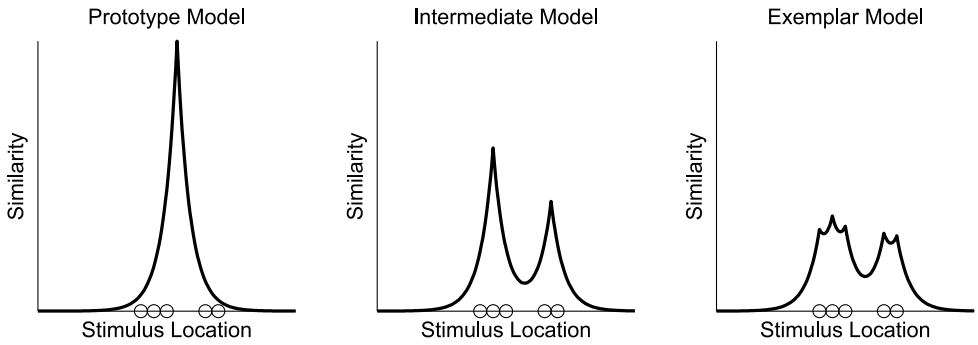
Figure 5. Probability of choosing category 1 for the stimuli from the first experiment of Medin & Schaffer (1978). The test stimuli (listed in order of human preference in the legend) are along the horizontal axis. In the first row only the first six trials are presented, while in the second row ten blocks of six trials each are presented. The three lines in each panel correspond to three different coupling parameters: $c = 0.25, 0.45,$ or 0.75. Correlations between the human data and the simulation data are displayed on each plot for each value of the coupling parameter.

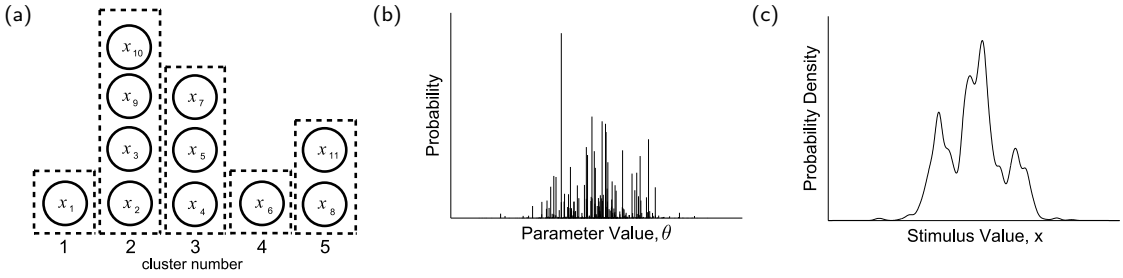
Figure 6. Illustration of the HDP prior. The prior probability for each cluster at the lower level is based on the number of category examples in that cluster. If a cluster is selected from the higher level, the prior probability of clusters is based on the number of categories by which they have been selected. Completely new clusters can only be created at the

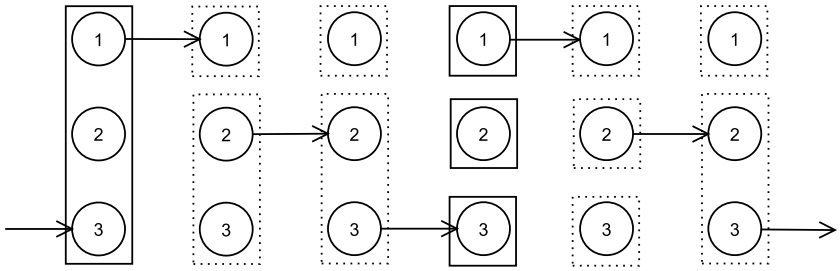
higher level.

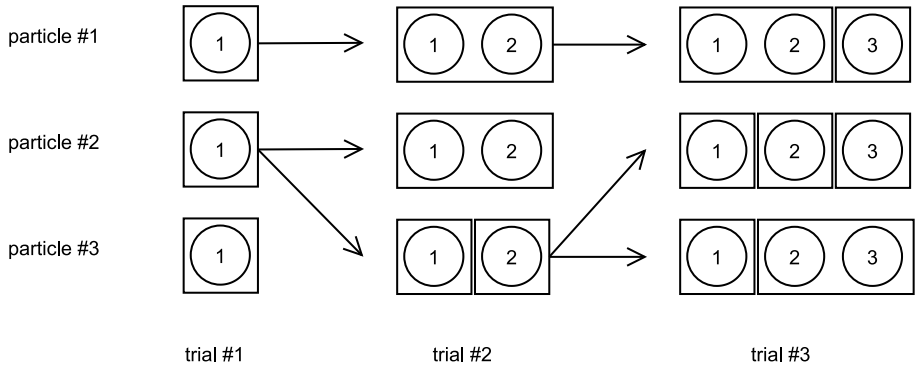
Figure 7. Structural assumptions underlying different parameterizations of the $\text{HDP}_{\alpha,\gamma}$ model. The unfilled circles are clusters, the filled circles are exemplars, and the boxes indicate which exemplars belong to the same categories. Descriptions of the properties of these six models and their correspondence to existing models are given in the text.

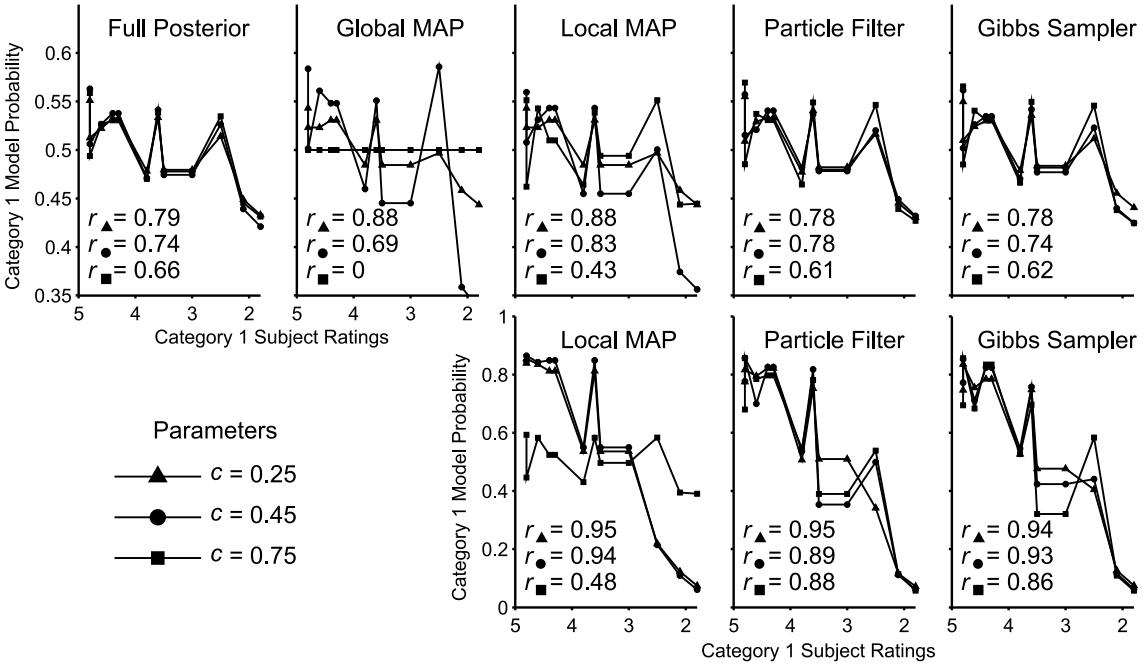
Figure 8. Human data and model predictions. (a) Results of Smith and Minda (1998, Experiment 2). (b) Prototype model, $\text{HDP}_{\infty,0}$. (c) Exemplar model, $\text{HDP}_{\infty,\infty}$. (d) $\text{HDP}_{+,\infty}$. For all panels, white plot markers are stimuli in Category A, and black are in Category B. Triangular markers correspond to the exceptions to the prototype structure (111101 and 000100 respectively).

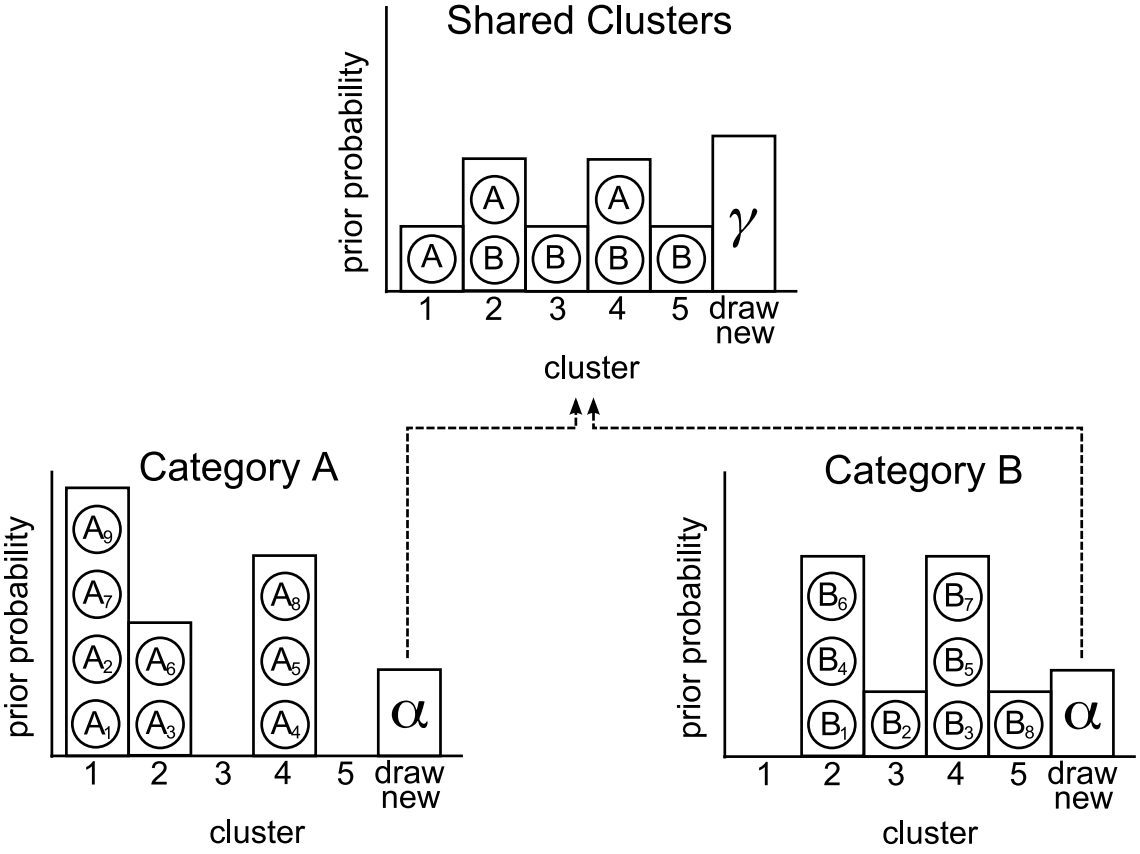

Prototype Model
Similarity
Stimulus Location
Intermediate Model
Similarity
Stimulus Location
Exemplar Model
Similarity
Stimulus Location

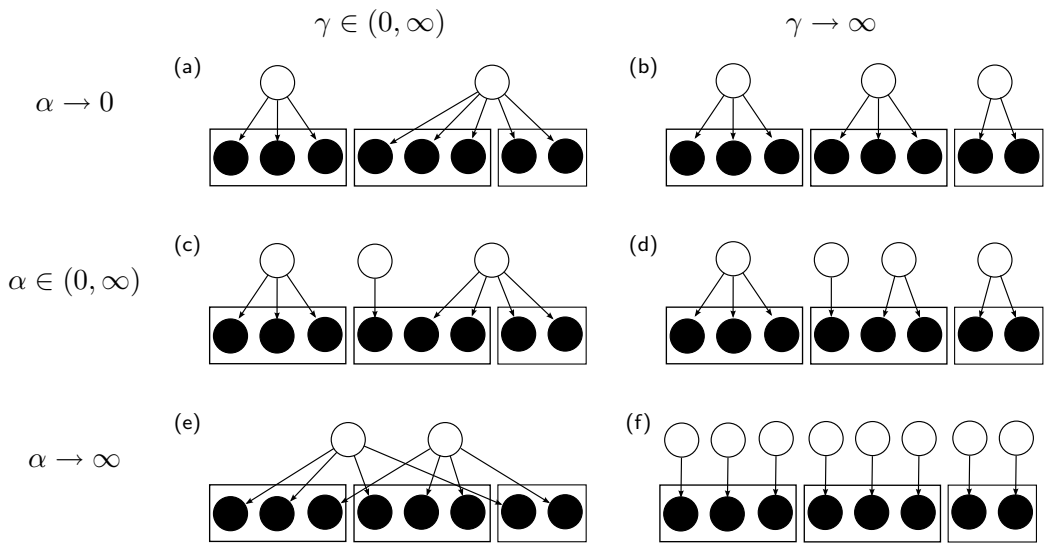

(a)
x_{10}
x_9
x_3
x_2
x_1
x_7
x_5
x_4
x_6
x_{11}
x_8
1
2
3
4
5
cluster number
(b)
Probability
Parameter Value, θ
(c)
Probability Density
Stimulus Value, x

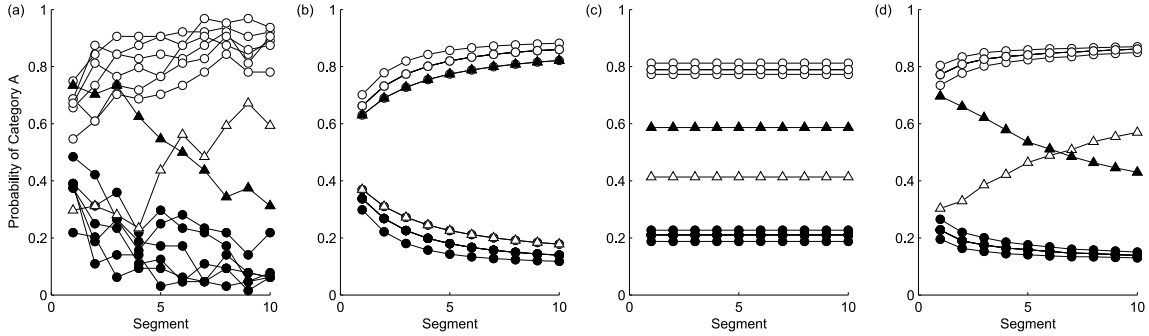

1
2
3
1
2
3
1
2
3
1
2
3
1
2
3


particle #1
particle #2
particle #3
1
1
2
1
2
3
1
1
2
1
2
3
1
1
2
1
2
3
trial #1
trial #2
trial #3


Full Posterior
Global MAP
Local MAP
Particle Filter
Gibbs Sampler
Category 1 Model Probability
Category 1 Subject Ratings
$r_▲$ = 0.79
$r_●$ = 0.74
$r_■$ = 0.66
$r_▲$ = 0.88
$r_●$ = 0.69
$r_■$ = 0
$r_▲$ = 0.88
$r_●$ = 0.83
$r_■$ = 0.43
$r_▲$ = 0.78
$r_●$ = 0.78
$r_■$ = 0.61
$r_▲$ = 0.78
$r_●$ = 0.74
$r_■$ = 0.62
Parameters
c = 0.25
c = 0.45
c = 0.75
Local MAP
Particle Filter
Gibbs Sampler
$r_▲$ = 0.95
$r_●$ = 0.94
$r_■$ = 0.48
$r_▲$ = 0.95
$r_●$ = 0.89
$r_■$ = 0.88
$r_▲$ = 0.94
$r_●$ = 0.93
$r_■$ = 0.86


Shared Clusters
prior probability
A
A
B
B
B
B
A
A
B
γ
1
2
3
4
5
draw new
cluster
Category A
prior probability
A9
A7
A2
A1
A6
A3
A8
A5
A4
α
1
2
3
4
5
draw new
cluster
Category B
prior probability
B6
B4
B1
B2
B7
B5
B3
B8
α
1
2
3
4
5
draw new
cluster


$\gamma \in (0, \infty)$
$\gamma \to \infty$
$\alpha \to 0$
$\alpha \in (0, \infty)$
$\alpha \to \infty$
(a)
(b)
(c)
(d)
(e)
(f)


(a)
(b)
(c)
(d)
Probability of Category A
Segment
0
0.2
0.4
0.6
0.8
1
5
10