# A More Rational Model of Categorization

**Adam N. Sanborn (asanborn@indiana.edu)**
Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN 47405, USA

**Thomas L. Griffiths (tom_griffiths@brown.edu)**
Department of Cognitive and Linguistic Sciences, Brown University, Providence, RI 02912, USA

**Daniel J. Navarro (daniel.navarro@adelaide.edu.au)**
School of Psychology, University of Adelaide, Adelaide SA 5005, Australia

**Abstract**

The rational model of categorization (RMC; Anderson, 1990) assumes that categories are learned by clustering similar stimuli together using Bayesian inference. As computing the posterior distribution over all assignments of stimuli to clusters is intractable, an approximation algorithm is used. The original algorithm used in the RMC was an incremental procedure that had no guarantees for the quality of the resulting approximation. Drawing on connections between the RMC and models used in nonparametric Bayesian density estimation, we present two alternative approximation algorithms that are asymptotically correct. Using these algorithms allows the effects of the assumptions of the RMC and the particular inference algorithm to be explored separately. We look at how the choice of inference algorithm changes the predictions of the model.

Category learning is one of the most extensively studied aspects of human cognition, with computational models that range from strict prototypes (e.g., Reed, 1972) to full exemplar models (e.g., Medin & Schaffer, 1978; Nosofsky, 1986). Recent work has emphasized the "rational" statistical basis of these models (Ashby & Alfonso-Reese, 1995), noting that prototype and exemplar models correspond to different approaches to the "density estimation" problem, in which one infers the probability distribution over stimuli associated with a category. These connections help to explain the success of the models and suggest new directions in which they can be extended. In this paper we discuss the statistical foundations of Anderson's (1990) model, one of the first explicitly rational approaches to category learning, articulating the relationship between this model and nonparametric Bayesian density estimation. Recognizing this relationship provides the opportunity to explore variations on the original model.

The rational model of categorization (RMC; Anderson, 1990, 1991) accounts for many of the basic categorization phenomena, although it is not without flaws (e.g., Murphy & Ross, 1994). The RMC uses a flexible representation that can interpolate between prototypes and exemplars by clustering stimuli into groups,[1] adding new clusters to the representation 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. The representation can thus grow 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.

There are two steps involved in defining any rational model of cognition: first, identifying the underlying computational problem, and second, showing how people might solve that problem given cognitive constraints. When Anderson (1990, 1991) introduced the RMC, he assumed two strong cognitive constraints: that stimuli are assigned to clusters sequentially, and that these assignments are fixed once they are made. He then introduced an algorithm for assigning stimuli to clusters that satisfied these constraints. However, without considering other algorithms for solving this problem it is impossible to tell whether the model's predictions result from casting categorization as a Bayesian inference about the clustering of objects, or from the assumptions about the way in which people perform this inference.

Connections between the RMC and nonparametric Bayesian density estimation provide a way of defining alternative algorithms for assigning stimuli to clusters. The two algorithms we present here both asymptotically approximate the Bayesian posterior distribution over assignments of stimuli to clusters, thus resulting in a "more rational" model of categorization. With these algorithms, the assumptions of the statistical model used in the RMC are no longer conflated with cognitive constraints, and can be tested directly. These algorithms also suggest a novel class of psychologically plausible procedures for performing approximate Bayesian inference under a range of cognitive constraints. To evaluate these ideas, we examine how well the different algorithms approximate both the true posterior distribution and human judgments using two data sets: the classic experiment of Medin and Schaffer (1978) and results on order-sensitivity reported by Anderson (1990).

## The rational model of categorization

According to the RMC, categorization is a special case of feature induction, in which the learner uses the observed features of a stimulus to predict its unobserved features, using the previous stimuli to guide the prediction. Since the model treats category labels as features, these labels are the obvious features to predict, but other features can be predicted as well. It is assumed that each stimulus belongs to a single cluster, and that the features of a stimulus are generated by the cluster to which it belongs.

[1] Anderson (1990, 1991) refers to these groupings of stimuli as "categories", but since they do not necessarily correspond to the category labels we will refer to them as "clusters".

If $x_n$ denotes a partition of the $n$ stimuli into clusters, and $F_n$ denotes all observed features of these $n$ stimuli, the probability that the (unobserved) target feature for the $n$th object has value $j$ is computed by summing over all partitions,

$$P(j|F_n) = \sum_{x_n} P(j|x_n, F_n)P(x_n|F_n) \quad (1)$$

where $P(x_n|F_n)$ is the *posterior probability* of a partition $x_n$ given $F_n$. This posterior probability can be obtained via Bayes' rule, with

$$P(x_n|F_n) = \frac{P(F_n|x_n)P(x_n)}{\sum_{x'_n} P(F_n|x'_n)P(x'_n)} \quad (2)$$

where $P(F_n|x_n)$ is the *likelihood*, the probability of the set of observed features given the partition $x_n$, and $P(x_n)$ is the prior probability of that partition. The sum in Equation 1 and the denominator of Equation 2 are intractable for large $n$, as the number of partitions grows rapidly with the number of stimuli.[2] Consequently, an approximate inference algorithm is needed.

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 (i.e., the maximum *a posteriori* or MAP cluster) given only the previous assignments.

Computing the posterior probability of a cluster assignment for a new stimulus, given the assignments of the previous stimuli, is straightforward. Using the notation from Anderson (1991), the posterior probability that stimulus $i+1$ was generated from cluster $k$ is

$$P(k|F_{i+1}) = \frac{P(F_{i+1}|k)P(k)}{\sum_k P(F_{i+1}|k)P(k)} \quad (3)$$

In this expression $P(F_{i+1}|k)$ is the probability of the set of observed features given the assignment of the stimulus to cluster $k$, $P(k)$ is the prior probability that the stimulus was generated from cluster $k$, and all probabilities are implicitly conditioned on the cluster assignments for the previous stimuli. We discuss the likelihood in greater detail below, and focus here on the prior $P(k)$. In addition to placing a distribution over existing clusters, the prior used in the RMC allows a new stimulus to be generated from a new cluster. Specifically, the prior probability of cluster $k$ is

$$P(k) = \begin{cases} \frac{cn_k}{(1-c)+ci} & n_k > 0 \text{ (i.e., } k \text{ is old)} \\ \frac{(1-c)}{(1-c)+ci} & n_k = 0 \text{ (i.e., } k \text{ is new)} \end{cases} \quad (4)$$

where $n_k$ is the number of stimuli in cluster $k$, and $c$ is the probability that any two stimuli belong to the same cluster, which Anderson (1990, 1991) calls the *coupling probability*. If we imagine each cluster assignment being drawn sequentially from this prior, it can be shown that the resulting distribution on partitions of $n$ stimuli gives each partition $x_n$ probability

$$P(x_n) = \frac{(1-c)^s c^{n-s}}{\prod_{i=0}^{n-1}[(1-c)+ci]} \prod_{k=1}^{s}(n_k - 1)! \quad (5)$$

where $s$ is the number of clusters in the partition.

## Dirichlet process mixture models

The problem of predicting an arbitrary feature of a stimulus can be solved by estimating the joint probability of the features of a set of stimuli. This is the statistical problem of *density estimation*. In Bayesian statistics, this problem is addressed by defining a prior distribution over a set of possible densities, and then updating this distribution with the observed data to obtain a posterior distribution over densities. In *nonparametric* Bayesian statistics, the goal is to define a prior that includes as broad a range of densities as possible, so that complex densities can be inferred if they are warranted by the data. The standard model used to solve this problem is called the *Dirichlet process mixture model* (DPMM; Antoniak, 1974; Neal, 1998).

The key idea behind the DPMM is to assume that observations are partitioned into clusters, with the probability of their features depending only on their cluster membership. The prior probability of a partition is

$$P(x_n) = \frac{\alpha^s}{\prod_{i=0}^{n-1}[\alpha+i]} \prod_{k=1}^{s}(n_k - 1)! \quad (6)$$

where $\alpha$ is the concentration parameter of the Dirichlet process. This distribution over partitions can be produced by a simple sequential stochastic process (Blackwell & MacQueen 1973). If observations are assigned to clusters one after another and the probability that observation $i+1$ is assigned to cluster $k$ is

$$P(k) = \begin{cases} \frac{n_k}{i+\alpha}, & n_k > 0 \text{ (i.e., } k \text{ is old)} \\ \frac{\alpha}{i+\alpha}, & n_k = 0 \text{ (i.e., } k \text{ is new)} \end{cases} \quad (7)$$

we obtain Equation 6 for the probability of the resulting partition. This distribution has a number of nice properties, including *exchangeability*: the probability of a partition is unaffected by the order in which the observations are received (Aldous, 1985).

It should be apparent from our description of the DPMM that it is similar in spirit to the probabilistic

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

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 5 and 6 are equivalent, as are Equations 4 and 7. Anderson (1990, 1991) (impressively) thus independently discovered one of the most celebrated models in nonparametric Bayesian statistics, deriving this distribution from first principles. Recognizing the connection between the DPMM and the RMC makes it possible to go beyond the assumptions behind the RMC. In particular, we can explore alternatives to the local MAP algorithm. In the remainder of the paper, we draw on the extensive literature on inference for the DPMM to offer two alternative algorithms for the RMC that offer asymptotically accurate approximations to Equation 1.

## Alternative inference algorithms

Equation 1 gives the complete Bayesian solution to the problem of prediction under the DPMM. One way to approximate the intractable sum over partitions is to use Monte Carlo methods, with

$$\sum_{x_n} P(j|x_n, F_n)P(x_n|F_n) \approx \frac{1}{m}\sum_{\ell=1}^{m} P(j|x_n^{(\ell)}, F_n) \quad (8)$$

where $x_n^{(1)}, \ldots, x_n^{(m)}$ are $m$ samples from $P(x_n|F_n)$, 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 $P(x_n|F_n)$ is not straightforward – even computing the posterior distribution requires an intractable sum – the two algorithms use more sophisticated Monte Carlo methods to generate a set of samples.

### Gibbs sampling

The approximate inference algorithm most commonly used for the DPMM is Gibbs sampling, a Markov chain Monte Carlo 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 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.

To describe this algorithm in more detail, we need to introduce some new notation. Let $Z_n = (z_1, \ldots, z_n)$ be a vector of cluster assignments for a set of $n$ stimuli, with each stimulus being assigned to one of $s$ clusters. Any vector of cluster assignments corresponds to a partition, $x_n$, so we can define our algorithm directly in terms of $z_1, \ldots, z_n$. The conditional probability of the assignment of stimulus $i$ given the assignments of all other stimuli and all observed features is

$$P(z_i|Z_{-i}, F_n) \propto P(f_i|z_i, Z_{-i}, F_{-i})P(z_i|Z_{-i}) \quad (9)$$

where $Z_{-i}$ is the assignments of all stimuli other than stimulus $i$, $f_i$ are the observed features of $i$, and $F_{-i}$ are the observed features of all other stimuli besides $i$.

The interesting term in Equation 9 is $P(z_i|Z_{-i})$. Due to exchangeability, the order of the observations can be rearranged so that any particular observation is considered the last observation. Hence, we can use Equation 7 to compute $P(z_i|Z_{-i})$, with old clusters receiving probability in proportion to their popularity, and a new cluster being chosen with probability determined by $\alpha$ (or, equivalently, $c$). The other term, $P(f_i|z_i, Z_{-i}, F_{-i})$, is the probability of the features of stimulus $i$ under the partition that results from this choice of $z_i$, and depends on the nature of the features. We discuss this in greater detail later in the paper.

The Gibbs sampling algorithm for the DPMM (Neal, 1998) is now straightforward. First, an initial assignment of stimuli to clusters is chosen. In the simulations, we simply assign all stimuli to a single cluster. Next, we cycle through all stimuli, sampling a cluster assignment from the distribution specified by Equation 9. This step is repeated, with each cycle potentially producing a new partition of the stimuli. Since the probability of obtaining a particular partition after each cycle depends only the previous cycle, this is a Markov chain. After enough cycles for the Markov chain to converge, we begin to save the partitions it produces. One cycle is not independent of the next, so some cycles are discarded to approximate independence. The partitions generated by the Gibbs sampler can be used in the same way as samples $x_n^{(\ell)}$ in Equation 8. The resulting approximation becomes exact as $m \to \infty$ (Gilks et al., 1996).

The Gibbs sampler provides an effective means of approximating the sum in Equation 1, 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. This is both a strength and a weakness. The strength is that the Gibbs sampler is an excellent algorithm to model experiments where people do not receive stimuli one after another, but instead receive the full set of stimuli simultaneously. The weakness is that 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 provides a discrete approximation to a posterior distribution that can be updated with new data (Doucet, de Freitas, & Gordon, 2001). Each "particle" is a partition $x_i^{(\ell)}$ 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

$$P(x_i|F_i) \approx \frac{1}{m}\sum_{\ell=1}^{m} \delta(x_i - x_i^{(\ell)}) \quad (10)$$

where $\delta(\cdot)$ is 1 when its argument is 0, and 0 otherwise.

Using Equation 10 as an approximation to the posterior distribution over partitions for $i$ trials, we can approximate the prior distribution for partitions of the first $i+1$ trials with

$$\begin{aligned} P(x_{i+1}|F_i) &= \sum_{x_i} P(x_{i+1}|x_i)P(x_i|F_i) \\ &\approx \sum_{x_i} P(x_{i+1}|x_i)\frac{1}{m}\sum_{\ell=1}^{m}\delta(x_i - x_i^{(\ell)}) \\ &= \frac{1}{m}\sum_{\ell=1}^{m} P(x_{i+1}|x_i^{(\ell)}) \end{aligned} \quad (11)$$

where $P(x_{i+1}|x_i)$ is given by Equation 7. We can then approximate the posterior for the first $i+1$ trials with

$$\begin{aligned} P(x_{i+1}|F_{i+1}) &\propto \sum_{x_i} P(f_{i+1}|x_{i+1}, F_i)P(x_{i+1}|F_i) \\ &\approx \frac{1}{m}\sum_{\ell=1}^{m} P(f_{i+1}|x_{i+1}, F_i)P(x_{i+1}|x_i^{(\ell)}) \end{aligned} \quad (12)$$

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 RMC is initialized with the first stimulus assigned to the first cluster for all $m$ particles. On each following trial, the distribution in Equation 12 is calculated, based on the particles sampled in the last trial. On any trial, these particles provide an approximation to the posterior distribution on 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 will likely change. As $m \to \infty$, the assignment will not appear to be fixed as the particle filter produces exactly the correct answer. When $m = 1$, the the probability of previous assignments cannot change, and the 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. However, instead of selecting the most likely cluster for the new stimulus, a cluster is sampled based its posterior probability.

## Comparing the algorithms

The existence of alternative algorithms that approximate the posterior distribution over partitions makes it possible to tease the predictions of the RMC that stem from the underlying statistical model apart from those that result from the local MAP algorithm. We do so in two stages. First, we evaluate the accuracy with which the different algorithms approximate the actual predictions produced by Bayesian inference, using a classic data set from Medin and Schaffer (1978). Second, we examine how well the predictions of the algorithms correspond to human judgments. Due to space constraints, we do not reproduce all of the modeling results from Anderson (1990). Instead, we focus on two data sets: the experiment by Medin and Schaffer (1978) mentioned above, and order sensitivity data reported by Anderson (1990).

To apply the algorithms to any dataset, a measure of the probability of a set of features given a partition of the stimuli needs to be introduced. The RMC assumes that the features 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

$$P(f_{i+1}|x_{i+1}, F_i) = \prod_d P(f_{i+1,d}|x_{i+1}, F_i)$$

where $f_{i+1,d}$ is the value of the $d$th feature. Anderson (1991) gives probabilities for both discrete and continuous features, but we only consider binary features here. Given the cluster, the value on each feature is assumed to have a Bernoulli distribution. Integrating out the parameter of this distribution with a Beta$(\beta_0, \beta_1)$ prior gives

$$P(f_{i+1,d} = j|x_{i+1}, F_i) = \frac{b_j + \beta_j}{b_\cdot + \beta_0 + \beta_1}$$

where $b_j$ is the number of stimuli with value $j$ on the $d$th feature in the cluster that partition $x_{i+1}$ assigns $f_{i+1,d}$. The term $b_\cdot$ denotes the number other stimuli in the same cluster. We use $\beta_0 = \beta_1 = 1$ in all simulations.

### Making accurate predictions

The local MAP algorithm, Gibbs sampler, and particle filter all give approximations to Equation 1. We now compare the accuracy of these approximations using the first experiment of Medin and Schaffer (1978). There were six training stimuli in this experiment with 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 solution to Equation 2 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 last feature of a set of test stimuli 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 1100 cycles on a single training order. The first 100 cycles were discarded and only every 10th cycle was kept for a total of 100 samples. The particle filter was run with 100 particles on a single training order.

The results shown in the top row of Figure 1 illustrate that the coupling parameter does not have a large effect on the exact solution of Equation 1. The particle filter and Gibbs sampler do a good job of approximating this solution, while the local MAP algorithm depends much

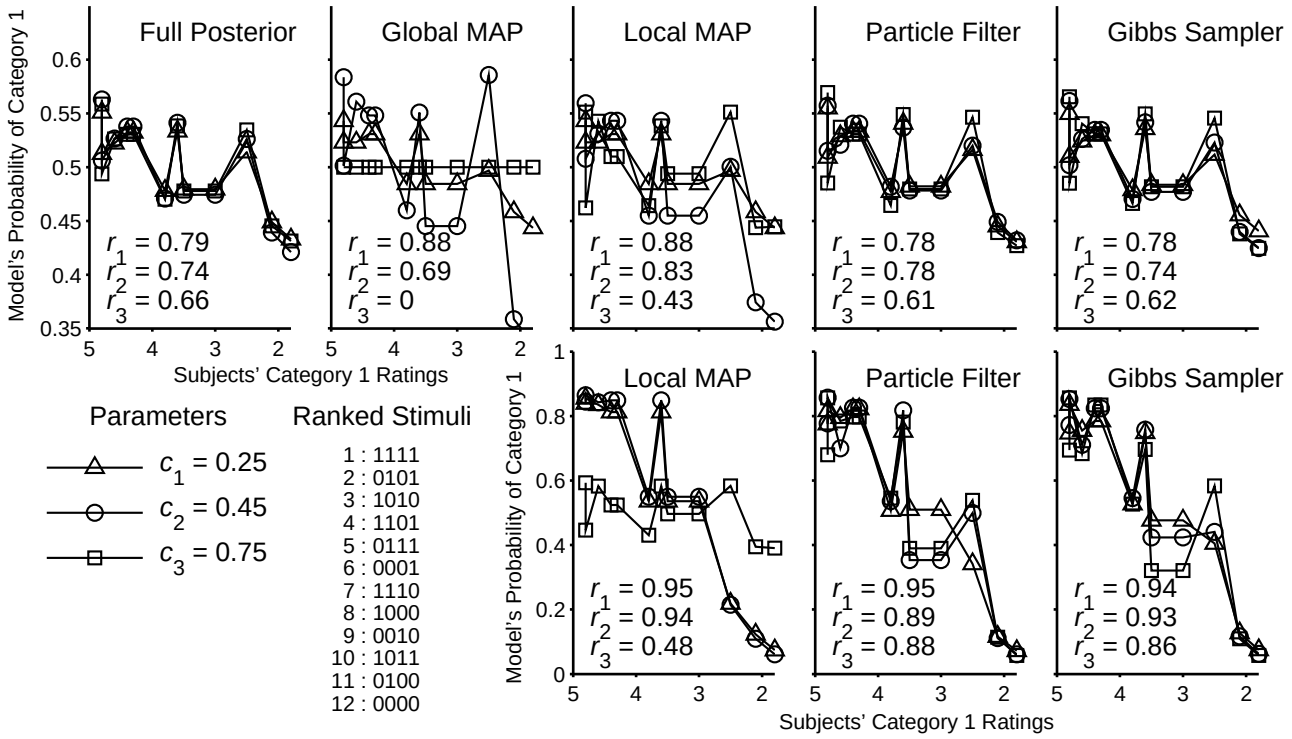


Figure 1: Probability of choosing category 1 for the stimuli from the first experiment of Medin & Schaffer (1978). The ratings of the test stimuli (converted to a single six-point scale) are along the horizontal axis. In the first row only the first six trials are presented, while in the second row ten blocks of six trials each are presented. The three lines in each panel correspond to three different coupling parameters: $c_1 = 0.25$, $c_2 = 0.45$, and $c_3 = 0.75$. Correlations between the human data and the simulation data are displayed on each plot for each value of the coupling parameter (e.g., correlation $r_1$ corresponds to parameter $c_1$).

more on the coupling parameter. The global MAP solution, which the local MAP algorithm attempts to discover, is not a very good approximation of the full posterior. Overall, these results indicate that the predictions of the model can be quite strongly affected by the choice of algorithm.

## Fitting human data

Linear correlations with the human confidence ratings reported by Medin and Schaffer (1978) were computed for all algorithms described in the previous section, and are shown in Figure 1. The fits to the human data for all three approximation algorithms improve when they are trained on ten blocks of the six stimuli, which is not surprising given that this more closely resembles the training given to human participants. This is illustrated in the second row of Figure 1. With ten blocks of training, the alternative algorithms predict human ratings equally as well or better than the local MAP.

The predictions of the local MAP algorithm depend strongly on the presentation order of the stimuli, since cluster assignments are made sequentially and fixed. Order effects are found in human cognition (Medin & Bettger 1994), but are not predicted by the DPMM because of exchangeability. Using data collected by Anderson and Matessa (Anderson, 1990), we explored the strength of the order effects produced by local MAP and the alternative algorithms introduced above.

In Anderson and Matessa's experiment, subjects were presented with a set of 16 stimuli in one of two orders, shown in Table 1. These stimuli were designed to either emphasize the first two features ("front-anchored stimuli") or the last two features ("end-anchored stimuli") in the first eight trials. Subjects were trained in one of the two orders. Following the training phase, subjects were shown the full set of stimuli on a sheet of paper and asked to divide the stimuli into two categories of eight stimuli each. The second column of Table 2 shows the probability of subjects using one of the first two features to split the stimuli into two categories. The stimuli could be split along any of the four features.

We compared order effects produced by the three approximation algorithms to the human data. For all three algorithms, $c = 0.5$, the value used for the local MAP by Anderson and Matessa (Anderson, 1990). The local MAP algorithm produces the same result each time it is run on these stimuli. The Gibbs sampler was run for 20200 cycles. The first 200 cycles were discarded and every 20th cycle kept for a total of 1000 samples. The particle filter was run 1000 times with either 1 or 100 particles. The results were restricted to allow only partitions that split the stimuli into two equal-sized groups

Table 1: Presentation Order of Anderson & Matessa Training Stimuli (from Anderson, 1990)

| Order Type | Stimuli |
|---|---|
| Front-Anchored | 1111, 1101, 0010, 0000, 0011, 0001, 1110, 1100, 0111, 1010, 1000, 0101, 0110, 1011, 1001, 0100 |
| End-Anchored | 0100, 0000, 1111, 1011, 0011, 0111, 1000, 1100, 1010, 0001, 0101, 1110, 1001, 0010, 0110, 1101 |

based on one of the features. The Adjusted Rand Index (Hubert & Arabie, 1985), a standard measure of distance between partitions, was used to find the similarity of the RMC samples to each of the four partitions that split the stimuli along a single feature. The single-feature-based partition that had the highest Adjusted Rand Index was selected as the partition for that sample. If there was a tie, one of the best was selected with equal probability.

The results of the simulations are shown in Table 2. The local MAP results illustrate a perfect bias for splitting the categories along the highlighted features: for the front-anchored stimuli, one of the first two features will always be used, and for the end-anchored stimuli, one of the last two features will always be used. Subjects showed a bias for the highlighted features, but not as strong a bias as predicted by the local MAP algorithm. Consistent with the DPMM, the particle filter with 100 particles and the Gibbs sampler do not show an effect of the ordering of the stimuli. Reducing the number of particles in the particle filter results in an increased order bias. A particle filter using one particle produces a softer bias that is more in line with the human data.

## Conclusion

Models of human categorization have assumed many different types of representations. The probabilistic model underlying the rational model of categorization (Anderson, 1990, 1991) is equivalent to the Dirichlet process mixture model used in nonparametric Bayesian statistics. However, exactly calculating the posterior distribution over assignments of stimuli to clusters in this model becomes impractical for any reasonable number of stimuli, making approximation algorithms necessary. We showed that the local MAP algorithm proposed by Anderson does not approximate the true posterior distribution well in all situations. The Gibbs sampler and particle filter, asymptotically correct algorithms that are more widely used in Bayesian statistics, produced closer approximations. These alternative algorithms thus allow us to directly test Anderson's assumptions about the computational problem underlying categorization.

Part of the motivation for Anderson's (1990, 1991) local MAP algorithm was a desire for a procedure that could plausibly be used by people. The particle filter provides a nice alternative to the local MAP algorithm, having the same psychologically plausible properties, but also providing asymptotic performance guarantees. A large number of particles will produce an accurate approximation of the posterior, while a small number of particles can capture both the variability and the order-sensitivity that people show when considering a sequence of stimuli. Varying the number of particles provides a way to explore the interaction between cognitive constraints and statistical inference, and a natural framework in which to define models that are rational not just in their construal of a computational problem, but in their approximate solution. More research is needed to test the predictions produced by these algorithms, but a particle filter with an intermediate number of particles is a promising candidate for explaining how people perform approximate Bayesian inference in a range of settings.

Table 2: Probability of Clustering Stimuli Along Either of the First Two Features in Anderson & Matessa Data

| Method | Order Type | |
|---|---|---|
| | Front-Anchored | End-Anchored |
| Experimental Data | 0.55 | 0.30 |
| Local MAP | 1.00 | 0.00 |
| Gibbs Sampler | 0.48 | 0.49 |
| Particle Filter (100) | 0.50 | 0.50 |
| Particle Filter (1) | 0.59 | 0.38 |

**Acknowledgments** The authors thank Jonathan Nelson and three anonymous reviewers for helpful comments and Matthew Loper for running preliminary simulations using particle filters in the RMC. Adam Sanborn was supported by an NSF Graduate Research Fellowship.

## References

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

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

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. and Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, 39:216–233.

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

Doucet, A., de Freitas, N., and Gordon, N. (2001). *Sequential Monte Carlo Methods in Practice.* Springer, New York.

Gilks, W., Richardson, S., and Spiegelhalter, D. J., editors (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall, Suffolk.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2:193–218.

Medin, D. L. and Bettger, J. G. (1994). Presentation order and recognition of categorically related examples. *Psychonomic Bulletin & Review*, 1:250–254.

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

Murphy, G. L. and Ross, B. H. (1994). Predictions from uncertain categorizations. *Cognitive Psychology*, 27:148–193.

Neal, R. M. (1998). Markov chain sampling methods for Dirichlet process mixture models. Technical Report 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.

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