Chapter 9

Capturing the growth of knowledge with nonparametric Bayesian models Joseph Austerweil, Adam N. Sanborn, Christopher Lucas & Thomas L. Griffiths

Mental representations are often simplifications. When we represent the objects in the world as members of a set of discrete categories, we are imposing a simple structure on the complexity of our experience. Postulating that a set of objects can be described by a small set of features makes thinking about those objects simpler. Likewise, assuming that the causal relationship between two variables can be represented as a smooth function simplifies the underlying reality. Psychologists usually specify the amount of structure in a mental representation – the number of categories, features, or causes – and tend to decide in advance how much complexity the representation can capture. However, we ideally want to simplify the world just enough, imposing no more structure than is needed to make accurate predictions and form reasonable explanations. In learning categories, we don't want to make the unrealistic assumption that there is a fixed number of kinds of things in the world. In identifying features, we don't want to assume that there are only so many features to go around. In learning causal relationships, we don't want to assume that every relationship is linear. The world is boundlessly complex, and we want to be able to emulate that complexity when the available data warrant it.

This need to accommodate potentially unlimited complexity is in tension with the practical challenges of probabilistic inference. Human brains are finite, so if they can form representations of arbitrary complexity this needs to be done in a way that remains tractable. Likewise, in making probabilistic models of human cognition, we need to be able to perform calculations using those models on a computer. We thus need to be able to define probabilistic models that allow us to work with infinite hypothesis spaces using finite representations and computations.

A natural way for a learner to strike a balance between complexity and tractability is to start with simple representations and add complexity incrementally upon making new observations that require it. This kind of approach is expressed in Piaget's (1954) characterization of cognitive development as a process of **assimilation** and **accommodation**. In Piaget's view, when presented with new information the child has two options: to assimilate that information to their current understanding of the world, or to change that understanding to accommodate the new information. Despite starting with a simple model of the world, these accommodations accumulate complexity that deepens and enriches the child's representations. Some researchers have tried to capture this process in models that can accumulate complexity in a similar way, such as cascade correlation neural networks (Shultz, Mareschal, & Schmidt, 1994).

This intuitive approach to solving the representational and computational challenges of capturing complexity is made precise in a set of tools developed in **nonparametric Bayesian statistics** (e.g., Muller & Quintana, 2004; Hjort, Holmes, Müller, & Walker, 2010). These tools exploit the methods for approximate inference described in previous chapters to work with models that can accommodate unlimited complexity. This approach is "nonparametric" in the sense that it provides a way to work with models that go beyond the simple parametric families that are commonly encountered in statistics. More formally, it covers situations where the complexity of the models increase with the data – for example, the effective number of parameters that need to be estimated grows as more data are observed. This is in contrast to parametric models, which have a fixed set of parameters, and is more akin to nonparametric frequentist methods such as kernel density estimation, which was briefly mentioned in Chapter 5.

To give a concrete example, consider the problem faced by an explorer visiting a new continent. This explorer is familiar with the animals on her own continent, and has already organized them into species. Pushing aside a tree branch, she encounters her first animal: Is this a new kind of animal, deserving a new species of its own? Or can its properties be explained by the existing species of animals she has previously encountered? While her representation at any point in time is finite – she can never postulate more species than individual animals she has seen – there is no upper limit to the number of species the world might contain. This is exactly the assumption behind the models we consider in this chapter.

In principle, nonparametric Bayesian models have an infinite amount of structure – an infinite number of categories, an infinite number of features, infinite degrees of freedom in a function – but they effectively only instantiate a finite amount of structure in response to any finite observed data set. They postulate no more complexity than is necessary, guided by a version of the Bayesian Occam's razor introduced in Chapter 3, and expand only as much as is necessary to explain the data. We will consider how this approach can be applied in three settings: categorization, feature learning, and function learning.

9.1 Infinite models for categorization

In Chapter 5 we saw that psychological models of categorization can be given a probabilistic interpretation. Specifically, these models can be thought of as corresponding to schemes for estimating a probability distribution over objects associated with a category, and how mixture models can be used for this purpose. In this section, we use this formulation of the problem of categorization to introduce one of the most common nonparametric Bayesian models – the **infinite mixture model**. We begin with a more formal treatment of finite mixture models, setting up the mathematical ideas that are then used to generalize this to the infinite mixture model. We then spend a little more time on the key idea that makes it possible to define an infinite model – the **Chinese restaurant process** – and discuss how it is possible to do inference in an infinite model with only finite means.

Assume we have n objects, with the ith object having d observable properties represented by a row vector $\mathbf{x}_i$. In a mixture model, each object is assumed to belong to a single cluster, z_i, and the properties $\mathbf{x}_i$ are generated from a distribution determined by that cluster. Using the matrix $\mathbf{X} = \left[\mathbf{x}_1^T \, \mathbf{x}_2^T \, \cdots \, \mathbf{x}_n^T\right]^T$ to indicate the properties of all n objects, and the vector $\mathbf{z} = [z_1 \; z_2 \; \cdots \; z_n]^T$ to indicate their cluster assignments, the model is specified by a prior over assignment vectors $P(\mathbf{z})$, and a distribution over property matrices conditioned on those assignments, $p(\mathbf{X}|\mathbf{z})$. These two distributions can be dealt with separately: $P(\mathbf{z})$ specifies the number of clusters and their relative probability, while $p(\mathbf{X}|\mathbf{z})$ determines how these clusters relate to the properties of objects. We will focus on the prior over assignment vectors, $P(\mathbf{z})$, showing how such a prior can be defined without placing an upper bound on the number of clusters.

9.1.1 Finite mixture models

Mixture models assume that the assignment of an object to a cluster is independent of the assignments of all other objects. Assume that there are k clusters, θ is a Discrete distribution over those clusters, and θ_j is the probability of cluster j under that distribution. Under this assumption, the probability of the properties of all n objects $\mathbf{X}$ can be written as

$$p(\mathbf{X}|\theta) = \prod_{i=1}^{n} \sum_{j=1}^{k} p(\mathbf{x}_i, z_i = j|\theta_j) = \prod_{i=1}^{n} \sum_{j=1}^{k} p(\mathbf{x}_i|z_i = j)\, \theta_j. \tag{9.1}$$

The distribution from which each $\mathbf{x}_i$ is generated is thus a mixture of the k cluster distributions $p(\mathbf{x}_i|z_i = j)$, with θ_j determining the weight of cluster j.

The mixture weights θ can either be treated as a parameter to be estimated, or a variable with prior distribution $p(\theta)$. In Bayesian approaches to mixture modeling, a standard choice for $p(\theta)$ is a symmetric Dirichlet distribution, introduced in Chapter 3. The probability of any Discrete distribution θ is given by

$$p(\theta) = \frac{\prod_{j=1}^{k} \theta_j^{\alpha_j - 1}}{D(\alpha_1, \alpha_2, \ldots, \alpha_k)} \tag{9.2}$$

in which $D(\alpha_1, \alpha_2, \ldots, \alpha_k)$ is the Dirichlet normalizing constant

$$\begin{aligned} D(\alpha_1, \alpha_2, \ldots, \alpha_k) &= \int_{\Delta_k} \prod_{j=1}^{k} \theta_j^{\alpha_j - 1} \, d\theta & (9.3) \\ &= \frac{\prod_{j=1}^{k} \Gamma(\alpha_j)}{\Gamma(\sum_{j=1}^{k} \alpha_j)} & (9.4) \end{aligned}$$

where Δ_k is the simplex of (all possible) Discrete distributions over k clusters, and $\Gamma(\cdot)$ is the generalized factorial or Gamma function, with $\Gamma(m) = (m-1)!$ for any positive integer m. In a **symmetric Dirichlet distribution**, all α_j are equal. For example, we could take $\alpha_j = \frac{\alpha}{k}$ for all j. In this case, Equation 9.4 becomes

$$D(\tfrac{\alpha}{k}, \tfrac{\alpha}{k}, \ldots, \tfrac{\alpha}{k}) = \frac{\Gamma(\frac{\alpha}{k})^k}{\Gamma(\alpha)} \tag{9.5}$$

and the mean of θ is the distribution that is uniform over all clusters.

The probabilistic model that we have defined is

$$\begin{aligned} \theta \,|\, \alpha &\sim \text{Dirichlet}(\tfrac{\alpha}{k}, \tfrac{\alpha}{k}, \ldots, \tfrac{\alpha}{k}) & (9.6) \\ z_i \,|\, \theta &\sim \text{Discrete}(\theta) & (9.7) \end{aligned}$$

Having defined a prior on θ, we can simplify this model by integrating over all values of θ (ie. the simplex Δ_k) rather than representing them explicitly. The marginal probability of an assignment vector $\mathbf{z}$, integrating over all values of θ, is

$$\begin{aligned} P(\mathbf{z}) &= \int_{\Delta_k} \prod_{i=1}^{n} P(z_i|\theta) p(\theta) \, d\theta & (9.8) \\ &= \int_{\Delta_k} \frac{\prod_{j=1}^{k} \theta_j^{m_j + \alpha_j - 1}}{D(\alpha_1, \alpha_2, \ldots, \alpha_k)} \, d\theta & (9.9) \\ &= \frac{D(m_1 + \frac{\alpha}{k}, m_2 + \frac{\alpha}{k}, \ldots, m_k + \frac{\alpha}{k})}{D(\frac{\alpha}{k}, \frac{\alpha}{k}, \ldots, \frac{\alpha}{k})} & (9.10) \\ &= \frac{\prod_{j=1}^{k} \Gamma(m_j + \frac{\alpha}{k})}{\Gamma(\frac{\alpha}{k})^k} \frac{\Gamma(\alpha)}{\Gamma(n + \alpha)} & (9.11) \end{aligned}$$

where m_j is the number of objects assigned to cluster j. The tractability of this integral is a result of the fact that the Dirichlet is conjugate to the multinomial.

Equation 9.11 defines a probability distribution over the cluster assignments $\mathbf{z}$ as an ensemble. Individual cluster assignments are no longer independent. Rather, they are **exchangeable** (Box & Tiao, 1992), with the probability of an assignment vector remaining the same when the indices of the objects are permuted. Exchangeability is a desirable property in a distribution over cluster assignments, because the indices labelling objects are typically arbitrary. However, the distribution on assignment vectors defined by Equation 9.11 assumes an upper bound on the number of clusters of objects, since it only allows assignments of objects to up to k clusters.

9.1.2 Infinite mixture models

Intuitively, defining an infinite mixture model means that we want to specify the probability of $\mathbf{X}$ in terms of infinitely many clusters, modifying Equation 9.1 to become

$$p(\mathbf{X}|\theta) = \prod_{i=1}^{n} \sum_{j=1}^{\infty} p(\mathbf{x}_i | z_i = j) \, \theta_j, \tag{9.12}$$

where θ is an infinite-dimensional multinomial distribution. In order to repeat the argument above, we would need to define a $p(\theta)$ on infinite-dimensional multinomials, and compute the probability of $\mathbf{z}$ by integrating over θ. Taking this strategy provides an alternative way to derive infinite mixture models, resulting in something known as a **Dirichlet process mixture model** (Antoniak, 1974; Ferguson, 1983). Instead, we will work directly with the distribution over assignment vectors given in Equation 9.11, considering its limit as the number of clusters approaches infinity (Green & Richardson, 2001; Neal, 1998).

Expanding the gamma functions in Equation 9.11 using the recursive law $\Gamma(x) = (x-1)\Gamma(x-1)$ and cancelling terms produces the following expression for the probability of an assignment vector $\mathbf{z}$:

$$P(\mathbf{z}) = \left(\tfrac{\alpha}{k}\right)^{k_+} \left(\prod_{j=1}^{k_+} \prod_{\ell=1}^{m_j-1} \ell + \tfrac{\alpha}{k} \right) \frac{\Gamma(\alpha)}{\Gamma(N+\alpha)}, \tag{9.13}$$

where k_+ is the number of clusters for which $m_j > 0$, and we have re-ordered the indices such that $m_j > 0$ for all $k \leq k_+$ ($m_j = 0$ for all $j > k_+$). There are k^n possible values for $\mathbf{z}$, which diverges as $k \to \infty$. As this happens, the probability of any single set of cluster assignments goes to 0. Since $k_+ \leq n$ and n is finite, it is clear that $P(\mathbf{z}) \to 0$ as $k \to \infty$, since $\frac{1}{k} \to 0$. Consequently, we will define a distribution over **equivalence classes** of assignment vectors – sets of vectors that have the same properties – rather than the vectors themselves.

Specifically, we will define a distribution on **partitions** of objects. In our setting, a partition is a division of the set of n objects into subsets, where each object belongs to a single subset and the ordering of the subsets does not matter. Two assignment vectors that result in the same division of objects correspond to the same partition. For example, if we had three objects, the cluster assignments $\{z_1, z_2, z_3\} = \{1, 1, 2\}$ would correspond to the same partition as $\{2, 2, 1\}$, since all that differs between these two cases is the labels of the clusters. A partition thus defines an equivalence class of assignment vectors, $[\mathbf{z}]$, with two assignment vectors belonging to the same equivalence class if they correspond to the same partition. A distribution over partitions is sufficient to allow us to define an infinite mixture model, since these equivalence classes of cluster assignments are the same as those induced by identifiability: $p(\mathbf{X}|\mathbf{z})$ is the same for all assignment vectors $\mathbf{z}$ that correspond to the same partition, so we can apply statistical inference at the level of partitions rather than the level of assignment vectors.

Assume we have a partition of n objects into k_+ subsets, and we have $k \geq k_+$ cluster labels that can be applied to those subsets. Then there are $\frac{k!}{(k-k_+)!}$ assignment vectors $\mathbf{z}$ that belong to the equivalence class defined by that partition, $[\mathbf{z}]$. We can define a probability distribution over partitions by summing over all cluster assignments that belong to the equivalence class defined by each partition. The probability of each of those cluster assignments is equal under the distribution specified by Equation 9.13, so we obtain

$$\begin{aligned} P([\mathbf{z}]) &= \sum_{\mathbf{z} \in [\mathbf{z}]} P(\mathbf{z}) & (9.14) \\ &= \frac{k!}{(k-k_+)!} \left(\tfrac{\alpha}{k}\right)^{k_+} \left(\prod_{j=1}^{k_+} \prod_{\ell=1}^{m_j-1} \ell + \tfrac{\alpha}{k} \right) \frac{\Gamma(\alpha)}{\Gamma(n+\alpha)}. & (9.15) \end{aligned}$$

Rearranging the first two terms, we can compute the limit of the probability of a partition as $k \to \infty$, which is

$$\begin{aligned} P([\mathbf{z}]) = \lim_{k \to \infty} \alpha^{k_+} \cdot \frac{\prod_{j=1}^{k_+} (k-j+1)}{k^{k_+}} \cdot \left(\prod_{j=1}^{k_+} \prod_{\ell=1}^{m_j-1} \ell + \tfrac{\alpha}{k} \right) \cdot \frac{\Gamma(\alpha)}{\Gamma(n+\alpha)} & & (9.16) \\ = \quad \alpha^{k_+} \cdot \quad 1 \quad \cdot \left(\prod_{j=1}^{k_+} (m_j - 1)! \right) \cdot \frac{\Gamma(\alpha)}{\Gamma(n+\alpha)}. & & (9.17) \end{aligned}$$

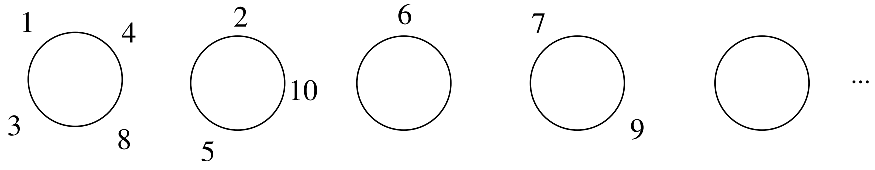


Figure 9.1: A partition induced by the Chinese restaurant process. Numbers indicate customers (objects), circles indicate tables (clusters).

These limiting probabilities define a valid distribution over partitions, and thus over equivalence classes of cluster assignments, providing a prior over cluster assignments for an infinite mixture model. Objects are exchangeable under this distribution, just as in the finite case: the probability of a partition is not affected by the ordering of the objects, since it depends only on the counts m_j.

As noted above, the distribution over partitions specified by Equation 9.17 can be derived in a variety of ways – by taking limits (Green & Richardson, 2001; Neal, 1998), from the Dirichlet process (Blackwell & MacQueen, 1973), or from other equivalent stochastic processes (Ishwaran & James, 2001; Sethuraman, 1994). We will briefly discuss a simple process that produces the same distribution over partitions: the Chinese restaurant process.

9.1.3 The Chinese restaurant process

The Chinese restaurant process (CRP) was named by Jim Pitman and Lester Dubins, based upon a metaphor in which the objects are customers in a restaurant, and the clusters are the tables at which they sit (the process first appears in Aldous, 1985, where it is attributed to Pitman). Imagine a restaurant with an infinite number of tables, each with an infinite number of seats.[1] The customers enter the restaurant one after another, and each choose a table at random. In the CRP with parameter α, each customer chooses an occupied table with probability proportional to the number of occupants, and chooses the next vacant table with probability proportional to α. For example, Figure 9.1 shows the state of a restaurant after 10 customers have chosen tables using this procedure. The first customer chooses the first table with probability $\frac{\alpha}{\alpha} = 1$. The second customer chooses the first table with probability $\frac{1}{1+\alpha}$, and the second table with probability $\frac{\alpha}{1+\alpha}$. After the second customer chooses the second table, the third customer chooses the first table with probability $\frac{1}{2+\alpha}$, the second table with probability $\frac{1}{2+\alpha}$, and the third table with probabililty $\frac{\alpha}{2+\alpha}$. This process continues until all customers have seats, defining a distribution over allocations of people to tables, and, more generally, objects to clusters. Extensions of the CRP and connections to other stochastic processes are pursued in depth by Pitman (2002).

The distribution over partitions induced by the CRP is the same as that given in Equation 9.17. If we assume an ordering on our n objects, then we can assign them to clusters sequentially using the method specified by the CRP, letting objects play the role of customers and clusters play the role of tables. The ith object would be assigned to the jth cluster with probability

$$P(z_i = k | z_1, z_2, \ldots, z_{i-1}) = \begin{cases} \frac{m_j}{i-1+\alpha} & j \leq k_+ \\ \frac{\alpha}{i-1+\alpha} & \text{otherwise} \end{cases} \tag{9.18}$$

where m_j is the number of customers currently sitting at table j, and k_+ is the number of tables that are currently occupied (i.e., for which $m_j > 0$). If all N objects are assigned to clusters via this process, the probability of a partition of objects $\mathbf{z}$ is that given in Equation 9.17. The CRP thus provides an

[1] Pitman and Dubins, both probability theorists at Berkeley, were inspired by the apparently infinite capacity of Chinese restaurants in San Francisco when they named the process.

intuitive means of specifying a prior for infinite mixture models, as well as revealing that there is a simple sequential process by which exchangeable cluster assignments can be generated.

9.1.4 Inference by Gibbs sampling

Inference in an infinite mixture model is only slightly more complicated than inference in a mixture model with a finite, fixed number of clusters. The standard algorithm used for inference in infinite mixture models is Gibbs sampling (Escobar & West, 1995; Neal, 1998). As discussed in Chapter 6, Gibbs sampling is a Markov chain Monte Carlo method in which variables are successively sampled from their distributions when conditioned on the current values of all other variables (Geman & Geman, 1984). This process defines a Markov chain, which ultimately converges to the distribution of interest (see Gilks, Richardson, & Spiegelhalter, 1996).

Implementing a Gibbs sampler requires deriving the conditional distribution for each variable conditioned on all other variables. In a mixture model, these variables are the cluster assignments $\mathbf{z}$. The relevant full conditional distribution is $P(z_i|\mathbf{z}_{-i}, \mathbf{X})$, the probability distribution over z_i conditioned on the cluster assignments of all other objects, $\mathbf{z}_{-i}$, and the data, $\mathbf{X}$. By applying Bayes' rule, this distribution can be expressed as

$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{X}) \propto p(\mathbf{X}|\mathbf{z})P(z_i = j|\mathbf{z}_{-i}) \tag{9.19}$$

where only the second term on the right hand side depends upon the distribution over cluster assignments, $P(\mathbf{z})$.

In a finite mixture model with $P(\mathbf{z})$ defined as in Equation 9.11, we can compute $P(z_i = j|\mathbf{z}_{-i})$ by integrating over θ, obtaining

$$\begin{aligned} P(z_i = j|\mathbf{z}_{-i}) &= \int P(z_i = j|\theta)p(\theta|\mathbf{z}_{-i})\, d\theta \\ &= \frac{m_{-i,j} + \frac{\alpha}{k}}{n - 1 + \alpha}, \end{aligned} \tag{9.20}$$

where $m_{-i,j}$ is the number of objects assigned to cluster j, not including object i, with $\mathbf{z}_{-i}$ defined analogously. This is the posterior predictive distribution for a multinomial distribution with a Dirichlet prior.

In an infinite mixture model with a distribution over cluster assignments defined as in Equation 9.17, we can use exchangeability to find the full conditional distribution. Since it is exchangeable, $P(\mathbf{z})$ is unaffected by the ordering of objects. Thus, we can choose an ordering in which the ith object is the last to be assigned to a cluster. It follows directly from the definition of the Chinese restaurant process that

$$P(z_i = j|\mathbf{z}_{-i}) = \begin{cases} \frac{m_{-i,j}}{n-1+\alpha} & m_{-i,j} > 0 \\ \frac{\alpha}{n-1+\alpha} & j = k_{-i,+} + 1 \\ 0 & \text{otherwise} \end{cases}, \tag{9.21}$$

where $k_{-i,+}$ is the number of clusters for which $m_j > 0$, not including the assignment of object i. The same result can be found by taking the limit of the full conditional distribution in the finite model, given by Equation 9.20 (Neal, 1998).

When combined with some choice of $p(\mathbf{X}|\mathbf{z})$, Equations 9.20 and 9.21 are sufficient to define Gibbs samplers for finite and infinite mixture models respectively. Demonstrations of Gibbs sampling in infinite mixture models are provided by Neal (1998) and Rasmussen (2000). Similar Markov chain Monte Carlo algorithms are presented in Bush and MacEachern (1996), West, Muller, and Escobar (1994), Escobar and West (1995), and Ishwaran and James (2001). Algorithms that go beyond the local changes in cluster assignment allowed by a Gibbs sampler are given by Jain and Neal (2004) and Dahl (2003).

9.1.5 Modeling human category learning

Infinite mixture models provide a way to solve one of the problems that came up when we first considered mixture models as a tool for understanding human category learning: they indicate how a learner could select what kind of representation to use for a given set of objects. If the objects are well-characterized as belonging to a single cluster, then the model might form a representation dominated by a single cluster – a prototype model. If the objects are so dispersed as to have nothing in common, the number of clusters might end up being closer to the number of objects – an exemplar model. Normally, the infinite mixture model will produce a result somewhere between these two extremes. Consequently, it seems that exploring this class of nonparametric Bayesian models might provide some insight into the flexibility of human category learning.

Interestingly, infinite mixture models were proposed as an account of human category learning before they became widespread in statistics and machine learning. Anderson (1990) proposed a model of categorization in which people assigned objects to clusters, with the possibility of increasing the number of clusters if there was a poor match between the object and all existing clusters. The probabilistic model at the heart of this account was an infinite mixture model, as pointed out by (Neal, 1998). Recognizing this relationship, Sanborn, Griffiths and Navarro (2006, 2010) showed that use of the more sophisticated inference algorithms subsequently developed in statistics and machine learning could improve the predictions made by this model.

Anderson's categorization model has been applied to understand how the category representations people learn change as they age. Figure 9.2A shows a set of classic category structures introduced by Shepard, Hovland, and Jenkins (1961). These category structures are defined eight stimuli that differed on three binary dimensions (e.g., shape, color, and size). The stimuli are divided into two categories, with four stimuli per category. This results in six distinct category structures, labeled Type I to Type VI. Shepard et al. (1961) found that these structures varied in how easy they were to learn, with Type I being easiest, Type II harder, Types III, IV, and V harder still, and Type VI hardest. These structures are also represented differently by mixture models: Type IV problems can be accurately represented by a single cluster per category, while Type II problems need two clusters per category to be accurately represented.

While Shepard et al. (1961) found that Type II problems are easier to learn than Type IV problems, this effect was found with the usual experimental population of young adults. When older adults are faced with this task, they show the opposite pattern: Type IV is easier than Type II (see Figure 9.2B; Badham, Sanborn, & Maylor, 2017; Rabi & Minda, 2016). This pattern is reproduced by Anderson's categorization model assuming that older adults have a substantially lower α parameter than young adults, so they produce fewer clusters (see Figure 9.2C). Further support comes from Davis, Love, and Maddox (2012), who found a similar result for Anderson's model for different categorization problems. Interestingly, the difference in α may be related to lower cognitive capacity in older adults: Dasgupta and Griffiths (2022) showed that higher values of α are consistent with a higher cognitive cost for representing a probability distribution. It may be that older adults have fewer representational resources than young adults, explaining why they appear to use fewer clusters in categorization tasks.

Infinite mixture models can be extended in a variety of ways to capture different aspects of categorization. For example, Anderson's categorization model assumes that the category label is just another feature of an object. Clusters are hence shared across categories, with the distribution over clusters associated with each category being obtained by conditioning on the feature corresponding to the category label. An alternative approach is to explicitly associate each category with a distinct distribution over clusters, but allow for the possibility that some of those clusters are shared. This assumption can be captured via the **hierarchical Dirichlet process** (Griffiths, Canini, Sanborn, & Navarro, 2007).

Another generalization of the infinite mixture model allows for the possibility that different clustering schemes might be appropriate for explaining the distribution of distinct subsets of observed features. For

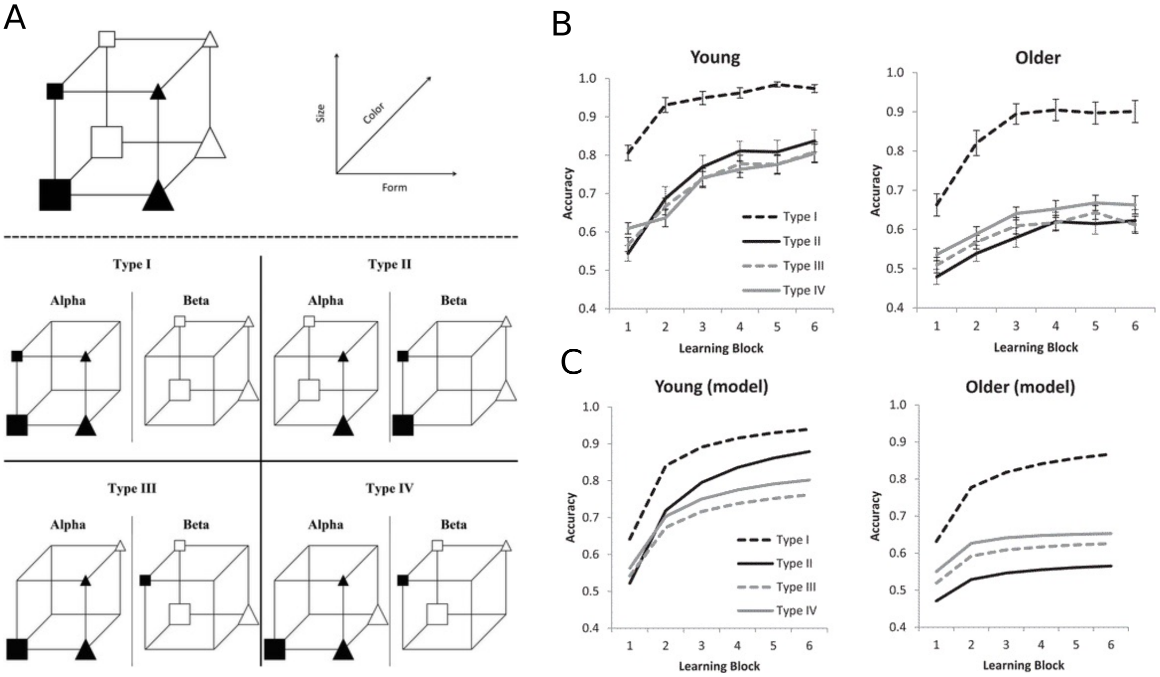


Figure 9.2: Category learning in young and older adults. (a) Illustration of four different category types from Shepard et al. (1961). For each type participants had to learn which stimuli had category label Alpha and which had category label Beta, The stimuli with each label are illustrated on each cube. While the Type IV structure can be captured by one cluster per category, the Type II structure needs two clusters per category. (b) Accuracy over learning for young and older adults for each type. (c) Anderson's categorization model fit to each group of participants. Adapted from Figures 1, 2, and 6 of Badham et al. (2017).

example, a piece of furniture could have features that describe its shape – having legs, a large flat surface – or the materials used to construct it – maple wood. The first set of features supports clusters based on function – tables, chairs, and so on – while the second set of features supports clusters based on material. This kind of distinction can be captured in a generative model that clusters the features themselves, and then clusters objects for each cluster of features (the CrossCat model, discussed in Chapter 1; Shafto, Kemp, Mansinghka, Gordon, & Tenenbaum, 2006; Mansinghka et al., 2016).

A different generalization of the infinite mixture model can help explain the strong effects that separable dimensions, those that are easily identifiable from the stimuli, have on the category representations. Categories that are aligned with separable dimensions tend to be easy to learn, while those that are not are more difficult. This can be explained as a prior over the shape of the clusters in the mixture: clusters are expected to be aligned with the separable dimensions, and category structures that match this prior are easier to learn (e.g., Shepard, 1987; Austerweil, Sanborn, & Griffiths, 2019). But how can this prior over cluster alignments itself be learned? This can be done using another infinite mixture: if the prior on the shape of each cluster is itself an infinite mixture, then different "types" of clusters that correspond to the separable dimensions can be learned across a lifetime of experience. A two-level infinite mixture model can thus explain a wide range of dimensional biases (Sanborn, Heller, Austerweil, & Chater, 2021).

Finally, infinite mixture models can also be extended to settings where objects are described not just by the features they possess, but by the relations they participate in with other objects. For example, when trying to make sense of a new social environment, you may pay attention to which pairs of people seem to be friends. Based on these relations, you could try to infer an underlying cluster structure, where the probability that any two people are friends depends only on the clusters they belong to. More formally, people a and b belong to clusters z_a and z_b, with the probability that they are friends being given by $\eta_{z_a z_b}$. In statistics, this kind of model is known as a **stochastic blockmodel**, and it is the relational equivalent of a mixture model. Defining a prior distribution over cluster memberships using the Chinese restaurant process results in the **infinite relational model**, which has been used to explain aspects of how humans learn relational theories (Kemp, Tenenbaum, Griffiths, Yamada, & Ueda, 2006).

9.1.6 Beyond categorization

Chinese restaurant processes and related distributions have applications in cognitive science that go well beyond categorization. They can be used as prior distributions in any setting where inferences are being made about a latent variable that has a discrete but potentially infinite set of values.

One setting where the Chinese restaurant process has been used successfully is in making inferences about the latent causes that might explain observed events. For example, imagine you went to a cafe, ordered a drink, and enjoyed it. A week later you returned to the same cafe and order the same drink, but this time it is terrible. One way you could make sense of this experience it is by inferring that something changed at the cafe – perhaps the coffee beans were different on the two occasions that you visited. In doing so, you're postulating a latent cause for the phenomenon. As you have more experiences, you might infer more latent causes as you hypothesize the cafe uses several different kinds of beans with different flavors.

This "latent cause" perspective has been used to explain patterns of results and animal conditioning, where suddenly removing a reward is less effective at reducing a behavior than gradually reducing the rate at which the reward is provided (Gershman, Blei, & Niv, 2010). Intuitively, the sudden change suggests a different latent cause should be inferred, and the animal learns that while that cause is present the action no longer produces the reward. The original relationship between the action and reward is thus preserved, and can manifest again if the animal thinks the environment has reverted to the original latent cause. Gradually reducing the rate of reward doesn't result in a change in the inferred latent cause, and the relationship is eliminated.

Another setting where unknown numbers of discrete latent variables appear is in language. Phonemes, words, and syntactic categories are all discrete sets that need to be inferred from the environment. In these cases, the Chinese restaurant process can be useful in defining prior distributions. For example, Goldwater, Griffiths, and Johnson (2006a) defined a probabilistic model of word segmentation – explaining how a child may go from hearing a continuous stream of phonemes to recognizing discrete words within that stream – where the Chinese restaurant process was used to define the prior distribution on words.

Variants of the Chinese restaurant process can be used to more precisely capture the probability distributions that appear in language. In the original Chinese restaurant process, as the number of customers in the restaurant increases the number who sit at each table follows a power-law distribution, with $P(m_j) \propto m_j^{-1}$. This is a "heavy-tailed" distribution, where a small number of tables and up with very large numbers of customers. Power-law distributions arise often in language – for example, the frequencies with which different words are used follow a power-law distribution (Zipf, 1932). However, in the Chinese restaurant process the exponent of the power law (the negative power to which m_j is raised) is 1, while linguistic power laws often have exponents closer to 2.

By introducing additional parameters into the Chinese restaurant process, it is possible to define models that produce power laws with a range of exponents. In particular, in the **two-parameter**

Pitman-Yor process the ith customer would be assigned to the jth table with probability

$$P(z_i = j | z_1, z_2, \ldots, z_{i-1}) = \begin{cases} \frac{m_j - a}{i-1+b+ak_+} & j \leq k_+ \\ \frac{b+ak_+}{i-1+b+ak_+} & \text{otherwise} \end{cases} \tag{9.22}$$

where a and b are parameters of the process. The resulting distribution in the number of customers per table is a power-law with an exponent of $1 + a$. This model can thus be used to better capture the distributions that arise in language, and models based on the Pitman-Yor process have been shown to have deep connections with sophisticated smoothing schemes used in estimating probability distributions over words (Goldwater, Griffiths, & Johnson, 2006b; Teh, 2006). In fact, the distribution induced by the Pitman-Yor process is the most general distribution over exchangeable partitions (Pitman, 2002).

Applications of ideas from nonparametric Bayesian statistics to language don't stop at the level of words. A standard problem that arises in natural language processing is estimating the probability distributions associated with the rules of probabilistic grammars (see Chapter 16). In these grammars, a rule identifies a discrete set of possible ways in which a symbol can be rewritten, each associated with a probability. Using distributions based on the Chinese restaurant process to represent these probabilities has the consequence of "caching" the outcomes of previous applications of a rule: at each point where the rule is applied, you can choose to use a previously generated outcome or create a new one (Johnson, Häubl, & Keinan, 2007). This property makes it possible to capture some of the rich dependencies in language that are otherwise missing from simple grammars. A similar approach has been used as a form of **stochastic memoization** in probabilistic programming languages (Goodman, Mansinghka, Roy, Bonawitz, & Tenenbaum, 2008), which we discuss in more detail in Chapter 18.

9.2 Infinite models for feature representations

The previous section showed how methods from nonparametric Bayesian statistics could be used to define models of category learning that do not require assuming that there is a fixed set of kinds of things in the world. The same strategy can be applied to feature learning. In feature learning the goal is to identify the latent features that explain the observed properties of a set of objects. In the simplest case – which we will focus on here – the assignment of features to objects is binary, with a latent variable indicating whether or not an object possesses each feature. The challenge is in simultaneously deciding what features the objects have and how many features should be used to represent the set of objects, just as in the category learning case the challenge is inferring both the category assignments and their number.

Feature learning can be thought of as a similar problem to category learning. If we imagine each object being associated with a binary vector of features, category learning corresponds to the case where there is a constraint that each row can contain only one non-zero entry. Feature learning is the general case, in which there can be multiple non-zero entries per object. In other words, one discrete unit is associated with each data point in category learning, whereas zero or more discrete units are associated with each data point in feature learning. In this section we introduce a distribution that is similar to the Chinese restaurant process, but covers this more general case. This distribution can then be used as a prior in models of feature learning, or other cases where we seek to infer a binary vector but do not wish to limit its length. As in the previous section, we derive this infinite binary prior as the limit of a finite distribution.

9.2.1 A finite feature model

We have n objects and k features, and the possession of feature j by object i is indicated by a binary variable z_{ij}. Each object can possess multiple features. The z_{ij} thus form a binary $n \times k$ feature matrix,

Z. We will assume that the entries in this matrix are generated by the model

$$\begin{aligned} z_{ij} &\sim \text{Bernoulli}(\theta_j) \\ \theta_j &\sim \text{Beta}(\frac{\alpha}{k}, 1), \end{aligned}$$

where θ_j is the probability that any object has feature j. Each z_{ij} is independent of all other assignments, conditioned on θ_j. The θ_j are independent, so each z_{ij} is dependent only upon whether other objects possess feature k.

We can compute the joint probability of all assignments of features to objects using this model, defining a probability distribution over matrices **Z**,

$$\begin{aligned} P(\mathbf{Z}) &= \prod_k \int \left(\prod_i P(z_{ij}|\theta_j) \right) p(\theta_j) \; d\theta_j \\ &= \prod_k \frac{B(m_j + \frac{\alpha}{k}, n - m_j + 1)}{B(\frac{\alpha}{k}, 1)} \\ &= \prod_j \frac{\Gamma(m_j + \frac{\alpha}{k})\Gamma(n - m_j + 1)}{\Gamma(\frac{\alpha}{k})} \frac{\Gamma(1 + \frac{\alpha}{k})}{\Gamma(n + 1 + \frac{\alpha}{k})} \end{aligned} \tag{9.23}$$

where $B(r, s)$ is the standard beta function, and m_j is the number of objects in cluster j. Again, the result follows from conjugacy, this time between the binomial and beta distributions. This distribution is exchangeable, depending only on the counts m_j.

It is straightforward to compute the full conditional distribution for any z_{ij},

$$\begin{aligned} P(z_{ij} = 1|\mathbf{Z}_{-i,j}) &= \int_0^1 P(z_{ij}|\theta_j) p(\theta_j|\mathbf{z}_{-i,j}) \; d\theta_j \\ &= \frac{m_{-i,j} + \frac{\alpha}{k}}{n + \frac{\alpha}{k}}, \end{aligned} \tag{9.24}$$

where $\mathbf{Z}_{-i,j}$ is the set of assignments of other objects, not including i, for feature j, and $m_{-i,j}$ is the number of objects possessing feature j, not including i.

9.2.2 Taking the infinite limit

We can now examine the consequences of taking $k \to \infty$. The use of $\frac{\alpha}{k}$ in defining the above model guarantees that the resulting infinite matrices remain sparse. As with the CRP, we need to define a distribution on equivalence classes of matrices, since the probability of any particular matrix will go to zero as $k \to \infty$. In this case, we calculate the probability of the equivalence class of matrices that are the same up to the order of their columns (for details, see Griffiths & Ghahramani, 2005). Taking the limit of Equation 9.23 gives

$$\lim_{k \to \infty} P([\mathbf{Z}]) = \exp\{-\alpha H_n\} \frac{\alpha^{k_+}}{\prod_{h>0} k_h!} \prod_{j \leq k_+} \frac{(n - m_j)!(m_j - 1)!}{n!}. \tag{9.25}$$

Again, this distribution is exchangeable: neither the number of identical columns nor the column sums are affected by the ordering on objects.

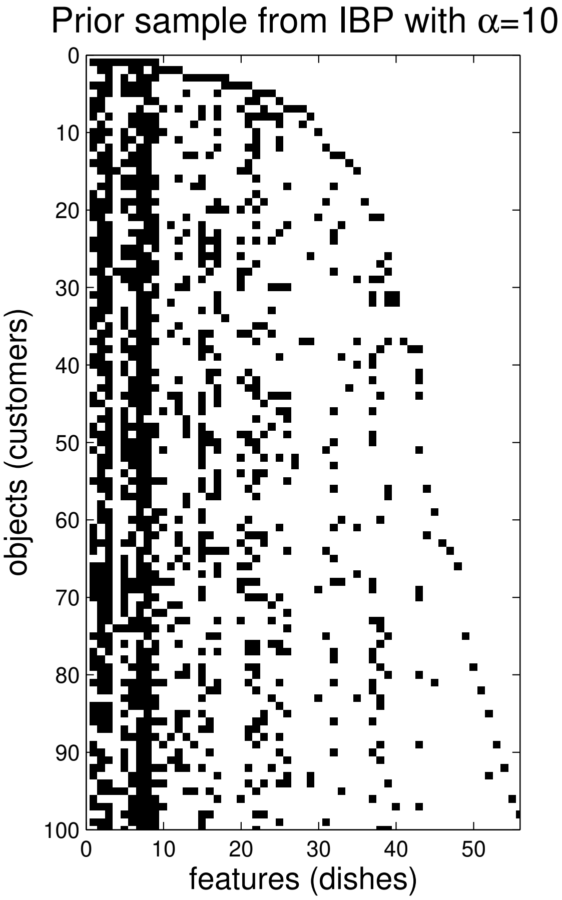


Figure 9.3: A binary matrix sampled from the Indian buffet process with $\alpha = 10$.

9.2.3 The Indian buffet process

The joint probability given in Equation 9.25 is not immediately intuitive, but can be produced by a simple generative process known as the **Indian buffet process (IBP)** (Griffiths & Ghahramani, 2005). As with the CRP, this process assumes an ordering on the objects, generating the matrix sequentially using this ordering, and objects correspond to customers in a restaurant. Many Indian restaurants in London offer lunchtime buffets with an apparently infinite number of dishes. The first customer starts at the left of the buffet, and takes a serving from each dish, stopping after a number of dishes drawn from a Poisson(α) distribution as her plate becomes overburdened. The ith customer moves along the buffet, sampling dishes in proportion to their popularity, serving herself with probability $\frac{m_j}{i}$, and trying a Poisson($\frac{\alpha}{i}$) number of new dishes. The customer-dish matrix $\mathbf{Z}$ is our feature matrix, with customers along the rows, dishes along the columns, and entries indicating which dishes were sampled by which customers. The probability of producing a member of each equivalence class is just the probability given in Equation 9.25. An example of a matrix sampled from this process is shown in Figure 9.3.

Inference in a model that uses the IBP as a prior may require full conditional distributions. If we care about the identities of the columns, as when they are associated with different parameters in a hierarchical model, the best way to derive these distributions is in terms of the second generative process outlined above, drawing each z_{ij} as a Bernoulli trial for j such that $m_j > 0$, and then re-sorting the columns. Using the fact that the distribution is exchangeable, we treat the ith object as the nth in the generative process outlined above, to give

$$P(z_{ij} = 1|\mathbf{z}_{-i,j}) = \frac{m_{j,-i}}{n}. \tag{9.26}$$

The same result can be obtained by taking the limit of Equation 9.24 as $k \to \infty$. By the same set of assumptions, the number of new clusters should be drawn from a Poisson($\frac{\alpha}{n}$) distribution. This can also be derived from Equation 9.24, using the same kind of limiting argument as that presented above to obtain the terms of the Poisson.

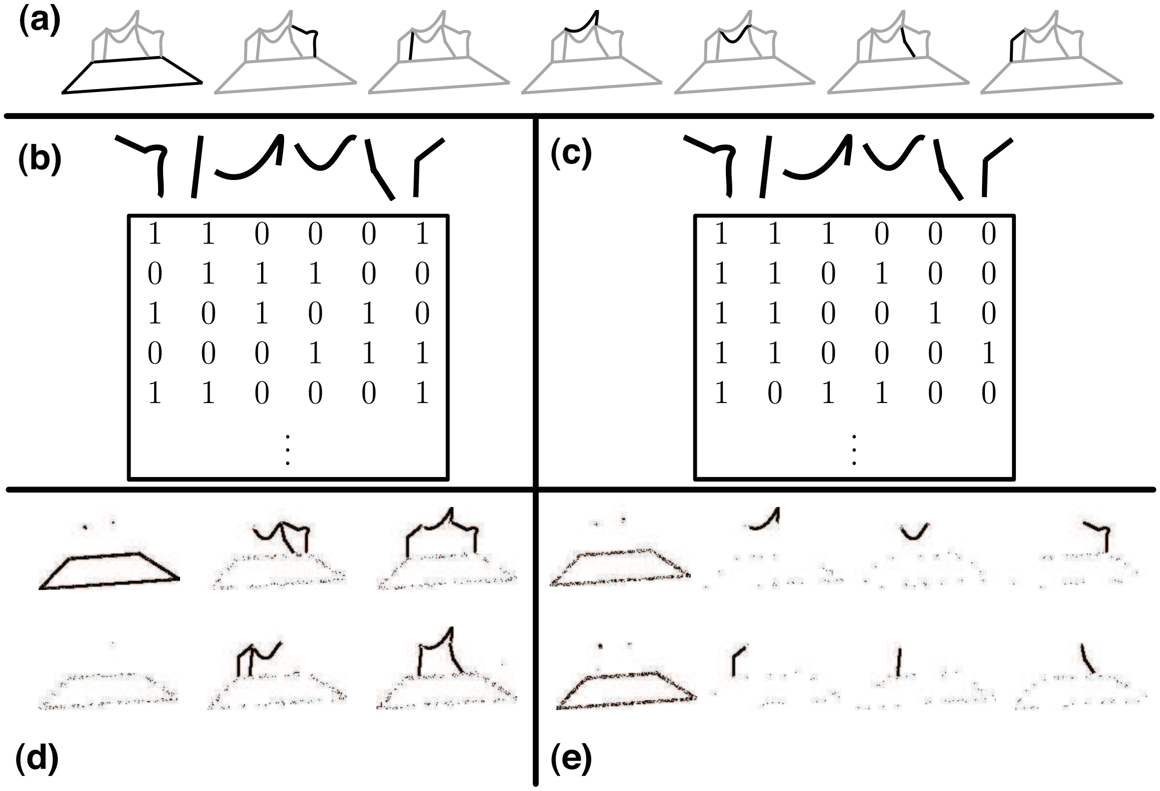
9.2.4 Modeling human feature learning

The IBP provides a simple way to define probabilistic models that can identify the features that should be used to represent a set of objects. This potentially provides an account of how people form feature representations, and how those representations depend on context (in particular, the other objects that a person is familiar with). Austerweil and Griffiths (2011) explored the predictions of this kind of account, showing that people seemed to form different representations of objects depending on the distributional properties of the set of objects in which they appear. Figure 9.4 shows how two sets of objects were generated from the same set of parts. Each object had three of the six parts, making for twenty possible combinations of parts. One set of objects, the *correlated* set, repeated the same four combinations of three parts four times each. The other set of objects, the *independent* set contained sixteen of the twenty unique combinations. When shown the *correlated* set, a probabilistic model based on the IBP forms a representation in which each repeated combination is a single feature. When shown the *independent* set, the six parts from which the objects were actually constructed are identified as features. People seem to form different representations of the objects when shown these two sets as well: when asked whether the four unobserved combinations of objects are likely to appear with the others, participants shown the *correlated* set were far less willing to generalize to these new objects than participants shown the *independent* set.

Analogous to the CRP and the associated Dirichlet Process, there have been many extensions and generalizations of the IBP and its associated continuous stochastic process, the **Beta process** (Hjort, 1990). One technique for doing so is to elaborate the culinary metaphor of the IBP. For example, transformation-invariant feature learning models have been produced by having customers take a “spice” that is applied to each dish, which transforms the taste of the dish (Austerweil & Griffiths, 2013). Recent work has also explored ways to combine the IBP with neural networks, defining a prior that can be used to help neural networks learn distinct representations of related tasks over time (Kessler, Nguyen, Zohren, & Roberts, 2021).

9.3 Infinite models for function learning

So far, we have focused on cases where the latent structure to be inferred is discrete – either a category or a set of features. However, a similar problem of wanting to accommodate infinite complexity while maintaining simplicity arises in other settings. One of the most prominent examples is **function learning** – learning a relationship between two (or more) continuous variables. This is a problem that people often solve without even thinking about it, as when learning how hard to press the pedal to yield a certain amount of acceleration when driving a rental car. Nonparametric Bayesian methods also provide a way to solve this problem that makes it possible to learn complex functions in a way that remains tractable and favors simple solutions.



$$\begin{array}{cccccc} 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ & & \vdots & & & \end{array}$$

$$\begin{array}{cccccc} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ & & \vdots & & & \end{array}$$

Figure 9.4: Inferring different feature representations depending on the distributional information. (a) The bias (on left) and the six features used to generate both object sets. (b) - (c) The feature membership matrices for (b) *correlated* and (c) *independent* sets respectively. (d) - (e) The feature representations inferred by model for (d) *correlated* and (e) *independent* sets respectively, here represented by single samples drawn from the posterior distribution. Reproduced with permission from Austerweil and Griffiths (2011).

Viewed abstractly, the computational problem behind function learning is to learn a function f mapping from x to y from a set of real-valued observations $\mathbf{x}_n = (x_1, \ldots, x_n)$ and $\mathbf{t}_n = (t_1, \ldots, t_n)$, where t_i is assumed to be the true value $y_i = f(x_i)$ obscured by some kind of additive noise. In machine learning and statistics, this is referred to as a **regression** problem. In this section, we discuss how this problem can be solved using Bayesian statistics, and how the result of this approach is related to a class of nonparametric Bayesian models known as Gaussian processes. Our presentation follows that in Lucas, Griffiths, Williams, and Kalish (2015).

9.3.1 Bayesian linear regression

Ideally, we would seek to solve our regression problem by combining some prior beliefs about the probability of encountering different kinds of functions in the world with the information provided by $\mathbf{x}_n$ and $\mathbf{t}_n$. We can do this by applying Bayes' rule, with

$$p(f|\mathbf{x}_n, \mathbf{t}_n) = \frac{p(\mathbf{t}_n|f, \mathbf{x}_n)p(\mathbf{f})}{\int_{\mathcal{F}} p(\mathbf{t}_n|f, \mathbf{x}_n)p(f)\, df}, \tag{9.27}$$

where $p(f)$ is the prior distribution over functions in the hypothesis space $\mathcal{F}$, $p(\mathbf{t}_n|f, \mathbf{x}_n)$ is the probability of observing the values of $\mathbf{t}_n$ if f were the true function (the likelihood), and $p(f|\mathbf{x}_n, \mathbf{t}_n)$ is the posterior distribution over functions given the observations $\mathbf{x}_n$ and $\mathbf{t}_n$. In many cases, the likelihood is defined by assuming that the values of t_i are independent given f and x_i, and each follow a Gaussian distribution with mean $y_i = f(x_i)$ and variance σ_t^2. Predictions about the value of the function f for a new input x_{n+1} can be made by integrating over this posterior distribution,

$$p(y_{n+1}|x_{n+1}, \mathbf{t}_n, \mathbf{x}_n) = \int_f p(y_{n+1}|f, x_{n+1}) p(f|\mathbf{x}_n, \mathbf{t}_n)\, df, \tag{9.28}$$

where $p(y_{n+1}|f, x_{n+1})$ is a delta function placing all of its mass on $y_{n+1} = f(x_{n+1})$.

Performing the calculations outlined in the previous paragraph for a general hypothesis space $\mathcal{F}$ is challenging, but becomes straightforward if we limit the hypothesis space to certain specific clusters of functions. If we take $\mathcal{F}$ to be all linear functions of the form $y = b_0 + xb_1$, then our problem takes the familiar form of linear regression. To perform Bayesian linear regression, we need to define a prior $p(f)$ over all linear functions. Since these functions can be expressed in terms of the parameters b_0 and b_1, it is sufficient to define a prior over the vector $\mathbf{b} = (b_0, b_1)$, which we can do by assuming that $\mathbf{b}$ follows a multivariate Gaussian distribution with mean zero and covariance matrix $\mathbf{\Sigma}_b$. Applying Equation 9.27 then results in a multivariate Gaussian posterior distribution on $\mathbf{b}$ (see Bernardo & Smith, 1994 for details) with

$$E[\mathbf{b}|\mathbf{x}_n, \mathbf{t}_n] = \left(\sigma_t^2 \mathbf{\Sigma}_b^{-1} + \mathbf{X}_n^T \mathbf{X}_n\right)^{-1} \mathbf{X}_n^T \mathbf{t}_n \tag{9.29}$$

$$\text{cov}[\mathbf{b}|\mathbf{x}_n, \mathbf{y}_n] = \left(\mathbf{\Sigma}_b^{-1} + \frac{1}{\sigma_t^2} \mathbf{X}_n^T \mathbf{X}_n\right)^{-1} \tag{9.30}$$

where $\mathbf{X}_n = [\mathbf{1}_n \ \mathbf{x}_n]$ (ie. a matrix with a vector of ones horizontally concatenated with $\mathbf{x}_{n+1}$) Since y_{n+1} is simply a linear function of $\mathbf{b}$, applying Equation 9.28 yields a Gaussian predictive distribution, with y_{n+1} having mean $[1 \ x_{n+1}] E[\mathbf{b}|\mathbf{x}_n, \mathbf{t}_n]$ and variance $[1 \ x_{n+1}] \text{cov}[\mathbf{b}|\mathbf{x}_n, \mathbf{t}_n] [1 \ x_{n+1}]^T$. The predictive distribution for t_{n+1} is similar, but with the addition of σ_t^2 to the variance.

While considering only linear functions might seem overly restrictive, linear regression actually gives us the basic tools we need to solve this problem for more general clusters of functions. Many clusters of functions can be described as linear combinations of a small set of basis functions. For example, all kth degree polynomials are linear combinations of functions of the form 1 (the constant function), x, x^2, ..., x^k. Letting $\phi^{(1)}, \ldots, \phi^{(k)}$ denote a set of functions, we can define a prior on the class of functions that are linear combinations of this basis by expressing such functions in the form $f(x) = b_0 + \phi^{(1)}(x)b_1 + \ldots + \phi^{(k)}(x)b_k$ and defining a prior on the vector of weights $\mathbf{b}$. If we take the prior to be Gaussian, we reach the same solution as outlined in the previous paragraph, substituting $\mathbf{\Phi} = [\mathbf{1}_n \ \boldsymbol{\phi}^{(1)}(\mathbf{x}_n) \ \ldots \ \boldsymbol{\phi}^{(k)}(\mathbf{x}_n)]$ for $\mathbf{X}$ and $[1 \ \phi^{(1)}(x_{n+1}) \ \ldots \ \phi^{(k)}(x_{n+1})$ for $[1 \ x_{n+1}]$, where $\phi(\mathbf{x}_n) = [\phi(x_1) \ \ldots \ \phi(x_n)]^T$.

9.3.2 Gaussian processes

If our goal were merely to predict y_{n+1} from x_{n+1}, $\mathbf{y}_n$, and $\mathbf{x}_n$, we might consider a different approach, simply defining a joint distribution on $\mathbf{y}_{n+1}$ given $\mathbf{x}_{n+1}$ and conditioning on $\mathbf{y}_n$. One surprisingly general and powerful way to do this is to take the $\mathbf{y}_{n+1}$ to be jointly Gaussian, with covariance matrix

$$\mathbf{K}_{n+1} = \begin{pmatrix} \mathbf{K}_n & \mathbf{k}_{n,n+1} \\ \mathbf{k}_{n,n+1}^T & k_{n+1} \end{pmatrix} \tag{9.31}$$

where $\mathbf{K}_n$ depends on the values of $\mathbf{x}_n$, $\mathbf{k}_{n,n+1}$ depends on $\mathbf{x}_n$ and x_{n+1}, and k_{n+1} depends only on x_{n+1}. If we condition on $\mathbf{y}_n$, the distribution of y_{n+1} is Gaussian with mean $\mathbf{k}_{n,n+1}^T \mathbf{K}_n^{-1} \mathbf{y}$ and variance

$k_{n,n+1} - \mathbf{k}_{n,n+1}^T \mathbf{K}_n^{-1} \mathbf{k}_{n,n+1}$. This approach to prediction is often referred to as using a **Gaussian process**, since it assumes a stochastic process that induces a Gaussian distribution on $\mathbf{y}$ based on the values of $\mathbf{x}$. This approach can also be extended to allow us to predict t_{n+1} from x_{n+1}, $\mathbf{t}_n$, and $\mathbf{x}_n$ by adding $\sigma_t^2 \mathbf{I}_n$ to $\mathbf{K}_n$, where $\mathbf{I}_n$ is the $n \times n$ identity matrix, to take into account the additional variance associated with the observations $\mathbf{t}_n$.

The covariance matrix $\mathbf{K}_{n+1}$ is specified using a two-place function in x known as a **kernel**, with $K_{ij} = K(x_i, x_j)$. Any kernel that results in an appropriate (symmetric, positive-definite) covariance matrix for all $\mathbf{x}$ can be used. Common kernels include a radial basis function, with

$$K(x_i, x_j) = \theta_1^2 \exp(-\frac{1}{\theta_2^2}(x_i - x_j)^2) \tag{9.32}$$

indicating that values of y for which values of x are close are likely to be highly correlated, where θ_1 and θ_2 are free parameters of the kernel controlling the overall level of covariation and the speed with which it falls off as a function of the distance between points. Gaussian processes thus provide an extremely flexible approach to regression, with the kernel being used to define which values of x are likely to have similar values of y.

9.3.3 Taking the infinite limit

Bayesian linear regression and Gaussian processes appear to present two quite different approaches to the problem of regression. In Bayesian linear regression, an explicit hypothesis space of functions is identified, a prior on that space is defined, and predictions are formed by computing the posterior distribution over functions and then averaging over that distribution. In contrast, Gaussian processes simply use the similarity between different values of x, as expressed through a kernel, to predict correlations in the corresponding values of y. It might thus come as a surprise to know that these two approaches are equivalent: continuing the theme of this chapter, we can derive standard Gaussian process models as the infinite limit of Bayesian linear regression.

Showing that a Bayesian linear regression model of the kind outlined above is a form of Gaussian process prediction is straightforward. The assumption of linearity means that the vector $\mathbf{y}_{n+1}$ is equal to $\mathbf{X}_{n+1}\mathbf{b}$. It follows that $p(\mathbf{y}_{n+1}|bfx_{n+1})$ is a multivariate Gaussian distribution with mean zero and covariance matrix $\mathbf{X}_{n+1}\mathbf{\Sigma}_b\mathbf{X}_{n+1}^T$. Bayesian linear regression thus corresponds to prediction using Gaussian processes, with this covariance matrix playing the role of $\mathbf{K}_{n+1}$ above. Using a richer set of basis functions corresponds to taking $\mathbf{K}_{n+1} = \mathbf{\Phi}_{n+1}\mathbf{\Sigma}_b\mathbf{\Phi}_{n+1}^T$. Thus, Bayesian linear regression corresponds to using the kernel function $K(x_i, x_j) = [1\ x_i][1\ x_j]^T$, and richer basis functions simply make this $K(x_i, x_j) = [1\ \phi^{(1)}(x_i)\ \ldots\ \phi^{(k)}(x_i)][1\ \phi^{(1)}(x_i)\ \ldots\ \phi^{(k)}(x_i)]^T$.

It is also possible to show that Gaussian process prediction can always be interpreted as Bayesian linear regression, albeit with potentially infinitely many basis functions. Just as we can express a covariance matrix in terms of its eigenvectors and eigenvalues, we can express a given (positive definite) kernel $K(x_i, x_j)$ in terms of its eigenfunctions $\boldsymbol{\phi}$ and eigenvalues $\boldsymbol{\lambda}$, with

$$K(x_i, x_j) = \sum_{k=1}^{\infty} \lambda_k \phi^{(k)}(x_i)\phi^{(k)}(x_j) \tag{9.33}$$

for any x_i and x_j. Using the results from the previous paragraph, any kernel can be viewed as the result of performing Bayesian linear regression with a set of basis functions corresponding to its eigenfunctions, and a prior with covariance matrix $\mathbf{\Sigma}_b = \text{diag}(\boldsymbol{\lambda})$.

These equivalence results establish an important duality between Bayesian linear regression and Gaussian processes: for every prior on functions, there exists a kernel that defines the similarity between values of x, and for every (positive-definite) kernel, there exists a corresponding prior on functions that yields the same predictions. Bayesian linear regression and prediction with Gaussian processes are thus just two views of the same class of solutions to regression problems.

9.3.4 Modeling human function learning

The duality between Bayesian linear regression and Gaussian processes provides a novel perspective on human function learning. Previously, theories of function learning had focused on the roles of different psychological mechanisms. One class of theories (e.g., Carroll, 1963; Brehmer, 1974; Koh & Meyer, 1991) suggests that people are learning an explicit function from a given cluster, such as the polynomials of degree k. This approach attributes rich representations to human learners, but has traditionally given limited treatment to the question of how such representations could be acquired. A second approach (e.g., DeLosh, Busemeyer, & McDaniel, 1997; Busemeyer, Byun, DeLosh, & McDaniel, 1997) emphasizes the possibility that people could simply be forming associations between similar values of variables. This approach has a clear account of the underlying learning mechanisms, but faces challenges in explaining how people generalize beyond their experience. More recently, hybrids of these two approaches have been proposed (e.g., McDaniel & Busemeyer, 2005; Kalish, Lewandowsky, & Kruschke, 2004), with explicit functions being represented, but associative learning.

Bayesian linear regression resembles explicit rule learning, estimating the parameters of a function, while the idea of making predictions based on the similarity between predictors (as defined by a kernel) that underlies Gaussian processes is more in line with associative accounts. The fact that, at the computational level, these two ways of viewing regression are equivalent suggests that these competing mechanistic accounts may not be as far apart as they once seemed. Just as viewing category learning as density estimation helps to understand the common statistical basis of prototype and exemplar models, viewing function learning as regression reveals the shared assumptions behind rule learning and associative learning.

Gaussian process models also provide a good account of human performance in function learning tasks. Griffiths et al. (2008) compared a Gaussian process model with a mixture of kernels (linear, quadratic, and radial basis) to human performance (see also Lucas et al., 2015). Figure 9.5 shows mean human predictions when trained on a linear, exponential, and quadratic function (from DeLosh et al., 1997), together with the predictions of the Gaussian process model. The regions to the left and right of the vertical lines represent extrapolation regions, being input values for which neither people nor the model were trained. Both people and the model extrapolate near optimally on the linear function, and reasonably accurate extrapolation also occurs for the exponential and quadratic function. However, there is a bias towards a linear slope in the extrapolation of the exponential and quadratic functions, with extreme values of the quadratic and exponential function being overestimated.

Subsequent work using Gaussian processes to model human function learning has dug more deeply into the kinds of kernel functions needed to capture human expectations about functions. Wilson, Dann, Lucas, and Xing (2015) directly estimated kernels from human function learning data, and found that humans tend to prefer smoother functions than those assumed in typical machine learning approaches. Schulz, Tenenbaum, Duvenaud, Speekenbrink, and Gershman (2017) explored how different kernels could be combined to capture the compositional structure of functions, using a simple grammar to define a distribution over kernels that allow different properties of functions (such as being linear, or periodic) to be combined together.

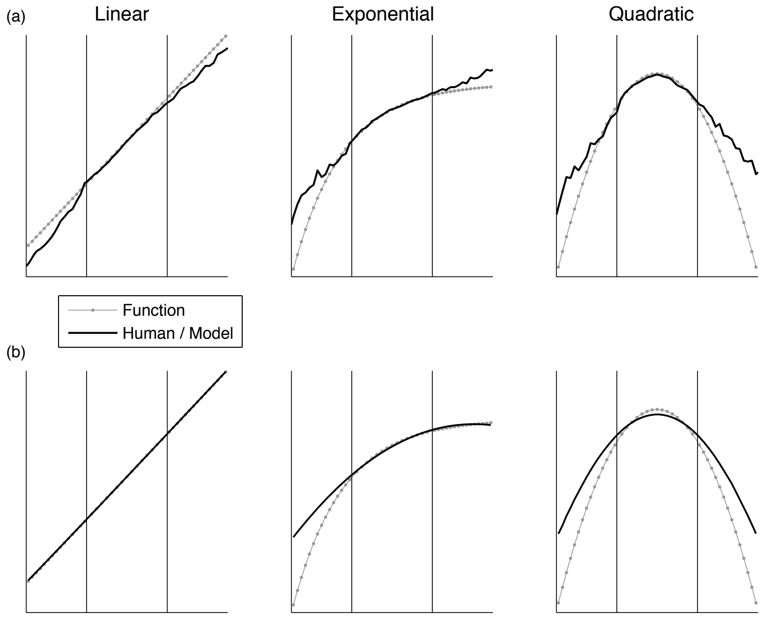


Figure 9.5: Extrapolation performance. (a)-(b) Mean predictions on linear, exponential, and quadratic functions for (a) human participants (from Delosh et al., 1997) and (b) a Gaussian process model with Linear, Quadratic, and Nonlinear kernels. Training data were presented in the region between the vertical lines, and extrapolation performance was evaluated outside this region. Reproduced with permission from Griffiths et al. (2008).

9.4 Future directions

While nonparametric Bayesian methods have been used to study a variety of topics in cognitive science, most of these applications have used the small family of tools presented in this chapter – the Chinese restaurant process, the Indian buffet process, and Gaussian processes. The literature on nonparametric Bayesian statistics covers a much wider range of topics and continues to expand, creating other opportunities for cognitive science. For example, methods similar to those used to define the Chinese restaurant and Indian buffet processes can be used to define probability distributions on infinite ranked sequences (Caron & Teh, 2012) and graphs (Caron, 2012).

While our focus in this chapter (and in the book more broadly) has been on Monte Carlo methods, variational inference can also be used for inference in nonparametric Bayesian models (e.g., Blei & Jordan, 2006). While Monte Carlo methods emphasize the discrete structure of the Chinese restaurant and Indian buffet processes, variational methods turn inference into a continuous optimization problem. As a consequence, these methods potentially have a different interpretation in terms of the underlying cognitive processes, and have the potential for establishing stronger links to methods based on artificial neural networks.

In general, integrating nonparametric Bayesian models with deep learning potentially offers a new way to think about the trade-off between structure and flexibility that is intrinsic to human cognition. For example, artificial neural networks are known to suffer from **catastrophic forgetting**, where training on one task replaces the knowledge acquired when performing a previous task (McCloskey & Cohen, 1989). The discrete structure offered by the Chinese restaurant process is potentially a way to prevent this: if the system is capable of recognizing that a task is different from what it was previously doing, it can lead to perform that task without modifying the representation that had previous tasks (Jerfel, Grant, Griffiths, & Heller, 2019). Likewise, the Indian buffet process has been used to define a structured prior to support continual learning in neural networks (Kessler et al., 2021). The integration of a capacity to recognize discrete distinctions in the environment with continuous learning suggests a path towards systems that appropriately balance structure and flexibility.

9.5 Conclusion

Human minds have to grapple with a world that contains unknown numbers of clusters, features, and causes, and unknown forms of relationships between variables. Nonparametric Bayesian models provide a way to define meaningful prior distributions for such a world, allowing us to model how people assimilate information into existing representations and modify those representations to accomodate inconsistent results. This capacity can be used as a component of more complex Bayesian models, with the prior distributions discussed in this chapter being useful in any situation where there is uncertainty over the dimensionality or complexity of latent variables.

We opened the chapter with the example of an explorer encountering a new kind of animal – a case that is readily addressed by the models we have described. But being able to postulate that something is of a kind we haven't seen before isn't the sole province of explorers. It's a problem faced by scientists who push the boundaries of their knowledge, and by every human child. Piaget highlighted assimilation and accommodation as essential forces of cognitive development because so much of our early experience requires expanding what we know in different ways. Nonparametric Bayesian models give us a way to understand these forces – a precise account of when to assimilate and when to accommodate. Using these models, we can capture a part of what it means to grow up in a world of boundless complexity.

References

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

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

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

Austerweil, J. L., & Griffiths, T. L. (2011). A rational model of the effects of distributional information on feature learning. *Cognitive Psychology*, *63*(4), 173–209.

Austerweil, J. L., & Griffiths, T. L. (2013). A nonparametric Bayesian framework for constructing flexible feature representations. *Psychological Review*, *120*(4), 817–851.

Austerweil, J. L., Sanborn, S., & Griffiths, T. L. (2019). Learning how to generalize. *Cognitive Science*, *43*(8), e12777.

Badham, S. P., Sanborn, A. N., & Maylor, E. A. (2017). Deficits in category learning in older adults: Rule-based versus clustering accounts. *Psychology and Aging*, *32*(5), 473-488.

Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian theory.* Wiley.

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

Blei, D. M., & Jordan, M. I. (2006). Variational inference for Dirichlet process mixtures. *Bayesian analysis*, *1*(1), 121–143.

Box, G. E. P., & Tiao, G. C. (1992). *Bayesian inference in statistical analysis.* Wiley.

Brehmer, B. (1974). Hypotheses about relations between scaled variables in the learning of probabilistic inference tasks. *Organizational Behavior and Human Decision Processes*, *11*, 1-27.

Busemeyer, J. R., Byun, E., DeLosh, E. L., & McDaniel, M. A. (1997). Learning functional relations based on experience with input-output pairs by humans and artificial neural networks. In K. Lamberts & D. Shanks (Eds.), *Concepts and categories* (p. 405-437). MIT Press.

Bush, C. A., & MacEachern, S. N. (1996). A semi-parametric Bayesian model for randomized block designs. *Biometrika*, *83*, 275-286.

Caron, F. (2012). Bayesian nonparametric models for bipartite graphs. *Advances in Neural Information Processing Systems*, *25*.

Caron, F., & Teh, Y. (2012). Bayesian nonparametric models for ranked data. *Advances in Neural Information Processing Systems*, *25*.

Carroll, J. D. (1963). *Functional learning: The learning of continuous functional mappings relating stimulus and response continua.* Educational Testing Service.

Dahl, D. B. (2003). *An improved merge-split sampler for conjugate Dirichlet process mixture models* (Tech. Rep. No. 1086). Department of Statistics, University of Wisconsin.

Dasgupta, I., & Griffiths, T. L. (2022). Clustering and the efficient use of cognitive resources. *Journal of Mathematical Psychology*, *109*, 102675.

Davis, T., Love, B. C., & Maddox, W. T. (2012). Age-related declines in the fidelity of newly acquired category representations. *Learning & Memory*, *19*(8), 325–329.

DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non of abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 968-986.

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

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

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

Gershman, S. J., Blei, D. M., & Niv, Y. (2010). Context, learning, and extinction. *Psychological Review*, *117*(1), 197.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice.* Chapman & Hall.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2006a). Contextual dependencies in unsupervised word segmentation. In *Proceedings of COLING/ACL 2006.*

Goldwater, S., Griffiths, T. L., & Johnson, M. (2006b). Interpolating between types and tokens by estimating power-law generators. *Advances in Neural Information Processing Systems 18*, 459-466.

Goodman, N. D., Mansinghka, V. K., Roy, D. M., Bonawitz, K., & Tenenbaum, J. B. (2008). Church: a language for generative models. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence.*

Green, P., & Richardson, S. (2001). Modelling heterogeneity with and without the Dirichlet process. *Scandinavian Journal of Statistics*, *28*, 355-377.

Griffiths, T., Lucas, C., Williams, J., & Kalish, M. (2008). Modeling human function learning with gaussian processes. In *Advances in Neural Information Processing Systems 21.*

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 Meeting of the Cognitive Science Society.*

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

Hjort, N. L. (1990). Nonparametric Bayes estimators based on Beta processes in models for life history data. *Annals of Statistics*, *18*, 1259-1294.

Hjort, N. L., Holmes, C., Müller, P., & Walker, S. G. (2010). *Bayesian nonparametrics.* Cambridge University Press.

Ishwaran, H., & James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, *96*, 1316-1332.

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

Jerfel, G., Grant, E., Griffiths, T. L., & Heller, K. A. (2019). Reconciling meta-learning and continual learning with online mixtures of tasks. In *Advances in Neural Information Processing systems 32.*

Johnson, E. J., Häubl, G., & Keinan, A. (2007). Aspects of endowment: a query theory of value construction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(3), 461–474.

Kalish, M., Lewandowsky, S., & Kruschke, J. (2004). Population of linear experts: Knowledge partitioning and function learning. *Psychological Review*, *111*, 1072-1099.

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T., & Ueda, N. (2006). Learning systems of concepts with an infinite relational model. In *Proceedings of the 21st National Conference on Artificial Intelligence.*

Kessler, S., Nguyen, V., Zohren, S., & Roberts, S. J. (2021). Hierarchical Indian buffet neural networks for Bayesian continual learning. In *Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence* (pp. 749–759).

Koh, K., & Meyer, D. E. (1991). Function learning: Induction of continuous stimulus-response relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 811-836.

Lucas, C. G., Griffiths, T. L., Williams, J. J., & Kalish, M. L. (2015). A rational model of function learning. *Psychonomic Bulletin & Review*, *22*(5), 1193–1215.

Mansinghka, V., Shafto, P., Jonas, E., Petschulat, C., Gasner, M., & Tenenbaum, J. B. (2016). CrossCat: a fully Bayesian nonparametric method for analyzing heterogeneous, high dimensional data. *Journal of Machine Learning Research*, *17*(1), 4760–4808.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation* (Vol. 24, pp. 109–165).

McDaniel, M. A., & Busemeyer, J. R. (2005). The conceptual basis of function learning and extrapolation: Comparison of rule-based and associative-based models. *Psychonomic Bulletin and Review*, *12*, 24-42.

Muller, P., & Quintana, F. A. (2004). Nonparametric Bayesian data analysis. *Statistical Science*, *19*, 95-110.

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

Piaget, J. (1954). *The construction of reality in the child.* Basic Books.

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

Rabi, R., & Minda, J. P. (2016). Category learning in older adulthood: A study of the Shepard, Hovland, and Jenkins (1961) tasks. *Psychology and Aging*, *31*(2), 185–197.

Rasmussen, C. E. (2000). The infinite Gaussian mixture model. In *Advances in Neural Information Processing Systems 12.*

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.*

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*(4), 1144–1167.

Sanborn, A. N., Heller, K., Austerweil, J. L., & Chater, N. (2021). Refresh: A new approach to modeling dimensional biases in perceptual similarity and categorization. *Psychological Review*, *128*(6), 1145–1186.

Schulz, E., Tenenbaum, J. B., Duvenaud, D., Speekenbrink, M., & Gershman, S. J. (2017). Compositional inductive biases in function learning. *Cognitive Psychology*, *99*, 44–79.

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

Shafto, P., Kemp, C., Mansinghka, V., Gordon, M., & Tenenbaum, J. B. (2006). Learning cross-cutting systems of categories. In *Proceedings of the 28th Annual Meeting of the Cognitive Science Society* (p. 2151-2156).

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

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs*, *75.* (13, Whole No. 517)

Shultz, T. R., Mareschal, D., & Schmidt, W. C. (1994). Modeling cognitive development on balance scale phenomena. *Machine learning*, *16*(1), 57–86.

Teh, Y. W. (2006). A hierarchical Bayesian language model based on Pitman-Yor processes. In *Proceedings of COLING/ACL* (pp. 985–992).

West, M., Muller, P., & Escobar, M. (1994). Hierarchical priors and mixture models, with application in regression and density estimation. In P. Freeman & A. Smith (Eds.), *Aspects of uncertainty* (p. 363-386). Wiley.

Wilson, A. G., Dann, C., Lucas, C., & Xing, E. P. (2015). The human kernel. In *Advances in Neural Information Processing Systems 28.*

Zipf, G. (1932). *Selective studies and the principle of relative frequency in language.* Harvard University Press.