Running head: NONPARAMETRIC BAYESIAN MODELS OF CATEGORIZATION

Nonparametric Bayesian models of categorization

Thomas L. Griffiths

Department of Psychology

University of California, Berkeley

Adam N. Sanborn

Gatsby Computational Neuroscience Unit

University College London

Kevin R. Canini

Computer Science Division

University of California, Berkeley

Daniel J. Navarro

School of Psychology

University of Adelaide

Joshua B. Tenenbaum

Department of Brain and Cognitive Sciences

Massachusetts Institute of Technology

Nonparametric Bayesian models of categorization

Motivation

Models of human categorization often focus on the psychological processes by which people form and use knowledge of categories, appealing to concepts such as memory traces, activation, and similarity. An alternative approach is to take a step back from these psychological processes, and instead consider the abstract computational problem being solved when we learn categories, exploring how ideal solutions to that problem might shed light on human behavior. This kind of investigation – conducted at Marr's (1982) computational level, or via Anderson's (1990) principles of rational analysis – has been particularly successful for categorization, identifying some surprising connections between psychological process models and methods used in machine learning and statistics. This chapter explores some of these connections in detail, and may present technical ideas that are new to many readers. Those who are interested in the mathematical details can find readable introductions from the perspectives of machine learning and cognitive science in Bishop (2006) and Griffiths, Kemp, and Tenenbaum (2008) respectively.

Categorization is an instance of an inductive problem, requiring category membership to be inferred from the limited information provided by the features of a stimulus. As such, an ideal solution to this problem is provided by Bayesian inference, and in particular by computing a probability distribution over categories given the stimulus. If the joint probability of the features x and category label c of a stimulus is $p(x, c)$, then the probability that x belongs to category c is given by

$$p(c|x) = \frac{p(x,c)}{\sum_{c'} p(x,c')} \qquad (1)$$

where the sum in the denominator ranges over all categories. From this perspective, learning a category reduces to estimating the joint probability distribution $p(x, c)$, indicating the probability of observing an object x that belongs to category c. Rational analyses of category learning thus agree that it is fundamentally a problem of *density estimation*, although they differ in whether they focus on estimating the joint distribution $p(x, c)$ directly (e.g., Anderson, 1990) or they consider how conditional distributions $p(x|c)$ could be estimated for each category separately (e.g., Ashby & Alfonso-Reese, 1995; Nosofsky, 1998; Rosseel, 2002).

Traditional statistical solutions to the problem of density estimation are of two types: parametric and nonparametric (Silverman, 1986). In parametric density estimation, a probability distribution is assumed to be of a known form, such as a Gaussian, and density estimation consists of determining the parameters of that distribution. In traditional nonparametric density estimation schemes, a probability distribution is approximated as the sum of a set of “kernels” – functions which fall off with distance from a central point – where the kernels are centered on points sampled from the distribution. When used to estimate the conditional distribution on features associated with each category, $p(x|c)$, these two approaches correspond to the two main classes of psychological process models: prototype and exemplar models (Ashby & Alfonso-Reese, 1995). Prototype models, corresponding to parametric density estimation, assume that a category is associated with a single prototype and that categorization involves comparing new stimuli to these prototypes (e.g., Reed, 1972). Exemplar models, corresponding to kernel-based nonparametric density estimation, assume that a category is represented by a set of stored exemplars and that categorizing new stimuli involves comparing these stimuli to the set of exemplars in each category (e.g., Medin & Schaffer, 1978; Nosofsky, 1986).

Traditional parametric and nonparametric density estimation methods have different advantages and disadvantages: the greater flexibility of nonparametric methods comes at the cost of requiring more data to estimate a distribution. Consequently, there is not a clear argument in favor of one of these approaches from rational grounds, and statisticians have begun to explore more sophisticated density estimation techniques that combine the strengths of both approaches by supporting representations that interpolate between using a single parametric distribution and having a kernel associated with each stimulus. Many of these approaches are based on *mixture models*, in which a distribution is assumed to be a mixture of a set of parametric densities (McLachlan & Basford, 1988). This is an idea that resonates with work in psychology that has explored process models in which categories are represented using clusters of several exemplars, with each cluster having its own prototype (e.g., Love, Medin, & Gureckis, 2004; Vanpaemel & Storms, 2008). The potential for mixture models to capture representations that lie between prototypes and exemplars has been recognized in the psychological literature (Anderson, 1990; Rosseel, 2002).

Recently, models of human categorization have begun to draw on another basic advance in density estimation techniques from statistics and machine learning. *Nonparametric Bayesian* methods for density estimation (e.g., Escobar & West, 1995; Neal, 1998) provide ways to adaptively select the effective number of clusters to use in representing a distribution, while allowing the number of possible clusters to remain unbounded. These models are particularly interesting in the context of understanding human category learning, as they offer an alternative to the idea that a single fixed representational strategy (such as forming prototypes or remembering exemplars) is necessary. Nonparametric Bayesian models illustrate how a rational learner could adaptively form a representation based on the distributional properties of the observed stimuli.

The most basic nonparametric Bayesian density estimation method is based on the *Dirichlet process mixture model* (DPMM; Antoniak, 1974). The DPMM assumes that a probability distribution can be represented as an infinite mixture of parametric densities, where the parameters of those densities are generated from a stochastic process known as the Dirichlet process (Ferguson, 1973). While the mathematical definition of the Dirichlet process is complex (for details, see Navarro, Griffiths, Steyvers, & Lee, 2006), its implications are straightforward. When the first stimulus is observed, a cluster (with an associated parametric distribution) is created to represent that stimulus. Each subsequent stimulus is then assigned to either an existing cluster (with probability proportional to the number of stimuli already assigned to that density), or is represented by a new cluster. The result of this process is a probability distribution over partitions of the stimuli into clusters that are each modeled with a single parametric density. This partitioning of the data is equivalent to the assumption in psychological process models that people might represent categories in terms of several clusters that can each be summarized by a prototype (e.g., Kruschke, 1990; Love et al., 2004; Vanpaemel & Storms, 2008).

As with other density estimation methods, the DPMM has a connection to a psychological model. However, in this case the connection is not to a process model, but to a rational model: Anderson's (1990, 1991) rational model of categorization. Anderson considered how the probability distribution associated with a set of categories could be estimated, and independently developed a solution to this problem that was equivalent to the DPMM (the equivalence was first pointed out by Neal, 1998). Recognizing this equivalence makes it possible to use algorithms developed for the DPMM to better approximate optimal performance in Anderson's rational model (Sanborn, Griffiths, & Navarro, 2006) and to develop psychological models that draw on recent generalizations of the Dirichlet process that have been developed in machine learning and statistics, such as the hierarchical Dirichlet process (Teh, Jordan, Beal, & Blei, 2004).

The key idea behind nonparametric Bayesian models – that learners can adaptively form a representation that captures the structure expressed in the observed data – is also applicable in cases that go beyond simple clustering of stimuli based on their features. Shafto, Kemp, Mansinghka, Gordon, and Tenenbaum (2006) introduced a model for learning cross-cutting systems of categories in which a similar principle was used to simultaneously decide how many systems of categories might be relevant to understanding the features of stimuli, and which category each stimulus belongs to within each system. Kemp, Tenenbaum, Griffiths, Yamada, and Ueda (2006) showed how the key ideas behind the DPMM could be extended to data that reflect the relations that exist among a set of stimuli, rather than the features that those stimuli express. This model could discover clusters of stimuli that behaved similarly in relation to other clusters of stimuli, forming abstractions about relational roles that might form a first step towards learning more complex relational systems such as folk theories (Kemp, Tenenbaum, Niyogi, & Griffiths, in press).

Our goal in this chapter is to provide a basic introduction to the ideas behind nonparametric Bayesian models in the context of category learning. To this end, we first give a more formal description of some nonparametric Bayesian models – the Dirichlet process mixture model, the hierarchical Dirichlet process, and related extensions. We then discuss how algorithms for inference in these models can be implemented, describing algorithms proposed both in psychology and in statistics. Finally, we present simple examples illustrating the operation of different algorithms and the prediction of human behavior.

Description

Our description of nonparametric Bayesian models of categorization begins with the Dirichlet process mixture model, since this model provides the simplest illustration of the principles on which other models are based. We then define the hierarchical Dirichlet process, and summarize some of the other ways in which this model has been extended.

The Dirichlet process mixture model

In order to compute the conditional probability distribution over categories specified in Equation 1, we need to estimate a probability distribution over features x and category labels c, $p(x, c)$. For simplicity, we will drop c, since it can be considered another feature of the stimulus (albeit one that should be given greater weight than other features), and just consider how we can estimate the joint distribution of a sequence of stimuli $\mathbf{x}_N = (x_1, \ldots, x_N)$.[1] Like other mixture models, the Dirichlet process mixture model assumes that each x_i was generated from a mixture component that is a parametric density. Intuitively, each mixture component corresponds to a cluster of stimuli that go together. We will use z_i to denote the index of the mixture component from which x_i was generated, and $\mathbf{z}_N = (z_1, \ldots, z_N)$ to indicate the vector of component assignments for all stimuli. Each z_i is just a nominal variable picking out the cluster to which x_i belongs, so $\mathbf{z}_N$ partitions the stimuli into clusters. A simple example helps to clarify the notation. In the box in Figure 1, we have three stimuli that are observed sequentially. Each stimulus has three binary features, but their features will not be important for now. Let us assume that an observer has already seen the first two stimuli, but has not yet seen the third stimulus. $\mathbf{x}_2$ is thus the set of the first two stimuli, x_1 and x_2. If these stimuli are assigned to the same component, the z_i values will be equal, for example $z_1 = z_2 = 1$. The only other alternative for two stimuli is that they are assigned to different components, in which case $z_1 = 1$ and $z_2 = 2$.

Usually, we just observe the stimuli $\mathbf{x}_N$ without being told which clusters they belong to. The joint distribution $p(\mathbf{x}_N)$ is thus obtained by averaging over all possible assignments of stimuli to clusters, with

$$p(\mathbf{x}_N) = \sum_{\mathbf{z}_N} p(\mathbf{x}_N|\mathbf{z}_N)p(\mathbf{z}_N) \tag{2}$$

where $p(\mathbf{x}_N|\mathbf{z}_N)$ indicates the probability of the stimuli under the assignments $\mathbf{z}_N$, and $p(\mathbf{z}_N)$ is a distribution that reflects our ignorance about the cluster assignments. If $p(\mathbf{x}_N|\mathbf{z}_N)$ depends only on which stimuli are assigned to the same clusters, then all vectors of assignments $\mathbf{z}_N$ that result in the same partition of the stimuli will give the same probability to $\mathbf{x}_N$. Consequently, we can define $p(\mathbf{z}_N)$ by specifying a probability distribution over partitions of N objects.

In the DPMM, this distribution over partitions corresponds to a stochastic process known as the *Chinese restaurant process* (CRP; Aldous, 1985; Pitman, 2002). The CRP defines a distribution that has two desirable properties: it makes it possible for any number of clusters to appear in the data, and it gives each partition that has clusters of the same sizes the same probability. Under this process, we imagine that stimuli are assigned to clusters one after another and the probability that stimulus $i + 1$ is assigned to cluster k is

$$P(z_i = k|\mathbf{z}_{i-1}) = \begin{cases} \frac{M_k}{i-1+\alpha} & \text{if } M_k > 0 \text{ (i.e., } k \text{ is old)} \\ \frac{\alpha}{i-1+\alpha} & \text{if } M_k = 0 \text{ (i.e., } k \text{ is new)} \end{cases} \tag{3}$$

where M_k is the number of stimuli that have already been assigned to cluster k, and α is a parameter of the process that determines the probability of generating new clusters. The process gets its unusual name from thinking of stimuli as customers entering a large Chinese restaurant, where the table they choose to sit at corresponds to their cluster assignment. Following this process for all N stimuli results in a distribution $p(\mathbf{z}_N)$ that has an interesting property: the probability of a partition $\mathbf{z}_N$ does not depend on the order in which we assigned stimuli to clusters. This property is known as *exchangeability* (Aldous, 1985).

To illustrate the use of the CRP, assume that the first two stimuli shown in Figure 1 are assigned to the same cluster, $\mathbf{z}_2 = (1, 1)$, when the third stimulus, x_3, is observed. We can then use the CRP to calculate the prior distribution over the cluster assignment for that object, z_3. The CRP gives the prior probability of x_3 belonging to the cluster with the other two stimuli and the prior probability that it is assigned to its own cluster. If $\alpha = 0.5$, then $P(z_3 = 1|\mathbf{z}_2) = \frac{2}{3-1+0.5} = 0.8$ and $P(z_3 = 2|\mathbf{z}_2) = \frac{0.5}{3-1+0.5} = 0.2$. The probability of a new cluster being created increases as α increases.

Up until this point, the features of the stimuli have been unimportant, but to fully specify the DPMM, we also need to define the distribution $p(\mathbf{x}_N|\mathbf{z}_N)$ that links partitions to stimuli (often referred to as the *likelihood*, with the Dirichlet process providing the *prior* on partitions $\mathbf{z}_N$). This is done by assuming that each cluster is associated with a probability distribution over stimuli. The choice of this distribution depends on the properties of the data: with continuous features, a Gaussian distribution may be appropriate, capturing the mean and variance of those features in each cluster; with discrete features, a multinomial distribution can be used to specify the probability of each feature value in each cluster. The distribution for each cluster is characterized by a set of parameters that can be estimated from the stimuli assigned to that cluster, or simply integrated out of the probabilistic model. The stimuli in the simple example shown in Figure 1 are parameterized by three binary features and the likelihoods $p(\mathbf{x}_3|\mathbf{z}_3)$ are calculated using separate multinomial distributions for each cluster. These multinomial distributions are independent for each feature (for details, see Anderson, 1990; Neal, 1998; Sanborn et al., 2006).

When using the DPMM for categorization, we need to be able to compute the probability distribution over features (including the category label) for a novel object. The distribution $p(x, c)$ required to apply Equation 1 is taken to be the *posterior predictive distribution* generated by the DPMM. This distribution is used because we do not know what the appropriate cluster assignments $\mathbf{z}_N$ are, so we have to average over the posterior distribution on cluster assignments, just as we averaged over the prior in computing $p(\mathbf{x}_N)$ in Equation 2. The posterior predictive distribution is obtained by computing the posterior probability of each partition $\mathbf{z}_N$ of the stimuli $\mathbf{x}_N$ that belong to the category, and then averaging the probability of a new stimulus x over the resulting distribution. More formally, dropping c again for convenience, we have

$$p(x|\mathbf{x}_N) = \sum_{\mathbf{z}_N} \sum_{z} p(x|z, \mathbf{z}_N, \mathbf{x}_N) p(z|\mathbf{z}_N) p(\mathbf{z}_N|\mathbf{x}_N) \tag{4}$$

where $p(x|z, \mathbf{z}_N, \mathbf{x}_N)$ is the probability of the stimulus x under the distribution associated with cluster z given the other stimuli in $\mathbf{x}_N$ assigned to z by the partition $\mathbf{z}_N$, $p(z|\mathbf{z}_N)$ is the probability of a stimulus being generated from cluster z given the partition $\mathbf{z}_N$, and $p(\mathbf{z}_N|\mathbf{x}_N)$ is the posterior probability of the partition $\mathbf{z}_N$ given the stimuli $\mathbf{x}_N$. Of the quantities on the right hand side of this equation, $p(x|z, \mathbf{z}_N, \mathbf{x}_N)$ can be computed directly from the specification of the distribution associated with each cluster. The quantity $p(z|\mathbf{z}_N)$ follows directly from the CRP.

In the example shown in Figure 1, computing the posterior predictive distribution $p(x_3|\mathbf{x}_2)$ requires summing over all possible partitions $\mathbf{z}_2$ which are $\mathbf{z}_2 = (1, 1)$ and $\mathbf{z}_2 = (1, 2)$ and summing over all possible assignments of x_3 to clusters given $\mathbf{z}_2$. Though this is straightforward in this case, computing the posterior predictive distribution is computationally expensive for larger numbers of stimuli. The main challenge of performing probabilistic inference using the DPMM is calculating the posterior distribution over partitions $\mathbf{z}_N$ given $\mathbf{x}_N$, because the number of partitions increases rapidly with N. We return to this problem later in the chapter.

The hierarchical Dirichlet process

In Anderson's (1990, 1991) original presentation of the RMC, category labels were taken to be just another feature of the stimuli, as in our presentation of the DPMM. However, other rational analyses of categorization have focused on estimating the conditional distribution over features x for each category c, $p(x|c)$ (e.g., Ashby & Alfonso-Reese, 1995; Nosofsky, 1998; Rosseel, 2002). It is straightforward to use a DPMM to estimate these conditional distributions, but taking a separate mixture model for each category means that the clusters that comprise those categories are taken to be completely disjoint. However, in some cases it may make sense to share those clusters between categories, providing a common vocabulary at a higher level than raw stimuli in which the structure of categories can be expressed. For example, a cluster of tabby cats might be useful in learning the categories of both cats and striped objects. This kind of sharing of clusters between categories can be achieved using the hierarchical Dirichlet process (HDP).

The HDP, introduced by Teh, Jordan, Blei, and Beal (2004), is a straightforward generalization of the basic Dirichlet process. Stimuli are divided into categories, and each category is modeled using a Dirichlet process mixture model (with parameter α). A new stimulus is first compared to all of the clusters in its category, with the prior probability of each cluster determined by Equation 3. If the stimulus 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. The probability of generating a new cluster in this higher-level Dirichlet process is governed by the 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 category, instead of the number of stimuli in each cluster. The new stimulus 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 2.

The HDP provides a way to model probability distributions across categories. Each distribution is a mixture of an unbounded number of clusters, but the clusters can be shared between categories. 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 category and the extent to which clusters are shared between categories are determined by the parameters α and γ. When α is small, each category will have few clusters, but when α is large, the number of clusters will be closer to the number of stimuli. When γ is small, categories are likely to share clusters, but when γ is large, the clusters in each category are likely to be unique.

The extra flexibility provided by the capacity to share clusters between categories means that the HDP can be used to characterize a variety of different schemes for representing the structure of categories as α and γ are varied. When $\alpha \to \infty$ and $\gamma \to \infty$ we obtain an exemplar model, with one cluster per stimulus and no sharing of clusters. When $\alpha \to 0$ and $\gamma \to \infty$ we obtain a prototype model, with one cluster per category and no sharing of clusters. When $\alpha \to \infty$ and γ is free to vary, we obtain a model where each stimulus comes from its own cluster but those clusters are drawn from a Dirichlet process shared between categories, similar to the original scheme for representing categories introduced by Anderson (1990, 1991). The hierarchical Dirichlet process has thus been proposed as a unifying rational model of categorization, containing these other models as special cases and allowing the learner to adaptively select an appropriate representation through estimation of α and γ in a given domain (Griffiths, Canini, Sanborn, & Navarro, 2007).

Other nonparametric Bayesian models

The basic principles behind the nonparametric Bayesian models outlined in this section can be used in any probabilistic model in which categories can be represented in terms of underlying clusters. This means that the nonparametric Bayesian approach can be extended to capture learning from different kinds of data, and forming richer representations of category structure. We will briefly summarize two models that provide examples of such extensions: learning from relations, and learning cross-cutting systems of categories.

Learning cross-cutting categories.

Most approaches to categorization, including the methods we describe above, assume that there is one best way to organize the entities in a given semantic domain. Most natural domains, however, can be represented in multiple ways: animals may be thought of in terms of their taxonomic groupings or their ecological niches, foods may be thought of in terms of their nutritional content or social role; products may be thought of in terms of function or brand; movies may be thought of in terms of their genre or star quality. Another nonparametric Bayesian model, CrossCat (Shafto et al., 2006), discovers multiple systems of categories given information about a domain of entities and their attributes. Each system of entity-categories accounts for a distinct and coherent subset of the observed attributes.

As with the other models we have discussed, CrossCat uses Dirichlet processes as priors on how to partition entities into categories within each system, and how to allocate attributes across systems. The nonparametric formulation allows CrossCat to find appropriate tradeoffs between two kinds of simplicity that are both desirable in a domain theory but tend to compete with each other: minimizing the number of category systems, and minimizing the number of categories within each system. Building an overly simplified model at either of these levels will lead to an overly complex model at the other. CrossCat naturally prefers the theory that is most compact overall, splitting up a category system into two if it will lead to many fewer categories per system, or splitting up a category within a system if it will substantially increase the number of attributes that system can explain.

CrossCat has been shown to discover meaningful semantic concepts in several kinds of data. For instance, given a data set of animal species and their attributes, CrossCat finds three ways to categorize the species: a system of taxonomic classes that accounts for anatomical and physiological properties, a system of ecological classes that accounts for behavioral features (relevant to being a predator or prey, or living in the land, air or sea); and a third system in which almost all species belong to the same class and which explains features that do not vary much, or vary idiosyncratically over this domain (e.g., the color or size of an animal). The model also finds cross-cutting systems of categories that match those identified by human learners in laboratory experiments (Shafto et al., 2006).

Learning from relations.

Traditional approaches to categorization treat each entity individually, simply in terms of the features it has. Richer semantic structure can be found if we can develop methods for effectively learning from complex forms of relational data, where categories are defined in terms of the relations between one another. Nonparametric Bayesian models can be used to solve this problem, discovering systems of related concepts from heterogeneous data sources. One such model, the infinite relational model (Kemp et al., 2006), identifies clusters of objects that not only share similar features, but also participate in similar relations. Given data involving one or more types of entities, their attributes, and relations among them, this model can discover the kinds of entities in each set and the relations between kinds that are possible or likely. For instance, a data set for consumer choice could be characterized in terms of these relations: which consumers bought which products, which features are present in which products, which demographic attributes characterize which users, and so on. The model simultaneously discovers how to cluster each type of entity as well as the regularities in how these clusters are related (e.g., consumers in class X tend to buy products in class Y).

The nonparametric nature of the infinite relational model allows it to automatically discover the appropriate number of categories to be used in describing each type of entity, and to grow the complexity of these categorization systems as new data warrant. This ability to grow representations of appropriate complexity as the observed data grow is especially important in relational settings. When the data concern how entities of different types interact, a choice about how finely to group entities of one type interacts with the analogous choices for all other types those interact with. For example, grouping one type too coarsely may lead to overly fine-grained representation of another type it interacts with. The automatic discovery of clusters of the appropriate granularity produced by this model also provides a way to explain how people might form categories of objects based on the causal relations that hold between them, providing a basic step towards learning a more sophisticated relational theory of a domain (Kemp et al., in press).

Implementation

Nonparametric Bayesian models present a basic challenge for the learner and the modeler: performing probabilistic inference about the values of the latent variables in the model (such as the partitions of stimuli used in the DPMM). Psychological research using these models has explored three algorithms for probabilistic inference. One algorithm, which we call the *local MAP* algorithm, was introduced by Anderson (1990) and motivated by psychological considerations. The two other algorithms – *Gibbs sampling* and *particle filtering* – draw on the statistics literature, and were first applied in a psychological setting by Sanborn, Griffiths, and Navarro (2006). For simplicity, we present these three algorithms just for the DPMM, but the same principles apply when they are used with other models.

The local MAP algorithm

The local MAP algorithm (short for local maximum a posteriori probability) approximates the sum in Equation 4 with just a single partition of the N objects, $\mathbf{z}_N$. This partition is selected by assigning each object to the highest probability cluster as it is observed. The posterior probability that stimulus i was generated from cluster k given the features of all stimuli, along with the cluster assignments $\mathbf{z}_{i-1}$ for the previous $i-1$ stimuli is

$$p(z_i = k|x_i, \mathbf{z}_{i-1}, \mathbf{x}_{i-1}) \quad \propto \quad p(x_i|z_i = k, \mathbf{z}_{i-1}, \mathbf{x}_{i-1})p(z_i = k|\mathbf{z}_{i-1}) \tag{5}$$

where $p(z_i = k|\mathbf{z}_{i-1})$ is given by Equation 3. Under the local MAP algorithm, x_i is assigned to the cluster k that maximizes Equation 5. 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.

To illustrate the local MAP algorithm, we applied it to the simple example of sequentially presented stimuli in Figure 1. As mentioned above, each stimulus is parameterized by three binary features and the likelihood $p(\mathbf{x}|\mathbf{z})$ is calculated using multinomial distributions that are independent for each feature, which is the standard approach for modeling binary data (for details, see Anderson, 1990; Neal, 1998; Sanborn et al., 2006). The local MAP algorithm initially assigns the first observed stimulus to its own cluster. When the second stimulus is observed, the algorithm generates each possible partition: either it is assigned to the same cluster as the first stimulus or to a new cluster. The posterior probability of each of these partitions is calculated and the partition with the highest posterior probability is always chosen as the representation. After the third stimulus is observed, the algorithm produces all possible partitions involving the third stimulus, assuming that the first two stimuli are part of the same cluster. Note that not all possible partitions of the three stimuli are considered, because the algorithm makes an irrevocable choice for the partition of the first two stimuli and the possible partitions on later trials have to be consistent with this choice. The local MAP algorithm will always produce the same final partition for a given sequential order of the stimuli, assuming there are no ties in the posterior probability.

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 discussed above. Figure 4 shows that the posterior distribution over partitions produced by the local MAP is very different from the distribution it attempts to approximate. The local MAP results in a single outcome, while the exact posterior distribution has some non-zero probability for every outcome. The partition the local MAP selects depends on the order in which the stimuli are observed, and this order dependence is perhaps stronger than order dependence human participants display (see Sanborn et al., 2006).

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

$$p(z_i = k|x_i, \mathbf{z}_{-i}, , \mathbf{x}_{-i}) \quad \propto \quad p(x_i|z_i = k, \mathbf{z}_{-i}, \mathbf{x}_{-i})p(z_i = k|\mathbf{z}_{-i}) \tag{6}$$

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

Equation 6 is extremely similar to Equation 5, although it gives the probability of a cluster based on 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 stimuli can be rearranged so that any particular stimulus is considered the last stimulus. Hence, we can use Equation 3 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. 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) and is illustrated for the simple example in Figure 1. First, an initial assignment of stimuli to clusters is chosen, with a convenient choice being all stimuli assigned to a single cluster. Unlike the local MAP algorithm, Gibbs sampling is not a sequential algorithm; all stimuli must be observed before it can be used. Next, we choose a single stimulus and consider all possible reassignments of that stimulus to clusters, including not making a change in assignments or assigning the stimulus to a new cluster. Equation 6 gives the probability of each partition and one of the partitions is sampled based on its posterior probability, making this algorithm stochastic, unlike the local MAP. The stochastic nature of the algorithm is evident in the example in Figure 3, because the final circled assignment has lower probability than the alternatives. The example shows a single iteration of Gibbs sampling, in which each stimulus is cycled through and reassigned. The algorithm goes through many iterations, with the output of one iteration the input to the next. 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 from the posterior distribution $p(\mathbf{z}_N|\mathbf{x}_N)$. Averaging over these samples thus provides a way to approximate the sum over $\mathbf{z}_N$ in Equation 4 without needing to calculate the posterior probability of all partitions. Over 10,000 iterations the Gibbs sampler produces a good approximation to the exact posterior, as shown in Figure 4.

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, where ℓ is the index of the particle. 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.

$$p(\mathbf{z}_i|\mathbf{x}_i) \approx \frac{1}{m}\sum_{\ell=1}^{m} \delta(\mathbf{z}_i, \mathbf{z}_i^{(\ell)}) \tag{7}$$

where $\delta(\mathbf{z}, \mathbf{z}')$ is 1 when $\mathbf{z} = \mathbf{z}'$, and 0 otherwise. If Equation 7 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 stimuli $\mathbf{x}_i$ in the following manner:

$$\begin{aligned} p(\mathbf{z}_{i+1}|\mathbf{x}_i) &= \sum_{\mathbf{z}_i} p(\mathbf{z}_{i+1}|\mathbf{z}_i)p(\mathbf{z}_i|\mathbf{x}_i) \\ &\approx \sum_{\mathbf{z}_i} p(\mathbf{z}_{i+1}|\mathbf{z}_i)\frac{1}{m}\sum_{\ell=1}^{m}\delta(\mathbf{z}_i, \mathbf{z}_i^{(\ell)}) \\ &= \frac{1}{m}\sum_{\ell=1}^{m} p(\mathbf{z}_{i+1}|\mathbf{z}_i^{(\ell)}) \end{aligned} \tag{8}$$

where $p(\mathbf{z}_{i+1}|\mathbf{z}_i)$ is given by Equation 3. We can then incorporate the information conveyed by the features and label of stimulus $i+1$, arriving at the approximate posterior probability

$$\begin{aligned} p(\mathbf{z}_{i+1}|\mathbf{x}_{i+1}) &\propto p(x_{i+1}|\mathbf{z}_{i+1}, \mathbf{x}_i)p(\mathbf{z}_{i+1}|\mathbf{x}_i) \\ &\approx \frac{1}{m}\sum_{\ell=1}^{m} p(x_{i+1}|\mathbf{z}_{i+1}, \mathbf{x}_i)p(\mathbf{z}_{i+1}|\mathbf{z}_i^{(\ell)}) \end{aligned} \tag{9}$$

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.

The particle filter for the simple example from Figure 1 is illustrated in Figure 3. The particle filter for the DPMM is initialized with the first stimulus assigned to the first cluster for all m particles, in this case $m = 2$. On observing each new stimulus, the distribution in Equation 9 is calculated, based on the particles sampled in the last trial. Like the local MAP, the particle filter updates the partition as each new stimulus is observed, and like the local MAP, only new partitions that are consistent with the previous choices made by the algorithm are considered. This consistency can be seen in the potential partitions when the third stimulus is observed in Figure 3: each descendant is consistent with the partition choices made by its ancestor. The particle filter differs in two ways from the local MAP algorithm. The first is that the choice of new partitions is stochastic instead of deterministic. The particle filter algorithm samples new partitions based on their posterior probabilities instead of always selecting the partition with the maximum probability. Stochastic selection generally produces more accurate approximation of the exact distribution, which can be seen in Figure 4. A particle filter with $m = 1$ particles is equivalent to the local MAP algorithm, except that the new partition is sampled instead of deterministically selected. Over 1,000 runs of the algorithm, the single-particle particle filter produces a far closer approximation to the exact distribution than the local MAP. The second difference is that multiple particles means that multiple partitions can be used instead of the single partition passed forward by the local MAP. The m partitions are selected without regard for ancestry, allowing a partition that was selected for the early observations to die out as the descendants of other partitions replace it. A large enough set of particles will prevent this algorithm from being sent down the wrong track, which is a danger for the local MAP. The particle filter in Figure 4 used 10 particles with 1,000 repetitions. The results of this algorithm are indistinguishable from the exact solution, and are both a better approximation than the local MAP or the single-particle particle filter.

Example

We present the following example to demonstrate how nonparametric models can capture certain aspects of human categorization that are not well-explained by either prototype or exemplar models. A more detailed analysis of this example can be found in Griffiths et al. (2007). In an experiment conducted by Smith and Minda (1998), a prototype model was found to provide a better explanation for human performance on a categorization task during the early stages of learning, while an exemplar model was found to be a better fit to the later stages. The authors presented these findings to dispute the sentiment that exemplar models provided a dominant account of human categorization over prototype models. Instead, they argued, people seemed to use strategies corresponding to both of these models, perhaps shifting between them over time. Due to the natural ability of the HDP to interpolate between exemplar- and prototype-style representations as warranted by the observed data, it seems a natural candidate to explain the results found by Smith and Minda.

We focus on the non-linearly separable structure explored in Experiment 2 of Smith and Minda (1998). The experiment consisted of a series of trials in which 16 participants were asked to guess whether six-letter nonsense words belonged to Category A or Category B. Each letter of the words took on one of two possible values, producing the binary feature representations of the two categories shown in Table 1. Each category contains one prototypical stimulus (000000 vs. 111111), five stimuli with five features in common with the prototype, and one stimulus with only one feature in common, which we refer to as an "exception". No linear function of the features can correctly classify every stimulus, meaning that a prototype model will not be able to distinguish between the categories exactly. Participants received feedback after each trial and were tested a total of 40 times on each of the 14 stimuli. The trials were split into 10 segments, each consisting of 4 trials of each of the 10 stimuli. The results of the experiment are shown in Figure 5 (a). The exceptions were initially identified as belonging to the wrong category, with performance improving over time.

As in the previous example, in applying the computational models, we can use a Beta prior distribution for each dimension of each cluster, allowing us to integrate out the parameters of each cluster to obtain a tractable distribution for $p(x|z)$. The three models that were tested, a prototype model, an exemplar model, and the DPMM, were exposed to the same training stimuli as the human participants and used to categorize each stimulus after each segment of four exposures to the stimuli. The inferences of the prototype and exemplar models can be computed exactly, but because the DPMM involves a summation over all the partitions of the stimuli into clusters, we resorted to the Gibbs sampling procedure described above. The results of the three models are plotted in Figure 5. Only the DPMM captures the cross-over effect for the exception stimuli, where they are categorized incorrectly at first and are learned gradually over time. This effect is due to the DPMM's ability to shift its representation between the prototype and exemplar styles. At first, it is more likely to use a single cluster to represent each category, as a prototype model would. After repeated exposure to the stimuli, the DPMM becomes more likely to split the exception stimuli into their own individual clusters, moving closer to an exemplar representation.

Relationship to other models

The relationships between Bayesian and other approaches can be described in terms of a few key concepts. In the first instance, the Bayesian framework described sits at the computational level of analysis. As with the machine learning approach suggested by Iba and Langley, we chose to phrase the problem in probabilistic language, but the model can be converted into an MDL style description (Pothos, Chater & Hines) without difficulty. The important similarity here is that all three models share the goal of expressing human inferences primarily in terms of the statistical characteristics of the problem to be solved. This is in sharp contrast to Ashby, Paul and Maddox, for instance, who describe the only model in this volume that has strong ties to the neural implementation level of analysis. However, the majority of the other chapters operate at the algorithmic level, and rely heavily on mechanistic psychological processes. Process assumptions can operate at a representational level, as is the case for prototype abstraction (Minda & Smith) and exemplar storage (Nosofsky), but can also describe learning rules such as backpropagation (Kruschke) and Hebbian learning (Harris & Rehder). Bayesian theories have typically stayed away from making strong commitments at the algorithmic level, or exploring the cognitive consequences of different algorithms for Bayesian learning and inference, but this is changing. As described in the section on implementation, we suggest that particle filtering, importance sampling and other methods may provide new possibilities for algorithmic-level Bayesian modeling that can bridge the gap between rational analyses and psychological processes.

Several of the contributions relate to our chapter through variations on the theme of hierarchical learning. The HDP model is hierarchical in the sense that it infers a generic clustering of entities in the environment in addition to the various category-specific distributions. It is via this mechanism that the HDP unifies prototypes (Minda & Smith) and exemplars (Nosofsky) with the original rational model (Anderson, 1991) and other mixture models, and the clustering of stimuli within categories is similar to other models of categorization such as SUSTAIN (Love et al., 2004). However, the idea has considerably more generality. In a related line of work, Bayesian theories exploit the fact that the representations learned at the top level act as priors over category distributions at the low level. In this respect, there is a strong link to the learned hierarchies in the Iba and Langley contribution and to the prior knowledge constraints described by Harris and Rehder: there are now several Bayesian models that focus on how such constraints can be learned from experience (e.g., Kemp, Perfors, & Tenenbaum, 2007; Kemp &

Tenenbaum, 2008; Shafto et al., 2006). As an example, Kemp et al. (2007) show how hierarchical learning provides an explanation for how children learn inductive biases about object categories. Over shorter time scales, learned selective attention (Kruschke) can also be considered to exemplify the same kind of learning, insofar as people learn the assignments of items to specific categories, as well as general knowledge about the extensions of categories along different dimensions. Hierarchical learning accounts for standard attentional phenomena such as the condensation-filtration effect (Navarro, 2006), as well as how people can make inferences about entirely novel categories (Perfors & Tenenbaum, 2009). Within the Bayesian framework, these things are all characterized as different examples of hierarchical learning, though defined over different domain sizes, complexities and time scales.

Other contributions raise different issues. The multiple system model described by Ashby, Paul and Maddox raises the question of how best to unify rule-based learning with graded, probabilistic categorization models from a computational perspective. The HDP model described here is probabilistic, though other Bayesian models have a rule-like character through their reliance on deterministic "consequential sets" (Shepard, 1987; Tenenbaum & Griffiths, 2001; Navarro, 2006), and others use formal grammars to place priors over logical rules (Goodman, Tenenbaum, Feldman, & Griffiths, 2008). An open question in this context is how to integrate these different kinds of mental representation, or whether deterministic rules and probabilistic categories should be left as two fundamentally distinct learning systems. Some prospects for unification exist. In a different context Maas and Kemp (2009) use priors constructed to induce a bias towards determinism, while Jaynes (2003) argues that some "improper" priors mimic the kind of one-shot learning of logical rules that occurs in the physical sciences. Nevertheless, this remains an avenue for future work.

The simplicity model (Pothos, Chater & Hines) raises a different question in relation to the HDP model, regarding the nature of Ockham's razor. The scheme used to index partitions is a two-part code, encoding the number of categories k using the same number of bits regardless of the value of k. Via the equivalence between codelength functions and probability distributions (e.g. Grünwald, 2007) this translates to a uniform prior over the number of categories k. Similarly, all partitions of the same cardinality are equally likely. This is very different to the prior over partitions in the Dirichlet process, which has a bias toward small k, and is non-uniform even for given k. The Dirichlet process supplies an explicit Ockham's razor (simplicity through the prior), whereas the simplicity model prefers fewer categories only to the extent that a simpler clustering provides the better account of the data (simplicity through the likelihood). Although in practice the two approaches often behave similarly (e.g., compare Lee & Navarro, 2005 to Navarro & Griffiths, 2008), there is a deeper theoretical question at stake. Do we possess genuine pre-existing biases to prefer simple categorizations, or do our preferences emerge because simple models just work better? This is another open question for future research.

Future directions

Nonparametric Bayesian models provide a way to analyze how an ideal learner would solve categorization problems that unifies previous models and creates the opportunity to define new models of human category learning. The "nonparametric" aspect of this approach provides flexibility, making it possible to entertain hypotheses about the structure of categories that have unbounded complexity, while the "Bayesian" aspect provides a framework for making statistical inferences about how much complexity is warranted given the observed data. We anticipate two important future directions for this approach to category learning, each building on one of these two aspects of the approach.

The first future direction is expanding the scope of nonparametric Bayesian models. Many of the models discussed in this chapter focus on the traditional task of learning to classify objects as belonging to a small number of non-overlapping categories. Despite the long history of research into human categorization, there still exist many behaviors and techniques used by people in category learning settings that have yet to be formally studied and modeled. These include the transfer of information from one category to another to increase learning rate, the automatic inference of hierarchically-organized category taxonomies, and the exploitation of logical (AND/OR) relationships between categories. In future work, we intend to experimentally study each of these behaviors and model them using extensions of the nonparametric Bayesian models considered here. For example, the HDP model with both α and γ positive and finite allows each category to be modeled as a Dirichlet process mixture model with the underlying clusters being shared among all categories, serving as a model for transfer learning. Combining a prior distribution over tree structures with the natural formulation of the HDP model on trees could serve as a model for the automatic inference of taxonomies. Finally, we envision an extension of the HDP which allows each category to be algebraically related to the others; this model could capture the effects of telling a learner that categories have certain logical relationships to each other.

The second future direction is capturing the effects of prior knowledge on category learning.

Psychological research has shown that people are strongly affected by their knowledge of the world when learning new categories, with categories that are consistent with prior knowledge being easier to learn. These effects are obtained in experiments that use meaningful stimuli that draw on the real-world knowledge of human learners, such as intuitions about the factors that influence the inflation of balloons (Pazzani, 1991), the properties of different types of buildings (Heit & Bott, 2000), the definition of honesty (Wattenmaker, Dewey, Murphy, & Medin, 1986), and the properties of vehicles (Murphy & Allopenna, 1994). Only a small number of computational models of knowledge effects in category learning exist (Rehder & Murphy, 2003; Heit & Bott, 2000), and these models have been developed with the more traditional psychological goals of understanding the mechanisms underlying this process. Developing probabilistic models that can account for knowledge effects in category learning provides the opportunity to discover how such knowledge *should* be used, and to generalize the resulting insights to develop better machine learning systems. Thinking about categorization in terms of density estimation lays the foundation for exploring these deeper questions about human cognition, and the opportunity to draw on tools from artificial intelligence and statistics in formalizing the prior knowledge that guides category learning.

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.

Bishop, C. M. (2006). *Pattern recognition and machine learning.* New York: Springer.

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.

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

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*, 108-154.

Griffiths, T. L., Canini, K. R., Sanborn, A. N., & Navarro, D. J. (2007). Unifying rational models of categorization via the hierarchical Dirichlet process. In *Proceedings of the Twenty-Ninth Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum.

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun (Ed.), *Cambridge handbook of computational cognitive modeling.* Cambridge: Cambridge University Press.

Griffiths, T. L., Sanborn, A. N., Canini, K. R., & Navarro, D. J. (2008). Categorization as nonparametric Bayesian density estimation. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind.* Oxford: Oxford University Press.

Grünwald, P. D. (2007). *The minimum description length principle.* Cambridge, MA: MIT Press.

Heit, E., & Bott, L. (2000). Knowledge selection in category learning. In D. L. Medin (Ed.), *The psychology of learning and motivation* (Vol. 39, p. 163-199). San Diego, CA: Academic Press.

Jaynes, E. T. (2003). *Probability theory: The logic of science.* Cambridge, UK: Cambridge University Press.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical Bayesian models. *Developmental Science*, *10*, 307-321.

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, *105*, 10687-10692.

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T., & Ueda, N. (2006). Learning systems of concepts with an infinite relational model. In *In proceedings of the 21st national conference on artificial intelligence.*

Kemp, C., Tenenbaum, J. B., Niyogi, S., & Griffiths, T. L. (in press). A probabilistic model of theory formation. *Cognition*.

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

Lee, M. D., & Navarro, D. J. (2005). Minimum description length and psychological clustering models. In P. D. Grünwald, I. J. Myung, & M. A. Pitt (Eds.), *Advances in minimum description length: Theory and applications* (p. 355-384). Cambridge, MA: MIT Press.

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

Maas, A. L., & Kemp, C. (2009). One-shot learning with Bayesian networks. *Proceedings of the 31st Annual Conference of the Cognitive Science Society.*

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

McLachlan, G. J., & Basford, K. E. (1988). *Mixture models.* New York: Marcel Dekker.

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

Murphy, G. L., & Allopenna, P. D. (1994). The locus of knowledge effects in concept learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 904-919.

Navarro, D. J. (2006). From natural kinds to complex categories. *Proceedings of the 28th Annual Conference of the Cognitive Science Society.*

Navarro, D. J., & Griffiths, T. L. (2008). Latent features in similarity judgments: A nonparametric Bayesian approach. *Neural Computation*, *20*, 2597-2628.

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. (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.

Pazzani, M. J. (1991). Influence of prior knowledge on concept acquisition: Experimental and computational results. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 416-432.

Perfors, A., & Tenenbaum, J. B. (2009). Learning to learn categories. *Proceedings of the 31st Annual Conference of the Cognitive Science Society.*

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

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

Rehder, B., & Murphy, G. L. (2003). A knowledge-resonance (KRES) model of category learning. *Psychonomic Bulletin & Review*, *10*, 759-784.

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.

Shafto, P., Kemp, C., Mansinghka, V., Gordon, M., & Tenenbaum, J. B. (2006). Learning cross-cutting systems of categories. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society.* Mahwah, NJ: Erlbaum.

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

Silverman, B. W. (1986). *Density estimation for statistics and data analysis.* 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. (2008). In search of abstraction: The varying abstraction model of categorization. *Psychonomic Bulletin & Review*, *15*, 732-749.

Wattenmaker, W. D., Dewey, G. I., Murphy, T. D., & Medin, D. L. (1986). Linear separability and concept learning: Context, relational properties, and concept naturalness. *Cognitive Psychology*, *18*, 158-194.

Author Note

TLG and KRC were supported by grant number FA9550-07-1-0351 from the Air Force Office of Scientific Reesarch. ANS was supported by a Royal Society Fellowship and the Gatsby Charitable Foundation. DJN was supported by an Australian Research Fellowship (ARC grant DP-0773794). This chapter summarizes and expands on material previously presented in Griffiths, Sanborn, Canini, and Navarro (2008).

Footnotes

[1] Throughout this chapter, we use boldface to indicate that a variable is a vector (e.g., $\mathbf{x}_N$), and and italics to indicate that a variable is a scalar (e.g., x_i).

Table 1
Categories A and B from Smith & Minda (1998).

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

Figure Captions

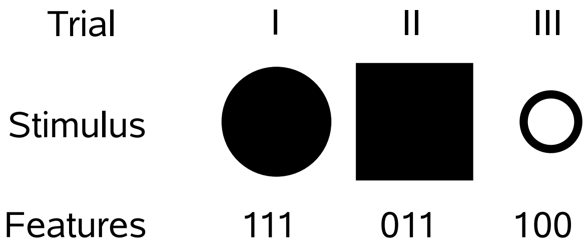
Figure 1. Example stimuli for a categorization experiment. Each stimulus is presented on one of three trials, and possesses three binary features.

Figure 2. Illustration of the hierarchical Dirichlet process. 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.

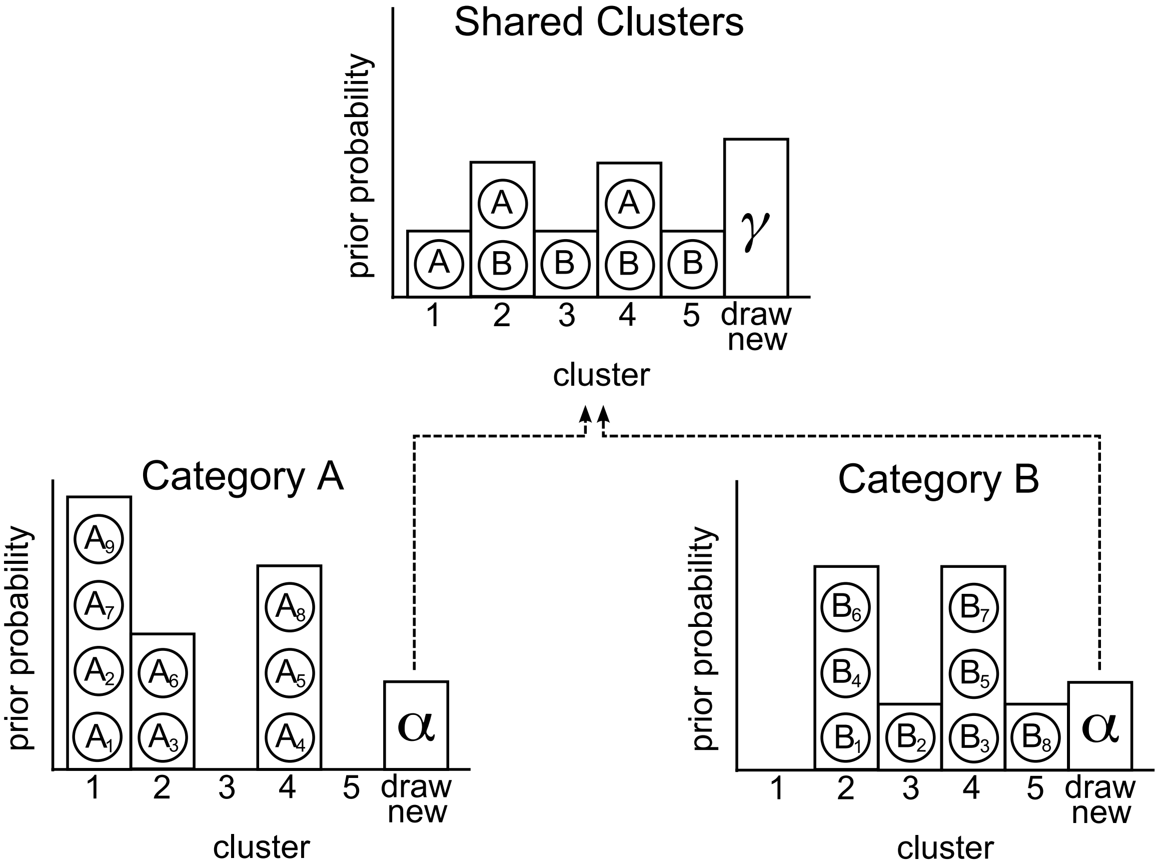
Figure 3. Illustration of the local MAP, particle filtering, and Gibbs sampling approximation algorithms. All three algorithms are applied to the stimuli shown in Figure 1. Each algorithm starts on the left side with an initial partition of the stimuli. Each box is a partition that contains one or more stimuli and the presence of a separating vertical line indicates that the stimuli belong to different clusters. The children of the initial (leftmost) partition are the partitions under consideration in the next step of each algorithm. These children partitions are all the possible reassignments of the stimulus marked by the arrow. Numbers underneath each partition show the posterior probability of that partition. Not all possible paths are followed, with Gibbs sampling and particle filtering choosing partitions to continue stochastically, while local MAP always chooses the partition with the maximum posterior probability. The partitions circled in red are the algorithms' outcomes.

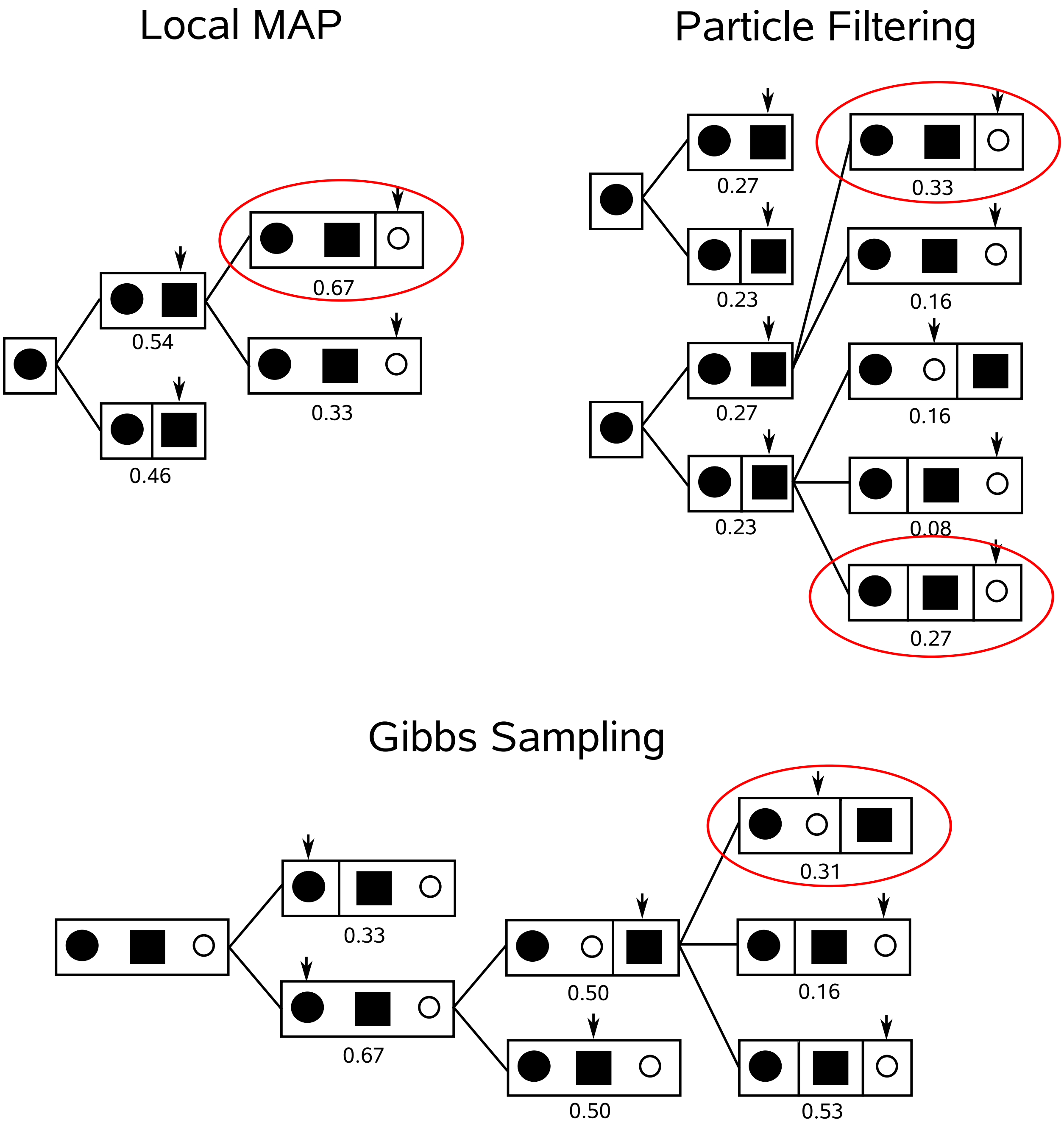
Figure 4. Results of the approximation algorithms compared to the exact posterior. The five bar groupings correspond to the five possible partitions of the three stimuli in Figure 3. The bars within each grouping correspond to the approximation algorithms outlined in the text. Standard error bars are provided for the Gibbs sampling, particle filter, and single-particle particle filter algorithms.

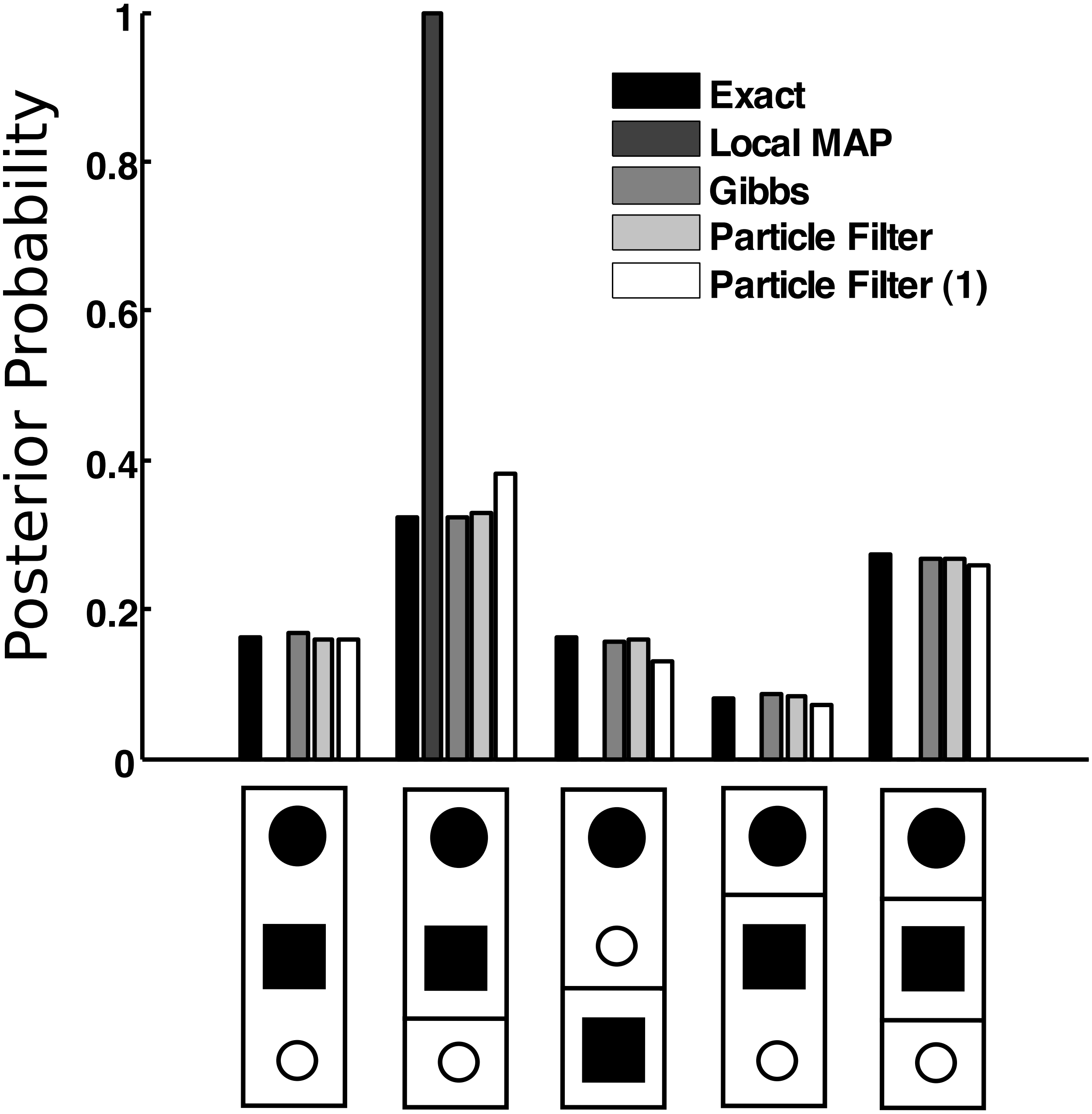
Figure 5. Human data and model predictions for Smith & Minda (1998, Experiment 2). (a) Human performance. (b) Prototype model. (c) Exemplar model. (d) DPMM. For all panels, white plot markers are stimuli in Category A, and black are in Category B. Triangular markers correspond to the exceptions to the prototype structure, i.e., stimuli 111101 and 000100, respectively.

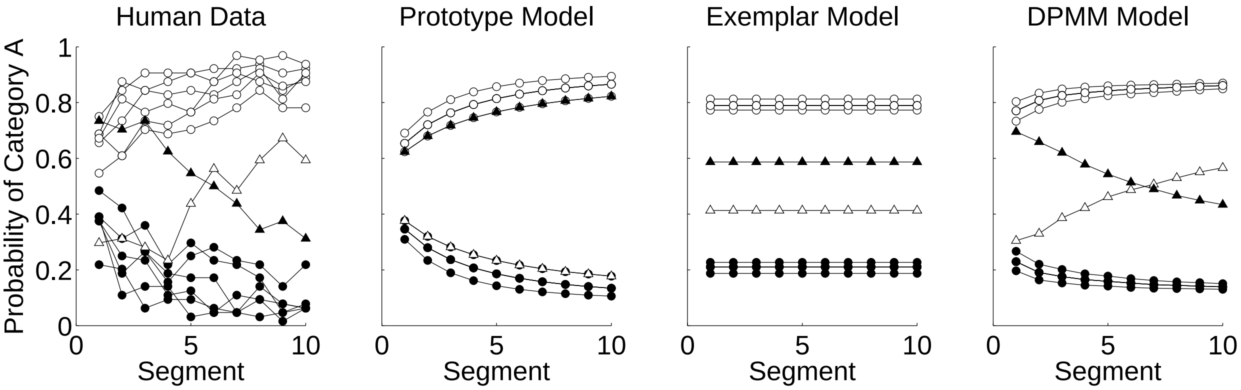


Trial	I	II	III
Stimulus	●	■	○
Features	111	011	100


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


Local MAP
0.54
0.46
0.67
0.33
Particle Filtering
0.27
0.23
0.27
0.23
0.33
0.16
0.16
0.08
0.27
Gibbs Sampling
0.33
0.67
0.50
0.50
0.31
0.16
0.53


Posterior Probability
0
0.2
0.4
0.6
0.8
1
Exact
Local MAP
Gibbs
Particle Filter
Particle Filter (1)


Human Data
Prototype Model
Exemplar Model
DPMM Model
Probability of Category A
0
0.2
0.4
0.6
0.8
1
0
5
10
Segment